UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS E TECNOLOGIA
CURSO DE MESTRADO EM ENGENHARIA CIVIL

MÉTODOS ESTATÍSTICOS MULTIVARIADOS APLICADOS À ANÁLISE DE TRANSPORTES

por

WALTER SANTA CRUZ

CAMPINA GRANDE - PB
MAIO DE 1983

MÉTODOS ESTATÍSTICOS MULTIVARIADOS APLICADOS À ANÁLISE DE TRANSPORTES

DIGITALIZAÇÃO:
UFCG
SUPRA OMNES LUX LUCIS
SISTEMOTECA UFCG

C957m Cruz, Walter Santa.
Métodos estatísticos multivariados aplicados à análise de transportes / Walter Santa Cruz. - Campina Grande, 1983.
92 f.

Dissertação (Mestrado em Ciências) - Universidade Federal da Paraíba, Centro de Ciências e Tecnologia, 1983.
"Orientação : Prof. Dr. José Eugênio Leal".
Referências.

1. Transportes - Planejamento e Organização. 2. Transportes - Métodos Estatísticos. 3. Técnicas Estatísticas Multivariadas. 4. Dissertação - Ciências. I. Leal, José Eugênio. II. Universidade Federal da Paraíba - Campina Grande (PB). III. Título

CDU 656.07:519.2(043)

MÉTODOS ESTATÍSTICOS MULTIVARIADOS APLICADOS À ANÁLISE DE TRANSPORTES

WALTER SANTA CRUZ
ENGENHEIRO CIVIL

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO E PESQUISA DO CENTRO DE CIÊNCIAS E TECNOLOGIA DA UNIVERSIDADE FEDERAL DA PARAIBA COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS (M.Sc.).

Aprovado por:

COMISSÃO

JOSÉ EUGÊNIO LEAL - Dr.Ing.
-Presidente-

ÉLIO SANTANA FONTES - M.Sc.
-Examinador Interno-

MOACIR GUILHERMINO DA SILVA - M.Sc.
-Examinador Externo-

CAMPINA GRANDE
ESTADO DA PARAÍBA-BRASIL
MAIO - 1983

À meus pais, irmãos, esposa e filhas.

AGRADECIMENTOS

Ao orientador, Professor Dr. José Eugênio Leal pelo incentivo, assistência e orientação prestadas na elaboração deste trabalho.

Ao Professor Sebastião Guimarães Vieira, Pró-Reitor para Assuntos do Interior, pelo apoio e confiança depositada no autor.

Ao Professor Jean Pierre Demartinecourt pela valiosa ajuda na parte matemática deste trabalho.

Ao Professor Dr. Edmilson de Vasconcelos Pontes do CCEN/UFAL, pela orientação acadêmica prestada, durante a graduação do autor.

À equipe do Núcleo de Processamento de Dados do Centro de Ciências e Tecnologia - Campina Grande, pela ajuda prestada nos trabalhos de computação.

A todos os outros que, direta ou indiretamente, contribuiram para a realização deste trabalho.

MÉTODOS ESTATÍSTICOS MULTIVARIADOS APLICADOS À ANÁLISE DE TRANSPORTES

DISSERTAÇÃO DE MESTRADO

POR

WALTER SANTA CRUZ

R E S U M O

Este trabalho tem como objetivo verificar a aplicabilidade das técnicas estatísticas multivariadas, Análise de Regressão Linear Múltipla, Análise Fatorial e Análise Discriminante em estudos de análise de transportes.

São mostrados os enfoques matemáticos para cada uma dessas técnicas, e realizadas aplicações e discussões dos resultados em diversos problemas da análise de transportes.

Multivariate Statistical Methods Applied to Transportation Analysis

M.Sc. Dissertation

BY

Walter Santa Cruz

A B S T R A C T

The objective of this work is to verify the applicability of the multivariate statistical techniques as Multiple Linear Regression Analysis, Factor Analysis and Discriminant Analysis in studies of transportation analysis.

It has been showed the mathematical approaches for each of these techniques as well as applications and discussions of the results for several problems of transportation analysis.

ÍNDICE

LISTA DE TABELAS

LISTA DE QUADROS

CAPÍTULO I

INTRODUÇÃO

1.1 - Considerações Gerais

Na década de 50, o déficit da ferrovia brasileira e a criação da indústria automobilística, no Brasil, com perspectivas de desenvolvimento a curto prazo, contribuiram para o país relegar a segundo plano o sistema ferroviário e voltar a atenção para a construção de uma infra-estrutura viária capaz de absorver o tráfego de veículos automotores e permitir condições de fluxos, aceitáveis pelos usuários.

Com isso, todos os esforços se concentraram nas capacidades das rodovias e o planejamento dos transportes tinha, como um dos objetivos fundamentais, a determinação da demanda atual de tráfego, nas rodovias mais importantes, e o cálculo de coeficientes que permitissem estimar o tráfego futuro. O que se procurava, na realidade, era identificar o volume de tráfego por automóvel e, assim, toda estimativa futura era para prover as rodovias de condições adequadas ao uso do transporte individual.

1.2 - Apresentação do Problema

No Brasil, as cidades de médio e pequeno portes, ge-

ralmente, carecem de um planejamento sistêmico dos transportes. Ora os estudos realizados não são implantados devido a limitações de recursos, dos mais variados tipos, ora a deficiência na obtenção dos dados ou o uso de modelos inadequados produz resultados imprecisos.

No que concerne a parte de modelagem, tem-se, na fase de geração de viagens, os modelos à base do cálculo da correlação linear como os mais usados para explicar as viagens geradas pelas zonas de tráfego. Considera-se determinadas variáveis, em geral, agregadas a nível de zona, como explicativas das ocorrências de viagens zonais e busca-se, através do método dos mínimos quadrados, uma equação de regressão linear múltipla que associe as viagens geradas às variáveis do problema.

Encontrada a equação de regressão linear múltipla, o passo seguinte é decidir se essa equação se presta para estimar a demanda futura de viagens. Essa decisão deve ser tomada com base na análise do coeficiente de determinação múltipla, no erro padrão de estimativa e na significância estatística dos coeficientes de regressão parcial. É comum, no entanto, decidir-se sobre a qualidade de uma equação de regressão linear múltipla baseando-se, apenas, no valor do coeficiente de determinação múltipla. Isso deve ser evitado, pois, pode conduzir a erros.

Kassoff & Deutschman [9] estudam a alteração sofrida por uma equação de regressão linear proveniente de dois aspectos das variáveis observadas: o primeiro, utiliza os valores totais das variáveis, para o nível de agregação considerado; o

segundo, utiliza os valores relativos das variáveis, isto é, os valores da razão entre os totais agregados de cada variável e o total agregado de uma variável específica. Eles mostram que as equações com as variáveis dadas por seus valores totais agregados são, apesar de possuirem o coeficiente de determinação múltipla maior do que o das equações com as variáveis dadas por seus valores relativos, qualitativamente, inferiores a estes últimos tipos de equações. Isso comprova que se pode incorrer em erro quando se julga uma equação de regressão linear múltipla com base, somente, no seu coeficiente de determinação.

Se a equação de regressão linear obtida possui um coeficiente de determinação múltipla próximo da unidade, um pequeno erro padrão de estimativa e os coeficientes de regressão parcial, estatisticamente significantes a um nível de significância estabelecido, adota-se essa equação como equação de previsão de viagens.

De posse da equação para previsão de viagens, as viagens no ano meta são calculadas, simplesmente, atribuindo-se às variáveis independentes os seus respectivos valores futuros. Em assim procedendo, admite-se a hipótese de que os coeficientes da equação são imutáveis, com o tempo, o que não parece razoável. Em outras palavras, significaria dizer que, dispondo dos valores reais das variáveis dependente e independentes, no ano meta, e calculando-se uma nova equação de regressão linear múltipla, essa equação não diferiria, substancialmente, da obtida no ano base. A inalteração dos coeficientes de regressão parcial, para o ano meta, mostra a

insensibilidade, às mudanças que possam ocorrer em uma área de estudo, dos modelos de regressão linear para previsão de viagens.

Por outro lado, reportando-se novamente ao trabalho de Kassoff & Deutschman [9], eles alertam para um outro tipo de problema: o da escolha das variáveis.

Não basta que as variáveis envolvidas numa equação de regressão linear múltipla satisfaçam as premissas básicas da técnica da regressão linear. É necessário que haja uma relação de causa-efeito entre cada variável independente e a variável dependente. Torna-se evidente, pois, que a escolha das variáveis, para explicar um dado fenômeno, é fundamental.

Não somente a escolha das variáveis, mas também, o nível de agregação em que elas se encontram, isto é, se os valores das variáveis são obtidos a nível de zona, domicílio, etc., afeta a qualidade de uma equação de regressão linear múltipla.

// McCarthy [7] mostra que, à medida que o nível de agregação das variáveis diminui, a qualidade de uma equação de regressão linear aumenta, porque a parcela correspondente à variância interna de cada variável vai se tornando menor e, consequentemente, maior é a contribuição da variância, entre as variáveis, para a variância total dos dados. Sabe-se que a variância total de um conjunto é igual a variância interna de cada elemento somada com a variância existente entre os elementos. //

Portanto, pode-se questionar o uso da Análise de Regressão Linear Múltipla em estudos de demanda de transportes porque as variáveis observadas são, geralmente, agregadas a

nível de zona de tráfego e com os seus valores representando os totais agregados àquele nível de agregação.

Em um processo de planejamento dos transportes, composto das fases de geração, distribuição, divisão modal e alocação de viagens, em que a primeira fase fornece os modelos de regressão linear para a previsão de viagens, é de se esperar que os resultados obtidos careçam de precisão. A geração de viagens fornece os dados de entrada para os modelos, normalmente, usados na fase de distribuição, e esta, por sua vez, tem seus resultados como elementos de entrada nos modelos das fases posteriores. Assim, as falhas na geração de viagens se ampliam nas fases subseqüentes.

Então, necessário se faz o estudo e aplicação de técnicas estatísticas mais refinadas a estudos de análise de transportes.

Além da Análise de Regressão Linear Múltipla, utiliza se, neste trabalho, as técnicas multivariadas Análise Fatorial e Análise Discriminante. Estas duas últimas técnicas, quando comparadas com a primeira, apresentam vantagens, a saber:

- Abordam a complexidade do processo de viagem.
- Incorporam a variância produzida, tanto por fatores sociais (internos à zona), como por fatores locacionais (entre as zonas).
- Representam a viagem de um indivíduo não em termos de uma variável isolada (análise univariada), mas de uma série de variáveis (análise multivariada) representativas do comportamento do indivíduo.

- Permitem uma análise mais aprofundada do fenômeno dos transportes, em vista de analisar os fatores-causas da diferença de comportamento do indivíduo, em relação aos transportes.

1.3 - Objetivos

Os objetivos desse trabalho são:

1 - Apresentar as técnicas estatísticas multivariadas, Análise de Regressão Linear Múltipla, Análise Fatorial e Análise Discriminante, com os seus respectivos desenvolvimentos matemáticos. Essa abordagem matemática objetiva dar uma visão dos métodos utilizados e permite uma melhor compreensão de cada método. Não se pretende, neste trabalho, esgotar a formulação matemática dos métodos tratados. O tratamento matemático formal pode ser visto na literatura especializada, citada na bibliografia.

2 - Propiciar uma visão crítica do uso da Análise de Regressão Linear Múltipla em estudos de demanda de viagens.

3 - Aplicar as técnicas estatísticas multivariadas, Análise Fatorial e Análise Discriminante, a estudos de análise de transportes. Esse constitui o principal objetivo desse trabalho.

CAPÍTULO II

AS TÉCNICAS ESTATISTÍCAS MULTIVARIADAS: ANÁLISE DE REGRESSÃO LINEAR MÚLTIPLA, ANÁLISE FATORIAL E ANÁLISE DISCRIMINANTE

2.1 - Introdução

Análise Multivariada compreende um conjunto de procedimentos estatísticos, dentre estes a Análise de Regressão Linear Múltipla, Análise Fatorial e Análise Discriminante, que consistem em analisar propriedades distintas, observadas simultâneamente, acerca de cada elemento de uma amostra ou de uma população.

Este capítulo trata do desenvolvimento matemático das técnicas estatísticas multivariadas, Análise de Regressão Linear Múltipla, Análise Fatorial e Análise Discriminante.

Na seção referente às considerações iniciais sobre cada técnica, dá-se uma visão geral da técnica apresentada. Em seguida, apresenta-se o enfoque matemático e os testes estatísticos, quando houver, para cada técnica.

2.2 - Análise de Regressão Linear Múltipla

2.2.1 - Considerações iniciais sobre a Análise de Regressão Linear Múltipla.

Os estudos de demanda para o planejamento dos transportes são, usualmente, baseados em modelos compostos de quatro etapas principais, a saber: geração, distribuição, divisão modal e alocação de viagens.

Tem-se, na fase de geração de viagens, os modelos à base do cálculo da correlação linear como os mais usados para explicar as viagens geradas pelas zonas de tráfego. Considera-se determinadas variáveis, em geral, agregadas a nível de zona, como explicativas das ocorrências de viagens zonais e busca-se, através do método dos mínimos quadrados, uma equação de regressão linear múltipla que associe as viagens geradas às variáveis do problema.

Assim, o que se obtem é uma tentativa de descrição das viagens atuais por uma expressão do tipo $Y=a+b_1X_1+b_2X_2+\ldots+b_pX_p$ onde, Y é a variável dependente que representa o número de viagens atuais, a é a parcela de Y não explicada pela equação de regressão, b_i é o i-ésimo coeficiente de regressão parcial e X_i é a i-ésima variável explicatica ou independente, $i = 1,2,\ldots,p$.

Se $Y = a + b_1X_1 + b_2X_2 +\ldots+ b_pX_p$ possui um coeficiente de determinação, R^2, próximo da unidade, um pequeno erro padrão de estimativa, S_e, e os coeficientes de regressão parcial, b_i, estatisticamente significantes, a um nível de significância estabelecido, adota-se essa equação como equação de previsão de viagens e, dessa forma, as viagens no ano meta são calculadas, pela equação de regressão, simplesmente, atribuindo-se às variáveis independentes os seus respectivos valores futuros.

Assim, os coeficientes de regressão parcial são supostamente imutáveis, com o tempo, o que equivale a se supor não haver modificações no comportamento espacial da população, no ano meta, que possa alterar os coeficientes de regressão.

Análise de Regressão Linear Múltipla, também, tem aplicações em outras fases do planejamento dos transportes, como na divisão modal, na calibração do modelo gravitacional, etc.

2.2.2 - Enfoque Matemático da Análise de Regressão Linear Múltipla.

Análise de regressão linear múltipla é uma técnica estatística cujo principal objetivo é analisar a relação entre uma variável e um conjunto de variáveis. A primeira é chamada variável dependente e as últimas, independentes.

Uma equação de regressão linear múltipla é uma expressão da forma:

$$\hat{X}_{kn} = b_o + b_1 X_{k1} + b_2 X_{k2} + \ldots + b_{n-1} X_{k\,n-1} ,$$

$$k = 1,2 \ldots, m$$

onde: m = número de observações de cada variável

n = número total de variáveis

$\hat{X}_{kn}$ = valor estimado de cada observação da variável dependente X_n.

X_{ki} = k-ésima observação da variável independente X_i, $i = 1,2 \ldots, n-1$.

b_o = constante não explicada pelas variáveis inde

pendentes X_i.

b_i = coeficiente de regressão parcial da variável independente X_i.

Utilizando-se o método dos mínimos quadrados, procura-se encontrar o conjunto de valores $b_1 \ldots\ldots, b_{n-1}$ tal que a soma dos quadrados dos desvios de X_n em relação ao valor estimado de X_n ($\hat{X}_n$) seja mínima. Em outras palavras, quer-se achar o Min $\Sigma(X_{kn} - \hat{X}_{kn})^2$.

Seja $S = \sum_{k=1}^{m} (X_{kn} - \hat{X}_{kn})^2$

O valor de S será mínimo quando as derivadas parciais de S em relação a $b_1, \ldots, b_{n-1}$ forem nulas. Essas expressões derivadas e arrumadas, convenientemente, fornecem o seguinte sistema de equações.

$$\begin{cases} b_1\ SQ_1 + b_2\ SP_{12} + \ldots + b_{n-1}\ SP_{1\,n-1} = SP_{1\,n} \\ b_1\ SP_{21} + b_2\ SQ_2 + \ldots + b_{n-1}\ SP_{2\,n-1} = SP_{2\,n} \\ \text{----------} \\ \text{----------} \\ b_1\ SP_{n-1\,1} + b_2\ SP_{n-1\,2} + \ldots + b_{n-1}\ SQ_{n-1} = SP_{n-1\,n} \end{cases} \quad (1)$$

onde: $SQ_i = \sum_{k=1}^{m} (X_{ki} - \bar{X}_i)^2$ = soma dos quadrados dos desvios de cada variável independente X_i em relação à sua média $\bar{X}_i$.

$SP_{ij} = \sum_{k=1}^{m} (X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j)$ = soma dos produtos cruzados entre as variáveis X_i e X_j.

A resolução do sistema de equações (1) conduz aos va-

lores de $b_1, b_2, \ldots, b_{n-1}$.

O valor de b_o é dado por:

$$b_o = \bar{X}_n - b_1\bar{X}_1 - \ldots - b_{n-1}\bar{X}_{n-1}$$

sendo $\bar{X}_i = \frac{\sum_{k=1}^{m} X_{ki}}{m}$, i = 1, 2, 3,..., n .

O sistema (1) em termos matriciais, torna-se:

$$\begin{pmatrix} SQ_1 & SP_{12} & \cdots & SP_{1\,n-1} \\ SP_{21} & SQ_2 & \cdots & SP_{2\,n-1} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ SP_{n-1\,1} & SP_{n-1\,2} & \cdots & SQ_{n-1} \end{pmatrix} \cdot \begin{pmatrix} b_1 \\ b_2 \\ - \\ - \\ b_n \end{pmatrix} = \begin{pmatrix} SP_{1\,n} \\ SP_{2\,n} \\ - \\ - \\ SP_{n-1\,n} \end{pmatrix}$$

$$SQ.\ b = SP_n$$

$$b = SQ^{-1}.\ SP_n$$

onde: b = vetor coluna (n-1 x 1) dos coeficiente parciais de regressão

SQ = matriz (n-1 x n-1) da soma dos quadrados e produtos cruzados

SP_n = vetor coluna (n-1 x 1) da soma dos produtos cruzados entre cada variável independente e a variável dependente.

Transformando-se as variáveis X_i nos seus valores padronizados, Z_i, i=1, 2,... n, a equação de regressão linear múltipla torna-se:

$$\hat{Z}_{kn} = B_1 Z_{k1} + B_2 Z_{k2} + \ldots + B_{n-1} Z_{k\,n-1}$$

sendo:

$$Z_{ki} = \frac{X_{ki} - \bar{X}_i}{\sigma_{X_i}} \quad , \; i = 1, 2, \ldots, n$$

σ_{X_i} = desvio padrão da variável X_i.

A função a ser minimizada passa a ser:

$$S_1 = \sum_{k=1}^{m} (Z_{kn} - \hat{Z}_{kn})^2$$

Derivando-se parcialmente S_1 em relação a B_1, B_2,..., B_{n-1} e igualando essas derivadas a zero, obtem-se um sistema de (n-1) equações normais com (n-1) incógnitas. Estas equações normais são da forma

$$\begin{cases} B_1 + r_{12} B_2 + r_{13} B_3 + \ldots + r_{1\,n-1} B_{n-1} = r_{1n} \\ r_{21} B_1 + B_2 + r_{23} B_3 + \ldots + r_{2\,n-1} B_{n-1} = r_{1n} \\ \text{---------- ---} \\ r_{n-1\,1} B_1 + r_{n-1\,2} B_2 + r_{n-1\,3} B_3 + \ldots + B_{n-1} = r_{n-1\,n} \end{cases} \qquad (2)$$

onde: r_{ij} = coeficiente de correlação entre as variáveis Z_i e Z_j

B_i = coeficiente de regressão padronizado da variável Z_i, i = 1,2,..., n-1

Seja R =

$$\left[\begin{array}{ccc|c} r_{11} & r_{12} & \cdots\; r_{1\,n-1} & r_{1n} \\ r_{21} & r_{22} & \cdots\; r_{2\,n-1} & r_{2n} \\ \multicolumn{3}{c|}{\text{- - - - - - - - - - - - - - - - -}} & \text{- - - -} \\ r_{n-1\,1} & r_{n-1\,2} & \cdots\; r_{n-1\,n-1} & r_{n-1\,n} \\ \hline r_{n1} & r_{n2} & \cdots\; r_{n\,n-1} & r_{nn} \end{array}\right]$$

dividida em blocos

Sejam $R_{11} = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n-1} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ r_{n-11} & r_{n-12} & \cdots & r_{n-1\,n-1} \end{pmatrix}$, $R_{12} = \begin{pmatrix} r_{1n} \\ r_{2n} \\ \cdot \\ \cdot \\ \cdot \\ r_{n-1n} \end{pmatrix}$

$R_{21} = (r_{n1}, r_{n2}, \ldots, r_{n\,n-1})$ e $R_{22} = r_{nn}$. Então

$R = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix}$ e o sistema de equações (2)

pode ser escrito como

$$R_{11} \cdot B = R_{12}$$

$$B = R_{11}^{-1} \cdot R_{12}$$

onde: B = vetor coluna (n-1 x 1) dos coeficientes de regressão padronizados

R_{11}^{-1} = inversa da matriz de correlação entre as variáveis independentes.

R_{12} = vetor coluna (n-1x1) dos coeficientes de correlação entre cada variável independente e a variável dependente.

Pode se obter os coeficientes de regressão não padronizados $b_1, b_2, \ldots, b_{n-1}$ a partir dos coeficientes de regressão padronizados, calculando-se

$$b = AB$$

onde:

b = vetor coluna (n-1 x 1) dos coeficientes de regressão não padronizados

A = matriz diagonal em que

$a_{ii} = \frac{\sigma_{X_n}}{\sigma_{X_i}}$, i = 1, 2,..., n-1, σ_{X_n} é o desvio padrão da variável dependente X_n e σ_{X_i} é o desvio padrão da variável independente X_i. As variáveis X_j, j=1,... n não estão na forma padronizada.

B = vetor coluna (n-1 x 1) dos coeficientes de regressão padronizados.

O valor de b_o é dado por:

$$b_o = \bar{X}_n - (b^t \cdot M)$$

onde:

$\bar{X}_n$ = média aritmética da variável dependente

b^t = vetor transposto de b

M = vetor coluna (n-1 x 1) das médias das variáveis independentes.

2.2.3 - Interpretação dos coeficientes de regressão e da parte constante de uma equação de regressão linear múltipla

Os coeficientes de regressão parcial b_i, associados a cada variável independente X_i, i = 1, 2,..., n-1, representam a variação ocorrida no valor observado da variável dependente X_n quando a variável X_i varia de uma unidade, des

de que as demais variáveis independentes permaneçam constantes.

Para a obtenção dos coeficientes de regressão padronizados, transforma-se as variáveis X_j nas variáveis Z_j, j=1, 2,..., n, com média zero e desvio padrão igual a 1. Isso elimina o efeito das dimensões diferenciadas entre as variáveis X_j e propicia uma análise mais realista da contribuição de cada variável independente para o valor estimado da variável dependente.

Nesse caso, o coeficiente de regressão na forma padronizada B_i indica a variação sofrida por X_n, quando ocorre uma variação de um desvio padrão na variável independente X_i, mantendo-se constante as outras variáveis independentes. Isso possibilita a identificação daquelas variáveis independentes que mais contribuem para a variação da variável dependente.

A parte constante de uma equação de regressão linear múltipla reflete a incapacidade das variáveis independentes para explicar a variável dependente.

Para se ter idéia da influência da magnitude da constante na qualidade de uma equação de regressão linear, geralmente, compara-se o seu valor com a média da variável dependente. Em termos práticos, pode-se considerar que a constante tem um valor razoável quando a porcentagem que ela representa da média da variável dependente for inferior a 20%.

2.2.4 - Análise da qualidade de uma equação de regressão linear múltipla.

Obtida a equação de regressão linear múltipla, $\hat{X}_{kn} = b_o + b_1 X_{k1} + \ldots + b_{n-1} X_{kn-1}$, a etapa seguinte é analisar a sua validade, como modelo para previsão de viagens.

Uma equação de regressão linear múltipla pode ser avaliada através dos seguintes critérios:

- Análise numérica da equação;
- Análise qualitativa da equação;
- Testes estatísticos.

2.2.4.1 - Análise numérica da equação

A análise numérica de uma equação de regressão linear múltipla é feita, normalmente, com base na análise de dois parâmetros: o erro padrão de estimativa e o coeficiente de determinação.

O erro padrão de estimativa S_e e o coeficiente de determinação R^2, podem ser calculados, respectivamente, por:

$$Se = \sqrt{\frac{\sum_{i=1}^{N} (X_{in} - \hat{X}_{in})^2}{N - K - 1}}$$

$$R^2 = \frac{\sum_{i=1}^{N} (\hat{X}_{in} - \bar{X}_n)^2}{\sum_{i=1}^{N} (X_{in} - \bar{X}_n)^2}$$

onde:

X_{in} = i-ésima observação da variável dependente X_n

$\hat{X}_{in}$ = i-ésima estimativa da variável dependente X_n

$\bar{X}_n$ = média aritmética da variável dependente X_n

N = número de observações

K = número de variáveis independentes

Em termos matriciais R^2 é dado por:

$R^2 = B^t . R_{12}$ sendo B^t o vetor transposto de B.

Quanto mais próximas forem as observações da variável dependente dos seus valores estimados, menores serão os desvios ou resíduos entre esses valores e, consequentemente, menor será o valor de S_e e mais próximo de 1 é o valor de R^2.

O valor de R^2, expresso em porcentagem, indica a quantidade da variância da variável dependente explicada pelas variáveis independentes.

Em uma equação de regressão, a soma dos quadrados dos desvios das observações da variável dependente X_n, em relação a sua média, $\bar{X}_n$, é igual a soma dos quadrados dos desvios dos valores estimados da variável dependente, $\hat{X}_n$, em relação a média $\bar{X}_n$, mais a soma dos quadrados dos desvios das observações da variável dependente, em relação aos valores estimados, $\hat{X}_n$.

Em termos matemáticos,

$$\sum_{i=1}^{N} (X_{in} - \bar{X}_n)^2 = \sum_{i=1}^{N} (\hat{X}_{in} - \bar{X}_n)^2 + \sum_{i=1}^{N} (X_{in} - \hat{X}_{in})^2$$

ou

$$SS_{X_n} = SS_{\hat{X}_n} + SS_{resíduo}$$

$$R^2 = \frac{SS_{\hat{X}_n}}{SS_{X_n}} = \frac{SS_{X_n} - SS_{resíduo}}{SS_{X_n}}$$, o que permite ver o intervalo de variação de R^2. Quando o resíduo é uma parcela

substancial de SS_{X_n}, $SS_{X_n} - SS_{resíduo}$ tende para zero e R^2, também. Se $SS_{resíduo}$ é, aproximadamente, igual a zero, $SS_{X_n} - SS_{resíduo} = SS_{X_n}$ e R^2 é igual a 1. Portanto,

$$0 \leq R^2 \leq 1.$$

O coeficiente de correlação múltipla $R = \sqrt{R^2}$ pode ser visualizado como o coeficiente de correlação simples r entre X_n e $\hat{X}_n$, porque $\hat{X}_n$ pode ser considerada como uma variável independente construída a partir da equação de regressão [2].

Uma boa equação de regressão deve ter o menor valor possível para S_e e um valor de R^2 próximo da unidade.

2.2.4.2 - Análise qualitativa da equação.

A qualidade de uma equação de regressão linear está, intimamente, ligada à qualidade das variáveis independentes. Se as variáveis independentes possuem uma relação lógica de causa-efeito com a variável dependente, a equação obtida passa a ter um sentido lógico definido e, dessa forma, pode-se analisar os coeficientes de regressão quanto a sua coerência, em termos de magnitude e sinal.

É de se esperar que uma equação de regressão linear construída sem uma relação linear bem definida entre cada variável independente e a dependente não produza bons resultados.

Um outro aspecto importante, é que entre as variáveis independentes não deve haver relação linear. Caso contrário, haverá o aparecimento de multicolinearidade entre as

variáveis independentes, e isso pode implicar em distorção na magnitude e no sinal dos coeficientes de regressão, afetando a interpretação lógica da equação.

Portanto, a análise qualitativa de uma equação de regressão linear deve abranger os seguintes passos:

- Análise das relações lineares entre as variáveis independentes e entre cada variável independente e a dependente.
- Análise dos coeficientes de regressão na tentativa de detectar alguma incoerência, quanto aos seus valores e sinais.
- Análise, do ponto de vista lógico, da relação de causa e efeito entre cada variável independente e a variável dependente bem como análise lógica da equação como um todo.

2.2.4.3 - Testes estatísticos

Para a verificação das qualidades estatísticas de uma equação de regressão linear múltipla deve-se analisar do ponto de vista estatístico o coeficiente de correlação multipla e os coeficientes de regressão parcial.

Para o teste do coeficiente de correlação múltipla, adota-se a hipótese nula de que o coeficiente de correlação é nulo e testa-se essa hipótese contra a hipótese alternativa de que ele não é nulo.

Computa-se, então,

$$F_{calculado} = \frac{R^2/K}{(1-R^2)/(N-K-1)}$$ e compara-se esse valor

com o $F_{crítico}$, de uma tabela de distribuição F, para K e (N-K-1) graus de liberdade, e um nível de significância α.

Se $F_{calculado} > F_{crítico}$, rejeita-se a hipótese nula e, em consequência, aceita-se a hipótese alternativa, ao nível de significância α. Esse teste para R é equivalente a se testar a nulidade de todos os coeficientes de regressão parcial. A rejeição da hipótese nula assegura, apenas, que um ou alguns coeficientes de regressão parcial, mas não necessariamente todos, são diferentes de zero. Esse fato mostra a necessidade de se testar individualmente cada coeficiente de regressão parcial. A hipótese nula, agora, é a de que cada coeficiente de regressão parcial é nulo e a alternativa é a de que cada um deles é diferente de zero.

Calcula-se F através da expressão

$$F_{calculado} = \frac{\text{incremento de } R^2 \text{ devido à entrada da variável ind. } X_i \text{ na eq./1}}{(1-R^2)/(N-K-1)}$$

Se $F_{calculado} > F_{crítico}$, para 1 e (N-K-1) graus de liberdade e um nível de significância α, rejeita-se a hipótese nula referente ao coeficiente de X_i. Nesse caso diz-se que esse coeficiente é estatisticamente significante ao nível de significância α. Em outras palavras, isto significa que a probabilidade de o coeficiente analisado ser nulo é menor do que α.

Se $F_{calculado} < F_{crítico}$ aceita-se a hipótese nula e, portanto, aquele coeficiente de regressão parcial não é estatisticamente significante ao nível de significância α.

A questão que se apresenta agora é a de considerar como válida ou não uma equação onde um ou vários coeficientes

não são estatisticamente significantes.

Um critério geral a adotar é o de rejeitar toda equação que tenha os coeficientes não significantes. Esse critério pode, no entanto, ocasionalmente, contradizer o critério da qualidade de uma equação baseado no valor de R^2 e no erro padrão de estimativa.

Pode ocorrer, por exemplo, uma situação em que uma equação de regressão linear múltipla possua um coeficiente de determinação, aproximadamente, igual a 1, um pequeno erro padrão de estimativa, e, no entanto, um dos coeficientes de regressão parcial, associado a uma variável com uma parcela de contribuição, relativamente, pequena para a variável dependente, não seja estatisticamente significante, ao nível de significância adotado. Nesse caso, deveria se construir uma nova equação excluindo aquela variável não importante e refazer toda a análise sobre a qualidade dessa nova equação.

Em suma, o critério recomendado neste trabalho para avaliar uma equação de regressão linear múltipla como modelo de previsão é que a equação tenha um menor erro padrão de estimativa possível, um coeficiente de determinação próximo de 1 e todos os coeficientes de regressão sejam estatisticamente significantes ao nível de significância adotado.

2.3 - Análise Fatorial

2.3.1 - Considerações iniciais sobre a Análise Fatorial.

Segundo Cooley [4], Análise Fatorial vem se tornando um

termo genérico para uma variedade de procedimentos desenvolvidos com a finalidade de analisar as interrelações dentro de um conjunto de variáveis.

A característica mais importante da Análise Fatorial é sua capacidade de transformar o conjunto de dados em um outro conjunto menor de fatores que, em geral, são não correlacionados, entre si, e são extraídos do conjunto de variáveis originais de tal forma que o primeiro fator explique a maior quantidade de variância existente dentro do conjunto de dados, o segundo fator explique a maior quantidade da variância remanescente, isto é, variância não explicada pelo primeiro, o terceiro fator explique parte da variância não explicada pelos dois fatores anteriores, e, assim, sucessivamente.

O método de extração dos fatores, explicado resumidamente acima, é chamado método dos componentes principais e faz parte do programa de Análise Fatorial, da bateria de programas SPSS (Statistical Package for the Social Sciences), utilizado neste trabalho.

A Análise Fatorial tem um vasto campo de aplicações. Além do seu amplo emprego em áreas concernentes à Psicologia e Sociologia, essa técnica da Análise Multivariada, também, é bastante utilizada na Geografia e nos diversos tipos de planejamento, como planejamento do uso do solo, dos transportes, regional, etc.

2.3.2 - Enfoque Matemático da Análise Fatorial

Seja X uma matriz do tipo (N x m) onde cada coluna cor

N-dimensional, na forma padronizada, no qual se quer projetar X.

Como em geral, X é de posto m, pode-se encontrar um vetor g(mxl) tal que

$$Xg = y \qquad (1)$$

Além disso,

$$\frac{y'y}{N} = g' \frac{X'X}{N} g = g'Rg \qquad (2)$$

onde:

$\frac{X'X}{N}$ = R = matriz de correlação entre os m vetores coluna de X do tipo (mxm) e y', g'e X' são os transpostos de y, g e X, respectivamente.

Por outro lado, $\frac{y'y}{N}$ é a variância de y e, portanto, pelo fato de y ser um vetor na forma padronizada

$$\frac{y'y}{N} = 1 \qquad (3)$$

Comparando-se as expressões (2) e (3), obtem-se:

$$g'R\,g = 1 \qquad (4)$$

A projeção de X sobre y é dada pelo produto escalar de X' pelo vetor unitário da direção y, $\frac{y}{\sqrt{N}}$. Entretanto, trabalhar-se com os vetores coluna da matriz $\frac{X}{\sqrt{N}}$, matriz em que seus vetores coluna são de comprimento unitário, equivale, simplesmente, a transformar a escala do sistema de referência, e as projeções dos vetores de $\frac{X}{\sqrt{N}}$ sobre o vetor y isto é, $\frac{X'}{\sqrt{N}} \cdot \frac{y}{\sqrt{N}}$, fornecem as projeções da variância unitária

das variáveis originais. Isso vai ao encontro da estratégia da Análise Fatorial de buscar os vetores y ortogonais, entre si, sobre cada um dos quais a projeção da variância é máxima. Como os vetores $\frac{X}{\sqrt{N}}$ e $\frac{y}{\sqrt{N}}$ são unitários a projeção se torna igual ao cosseno do ângulo formado por cada vetor coluna de X e o vetor y, e esse cosseno é o coeficiente de correlação linear entre os vetores considerados.

Seja f a projeção procurada.

$$f = \frac{X'y}{N}$$

De (1), $y = Xg$. Então:

$$f = \frac{X'Xg}{N} = Rg \qquad (5)$$

O objetivo, agora, é maximizar a soma dos quadrados das projeções f. Em termos vetoriais, a soma dos quadrados dessas projeções é dada por f'f, onde f' é o vetor transposto de f.

$$f'f = g'R'Rg = g'R^2g \qquad (6)$$

Quer-se, portanto, maximizar f'f sujeita à condição de $g'Rg = 1$.

Usando o método dos multiplicadores indeterminados de Lagrange, obtém-se:

$$Z = g'R^2g - \lambda\ (g'Rg - 1)$$

Derivando-se Z em relação a g' e igualando-se o resultado a zero:

$$\frac{\partial Z}{\partial g'} = 2\ R^2g - 2\lambda Rg = 0 \ \therefore\ R^2g = \lambda\ Rg \qquad (7)$$

Ainda,

$R^2g = R.Rg = \lambda Rg$

De (5), $Rg = f$. Tem-se, pois:

$Rf = \lambda f$ (8)

A expressão (8) mostra f como um autovetor correspondente ao autovalor λ, de R. Se R possui m autovalores reais não nulos e distintos, a matriz F, formada pelos autovetores f_i, dispostos em colunas, correspondentes aos autovalores λ_i de R, é de posto m. Generalizando a expressão (8), tem-se:

$RF = F\Lambda$ (9)

onde:

$F = (f_{ij})$ = matriz dos carregamentos dos fatores

Λ = matriz diagonal formada pelos autovalores de R, em ordem decrescente, isto é, $\lambda_{11} > \lambda_{22} > \ldots > \lambda_{mm}$.

Multiplicando-se ambos os membros de $R^2g = \lambda Rg$, à esquerda por g', vem:

$g'R^2g = g'\lambda Rg = \lambda g'Rg$

De (4) e (6), tem-se, respectivamente,

$g'Rg = 1$ e $g'R^2g = f'f$. Então,

$f'f = \lambda$ (10)

Portanto, para se obter a maior soma dos quadrados das projeções f, basta tomar o maior autovalor λ, de R.

Generalizando a expressão (10), vem:

$$F'F = \Lambda \qquad (11)$$

Substituindo a expressão (11) em (9), resulta:

$$RF = FF'F$$

Como a matriz F é inversível,

$$RFF^{-1} = FF'FF^{-1} \quad \therefore$$

$$R = FF' \qquad (12)$$

A expressão (7) pode, também, ser escrita como:

$$R^2g - \lambda Rg = 0$$

$$R.Rg - R\lambda g = 0$$

Como $R = \frac{X'X}{N}$ tem posto m, R^{-1} existe. Assim:

$$R^{-1}R.Rg - R^{-1}R\lambda g = 0$$

$$Rg = \lambda g$$

De (5), Rg = f. Então:

$$f = \lambda g \quad \therefore$$

$$g = f\lambda^{-1}$$

Generalizando o vetor g para uma matriz G, tem-se:

$$G = F\Lambda^{-1} \qquad (13)$$

A matriz Y, formada pelos vetores y é dada por:

$$Y = XG = XF\Lambda^{-1} \qquad (14)$$

$$\frac{Y'Y}{N} = \frac{G'X'XG}{N} = G'RG = \Lambda^{-1}F'RF\ \Lambda^{-1} =$$

$$= \Lambda^{-1}F'\ F\ \Lambda^{-1} = \Lambda^{-1}\Lambda\Lambda\Lambda^{-1} = I.I = I$$

Portanto, Y é ortogonal

De $R = FF'$,

$$R = f_1f'_1 + f_2f'_2 + \ldots + f_m\ f'_m$$

A parcela $f_jf'_j$ representa a contribuição do fator j para a matriz de correlação.

De (14) pode-se escrever:

$$X\ F = Y\ \Lambda$$

$$X\ F - Y\ \Lambda = 0$$

De (11), $\Lambda = F'F$. Daí,

$$XF - Y\ F'F = 0$$

Como a matriz F possui inversa:

$$X\ FF^{-1} - YF'FF^{-1} = 0$$

$$X = YF'$$

Assim, para qualquer vetor coluna da matriz X tem-se:

$$X_i = Y\ \ F_i' \qquad (15)$$

onde F_i' é o transposto do i-ésimo vetor linha F_i da matriz F.

Multiplicando ambos os membros de (15) por $\frac{X'_i}{N}$, vem:

$$\frac{X'_i\ X_i}{N} = \frac{F_i\ Y'_i\ Y_i\ F'_i}{N} \qquad (16)$$

Como X possui os seus vetores coluna na forma padronizada, $\frac{X'_i\ X_i}{N}$ representa a variância do vetor coluna X_i e, portanto, é igual a 1. Ainda, conforme mostrado anteriormente, $\frac{Y'Y}{N}$ é igual a matriz iden-

tidade I. Dessa forma:

$$1 = F_i F_i' = f_{i1}^2 + \ldots + f_{im}^2$$

Portanto, os termos f_{ik}^2, k=1,2,..., m, são os componentes da variância unitária de X_i.

Se a matriz F final é de dimensão m,

$$\sum_{k=1}^{m} f_{ik}^2 = 1$$

Se F final é de dimensão p<m,

$\sum_{k=1}^{p} f_{ik}^2 < 1$, e esse somatório é chamado comunalidade da variável X_i.

$f'f = \lambda$ equivale a $\sum_{i=1}^{N} f_{ij}^2 = \lambda j$ e representa a contribuição do fator j para a variância total das variáveis.

Como a variância total é igual ao número de variáveis m, $\frac{\lambda j}{m}$ x 100 fornece a porcentagem da variância total dos dados explicada pelo fator j.

2.3.3 - Rotação dos fatores

Partindo-se da matriz dos fatores iniciais quer-se fazer uma mudança nos carregamentos das variáveis de modo que os novos valores dos carregamentos sejam as coordenadas das variáveis em um novo sistema de eixos ortogonais, obtidos de uma rotação dos eixos antigos.

Essa rotação possibilita uma interpretação lógica dos fatores. Em geral, é difícil a interpretação dos fatores como eles se apresentam na matriz dos fatores iniciais.

Em termos matriciais, se $F = (f_{ij})$ é a matriz dos fa-

tores iniciais $F^r = (f^r_{ij})$, matriz dos fatores após a rotação, é definida por

$$F^r = F.T$$

onde:

T = matriz de rotação final

A matriz T é o produto de sucessivas rotações que podem ser organizadas em uma ordem conveniente de forma que cada eixo seja rotacionado com cada um dos outros eixos apenas uma vez.

O processo de rotação apresentado pode ser generalizado para um espaço de dimensão p [8]. Nesse caso, a matriz de rotação T torna-se:

$$T = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1p} \\ \alpha_{21} & \alpha_{22} & & \alpha_{2p} \\ \multicolumn{4}{c}{\text{----------}} \\ \alpha_{p1} & \alpha_{p2} & & \alpha_{pp} \end{pmatrix}$$

As colunas da matriz acima são os cossenos diretores dos eixos F^r_1, F^r_2, ..., F^r_p em relação a F_1, F_2, F_3, ..., F_p.

Como a rotação é ortogonal, os cossenos diretores devem satisfazer a condição

$$\sum_{r=1}^{p} \alpha_{rk}\, \alpha_{r\ell} = \delta_{k\ell} \quad (K, \ell = 1,2,\ldots, p;\ K \leq \ell)$$

onde:

$\delta_{k\ell}$ é o delta de Kronecker, isto é,

$\delta_{k\ell} = 0$, se $K \neq \ell$

$\delta_{k\ell} = 1$, se $k = \ell$

Nesse trabalho é usado o método de rotação ortogonal VARIMAX. Esse método tem como principal finalidade rotacionar os eixos de modo a maximizar a variância dos quadrados dos carregamentos de cada coluna da matriz F.

2.4 - Análise Discriminante

2.4.1 - Considerações iniciais sobre a Análise Discriminante.

As N observações de um conjunto de m variáveis formam uma matriz X(N x m).

Se essas N linhas podem ser arrumadas de forma a constituirem grupos de observações segundo um critério estabelecido, isto é, se há uma classificação prévia dos elementos de X, essa matriz fica dividida em matrizes de dimensões menores formada pelos grupos de observações, onde a soma do número de linhas de todas essas matrizes é igual a N.

Feita essa classificação inicial, tem-se a matriz X preparada para a aplicação de uma Análise Discriminante.

Análise Discriminante é uma técnica estatística multivariada que tem como objetivo principal construir combinações lineares das variáveis originais que maximizem a separação entre os grupos. Essa técnica, também, pode ser usada para testar se os grupos são, realmente, distintos, determinar o poder discriminatório das variáveis iniciais e identificar aquelas que mais contribuem para a separação dos grupos, avaliar a dispersão interna dos grupos, identificar elementos indevidamente classificados, alocar elementos a um dos grupos existentes e melhorar a qualidade da classificação ini

cial por meio de iterações que tem como finalidade minimizar a variância interna dos grupos e maximizar a variância entre os grupos.

Nesse trabalho será usado o programa computacional de Análise Discriminante, da bateria de programas SPSS (Statistical Package for the Social Sciences).

2.4.2 - Enfoque matemático da Análise Discriminante.

Seja X uma matriz do tipo (Nxm), formada por grupos de observações, em que cada vetor coluna representa uma variável específica com seus valores dados em termos de desvios da sua média.

Então, cada observação de X pode ser escrita como:

$$x_{ijk} = m_{ij} + C_{ijk} \qquad (17)$$

onde:

x_{ijk} = k-ésima observação da variável X_i no grupo j.

m_{ij} = média de X_i no grupo j.

C_{ijk} = k-ésimo componente do erro da variável X_i no grupo j.

Generalizando a expressão (17), vem:

$$X = M + E, \qquad (18)$$

onde as linhas da matriz M, em cada grupo, são iguais.

$$\text{Seja } T = X'X \qquad (19)$$

T é chamada matriz da soma dos quadrados e produtos cruzados.

$$T = X'X = (M+E)'(M+E) = M'M + E'E + M'E + E'M \qquad (20)$$

Adotando a hipótese de não haver correlação entre as matrizes dos erros e as das médias, $M'E = E'M = 0$ e

$$T = M'M + E'E = B + W \qquad (21)$$

As matrizes B, W e T podem, também, ser definidas por:

$$B = (b_{ij}),\ b_{ij} = \sum_{k=1}^{G} Ng(\bar{X}_{ik} - \bar{X}_i)(\bar{X}_{jk} - \bar{X}_j)$$

$$W = (w_{ij}),\ w_{ij} = \sum_{k=1}^{G} \sum_{n=1}^{Ng} \left[(X_{ikn} - \bar{X}_{ik})(X_{jkn} - \bar{X}_{jk})\right]$$

$$T = (t_{ij}),\ t_{ij} = \sum_{p=1}^{N} (X_{ip} - \bar{X}_i)(X_{jp} - \bar{X}_j)$$

onde:

N = número total de observações

G = número de grupos

Ng = número de elementos no grupo g

m = número de variáveis originais

B = matriz de dispersão entre grupos

W = matriz de dispersão interna dos grupos

T = B + W = matriz de dispersão total,

$\bar{X}_{ik}$ = média aritmética da variável X_i no grupo k

i = 1, 2,..., m e k = 1, 2,..., G

$\bar{X}_i$ = média aritmética da variável X_i

X_{ikn} = n-ésima observação da variável X_i no grupo k

n = 1, 2,..., Ng

X_{ip} = p-ésima observação da variável X_i, p = 1, 2,...,N

Com a finalidade de maximizar a razão da dispersão en

tre os grupos para a dispersão interna dos grupos, pode-se projetar a matriz X em um vetor k, cujas componentes são os seus cossenos diretores.

Seja y = XK a projeção de X sobre K.

$$y'y = K'X'XK = K'TK = K'(B+W)K = K'BK + K'WK \qquad (22)$$

onde:

y'y = soma dos quadrados das projeções de X sobre K.

K'BK = parcela de y'y correspondente à dispersão entre os grupos.

K'WK = parcela de y'y correspondente à dispersão interna dos grupos.

Quer-se, portanto, maximizar a razão $\frac{K'BK}{K'WK}$.

$$\text{Seja } \lambda = \frac{K'BK}{K'WK} \qquad (23)$$

$$\log \lambda = \log(K'BK) - \log(K'WK)$$

Derivando-se log λ em relação a K e igulando-se o resultado a zero, tem-se:

$$\frac{\partial \log \lambda}{\partial K} = \frac{2BK}{K'BK} - \frac{2WK}{K'WK} = 0 \qquad (24)$$

Multiplicando-se os termos da expressão (24) por K'BK, vem:

$$2BK - 2WK\,(K'BK)/K'WK = 0$$

De (23), $(K'BK)/K'WK = \lambda$. Então:

$$BK - \lambda WK = 0 \qquad (25)$$

Multiplicando-se a expressão (25) à esquerda por W^{-1},

inversa de W, resulta.

$$W^{-1}BK - \lambda K = 0 \quad \text{ou}$$

$$(W^{-1}B - \lambda I)K = 0 \qquad (26)$$

Dessa forma, K é um autovetor associado ao autovalor λ de $W^{-1}B$.

Para se obter K deve-se encontrar as soluções não nulas do sistema de equações homogêneo (26).

A solução **não trivial do sistema de equações homogeneo** $(W^{-1}B - \lambda I)K=0$ requer:

$$\left|W^{-1}B - \lambda I\right| = 0$$

onde:

$$\left|W^{-1}B - \lambda I\right| = \text{determinante da matriz } (W^{-1}B - \lambda I)$$

Para cada autovalor de $W^{-1}B$ obtém-se um autovetor correspondente, K, em que suas componentes são os coeficientes de uma combinação linear das variáveis originais e essas combinações lineares são chamadas funções discriminantes.

Como o número de autovalores reais é igual ao mínimo de $\left[(Ng - 1), m\right]$, isso implica em que se tenha esse mesmo número de autovetores de $W^{-1}B$ e, consequentemente, esse mesmo número de funções discriminantes.

A forma geral de uma função discriminante é:

$$D_i = d_{i1} Z_1 + d_{i2} Z_2 + \ldots + d_{ip} Z_p, \; p \leq m$$

D_i = i-ésima função discriminante i=1,...,
min $\left[(Ng -1), m\right]$

Z_j = variável discriminante j, na forma padronizada, j = 1,..., p

d_{ij} = coeficiente da variável discriminatória j, na i-ésima função discriminante.

m = número de variáveis originais

p = número de variáveis discriminatórias.

As funções discriminantes são obtidas em ordem decrescente de importância. A primeira função explica o máximo da variância entre os grupos; a segunda função, ortogonal à primeira, o máximo da variância remanescente, e assim, sucessivamente, até se esgotar a variância existente na matriz $W^{-1}B$.

Um dos critérios, normalmente usados para se avaliar, qualitativamente uma função discriminante é a relação entre o seu autovalor e o somatório dos autovalores para todas as funções, expresso em porcentagem, $\frac{\lambda j}{\sum_{i=1}^{S} \lambda_i} \times 100$,

onde:

$S = \min \left[(Ng - 1), m \right]$

λ_i = i-ésimo autovalor de $W^{-1}B$

O poder discriminatório do conjunto das variáveis é geralmente, verificado através do lambda de Wilks.

$$\Lambda = \frac{|W|}{|T|} \quad , \quad 0 \leqslant \Lambda \leqslant 1$$

Quanto maior o poder discriminatório das variáveis menor o valor de Λ.

Se, no entanto, se pretende analisar o poder discriminatório de uma variável, tomada isoladamente, pode-se exami

nar o seu coeficiente na função discriminante em que a variável aparece, pela primeira vez, e baseando-se na importância da função, tem-se uma idéia do poder discriminatório dessa variável.

Uma outra alternativa para se analisar o poder discriminatório de cada variável é realizar o teste F de significância para cada variável.

2.4.3 - Interpretação Bayesiana da Análise Discriminante.

O objetivo principal da interpretação Bayesiana da Análise Discriminante é, conhecida uma observação x, alocar essa observação para um dos grupos anteriormente definidos. Essa alocação é feita para o grupo ao qual x tem a maior probabilidade de pertencer.

Sejam H_1, H_2,..., H_r, r hipóteses mutuamente exclusivas e $p(H_1)$, $p(H_2)$,..., $p(H_r)$ as probabilidades associadas a essas hipóteses.

$p(H_g)$, g = 1, 2,..., r, pode ser interpretada como a probabilidade de ser correta a sentença: H_g é verdadeira para a observação x.

$$p(H_g/x) = \frac{p(Hg) \cdot p(x/Hg)}{p(x)} \qquad (27)$$

onde:

$p(H_g/x)$ = probabilidade de H_g ser verdadeira se x é observado

$p(x/H_g)$ = probabilidade condicional de x ser observado se H_g é verdadeira.

$p(x)$ = probabilidade associada à observação x.

Então, se o conjunto das hipóteses contém todas as alternativas possíveis, uma delas deve ser verdadeira. Tem-se, pois:

$$\sum_i p(H_i/x) = 1 \qquad (28)$$

Portanto, o somatório de (27) para todo i variando de de 1 a r, torna-se:

$$\sum_i p(H_i/x) = \frac{\sum_i \left[p(H_i) \cdot p(x/H_i)\right]}{p(x)}$$

De (28), $\sum_i p(H_i/x)=1$. Resulta, então:

$$p(x) = \sum_i \left[p(H_i) \cdot p(x/H_i)\right] \qquad (29)$$

Substituindo-se (29) em (27), vem:

$$p(H_g/x) = \frac{p(H_g) \cdot p(x/H_g)}{\sum_i \left[p(H_i) \cdot p(x/H_i)\right]} \qquad (30)$$

Similarmente, para uma outra hipótese H_f, tem-se:

$$p(H_f/x) = \frac{p(H_f) \cdot p(x/H_f)}{\sum_i \left[p(H_i) \cdot p(x/H_i)\right]} \qquad (31)$$

Dividindo-se (30) por (31), chega-se a:

$$\frac{p(H_g/x)}{p(H_f/x)} = \frac{p(H_g)}{p(H_f)} \cdot \frac{p(x/H_g)}{p(x/H_f)} = \frac{p(H_g)}{p(H_f)} \cdot L \qquad (32)$$

onde:

$\frac{p(H_g/x)}{p(H_f/x)}$ = razão das probabilidades a posterior

$\frac{p(H_g)}{p(H_f)}$ = razão das probabilidades a priori

$\frac{p(x/H_g)}{p(x/H_f)}$ = L = razão de verossimilhança

Uma estimativa da razão das probabilidades a priori, $p(H_g)/p(H_f)$, pode ser obtida de N_1/N_2.

onde:

N_1 = número de observações do grupo 1

N_2 = número de observações do grupo 2

A expressão (32) mostra L como um fator de correção da razão das probabilidades a priori para se obter a razão das probabilidades a posteriori. Isso permite, dada uma observação, inferir que hipótese é a mais provável.

2.4.3.1 - Análise discriminante para dois grupos. Uso da razão de verossimilhança.

Sejam duas hipóteses H_1 e H_2 para as quais se conheça as distribuições de probabilidade de uma observação x, $f_1(x)$ e $f_2(x)$.

Assim,

$$\frac{f_1(x)}{f_2(x)} = L \qquad (33)$$

Em geral, torna-se mais simples, em vez de se considerar L, considerar-se o logaritmo de L, isto é, log L, principalmente se $f_1(x)$ e $f_2(x)$ são distribuições normais ou multinormais.

No caso multinormal,

$$\log\left[f_1(x)\right] = -\frac{1}{2} m \log(2\pi) - \frac{1}{2} \log\left|V_1\right| - \frac{1}{2} (x-\bar{x}_1)' V_1^{-1} (x-\bar{x}_1)$$

$$\log \left[f_2(x)\right] = -\frac{1}{2} m \log (2\pi) - \frac{1}{2} \log \left| V_2 \right| - \frac{1}{2} (x-\bar{x}_2)' V_2^{-1} (x-\bar{x}_2)$$

onde:

$\left|V_1\right|$ e $\left|V_2\right|$ = determinantes das matrizes de variância-covariância interna dos grupos 1 e 2, respectivamente.

x = vetor cujas componentes são os valores das variáveis correspondentes a uma observação considerada.

$\bar{x}_1$ e $\bar{x}_2$ = vetores cujas componentes são os valores médios das observações das variáveis para os grupos 1 e 2, respectivamente.

$$\log \left[f_1(x)/f_2(x)\right] = \log L = -\frac{1}{2} \log V_1 + \frac{1}{2} \log \left|V_2\right| - \frac{1}{2} (x-\bar{x}_1)' V_1^{-1} (x-\bar{x}_1) + \frac{1}{2} (x-\bar{x}_2) V_2^{-1} (x-\bar{x}_2) \qquad (34)$$

Se $V_1 = V_2 = V$, (34) torna-se:

$$\text{Log } L = -\frac{1}{2} (x-\bar{x}_1)' V^{-1} (x-\bar{x}_1) + \frac{1}{2} (x-\bar{x}_2)' V^{-1} (x-\bar{x}_2)$$

$$\text{Log } L = x' V^{-1} (\bar{x}_1-\bar{x}_2) - \frac{1}{2} (\bar{x}_1' V^{-1} \bar{x}_1 - \bar{x}_2' V^{-1} \bar{x}_2) \qquad (35)$$

A matriz V de variância-covariância interna dos grupos pode ser estimada através de [10]:

$$V = \frac{1}{N_1 + N_2 - 2} \left[\sum_i (x_i-\bar{x}_1)(x_i-\bar{x}_1)' + \sum_i (x_i-\bar{x}_2)(x_i-\bar{x}_2)'\right]$$

Quando log L = 0, L = 1 e com base em (33), $f_1(x) = f_2(x)$. Tem-se, portanto, da expressão (32),

$$\frac{p(H_g/x)}{p(H_f/x)} = \frac{p(H_g)}{p(H_f)} \qquad (36)$$

Nesse caso, (36) indica que a razão das probabilida-

des a priori é igual a razão das probabilidades a posteriori, implicando em que não se pode estabelecer se a observação x provêm do primeiro grupo ou do segundo.

A curva associada a log L é chamada curva de verossimilhança. Se log L = 0, ou seja, L = 1, obtém-se a curva discriminante, isto é, aquela que representa o lugar geométrico das observações com a mesma distribuição de probabilidade para as hipóteses H_1 e H_2.

2.4.3.2 - Análise Discriminante para mais de dois grupos. Uso da razão de verossimilhança.

A abordagem da Análise Discriminante para mais de dois grupos é, apenas, uma generalização do caso de dois grupos. Para três grupos, por exemplo, adota-se as hipóteses H_1, H_2 e H_3 e calcula-se as curvas de verossimilhança para os grupos, tomados dois a dois, como segue:

$$\begin{aligned} \log L_{12} &= \log \left[f_1(x)/f_2(x) \right] = \log \left[f_1(x) \right] - \log \left[f_2(x) \right] \\ \log L_{23} &= \log \left[f_2(x)/f_3(x) \right] = \log \left[f_2(x) \right] - \log \left[f_3(x) \right] \\ \log L_{31} &= \log \left[f_3(x)/f_1(x) \right] = \log \left[f_3(x) \right] - \log \left[f_1(x) \right] \end{aligned} \qquad (37)$$

Quando $\log L_{12} = \log L_{31} = 0$, o espaço das observações fica dividido em regiões delimitadas pelos hiperplanos $\log L_{23}=0$ e $\log L_{31} = 0$, e pode-se verificar que a terceira equação do sistema (37) é o resultado da soma das duas primeiras. Assim, despreza-se a última equação e as probabilidades a posteriori, dada uma observação de x, para as hipó-

teses H_1, H_2 e H_3 podem ser dadas por:

$$\frac{p\ (H_1/x)}{p\ (H_2/x)} = \frac{p\ (H_1)}{p\ (H_2)} \cdot L_{12}$$

$$\frac{p\ (H_2/x)}{p\ (H_3/x)} = \frac{p\ (H_2)}{p\ (H_3)} \cdot L_{23}$$

Nesse caso, a observação x será alocada à hipótese que tenha a maior probabilidade de ocorrer.

CAPÍTULO III

APLICAÇÕES E RESULTADOS

3.1 - Introdução

Neste capítulo são apresentadas aplicações dos métodos estatísticos multivariados: Análise de Regressão Linear Multipla, Análise Fatorial e Análise Discriminante.

Os dados utilizados nas aplicações para Campina Grande provém das seguintes fontes: de uma pesquisa domiciliar, realizada em 1978, pelo GEIPOT e de um trabalho realizado pelo professor Masayuki Doi[11], sob o título "The land use and zoning study in relation to systematic transportation study".

Os resultadcs das aplicações foram obtidos através do uso do pacote computacional SPSS [2] no computador IBM-370, do Núcleo de Processamento de Dados da Universidade Federal da Paraíba, Campus de Campina Grande.

3.2 - Análise de Regressão Linear Múltipla. Aplicações e resultados.

3.2.1 - Aplicação 1

Esta aplicação tem por finalidade verificar se o nível de agregação das variáveis envolvidas numa equação de regressão linear tem influência na qualidade dessa equação.

Mc Carthy [7] construiu equações de regressão linear com as variáveis agregadas a nível de distrito, setor, zona e domicílio e verificou que essas equações tornavam-se melhores à medida que o nível de agregação diminuía.

Para esta aplicação, utilizou-se os dados da pesquisa domiciliar em Campina Grande realizada pelo Geipot, em 1978, e considerou-se essa cidade dividida a nível de setor, de zona e de domicílio.

Os setores foram definidos como o agrupamento de determinadas zonas e a Tabela 1 mostra as zonas que compõem cada setor. O zoneamento adotado foi o utilizado pelo Geipot na pesquisa domiciliar em 1978, em que a cidade foi dividida em 23 zonas de tráfego.

Tabela 1 - Setores e Suas Zonas

Setor	Zonas
1	1, 2, 3, 4
2	5, 6, 7
3	8, 9, 10, 11, 12
4	13, 14, 15
5	16, 17, 18, 19, 20
6	21, 22, 23

Para cada nível computou-se as seguintes variáveis:

X_1 = viagens diárias com base domiciliar por domicílio.

X_2 = renda por domicílio

X_3 = tamanho da família por domicílio

X_4 = número de veículos por domicílio

X_5 = pessoas maiores ou iguais a 5 anos por domicílio.

Essas variáveis foram as mesmas empregadas por Mc Carthy[7].

Com base no raciocínio de se considerar as variáveis em diversos níveis de agregação, empregado por Mc Carthy[7], e nos coeficientes de correlação linear entre as variáveis, mostrados nas Tabelas 2, 3 e 4, construíu-se as equações de regressão linear de X_1 em X_4 e X_2 e de X_1 em X_4, tanto para os níveis de setor como para o de zona, e as equações de X_1 em X_5 e X_3 e de X_1 em X_5, para o nível de domicílio.

Tabela 2 - Matriz de correlação entre as variáveis a nível de setor

	X_1	X_2	X_3	X_4	X_5
X_1	1,0000	0,2369	-0,5536	0,5666	-0,0861
X_2		1,0000	0,1375	-0,0956	0,1252
X_3			1,0000	-0,6325	0,8503
X_4				1,0000	-0,3722
X_5					1,0000

Tabela 3 - Matriz de correlação entre as variáveis a nível de zona

	X_1	X_2	X_3	X_4	X_5
X_1	1,0000	0,5464	-0,5117	0,6911	-0,5393
X_2		1,0000	-0,3778	0,7822	-0,2409
X_3			1,0000	-0,5861	0,9480
X_4				1,0000	-0,4991
X_5					1,0000

Tabela 4 - Matriz de correlação entre as variáveis a nível de domicílio

	X_1	X_2	X_3	X_4	X_5
X_1	1,0000	0,2304	0,5667	0,2561	0,6140
X_2		1,0000	0,0730	0,3820	0,0917
X_3			1,0000	0,1021	0,9535
X_4				1,0000	0,1150
X_5					1,0000

As equações com os seus respectivos parâmetros são:

3.2.1.1 - Nível de setor

Número de observações = 6

$$X_1 = 4,5046\ X_4 + \frac{0,1709}{10^4} X_2 + 5,7191 \qquad (38)$$

$B_{X_4} = 0,5947$; $B_{X_2} = 0,2937$

$R^2 = 0,4066$; Se = 1,0225; F = 1,0277; F_3^2 (5%) = 9,55

$F_{X_4} = 1,7772$; $F_{X_2} = 0,432$; F_3^1 (5%) = 10,13

onde:

B_{X_4} e B_{X_2} = coeficientes de regressão padronizados de X_4 e X_2, respectivamente

R^2 = coeficiente de determinação múltipla

Se = erro padrão de estimativa

F = $F_{calculado}$ para a equação

F_{X_4} e F_{X_2} = $F_{calculado}$ para, respectivamente, os coeficientes de X_1 e X_2

$F^{GLN}_{GLD}(\alpha)$ = $F_{crítico}$, isto é, tirado de uma tabela de distribuição F, para GLN graus de liberdade do numerador, GLD graus de liberdade do denominador e um nível de significância α.

Pelo critério recomendado neste trabalho de se adotar como equação de previsão aquelas equações possuidoras de um coeficiente de determinação R^2 próximo de 1, um erro padrão de estimativa o menor possível, e os coeficientes de regressão parcial estatisticamente significantes a um nível de significância, previamente, adotado, a equação (38) não poderia ser usada como equação de previsão de demanda de viagens, pois, não apresenta o coeficiente de correlação múltipla R e os coeficientes de regressão parcial, estatisticamente significantes ao nível de significância de 5% porque $F < F^2_3(5\%)$ e F_{X_4} e F_{X_2} ambos são menores do que o $F^1_3(5\%)$.

A outra equação de regressão obtida, ainda, para o nível de setor foi:

$$X_1 = 4{,}2919\ X_4 + 6{,}0381 \qquad (39)$$

$$B_{X_4} = 0{,}5666$$

$$R^2 = 0{,}3211;\ S_e = 0{,}9472;\ F = 1{,}8918;\ F^1_4(5\%) = 7{,}71$$

Como a equação (39) é uma equação de regressão linear simples, o $F_{calculado}$ para o coeficiente de regressão é igual ao $F_{calculado}$ para a equação e, portanto, desnecessário escrevê-lo.

Nota-se que a equação (39) também não possui o seu coeficiente de regressão estatisticamente significante a 5% de

nível de significância, além de um pequeno valor para R^2.

Portanto, essa equação, também, não satisfaz às condições exigidas, nesse trabalho, para uma equação de previsão.

3.2.1.2 - Nível de zona

Número de observações = 19

Apesar de Campina Grande estar dividida em 23 zonas de tráfego, as zonas 2, 4, 6 e 17 foram eliminadas porque não possuem domicílios. Portanto, restaram 19 zonas que correspondem ao número de observações utilizadas nos cálculos das equações de regressão linear, a esse nível de agregação das variáveis.

O processo de obtenção das equações foi o "Stepwise" onde a entrada das variáveis independentes na equação se dá em ordem decrescente de sua capacidade de explicar a variância da variável dependente. Pode ocorrer, pois, que uma ou mais variáveis independentes fiquem fora da equação de regressão final, em virtude de essas variáveis não darem contribuição adicional para a explicação da variância da variável dependente.

Para o nível de zona, na equação de regressão linear de X_1 em X_4 e X_2, a variável X_2 foi eliminada; com isso se obteve, somente, a equação de regressão linear, dada por:

$$X_1 = 9{,}1482\ X_4 + 3{,}8220 \qquad (40)$$

$B_{X4} = 0{,}6911$

$R^2 = 0{,}4776$; Se = 1,8031; F = 15,5449; F_{17}^{1} (5%)=4,45

A equação (40) possui o valor de R^2 pequeno e, embora

o coeficiente de regressão seja estatisticamente significante a 5% de nível de significância, o valor do coeficiente de regressão parece excessivo, visto que ele indica um acréscimo de 9 viagens, aproximadamente, para um acréscimo de um automóvel por domicílio. Essa equação, também, não deveria ser usada como modelo de previsão de demanda de viagens.

3.2.1.3 - Nível de domicílio

Número de observações = 1940

As equações obtidas a nível de domicílio foram:

$$X_1 = 2,1455\ X_5 - 0,5258\ X_3 + 0,6341 \qquad (41)$$

$B_{X_5} = 0,8112$; $B_{X_3} = -0,2068$

$R^2 = 0,3809$; $S_e = 4,8466$; $F = 595,8394$; $F^2_{1937}(5\%) = 4,61$

$F_{X_5} = 187,007$; $F_{X_3} = 12,159$; $F^1_{1937}(5\%) = 6,63$

A equação (41) possue um valor de R^2, ainda, pequeno porém a equação e os coeficientes de regressão são estatisticamente significantes ao nível de significância de 5%.

Vale notar, por exemplo, que o coeficiente de X_3 não possue o sinal coerente com a definição dessa variável. Tem se, nesse caso, um decréscimo de 0,52 viagens para uma acréscimo de uma pessoa no tamanho da família, mantendo-se constante a variável X_5.

Inicialmente, era de se esperar que a equação (41) produzisse melhores resultados. Isto porque os coeficientes de correlação linear entre X_1 e X_5 e entre X_1 e X_3, apesar de não serem altos, em termos absolutos, podem ser aceitos como, relativamente, altos, quando comparados com os demais

coeficientes de correlação de X_1 com as outras variáveis, mostrados na Tabela 4. Entretanto, a matriz de correlação entre as variáveis envolvidas em uma equação de regressão linear pode antecipar alguma indicação sobre um possível resultado. Esse é o caso da Tabela 4 que mostra o coeficiente linear entre X_5 e X_3 igual a 0,9535, portanto, muito alto. Surge, então,o problema da multicolineridade e, podia-se ter suspeitado, desde antes de se obter a equação (41), da qualidade da equação de regressão linear de X_1 em X_5 e X_3.

A outra equação para o nível de domicílio é dada por:

$$X_1 = 1,6239\ X_5 + 0,4068 \qquad (42)$$

$$B_{X_5} = 0,61401$$

$$R^2 = 0,3770;\ S_e = 4,8605;\ F = 1172,7676;\ F^1_{1938}(5\%) = 6,64$$

Conforme mostra a equação (42), o valor do coeficiente de determinação múltipla, R^2, é pequeno. Portanto, a equação (42) não possui todas as qualidades, para ser usada como modelo de previsão de viagens, recomendadas neste trabalho.

3.2.1.4 - Considerações sobre a aplicação 1.

Comparando-se os resultados das equações obtidas para os três níveis de agregação das variáveis, pode-se ser levado a supor que a equação obtida para o nível de zona é melhor do que as obtidas para o nível de domicílio, porque apresentam um R^2 ligeiramente maior e um erro padrão de estimativa menor.

No entanto, Mc Carty[7] alerta para o significado do

UFPb/BIBLIOTECA/PRAI

coeficiente de determinação. O coeficente de determinação, R^2, mede a variância da variável dependente que é explicada pela combinação linear das variáveis independentes. Assim, quanto maior o valor de R^2 maior é a relação de linearidade entre as variáveis dependente e independentes, não significando, porém, uma maior explicação da relação de causa-efeito, por parte da equação.

Mc Carthy[7], através da análise de variância mostrou que somente os fatores entre as unidades de agregação são levados em conta nos modelos de regressão linear, deixando-se de lado os fatores internos. Isso mostra a limitação dos modelos à base da correlação linear e a necessidade de se considerar modelos que incorporem esses fatores internos.

É evidente, porém, que à falta de dados que permitam um estudo desagregado do problema da demanda de viagens pode-se utilizar modelos à base da correlação linear embora deva-se estar consciente das limitações dos modelos empregados.

Nesta aplicação, foi seguido o raciocínio empregado por Mc Carthy[7] e os resultados encontrados mostraram uma tendência similar à verificada pelos resultados dele.

3.2.2 - Aplicação 2

A finalidade dessa aplicação é obter uma equação de regressão linear múltipla que relacione o número de automóveis em cada zona de tráfego com o número de famílias em cada nível de renda, por zona.

Inicialmente, considerou-se a renda familiar das pes

soas residentes em Campina Grande, em 1978, dividida em 20 classes de amplitude igual a Cr$ 1.000,00, e determinou-se, para cada classe, o quociente automóveis/família, através da razão NA(J)/NF(J),

onde:

NA(J) = número total de automóveis pertencentes às famílias de classe de renda J, J=1,2,...,20.

NF(J) = número total de famílias de classe de renda J.

Isso possibilitou a verificação e agrupamento daquelas classes que apresentaram o quociente automóveis/família, aproximadamente, igual e tornou-se possível reduzir, de 20 para 4, o número de classes. A tabela 5 mostra os quatro níveis de renda considerados, as classes de renda correspondentes, e o número de automóveis/família, para cada uma dessas classes.

Tabela 5 - Níveis de renda e número médio de automóveis por família em cada classe.

Níveis de renda	Classes	Automóveis/família
1	0 - 6.999	0,174
2	7.000 - 13.999	0,701
3	14.000 - 20.999	1,010
4	⩾ 21.000	1,232

Com os níveis de renda definidos, computou-se, por zona, o número de automóveis e o número de famílias, em cada

nível de renda. A Tabela 6 mostra esse resultado.

Com os dados dessa Tabela, exceto os referentes às zonas 2,4,6,11 e 17, calculou-se a equação de regressão linear múltipla $Y = f(X_1, X_2, X_3, X_4)$,

onde:

Y = número de automóveis por zona

X_1 = número de famílias de nível de renda 1, por zona

X_2 = número de famílias de nível de renda 2, por zona

X_3 = número de famílias de nível de renda 3, por zona

X_4 = número de famílias de nível de renda 4, por zona

A Tabela 7 mostra a matriz de correlação entre essas variáveis.

Tabela 7 - Matriz de correlação

	Y	X_1	X_2	X_3	X_4
Y	1,0000	0,4942	0,9411	0,8248	0,7952
X_1		1,0000	0,4430	0,1348	0,0732
X_2			1,0000	0,8073	0,7171
X_3				1,0000	0,8864
X_4					1,0000

A equação obtida foi:

$$Y=0,1206X_1+1,3866X_2+0,0729X_3+2,6269X_4-6,0240 \qquad (43)$$

$B_{X_1}=0,2043$; $B_{X_2}=0,5927$; $B_{X_3}=0,0181$; $B_{X_4}=0,3392$

$R^2 = 0,9434$; $S_e=7,9591$; $F=54,1583$; $F^4_{14}(5\%)= 3,34$

$F_{X_1}= 6,237$; $F_{X_2}=18,804$; $F_{X_3}=0,011$; $F_{X_4}=5,569$; $F^1_{14}(5\%)=4,60$

A equação (43) possui um valor alto para R^2 e o coefi

Tabela 6 - Número de automóveis e número de famílias, em cada nível de renda, por zona.

Zona	Autos	Número de famílias em cada nível de renda.			
		1	2	3	4
1	51	43	26	14	11
2	0	0	0	0	0
3	96	91	50	24	9
4	0	0	0	0	0
5	89	97	28	12	10
6	0	0	0	0	0
7	30	70	11	7	3
8	28	164	12	2	2
9	18	23	6	1	3
10	73	121	22	20	11
11	0	1	0	0	0
12	12	65	4	1	1
13	7	19	5	0	4
14	59	124	25	5	5
15	20	101	9	2	1
16	40	170	18	1	2
17	0	0	0	0	0
18	3	35	5	1	1
19	12	69	5	4	3
20	6	23	3	0	0
21	12	38	4	1	1
22	58	160	26	2	2
23	19	92	9	0	0

ciente de correlação múltipla, R, e os coeficientes de regressão, exceto o da variável X_3, são estatisticamente significantes ao nível de significância de 5%.

Entretanto, se os coeficientes de regressão da equação (43) representassem o número médio de automóveis por família, para cada nível de renda, como era de se esperar, a equação (43) se transformaria em:

$$Y = 0,174X_1+0,701X_2+1,010X_3+1,232X_4 \qquad (44)$$

Nota-se, portanto, grandes discrepâncias entre os coeficientes das equações (43) e (44). A ocorrência dessas discrepâncias pode ser, principalmente, devida ao efeito da multicolinearidade, pois, as variáveis X_2, X_3 e X_4. tomadas duas a duas, possuem o coeficiente de correlação linear, relativamente, alto, conforme mostrado na Tabela 7.

Na equação (43) o coeficiente da variável X_3 é, aproximadamente, nulo. Então, eliminou-se essa variável e calculou-se uma equação de regressão linear múltipla de Y em X_1, X_2, e X_4.

A equação obtida foi:

$$Y = 0,1195X_1+1,4049X_2+2,7087X_4 - 6,1294 \qquad (45)$$

$$B_{X_1} = 0,2025;\ B_{X_2}=0,6005;\ B_{X_4}=0,3497;\ F=54,1583;\ F^{3}_{15}(5\%)=3,34$$

$$R^2 = 0,9434;\ S_e = 7,9591$$

$$F_{X1} = 6,897;\ F_{X_2}=29,627;\ F_{X_4}=12,435;\ F^{1}_{15}(5\%)= 4,60$$

Analisando-se a equação (44) constata-se que ela possui qualidades que permitem o seu uso como equação de previsão.

Por outro lado, comparando-se os resultados das equações (43) e (45) verifica-se que elas possuem o mesmo valor para R^2, o mesmo S_e e os valores dos coeficientes de regressão são, praticamente, iguais. Portanto, pode-se usar, também, a equação (43) como equação de previsão, sem deixar, no entanto, de se reconhecer as suas limitações para esse fim.

As equações (43), (44) e (45) podem ser usadas, por exemplo, para se estimar o número de automóveis em uma zona, desde que se conheça o número de famílias em cada nível de renda, por zona de tráfego. Essa aplicação pode ser feita tanto para novas zonas de tráfego como para as zonas já existentes, em situações futuras.

3.3 - Análise Fatorial. Aplicações e Resultados

3.3.1 - Aplicação 1

Essa aplicação utiliza variáveis que descrevem o uso do solo dos bairros de Campina Grande e tem como objetivo reduzir o número dessas variáveis, sem perda significante de informações. Para isso, empregou-se a Análise Fatorial do tipo R.

As variáveis originais utilizadas representam as frações da área total, de cada bairro, destinadas a cada tipo de uso do solo.

De posse da divisão em 54 bairros feita pela Companhia de Desenvolvimento de Campina Grande - COMDECA, Masayuki Doi[11] fez a classificação do uso do solo de Campina Grande conforme mostra o Quadro 1.

Quadro 1 - Classificação do uso do solo

---

1 - Residencial

1.1 - Residencial unifamiliar

RU_1 - multicômodo isolado

RU_2 - multicômodo conjugado

RU_3 - unicômodo isolado, geminado, conjugado

1.2 - Residencial multifamiliar

RM_1 - edifício multicômodo

RM_2 - edifício unicômodo

2 - Comercial

2.1 - Região comercial (atende a população de todos os bairros)

CR_1 - compras grandes (movelaria, implementos agrícolas, etc.)

CR_2 - compras pequenas (camisas, sapatos, etc.)

CR_3 - serviços (restaurante, postos de gasolina , cinema, etc.)

2.2 - Comercial de bairro (atende, principalmente, a população do bairro)

CB - comércio de bairro (bares, etc.)

2.3 - Comercial de manufatura

CM - comércio de manufatura (oficinas, atacadistas, etc.)

3 - Industrial

I - Indústria

4 - Agricultura

A - agricultura, fazendas, etc.

Continuação do Quadro 1

5 - Público

P - igrejas, escolas, prefeituras, etc.

6 - Espaço Aberto

EA - Campos de futebol, parques, etc.

7 - Utilidade (Equipamentos de utilidade pública de grande consumo de solo)

U - Cagepa, aeroporto, estação ferroviária, etc.

8 - Não Desenvolvido

ND - lotes desocupados.

O Quadro 2 mostra a denominação dos bairros de Campina Grande.

Quadro 2 - Numeração e nomes dos bairros.

Número	Bairro	Número	Bairro
1	São José	28	Conceição
2	Açude Novo	29	Louzeiro
3	Açude Velho	30	Pirineus
4	Centro	31	Bela Vista
5	Ariús	32	Monte Santo
6	Acauã	33	Geremias
7	Vila Cabral	34	Redentorista
8	Sandra Cavalcanti	35	Araxá
9	Catolé	36	Cidade Universitária
10	Prado	37	Centenário
11	Sandra	38	Casa de Pedra
12	Provisão	39	Dona Merquinha

Continuação do Quadro 2

Número	Bairro	Número	Bairro
13	Passa Tempo	40	Produção Mineral
14	Tambor	41	Quarenta
15	Santíssima	42	Santana
16	Cachoeira	43	Santa Rosa
17	Monte Castelo	44	Jardim Nordeste
18	Nova Brasília	45	Moita
19	Vila Castelo Branco	46	Cruzeiro
20	Santo Antonio	47	Liberdade
21	Tavares	48	Graças
22	Lauritzen	49	Jardim Paulistano
23	Alto Branco	50	Três Irmães
24	José Pinheiro	51	Distrito Industrial
25	Palmeira	52	Adrianópolis
26	Areias	53	Bodocongó
27	Prata	54	Zoobotânico.

Masayuki Doi[11] contém os valores das variáveis para cada uso do solo mencionado no Quadro 1, em m^2.

Definiu-se, para cada bairro, as variáveis:

XX1 = $(RU_1 + RU_2 + RU_3 + RM_1 + RM_2)/TOTAL$

XX2 = $CR_1/TOTAL$

XX3 = $(CR_2 + CB)/TOTAL$

XX4 = $CR_3/TOTAL$

XX5 = CM/TOTAL

XX6 = I/TOTAL

XX7 = A/TOTAL

XX 8 = P/TOTAL

XX 9 = EA/TOTAL

XX10 = U/TOTAL

XX11 = ND/TOTAL

Construíu-se, então a matriz X do tipo (54x11)e fez-se uma Análise Fatorial tipo R. A Tabela 8 é um extrato da listagem de computador para essa Análise Fatorial.

Tabela 8 - Numeração dos fatores, autovalores e porcentagem da variância do conjunto das variáveis explicada por cada fator

Fator Fi	Autovalor λ_i	% da variância	% Acumulada da Variância
1	3.14453	28.6	28.6
2	1.67022	15.2	43.8
3	1.32993	12.1	55.9
4	1.21150	11.0	66.9
5	1.02206	9.4	76.2
6	0.92752	8.4	84.7
7	0.84164	7.7	92.3
8	0.66189	6.0	98.3
9	0.16462	1.5	99.8
10	0.01337	0.1	100.0
11	0.00357	0.0	100.0

Conforme visto no enfoque matemático, $\frac{\lambda_i}{m}$ x 100 fornece a porcentagem da variância do conjunto de dados explicada pelo fator i, F_i. Portanto, o primeiro elemento da terceira coluna, 28,6, é igual a $\frac{3,14453}{11}$ x 100. A quarta coluna é obtida somando-se cada elemento da terceira coluna à soma dos valores da terceira coluna, anteriores a esse elemento.

Como se quer reduzir o número de variáveis, o número de fatores da solução final deve ser menor do que 11. Para isso, adota-se um autovalor como mínimo e despreza-se todo

Com base na Tabela 9, pode-se escrever qualquer variável como uma combinação linear dos fatores. Tem-se, por exemplo,

$$XX1 = 0,17549F_1+0,05183F_2+0,86013F_3+0,05274F_4-0,21618F_5 \quad (46)$$

Desde que se conheça os componentes de cada fator pode-se calcular os correspondentes valores da variável original XX1, dada por (46). As componentes ou escores de um fator podem ser obtidas através da Tabela 11.

Tabela 12 - Matriz dos coeficientes para o cálculo dos escores dos fatores.

	Fator 1	Fator 2	Fator 3	Fator 4	Fator 5
XX1	-0.02061	-0.12757	0.57440	0.05882	-0.17401
XX2	0.45385	-0.07627	-0.09835	-0.07508	-0.00309
XX3	0.43617	-0.15742	0.01315	-0.04920	-0.00293
XX4	0.18454	0.36471	-0.13836	0.02260	0.03849
XX5	0.01644	0.41383	-0.03538	0.12210	-0.00747
XX6	-0.08155	0.18342	-0.06983	0.04638	0.48682
XX7	-0.00747	-0.10301	-0.31164	0.48930	0.06015
XX8	-0.08045	-0.05900	0.43667	0.09898	0.28220
XX9	-0.16048	0.49716	-0.05978	-0.12202	-0.06567
XX10	-0.07955	0.20973	-0.08250	0.04598	-0.72613
XX11	0.06922	-0.01824	-0.19821	-0.63844	0.04127

Os escores de F_1 podem ser dados por:

$$F_1 = -0,02061XX1+0,45385XX2+0,43617XX3+...+0,06922XX11 \quad (47)$$

Entrando-se com os valores padronizados de cada observação das variáveis originais em (47), pode-se calcular os escores de F_1. De forma semelhante, calcula-se os escores dos outros fatores e, construindo-se uma matriz onde os vetores coluna sejam os escores de cada fator, a correlação linear entre cada par de vetores-coluna é zero, visto que os

fator correspondente a um autovalor menor do que o especificado como mínimo.

A menos que se diga o contrário, o autovalor adotado como mínimo, nas aplicações de Análise Fatorial neste trabalho, é 1.

Assim, apenas, os cinco primeiros fatores da Tabela 9 permanecerão na solução final, passando-se de 11 variáveis originais para 5 variáveis novas, chamadas fatores.

A matriz dos carregamentos dos fatores após a rotação é:

Tabela 09 - Matriz de carregamentos dos fatores após a rotação varimax.

	Fator 1	Fator 2	Fator 3	Fator 4	Fator 5
XX1	0.1754[illegible]	0.051[illegible]	-0.86013	0.0527[illegible]	[illegible].2141[illegible]
XX2	[illegible]	[illegible]	0.02619	0.0[illegible]619	[illegible]
XX3	[illegible]	[illegible]	0.15617	0.0472[illegible]	-0.0734[illegible]
XX4	[illegible]	-0.67[illegible]	0.[illegible]5714	0.12879	0.[illegible]513
XX5	0.[illegible]	0.71621	0.16461	0.22542	0.14711
XX6	-0.1[illegible]	0.[illegible]	-0.35726	0.05051	0.59115
XX7	-0.[illegible]	-0.29[illegible]	-0.58494	0.7166[illegible]	0.04[illegible]
XX8	-0.[illegible]	0.12494	-0.63784	0.09052	0.91359
XX9	-0.1[illegible]	0.7[illegible]	0.06931	-0.17806	-0.00191
XX10	-0.09277	0.1712[illegible]	-0.06836	0.08417	0.90226
XX11	-0.13[illegible]	-0.14919	-0.27769	0.91350	0.06005

A Tabela 10 fornece as comunalidades das variáveis

Tabela 10 - Comunalidade das Variáveis

Variável	Comunalidade
XX1	[illegible]
XX2	[illegible]
XX3	[illegible]
XX4	[illegible]
XX5	[illegible]
XX6	[illegible]
XX7	0.94[illegible]
XX8	0.529[illegible]
XX9	0.56471
XX10	0.69332
XX11	0.95714

fatores são ortogonais.

Analisando-se a Tabela 9, pode-se destacar em que fator cada variável tem projetada a maior quantidade de sua variância. Isso é mostrado na Tabela 12.

Tabela 12 - Fatores mais importantes na composição da variância das variáveis

Fator 1	Fator 2	Fator 3	Fator 4	Fator 5
XX2	XX5	XX1	XX11	XX10
XX3	XX9	XX8	XX 7	XX 6
	XX4			

Com base na tabela 12, pode-se, então, considerar as 5 variáveis da solução final como uma junção das variáveis originais, como mostra a Tabela 13.

Tabela 13 - Fatores e suas definições em termos das variáveis originais dadas em valores relativos

Fatores	Variáveis Originais	Característica dos Fatores
F_1	CR_1+CR_2+CB	Variável que caracteriza usos do solo destinados a comércio.
F_2	$CM + EA + CR_3$	Variável que caracteriza usos do solo destinados a serviços e lazer
F_3	$RU_1+RU_2+RU_3+RU_4+RU_5+P$	Variável que caracteriza usos do solo destinados a residências e a equipamentos de uso público
F_4	A + ND	Variável que caracteriza usos do solo, destinados a agricultura e solos não ocupados.
F_5	I + U	Variável que caracteriza usos do solo destinados a indústria e a equipamentos de utilidade pública.

As variáveis F_1, F_2, F_3, F_4 e F_5, mostradas na Tabela 13, são não correlacionadas, entre si, e, portanto, poderiam ser usadas, por exemplo, como variáveis independentes em modelos à base do cálculo da correlação linear.

Dessa forma, se evitaria o problema da multicolinearidade.

Exemplificando, uma aplicação utilizando F_1, F_2, F_3, F_4 e F_5 como variáveis independentes poderia ser a construção de um modelo de regressão para geração e atração de viagens.

3.3.2 - Aplicação 2

Para essa aplicação utilizou-se as mesmas variáveis e os mesmos bairros da aplicação 3. O objetivo, agora, é usar a Análise Fatorial tipo Q para agrupar os bairros de Campina Grande que apresentem padrões semelhantes de uso de solo. O resultado dessa aplicação pode ser usado como um critério para definir um novo zoneamento da cidade, segundo o padrão do uso do solo.

Sejam:

B1 = vetor-coluna dos dados do bairro 1

B2 = vetor-coluna dos dados do bairro 2

⋮ ⋮

B54= vetor-coluna dos dados do bairro 54

Para essa Análise Fatorial, utilizou-se as variáveis:

BB 1 = B 1/TOTAL 1

BB 2 = B 2/TOTAL 2

⋮ ⋮

BB54 = B54/TOTAL 54

As grandezas TOTAL 1, TOTAL 2,..., TOTAL 54 significam o valor da área total de cada bairro.

Os resultados dessa Análise Fatorial são mostrados nas Tabelas 14 a 17.

A Tabela 14 mostra que com os onze fatores iniciais já foi possível explicar 100% da variância dos dados originais. Entretanto, somente os seis fatores iniciais tomarão parte na solução final, em virtude de corresponderem a autovalores maiores que o adotado como mínimo. A Tabela 14 mostra, ainda, que os fatores da solução final explicam 95,6% da variância total dos dados.

Após a rotação varimax, tem-se os carregamentos dos fatores, conforme a Tabela 15.

As comunalidades das variáveis são mostradas na Tabela 16.

Para o cálculo dos escores dos fatores usa-se os elementos da Tabela 17.

Partindo-se da Tabela 15 pode-se construir a Tabela 18.

Tabela 18 - Fatores mais importantes na composição da variância das variáveis.

F_1	BB3, BB5, BB8, BB9, BB11, BB13, BB15, BB16, BB17, BB18, BB19, BB21, BB23, BB29, BB30, BB31, BB33, BB39, BB40, BB41, BB42, BB43, BB 44, BB48, BB49, BB51, BB53.
F_2	BB6, BB7, BB12, BB14, BB34, BB35, BB36, BB38, BB45, BB46, BB50, BB52, BB54.
F_3	BB1, BB2, BB4, BB20, BB22, BB25, BB26, BB27.
F_4	BB24, BB28, BB32, BB37.
F_5	BB10.

Tabela 14 - Numeração dos fatores, autovalores e porcentagem da variância do conjunto das variáveis explicada por cada fator.

Fator	Autovalor	%da Variância	%Acumulada da Variância
1	27.11[illegible]55	50.2	50.2
2	11.77057	21.9	72.[illegible]
3	6.63826	12.3	84.[illegible]
4	3.12819	5.8	90.1
5	1.97277	3.7	93.8
6	1.00546	1.9	95.6
7	0.76839	1.4	97.1
8	0.71[illegible]98	1.3	98.4
9	0.41943	0.8	99.2
10	0.30569	0.6	99.7
11	0.1[illegible]642	0.3	100.0
12	0.00758	0.0	100.0
13	0.00182	0.0	100.0
14	0.00003	0.0	100.0
15	0.00078	0.0	100.0
16	0.00002	0.0	100.0
17	0.00001	0.0	100.0
18	0.00001	0.0	100.0
19	0.00000	0.0	100.0
20	0.00000	0.0	100.0
21	0.00000	0.0	100.0
22	0.00000	0.0	100.0
23	0.00000	0.0	100.0
24	0.00000	0.0	100.0
25	0.00000	0.0	100.0
26	0.00000	0.0	100.0
27	0.00000	0.0	100.0
28	0.00000	0.0	100.0
29	0.00000	0.0	100.0
30	0.00000	0.0	100.0
31	-0.00000	-0.0	100.0
32	-0.00000	-0.0	100.0
33	-0.00000	-0.0	100.0
34	-0.00000	-0.0	100.0
35	-0.00000	-0.0	100.0
36	-0.00000	-0.0	100.0
37	-0.00000	-0.0	100.0
38	-0.00000	-0.0	100.0
39	-0.00000	-0.0	100.0
40	-0.00000	-0.0	100.0
41	-0.00000	-0.0	100.0
42	-0.00000	-0.0	100.0
43	-0.00000	-0.0	100.0
44	-0.00000	-0.0	100.0
45	-0.00000	-0.0	100.0
46	-0.00000	-0.0	100.0
47	-0.00000	-0.0	100.0
48	-0.00000	-0.0	100.0
49	-0.00000	-0.0	100.0
50	-0.00000	-0.0	100.0
51	-0.00000	-0.0	100.0
52	-0.00001	-0.0	100.0
53	-0.00001	-0.0	100.0
54	-0.00003	-0.0	100.0

Tabela 15 - Matriz dos carregamentos dos fatores após rotação varimax.

Variável	Fator 1	Fator 2	Fator 3	Fator 4	Fator 5	Fator 6
B01	-0.03042	-0.11603	0.87089	-0.09345	0.03902	0.38394
B02	0.26271	-0.07136	0.03944	0.18906	-0.11962	-0.13190
B03	-0.62854	-0.01093	0.20217	0.45405	0.54492	0.17713
B04	-0.11934	-0.13834	0.81986	0.33470	-0.00542	0.08374
B05	-0.92156	-0.00365	0.16634	0.24804	0.06547	0.15932
B06	0.57959	0.58241	0.01229	0.19126	0.53559	0.04824
B07	-0.02532	0.97517	-0.09258	0.12214	-0.02512	0.00305
B08	0.88804	0.44519	0.02011	0.05190	0.03765	-0.01542
B09	0.93744	-0.00079	0.06690	0.11826	0.05682	-0.01137
B010	0.07046	0.08084	0.15187	0.08917	0.92583	-0.01955
B011	0.85041	-0.02333	0.05956	0.33038	0.37762	-0.07597
B012	0.52902	0.81686	0.16414	-0.08981	0.02016	-0.00815
B013	0.90594	-0.00284	0.02742	0.14660	-0.04209	-0.15475
B014	0.56801	0.81795	-0.03020	0.00273	0.03169	0.00871
B015	0.72370	-0.07021	0.15760	0.41266	0.43087	0.08791
B016	0.98365	0.02950	-0.01833	0.10163	-0.02899	0.00271
B017	0.66059	-0.01419	0.35031	0.64165	-0.07479	0.03214
B018	0.94899	0.01265	0.09112	0.23183	0.11646	-0.03562
B019	0.99504	0.03824	0.03784	0.04910	0.02184	0.00653
B020	0.33057	-0.01668	0.91310	0.10204	-0.00195	-0.07469
B021	0.99708	0.06053	0.01476	0.01462	0.00081	0.00512
B022	0.46711	-0.01901	0.80255	0.32776	0.04525	-0.06667
B023	0.86678	0.08719	0.32212	0.34645	0.08318	-0.02693
B024	0.23398	-0.06363	0.18768	0.93271	0.00304	0.07263
B025	0.26737	-0.05359	0.89703	0.23876	0.10265	-0.05311
B026	-0.12553	-0.16678	0.75100	0.09373	0.30129	-0.05250
B027	-0.01723	-0.05519	0.96831	0.15036	0.01281	0.13797
B028	0.03246	-0.08810	0.34018	0.87919	0.08116	0.57611
B029	0.87477	0.43614	0.12836	0.14285	0.03725	0.02047
B030	0.88208	0.08030	0.43501	0.08174	0.01283	-0.02835
B031	0.65899	0.05349	0.42320	0.55642	0.07161	0.16896
B032	0.37512	0.11319	0.23886	0.87747	0.12269	0.04416
B033	0.01868	-0.00331	0.09283	0.54755	-0.01740	-0.06999
B034	-0.05242	0.99136	-0.06564	-0.03924	0.02545	0.07175
B035	-0.10685	0.99345	-0.07461	-0.02556	0.01343	0.00936
B036	0.45771	0.88265	-0.04784	-0.04091	0.04116	0.02049
B037	0.40621	-0.02907	0.27013	0.86468	0.03045	-0.03965
B038	0.47452	0.87593	-0.00972	-0.02441	0.03337	-0.00789
B039	0.99654	0.03717	0.02167	-0.00655	0.01986	-0.00691
B040	0.91264	0.00992	-0.04990	0.13193	-0.03583	0.53489
B041	0.97648	0.15084	0.00811	0.11837	0.01366	0.01574
B042	0.87832	0.02665	0.18819	0.00321	-0.00951	0.02668
B043	0.72139	0.51924	0.06620	0.41261	-0.02214	0.02322
B044	0.96515	0.00413	0.07307	0.22342	0.02751	-0.03905
B045	0.99622	0.00944	-0.00444	0.02760	-0.01072	0.00893
B046	-0.10654	0.99941	-0.07933	-0.02603	0.01123	0.00712
B047	-0.01613	-0.03607	0.44511	0.88746	0.14953	-0.16094
B048	0.92260	-0.01547	0.12563	0.22903	0.08541	0.11547
B049	0.99491	0.05327	0.06092	0.01764	0.03010	0.01141
B050	0.03926	0.99135	-0.05938	0.05256	0.01741	0.00330
B051	0.92724	-0.00128	-0.05771	-0.07789	0.53217	0.03452
B052	-0.11451	0.78041	-0.10045	-0.08352	-0.11746	-0.21925
B053	0.92237	0.33152	0.05714	0.07422	0.14420	0.06373
B054	-0.04739	0.98403	-0.08492	-0.05457	0.13561	0.02013

Tabela 16 - Comunalidade das Variáveis.

Variável	Comunalidade
B01	0.93340
B02	0.94431
B03	1.00070
B04	0.91888
B05	0.97144
B06	0.98733
B07	0.97874
B08	0.98157
B09	0.90344
B010	0.98036
B011	0.98481
B012	0.98740
B013	0.97022
B014	0.99370
B015	0.91697
B016	0.97006
B017	0.95745
B018	0.99504
B019	0.99543
B020	0.99465
B021	0.99035
B022	0.97657
B023	0.98452
B024	0.97204
B025	0.97500
B026	0.73960
B027	0.99278
B028	0.92423
B029	0.99589
B030	0.93141
B031	0.95081
B032	0.99753
B033	0.97620
B034	0.99710
B035	0.99351
B036	0.99466
B037	0.98389
B038	0.99423
B039	0.99543
B040	0.96777
B041	0.99077
B042	0.99405
B043	0.96132
B044	0.99613
B045	0.99599
B046	0.99753
B047	0.99693
B048	0.94048
B049	0.99773
B050	0.99709
B051	0.97009
B052	0.70111
B053	0.99439
B054	0.99554

Tabela 17 - Matriz dos coeficientes para o cálculo dos escores dos fatores.

VARIÁVEL	FATOR 1	FATOR 2	FATOR 3	FATOR 4	FATOR 5	FATOR 6
B81	-0.03757	0.01634	0.15633	-0.12156	-0.01210	0.36701
B82	0.03311	0.03642	0.14981	-0.03189	-0.03749	-0.15694
B83	-0.03459	-0.01181	-0.02167	0.04209	-0.21955	0.11146
B84	-0.12675	0.00764	0.12231	0.02136	-0.03039	-0.00188
B85	0.04069	-0.01191	-0.00234	-0.00486	-0.02128	0.13661
B86	-0.00383	0.03803	-0.01624	-0.00983	0.23151	0.01210
B87	-0.02371	0.09168	-0.00786	0.04787	-0.02760	0.03821
B88	0.04341	0.02572	0.00109	-0.02766	-0.01966	-0.02530
B89	0.03053	-0.01615	-0.00547	-0.02108	-0.03093	-0.03155
B810	-0.03434	-0.00242	0.00194	-0.02946	0.45386	-0.07238
B811	0.02519	-0.02034	-0.03035	0.03083	0.14847	-0.12204
B812	0.02032	0.06967	0.05029	-0.05802	-0.01845	-0.00543
B813	0.05245	-0.01640	-0.00976	0.00443	-0.05113	-0.16362
B814	0.01924	0.04577	0.00515	-0.02132	-0.01426	0.00920
B815	0.00809	-0.01926	-0.02387	0.03819	0.16766	0.03075
B816	0.05611	-0.01467	-0.01815	-0.01647	-0.05299	-0.01033
B817	0.01022	-0.00747	-0.04546	0.13518	-0.03943	-0.00936
B818	0.04127	-0.01477	-0.01490	0.02091	0.01451	-0.00965
B819	0.05799	-0.01436	-0.00439	-0.03794	-0.02724	-0.00787
B820	0.01973	0.01011	0.16096	-0.06248	-0.00323	-0.10305
B821	0.06001	-0.01223	-0.00487	-0.04449	-0.03486	-0.00172
B822	0.00731	0.00615	0.11685	-0.05197	-0.03134	-0.10927
B823	0.03263	-0.00002	0.02594	0.02403	-0.03477	-0.06298
B824	-0.02276	-0.00364	-0.04906	0.21410	-0.05225	0.01079
B825	-0.00560	0.00756	0.13906	-0.01433	0.01376	-0.09682
B826	-0.02460	-0.00249	0.12465	-0.04385	0.13948	-0.08546
B827	-0.01603	0.01612	0.16401	-0.05063	-0.02546	0.10504
B828	-0.03820	0.00586	-0.01073	0.11167	-0.02277	0.51628
B829	0.03707	0.02814	0.01222	-0.01280	-0.02740	0.00219
B830	0.04630	-0.00003	0.06985	-0.05524	-0.03595	-0.05091
B831	0.02675	0.02509	0.02265	0.06748	-0.02325	0.11571
B832	-0.02624	0.01241	-0.03480	0.13655	0.00318	-0.02077
B833	0.02576	-0.01065	-0.03841	0.10311	-0.05828	-0.10975
B834	-0.02073	0.05324	0.00862	-0.00186	0.00122	0.03256
B835	-0.02386	0.09566	0.00800	0.00433	-0.00325	0.02159
B836	0.01292	0.07328	0.00720	-0.02585	-0.00506	0.02494
B837	-0.01728	-0.00058	-0.02651	0.13686	-0.03833	-0.10045
B838	0.01365	0.07304	0.01346	-0.02362	-0.00961	-0.00467
B839	0.06093	-0.01476	-0.00226	-0.05044	-0.02502	-0.01745
B840	0.04056	-0.00848	-0.03736	-0.02616	-0.07648	0.51444
B841	0.05183	-0.00299	-0.01542	-0.01577	-0.03498	-0.00049
B842	0.05800	-0.01140	0.02759	-0.06197	-0.04464	0.01292
B843	0.01657	0.03874	-0.02910	0.07715	-0.06126	-0.00241
B844	0.04702	-0.00778	-0.01455	0.00911	-0.02982	-0.06310
B845	-0.01915	0.09181	-0.00052	0.02077	-0.01801	0.01860
B846	-0.02353	0.09349	-0.00724	0.00983	-0.00111	0.01962
B847	-0.04500	0.00700	0.01411	0.17861	0.04040	-0.22218
B848	0.04197	-0.01471	-0.00754	-0.00386	-0.00640	0.08439
B849	0.05059	-0.01227	0.00272	-0.04871	-0.02289	-0.00199
B850	-0.01466	0.09092	0.00309	0.02076	-0.01337	0.00916
B851	0.03736	-0.02640	-0.02337	-0.07858	0.24090	0.00558
B852	-0.01059	0.07261	0.01196	0.01392	-0.04877	-0.19530
B853	0.04184	0.01454	-0.00746	-0.03617	0.02942	0.04412
B854	-0.02278	0.09009	0.00506	-0.00594	0.05984	0.02756

Uma análise detalhada das características dos fatores da solução final revela as variáveis originais que predominam na formação dos fatores. A Tabela 19 apresenta os fatores e suas características.

Tabela 19 - Fatores e suas definições em termos das variáveis originais predominantes, em valores relativos.

Fatores	Variáveis Originais Predominantes	Características
F_1	ND	bairros já loteados com baixo índice de uso do solo
F_2	ND + A	bairros com baixo índice de uso do solo, parcialmente loteado e com uso do solo para agricultura
F_3	R + CR + P	bairros residencial e/ou comercial e/ou de uso público
F_4	R + (P+EA) + ND	bairro residencial com área para lazer e com loteamento
F_5	I	bairro indústrial

A Tabela 19 mostra o resultado do agrupamento dos bairros com padrões semelhantes de uso do solo.

Na Tabela 18, tem-se os bairros pertencentes a cada um dos grupos e, identificando-se aqueles bairros, espacialmente vizinhos, pertencentes a um mesmo grupo, pode-se juntá-los para formar novas unidades de agregação, de mesmo padrão de uso do solo. Então, pode-se usar esse resultado para, por exemplo, estudar o processo de geração de viagens dessas novas unidades de agregação.

3.4. - Análise Discriminante. Aplicações e resultados

3.4.1. - Aplicação 1

O objetivo dessa aplicação é, dados três grupos de observações formados por pessoas que exercem atividade trabalho no setor secundário, no terciário e em ambos, no secundário e terciário, verificar se esses grupos apresentam comportamentos distintos, usando a Análise Discriminante.

Para essa aplicação, observou-se em que setor da economia era exercida a atividade trabalho das pessoas ocupadas, residentes em Campina Grande, e agrupou-se, separadamente, os indivíduos com atividade trabalho no setor secundário, no terciário e aqueles com atividade trabalho nos setores secundário e terciário.

Em seguida, dividiu-se o período de 24 horas em cinco intervalos de tempo, 5hs às 7hs, 7hs às 9hs, 9hs às 13hs, 13hs às 15hs e 15hs às 24 hs, e considerou-se os seguintes modos de transporte: ônibus, carro-motorista, carro-acompanhante, taxi, a pé e bicicleta. Os modos a pé e bicicleta foram agrupados passando a formar um único modo.

Para cada indivíduo, construíu-se um perfil do seu comportamento, em relação aos transportes. Esse perfil era representado por um vetor-linha de 30 componentes, isto é, 30 variáveis, onde as primeiras 25 variáveis correspondiam aos cinco intervalos de tempo para cada modo de transporte, e as 5 últimas à duração da atividade trabalho do indíviduo, em cada intervalo. Inicialmente, o vetor perfil era zerado e, para cada viagem casa-trabalho realizada, alocava-se o

número 1 à posição correspondente, simultaneamente, ao modo de transporte utilizado e ao intervalo de tempo no qual a viagem teve início. Ao intervalo de tempo no qual o indivíduo começou a sua atividade, alocou-se o valor da duração daquela atividade. Dessa forma obteve-se os três grupos de observação.

A Tabela 20 mostra o número de casos em cada grupo.

Tabela 20 - Número de casos por grupo

Grupo	Número de casos
1	249
2	2208
3	7
Total	2464

Com base no número de casos apresentados na Tabela 20, pode-se questionar a representatividade do grupo 3, quanto ao comportamento das pessoas com atividade trabalho, nos setores secundário e terciário, visto que esse grupo contém um número pequeno de casos.

Ao término do processo da Análise Discriminante, verificou-se, com base na Tabela 21, que a atividade trabalho é um fator significante para a distinção dos grupos.

Tabela 21 - Lambda de Wilks e valor do $F_{equivalente}$

		Graus de Liberdade	significância
Lambda de Wilks	0,9673821		
$F_{equivalente}$	3,725320	20 e 4904	0.0000

O lambda de Wilks ou o $F_{equivalente}$ calculado a partir desse vetor dá indicativos sobre o poder discriminatório do conjunto de variáveis incluídas na análise. Para saber se o valor de F é significativo, testa-se com o $F_{crítico}$ para os números de graus de liberdade fornecidos na tabela 21 e para um nível de significância estabelecido.

Para esta aplicação, foi encontrado que o valor de F é estatisticamente significante a 5% de nível de significância, em virtude de o $F_{equivalente}$ ser maior do que o $F_{crítico}$ para 20 e 4904 graus de liberdade. Isso, em outras palavras, equivale a dizer que o critério usado na construção dos grupos, realmente, produz diferença significativa nos grupos.

Pode-se, também, verificar a significância estatística da diferença entre pares de grupos, e isso é mostrado na Tabela 22.

Tabela 22 - Valores de F e significância entre pares de grupos. Cada valor F tem 11 e 2451 graus de liberdade.

Grupos	1	2
2	6,2743*	
	0,0000**	
3	1,5003	1,1880
	0,1242	0,2895

* - Valor de F

** - Significância entre pares de grupos

O valor de F apresentado na Tabela 22 é o $F_{calculado}$

para cada par de grupos. Admite-se as hipóteses:

H_0: Os grupos não são diferentes

H_1: Os grupos são diferentes

Compara-se o $F_{calculado}$ com o $F_{crítico}$ tirado de uma tabela F para 11 e 2451 graus de liberdade e um nível de significância desejado; nesse trabalho está sendo considerado um nível de significância de 5%. Se $F_{calculado} > F_{crítico}$, rejeita-se H_0 e tem-se que os grupos são diferentes; se $F_{calculado} < F_{crítico}$, aceita-se H_0 e, nesse caso, os grupos não são significativamente diferentes.

Nessa aplicação tem-se $F_{crítico} = 1,79$ e, portanto, apenas os grupos 1 e 2 são significativamente diferentes, entre si.

A significância entre pares de grupos na Tabela 22 indica o nível de significância a partir do qual se tem $F_{calculado} > F_{crítico}$.

A Análise Discriminante selecionou como variáveis discriminatórias as variáveis originais mostradas na Tabela 23.

Na Tabela 23 não aparecem variáveis que representam o uso de modos de transporte no 2º intervalo de tempo. No entanto, esse tipo de variável é importante no comportamento dos grupos porque se trata de valores para o intervalo entre 7hs e 9hs, horário em que há um grande número de viagens para o trabalho, principalmente, para o setor terciário.

O fato de a Análise Discriminante Stepwise trabalhar com o objetivo de encontrar a máxima discriminação entre os grupos, excluindo as variáveis que não contribuem significativamente para tal objetivo, fez com que fosse excluida essa

Tabela 23 - Variáveis discriminatórias

Variável	Modo	Intervalo de Tempo
X_3	Ônibus	3º
X_6	Carro-Motorista	1º
X_8	Carro-Motorista	3º
X_{11}	Carro-Motorista	1º
X_{16}	Taxi	1º
X_{21}	Pé+Bicicleta	1º
X_{23}	Pé+Bicicleta	3º
X_{26}	Duração	1º
X_{27}	Duração	2º
X_{29}	Duração	4º
X_{30}	Duração	5º

variável, importante no comportamento dos grupos. Essa é uma limitação do uso da análise discriminante Stepwise nêsse tipo de aplicação. A garantia da inclusão dessas variáveis na Análise Discriminante pode ser obtida, pela realização de uma Análise Discriminante, Método Direto, forçando-se a entrada de todas as variáveis na solução final.

A Análise Discriminante realizada nessa aplicação deve fornecer, no máximo, duas funções discriminantes. Essas funções são combinações lineares das variáveis discriminatórias e o valor numérico ou escore de cada observação, em termos de função discriminante, é obtido entrando-se com os valores observados das variáveis discriminatórias, na forma padronizada, nas funções discriminantes. A média aritmética

dos escores das observações de um grupo, em uma função, é chamada média do grupo relativo aquela função. Para um mesmo grupo, a média dos escores em todas as funções é chamada centróide do grupo.

A Tabela 24 dá algumas informações sobre as funções discriminantes obtidas, e a Tabela 25 mostra os coeficientes das variáveis discriminatórias, em cada função discriminante.

Tabela 24 - Funções discriminantes

	Função	
	1	2
Autovalor	0.02842	0.00515
% da Variância	84.66	13.34
% Acumulada da Variância	84.66	100.00
Lambda de Wilks	0.9673821	0.0048769
Qui-quadrado	81.445	12.615
Graus de Liberdade	22	10
Significância	0.0000	0.2460

Tabela 25 - Coeficientes, na forma padronizada, das funções discriminantes.

Variável	Função 1	Função 2
X_3	-0.14137	0.24622
X_6	-0.49900	0.34869
X_8	-0.09623	-0.58622
X_{11}	-0.18422	0.10923
X_{16}	-0.21263	0.05284
X_{21}	-0.33142	0.11730
X_{23}	0.46099	0.12840
X_{26}	0.98837	0.06180
X_{27}	-0.25150	0.28344
X_{29}	0.06199	0.41729
X_{30}	0.06315	-0.45406

Da Tabela 25, tem-se a função discriminante 1 como:

$$D_1 = -0,14137X_3 - 0,49900X_6 - 0,09623X_8 - \ldots + 0,06315X_{30} \qquad (48)$$

A função discriminante 2 é obtida de maneira semelhante à (48).

Através dos coeficientes de (48) pode-se avaliar a importância de cada variável discriminatória como expressão da diferença de comportamento entre os grupos. Aquela variável de maior coeficiente em (48) dá maior contribuição para a discriminação dos grupos, isto é, expressa melhor a diferença de comportamento entre os grupos.

Conforme mostrado no enfoque matemático da Análise Discriminante, a interpretação Bayesiana permite alocar uma dada observação a um dos grupos já definidos. A alocação se dá para o grupo no qual a observação tem a maior probabilidade de pertencer.

Uma outra forma de se fazer a classificação de uma observação é com o uso das funções de classificação, também chamadas de funções discriminantes lineares de Fisher.

A Tabela 26 mostra os coeficientes das funções de classificação, para cada grupo.

Tabela 26 - Coeficientes das funções de classificação.

A função de classificação para o grupo 1 é dada por:

$$FC_1 = 1{,}2513X_3 - 1{,}3961X_6 + 1{,}4584X_8 - \ldots - 2{,}0178 \qquad (49)$$

As demais funções de classificação FC_2 e FC_3 são construídas de maneira similar à (49)

Entrando-se com os valores das variáveis discriminatórias de uma observação nas funções FC_1, FC_2 e FC_3 obtem-se

os escores daquela observação nessas funções. A observação, então, será alocada para o grupo correspondente à função de mais alto escore.

Sob a hipótese de uma distribuição normal multivariada, os escores de classificação podem ser convertidos em probabilidades de a observação pertencer a um grupo. Assim, alocar a observação para o grupo de mais alto escore de classificação é equivalente a se alocar a observação para o grupo no qual ela tem a maior probabilidade de pertencer. Isso vai ao encontro da abordagem Bayesiana na Análise Discriminante. É sempre conveniente usar a abordagem Bayesiana na classificação das observações quando os prejuízos causados por uma classificação errada são muito grandes, ou quando os grupos são de tamanhos muito diferentes ou, ainda, quando se deseja tirar proveito do conhecimento das probabilidades a priori de uma observação pertencer a um grupo[2].

Para essa aplicação é conveniente usar a abordagem Bayesiana na classificação dos grupos devido a diferença acentuada nos tamanhos dos grupos.

A Tabela 27 mostra a classificação Bayesiana para uma parte das observações.

Como resultado da classificação de todas as observações, tem-se:

Tabela 27 - Classificação das observações

Arquivo	Nº do caso	Grupo real		Maior Probabilidade Grupo	P(X/G)	P(G/X)	Segunda maior Probabilidade Grupo	P(G/X)	Escores Discriminante	
CM4T81.D	241	1	***	2	0.8796	0.3911	1	0.3062	-0.2390	0.4658
CM4T81.D	242	1		1	0.9939	0.4729	2	0.3956	0.5437	0.0026
CM4T81.D	243	1		1	0.2951	0.6741	2	0.2444	2.0334	0.3208
CM4T81.D	244	1	***	2	0.8796	0.3911	1	0.3062	-0.2390	0.4658
CM4T81.D	245	1	***	2	0.8796	0.3911	1	0.3062	-0.2390	0.4658
CM4T81.D	246	1		1	0.5617	0.6282	2	0.3317	1.3983	-0.5700
CM4T81.D	247	1	***	2	0.7069	0.5492	1	0.3493	-0.5610	-0.6532
CM4T81.D	248	1	***	2	0.8796	0.3911	1	0.3062	-0.2390	0.4658
CM4T81.D	249	1	***	3	0.7350	0.7564	2	0.1260	0.0047	2.0421
CM4T81.D	250	2		2	0.8796	0.3911	1	0.3062	-0.2390	0.4658
CM4T81.D	251	2		2	0.8796	0.3911	1	0.3062	-0.2390	0.4658
CM4T81.D	252	2		2	0.6083	0.5548	1	0.3754	-0.4373	-0.9274
CM4T81.D	253	2		2	0.6689	0.5465	1	0.3595	-0.4453	-0.7805
CM4T81.D	254	2		2	0.8796	0.3911	1	0.3062	-0.2390	0.4658
CM4T81.D	255	2		2	0.5974	0.5607	1	0.3646	-0.4044	-0.9192
CM4T81.D	256	2		2	0.6979	0.5467	1	0.3619	-0.4016	-0.7476
CM4T81.D	257	2		2	0.4029	0.5638	1	0.3963	-0.3550	-1.3210
CM4T81.D	258	2		2	0.5216	0.5715	1	0.3850	-0.5360	-1.0411
CM4T81.D	259	2		2	0.5687	0.5630	1	0.3596	-0.5561	-0.9434
CM4T81.D	260	2		2	0.8796	0.3911	1	0.3062	-0.2390	0.4658
CM4T81.D	261	2		2	0.6733	0.5557	1	0.3469	-0.5921	-0.7145
CM4T81.D	262	2		2	0.8796	0.3911	1	0.3062	-0.2390	0.4658
CM4T81.D	263	2	***	1	0.6733	0.5893	2	0.3888	0.9439	-0.7507
CM4T81.D	264	2		2	0.8796	0.3911	1	0.3062	-0.2390	0.4658
CM4T81.D	265	2		2	0.8796	0.3911	1	0.3062	-0.2390	0.4658
CM4T81.D	266	2		2	0.8796	0.3911	1	0.3062	-0.2390	0.4658
CM4T81.D	267	2		2	0.7446	0.5375	1	0.3684	-0.4254	-0.6794
CM4T81.D	268	2	***	3	0.0206	0.9740	1	0.0140	0.3146	4.3462
CM4T81.D	269	2		2	0.8796	0.3911	1	0.3062	-0.2390	0.4658
CM4T81.D	270	2		2	0.8663	0.4644	1	0.2975	-0.9831	0.0865
CM4T81.D	271	2		2	0.8925	0.4447	1	0.3006	-0.4905	0.1385
CM4T81.D	272	2		2	0.5849	0.5657	1	0.3595	-0.5507	-0.9159
CM4T81.D	273	2		2	0.9262	0.4817	1	0.4072	-0.0612	-0.3979
CM4T81.D	274	2		2	0.1531	0.5804	1	0.4008	-0.3550	-1.9204
CM4T81.D	275	2		2	0.7993	0.5296	1	0.3609	-0.4401	-0.5541
CM4T81.D	276	2		2	0.8496	0.4732	1	0.2959	-0.6243	0.0411
CM4T81.D	277	2		2	0.4750	0.5979	1	0.3023	-0.9637	-0.8212
CM4T81.D	278	2		2	0.8796	0.3911	1	0.3062	-0.2390	0.4658
CM4T81.D	279	2		2	0.6218	0.5629	1	0.3528	-0.5799	-0.8281
CM4T81.D	280	2		2	0.1983	0.5124	1	0.4706	-0.1525	-1.7932
CM4T81.D	281	2		2	0.8796	0.3911	1	0.3062	-0.2390	0.4658
CM4T81.D	282	2		2	0.8796	0.3911	1	0.3062	-0.2390	0.4658
CM4T81.D	283	2		2	0.5933	0.5574	1	0.3740	-0.4515	-0.9430
CM4T81.D	284	2	***	3	0.0073	0.9911	2	0.0052	-0.8809	4.4285
CM4T81.D	285	2		2	0.8857	0.4675	1	0.4149	-0.0441	-0.4939
CM4T81.D	286	2		2	0.8097	0.4991	1	0.2925	-0.7038	-0.3465
CM4T81.D	287	2		2	0.8796	0.3911	1	0.3062	-0.2390	0.4658
CM4T81.D	288	2		2	0.9678	0.4781	1	0.3731	-0.2141	-0.2068

Tabela 28 - Resultados da classificação

Grupo real	Número de casos	Grupo estimado		
		1	2	3
1	249	89	144	16
		35,7%	57,8%	6,4%
2	2208	421	1650	137
		19,1%	74,7%	6,2%
3	7	0	5	2
		0%	71,4%	28,6%

Com esse resultado verificou-se que existem pessoas de um grupo que se comportam como pertencentes a outro grupo.

Para essa aplicação o número total de casos corretamente classificados foi de 70,66%.

3.4.2 - Aplicação 2

O objetivo dessa aplicação é, para as pessoas com atividade trabalho no setor secundário e utilizando-se as mesmas variáveis da aplicação 3.4.1, fazer uma Análise Discriminante para verificar se o nível de renda é um fator da distinção significativa dos grupos.

Nesse caso, considerou-se quatro níveis de renda, em valores de 1978, e os grupos ficaram constituídos por:

Grupo 1 - NR_1: Renda < 1111

Grupo 2 - NR_2: 1111 < Renda < 2222

Grupo 3 - NR_3: 2222 < Renda < 3333

Grupo 4 - NR_4: Renda > 3333

onde NR_1, NR_2, NR_3 e NR_4 são os níveis de renda de 1 a 4, respectivamente.

O número de casos em cada grupo é mostrado na Tabela 29.

Tabela 29 - Número de casos por grupo

Nível de renda	Número de casos
1	26
2	69
3	48
4	97
TOTAL	240

Conforme mostrado na Tabela 29, não há grandes diferenças nos tamanhos dos grupos, tornando-se, praticamente, impossível avaliar-se inicialmente, se os grupos podem, ou não, ser representativos do comportamento das pessoas de cada nível de renda.

As informações acerca do critério nível de renda como fator da distinção dos grupos são dadas na Tabela 30.

Tabela 30 - Lambda de Wilks e $F_{equivalente}$.

		Graus de Liberdade	Significância
Lambda Wilks	0,8018216		
Fequivalente	2,191764	24 e 664,8	0,0009

A coluna referente à significância indica que F equivalente é maior do que qualquer valor de F com 24 e 664,8 graus de liberdade a um nível de significância maior do que 0.0009%. Portanto, como se está usando um nível de significância de 5%, concluí-se que o fator nível de renda distingue, significativamente, os grupos.

A tabela 31 mostra os valores de F e as significâncias entre pares de grupos.

Tabela 31 - Valores de F, para 8 e 229 graus de liberdade, e significância para os pares de grupos.

Grupo	1	2	3
2	1.6331 0.1164		
3	1.5783 0.1322	0.90482 0.5132	
4	2.0147 0.0457	4.0744 0.0001	2.0749 0.0392

Baseando-se na tabela 31, verifica-se que o grupo 4 difere dos demais grupos a 5% de nível de significância. Porém, para esse nível de significância, os outros pares de grupos não apresentam diferenças significantes. Esse fato poderia ser usado para se tentar redefinir os grupos, juntando-se aqueles com diferenças menos significantes. Por exemplo, poderia se juntar os grupos 2 e 3 dessa aplicação e fazer Análise Discriminante com os grupos 1, (2+3) e 4, isto

é, com três grupos.

As variáveis discriminatórias são mostrada na tabela 32.

Tabela 32 - Variáveis discriminatórias

Variável	Modo	Intervalo
X_1	Ônibus	1º
X_3	Ônibus	3º
X_6	Carro-Motorista	1º
X_7	Carro-Motorista	2º
X_{12}	Carro-Acompanhante	2º
X_{21}	Pé+Bicicleta	1º
X_{22}	Pé+Bicicleta	2º
X_{27}	Duração	2º

A tabela 32 mostra que as variáveis que discriminam o comportamento dos indívíduos, segundo o nível de renda, são os modos de transporte, exceto taxi, usados, principalmente, no período entre 5 e 9hs, bem como a duração da atividade no 2º intervalo. Esse resultado é coerente, pois, é de se supor que o modo de transporte utilizado pelas pessoas ocupadas no secundário, em Campina Grande, seja um indicador do nível de renda dessas pessoas.

A tabela 33 fornece os coeficientes das funções discriminantes.

Tabela 33 - Coeficientes, na forma padronizada, das funções discriminantes.

Variáveis	Função 1	Função 2	Função 3
X_1	0.14008	0.37929	0.83615
X_3	0.38398	0.08126	0.14618
X_6	0.30927	0.11965	0.14654
X_7	0.37472	0.27236	0.25502
X_{12}	0.41472	0.05356	0.28903
X_{21}	0.50813	0.58468	0.04371
X_{22}	0.15318	1.02007	0.38854
X_{27}	0.26781	0.96556	0.24579

Então, a função discriminante 1 é dada por:

$$D_1 = - 0,14008X_1+0,38398X_3+...+0,26781X_{27} \qquad (48)$$

A tabela 34 mostra a classificação Bayesiana para uma parte das observações.

Os resultados dessa classificação são mostrados na Tabela 35.

Tabela 34 - Classificação das observações

Arquivo	Nº do caso	Grupo		Maior Probabilidade Grupo	P(X/G)	P(G/X)	Segunda maior Probabilidade Grupo	P(G/X)	Escores Discriminantes		
[illegible]	1	1		1	0.9258	0.2876	2	0.2619	-0.1991	-0.1623	-0.6969
[illegible]	2	1		1	0.9258	0.2876	2	0.2619	-0.1991	-0.1623	-0.6969
[illegible]	3	1		1	0.3149	0.6364	3	0.1328	0.7646	-2.3650	-0.3381
[illegible]	4	1		1	0.9258	0.2876	2	0.2619	-0.1991	-0.1623	-0.6969
[illegible]	5	1		1	0.9258	0.2876	2	0.2619	-0.1991	-0.1623	-0.6969
[illegible]	6	1		1	0.3149	0.6364	3	0.1328	0.7646	-2.3650	-0.3381
[illegible]	7	1	***	3	0.3021	0.4854	2	0.1337	0.4417	-0.4919	2.1500
[illegible]	8	1		1	0.3263	0.6304	3	0.1355	0.7719	-2.3384	-0.3157
[illegible]	9	1		1	0.9258	0.2876	2	0.2619	-0.1991	-0.1623	-0.6969
[illegible]	10	1		1	0.9258	0.2876	2	0.2619	-0.1991	-0.1623	-0.6969
[illegible]	11	1		1	0.9258	0.2876	2	0.2619	-0.1991	-0.1623	-0.6969
[illegible]	12	1		1	0.9258	0.2876	2	0.2619	-0.1991	-0.1623	-0.6969
[illegible]	13	1	***	3	0.3021	0.4854	2	0.1337	0.4417	-0.4919	2.1500
[illegible]	14	1		1	0.9258	0.2876	2	0.2619	-0.1991	-0.1623	-0.6969
[illegible]	15	1		1	0.9258	0.2876	2	0.2619	-0.1991	-0.1623	-0.6969
[illegible]	16	1		1	0.3263	0.6304	3	0.1355	0.7719	-2.3384	-0.3157
[illegible]	17	1	***	2	0.6094	0.4500	3	0.3170	-1.5293	0.5634	0.6395
[illegible]	18	1	***	4	0.3407	0.5344	3	0.2034	2.1756	0.3215	0.0906
[illegible]	19	1		1	0.2785	0.6567	3	0.1245	0.7417	-2.4922	-0.4082
[illegible]	20	1		1	0.9258	0.2876	2	0.2619	-0.1991	-0.1623	-0.6969
[illegible]	21	1	***	2	.6094	0.4500	3	0.3170	-1.5293	0.5634	0.6395
[illegible]	22	1		1	0.3382	0.6237	3	0.1385	0.7801	-2.3078	-0.2905
[illegible]	23	1		1	0.9258	0.2876	2	0.2619	-0.1991	-0.1623	-0.6969
[illegible]	24	1		1	0.9258	0.2876	2	0.2619	-0.1991	-0.1623	-0.6969
[illegible]	25	1		1	0.9258	0.2876	2	0.2619	-0.1991	-0.1623	-0.6969
[illegible]	26	1		1	0.9258	0.2876	2	0.2619	-0.1991	-0.1623	-0.6969
[illegible]	27	2	***	3	0.3021	0.4854	2	0.1337	0.4417	-0.4919	2.1500
[illegible]	28	2	***	1	0.9258	0.2876	2	0.2619	-0.1991	-0.1623	-0.6969
[illegible]	29	2	***	3	0.3021	0.4854	2	0.1337	0.4417	-0.4919	2.1500
[illegible]	30	2		2	0.6094	0.4500	3	0.3170	-1.5293	0.5634	0.6395
[illegible]	31	2	***	1	0.9258	0.2876	2	0.2619	-0.1991	-0.1623	-0.6969
[illegible]	32	2		2	0.6094	0.4500	3	0.3170	-1.5293	0.5634	0.6395
[illegible]	33	2	***	1	0.9258	0.2876	2	0.2619	-0.1991	-0.1623	-0.6969
[illegible]	34	2	***	1	0.3959	0.5684	3	0.1511	0.8104	-2.1636	-0.1980
[illegible]	35	2	***	1	0.9258	0.2876	2	0.2619	-0.1991	-0.1623	-0.6969
[illegible]	36	2	***	3	0.3021	0.4854	2	0.1337	0.4417	-0.4919	2.1500
[illegible]	37	2	***	3	0.3021	0.4854	2	0.1337	0.4417	-0.4919	2.1500
[illegible]	38	2	***	1	0.9258	0.2876	2	0.2619	-0.1991	-0.1623	-0.6969
[illegible]	39	2		2	0.6094	0.4500	3	0.3170	-1.5293	0.5634	0.6395
[illegible]	40	2	***	1	0.9258	0.2876	2	0.2619	-0.1991	-0.1623	-0.6969
[illegible]	41	2		2	0.6094	0.4500	3	0.3170	-1.5293	0.5634	0.6395
[illegible]	42	2	***	1	0.9258	0.2876	2	0.2619	-0.1991	-0.1623	-0.6969
[illegible]	43	2	***	1	0.9258	0.2876	2	0.2619	-0.1991	-0.1623	-0.6969
[illegible]	44	2		2	0.6094	0.4500	3	0.3170	-1.5293	0.5634	0.6395
[illegible]	45	2	***	1	0.9258	0.2876	2	0.2619	-0.1991	-0.1623	-0.6969
[illegible]	46	2	***	1	0.9258	0.2876	2	0.2619	-0.1991	-0.1623	-0.6969
[illegible]	47	2		2	0.6094	0.4500	3	0.3170	-1.5293	0.5634	0.6395
[illegible]	48	2	***	1	0.9258	0.2876	2	0.2619	-0.1991	-0.1623	-0.6969

Tabela 35 - Resultado da classificação

Grupo Real	Número de Casos	Grupo estimado			
		1	2	3	4
1	26	21 80.8%	2 7.7%	2 7.7%	1 3.8%
2	69	39 56.5%	22 31.9%	7 10.1%	1 1.4%
3	48	20 41.7%	12 25.0%	11 22.9%	5 10.4%
4	97	53 54.6%	10 10.3%	5 8.2%	26 26.8%

A tabela 35 mostra que a maioria das observações dos grupos 2, 3 e 4 estão alocadas inadequadamente. Daí a porcentagem dos casos corretamente classificados ter sido 33,33%, isto é, muito baixa.

Portanto, em situações semelhantes a essa recomenda-se realizar uma Análise Discriminante para uma nova definição dos grupos, através da junção daqueles grupos que apresentem diferenças menos significantes, quando testados dois a dois.

CAPÍTULO IV

CONCLUSÕES E SUGESTÕES PARA PESQUISAS FUTURAS

4.1 - Conclusões

As conclusões tiradas com o desenvolvimento deste trabalho podem ser enumeradas por:

1 - O conhecimento das abordagens matemáticas das técnicas estatísticas multivariadas, usadas neste trabalho, propicia um mais rápido entendimento das aplicações práticas de cada técnica. No entanto, o analista de transportes não deve superestimar o enfoque matemático em detrimento das relações lógicas aparentes de cada caso. Deve, isto sim, tirar proveito do seu entendimento teórico, acerca da técnica usada, para facilitar a análise dos resultados e a elaboração de conclusões mais detalhadas.

2 - No tocante ao emprego da Análise de Regressão Linear Múltipla em estudos de demanda de viagens, comprova-se que o processo comumente usado, maior R^2 e menor Se, para a escolha de uma equação de regressão linear múltipla como modelo de previsão, não é suficiente. Deve-se, além disso, analisar a qualidade da equação com base nas relações lineares entre as variáveis mostradas na matriz

de correlação, na coerência quanto à magnitude e sinal dos coeficientes de regressão, no sentido lógico da equação, como um todo, e nos testes estatísticos tanto para a equação como para os coeficientes de regressão. A decisão sobre que equação usar como modelo de previsão, fica a critério do analista.

3 - O emprego de técnicas estatísticas multivariadas, como a Análise Fatorial e a Análise Discriminante em análise de transportes, proporcionam um alto grau de detalhamento na análise dos resultados, visto que, permitem, por exemplo, identificar indivíduos de comportamento semelhantes, em relação aos transportes, e agrupá-los segundo um comportamento padrão. Assim, é mais confiável tirar conclusões acerca de grupos, construídos a partir de indivíduos de comportamentos semelhantes, do que se tirar conclusões a respeito de cada indivíduo, partindo-se de grupos não homogêneos, como é feito nos modelos agregados que considera a média da zona como representativa do comportamento de todos os indivíduos da zona. Portanto, os modelos comportamentais ou desagregados para a análise de transportes abordam a complexidade do processo de viagem, pois, a análise das viagens de um indivíduo é feita em termos de uma série de variáveis, representativas do comportamento desse indivíduo, e não de uma só variável como se faz tradicionalmente. Além disso, permite incorporar as variâncias da amostra produzidas por fatores sociais ou locacionais da unidade de agregação considerada.

4.2 - Sugestões para pesquisas futuras

1 - Construir um modelo comportamental de geração- repartição modal integrado, para Campina Grande.

2 - Utilizar a Análise Fatorial e/ou Análise Discriminante para analisar os perfis de atividade das pessoas residentes em Campina Grande.

3 - Construir e avaliar modelos de escolha modal baseados na Análise de Regressão Linear Múltipla, na Análise Discriminante e na abordagem Multinomial Logit.

BIBLIOGRAFIA

01. Draper N.R. e H. Smith - Applied Regression Analysis.

02. Nie, Norman H. - Statistical Package for the Social Sciences. SPSS - Editora McGraw Hill - 1975.

03. Faissol, Esperidião - Tendências Atuais na Geografia Urbana/Regional - 1978.

04. Cooley, Willian W. e Paul R. Lohnes - Multivariate Procedures for the Behavioral Sciences.

05. Van De Geer - Introduction to Multivariate Analysis for the Social Sciences. Freemar e Company - 1971.

06. Rummel, R. J. - Applied Factor Analysis - 1970.

07. McCarthy, Gerald M. - Multiple - Regression Analysis of Household Trip Generation - A Critique.

08. Harman, H. H. - Modern Factor Analysis - 1967.

09. Kassoff e Deutschman - Trip Generation: A Critical Appraisal

10. Leal, José E. - Notas de aula do Curso de Modelos de Planejamento - 1981 (Não publicada)

11. Doi, Masayuki - The land Use and Zoning Study in Relation to Systematic Transportation Study in Campina Grande - Parte do projeto de planejamento de Transportes - 1978 - (Não publicado).

12. Cruz, Walter S. - A Necessidade de Novas Abordagens à Demanda de Viagens - Anais do Simpósio sobre Modelos Urbanos, Regionais e de Transportes - IPT - São Paulo - Março/1983 - pp 203 - 209.

13. Leal, José E. - Bases para Um Modelo Desagregado de Geração-Repartição - Anais do Simpósio sobre Modelos Urbanos, Regionais e de Transportes - IPT - São Paulo Março/1983.