TRATAMENTO DE DADOS AUSENTES PARA ANÁLISE FATORIAL DE INDICADORES DE SAÚDE

Antonio José Ribeiro Dias

TESE SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Aprovada por:

Prof. Claudio Thomás Bornstein, Dr.
(Presidente)

Prof. Flávio Fonseca Nobre, PhD.

Prof. Victor Hugo de Carvalho Gouvea, Dr. Ing.

RIO DE JANEIRO, RJ - BRASIL

ABRIL DE 1990

DIAS, ANTONIO JOSÉ RIBEIRO

Tratamento de dados ausentes para análise fatorial de indicadores de saúde [Rio de Janeiro] 1990

ix, 180 p. 29,7 cm (COPPE/UFRJ, M. Sc., Engenharia de Sistemas, 1990)

Tese - Universidade Federal do Rio de Janeiro, COPPE

1. Tratamento de dados ausentes para análise fatorial de indicadores de saúde I. COPPE/UFRJ II. Título (série).

...é sempre uma nova esperança
que a gente alimenta de sobreviver...

(Paulinho da Viola, em Amor à natureza)

Para minha Goretti

AGRADECIMENTOS

Quero agradecer ao Claudio Bornstein pela ajuda e pelas conversas que tivemos que não se relacionavam com este trabalho. Isso me levou a conhecer uma pessoa muito interessante.

Ao Flávio e demais pessoas do Programa de Biomédicas agradeço pelo acesso aos dados, que deram oportunidade para a feitura deste trabalho.

O Victor Hugo não será esquecido por ter aceito fazer parte da banca.

Neste parágrafo reservo meu abraço para todos os amigos, que me ajudaram ou não, que compartilham comigo, conscientemente ou não, todos os momentos de minha vida. Particularmente agradeço ao Edvaldo pelo apoio moral e material (o que seria de mim sem sua máquina milagrosa?!).

Finalmente, e sem comentários, me lembro da Goretti, da Beatriz e do Gabriel...

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para obtenção do grau de Mestre em Ciências (M. Sc.).

TRATAMENTO DE DADOS AUSENTES PARA ANÁLISE FATORIAL DE INDICADORES DE SAÚDE

Antonio José Ribeiro Dias

Abril, 1990

Orientador: Claudio Thomás Bornstein
Programa: Engenharia de Sistemas e Computação

Neste trabalho são apresentadas alternativas para se trabalhar com conjuntos de dados estatísticos onde existe o problema da ausência de informação em algumas células da matriz dos dados. Em seguida é apresentada a técnica de análise fatorial, bem como um exemplo de aplicação num conjunto de indicadores para determinar as dimensões (fatores) importantes a serem consideradas no estudo dos problemas relativos à saúde.

Abstract of Thesis presented to COPPE/UFRJ as partial fulfillment of the requeriments for the degree of Master of Science (M. Sc.)

TRATAMENTO DE DADOS AUSENTES PARA ANÁLISE FATORIAL DE INDICADORES DE SAÚDE

Antonio José Ribeiro Dias

Thesis Supervisor: Claudio Thomás Bornstein

Department: Engenharia de Sistemas e Computação

This work presents alternatives for handling statistical information with missing values for some cells of the data matrix. Furthermore factor analysis techniques are presented and aplyed to determining the most important dimensions (factors) for the study of health problems.

ÍNDICE

ANEXO H

CAPÍTULO I

I. Introdução

A idéia inicial deste trabalho consistia na análise de um conjunto de variáveis tradicionalmente tidas como relacionadas com a questão da saúde buscando definir as dimensões mais importantes e necessárias para o estudo e compreensão do problema.

Para isso contava-se com um arquivo, já em meio magnético, de dados para cinquenta e nove cidades brasileiras espalhadas por, praticamente, todo o território brasileiro, onde existiam sessenta e três variáveis relativas à demografia, saúde, mortalidade, infraestrutura urbana e rural, emprego, saneamento, etc, para o período que vai do ano de 1960 até 1982.

As informações foram compiladas a partir de diversas fontes independentes como publicações de orgãos oficiais de estatística, secretarias ou outros organismos ligados às diversas administrações estaduais referentes aos municípios de interesse.

As variáveis escolhidas são aquelas comumente usadas neste tipo de estudo, como pode ser verificado, por exemplo, em *BUSSAB E HO [1]*.

Os critérios para a escolha das cidades a serem incluídas no estudo, segundo *PANERAI [2]*, foram os seguintes:

- cidades com mais de 100000 habitantes representativas de centros urbanos potencialmente sujeitos a problemas relativos à saúde e alta taxa de crescimento populacional;
- representatividade da população brasileira segundo os estados e regiões geográficas; e
- facilidade para coleta das informações necessárias.

Após digitadas as informações sofreram um processo de depuração para eliminar possíveis erros de coleta e/ou entrada dos dados.

Em seguida os dados foram submetidos a uma normalização. Esse processo visava proporcionar maior comparabilidade entre as diversas cidades, principalmente diminuindo a influência do seu tamanho que variava, em 1980, entre pouco mais de 100.000 (Sumaré) e mais de 8.000.000 habitantes (São Paulo). Esse trabalho, bem como a descrição de algumas variáveis derivadas, é descrito em detalhes em *PANERAI [2].*

Depois de normalizados, os dados ainda foram analisados no sentido de serem localizados possíveis valores extremos (*"outliers"*) os quais foram conferidos e corrigidos quando necessário.

O Anexo A apresenta a descrição das variáveis e as fontes para obtenção dos dados; o Anexo B a lista das variáveis com as unidades de medida após a normalização e o Anexo C a lista das cidades estudadas.

À primeira vista parece que se tem em mãos um conjunto de dados ideal para se analisar. Ocorre, entretanto, que se verifica a ocorrência de uma quantidade bastante significativa de células vazias na matriz de dados.

Para as pessoas que estão habituadas a trabalhar com dados estatísticos isso, infelizmente, não chega a constituir surpresa, já que se sabe das inúmeras dificuldades que se encontra no trabalho de coleta de informações. O problema pode ocorrer em diversos planos. Quando se busca dados em pesquisas de campo, diretamente com o informante, pode-se deparar com a recusa por parte deste em prestar todas as informações desejadas. Às vezes os documentos de coleta (questionários), ou mesmo a própria pesquisa, podem ter falhas de planejamento e/ou execução que levem a perda, ou alteração da qualidade dos dados. A falta de treinamento adequado, ou até a má fé, dos entrevistadores também podem ser fontes de erros ou omissões.

Quando os dados são obtidos de outras fontes como arquivos, cadastros ou publicações de outras instituições, frequentemente ocorre que tais fontes em si já são incompletas. Se as fontes são múltiplas podem divergir quanto à definição das variáveis levando á dificuldades, ou mesmo impossibilidade, de compatibilização, o que pode acarretar, também, na perda de informações valiosas.

Acontece que as técnicas tradicionais de análise estatística são adequadas à aplicações em matrizes de dados sem falta de informações.

No Capítulo II desta dissertação procura-se discutir a questão da ausência de informações (*"missing values"*) sugerindo algumas maneiras de se tratar o problema, no sentido de se possibilitar alguma análise a partir dos dados disponíveis, mesmo que incompletos. Os métodos apresentados têm como referência principal o livro de *LITTLE E RUBIN [3]*, sendo que no Capítulo VI são oferecidas outras opções para consulta.

Como se percebe, o número de variáveis disponíveis no arquivo de dados descrito anteriormente é bastante grande, o que dificulta a visualização de seus efeitos. Esse fato remete imediatamente para o uso de alguma técnica de análise multivariada, para reduzir a dimensão do problema facilitando sua compreensão.

Pela observação das matrizes de correlações entre as variáveis vê-se que estas variam em magnitude sendo algumas consideravelmente altas enquanto outras quase nulas. Por outro lado as variáveis são todas numéricas o que indica o uso da análise fatorial (*WELLS E SHETH [4]*).

Esta técnica de análise multivariada procura descrever as relações de covariância entre um conjunto grande de variáveis através de um pequeno (o menor possível) número de fatores (variáveis aleatórias não diretamente observáveis). Cada uma das variáveis originais pode ser descrita como uma combinação linear dos fatores (ou vice versa) sendo que os coeficientes da função linear representam a correlação entre a variável e os fatores correspondentes. Desta forma é possível associar um significado particular a cada um dos fatores de acordo com o grupo de variáveis com as quais mais se relacionam (positiva ou negativamente).

No Capítulo III é apresentada uma visão geral sobre análise fatorial enfocando o modelo básico, os métodos de solução mais difundidos, escolha do número de fatores a serem considerados e o problema da rotação dos fatores. Embora seja inicialmente introduzido o modelo geral, a tônica do capítulo é centrada no modelo ortogonal.

A organização do texto sobre análise fatorial segue, básicamente a orientação de *HARMAN [5]* e *JOHNSON E WICHERN [6].*

Dedica-se o Capítulo IV do trabalho à apresentação dos resultados de um exercício de aplicação tanto das técnicas de tratamento da falta de informações como da análise fatorial.

Para o tratamento do problema dos valores ausentes através do algorítmo EM foi utilizado um programa desenvolvido originalmente por *SILVA [7]*, com algumas adaptações introduzidas pelo autor desta dissertação. O programa foi desenvolvido na linguagem do *SAS* (Statistical Analysis System), pois este *"software"* possui um módulo (Proc Matrix) para álgebra matricial o que torna o trabalho de programação razoavelmente simples.

A análise fatorial na fase inicial do trabalho foi desenvolvida com o uso do *SYSTAT*, que é um pacote estatístico, de uso geral, para computadores pessoais (a versão utilizada é para uso em microcomputadores da linha *IBM-PC* ou compatíveis), porém optou-se depois por usar o *SAS* devido às facilidades apresentadas por este pacote, tanto de programação como pelo fato dos resultados por ele apresentados possuirem muito mais elementos para análise. O *SAS* está disponível tanto para micros como para computadores de grande porte.

Finalmente são apresentados, no Capítulo V, alguns comentários sobre os resultados obtidos.

CAPÍTULO II

II.1 Tratamento dos valores ausentes

Todos os métodos conhecidos tradicionalmente para análise estatística de dados foram pensados em condições ideais, onde tem-se em mãos uma matriz de valores observados de p variáveis para n casos distintos. Na prática, porém, as coisas quase nunca ocorrem dessa maneira. O que se tem é uma matriz de dados onde frequentemente existe falta de informações (*"missing values"*) para algumas células e, ainda, alguns valores sobre os quais pode pesar a suspeita de que são portadores de erros de medida ou de aplicação incorreta de conceitos, de questionário, etc...

Na estatística clássica as diversas técnicas da teoria da amostragem se preocupam e resolvem de maneira satisfatória os chamados erros amostrais, que são inerentes ao processo pelo simples fato de se trabalhar com uma amostra, ou seja, apenas uma parte da população sobre o(s) efeito(s) do(s) fenômeno(s) estudado(s). Os métodos de tratamento dos erros amostrais podem ser vistos em vasta bibliografia, como, por exemplo, nos clássicos *COCHRAN [8] E HANSEN ET ALLI [9].*

Idealmente numa pesquisa de campo (ou mesmo numa compilação de dados de fontes conhecidas, como publicações ou arquivos magnéticos de dados) só haveriam os chamados erros amostrais. No plano real sabe-se que mesmo que a pesquisa seja censitária, não estará livre dos erros ditos não amostrais que podem ter como fontes diversos fatores:

- **erros de cobertura**: quando o sistema de referência (cadastro) da pesquisa possui falhas (falta ou duplicação de registros, por exemplo);

- **erros de conteúdo**: problemas de compreensão de conceitos, falhas de documento de coleta, má condução das entrevistas, má fé, erros introduzidos durante o processamento dos dados, etc...;

- **erros de não resposta**: impossibilidade de acesso às fontes, recusa de resposta por parte dos informantes, etc...

Há algum tempo existe a preocupação com o tratamento desses erros ditos não amostrais, tendência que sofre impulso a partir da

década de 70, principalmente pelas facilidades proporcionadas pelo avanço da informática, facilitando a implementação de métodos que eventualmente exigem cálculos praticamente impossíveis de serem feitos à mão ou em calculadoras convencionais.

Neste trabalho a preocupação se restringe apenas ao tratamento do problema dos valores ausentes da matriz de dados, supondo-se que os dados presentes estão livres de outros tipos de erros não amostrais citados. Um trabalho que se preocupa com os resultados que embora presentes possam carregar alguma suspeita de anomalia é, por exemplo, o de *SILVA [7].*

Para terminar a introdução deste capítulo deve-se lembrar que quaisquer que sejam as técnicas, por mais sofisticadas que possam ser, de preenchimento dos "buracos" de uma matriz de dados, estas não irão melhorar a qualidade desses dados mas apenas possibilitar maneiras de se trabalhar com o que se tem à mão, sempre levando em conta esse fato nas análises e conclusões, pois os dados verdadeiramente bons são aqueles que são originários das fontes.

II.2 Alguns métodos para tratamento de dados ausentes

Aqui se introduz algumas maneiras para se tratar os dados com ocorrência de *"missing values"*, em conjunto de dados relativos à variáveis numéricas. Basicamente o que se supõe é que os "buracos" na matriz de dados acontecem completamente ao acaso, não se conhecendo nenhum padrão de comportamento dessas ocorrências, o que os torna, em certo sentido, métodos gerais de imputação dos valores ausentes. Quando se conhece alguma tendência, ou padrão, de ocorrência das falhas existem métodos apropriados, que levam em conta essa informação (ver *LITTLE E RUBIN [3]*).

Geralmente os métodos de análise do tipo multivariado necessitam para a sua aplicação que se tenha boas estimativas do vetor de médias, μ, e da matriz de covariâncias, $\mathbb{V}$, das variáveis a serem analisadas. Para isso, na presença de *"missing values"*, pode-se trabalhar apenas com aquelas observações onde tais falhas não ocorrem, usar métodos que levam em conta todos os dados presentes sem se preocupar em preencher os "buracos" da matriz, ou, métodos que se preocupam em estimar valores para as células vazias baseando-se nos dados disponíveis.

Para uma análise fatorial, por exemplo, só é necessário ter os dados sumariados através da matriz de covariâncias (ou correlações), porém, quando se tem a possibilidade de estimar valores para substituir os ausentes isso pode ensejar uma maior variedade na escolha de métodos de análise.

II.2.1. Análise a Partir dos Casos Completos (Método Listwise)

Este é o método, certamente, mais simples para o tratamento de dados com ausência de informações, pois consiste simplesmente em abandonar os casos onde pelo menos uma das variáveis não estiver presente. Com isso o que se passa a ter é uma sub amostra dos dados originalmente desejados, formando, agora, uma matriz completa de dados. Em consequência todo o arsenal de análise estatística disponível poderá ser aplicado aos dados.

Para problemas em que se possa, efetivamente, supor que a ocorrência de células vazias na matriz de dados seja completamente aleatória, este método pode fornecer estimativas não viciadas para o vetor de médias e matriz de covariâncias.

Este método pode ser usado, e deve dar bons resultados, quando a quantidade de falta de informações for relativamente pequena. Nos problemas em que o número de variáveis é grande, geralmente a probabilidade de se ter casos sem nenhuma omissão é pequena, o que dificulta a aplicação do método, pois o número de casos completos provavelmente será pequeno.

Outro problema que se pode ter com tal processo é que mesmo que se tenha bons estimadores para os parâmetros μ e $\mathbb{V}$, ele não propicia estimativas para os próprios valores ausentes, o que seria desejável para análises posteriores dos dados.

Os pacotes computacionais estatísticos mais difundidos, como *SAS*, *BMDP*, *SYSTAT*, etc, dispõe dessa opção para o tratamento dos valores ausentes.

II.2.2 Análise a Partir de Todos os Dados Disponíveis (Método Pairwise)

Em oposição ao método anterior, este se preocupa em usar ao máximo os dados disponíveis não abandonando aquelas observações incompletas. Através de simulações mostrou-se que este método parece ser mais eficiente, que o *Listwise*, nos casos em que as correlações entre as variáveis são pequenas (*KIM e CURRY [10]*), ocorrendo o contrário quando as mesmas são altas (*AZEN E VAN GUILDER [11]*).

A estimativa do vetor de médias é feita estimando-se a média de cada uma das variáveis usando os dados disponíveis:

$$\mu_j^{(j)} = (1/n^{(j)}) \sum_{(j)} x_{ij} , \quad j = 1, 2, \ldots, p$$

onde: $n^{(j)}$ é o número de valores presentes para a variável X_j

$\mu_j^{(j)}$ é a estimativa da média da variável X_j baseada apenas nos valores presentes;

x_{ij} é valor da variável j para o elemento i ;

$\sum_{(j)}$ indica o somatório para todo o elemento i para os quais existe efetivamente o dado.

A estimação das matrizes de covariâncias (ou correlações) apresenta mais de uma alternativa baseadas sempre no emparelhamento dos dados existentes simultaneamente para as variáveis X_j e X_k correspondentes ao elemento s_{jk} da matriz de covariâncias (ou correlações).

Uma das alternativas para o cálculo da matriz de covariâncias seria:

$$s_{jk}^{(jk)} = \sum_{(jk)} (x_{ij} - \bar{x}_j^{(jk)})(x_{ik} - \bar{x}_k^{(jk)})/(n^{(jk)} - 1)$$

$$j, k = 1, 2, \ldots, p$$

onde: $s_{jk}^{(jk)}$ covariância entre as variáveis X_j e X_k baseada nos valores simultaneamente presentes para as duas variáveis;

$\bar{x}_j^{(jk)}$ (ou $\bar{x}_k^{(jk)}$) média da variável X_j (ou X_k) baseada nos valores presentes para X_j e X_k;

$n^{(jk)}$ número de observações com valores presentes simultaneamente para para X_j e X_k;

$\sum_{(jk)}$ indica o somatorio para todos os elementos i para os quais estão presentes os dados da variável X_j e X_k simultaneamente.

Outra possibilidade seria estimar a média de cada uma das variáveis considerando todos os valores presentes para cada uma delas, ou seja:

$$\tilde{s}_{jk}^{(jk)} = \sum_{(jk)} (x_{ij} - \bar{x}_j^{(j)})(x_{ik} - \bar{x}_k^{(k)})/(n^{(jk)} - 1)$$

$$j, k = 1, 2, \ldots, p$$

onde: $\bar{x}_j^{(j)}$ (ou $\bar{x}_k^{(k)}$) média da variável X_j (ou X_k) baseada nos valores presentes para X_j (ou X_k).

Para o cálculo das estimativas de correlações também se pode recorrer a alternativas diversas, usando tanto uma como outra maneira de se calcular as variâncias e covariâncias.

$$r_{jk}^{*} = s_{jk}^{(jk)} / \sqrt{s_{jj}^{(j)} s_{kk}^{(k)}} \quad , \; j, k = 1, 2, \ldots, p$$

$$r_{jk}^{(jk)} = s_{jk}^{(jk)} / \sqrt{s_{jj}^{(jk)} s_{kk}^{(jk)}} \quad , \; j, k = 1, 2, \ldots, p$$

A primeira opção pode resultar em correlações estimadas fora de intervalo [-1, 1] o que não faz sentido teórico.

O método *Pairwise* tem como mérito a tentativa de usar efetivamente toda a informação coletada, não desprezando aquelas observações onde exista falta de informação para alguma(s) variável(s), como no *Listwise*. Seu principal problema, no entanto, é a possibilidade de gerar matrizes de covariâncias (ou correlações) que não serão, necessariamente, positivas definidas (ou ao menos não negativas). Quando isso ocorre, a única maneira de solucionar o problema é por meio de ajustes feitos arbitrariamente na matriz calculada, o que não é sempre muito agradável.

LITTLE E RUBIN [3] mostram através de um exemplo artificial os problemas que podem ocorrer.

Seja a matriz da dados abaixo:

$$X' = \begin{bmatrix} 1 & 2 & 3 & 4 & 1 & 2 & 3 & 4 & - & - & - & - \\ 1 & 2 & 3 & 4 & - & - & - & - & 1 & 2 & 3 & 4 \\ - & - & - & - & 1 & 2 & 3 & 4 & 4 & 3 & 2 & 1 \end{bmatrix}$$

Tem-se que:

$$r_{12}^{(12)} = 1,\ r_{13}^{(13)} = 1 \text{ e } r_{23}^{(23)} = -1.$$

Mas, como se pode notar, isso é uma contradição já que:

$$Cov(X_i, X_j) = Cov(X_i, X_k) = 1 \neq Cov(X_j, X_k)$$

É também interessante notar que o método *Listwise* não pode ser aplicado ao exemplo, já que não há sequer uma observação completa.

II.2.3 O método das médias

Este método se diferencia, basicamente, dos apresentados anteriormente pelo fato de se preocupar em estimar (imputar) valores para os dados ausentes. O atrativo principal de tal estratégia é que se pode, após aplicá-la, usar os métodos disponíveis da análise estatística como se os dados fossem completos. Tal atitude pode, porém, ser perigosa se usada indiscriminadamente, pois, as estimativas produzidas a partir de dados imputados podem trazer vícios (*"bias"*) importantes.

O método das médias consiste, simplesmente, em se imputar os valores ausentes da variável X_j pela média aritmética dos seus valores presentes, $\bar{x}_j^{(j)}$.

É fácil de se perceber que com este tipo de estimação dos valores ausentes passa-se a conviver com uma sub estimativa da variabilidade dos dados.

Pode-se ver que a variância estimada da variável X_j será:

$$s_{jj} = [1/(n-1)] \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$

$$= [1/(n-1)] \left\{ \sum_{i=1}^{n} x_{ij}^2 - n\bar{x}_j^2 \right\}$$

Mas como os valores ausentes foram substituidos pela média dos valores presentes, tem-se que:

$$\sum_{i=1} x_{ij}^2 = (n - n^{(j)})\, \bar{x}_j^{(j)2} + \sum_{(j)} x_{ij}^2$$

$$= n\bar{x}_j^{(j)2} - n^{(j)}\, \bar{x}_j^{(j)2} + \sum_{(j)} x_{ij}^2$$

E, ainda:

$$\bar{x}_j = (1/n) \sum_{i=1}^{n} x_{ij}$$

$$= (1/n) \left[(n - n^{(j)})\, \bar{x}_j^{(j)} + \sum_{(j)} x_{ij} \right]$$

$$= (1/n) \left[n\bar{x}_j^{(j)} - n^{(j)}\, \bar{x}_j^{(j)} + n^{(j)} \bar{x}_j \right] = \bar{x}_j^{(j)}$$

Portanto:

$$\bar{x}_j = \bar{x}_j^{(j)} , \qquad j = 1, 2, \ldots, p$$

Logo:

$$s_{jj} = [1/(n-1)]\left\{n\bar{x}_j^{(j)2} - n^{(j)}\,\bar{x}_j^{(j)2} + \sum_{(j)} x_{ij}^2 - n\bar{x}_j^{2}\right\}$$

$$= [1/(n-1)]\left\{n\bar{x}_j^{(j)2} - n^{(j)}\,\bar{x}_j^{(j)2} + \sum_{(j)} x_{ij}^2 - n\bar{x}_j^{(j)2}\right\}$$

$$= [1/(n-1)]\left\{\sum_{(j)} x_{ij}^2 - n^{(j)}\bar{x}_j^{(j)2}\right\}$$

$$= [1/(n-1)]\sum_{(j)} (x_{ij} - \bar{x}_j^{(j)})^2$$

Mas,

$$s_{jj}^{(j)} = [1/(n^{(j)}-1)]\sum_{(j)} (x_{ij} - \bar{x}_j^{(j)})^2$$

Portanto:

$$s_{jj} = [(n^{(j)}-1)/(n-1)]s_{jj}^{(j)}, \qquad j = , 2, \ldots, p$$

Supondo-se que as faltas de informação ocorrem ao acaso, sabe-se que $s_{jj}^{(j)}$ é uma estimativa não viciada da variância de X_j, e portanto como o fator $(n^{(j)}-1)/(n-1)$ é menor que 1, desde que ao menos uma das observações seja um *"missing value"*, tem-se que s_{jj} subestima a variância de X_j.

Generalizando tem-se que:

$$s_{jk} = [(n^{(jk)}-1)/(n-1)]\tilde{s}_{jk}^{(jk)}, \quad j, k = 1,2,\ldots,p$$

Dessa maneira a matriz de covariâncias obtida dos dados com os "buracos" preenchidos pelas médias dos valores disponíveis será positiva semi definida. Então, apesar de se saber que existe um vício na estimativa da matriz de covariâncias, pode-se aplicar sem problemas as técnicas de análise estatística que a utilizam como entrada.

II.2.4 O Algorítimo EM

Aqui introduz-se a idéia de usar o algorítmo EM (*Expectation-Maximization*) como um instrumento de geração de estimativas a serem imputadas no lugar dos valores ausentes da matriz de dados.

Esta técnica, em dois passos e iterativa, é usada para calcular as estimativas de máxima verossimilhança da matriz de covariâncias, $\mathbb{V}$, e do vetor de médias, μ, de uma variável normal multivariada. Segundo *LITTLE E RUBIN [3]* essa hipótese de normalidade dos dados pode ser enfraquecida, desde que o algorítmo é capaz de fornecer estimativas consistentes para qualquer variável cuja distribuição possua o quarto momento finito e os dados estejam livres de valores espúrios (*"outliers"*).

Em caso de haver contaminação nos dados, *LITTLE E SMITH [12]* sugerem uma alternativa, que vem a ser uma modificação no segundo passo do algorítmo EM, denominada algorítmo ER, que usa a teoria de estatística robusta para ponderar as informações, diminuindo a influência de valores extremos.

A principal idéia do método EM é imputar os valores ausentes de uma dada observação através da regressão linear das variáveis correspondentes a estes dados sobre as variáveis que possuem valores presentes . Vê-se que a idéia é bastante simples, o que talvez explique o fato de referências ao método estarem presentes na literatura à bastante tempo. *MCKENDRICK [13]*, já em 1926, aplica esta idéia num problema de análise de dados em medicina. *DEMPSTER, LAIRD E RUBIN [14]* é que introduzem a denominação EM dando vários exemplos de aplicação e provando resultados gerais sobre o comportamento e convergência do método.

Toda a idéia do processo pode ser resumida da seguinte maneira:

- inicialização: determinar estimativas iniciais para o vetor de médias, μ, e matriz de covariâncias, $\mathbb{V}$;

- estimar os dados ausentes, supondo que as estimativas atuais dos parâmetros μ e $\mathbb{V}$ são corretas;

- calcular novas estimativas para μ e $\mathbb{V}$ e iterar o processo até sua convergência.

Uma grande vantagem teórica do algorítmo é ser assegurada a sua convergência sob condições gerais. A função de verossimilhança, $l(\theta/X_{obs})$, é crescente e, se é limitada, a sequência $l(\theta_{(t)}/X_{obs})$ converge para um valor estacionário.

Pode-se dizer que o algorítmo EM é um método eficiente de imputação de valores ausentes, pois, no seu passo E (*Expectation)* os *"missing values"* são substituídos pelos melhores preditores lineares baseados nas estimativas atuais de μ e $\mathbb{V}$ (ver por exemplo *SEARLE [15]* para a teoria sobre estimadores BLUE - *Best Linear Umbiased Estimators*).

Antes de apresentar o algorítmo propriamente dito será exposta uma maneira de se estimar os coeficientes de uma regressão linear conhecendo-se apenas a matriz de covariâncias e o vetor de médias da matriz de dados aumentada. Dá-se o nome de matriz de dados aumentada à matriz formada pelas variáveis dependentes e preditoras de uma regressão linear.

II.2.4.1 Regressão linear a partir da matriz de covariâncias e do vetor de médias de todas variáveis envolvidas

Foi dito que a idéia do método EM é imputar os valores ausentes pelo valor da regressão linear das variáveis faltantes sobre as variáveis cujos dados são disponíveis. O problema é que não dispondo dos valores da(s) variável(s) dependente(s) mas apenas das variáveis preditoras, não se pode calcular os parâmetros de uma regressão linear da forma tradicional. Nesse sentido é necessário apresentar uma maneirade calcular tais parâmetros prescindindo dos dados que não são conhecidos. Isso é possível desde que se disponha de estimativas do vetor de médias e matriz de covariâncias de todas as variáveis envolvidas. Ressalte-se, ainda, que este método pode ser usado sem problemas quando os dados são completos.

Seja a variável dependente **Y** e a matriz **X**, cujas colunas são p variáveis preditoras $\mathbf{X}_1$, $\mathbf{X}_2$,..., $\mathbf{X}_p$.

O modelo clássico de regressão linear múltipla é definido como:

$$\mathbf{Y} = \beta_0 + \mathbf{X}\beta + \varepsilon \quad .$$

onde: $\mathbf{Y}_{(n,1)}$ é o vetor formado pela variável dependente;
$\mathbf{X}_{(n,p)}$ é a matriz das variáveis preditoras;
β_0 é o parâmetro independente de **X** (*"intercept"*);
$\beta_{(p,1)}$ é o vetor dos parâmetros associados às variáveis componentes da matriz **X**;
$\varepsilon_{(n,1)}$ é o vetor dos erros aleatorios.

São ainda suposições do modelo que os erros são normalmente distribuídos com média zero e variância σ^2.

Sejam ainda conhecidos o vetor de médias, μ, e a matriz das covariâncias, $\mathbb{V}$, de todas as variáveis envolvidas, particionados como se mostra a seguir:

$$\mu = \left[\mu_Y \,\vdots\, \mu_X\right] \text{ e } \mathbb{V} = \left[\begin{array}{c|c} \mathbb{V}_{YY} & \mathbb{V}'_{XY} \\ \hline \mathbb{V}_{XY} & \mathbb{V}_{XX} \end{array}\right]$$

onde: μ_Y é a média de **Y**
μ_X é o vetor de médias da matriz **X**
$\mathbb{V}_{YY}$ é a variância de **Y**
$\mathbb{V}_{XX}$ é a matriz de covariâncias de **X**
$\mathbb{V}_{XY}$ são as covariâncias de **Y** com as variáveis de **X**

Sejam b_0 e **b**, respectivamente, os estimadores de β_0 e β.

Um preditor linear do modelo definido dessa maneira será calculado como:

$$\begin{aligned} \hat{\mathbf{Y}} &= b_0 + b_1 X_1 + \ldots + b_p X_p \\ &= b_0 + \mathbf{Xb} \end{aligned}$$

O erro de predição pode ser calculado pela diferença abaixo:

$$\mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - b_0 - \mathbf{Xb}$$

Uma maneira de se calcular b_0 e **b** é determinar seus valores de maneira que minimizem o erro quadrático médio da predição, que é definido como:

$$EQM = E[\mathbf{Y} - b_0 - \mathbf{Xb}]^2$$

Resultado: os valores de b_0 e **b** que minimizam o erro quadrático médio são dados por:

$$\mathbf{b} = \mathbb{V}_{XX}^{-1}\mathbb{V}_{XY}$$

$$b_0 = \mu_Y - \mu_X\mathbf{b}$$

e o valor mínimo do erro quadrático médio é atingido quando:

$$EQM = \mathbb{V}_{YY} - \mathbb{V}'_{XY}\mathbb{V}_{XX}^{-1}\mathbb{V}_{XY}$$

O preditor linear será dado substituindo os valores dos parâmetros em sua equação, ou seja:

$$\hat{\mathbf{Y}} = b_0 + \mathbf{X}\mathbf{b}$$

$$= \mu_Y - \mu_X \mathbb{V}_{XX}^{-1} \mathbb{V}_{XY} + \mathbf{X}\mathbb{V}_{XX}^{-1}\mathbb{V}_{XY}$$

$$= \mu_Y + (\mathbf{X} - \mu_X)\mathbb{V}_{XX}^{-1}\mathbb{V}_{XY}$$

Um fato importante a ser ressaltado é que sendo o preditor linear aqui definido, pelo menos sob a hipótese de normalidade dos dados, um estimador não viciado dos valores de **Y** (lembre-se que é *BLUE* de acordo com *SEARLE [15]*) o erro quadrático médio coincide com sua variância, ou seja:

$$\mathrm{Var}(\hat{\mathbf{Y}}) = \mathbb{V}_{YY} - \mathbb{V}'_{XY}\mathbb{V}_{XX}^{-1}\mathbb{V}_{XY}$$

Para provar a validade dos resultados apresentados basta calcular o erro quadrático médio a partir da definição do modelo. Tal prova é mostrada no Anexo D desta dissertação.

Para completar este item, pode-se verificar que as estimativas dos parâmetros do modelo de regressão calculadas da forma aqui apresentada, coincidem com os valores estimados pelo método dos mínimos quadrados. Tal verificação será feita por meio de um exemplo simples.

Seja o seguinte conjunto de dados:

Y	1	4	3	8	9
X	0	1	2	3	4

Sabe-se que pelo método dos mínimos quadrados (ver, por exemplo, *SEARLE [15]*) o estimador dos parâmetro da regressão é dado por:

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

Onde deve-se adicionar uma coluna de 1's à matriz **X** para

permitir a estimação do parâmetro β_0, que vem a ser o termo independente (*"intercept"*) da equação de regressão.

Com os dados acima tem-se:

$$\mathbf{X'X} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} = \begin{bmatrix} 5 & 10 \\ 10 & 30 \end{bmatrix}$$

$$\mathbf{X'Y} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} 1 \\ 4 \\ 3 \\ 8 \\ 9 \end{bmatrix} = \begin{bmatrix} 25 \\ 70 \end{bmatrix}$$

$$(\mathbf{X'X})^{-1} = \begin{bmatrix} 0,6 & -0,2 \\ -0,2 & 0,1 \end{bmatrix}$$

$$\hat{\beta} = (\mathbf{X'X})^{-1}\mathbf{X'Y} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \Longrightarrow \hat{\mathbf{Y}} = 1 + 2X$$

Agora, calculando pelo método apresentado, tem-se:

$$\begin{bmatrix} \mathbf{Y} & \mathbf{X} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 4 & 1 \\ 3 & 2 \\ 8 & 3 \\ 9 & 4 \end{bmatrix} \qquad \mu = \left[\begin{array}{c:c} 5 & 2 \end{array}\right]$$

$$\mathbb{V} = (1/5)\left[\begin{array}{c:c} 56 & 20 \\ \hdashline 20 & 10 \end{array}\right]$$

Portanto:

$$\mathbf{b} = \frac{5}{10} \quad \frac{20}{5}$$

$$= 2$$

$$b_0 = 5 - 2*2$$

$$= 1 \qquad \Longrightarrow \hat{\mathbf{Y}} = 1 + 2\mathbf{X}$$

Nota-se, portanto, que os dois métodos têm como resultado a mesma equação de regressão.

Para finalizar este item sugere-se uma ferramenta importante para ser usada no cálculo dos elementos do método apresentado, chamado operador *SWEEP*.

O operador *SWEEP* foi definido por *BEATON [16]* tendo sofrido algumas adaptações posteriores, sendo uma poderosa ferramenta para a regressão linear tanto para os casos em que os dados são completos como para quando se tem ausência de informações. No livro de *LITTLE E RUBIN [3]* há uma apresentação detalhada deste operador bem como outras referências bibliográficas sobre o assunto.

Pode ser visto nas referências citadas que este operador quando aplicado adequadamente à matriz aumentada dos dados fornece, também em forma matricial, praticamente todos os elementos necessários à uma análise de regressão. Particularmente existe no *SAS* uma implementação que fornece os elementos para a solução do problema de regressão como aqui foi exposto.

A implementação do *SWEEP* existente no *SAS* tem a seguinte forma:

Seja uma matriz simétrica, M, particionada adequadamente, ou seja:

$$M = \left[\begin{array}{c|c} M_{11} & M'_{21} \\ \hline M_{21} & M_{22} \end{array}\right]$$

Aplicando então o operador *SWEEP* tem-se o seguinte resultado:

$$SWEEP[M] = \left[\begin{array}{c|c} M_{11}^{-1} & M_{11}^{-1}M_{12} \\ \hline -M_{21}M_{11}^{-1} & M_{22} - M_{21}M_{11}M_{12} \end{array}\right]$$

Nota-se que aplicando este operador à matriz de covariâncias, $\mathbb{V}$, particionada adequadamente pode-se obter todos os elementos para a determinação de b_0, $\mathbf{b}$ e do erro e variância de estimativa da variável independente.

II.2.4.2 Os passos do algorítmo EM

Com os elementos expostos até o momento pode-se, então, apresentar o algorítmo EM.

Os dois passos centrais deste método consistem em primeiro lugar calcular estimativas para substituir os valores ausentes através da regressão linear das variáveis onde estes se localizam sobre as demais variáveis tomadas como preditoras, supondo corretas as estimativas do vetor de médias, μ, e da matriz de covariâncias, $\mathbb{V}$, disponíveis e, no segundo passo atualizar os valores de μ e $\mathbb{V}$ a partir dos dados onde os valores ausentes foram estimados.

Para formalizar as idéias apresentadas acima se faz necessário definir uma notação adequada. Tal notação levará em conta o processo de cálculo dos parâmetros de uma regressão linear da maneira apresentada no item II.2.4.1.

Seja a matriz dos dados, denotada por $\mathbf{X}$, composta de todas as p variáveis envolvidas no problema. Assim $\mathbf{X}$ terá n linhas e p colunas. Seja $\mathbf{X}_i$ o vetor correspondente a observação i, ou seja, a $i^{\underline{a}}$ linha da matriz $\mathbf{X}$, $i = 1, 2, \ldots, n$.

O vetor $\mathbf{X}_i$ assim definido é um vetor linha com p elementos e pode ser particionado da seguinte forma:

$$\mathbf{X}_i = \left[\mathbf{X}_{(a)i} \;\vdots\; \mathbf{X}_{(p)i} \right] \qquad i = 1, 2, \ldots, n$$

onde: $\mathbf{X}_{(a)i}$ é a partição correspondente aos valores ausentes na observação i;

$\mathbf{X}_{(p)i}$ é a partição correspondente aos valores presentes na observação i;

Define-se ainda o vetor de médias e a matriz de covariâncias correspondentes às variáveis da matriz $\mathbf{X}$ como a seguir.

Sejam:

$$\mu^{(t)} = \left[\mu^{(t)}_{(a)} \;\vdots\; \mu^{(t)}_{(p)} \right]$$

$$\mathbb{V}^{(t)} = \left[\begin{array}{c|c} \mathbb{V}^{(t)}_{aa} & \mathbb{V}^{(t)}_{ap} \\ \hline \mathbb{V}^{(t)}_{pa} & \mathbb{V}^{(t)}_{pp} \end{array} \right]$$

onde, de acordo com a metodologia exposta no item II.2.4.1:

$\mu^{(t)}_{(a)}$ são as médias correspondentes às variáveis com dados ausentes, na iteração t;

$\mu^{(t)}_{(p)}$ são as médias correspondentes às variáveis com dados presentes, na iteração t;

$\mathbb{V}^{(t)}_{aa}$ é a partição da matriz de covariâncias correspondentes às variàveis com dados ausentes, na iteração t;

$\mathbb{V}^{(t)}_{pp}$ é a partição da matriz de covariâncias correspondentes às variàveis com dados presentes, na iteração t;

$\mathbb{V}^{(t)}_{ap}$ é a partição da matriz das covariâncias entre as variàveis com dados ausentes e as variáveis com dados presentes, na iteração t;

$\mathbb{V}^{(t)}_{pa} = \mathbb{V}'^{(t)}_{ap}$ é a matriz transposta de $\mathbb{V}^{(t)}_{ap}$.

Como o processo é iterativo deve-se indicar a que iteração as estimativas se referem. A própria notação revela que os valores presentes ($\mathbf{X}_{(p)i}$) não são alterados pelo processo, como era de se esperar.

Com a notação definida pode-se representar os elementos da matriz de dados, **X**, na iteração t como:

$$X^{(t)}_{ij} = \begin{cases} X_{ij} & \text{se } X_{ij} \text{ é um valor presente} \\ \hat{X}^{(t)}_{ij} & \text{se } X_{ij} \text{ é um valor ausente} \end{cases}$$

Como o método EM supõe a normalidade de **X**, a média aritimética das observações e a matriz de covariâncias observadas são estatísticas suficientes, ou seja: toda a informação amostral sobre **X** está contida em $\overline{\mathbf{X}}$ e **S** (para uma definição mais formal de suficiência estatística pode-se consultar, por exemplo, *MOOD, GRAYBILL E BOES [17]*).

É necessário, então, calcular a soma das observações de cada uma das variáveis e a soma dos seus produtos cruzados. Para isso é necessário estimar os valores faltantes.

Com as notações e definições apresentadas pode-se, então, definir os passos do algorítmo EM.

Passo E (*Expectation*): dadas as estimativas atuais do vetor de médias, $\mu^{(t)}$, e da matriz de covariâncias, $\mathbb{V}^{(t)}$, pode-se estimar os valores ausentes pela regressão linear sobre as variáveis presentes:

$$\hat{\mathbf{X}}^{(t)}_{(a)i} = \mu_{(a)} + (\mathbf{X}_{(p)i} - \mu_{(p)})\ \mathbb{V}^{-1}_{pp}\mathbb{V}_{pa}$$

Na verdade $\hat{\mathbf{X}}^{(t)}_{(a)i}$ assim definido é a esperança condicional de $\mathbf{X}^{(t)}_{(a)i}$ dados como conhecidos os valores de $\mathbf{X}_{(p)i}$, o vetor das médias e a matriz das covariâncias. Em notação própria de esperança condicional pode-se escrever:

$$\hat{\mathbf{X}}^{(t)}_{(a)i} = \mathrm{E}\left[\mathbf{X}^{(t)}_{(a)i} / \mathbf{X}_{(p)i};\ \mu^{(t)}, \mathbb{V}^{(t)}\right]$$

A esperança condicional acima definida é denominada *curva ou função de regressão linear* (veja, por exemplo, *MOOD, GRAYBILL E BOES [17]* ou *JOHNSON E WICHERN [6]*).

Com os valores estimados para os dados ausentes pode-se calcular a soma necessária para se estimar a média aritimética, restando calcular sua contribuição para a soma dos produtos cruzados que serão usados para estimar as covariâncias.

Para isso é necessário introduzir o conceito de variância condicional.

Definição: a variância condicional de uma variável Y dada a variável Z conhecida é definida por:

$$Var(Y/\ Z) = E(Y^2/\ Z) - \left[E(Y/\ Z)\right]^2$$

Portanto o produto cruzado desejado será calculado, no caso dos dados ausentes, como:

$$\mathbf{X}^{\prime(t)}_{(a)i}\mathbf{X}^{(t)}_{(a)i} = E\left[\mathbf{X}^{\prime(t)}_{(a)i}\mathbf{X}^{(t)}_{(a)i}/\ \mathbf{X}_{(p)i};\ \mu^{(t)}, \mathbb{V}^{(t)}\right]$$

$$= Var\left[\mathbf{X}^{(t)}_{(a)i}/\ \mathbf{X}_{(p)i};\ \mu^{(t)}, \mathbb{V}^{(t)}\right] + E\left[\mathbf{X}^{(t)}_{(a)i}/\ \mathbf{X}_{(p)i};\ \mu^{(t)}, \mathbb{V}^{(t)}\right]$$

Mas,

$$E\left[\mathbf{X}^{(t)}_{(a)i}/\mathbf{X}_{(p)i};\ \mu^{(t)}, \mathbb{V}^{(t)}\right] = \hat{\mathbf{X}}^{(t)}_{(a)i}$$

E de acordo com o item II.2.4.1:

$$Var\left[\mathbf{X}^{(t)}_{(a)i}/\ \mathbf{X}_{(p)i};\ \mu^{(t)}, \mathbb{V}^{(t)}\right] = \mathbb{V}_{aa} - \mathbb{V}'_{pa}\,\mathbb{V}^{-1}_{pp}\,\mathbb{V}_{pa}$$

Portanto o produto cruzado fica:

$$\mathbf{X}^{\prime(t)}_{(a)i}\mathbf{X}^{(t)}_{(a)i} = \mathbb{V}^{(t)}_{aa} - \mathbb{V}^{\prime(t)}_{pa}\mathbb{V}^{-1}_{pp}\ \mathbb{V}^{(t)}_{pa} + \hat{\mathbf{X}}^{\prime(t)}_{(a)i}\hat{\mathbf{X}}^{(t)}_{(a)i}$$

O produto cruzado envolvendo variáveis com dados presentes e ausentes será calculado como:

$$\mathbf{X}^{\prime(t)}_{(a)i}\mathbf{X}_{(p)i} = E\left[\mathbf{X}^{\prime(t)}_{(a)i}\mathbf{X}_{(p)i}/\ \mathbf{X}_{(p)i};\ \mu^{(t)}, \mathbb{V}^{(t)}\right]$$

Como $\mathbf{X}^{(t)}_{(p)i}$ é um vetor de valores conhecidos, tem-se que:

$$\mathbf{X}^{\prime(t)}_{(a)i}\mathbf{X}_{(p)i} = \left\{E\left[\mathbf{X}^{(t)}_{(a)i}/\ \mathbf{X}_{(p)i};\ \mu^{(t)}, \mathbb{V}^{(t)}\right]\right\}' \mathbf{X}_{(p)i}$$

Dessa maneira:

$$\mathbf{X}^{\prime(t)}_{(a)i}\mathbf{X}_{(p)i} = \hat{\mathbf{X}}^{\prime(t)}_{(a)i}\mathbf{X}_{(p)i}$$

Pode-se, finalmente, calcular os valores necessários para as estatísticas suficiêntes, encerrando o passo E do algorítmo:

$$T_1^{(t)} = \sum_{i=1}^{n} X_i^{(t)}$$

$$T_2^{(t)} = \sum_{i=1}^{n} X_i^{,(t)} X_i^{(t)}$$

Deve-se ressaltar que os somatórios acima são somas de vetores no caso de $T_1^{(t)}$, e matrizes no caso de $T_2^{(t)}$.

Passo M (*Maximization*): neste passo são atualizados os valores das estimativas da matriz de covariâncias e do vetor de médias, que serão usadas na iteração seguinte, *t+ 1,* do algorítmo .Tais estimativas são calculadas usando-se os estimadores clássicos de máxima verossimilhança.

$$\mu^{(t+1)} = \frac{T_1^{(t)}}{n}$$

$$\mathbb{V}^{(t+1)} = \frac{T_2^{(t)}}{n} - n\,\mu^{,(t+1)}\mu^{(t+1)}$$

BEALE E LITTLE [18] sugerem a substituição do denominador de $\mathbb{V}^{(t+1}$ por *n-1* para se obter o estimador não viciado da variância.

Com os valores atualizados das estimativasde μ e $\mathbb{V}$ volta-se ao passo E e itera-se até que o critério de convergência seja atingido. O critério numérico sugerido é que a menor diferença relativa, em valor absoluto, entre as estimativas das médias e covariâncias entre os passos *t* e *t+1*, não ultrapasse um valor θ previamente fixado.

Sejam:

$$\theta_1 = \max_{j,k}\left\{\text{abs}\left[\left(s_{(t+1)jk} - s_{(t)jk}\right)/s_{(t)jk}\right]\right\}$$

$$j,\ k = 1,\ 2,\ldots,\ p$$

$$\theta_2 = \max_{j}\left\{\text{abs}\left[\left(\mu_{(t+1)j} - \mu_{(t)j}\right)/\mu_{(t)j}\right]\right\}$$

$$j = 1,\ 2,\ldots,\ p$$

Então a convergência se dá quando:

$$\max \left[\theta_1 , \theta_2 \right] < \theta$$

O número de iterações suficiêntes para que a condição acima seja satisfeita dependerá do tamanho da matriz de dados e da quantidade de céluas com falta de informação, já que o aumento dos "buracos" aumenta também o total de regressões que deverão ser estimadas e todas as estimativas deverão satisfazer a condição simultaneamente.

II.2.4.3 Escolha de estimativas iniciais para o vetor de médias e a matriz de covariâncias

Resta discutir a escolha dos valores iniciais para os parâmetros da distribuição, que serão usados na primeira iteração do algorítmo.

Pode-se optar por uma entre muitas alternativas de acordo com a matriz de dados a ser trabalhada. Quando o número de *"missing values"* for pequeno, ou , em outras palavras, quando o número de casos completos for suficientemente grande, parece que a melhor escolha é calcular as estimativas de μ e $\mathbb{V}$ a partir desses casos completos, o que pode gerar estimativas consistentes para esses parâmetros.

Quando o número de variáveis é consideravelmente grande, geralmente passa-se a não dispor de muitos casos completos o que prejudica o critério anterior. Nestes casos pode-se optar pelo método *Pairwise* para calcular os valores iniciais dos parâmetros.

Outra alternativa é imputar os dados ausentes pela média dos presentes e, em seguida, estimar a matriz de covariâncias, como se os dados fossem completos.

É bom relembrar as limitações, já discutidas, destas duas últimas alternativas propostas: a geração de matrizes não positivas definidas pelo método *Pairwise* e a sub estimação das covariâncias, no caso de se usar o método das médias. Note-se, entretanto que este último sempre pode oferecer uma estimativa inicial para a matriz de covariâncias que não deverá causar problemas númericos na aplicação do algorítmo EM.

CAPÍTULO III

III.1 A análise fatorial

Quando se busca a origem histórica da análise fatorial, volta-se ao início do nosso século aos trabalhos de *Karl Pearson* e *Charles Spearman* na tentativa de definir e medir a inteligência humana.

Alguns autores definem como "data de nascimento" da técnica de análise fatorial o ano de 1904 quando *Spearman* publica seu trabalho denominado *"General Intelligence, Objectivelly Determined and Mesured"* no *American Journal of Psychology*. Essa publicação marca o início de vasto trabalho do autor, aplicado ao desenvolvimento da teoria psicológica.

Antes disso, porém, em 1901, *Karl Pearson* já havia publicado seu trabalho *"The Principal Axes Method"*, que serve de base estatística ao trabalho de *Spearman* e é o marco inicial do estudo de Componentes Principais.

Spearman, baseado no trabalho de *Pearson* desenvolveu sua "Teoria dos Dois Fatores", onde descreve a inteligência humana através de um Fator Geral, comum a todos os indivíduos, embora variando de nível para cada pessoa, e um Fator Específico que depende de cada pessoa.

Com tal trabalho o autor dá início ao estudo das Variáveis Latentes ou não observáveis diretamente (os fatores) de grande utilidade para entender fenômenos em diversas áreas do conhecimento como: Psicologia, Sociologia, Economia, Biologia, Medicina, Geologia, Metereologia, etc. No livro de *HARMAN[5]* pode-se encontrar muitas referências de aplicações em todas as áreas acima relacionadas, dentre outras.

Para terminar este breve histórico da análise fatorial faltaria citar, ao menos, mais dois precursores.

O primeiro é *J. C. M. Garret* que em 1919 publica seu artigo *"On Certain Independent Factors in Mental Mesurament"* nos *Proceeding of the Royal Society.* Neste trabalho *Garret* contesta a "Teoria dos Dois Fatores" de *Spearman* e lança as bases da análise fatorial com múltiplos fatores.

Mais tarde, em 1930, *Harold Hotelling* sugere um método numérico satisfatório para a resolução do problema de Componentes Principais, onde ele incorpora idéias de otimização, já que leva em conta a maximização da variabilidade dos componentes.

III.2 Objetivos da análise fatorial

Como acontece com a maioria das técnicas de análise multivariada, a análise fatorial tem como objetivo sumariar informações sobre um determinado fenômeno de interesse em algum campo do conhecimento humano.

Em geral tal fenômeno pode ser observado, ou medido, por meio de um conjunto bastante numeroso de variáveis o que torna sua compreensão, ou visualização, às vezes muito difícil. A redução dessa dimensão serve para facilitar a análise do comportamento dos dados.

Existem, basicamente, duas situações onde a análise fatorial pode ser de grande utilidade para a análise de dados. A primeira delas ocorre nos casos em que os fenômenos a serem estudados são associados a um (ou mais) modelo matemático já conhecido, e neste caso a análise fatorial se presta para que se verifique a aderência dos dados ao modelo, ou teoria, proposto. O segundo tipo de aplicação aparece quando não se conhece, à *priori*, nenhum modelo para o fenômeno em questão. Aí a análise fatorial pode se prestar a uma análise exploratória dos dados coletados no sentido de se fazer conjecturas para que possam eventualmente indicar um caminho para que se proponha algum(s) modelo(s).

Alguns autores questionam a utilidade da análise fatorial devido a sua característica indeterminística, pela variedade dos métodos de derivação dos fatores, pela dependência do resultado em relação a escolha das variáveis a serem incluídas na análise. Por outro lado, o fato da análise fatorial não estar "amarrada" a nenhum modelo específico, e sim aos dados propriamente ditos, pode ser considerado como uma de suas qualidades.

Formalmente o propósito da análise fatorial é o de descrever satisfatoriamente a estrutura da matriz de covariâncias (correlações) de um conjunto grande de variáveis pelo menor número possivel de fatores subjacentes. Tais fatores podem ser considerados como variáveis aleatórias não observáveis diretamente, mas que podem ser expressas através de combinações lineares das variáveis originais (observáveis).

A intuição do modelo fatorial pode ser depreendida do seguinte argumento: as variáveis observadas são agrupadas de acordo com suas correlações de maneira que as que pertencem ao mesmo grupo são altamente correlacionadas entre si e possuam uma correlação baixa em relação as variáveis de outros grupos. Então é razoável se supor que cada um dos grupos de variáveis possam estar representando uma das dimensões (ou fator) do problema que está sendo estudado.

III.3 Alguns conceitos básicos

Nesta seção procura-se estabelecer uma notação a ser seguida, bem como definir os principais conceitos estatísticos básicos para o estudo da análise fatorial.

Normalmente a aplicação de alguma técnica de análise estatística é feita sobre um conjunto de dados onde são observadas p variáveis ou características em n indivíduos pertencentes a uma determinada população.

No caso as palavras indivíduo e população têm um significado mais amplo do que na linguagem corrente. População aqui se compreende como qualquer agregado sobre o qual se deseja fazer alguma inferência. Numa pesquisa sobre a agropecuária no estado do Rio de Janeiro, poder-se-ia definir população como o conjunto formado por todos os estabelecimentos que se ocupassem da exploração de algum ramo da atividade agrícola ou da pecuária. Outro exemplo seria a produção de parafusos de uma fábrica, sobre a qual se deseja estabelecer um controle de qualidade. Numa pesquisa demográfica a palavra população pode assumir seu significado corrente, ou seja, o conjunto de habitantes de uma determinada localidade.

A palavra indivíduo, então, serve para definir um determinado elemento de uma população, ou seja: um estabelecimento agropecuário, um parafuso, um habitante, etc... São também frequentemente usados para designar este conceito os termos unidade amostral ou, simplesmente, unidade.

Na aplicação a ser apresentada no presente trabalho, população é o conjunto formado pelos municípios brasileiros e os indivíduos são cada um dos municípios em questão.

Os dados a serem analisados podem ser representados por uma matriz $\mathbf{X}_{(n,p)}$, onde cada linha corresponde aos valores observados para as p variáveis de um dos n indivíduos estudados.

São apresentados a seguir, com a notação proposta, alguns dos conceitos estatísticos básicos que serão de utilidade no desenvolvimento do trabalho.

A média aritimética de determinada variável, X_j, para o presente conjunto de *n* indivíduos será dada por:

$$\bar{X}_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij} , \qquad j = 1, 2, \ldots, p$$

Às vezes é vantajoso se trabalhar com as observações centradas na média, ou seja:

$$x_{ij} = X_{ij} - \bar{X}_j, \qquad \begin{matrix} i = 1, 2, \ldots, n \\ j = 1, 2, \ldots, p \end{matrix}$$

A variância amostral de cada uma das variáveis é calculada por:

$$s_j^2 = \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2, \qquad j = 1, 2, \ldots, p$$

O estimador da variância calculado como acima é de máxima verossimilhança. Tal estimador é sabidamente viciado, sendo por isso usualmente substituído o denominador *n* por *(n-1)* para que o vício seja eliminado.

Para qualquer par de variáveis *j* e *k* a covariância entre as mesmas pode será calculada pela seguinte fórmula:

$$s_{jk} = \frac{1}{n} \sum_{i=1}^{n} x_{ij} x_{ik}, \qquad \begin{matrix} j = 1, 2, \ldots, p \\ k = 1, 2, \ldots, p \end{matrix}$$

A partir das covariâncias pode-se definir os coeficientes de correlação, ou seja:

$$r_{jk} = s_{jk} / s_j \, s_k \qquad \begin{matrix} j = 1, 2, \ldots, p \\ k = 1, 2, \ldots, p \end{matrix}$$

Tomando o desvio padrão amostral como unidade de medida para cada uma das respectivas variáveis envolvidas, tem-se as variáveis em

sua forma padronizada representadas por:

$$z_{ij} = x_{ij}/s_j \qquad \begin{matrix} i = 1, 2, \ldots, n, \\ k = 1, 2, \ldots, p \end{matrix}$$

Tradicionalmente adota-se letras gregas para representar os parâmetros populacionais correspondentes às estatísticas definidas anteriormente. Assim a média populacional da variável X_j será μ_j, a variância σ_j^2 e o coeficiente de correlação entre as variáveis X_j e X_k será denotado por ρ_{jk}.

Serão também usados alguns conceitos de álgebra matricial, que poderão ser vistos em *JOHNSON E WICHERN [6]*. Aqui vale a pena destacar dois de tais conceitos.

Definição 1: Seja $\mathbf{A}_{(k,k)}$ uma matriz quadrada e $\mathbf{I}_{(k,k)}$ uma matriz identidade. Os números denotados por $\lambda_1, \lambda_2, \ldots, \lambda_k$ que satisfaçam a equação :

$$\det(\mathbf{A} - \lambda \mathbf{I}) = 0$$

são chamados autovalores ou raízes características de $\mathbf{A}_{(k,k)}$. A equação $\det(\mathbf{A} - \lambda \mathbf{I}) = 0$ é chamada de equação característica da matriz $\mathbf{A}_{(k,k)}$.

Exemplo 1: seja a matriz

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 1 & 3 \end{bmatrix}$$

$$\det(\mathbf{A} - \lambda \mathbf{I}) = \begin{bmatrix} 1-\lambda & 0 \\ 1 & 3-\lambda \end{bmatrix} = 0$$

$$(1-\lambda)(3-\lambda) = 0$$

As raízes da equação característica da matriz dada são $\lambda_1 = 1$ e $\lambda_2 = 3$ e, portanto, tais números são os seus autovalores.

Definição 2: Seja $\mathbf{A}_{(k,k)}$ uma matriz quadrada e λ um de seus autovalores. Se $\mathbf{x}_{(k,1)}$ é um vetor não nulo tal que:

$$\mathbf{Ax} = \lambda \mathbf{x},$$

diz-se que $\mathbf{x}_{(k,1)}$ é um autovetor ou vetor característico da matriz $\mathbf{A}_{(k,k)}$.

Exemplo: seja a matriz dada no exemplo anterior e seja $\lambda = 3$

um de seus autovalores:

$$\begin{bmatrix} 1 & 0 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 3 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Então, tem-se oseguinte sistema de duas equações e duas incognitas:

$$\begin{cases} x_1 = 3\, x_1 \\ x_1 + 3x_2 = 3\, x_2 \end{cases}$$

Da primeira equação tem-se que $x_1 = 0$. Tomando-se $x_2 = 1$ (arbitrariamente), tem-se que $\mathbf{x}' = [0 \quad 1]$ é um autovetor ou vetor característico da matriz $\mathbf{A}_{(2,2)}$dada.

III.4 O modelo da análise fatorial

O modelo básico da análise fatorial deriva diretamente do objetivo principal desse tipo de técnica que é determinar a "melhor" representação das variáveis originais por meio de combinações lineares de $m << p$ fatores comuns, que são variáveis aleatórias não observáveis diretamente.

A "melhor" representação será aquela em que a matriz das covariâncias (ou correlações) calculada a partir do modelo seja mais próxima possivel da matriz calculada a partir dos dados originais.

Levando-se em conta as observações acima pode-se representar algebricamente o modelo da análise fatorial por:

$$z_i = \sum_{j=1}^{m} a_{ij} F_j + d_i U_i \qquad i = 1, 2, \ldots, p$$

Nota-se que cada uma das variáveis originais é descrita por uma combinação linear de m fatores comuns, F_j, $j = 1, 2, \ldots, m$, mais um fator que é específico à cada uma das variáveis.

Para maior facilidade, nas operações algébricas, trabalhar-se-á com as variáveis padronizadas, o que não leva a nenhuma perda de generalidade nos resultados obtidos. Considera-se também os F_j, $j = 1, 2, \ldots, m$, e os U_i, $i = 1, 2, \ldots, p$, variáveis aleatórias de média nula e variância unitária, sendo, ainda, os fatores comuns não correlacionados e os U_i independentes. A hipótese de não correlação dos fatores comuns será colocada visando apenas a facilidade de interpretação dos resultados a ser apresentados, sendo que de maneira geral pode-se estendê-los para o caso de fatores não ortogonais.

Assim colocado, o problema se resume à determinação dos valores dos coeficientes a_{ij} e d_i, $i = 1, 2, \ldots, p$ e $j = 1, 2, \ldots, m$, com os quais se consiga reproduzir da melhor maneira a matriz de covariâncias (ou correlações) dos dados originais.

Antes, porém, de atacar o problema do cálculo dos elementos acima (item III.7) serão mostrados alguns resultados importantes para o entendimento e utilização da análise fatorial como ferramenta.

Resultado 1: A variância da variável z_j pode ser expressa de acordo com a equação do modelo acima. Basta que se aplique a definição, lembrando que as variáveis são padronizadas:

$$s_j^2 = \frac{1}{n} \sum_{i=1}^{n} z_{ij}^2 \qquad j = 1, 2, \ldots, p$$

$$z_{ij} = \sum_{k=1}^{m} a_{jk} F_{ki} + d_j U_{ij} \qquad \begin{matrix} i = 1, 2, \ldots, n \\ j = 1, 2, \ldots, p \end{matrix}$$

$$z_{ij}^2 = \left(\sum_{k=1}^{m} a_{jk} F_{ki} \right)^2 + 2 d_j U_{ij} \sum_{k=1}^{m} a_{jk} F_{ki} + d_j^2 U_{ij}^2$$

$$s_j^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{k=1}^{m} a_{jk} F_{ki} \right)^2 + 2 d_j \sum_{k=1}^{m} a_{jk} \frac{1}{n} \sum_{i=1}^{n} F_{ki} U_{ij} + d_j^2 \frac{1}{n} \sum_{i=1}^{n} U_{ij}^2$$

Usando as hipóteses de não correlação, médias e variâncias dos fatores comuns e específicos, tem-se:

$$s_j^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{k=1}^{m} a_{jk} F_{ki} \right)^2 + d_j^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{k=1}^{m} a_{jk}^2 F_{ki}^2 + 2 \sum_{k<l=1}^{m} a_{jk} a_{jl} F_{ki} F_{li} \right) + d_j^2$$

$$= \sum_{k=1}^{m} a_{jk}^2 \frac{1}{n} \sum_{i=1}^{n} F_{ki}^2 + 2 \sum_{k<l=1}^{m} a_{jk} a_{jl} \frac{1}{n} \sum_{i=1}^{n} F_{ki} F_{li} + d_j^2$$

$$s_j^2 = \sum_{k=1}^{m} a_{jk}^2 + d_j^2 , \qquad j = 1, 2, \ldots, p$$

Sabe-se ainda que, se as variáveis são padronizadas, as variâncias são unitárias. Então:

$$\sum_{k=1}^{m} a_{jk}^2 + d_j^2 = 1$$

Desse modo a variância é decomposta em duas partes que representam respectivamente a proporção relativa à contribuição efetiva dos fatores comuns (*comunalidade*) e a variabilidade específica de cada variável.

Resultado 2: Define-se como comunalidade, h_j^2, $j = 1, 2, \ldots, p$, a variância da parte comum da expressão do modelo da análise fatorial, ou seja:

$$h_j^2 = \text{Var}\left(\sum_{k=1}^{m} a_{jk} F_k\right)$$

Como se trata de uma combinação linear de vetor de média nula, esta também tem média nula. Então:

$$h_j^2 = \sum_{k=1}^{m} a_{jk}^2 \text{Var}(F_k) + 2 \sum_{k<l=1}^{m} a_{jk} a_{jl} r_{F_k F_l}$$

$$h_j^2 = \sum_{k=1}^{m} a_{jk}^2$$

Como consequencia dos resultados acima tem-se que a proporção específica da variância de cada variável é dada por:

$$d_j^2 = 1 - h_j^2 \qquad j = 1, 2, \ldots, p$$

Resultado 3: A correlação entre duas variáveis z_j e z_l, $j, l = 1, 2, \ldots, p$, é expressa em função dos coeficientes dos fatores comuns como:

$$r_{z_j z_l} = \sum_{k=1}^{m} a_{ik} a_{lk} \qquad j, l = 1, 2, \ldots, p$$

Usando o fato de que as variáveis são padronizadas, calcula-se o coeficiente de correlação entre elas por:

$$r_{z_j z_l} = \frac{1}{n} \sum_{i=1}^{n} \left[\left(\sum_{k=1}^{m} a_{jk} F_{ki} + d_j U_{ij} \right) \left(\sum_{k=1}^{m} a_{lk} F_{ki} + d_l U_{il} \right) \right]$$

$$r_{z_j z_l} = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{k=1}^{m} a_{jk} F_{ki} \right) \left(\sum_{k=1}^{m} a_{lk} F_{ki} \right) + d_l \sum_{k=1}^{m} a_{jk} r_{F_k U_l} +$$

$$+ d_j \sum_{k=1}^{m} a_{lk} r_{F_k U_j} + d_j d_l r_{U_l U_j}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{k=1}^{m} a_{jk} a_{lk} F_{ki}^2 + 2 \sum_{k<h=1}^{m} a_{jk} a_{lh} F_{ki} F_{hi} \right)$$

$$= \sum_{k=1}^{m} a_{jk} a_{lk} \frac{1}{n} \sum_{i=1}^{n} F_{ki}^2 + 2 \sum_{k<h=1}^{m} a_{jk} a_{lh} \frac{1}{n} \sum_{i=1}^{n} F_{ki} F_{hi}$$

$$r_{z_j z_l} = \sum_{k=1}^{m} a_{jk} a_{lk}, \qquad j, l = 1, 2, \ldots, p$$

Resultado 4: a correlação entre uma dada variável X_j e um fator comum F_k é dada por:

$$r_{x_j F_k} = a_{jk}$$

Usando novamente a equação do modelo, tem-se:

$$r_{x_j F_k} = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{l=1}^{m} a_{jl} F_{li} + d_l U_{ij} \right) F_{ki}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{l=1}^{m} a_{jl} F_{li} F_{ki} + d_l U_{ij} F_{ki} \right)$$

$$= \sum_{l=1}^{m} a_{jl} \frac{1}{n} \sum_{i=1}^{n} F_{li} F_{ki} + d_l \frac{1}{n} \sum_{i=1}^{n} U_{ij} F_{ki}$$

$$= a_{jk} \frac{1}{n} \sum_{i=1}^{n} F_{ki}^2 + \sum_{l<>k}^{m} a_{jl} \frac{1}{n} \sum_{i=1}^{n} F_{li} F_{ki}$$

$$= a_{jk}, \qquad \begin{matrix} j = 1, 2, \ldots, p \\ k = 1, 2, \ldots, m \end{matrix}$$

Com estes quatro resultados pode-se estabelecer todas as relações para a interpretação dos resultados da análise fatorial.

III.5 O ajuste do modelo fatorial

Como foi visto, pode-se representar (ou reproduzir) as correlações observadas entre as variáveis estudadas por meio dos coeficientes do modelo fatorial. No caso de fatores ortogonais isso é obtido pela soma dos produtos dos coeficientes correspondentes às variáveis para as quais se deseja calcular a correlação. Denote-se por r_{ij} a correlação observada entre as variáveis z_i e z_j, e por r'_{ij} a mesma correlação reproduzida pelo modelo.

É certo que existe uma diferença numérica entre r_{ij} e r'_{ij}, visto que em quaisquer dados provenientes da observação de um experimento estão presentes ruídos quer sejam por questões ligadas aos problemas amostrais ou simplesmente erros de medida, ditos erros não amostrais. Por outro lado a introdução de um modelo sempre pressupõe algumas simplificações quer por facilidade de interpretação quer por desconhecimento de problemas adjacentes ao que está sendo estudado.

Dessa maneira pode-se verificar a qualidade do ajuste do modelo fatorial aos dados observados pela recomposição das correlações e verificação de quão próximos estão das correlações originalmente observadas.

Para isso define-se como correlação residual entre as variáveis z_i e z_j o seguinte valor:

$$\bar{r}_{ij} = r_{ij} - r'_{ij}$$

HARMAN[5] apresenta um teste simples para a análise deste resultado, que é baseado apenas no desvio padrão das correlações residuais e no número de observações que compõe o conjunto de dados analisado.

Por este teste tem-se que as correlações observadas e reproduzidas são próximas se:

$$\sigma_{\bar{r}} = 1/\sqrt{n}$$

onde: n é o número de observações do conjunto de dados;

$\sigma_{\bar{r}}$ é o desvio padrão das correlações residuais.

No caso de se ter um valor para $\sigma_{\bar{r}}$ maior que $1/\sqrt{n}$ pode-se admitir que é necessário adicionar mais fatores ao modelo e no caso de $\sigma_{\bar{r}}$ ser muito menor que $1/\sqrt{n}$ pode significar que o modelo tem fatores em excesso sendo considerados.

Outro aspecto a ser considerado, além da boa reprodução das correlações observadas, é o fato de que na análise fatorial não existe uma solução única para o problema mas sim uma variedade delas. Uma solução deve ser escolhida considerando-se vários fatores entre os quais uma boa interpretabilidade dos resultados e a simplicidade do modelo final, no qual uma característica desejável é que ele tenha um pequeno número de fatores comuns. Com a finalidade de se obter uma interpretação mais clara e adequada ao problema, após escolhido um método e calculada uma solução, pode-se, por meio de artifícios algébricos, rotacioná-la. Isso significa fixar uma posição desejada dos fatores no seu espaço *m-dimensional* que seja mais favorável a interpretação dos resultados.

III.6 Notação matricial

Até aqui foram apresentados o modelo fatorial bem como alguns resultados importantes para a análise fatorial, em notação algébrica comum o que ajuda na interpretação de tais resultados, à medida que pode-se isolar cada componente das respectivas fórmulas. A notação matricial que passará a ser usada doravante é mais compacta e muitas vezes facilita as operações na parte computacional, como, também, em algumas demonstrações.

Os conceitos básicos já vistos até o momento podem ser representados por meio de matrizes escolhidas adequadamente.

Para introduzir a notação matricial será considerado o caso de fatores ortogonais, sendo que o caso mais geral pode ser visto, por exemplo, em *HARMAN[5]*.

Seja o vetor aleatório $\mathbf{X}$ composto pelas variáveis $X_1, X_2, \ldots, X_p$ com vetor de médias μ e matriz de covariância $\mathbb{V}$.

Suponha ainda que estas variáveis são correlacionadas entre si e linearmente dependentes dos fatores $F_1, F_2, \ldots, F_m$, com $m \leq p$, e os fatores específicos $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$.

Então o modelo de análise fatorial pode ser escrito como:

$$
\begin{aligned}
X_1 &= a_{11}F_1 + a_{12}F_2 + \ldots a_{1m}F_m + \varepsilon_1 \\
X_2 &= a_{21}F_1 + a_{22}F_2 + \ldots a_{2m}F_m + \varepsilon_2 \\
&\vdots \\
X_p &= a_{p1}F_1 + a_{p2}F_2 + \ldots a_{pm}F_m + \varepsilon_p
\end{aligned}
$$

Ou mais sinteticamente:

$$\mathbf{X} - \mu = \mathbf{AF} + \varepsilon$$

onde: $\mathbf{X}_{(p,1)}$ é o vetor aleatório observável;

$\mu_{(p,1)}$ é o vetor de médias de $\mathbf{X}$;

$\mathbf{A}_{(p,m)}$ é a matriz dos coeficientes ou cargas fatoriais

$\mathbf{F}_{(m,1)}$ é o vetor dos fatores comuns; e

$\varepsilon_{(p,1)}$ é o vetor dos fatores específicos de cada variável.

Como já foi visto anteriormente os únicos elementos observáveis do modelo acima são as variáveis X_1, X_2,..., X_p. Porem com algumas hipóteses adicionais pode-se desenvolver métodos para a estimação da matriz **A** e do vetor de fatores específicos ε, além de se estabelecer relações para a análise da matriz de covariância.

Apresenta-se a seguir tais suposições na sua forma matricial:

$$E(\mathbf{F}) = \mathbf{0}_{(m,1)} \qquad Cov(\mathbf{F}) = \mathbf{I}_{(m,1)}$$

$$E(\varepsilon) = \mathbf{0}_{(p,1)} \qquad Cov(\varepsilon) = \Psi$$

$$\Psi = Diag(\psi_1, \psi_2, \ldots, \psi_p)$$

$$Cov(\varepsilon, \mathbf{F}) = \mathbf{0}_{(p,m)}$$

Dadas estas condições está completo o modelo para análise fatorial para *p* variáveis e *m* fatores ortogonais.

Ao invés de se trabalhar com o vetor **X**, pode-se alternativamente trabalhar com o vetor das variáveis padronizadas, **Z**, ou equivalentemente usar a matriz de correlações ρ no lugar da matriz de covariância $\mathbb{V}$.

Dessa maneira o modelo fatorial será representado por:

$$\mathbf{Z} = \mathbf{AF} + \varepsilon$$

Valem aqui as mesmas observações anteriores em relação ao vetor dos fatores comuns **F** e dos fatores específicos ε.

Os resultados derivados no item III.4 também podem ser mostrados matricialmente.

Resultado 1:

$$\begin{aligned}
Cov(\mathbf{X}) &= E(\mathbf{ZZ}') \\
&= E[(\mathbf{AF}+\varepsilon)(\mathbf{AF}+\varepsilon)'] \\
Cov(\mathbf{X}) &= E[(\mathbf{AF}+\varepsilon)(\mathbf{F}'\mathbf{A}'+\varepsilon')] \\
&= E(\mathbf{AFF}'\mathbf{A}'+\mathbf{AF}\varepsilon'+\varepsilon\mathbf{F}'\mathbf{A}'+\varepsilon\varepsilon') \\
&= \mathbf{A}E(\mathbf{FF}')\mathbf{A}'+\mathbf{A}E(\mathbf{F}\varepsilon')+E(\varepsilon\mathbf{F}')\mathbf{A}'+E(\varepsilon\varepsilon') \\
&= \mathbf{A}Cov(\mathbf{F})\mathbf{A}'+\mathbf{A}Cov(\mathbf{F},\varepsilon)+Cov(\varepsilon,\mathbf{F})\mathbf{A}'+Cov(\varepsilon) \\
&= \mathbf{AA}'+\Psi
\end{aligned}$$

Logo, para uma dada variável X_i, tem-se:

$$Var(X_i) = a_{i1} + a_{i2} + \ldots + a_{im} + \psi_i$$

O resultado 3 está mostrado, pois se $Cov(\mathbf{X}) = \mathbf{AA'} + \Psi$, então dadas as variáveis X_i e X_j a sua covariância será dada pelo produto interno das linhas *i* e *j* da matriz **A**, correspondentes às duas variáveis:

$$Cov(X_i, X_j) = a_{i1}a_{j1} + a_{i2}a_{j2} + \ldots + a_{im}a_{jm}$$

$$= \sum_{k=1}^{m} a_{ik}a_{jk}, \qquad i, j = 1, 2, \ldots, p$$

Lembre-se que Ψ é uma matriz diagonal e que pode-se, além disso, trabalhar com os dados padronizados onde as matrizes de correlação e covariância são identicas.

O resultado 2 diz respeito a variância da parte comum do modelo de análise fatorial.

$$\begin{aligned} Cov(\mathbf{AF}) &= E[(\mathbf{AF})(\mathbf{AF})'] \\ &= E(\mathbf{AFF'A'}) \\ &= \mathbf{A}E(\mathbf{FF'})\mathbf{A'} \\ &= \mathbf{AIA'} \\ &= \mathbf{AA'} \end{aligned}$$

Logo para uma dada variável X_i tem-se que sua comunalidade é dada pelo produto interno da linha *i* da matriz **A**:

$$h_i^2 = a_{i1}^2 + a_{i2}^2 + \ldots + a_{im}^2, \quad i = 1, 2, \ldots, p$$

O Resultado 4 apresenta a correlação entre uma dada variável , X_i, e um determinado fator comum, F_j.

$$\begin{aligned} Cov(\mathbf{X}, \mathbf{F}) &= E[(\mathbf{AF}+\psi)\mathbf{F'}] \\ &= E(\mathbf{AFF'}+\psi\mathbf{F'}] \\ &= E(\mathbf{AFF'})+E(\psi\mathbf{F'}) \\ &= \mathbf{A}E(\mathbf{FF'})+Cov(\psi, \mathbf{F}) \\ &= \mathbf{AI}+\mathbf{0} \\ &= \mathbf{A} \end{aligned}$$

Logo:

$$Cov(X_i, F_j) = a_{ij} \quad , \quad \begin{matrix} i = 1,2, \ldots, p \\ j = 1,2, \ldots, m \end{matrix}$$

Vê-se que através da notação matricial pode-se conseguir uma representação mais compacta e elegante de todos os resultados necessários à uma análise fatorial. No item que se segue passa-se a discutir os métodos de estimação usando-se, para tal, a notação aqui apresentada.

III.7 Métodos de estimação

Existe uma grande variedade de maneiras de se fatorar uma dada matriz de covariância (correlação) na busca de uma solução para o problema da análise fatorial. Alguns métodos necessitam que se tenha à *priori* uma estimativa inicial das comunalidades das variáveis em questão, enquanto que outros precisam que se saiba o número de fatores comuns que se deseja extrair. Cada um dos métodos pode ser aplicado a diversos tipos de problemas, sendo que alguns deles parecem contar com maior "simpatia" entre os usuários da análise fatorial. Pela própria leitura dos textos pode-se notar a preferência dos autores por determinados métodos. No livro de *HARMAN [5]* nota-se que o espaço dedicado e o número de aplicações dadas como exemplo, podem revelar a preferência do autor pelo Método do Fator Principal. Já no livro de *JOHNSON E WICHERN [6]* o mesmo ocorre em relação aos métodos da Maxima Verossimilhança e das Componentes Principais. Estes autores chegam mesmo a dizer textualmente que "em nossa opinião, os métodos de solução mais recomendados são o Método das Componentes Principais e o Método da Maxima Verossimilhança".

As razões para a escolha de um determinado método para a solução de um dado problema parecem ser de ordem prática ou as vezes até um tanto subjetivas. Uma questão de ordem prática é a disponibilidade de bons programas de computador, já que quaisquer um dos métodos possíveis de serem escolhidos requerem uma quantidade , e complexidade, de cálculos que os tornam praticamente impossiveis de serem aplicados sem ajuda de máquina. Por outro lado os autores mencionam a adequabilidade dos métodos aos problemas no sentido de produzirem resultados coerentes com a teoria subjacente, com uma interpretação razoável e clara desses resultados, e aí entra-se no campo da subjetividade. *HARMAN [5]* chega a relacionar determinados métodos como sendo adequados a solução de problemas em determinados campos do conhecimento.

Neste trabalho apresenta-se os três métodos que parecem ser os mais difundidos e de aplicação mais geral para a solução do problema da análise fatorial.

III.7.1 Método das componentes principais

O método das componentes principais, como o próprio nome indica, usa a teoria de Componentes Principais (ver por exemplo *JOHNSON E WICHERN [5], ANDERSON [19],* etc) para aproximar uma solução para o problema de análise fatorial. O modelo de componentes principais busca fazer uma rotação no sistema de coordenadas originais determinado pelo vetor aleatório $\mathbf{X} = [X_1, X_2, \ldots, X_p]$ de matriz de covariâncias $\mathbb{V}$. As novas coordenadas representam as direções de maior variabilidade e devem proporcionar uma descrição mais clara da estrutura de covariância do problema. As novas coordenadas são ortogonais e cada uma das p novas variávies por elas definidas são combinações lineares das p variáveis originais.

Usando a notação definida para a análise fatorial, e supondo as variáveis com média zero, o modelo é dado por:

$$\mathbf{X}_{(p,n)} = \mathbf{A}_{(p,p)} \mathbf{F}_{(p,n)}$$

Ou, em termos da matriz de covariância, tem-se que:

$$\mathbb{V}_{(p,p)} = \mathbf{A}_{(p,p)} \mathbf{A'}_{(p,p)}$$

Vê-se que, assim definido, o modelo é exato, não causando nenhuma redução na dimensão do problema nem dando nenhuma idéia sobre a estrutura dos fatores comuns e fatores específicos de cada variável.

O método das componentes principais aplicado a solução do problema de análise fatorial consiste em se trabalhar com apenas as $m < p$ primeiras componentes (a primeira componente principal é a de maior variabilidade, a segunda é a de maior variabilidade ortogonal à primeira, e assim sucessivamente) abandonando as últimas sob a hipótese de que a sua contribuição para a explicação da variabilidade pode ser considerada residual.

Para completar o modelo de análise fatorial define-se a matriz dos fatores específicos, Ψ, como sendo dada por:

$$\Psi = \begin{bmatrix} \psi_1 & & 0 \\ & \psi_2 & \\ & & \ddots \\ 0 & & \psi_p \end{bmatrix}$$

onde: $\psi_i = \sigma_{ii} - \sum_{j=1}^{m} a_{ij}^2 \qquad i=1,2,\ldots p$

A solução do problema de componentes principais (ver cap. 8 de *JOHNSON E WICHERN [6]*) é dada pela extração dos autovalores e autovetores da matriz de covariância (correlação) de modo que, pelo teorema da decomposição espectral, tem-se:

$$\mathbb{V} = \lambda_1 e_1 e_1' + \lambda_2 e_2 e_2' + \ldots + \lambda_p e_p e_p'$$

Onde: λ_i são os autovalores de $\mathbb{V}$, para $i = 1, 2, \ldots, p$;.

e_i são os autovetores ortonormais de $\mathbb{V}$, para $i = 1, 2, \ldots, p$.

Para adequar a solução ao problema da análise fatorial, basta definir a matriz dos coeficientes fatoriais, **A**, como:

$$\mathbf{A} = \left[\sqrt{\lambda_1}\, e_1,\ \sqrt{\lambda_2}\, e_2,\ \ldots,\sqrt{\lambda_p}\, e_p\right]$$

Desta maneira pode-se representar o modelo como:

$$\mathbb{V}_{(p,p)} \simeq \mathbf{A}_{(p,m)} \mathbf{A}'_{(m,p)} + \Psi_{(p,p)}$$

Em aplicações práticas o que se tem em mãos é a matriz de covariâncias observadas, **S**, ou correlações, **R**, quando se trabalha com as observações padronizadas, que são respectivamente os estimadores usuais dos parâmetros populacionais $\mathbb{V}$ e ρ. Serão mantidas as notações até aqui usadas para a matriz dos coeficiêntes ou cargas fatoriais, **A**, e dos erros específicos, Ψ.

Um elemento importante para a análise fatorial é saber da contribuição de cada um dos fatores na composição da variabilidade total

do problema. Para definir tal elemento lança-se mão do fato de que no método de solução por componentes principais os coeficiêntes de cada um dos fatores não se alteram quando o número de fatores considerados é aumentado, ou seja: ao ser considerado um modelo com m=1 a matriz dos coeficiêntes é dada por:

$$\mathbf{A} = [\sqrt{\lambda_1}\, e_1]$$

Se alternativamente resolve-se considerar m = 2, então:

$$\mathbf{A} = [\sqrt{\lambda_1}\, e_1,\ \sqrt{\lambda_2}\, e_2]$$

Em ambos os casos (λ_1, e_1) e (λ_2, e_2) são os pares compostos pelos primeiros autovalores e autovetores da matriz **S** (ou **R**).

Por outro lado a variância total do problema é dada pela soma:

$$s_{11} + s_{22} + \ldots + s_{pp} = \sum_{i=1}^{p} s_{ii}$$

Mas por definição (ver *JOHNSON E WICHERN [6]*) tem-se que:

$$tr(\mathbf{S}) = \sum_{i=1}^{p} s_{ii}$$

Tome-se então o modelo fatorial completo onde m=p, então:

$$\mathbf{S} = \mathbf{AA}'$$
$$\mathbf{S} = \lambda_1 e_1 e' + \lambda_2 e_2 e'_2 + \ldots + \lambda_p e_p e'_p$$

Então a matriz **S** pode ser escrita como:

$$\mathbf{S} = \mathbf{P}\Lambda\mathbf{P}'$$
onde: $\mathbf{P} = [e_1, e_2, \ldots, e_p]$
$\Lambda = diag[\lambda_1, \lambda_2, \ldots, \lambda_p]$

Então:

$$tr(\mathbf{S}) = tr(\mathbf{P}\Lambda\mathbf{P}')$$
$$tr(\mathbf{S}) = tr(\Lambda\mathbf{PP}')$$
$$= tr(\Lambda\mathbf{I})$$
$$= tr(\Lambda)$$
$$= \sum_{i=1}^{p} \lambda_{ii}$$

Logo:

$$\sum_{i=1}^{p} s_{ii} = \sum_{i=1}^{p} \lambda_{ii}$$

Para o caso de se trabalhar com a matriz de correlações, **R**, como a diagonal principal é composta de unidades, tem-se que a variabilidade total é igual ao número de variáveis, *p*, do problema.

A contribuição de um dos fatores comuns na variância total é dada por:

$$\sum_{i=1}^{p} a_{ij}^2 = \sum_{i=1}^{p} (\sqrt{\lambda_j}\, e_{ij})^2$$

$$= \lambda_j \sum_{i=1}^{p} e_{ij}^2$$

$$= \lambda_j$$

Portanto a contribuição relativa de cada um dos fatores para a variabilidade total do problema, quando se trabalha com a matriz de covariâncias, é dada por:

$$\lambda_j / \sum_{i=1}^{p} \lambda_i, \quad j = 1, 2, \ldots, p$$

Caso se trabalhe com as variáveis padronizadas, o cálculo se torna mais simples, ou seja:

$$\lambda_j / p, \qquad j = 1, 2, \ldots, p$$

III.7.1.1 A escolha do número de fatores

Não existe uma fórmula fechada para se determinar o número de fatores, m, que devem ser considerados numa análise fatorial, quando se extrai tais fatores pelo método das componentes principais. Em geral essa escolha pode ser baseada na teoria que envolve as variáveis do problema que está sendo estudado, em experiências anteriores ou de outras pessoas, ou, então, em regras práticas de uso bastante frequente. Aqui busca-se colocar algumas destas regras práticas.

A primeira delas é de se levar em conta a proporção da variância explicada pelos m primeiros fatores, ou seja: tomar m de modo que $\sum_{i=1}^{m}\lambda_i / \sum_{j=1}^{p}\lambda_j$ seja "grande". Obviamente que o conceito de "grande" depende do problema em questão e da sensibilidade de quem o está resolvendo.

Outra regra amplamente usada, que é de fácil implementação num programa de computador, é a de se considerar todos os fatores cujos respectivos autovalores sejam positivos no caso de se trabalhar com a matriz de covariâncias, ou maiores que 1 no caso de ser usada a matriz de correlações.

Pode-se, também, como já foi mencionado analisar a recomposição da matriz de correlações **R** (ou **S**), sendo que esta regra é a mais custosa de se aplicar até que se chegue ao resultado considerado satisfatório.

Nenhuma destas regras, ou qualquer outra que se possa conhecer, deve ser usada indiscriminadamente, mas sim levando-se em conta os três aspectos importantes numa análise fatorial, ou seja:

- ter um número pequeno de fatores;
- ter uma interpretação satisfatória e coerente do problema; e
- a parte da variância correspondente aos fatores abandonados, $\sum_{i=m+1}^{p}\lambda_i$, deve ser pequena.

III.7.2 Método do fator principal

O método do fator principal é uma forma de solução do problema da análise fatorial que consiste, basicamente, numa variação do método das componentes principais, onde se tem uma estimativa à *priori* do valor das comunalidades, h_i^2, $i = 1, 2, \ldots, p$.

Para se aplicar o método do fator principal deve-se trabalhar com os dados sumarizados pela matriz das correlações observadas, **R**. Supõe-se em seguida que seja possível obter-se uma estimativa das comunalidades das p variáveis do problema de modo que:

$$h_i^{*2} = 1 - \psi_i^*, \quad i = 1, 2, \ldots, p$$

Assim se obtém a matriz reduzida das correlações amostrais, subistituindo-se a diagonal principal da matriz **R** pelas correspondentes estimativas das comunalidades. A matriz reduzida das correlações terá, então, a seguinte forma:

$$\mathbf{R}_r = \begin{bmatrix} h_1^{*2} & r_{21} & \cdots & r_{p1} \\ r_{21} & h_2^{*2} & & \vdots \\ \vdots & & \ddots & \\ r_{p1} & r_{p2} & \cdots & h_p^{*2} \end{bmatrix}$$

Formulando o problema desta maneira todos os elementos da matriz $\mathbf{R}_r$ podem ser reproduzidos pelos fatores comuns, sendo que:

$$\mathbf{R}_r = \mathbf{A}_r^* \mathbf{A}_r^{*\prime}$$

A solução do problema é similar ao método descrito anteriormente, ou seja:

$$\mathbf{A}_r^* = \left[\sqrt{\lambda_1^*}\, e_1^*, \sqrt{\lambda_2^*}\, e_2^*, \ldots, \sqrt{\lambda_m^*}\, e_m^* \right]$$

$$\psi_i^* = 1 - \sum_{j=1}^{m} a_{ij}^{*2}$$

Os pares (λ_i^*, e_i^*) são os autovalores e correspondentes autovetores ortonormais da matriz reduzida das correlações amostrais, $\mathbf{R}_r$.

A discussão sobre a escolha do valor de *m*, aqui, é semelhante a feita para o método das componentes principais, sendo que deve ser levado em conta que pela substituição dos valores da diagonal principal de **R** pelas estimativas iniciais das comunalidades não há mais garantia de que os autovalores serão todos positivos. Sempre que o posto da matriz $\mathbf{R}_r$ possa ser determinado, esse valor pode ser assumido para *m*.

III.7.2.1 Escolha dos valores iniciais para as comunalidades

Para finalizar é necessário apresentar algumas maneiras de se estimar os valores iniciais das comunalidades para aplicação do método do fator principal. Não existem justificativas teóricas claras para a escolha desses valores iniciais, porém existem algumas estratégias práticas que costumam funcionar bem.

Uma dessas maneiras é tomar os valores iniciais de ψ_i^* como sendo o inverso do i-ésimo elemento da diagonal principal da matriz $\mathbf{R}^{-1}$. Dessa maneira tem-se que os valores das comunalidades serão estimados por:

$$h_i^{*2} = 1 - 1/r^{ii}, \quad i = 1, 2, \ldots, p$$

Esse valor coincide com o valor do quadrado do coeficiente de correlação múltipla da variável X_i em relação às outras p-1 variáveis.

SOUZA [20] apresenta um teorema interessante baseado no que ele chama de "regressão da imagem da resposta", para justificar esta maneira de escolher valores iniciais para as comunalidades.

Teorema: Seja a regressão da variável X_j , $j = 1, 2, \ldots, p$, sobre as demais p-1 variáveis restantes. Considere-se todas as variáveis padronizadas. Sob tais considerações:

$$h_j^{*2} \geq R_j^2 , \quad j = 1, 2, \ldots, p$$

Onde R_j^2 é o coeficiente de determinação (ou correlação múltipla) da respectiva regressão de X_j sobre as demais variáveis.

Uma outra maneira de se escolher os valores iniciais das comunalidades é extremamente mais simples não necessitando de cálculos adicionais pois já é dada pela própria matriz de correlações, ou seja:

$$h_i^{*2} = \max_{j \neq i} |r_{ij}| , \quad i = 1, 2, \ldots, p$$
$$j = 1, 2, \ldots, p$$

Quando se deseja trabalhar diretamente com a matriz de covariâncias observadas ao invés da matriz de correlações, pode-se substituir a diagonal principal de **S** pela diagonal principal da sua inversa $\mathbf{S}^{-1}$.

Uma outra possibilidade é usar as comunalidades calculadas através da aplicação do método das componentes principais como sendo os valores iniciais para o método do fator principal, entendendo esse procedimento como um refinamento da solução dada pelo primeiro método.

Independentemente da forma que se utilize para escolha dos valores iniciais das comunalidades, pode-se implementar o método do fator principal de forma iterativa onde a diagonal principal é substituida pelo valor das comunalidades calculadas no passo anterior. Um critério de convergência seria dado pela estabilidade das comunalidades resultantes. Para evitar a necessidade do "chute" inicial das comunalidades pode-se iniciar o processo com a matriz **R** (ao invés de $\mathbf{R}_r$) no primeiro passo, ou, em outras palavras, aplicar o método das componentes principais na primeira iteração.

HARMAN [5] dedica todo um capítulo de seu livro para a discussão do problema da comunalidade, incluindo várias maneiras para a sua estimação.

III.7.3 Método da maxima verossimilhança - preliminares

Até o momento, na apresentação dos métodos para solução do problema da análise fatorial, não foi preciso fazer nenhuma hipótese estatística sobre a distribuição das variáveis envolvidas. Com a introdução do método da maxima verossimilhançao isto se faz necessário. Em contrapartida a determinação do número de fatores a serem extraídos, até o momento, se baseou mais na intuição do que em fundamentos teóricos, o que não ocorre na presente método já que dele se pode derivar um teste para a hipótese de que a matriz de covariâncias (correlações) é satisfatoriamente recomposta pelo número, m, de fatores comuns levados em conta.

Antes de entrar no método propriamente dito é util relembrar o que vem a ser a função de verossimilhança para uma variável aleatória com distribuição de probabilidade Normal.

Seja X uma variável aleatória com distribuição normal $N(\mu,\sigma)$.

Sua função densidade de probabilidade é dada pela fórmula:

$$f(X;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\ \sigma} \exp\left\{-(\tfrac{1}{2}\sigma^{-2})(X - \mu)^2\right\}$$

A função de verossimilhança é definida como a distribuição conjunta de n obsevações indepemdentes e identicamente distribuídas da variável X, ou seja:

$$L(\mu,\sigma) = \prod_{i=1}^{n} f(X_i;\mu,\sigma)$$

$$L(\mu,\sigma) = \left[\frac{1}{\sqrt{2\pi}\ \sigma}\right]^n \exp\left\{-(\tfrac{1}{2}\sigma^{-2})\sum_{i=1}^{n}(X_i - \mu)^2\right\}$$

A máxima verossimilhança é atingida quando são encontrados estimadores de μ e σ que maximizem a função L. Um artifício algébrico usado para resolver este problema é tomar o logarítimo natural de L para se obter uma função linear para a qual a tarefa de maximização se torna menos trabalhosa e os resultados são equivalentes a quando se trabalha com a função original.

Então tem-se:

$$\ln L = -\frac{n}{2}\ln(2\pi) - n \ln \sigma - (1/2\sigma^2) \ln \sum_{i=1}^{n}(X_i - \mu)^2$$

Derivando-se parcialmente em relação a μ e σ, tem-se:

$$\frac{\partial \log L}{\partial \mu} = (\sigma^{-2}) \sum_{i=1}^{n}(X_i - \mu)$$

$$\frac{\partial \log L}{\partial \sigma} = -\left[\frac{n}{\sigma}\right] + (\sigma^{-3}) \sum_{i=1}^{n}(X_i - \mu)^2$$

Igualando-se a zero e resolvendo-se o sistema de equações resultante para μ e σ, tem-se os estimadores de máxima verossimilança para os dois parâmetros:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i \Rightarrow \hat{\mu} = \bar{X}$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n}(X_i - \bar{X})^2$$

Os estimadores de máxima verossimilahnça gozam de uma propriedade bastante importante e útil que é a chamada propriedade de invariância:

Propriedade: seja $\hat{\theta}$ o estimador de máxima verossimilhança do parâmetro θ de uma dada distribuição. Seja $h(\theta)$ uma função qualquer de θ. O estimador de máxima verossimilahnça da função $h(\theta)$ será dado por $h(\hat{\theta})$.

III.7.3.1 Método da máxima verossimilhança

Para enunciar o método da máxima verossimilhança para a solução do problema de análise fatorial faz-se então as hipóteses que tanto os fatores comuns F_i, $i = 1, 2, \ldots, m$, como os fatores específicos ε_j, $j = 1, 2, \ldots, p$, possuem distribuições normais.

Mas, pela construção do modelo de análise fatorial, as hipótese acima resultam que as váriáveis X_j, $j = 1, 2, \ldots, p$, sendo combinações lineares de variáveis normalmente distribuídas também possuem distribuição normal.

Admite-se, então, que a matriz **X** tem distribuição normal multivariada com vetor de médias μ e matriz de covariâncias $\mathbb{V}$.

Pode-se portanto escrever a função de verossimilhança, em sua forma matricial, como:

$$L(\mu,\mathbb{V})=K\exp\left\{-\frac{1}{2}tr\left[\mathbb{V}^{-1}\left[\sum_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x})'+n(\bar{x}-\mu)(\bar{x}-\mu)'\right]\right]\right\}$$

$$\text{Onde } K = \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}}$$

Algumas manipulações algébricas podem tornar a expressão um pouco mais simples.

$$L(\mu,\mathbb{V})=K\exp\left\{-\frac{1}{2}tr\left[\mathbb{V}^{-1}\sum_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x})'\right]\right\}\exp\left\{-\frac{n}{2}(\bar{x}-\mu)'\mathbb{V}^{-1}(\bar{x}-\mu)\right\}$$

$$L(\mu,\mathbb{V})=K\exp\left\{-\frac{n}{2}tr\left[\mathbb{V}^{-1}\ \mathbf{S}_n\right]\right\}\ \exp\left\{-\frac{n}{2}(\bar{x}-\mu)'\mathbb{V}^{-1}(\bar{x}-\mu)\right\}$$

$$\text{Onde } \mathbf{S}_n = \frac{1}{n}\sum_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x})' = \frac{n-1}{n}\mathbf{S}$$

S é a matriz de covariâncias observada

Portanto a função de verossimilhança pode ser escrita como:

$$L(\mu,\mathbb{V})=K\exp\left\{-\frac{n-1}{2}\ tr\left[\mathbb{V}^{-1}\ \mathbf{S}\right]\right\}\ \exp\left\{-\frac{n}{2}(\bar{x}-\mu)'\mathbb{V}^{-1}(\bar{x}-\mu)\right\}$$

Como $\mathbb{V} = \mathbf{AA'} + \Psi$ vê-se que a função de verossimilhança depende de $\mathbf{A}$ e Ψ.

Para que o modelo fique bem definido basta adicionar a condição que a matriz dada por $\mathbf{A'}\Psi^{-1}\mathbf{A}$ seja diagonal.

Dessa maneira o método da maxima verossimilhança para a resolução do problema da análise fatorial pode ser escrito da seguinte forma:

$$\text{Maximizar: } K\exp\left\{-\frac{n-1}{2}\,tr\left[\mathbb{V}^{-1}\,\mathbf{S}\right]\right\}\exp\left\{-\frac{n}{2}(\bar{x}-\mu)'\mathbb{V}^{-1}(\bar{x}-\mu)\right\}$$

$$\text{Sujeito a: } \mathbf{AA'} + \Psi = \mathbb{V}$$

$$\mathbf{A'}\,\Psi^{-1}\,\mathbf{A} = \Lambda, \text{ com } \Lambda \text{ matriz diagonal}$$

Note-se que não foi tomado o logarítimo da função $L(\mu,\mathbb{V})$, já que é possível resolver o problema por maximização numérica diretamente em $L(\mu,\mathbb{V})$. *JOHNSON E WICHERN [6]* discutem alguns aspectos computacionais deste problema e, também, remetem à bibliografia específica sobre o assunto.

Lançando mão da propriedade da invariância dos estimadores de máxima verossimilhança, pode-se então calcular os demais elementos necessários à análise fatorial. As comunalidades de cada uma das variáveis são função dos elementos da matriz $\mathbf{A}$, portanto:

$$\hat{h}_i^2 = \hat{a}_{i1}^2 + \hat{a}_{i2}^2 + \ldots + \hat{a}_{im}^2, \quad i = 1, 2, \ldots, p$$

Da mesma forma a variância explicada por cada um dos fatore comuns pode ser estimada por:

$$\frac{\hat{a}_{1j}^2 + \hat{a}_{2j}^2 + \ldots + \hat{a}_{pj}^2}{s_{11} + s_{22} + \ldots + s_{pp}}, \quad j = 1,2, \ldots, m$$

Até aqui foi apresentado o método da máxima verossimilhança trabalhando com a matriz de covariância $\mathbb{V}$. Como nos métodos anteriores o

mesmo pode ser feito usando-se a matriz de correlações observadas, em substituição à $n^{-1}(n-1)\mathbf{S}$. Os resultados assim obtidos são análogos aos calculados pela tranformação $\hat{\mathbf{A}}$ e $\hat{\Psi}$ conseguidos diretamente a partir da matriz **S**.

Sabe-se que a matriz de correlações é dada por:

$$\rho = \mathbf{V}^{-(1/2)}\mathbb{V}\mathbf{V}^{-(1/2)}$$

$$\text{Onde } \mathbf{V} = \text{diag}\left[\sigma_{11}, \sigma_{22}, \ldots, \sigma_{pp}\right]$$

Portanto:

$$\rho = \mathbf{V}^{-(1/2)}[\mathbf{AA}' + \psi]\mathbf{V}^{-(1/2)}$$

$$= \mathbf{V}^{-(1/2)}\mathbf{AA}'\mathbf{V}^{-(1/2)} + \mathbf{V}^{-(1/2)}\psi\ \mathbf{V}^{-(1/2)}$$

Mas:

$$\mathbf{V}^{-(1/2)} = \left[\mathbf{V}^{-(1/2)}\right]'$$

Então:

$$\rho = \mathbf{V}^{-(1/2)}\mathbf{A}\left[\mathbf{V}^{-(1/2)}\mathbf{A}\right]' + \mathbf{V}^{-(1/2)}\psi\ \mathbf{V}^{-(1/2)}$$

$$= \mathbf{A}_\rho\mathbf{A}_\rho' + \Psi_\rho$$

$$\text{Onde : } \mathbf{A}_\rho = \mathbf{V}^{-(1/2)}\mathbf{A}$$

$$\psi_\rho = \mathbf{V}^{-(1/2)}\psi\ \mathbf{V}^{-(1/2)}$$

Mais uma vez, pela invariância dos estimadores de máxima verossimilhança tem-se:

$$\hat{\mathbf{A}}_\rho = \hat{\mathbf{V}}^{-(1/2)}\hat{\mathbf{A}}$$

$$\hat{\psi}_\rho = \hat{\mathbf{V}}^{-(1/2)}\hat{\psi}\ \hat{\mathbf{V}}^{-(1/2)}$$

III.7.3.2 Teste para o número de fatores comuns

Como já foi dito na introdução do item 7.3, uma vantagem teórica que o método da máxima verossimilhança introduz é a possibilidade de se construir um teste para se testar a hipótese de que o número, *m*, de fatores comuns extraídos é estatisticamente adequado à solução do problema em pauta.

O teste usado se baseia no Teste da Razão de Máxima Verossimilhança, que pode ser visto com detalhes em *MOOD, GRAYBILL e BOES [17]* ou *MARDIA ET ALLI [21]*, em sua versão multivariada. Aqui será apresentada apenas as idéias gerais segundo a linha de *JOHNSON E WICHERN [6]*.

Seja o seguinte teste de Hipóteses:

$$H_0: \mathbb{V}_{(p,p)} = \mathbf{A}_{(p,m)}\mathbf{A}'_{(m,p)} + \Psi_{(p,p)}$$

H_1: $\mathbb{V}$ é uma matriz positiva definida qualquer

O teste da razão de máxima verossimilhança se baseia na razão definida por:

$$\lambda = \frac{\text{M. V. sob a hipotese}}{\text{M.V.}}$$

Então, para o caso em questão, sob a hipótese H_0 tem-se que a função de verossimilhança é proporcional à:

$$|\hat{\mathbf{A}}\hat{\mathbf{A}}' + \hat{\Psi}|^{-(n/2)} \exp\left\{-\frac{n}{2}\,\mathrm{tr}\left[\left[\hat{\mathbf{A}}\hat{\mathbf{A}}' + \hat{\Psi}\right]^{-1}\mathbf{S}_n\right]\right\}$$

Da mesma forma, sob H_1 temos a função de verossimilhança proporcional à:

$$|\mathbf{S}_n|^{-(n/2)} \exp\left\{-\frac{n}{2}\,\mathrm{tr}\left[\mathbf{S}_n^{-1}\mathbf{S}_n\right]\right\}$$

Onde $\mathbf{S}_n = \frac{n-1}{n}\mathbf{S}$ é o estimador de máxima verossimilhança da matriz de covariâncias.

Dessa maneira tem-se que:

$$\lambda = \frac{|\hat{A}\hat{A}' + \hat{\Psi}|^{-(n/2)} \exp\left\{-\frac{n}{2} \operatorname{tr}\left[\left(\hat{A}\hat{A}' + \hat{\Psi}\right)^{-1} S_n\right]\right\}}{|S_n|^{-(n/2)} \exp\left\{-\frac{n}{2} \operatorname{tr}\left[S_n^{-1} S_n\right]\right\}}$$

$$= \left[\frac{|\hat{A}\hat{A}' + \hat{\Psi}|}{|S_n|}\right]^{-(n/2)} \exp\left\{\frac{n}{2}\left[p - \operatorname{tr}\left[\left(\hat{A}\hat{A}' + \hat{\Psi}\right)^{-1} S_n\right]\right]\right\}$$

Com o intuito de simplificar o teste usualmente se calcula o valor de $-2\ln\lambda$, que resulta na seguinte expressão:

$$-2\ln\lambda = n \ln \frac{|\hat{A}\hat{A}' + \hat{\Psi}|}{|S_n|} + n\left\{\operatorname{tr}\left[\left(\hat{A}\hat{A}' + \hat{\Psi}\right)^{-1} S_n\right] - p\right\}$$

Mas *JOHNSON E WICHERN [6]* mostram que:

$$p = \operatorname{tr}\left[\left(\hat{A}\hat{A}' + \hat{\Psi}\right)^{-1} S_n\right]$$

$$-2\ln\lambda = n \ln \frac{|\hat{A}\hat{A}' + \hat{\Psi}|}{|S_n|}$$

BARTLETT [22] mostrou a seguinte aproximação para a distribuição de $-2\ln\lambda$:

$$[n-1\ (2p + 4m + 5/6)] \ln \frac{|\hat{A}\hat{A}' + \hat{\Psi}|}{|S_n|} \cong \chi^2_k(\alpha)$$

Onde: $k = [(p-m)^2 - p - m]/2$ é o número de graus de liberdade

α é o nível de significância do teste

Logo um teste para a hipótese H_0 com nível de significância α é dado por:

$$\begin{cases} \text{Rejeitar } H_0 \text{ se:} \\ \qquad [n-1\ (2p + 4m + 5/6)] \ln \dfrac{|\hat{A}\hat{A}' + \hat{\Psi}|}{|S_n|} > \chi^2_k(\alpha) \\ \text{Não rejeitar } H \text{ em caso contrário} \end{cases}$$

Para aplicação do teste acima devem ser observadas algumas condições a saber:

- os valores de n e $n-p$ devem ser "grandes";

-como o número de graus de liberdade de uma distribuição qui-quadrado é sempre positivo, a escolha do valor de m (número de fatores) deve acontecer de maneira que seja respeitada a seguinte relação:

$$\frac{1}{2}\,[(p-m)^2 - p - m] > 0$$

$$m < \frac{1}{2}\,(2p + 1 - \sqrt{8p + 1}\,)$$

III.8 Rotação dos fatores comuns

Muitas vezes após escolhido e aplicado um método para a solução do problema da análise fatorial chega-se a um conjunto de fatores comuns cuja interpretação não apresenta a desejada claresa. Podem ocorrer casos, por exemplo, onde uma ou mais variáveis apresentam coeficiêntes importantes para muitos fatores selecionados, dificultando dessa maneira que se possa atribuir significados distintos e claros a cada um deles, como seria desejável.

Com o intuito de dar uma solução a tal questão aparecem alguns métodos para que se faça uma rotação nos resultados iniciais buscando dar uma maior interpretabilidade aos fatores extraídos. Para *JOHNSON E WICHERN [6]* fazer uma rotação nos fatores equivaleria a ajustar o foco de um microscópio na tentativa de se observar melhor algum fenômeno.

Existem distintos métodos para rotação dos fatores inicialmente calculados que podem ser agrupados em duas classes de acordo com o ângulo de rotação: as rotações ortogonais e as obliquas. As rotações obliquas apresentam características que podem ser consideradas como complicadoras do ponto de vista da análise final dos resultados, pois, os fatores resultantes não sendo ortogonais não são mais não correlacionados e, consequentemente, a variância das variáveis originais já não podem mais ser obtidas diretamente pelos coeficiêntes fatoriais passando a depender, também, da correlação entre os fatores.

Devido ao exposto acima dedica-se mais atenção aos métodos de rotação ortogonal, através de suas duas variantes mais populares que são os métodos Varimax e Quartimax. Uma discussão bastante ampla e detalhada sobre a rotação obliqua pode ser vista no livro de *HARMAN [5].*

Antes de apresentar uma discussão de cada um dos métodos apresenta-se, brevemente, a base algébrica em que eles se fundamentam.

Seja **T** uma matriz quadrada ortogonal, o que implica que:

$$\mathbf{T}\mathbf{T}' = \mathbf{T}'\mathbf{T} = \mathbf{I}$$

onde **I** é a matriz identidade.

Pode-se então reescrever o modelo da análise fatorial, levando-se em conta que a matriz identidade é o elemento neutro da multiplicação de matrizes, como:

$$\begin{aligned} X &= ATT'F + \varepsilon \\ &= A^*F^* + \varepsilon \\ \text{onde } A^* &= AT \\ F^* &= T'F \end{aligned}$$

Algebricamente isso significa uma rotação rígida do sistema de cordenadas definido pelos fatores (eixos) iniciais, sendo que assim permanecem inalterados tanto as comunalidades como os fatores específicos de cada uma das variáveis.

É fácil de se ver que neste novo modelo as comunalidades não se alteram, pois estas são dadas pela diagonal principal da matriz produto da multiplicação da matriz dos coeficiêntes fatoriais pela sua tranposta e, portanto:

$$\begin{aligned} AA' &= AIA' \\ &= ATT'A' \\ &= A^*A^* \end{aligned}$$

Então:

$$\sum_{j=1}^{m} a_{ij}^{*2} = \sum_{j=1}^{m} a_{ij}^{2} \Longrightarrow h_i^{*2} = h_i^2, \quad i = 1, 2, \ldots, p$$

Os fatores específicos, ψ_i, associados a cada uma das variáveis também não se alteram, já que:

$$\begin{aligned} \psi_i^* &= \sigma_{ii} - \sum_{j=1}^{m} a_{ij}^{*2} \\ &= \sigma_{ii} - h_i^{*2}, \quad i = 1, 2, \ldots, p \end{aligned}$$

O problema então se resume a escolher adequadamente a matriz ortogonal **T**.

III.8.1 Rotação varimax

Este método de rotação foi proposto por *KAISER [23]* e seu objetivo é determinar uma matriz de rotação tal que a nova matriz de coeficiêntes resultante tenha sua estrutura simplificada no sentido das colunas, o que equivale a minimizar o número de variáveis importantes em cada um dos fatores comuns. De maneira geral, por este processo deve-se chegar a um grande número de elementos da matriz $\mathbf{A}^*$ com valores próximos a zero.

Como o que deve ser levado em conta é a grandeza de cada coeficiente, não importando o sinal, chega-se ao objetivo maximizando a variância dos quadrados dos coeficiêntes para cada um dos fatores:

$$V_j = (1/p) \sum_{i=1}^{p} [a_{ij}^{*2} - \bar{a}_j^{*2}]^2$$

Onde:

$$\bar{a}_j^{*2} = (1/p) \sum_{i=1}^{p} a_{ij}^{*2}$$

Expandindo a expressão de V_j e fazendo as substituições devidas chega-se a uma forma simplificada:

$$V_j = (1/p) \sum_{i=1}^{p} a_{ij}^{*4} - \left[(1/p) \sum_{i=1}^{p} a_{ij}^{*2} \right]^2$$

Para que todas a variáveis contribuam em igualdade para a solução do problema seria necessário que os coeficientes fossem normalizados. Mas a norma do vetor que descreve cada variável no novo espaço definido pelos *m* fatores comuns é dada pela raiz quadrada da respectiva comunalidade. Então basta dividir cada linha da matriz $\mathbf{A}^*$ por esse valor para obter-se todos os coeficientes normalizados.

Logo, a matriz **T** proposta como solução para o método varimax é aquela que maximiza a média da variação dos quadrados dos coeficientes, normalizados pela correspondente comunalidade, dentro de cada coluna, representada pela expressão abaixo:

$$V = \frac{1}{p} \sum_{j=1}^{m} \left[\sum_{i=1}^{p} d_{ij}^4 - \frac{1}{p} \left(\sum_{i=1}^{p} d_{ij}^2 \right)^2 \right]$$

$$\text{onde } d_{ij} = \frac{a_{ij}^*}{h_i}$$

A maioria dos "pacotes" estatísticos desenvolvidos para computador, que resolvem o problema de análise fatorial, têm implementados algorítmos para o cálculo da rotação ortogonal pelo método varimax.

III.8.2 Rotação quartimax

Este método de rotação tem como idéia básica fazer com que cada variável tenha o menor número possível de coeficientes altos nas matriz de cargas fatoriais. Em termos ideais, seria bom que cada uma das variáveis estudadas tivesse correlação alta apenas com um dos fatores comuns, fazendo com que $m - 1$ elementos da linha correspondente da matriz dos coeficientes fatoriais fossem iguais a zero. Na prática o que se busca é minimizar o número de coeficientes altos de cada linha da matriz ou, melhor dizendo, fazer com que cada variável seja descrita pelo menor número possível de fatores comuns.

Para se alcançar o objetivo acima deve-se levar em conta que se está interessado na grandeza dos coeficientes independentemente do sinal dos mesmos. Nesse sentido pode-se resolver o problema através da maximização da variância dos quadrados dos coeficientes. Assim uma solução para o método de rotação Quartimax é dada pela matriz de cargas fatoriais $\mathbf{A}^*$ que maximize a equação abaixo:

$$V = (1/mp) \sum_{i=1}^{p} \sum_{j=1}^{m} [a_{ij}^{*2} - \bar{a}^{*2}]^2$$

Onde:

$$\bar{a}^{*2} = (1/mp) \sum_{i=1}^{p} \sum_{j=1}^{m} a_{ij}^{*2}$$

Expandindo a expressão de V e fazendo as substituições de vidas chega-se, a seguinte simplificação:

$$V = (1/mp) \sum_{i=1}^{p} \sum_{j=1}^{m} a_{ij}^{*4} - (\bar{a}^{*2})^2$$

Já foi visto anteriormente que as comunalidades ($\sum_{j=1} a_{ij}^{*2}$) são invariantes para qualquer que seja a rotação, desde que ortogonal. Portanto a parcela dada por $(a_{ij}^{*2})^2$ será uma constante independente do padrão fatorial bastando, então, maximizar a soma definida por:

$$\sum_{i=1}^{p} \sum_{j=1}^{m} a_{ij}^{*4}$$

No trabalho de *KUBRUSLY [24]* são apresentadas com detalhes as soluções dos problemas de rotação Quartimax e Varimax.

CAPÍTULO IV

IV.1 Aplicação

Neste capítulo será apresentada uma aplicação das técnicas expostas nos capítulos II e III, a um conjunto de dados relacionados com o estudo dos problemas de saúde da população.

Os dados disponíveis para esse estudo foram coletados e organizados em arquivo magnético pelo Programa de Engenharia Biomédica da COPPE e dizem respeito à observações, para 59 cidades brasileiras, de 63 variáveis para um período que vai de 1960 até 1982. Nos Anexos A e B são apresentadas as variáveis, fontes de informação e as diversas unidades de medida usadas. No Anexo C está a lista das cidades consideradas.

Os critérios que orientaram a escolha das cidades, segundo *PANERAI [2]*, foram os seguintes:

- cidades com mais de 100000 habitantes, no Censo Demográfico de 1980, representativas dos grandes centros urbanos potencialmente sujeitos a problemas de saúde e que apresentam alta taxa de crescimento no período considerado;
- representatividade da população brasileira segundo os estados e grandes regiões geográficas; e
- facilidade para a coleta das informações.

Nestes critérios se encaixavam 60 municípios brasileiros, incluindo praticamente todas as capitais e cidades mais importantes do pais.

Dessas 60 cidades, inicialmente escolhidas, foi descartada uma pela impossibilidade prática que a mesma apresentava para a coleta de seus dados, resultando então as 59 já referidas.

Após digitados os dados foram submetidos a um processo de conferência visual para a descoberta e correção de possíveis erros de digitação. Em seguida fez-se uma normalização (não no sentido

estatístico) dos dados com a finalidade principal de amenizar a influência do número de habitantes. Para as cidades escolhidas a população total em 1980 variava de pouco mais de cem mil (Sumaré, SP) até cerca de oito e meio milhões de habitantes (São Paulo, SP).

Depois de normalizados verificou-se, ainda, a provável existência de valores suspeitos ("*outliers*"), os quais foram corrigidos se necessário.

Na aplicação aqui apresentada foram analisados os dados para o ano de referência de 1980.

Tais dados têm como referência o ano de 1980, mas na verdade consistem de valores médios das informações disponíveis para o período de cinco anos, compreendido pelos anos de 1978, 1979, 1980, 1981 e 1982. Tal procedimento se baseia em estudo referenciado por *PANERAI [2]*, que mostra ser razoável tal procedimento. *ALMEIDA [25]* em seu trabalho analisa este problema e conclui, para um conjunto de variáveis selecionadas do mesmo banco de dados aqui trabalhado, que de maneira geral a ordem das séries pode ser considerada como igual a 5, significando que ao se tomar a média aritmética de cinco anos consecutivos se obtém uma amostra independente da variável desejada.

IV.2 O Problema dos dados ausentes

Já foi mencionado na introdução desta dissertação que o principal problema para a análise dos dados acima descritos, como geralmente acontece na maioria das aplicações práticas das técnicas estatísticas, é a ausência de informações para determinadas células da matriz dos dados.

No sentido de se produzir alguma análise útil para a compreensão do problema em questão, procura-se então maneiras de se suprir essa ausência de dados com técnicas que levem em consideração os dados disponíveis.

Nos casos em que se trabalha com poucas variáveis, geralmente a ocorrência de *"missing values"* é pequena o que faz com que seja possível simplesmente abandonar aquelas observações que apresentam falhas e trabalhar apenas com as observções completas. Como se discutiu no item II.2.1, quando a falta de dados se dá completamente ao acaso, mesmo com a redução do tamanho da amostra representada por esta abordagem, pode-se ter estimativas de boa qualidade para os parâmetros desejados.

No caso aqui apresentado a opção acima fica prejudicada pois além de se ter um número considerável de variáveis, também o número de informações ausentes é bastante grande o que leva a uma quantidade de casos completos muito pequeno em relação ao número de variáveis.

Optou-se então pelo uso do algorítmo EM e do método das médias (veja os itens II.2.3 e II.2.4) para estimar os valores correspondente aos "buracos" da matriz de dados. O método EM apresenta algumas características desejáveis, ou seja: leva em conta toda a informação contida nos dados efetivamente observados já que as estimativas dos dados ausentes são, na verdade, a regressão linear sobre os dados presentes; propicia uma matriz de dados "cheia" para análises posteriores e produz estimativas de máxima verossimilhança para a matriz de covariâncias e para o vetor de médias. Contra o método está o fato de necessitar uma computação um tanto pesada, principalmente por se tratar de um processo iterativo.

Por sua vez o método das médias embora tenha a desvantagem de levar a uma subestimação da variabilidade das variáveis imputadas, pode ser muito util do ponto de vista prático já que sua aplicação é bastante simples, não necessitando de ferramentas computacionais sofisticadas.

Antes da aplicação efetiva dos dois processos de tratamento de valores ausentes relacionados acima, procedeu-se uma análise exploratória dos dados a qual é apresentada no item a seguir.

IV.2.1 Análise exploratória dos dados

Como já foi visto o algorítmo EM trabalha com a hipótese de que a matriz dos dados é formada por observações de uma variável normal multivariada. Nesse sentido é interessante verificar se, pelo menos, as distribuições marginais dessa variável não violam a hipótese de normalidade univariada(em outras palavras: se a distribuição de cada uma das variáveis isoladamente pode ser considerada normal).

Sabe-se que a normalidade das marginais não garante totalmente a hipótese de normalidade multivariada, porém pode fornecer dados razoáveis para as análises estatísticas usuais. *SILVA [7]* apresenta uma boa discussão sobre esta questão, remetendo à bibliografia adequada.

A análise exploratória dos dados visa, além da verificação da hipótese de normalidade, dar uma maior familiaridade aos dados, verificar a ocorrência de eventuais valores discrepantes e uma contagem dos *"missing values"*.

Para os dados referentes ao ano de 1980, após uma primeira análise, decidiu-se trabalhar apenas com aqueles indicadores que possuissem ao menos 30 valores presentes. Sob esta condição tem-se um conjunto de 49 variáveis. No Anexo B estão assinaladas as variáveis que foram selecionadas para este estudo, dentre as 63 originalmente existentes no arquivo.

Em seguida a matriz de dados foi submetida à PROC UNIVARIATE do *SAS* (*SAS INSTITUTE [26]*). Esse procedimento fornece uma série de elementos para uma análise exploratória dos dados como teste de normalidade, parâmetros de variabilidade, medidas de tendência central, vários percentis, valores mínimos e máximos, gráficos de ramo e folhas, *"box-plots"*, *"normal-plots"*, etc... No Anexo F é apresentada uma listagem padrão dessa *"procedure"* como ilustração. Um texto básico para entender os elementos da análise exploratória é o de *DACHS [27]*.

Através dessa análise verificou-se a ocorrência de alguns valores duvidosos para 6 variáveis (Quadro 1). Decidiu-se eliminar esses

pontos já que aparentemente estariam influindo negativamente no comportamento das variáveis.

Quadro 1

Pontos excluídos da análise para o ano de 1980

Variável	Cidade
Moradias com esgoto	Gravataí
Total de moradias	São Luiz
Moradias com água encanada	São Luiz e Uberaba
Número de telefones	Rio de Janeiro
Alfabetizados com mais de 5 anos	Sumaré
Número de estabelecimentos rurais	Mauá

A eliminação desses pontos se deve ao fato dos valores presentes estarem incompatíveis com a própria unidade de medida usada. No caso do número total de moradias para São Luiz, por exemplo, o valor registrado é 1811 enquanto o padrão usado é número de moradias por mil habitantes. Para a variável número de telefones no Rio de Janeiro tem-se registrado o valor de 850 telefones por mil habitantes, que embora não contrarie a unidade de medida parece um valor excessivamente alto já que para os outro municípios essa variável não ultrapassa o valor 220 telefones por mil habitantes.

Em seguida os dados foram novamente submetidos á análise exploratória.

O teste de normalidade univariada para cada uma das variáveis é dado pela estatística *D* de Kolmogorov quando existem mais de 50 observações, ou pela estatística *W* de *Shapiro e Wilks*, quando o número de casos não atinge 51 (veja *SAS INSTITUTE [26]*).

O Quadro 2 apresenta a lista das variáveis para as quais não se rejeita a hipótese de normalidade univariada, considerando as escalas originais em que foram medidas, mesmo para níveis de significância muito altos ($\alpha = 0,10$ por exemplo). No quadro vê-se que apenas a variável população de maiores de 65 anos de idade apresenta a significância do teste próxima de 10%, para as demais essa significância é muito maior.

Quadro 2

Variáveis para as quais não se rejeita a hipótese de normalidade na escala original

Variável	Descrição	PROB > *D*
V1	Nascidos vivos	>.15
V4	Mortalidade todas idades	>.15
V6	Mortalidade 5 a 19 anos	>.15
V7	Mortalidade 20 a 49 anos	>.15
V9	Mortalidade de maiores de 65 anos	>.15
V13	Mortalidade doenças cardio-vascular	>.15
V14	Mortalidade doenças respiratorias	>.15
V17	Mortalidade doenças neoplasicas	>.15
V18	Mortalidade acidentes de trafego	>.15
V27	População do município menores 1 ano	>.15
V28	População do município 1 a 4 anos	>.15
V29	População do município 5 a 19 anos	>.15
V30	População do município 20 a 49 anos	>.15
V31	População do município 50 a 65 anos	>.15
V32	População do munic. de 65 anos ou mais	0.12
V53	Pessoas com 8 ou mais anos de estudo	>.15
V54	Pessoas com 11 ou mais anos de estudo	>.15

No sentido de se alcançar a normalidade para as variáveis que não aparecem no Quadro 2 decidiu-se pela aplicação da transformação definida por *BOX E COX [28]*. Essa transformação é dada por:

$$X_{(\lambda)} = \begin{cases} (X^{\lambda} - 1)/\lambda & \text{para } \lambda \neq 0 \\ \log(X) & \text{para } \lambda = 0 \end{cases}$$

onde: $X_{(\lambda)}$ é o valor da variável X transformada
λ é o parâmetro da transformação

Para a aplicação da transformação de *Box e Cox* foi usada uma rotina desenvolvida por *SILVA [7]*, que busca um valor adequado para λ num intervalo dado pelo usuário, e fornece como saída um arquivo com os valores transformados das variáveis de entrada bem como outro arquivo com o valor de λ calculado para cada uma das variáveis.

Pela definição da transformação de *Box e Cox* dada acima, pode-se facilmente perceber que valores muito grandes de λ, principalmente quando a escala original da variável a ser tranformada também for grande, podem causar problemas numéricos quando se calcula a potência X^{λ}. Uma solução possível para esse problema é alterar a escala dividindo os valores originais por uma potência de dez.

No caso apresentado esse artifício foi usado para as variáveis V52 e V61 e mesmo assim os valores encontrados para λ foram razoavelmente grandes.

Para alguns casos pode-se não conseguir uma transformação satisfatória pelo método sugerido por *Box e Cox*, sendo então necessário buscar outro tipo de solução. Para algumas das variáveis aqui estudadas este problema ocorreu, e decidiu-se, então, pela aplicação da transformação logarítmica que, embora não tenha servido para torná-las normais, serviu ao menos para simetrizá-las.

No Quadro 3 são apresentados os resultados das transformações realizadas para cada uma das 32 variáveis que tiveram a hipótese de normalidade rejeitada em sua escala original.

Vê-se que para a grande maioria das variáveis a hipótese de normalidade não pode ser rejeitada mesmo para níveis de significância bastante altos com excessão de V8 cuja significância do teste de Kolmogorov de α = 6.5%. Para as variáveis assinaladas com um asterisco no Quadro 3, se conta com menos de 51 observações sendo então usada a estatística *K* de *Shapiro e Wilks*, sendo que a hipótese de normalidade não deve ser rejeitada para os valores do teste proximos de 1. Para as variáveis V11, V42 E V59 não se conseguiu a normalidade, porém a transformação logarítmica tornou as respectivas distribuições simétricas, condição considerada suficiente para aplicação do algorítmo EM.

Quadro 3

Resultado das transformações aplicadas a 32 variáveis selecionadas para o ano de 1980

Variáveis	Descrição	λ	PROB > *D*
V2	Trabalhadores em licenca	0.352	>.15
V3	Aposentadorias prematuras	0.448	>.15
V5	Mortalidade 1 a 4 anos	-0.039	>.15
V8	Mortalidade 50 a 65 anos	0.631	.065
V10	Mortalidade < de 1 ano	-0.083	>.15
V11	Obitos fetais	log	<.01
V12	Mort. d. infec. e paras.	-0.409	>.15
V15	Mort. d. resp. 0-1 ano	0.288	>.15
V16	Mort. d. resp. 1-4 anos	0.043	>.15
V19	Mort. causas violentas	0.040	>.15
V20	Mort. d. diarreicas	-0.015	>.15
V21	Mort. d. diarreicas 0-1 ano	log	>.15
V22	Num. de leitos	0.444	>.15
V33	Pop. rur. < de 1 ano	0.092	.82*
V34	Pop. rur. 1 a 4 anos	0.106	.92*
V35	Pop. rur. 5 a 19 anos	0.126	.49*
V36	Pop. rur. 20 a 49 anos	0.111	.60*
V37	Pop. rur. 50 a 65 anos	0.152	.36*
V38	Pop. rur. acima 65 anos	0.084	.64*
V41	Moradias c/ algum esgoto	2.438	>.15
V42	Total de moradias	log	<.01
V43	Moradias c/agua encanada	1.296	>.15
V44	Moradias c/ poco	0.232	>.15
V48	Num. de telefones	0.099	>.15
V49	Area do municipio	-0.073	>.15
V52	Alfabetizados > 5 anos	5.359	>.15
V55	Total de terra arável	log	>.15
V56	Total de terra cultivada	log	>.15
V59	Culturas resp. 80% da produção	log	<.01
V60	Num. estab. rurais	0.312	0.39*
V61	Num. de empregos reg.	3.742	>.15
V62	População total do mun.	-0.455	>.15

* Variáveis com menos de 51 observações: nestes casos a estatística do teste é a *W* e o teste é PROB < *W*

IV.2.2 Aplicação do algorítmo EM e do método das médias

Após a análise exploratória dos dados, apresentada anteriormente, a matriz dos dados composta pelas 17 variáveis cuja normalidade foi aceita em sua escala original e pelas 32 que necessitaram de transformações apresenta a configuração dada no Quadro 4, em relação aos valores ausentes.

Quadro 4

Número de valores ausentes por município

Valores Ausentes	Nº de Casos	% do Total	Nº Acumulado	% Acumulada
0	35	59.3	35	59.3
1	4	6.8	39	66.1
2	3	5.1	42	71.2
3	3	5.1	45	76.3
4	1	1.7	46	78.0
6	4	6.8	50	84.7
7	3	5.1	53	89.8
8	3	5.1	56	94.9
9	2	3.4	58	98.3
10	1	1.7	59	100.0

A matriz de dados com a configuração acima foi então submetida ao algorítmo EM e ao método das médias para que fossem estimados valores para preencher os "buracos" causados pela falta de dados.

Para aplicação do método EM o vetor de médias inicial foi estimado a partir dos dados disponíveis para cada uma das variáveis.

A estimativa inicial para a matriz de covariâncias foi calculada a partir dos dados completos pela aplicação do método das médias. Optou-se por esta alternativa, preencher as lacunas com as médias respectivas a cada uma das variáveis com dados ausentes, para evitar a possibilidade de uma matriz singular como entrada do algorítmo EM o que poderia ocorrer pelo método *Pairwise* ou mesmo se fossem usados apenas os casos completos, já que o número deles, 35, é menor que o número total de variáveis, 49 (ver *JOHNSON E WICHEN [6]*).

Considerando-se como critério de parada que a maior diferença relativa entre as estimativas das médias ou covariâncias não ultrapassasse 1,5% em relação a iteração anterior, foram necessárias mais de 850 iterações do algorítmo EM. Esse número elevado provavelmente se deve a quantidade muito grande de variáveis consideradas já que em outros teste feitos com o mesmo programa mas com poucas variáveis a convergência, mesmo para critérios um pouco mais rígidos, se deu com menos de 25 iterações. Por outro lado foram feitos testes rodando uma análise fatorial com resultados de 50 iterações do algorítmo EM, e as cargas fatoriais obtidas não tiveram alterações significantes em relação aos resultados que serão analisados neste texto.

No Anexo F.1 são apresentados os desvios padrão calculados para as variáveis após a imputação pelos dois métodos escolhidos para o tratamento dos valores faltantes, bem como a respectiva diferença relativa à mesma medida calculada a partir dos dados realmente observados, para que se possa avaliar as diferenças entre os resultados.

Verifica-se que a variabilidade das variáveis é sempre reduzida no caso do uso do método das médias, sendo tal fato mais acentuado para os casos onde o número de dados ausentes é maior. Tal resultado é bastante intuitivo visto que os valores perdidos são substituídos por um valor constante, não importando a informação que pode ser agregada a partir das demais variáveis que possuam valores observados.

Por seu lado o algoritmo EM ao imputar o *"missing value"* pela regressão linear sobre as variáveis com valores presentes está levando em conta tais informações adicionais.

O método EM, sob a condição de que a ocorrência de *"missing values"* é completamente aleatória, produz estimativas não viciadas para tais valores ausentes. Ainda assim nota-se que quanto maior o número de observações perdidas menor será a qualidade dessas estimativas, resultado que pode ser considerado coerente com a teoria da amostragem já que quando se trabalha com amostras maiores espera-se obter uma precisão mais elevada.

No Anexo E.2 são apresentados gráficos para que se possa visualizar a diferença nos resultados da aplicação dos dois métodos de tratamento de dados ausentes. Neles se pode ver claramente o efeito de se substituir os valores desconhecidos por uma constante, como no caso do método das médias.

Analisando tais resultados pode-se depreender, intuitivamente, que quando o número de dados ausentes for pequeno qualquer método de tratamento do problema de dados ausentes pode, pelo menos em termos práticos, ser empregado sem que se obtenha diferenças marcantes nas imputações feitas.

Pode-se recuperar a escala original das variáveis transformadas para a matriz de dados completa. Para isso deve-se lançar mão dos valores de λ que são armazenados durante o processo da tranformação de *Box e Cox*. No caso aqui apresentado optou-se por aplicar a análise fatorial sobre os dados transformados, já que a normalidade dos dados é uma propriedade desejável à maioria das técnicas de análise estatística de dados.

IV.3 Resultados da análise fatorial

Os dados completos pela aplicação do algorítmo EM e pelo método das médias foram então submetidos à análise fatorial.

Com a finalidade de se ter uma comparação da aplicação de técnicas variadas decidiu-se pelo método das componentes principais e método do fator principal em sua forma iterativa. Para ambos foram aplicadas as rotações Varimax e Quartimax.

Optou-se pelo uso do *SAS* para a realização da análise fatorial, já que este *Software* dispõe de várias alternativas de métodos de extração de fatores, além de propiciar farta impressão de resultados parciais, gráficos, etc, que facilitam muito o trabalho de análise. O *Systat,* que foi usado nas primeiras explorações feitas nos dados, é um pacote muito mais limitado, de utilização mais complicada, processamento muito lento e só possibilita a aplicação do método das componentes principais. O *SAS*, tanto na versão para computadores de grande porte como para micro computadores, pela sua abrangência e arquitetura proporciona um uso integrado de suas funções, possibilitando bastante agilidade no trabalho de análise.

IV.3.1 Aplicação do método das componentes principais

Para aplicação do método das componentes principais a escolha do número de fatores comuns foi pelo critério dos autovalores maiores que a unidade.

Esse critério, para a matriz dos dados completada pelo algoritmo EM, gerou um total de 9 fatores comuns a serem considerados. A matriz dos coeficientes fatoriais (matriz **A**, definida no capítulo III), após aplicada a rotação Varimax nos fatores extraídos originalmente, é mostrada no Quadro 5[1]. Foram omitidos os coeficientes menores que 0,5, em valores absolutos, para melhorar a visualização dos resultados, sendo que a matriz completa aparece no Anexo G.1. Também para facilitar a visualização os valores dos coeficientes foram multiplicados por 100 e arredondados para o inteiro mais próximo (*SAS INSTITUTE [29]*).

Analisando as correlações entre as variáveis e os fatores comuns pode-se associar um provável significado para cada um desses fatores, de acordo com o grupo de variáveis que possuam alta correlação (positiva ou negativa) com cada um deles.

O grupo das variáveis com correlação alta em relação ao fator de número 1 é composto basicamente pelos indicadores de mortalidade em idades adultas e suas causas de morte (cancer e doenças cardio vasculares). Em contraposição aparecem com correlações altamente negativas as variáveis que quantificam as populações mais jovens (idade menor que 20 anos). Escores altos para este fator devem significar municípios com uma população predominantemente mais idosa com grande ocorrência de mortes nas idades mais altas, por doenças cárdio-vasculares ou doenças neoplásicas. Também há uma variável que indica uma tendência de aumento do escore para as cidades maiores, que é a variável total de moradias. Uma variável com correlação moderada (43%) com este fator é mortalidade por doenças respiratórias.

[1] As variáveis são representadas por meneumônicos para facilitar a leitura da tabela.

Quadro 5

Coeficientes fatoriais para o ano de 1980
Método das componentes principais - rotação Varimax
(Algoritmo EM)

Variáveis	Fatores									h^2(%)
	1	2	3	4	5	6	7	8	9	
MORT50-65	95									91
DVASC	87									87
DNEOPL	85									88
POP50-65	83									95
POP65+	71				51					95
MORT20-49	71									84
TOTMORA	56									69
MORT65+	51									73
VIOLENT										76
POP1-4	-74									93
POP<1	-78									90
MORT1-4		92								90
DIARRE		83								90
DIAR<1		81								89
POP5-19	-61	67								95
MORT<1		67								85
DRES1-4		63		-52						84
DINFEC		61								77
MORTGER		57								87
EMPREGO										69
MORESG	50	-71								83
ALFABET		-76								93
POP20-49		-80								91
RUR50-65			99							99
RUR5-19			99							99
RUR65+			98							99
RUR20-49			98							99
RUR1-4			98							98
RUR<1			98							97
CULTIVA				90						89
ARAVEL				89						88
AREAMUN				72						76
CULT80%				62						66
ESTRUR				57						56
DRESP				-73						85
DRES<1				-81						86
ESTU11+					93					92
ESTU8+					90					91
LEITOS					67					78
TELEFO					64					83
AGENCAN					61					72
POPTOT					52					77
POÇO					-63					75
APOSPRE						85				86
TRABLIC						80				80
ACTRAF							87			83
MORT5-19		56					57			87
NASCVIV								81		70
OBIFET										73
% Var.	17	16	13	12	11	5	5	3	2	85

O segundo fator agrupa com correlações bastante significativas a mortalidade da população mais jovem com as causas de morte mais comuns para essa faixa etária (diarréias, doenças respiratórias e infecciosas e parasitárias). Por outro lado tem-se um indicador de alfabetização e outro de estrutura urbana (moradias com esgoto) com correlações inversas altas, o que parece ser bastante razoável, já que se pensa que tais causas de morte estão ligadas a fatores de subdesenvolvimento como falta de condições de saneamento. Um escore alto do fator dois, dessa forma pode significar municípios com mortalidade alta devido às causas citadas e em contrapartida condições básicas de saneamento deficitárias.

O fator 3 é altamente correlacionado com as variáveis que quantificam a população rural do município, não se destacando nenhuma variável com correlação negativa.

O quarto fator destaca o grupo de variáveis relativas à estrutura agrária dos municípios, ou seja: área total, área arável, área cultivada, além do número de estabelecimentos rurais e número de culturas responsáveis por 80% do valor da produção agrícola. Em contrapartida tem-se as variáveis sobre mortalidade por doenças respiratórias com correlações negativas, o que deve representar o fato dessas doenças estarem mais ligadas a fenômenos urbanos como a poluição.

No fator 5 nota-se a presença mais marcante das variáveis que se relacionam com infraestrutura urbana, como nível de escolaridade mais alto, número de telefones e de leitos hospitalares. Aparece com correlação negativa a variável número de moradias com poço, que pode indicar uma estrutura urbana mais pobre.

A interpretação dos demais fatores não parece ser de grande valia para o conhecimento do problema, já que cada um se relaciona com poucas variáveis sendo, mesmo, que o fator 9 não apresenta correlação alta (>0,5) com nenhuma das variáveis em questão. Talvez caiba destacar o fator 6 que agrupa as variáveis TRABLIC e APO, ambas relacionadas à said prematura (momentânea ou não) do mercado de trabalho.

Há que se destacar, também, o fato de algumas variáveis (óbitos fetais, mortalidade por causas violentas e número de empregos

regulares) não apresentarem correlações altas com nenhum dos 9 fatores em questão. Sabe-se que é tradicional, pelo menos em nosso país, a baixa qualidade do registro particularmente dos óbitos fetais e das mortes violentas, sendo comum se ter notícia pela imprensa da descoberta de cemitérios clandestinos.

Usando-se alternativamente o método de rotação Quartimax, obteve-se a matriz de cargas fatoriais mostrada no Quadro 6.

Parece não haver alterações significativas a nível de mudar a interpretação dos fatores comuns. Mesmo a distribuição dos pesos, no sentido da parcela explicada da variância por cada um dos fatores, não se altera de maneira muito intensa, havendo apenas um pequeno aumento de concentração nos primeiros fatores, no caso da rotação Quartimax.

Quanto a quantidade de variância explicada pode-se ver que, ao serem considerados todos os 9 fatores, é de cerca de 85%. No caso de se optar por trabalhar apenas com os oito primeiros fatores, já que o fator 9 não possui correlação alta com nenhuma variável, essa explicação se reduz para cerca de 82,5%.

A matriz dos coeficientes completa, para a rotação Quartimax, pode ser vista no Anexo G.2.

Usando os dados com imputação feita pelo método das médias pode-se dizer que, em termos práticos, não ocorreram grandes alterações nos resultados da análise fatorial tanto para a rotação Varimax quanto para a Quartimax.

Ao nível da interpretação dos fatores verifica-se que os significados a ele atribuídos anteriormente podem ser mantidos, apesar de alterações no valor das cargas fatoriais e na ordenção das variáveis dentro de cada um dos fatores.

Analisando a parcela da variação explicada pelos fatores retidos ve-se que totaliza cerca de 84%, sendo que há uma concentração maior nos primeiros fatores.

Quadro 6

Coeficientes fatoriais para o ano de 1980
Método das componentes principais - rotação Quartimax

Variáveis	Fatores									h^2(%)
	1	2	3	4	5	6	7	8	9	
MORT50-65	94									91
DNEOPL	89									88
POP50-65	88									95
DVASC	88									87
POP65+	76				51					95
MORT20-49	72									84
TOTMORA	59									69
MORT65+	52									73
VIOLENT										76
POP5-19	-69	62								95
POP1-4	-82									93
POP<1	-83									90
MORT1-4		92								90
DIARRE		82								90
DIAR<1		80								89
MORT<1		68								85
DRES1-4		65		-50						85
DINFEC		65								77
MORT5-19		63					54			87
MORTGER		62								87
EMPREGO										69
MORESG	57	-66								83
ALFABET	52	-71								93
POP20-49		-75								95
RUR50-65			99							99
RUR5-19			99							99
RUR65+			98							99
RUR20-49			98							98
RUR1-4			98							98
RUR<1			98							97
CULTIVA				90						89
ARAVEL				89						88
AREAMUN				73						76
CULT80%				62						66
ESTRUR				58						56
DRESP				-71						85
DRES<1				-80						86
ESTU11+					91					92
ESTU8+					87					91
TELEFO					63					83
LEITOS					62					78
AGENCAN					58					72
POPTOT										77
POÇO				-63					75	
APOSPRE						84				86
TRABLIC						79				80
ACTRAF		56					87			83
NASCVIV								81		71
OBIFET										73
% Var.	19	16	13	12	10	5	4	3	2	85

Como já foi dito anteriormente embora o método do algoritmo EM se fundamente em pressupostos teóricos mais bem definidos, o método das médias pode ser usado, desde que com o devido cuidado, já que pode propiciar resultados úteis em termos práticos, além de ser sua aplicação bastante simples e rápida.

No Anexo G são apresentadas as matrizes de coeficientes fatoriais relativas às aplicações da análise fatorial para os dados tratados pelo método das médias (G.5 até G.8).

IV.3.2 Aplicação do método do fator principal

Para a aplicação deste método de solução do problema de análise fatorial (veja o item III.7.2) optou-se pela sua forma iterativa, o que elimina a necessidade de estimação à *priori* da comunalidades. Com o intuito de poder comparar os resultados com aqueles obtidos pelo método das componentes principais decidiu-se fixar o número de fatores retidos, também, em $m = 9$. Caso se optasse por só trabalhar com os fatores correspondentes a autovalores maiores que 1, seriam retidos apenas 8 fatores comuns.

A matriz resumida dos coeficientes fatoriais gerada a partir da aplicação da rotação Varimax aos fatores originalmente obtidos é mostrada no Quadro 7. No Anexo G.3 é apresentada a matriz completa das cargas fatoriais.

Pode-se ver que aqui, também, não desponta nenhuma diferença marcante em relação as resultados já apresentados no que diz respeito a interpretação dos fatores comuns.

As diferenças ficam por conta de um maior número de variáveis sem apresentar alta correlação com os fatores comuns retidos. Além das já citadas não há coeficientes altos para: nascidos vivos, população total e mortalidade acima de 65 anos.

A variável população total mesmo quando aparece, no método das componentes principais, tem correlação relativamente baixa (52%) com o fator 5. Isso deve ocorrer pelo fato dessa variável ter sido usada na normalização de várias outras variáveis, diminuindo assim sua influência.

Por sua vez, a variável nascidos vivos quando aparece está isolada num dos fatores (fator 8), não se agrupando com nenhuma das demais variáveis.

Quadro 7

Coeficientes fatoriais para o ano de 1980
Método do fator principal - iterativo - rotação Varimax

Variáveis	Fatores									h^2(%)
	1	2	3	4	5	6	7	8	9	
MORT50-65	95									91
DVASC	86									86
DNEOPL	84									87
POP50-65	83									97
POP65+	71				52					98
MORT20-49	69									81
TOTMORA	53									60
MORT65+										68
VIOLENT										68
POP1-4	-74									94
POP<1	-77									90
MORT1-4		91								89
DIARRE		82								92
DIAR<1		80								91
POP5-19	-60	68								97
MORT<1		64								83
DRES1-4		61		-52						85
DINFEC		57								68
MORTGER		56								86
EMPREGO										52
MORESG		-71								81
ALFABET		-78								95
POP20-49		-82								97
RUR50-65			99							100
RUR5-19			99							100
RUR65+			98							99
RUR20-49			98							99
RUR1-4			98							98
RUR<1			97							97
CULTIVA				90						89
ARAVEL				89						87
AREAMUN				72						72
CULT80%				59						52
ESTRUR				53						46
DRESP				-71						84
DRES<1				-79						84
ESTU11+		-54			95					94
ESTU8+					92					94
LEITOS					63					70
TELEFO					62					82
AGENCAN					58					67
POPTOT										67
OBIFET										50
POÇO					-59					66
APOSPRE						81				81
TRABLIC						73				71
ACTRAF		56					81			73
MORT5-19							58			86
NASCVIV										27
% Var.	16	16	13	12	11	4	4	2	2	81

A rotação Quartimax (Quadro 8) também neste caso não altera nenhum ponto importante da análise, ficando a diferença por conta da variável mortalidade acima de 65 anos aparecer no grupo de variáveis correlacionadas com o fator 1, embora ainda com peso relativamente baixo (correlação de 51%). Também, em relação a quantidade de variância explicada, o que se nota é um pequeno aumento de concentração nos primeiros fatores.

O total de variância explicada considerando-se os 9 fatores extraídos é de cerca de 81%

A matriz completa dos coeficientes fatoriais para a rotação Quartimax está no Anexo G.4.

Quadro 8

Coeficientes fatoriais para o ano de 1980
Método do fator principal - iterativo - rotação Quartimax

Variáveis	Fatores									h^2(%)
	1	2	3	4	5	6	7	8	9	
MORT50-65	94									86
DNEOPL	89									87
POP50-65	88									97
DVASC	87									86
POP65+	77									98
MORT20-49	71									81
TOTMORA	57									60
MORT65+	51									68
VIOLENT										68
POP5-19	-69	63								97
POP1-4	-81									94
POP<1	-83									90
MORT1-4		92								89
DIARRE		82								92
DIAR<1		79								91
MORT<1		66								83
DRES1-4		64		-51						85
DINFEC		62								68
MORT5-19		61					54			86
MORTGER		61								86
EMPREGO										52
MORESG	57	-66								81
ALFABET	51	-72								95
POP20-49		-77								97
RUR50-65			99							100
RUR5-19			99							100
RUR65+			98							99
RUR20-49			98							99
RUR1-4			98							98
RUR<1			97							97
CULTIVA				91						89
ARAVEL				90						87
AREAMUN				73						72
CULT80%				58						52
ESTRUR				54						46
DRESP				-70						84
DRES<1				-79						84
ESTU11+					92					94
ESTU8+					89					94
TELEFO					61					82
LEITOS					59					70
AGENCAN					54					67
POPTOT										67
OBIFET										50
POÇO					-59					66
APOSPRE						80				81
TRABLIC		56				72				71
ACTRAF							81			73
NASCVIV										27
% Var.	19	15	13	12	10	4	4	2	2	81

IV.3.3 Análise das cidades em relação aos fatores

A análise realizada até o momento se restringiu à discussão sobre a dimensão do problema que diz respeito às variáveis envolvidas. Nesse sentido conseguiu-se uma redução da dimensão que facilitou a compreensão do problema, através da interpretação de um número de fatores comuns bem menor que o número de variáveis originais.

É, também, interessante que se analise o comportamento das observções (no caso presente as cidades) em relação a tais fatores comuns.

Isso pode ser feito calculando-se os escores fatoriais para cada um dos fatores identificados e considerados relevantes para o estudo que está sendo realizado. Os escores fatoriais nada mais são que os valores dos fatores comuns, considerados como variáveis aleatórias que não podem ser diretamente medidas, para cada uma das observações do conjunto de dados em questão. Como estas novas variáveis aleatórias têm a propriedade de independência estatística (fatores ortogonais) podem ser estudadas isoladamente o que vem a facilitar a compreensão do problema.

Para a aplicação aqui apresentada foi escolhido o resultado do método de componentes principais com rotação Varimax, com tratamento dos valores ausentes pelo algoritmo EM, para ser analisado.

Os escores calculados para os nove fatores retidos são apresentados no Anexo H.1. Tais valores foram calculados pelo método de regressão (veja *JOHNSON E WICHERN [6]*).

Ordenando os valores relativos ao fator de número 1 tem-se as cidades dispostas numa escala segundo a mortalidade das pessoas de idade mais avançada. Olhando para as extremidades dessa escala pode-se ver um que existe um grupo de cidades onde este fator assume valores muito pequenos (negativos) o que leva a crer que as variáveis relativas a mortalidade dos idosos aí têm peso pequeno, enquanto o quantitativo da população jovem é elevado. Estão nesse grupo cidades como Brasilia, Imperatriz, Foz do Iguaçu, Cascavel, Ipatinga, etc.

Na outra extremidade da escala tem-se cidades como Rio Grande, Rio de Janeiro, Pelotas Niteroi, Viamão, etc, o que indica que para estas localidades é importante a mortalidade nas idades mais avançadas e/ou a sua população pode ser considerada ponderavelmente idosa.

Fazendo o mesmo tipo de análise para o segundo fator temos as cidades de Brasilia, São José dos Campos, Novo Hamburgo, Porto Alegre, Joinville, Canoas, Ribeirão Preto, Campinas, etc, que são localidades situadas nas regiões do país sabidamente com um bom nivel de desenvolvimento, com valores baixos indicando baixa mortalidade nas idades mais jovens e um razoável desempenho no que diz respeito à saneamento e escolaridade.

Com valores altos para esse mesmo fator, do lado oposto da escala, tem-se importantes cidades da região nordeste que, sabe-se, não têm um desempenho satisfatório no que diz respeito ao desenvolvimento. As principais cidades deste grupo são Maceió, João Pessoa, Jaboatão, Aracaju, Recife, Olinda, Feira de Santana, Teresina, Fortaleza, São Luiz e Natal.

O fator 3 é aquele que agrupa as variáveis referentes à proporção da população rural de cada município, por faixa etária, em relação à população total. Para este fator têm destaque com valores positivos cidades como Niterói, Carapicuíba, Natal, São Gonçalo, São João do Meriti, etc, enquanto que com valores altamente negativos aparecem Diadema, Nova Iguaçu, Salvador, Santos, etc. Há que se ter cuidado com estas variáveis já que a população rural de um município é definida legalmente, deixando de ter importância o fato deste possuir ou não características rurais. Em outras palavras, uma localidade pode ter produção agropecuária sem que legalmente pussua área rural. Para o estado do Rio de Janeiro as cidades de Niteroi, São Gonçalo e Rio de Janeiro, por exemplo, não possuem área rural por definição.

O quarto fator contrapõem as variáveis que descrevem a estrutura da produção agrícola às variáveis relativas à mortalidade por doenças respiratórias. Dessa maneira na extremidade negativa deste eixo encontram-se cidades localizadas em áreas de alta concentração

industrial, sujeitas aos efeitos da poluição do ar e com uma estrutura de produção rural de pouca importância, ou seja: Osasco, Diadema, Carapicuíba, São João do Meriti, São Paulo, São Bernardo do Campo, etc. Por outro lado, as localidades de Pelotas, Cuiabá, Cascavel, Campo Grande, Joinville, Uberaba, Londrina, etc, se agrupam na extremidade positiva do eixo devido à sua condição de polos de produção agropecuária e de condições ambientais dentro de padrões mais desejáveis.

O quinto fator a ser considerado dá uma idéia da estrutura urbana ao levar em alta conta variáveis que medem a escolaridade da população, número de leitos hospitalares , telefones e moradias com água encanada, em contraposição ao número de moradias com poço. As cidades que assumem escores positivos neste fator são, por exemplo, Vitória, Florianópolis, Niteroi, Santos, Curitiba, Rio de Janeiro, São Paulo e Porto Alegre. Na extremidade oposta estão cidades que se destacam por não possuirem boas condições de saneamento ou pusuirem características mais rurais como: Viamão, Gravataí, Nova Iguaçu, Duque de Caxias, Sumaré, Diadema, São João do Meriti, Canoas, etc. É interessante notar que no lado positivo do fator se encontram cidades que são centro de grandes aglomerados urbanos ou regiões metropolitanas enquanto que no lado negativo estão os municípios periféricos dessas mesmas áreas, mostrando, talvez, uma concentração de recursos que faz com que a periferia não seja aquinhoada.

Para se ter uma idéia visual do comportamento dos municípios em relação aos fatores considerados, pode-se fazer um gráfico bidimensional onde cada eixo representa um fator. Tais gráficos para a presente aplicação estão no Anexo H.2.

Além do tipo de análise aqui apresentado pode-se aplicar aos fatores comuns calculados, toda uma gama de ferramentas de análise estatística. É usual se buscar identificar grupos de comportamento homogêneo em relação ao conjunto de fatores considerados, por meio da aplicação de técnicas de agrupamento, sendo que os grupos assim definidos podem sevir para definir parâmetros para modelos de classificação, como por exemplo a análise discriminante.

CAPÍTULO V

V. Alguns comentários e conclusões

Neste trabalho procurou-se fazer um apanhado geral do problema da análise de dados no que diz respeito à disponibilidade de informações estatísticas. Esse aspecto do problema preocupa grande parte das pessoas que se dedicam tanto ao trabalho de análise de problemas quantitativos nas mais diversas áreas, como aquelas que atuam na chamada área de estatísticas primárias, ou seja, que se dedicam a coletar os dados e torná-los disponíveis ao especialista ou usuário final.

Embora a discussão, aqui, tenha sido centrada no problema da falta ou omissão do dado, sabe-se que ele é muito mais amplo pois, às vezes, quando se dispõe de uma informação não se pode ter a devida confiança na mesma.

As fontes de erro ou omissões de dados podem ser as mais variadas, desde o planejamento e execução de uma pesquisa, até problemas de transcrição para arquivos magnéticos, passando pela, sempre possível, recusa do informante ou má fé do entrevistador.

Em países como o nosso ainda deve-se contar com a falta de tradição no registro de dados estatísticos, o que pode frustrar algumas iniciativas de análise, principalmente no aspecto temporal, já que raramente se pode contar com séries de dados um pouco mais longas.

Na aplicação apresentada no capítulo IV foi feita uma análise exploratória dos dados, onde pode-se notar alguns problemas não só de células vazias como de alguns valores que talvez pudessem ser considerados suspeitos e merecessem, ao menos, uma análise individual mais atenta. Decidiu-se pela exclusão de alguns pontos (Quadro 1), mas apenas naqueles casos em que os valores, inclusive, eram incompatíveis com as unidades de medida usadas.

Existem técnicas bastante poderosas para a localização desses valores discrepantes, que se baseiam na própria estrutura da matriz dos dados. Essas técnicas, no entanto, devem sempre ser aplicadas com a

consciência de que o fato de um dado ser discrepante nem sempre significa que esteja errado, e portanto deve ser conferido, sempre que possível, com a fonte de origem da informação.

Hoje as instituições que se dedicam à coleta e armazenamento de dados, geralmente se preocupam em incluir na sua rotina de apuração uma etapa de crítica das informações coletadas antes de colocá-las à disposição dos possíveis usuários. Quando a pesquisa dos dados for feita por amostragem a preocupação deve se estender, inclusive, no sentido de divulgar junto aos dados as medidas dos chamados erros amostrais.

É, de certa forma, intuitivo o fato de que as técnicas de tratamento dos dados ausentes não têm o poder de melhorar a qualidade do conjunto dos dados. Tais procedimentos são úteis quando encarados como uma forma de possibilitar o uso de informações incompletas, desde que não seja realmente possível obtê-las através de novas consultas às fontes de informação.

Quando o número de unidades pesquisadas com dados incompletos for muito pequeno, talvez seja preferível abandoná-las, em detrimento da diminuição da amostra, e trabalhar apenas com as observações completas. Infelizmente, à medida que o fenômeno a ser estudado é mais complexo, o número de variáveis envolvidas cresce tornando a possibilidade de se obter observações incompletas cada vez menor, o que torna impraticável a opção de abandoná-las.

As técnicas que permitem estimar os valores ausentes, completando a matriz de dados, parecem oferecer vantagens no sentido de permitir um leque maior de opções de uso de ferramentas de análise. Concretamente quando se usa técnicas com a finalidade de resumir as informações, ou reduzir a dimensão do problema, como a análise fatorial, bastaria existir um bom método para se estimar a matriz de covariâncias (ou correlações) a partir dos dados incompletos que o problema estaria sanado. Acontece que, geralmente, uma análise desse tipo é apenas um passo intermediário do processo, e para se ir adiante, para aplicar uma técnica de grupamento de casos, por exemplo, seria necessário que se tivesse os escores fatoriais para cada caso, onde voltaria o problema dos *"missing values"*.

O uso da análise fatorial nos problemas do tipo da aplicação aqui apresentada não se constitui em novidade. Em alguns trabalhos (*Panerai [2]* e *Bussab e Ho [2]*, por exemplo) antes de se aplicar a técnica, as variáveis são divididas préviamente em variáveis preditoras e variáveis resposta, sendo fatoradas separadamente. Aqui optou-se por fazer a análise fatorial a partir do conjunto de todas as variáveis consideradas, buscando nos dados os possíveis agrupamentos das variáveis.

A qualidade do resultado de uma análise fatorial, como foi discutido no capítulo III, deve ser julgado, por um lado, por medidas objetivas como a quantidade da variabilidade explicada, pela reconstituição da matriz de correlações a partir dos fatores comuns extraídos, etc..., e, por outro lado, pela possibilidade de se associar uma interpretação razoável aos fatores gerados.

Pelos resultados apresentados pode-se notar que não houveram diferenças sensíveis a partir da aplicação das várias técnicas de solução da análise fatorial. Apenas para orientar a discussão escolheu-se, aqui, o resultado da solução pelo método das componentes principais, com a rotação Varimax.

A solução por esse método propicia um agrupamento de variáveis onde se pode associar interpretações razoáveis aos fatores comuns (ao menos para os cinco primeiros) e pode-se considerar que a matriz das cargas fatoriais obedecem os critérios de julgamento sugeridos por *Thurstone* (ver *Kubrusly [24]*).

Os fatores podem ser classificados como:

Fator 1: mortalidade nas idades mais avançadas
Fator 2: mortalidade nas idades jovens (desenvolvimento)
Fator 3: população rural
Fator 4: estrutura agrária
Fator 5: urbanização

Como o problema envolve um número muito grande de variáveis, 49, era de se esperar que os fatores acabassem ficando, também, com muitas variáveis importantes para cada um deles e, ainda, com uma certa complexidade. Assim, é possivel olhar "por dentro" desses fatores para procurar entendê-los melhor.

Para o caso do fator associado à mortalidade das pessoas mais idosas pode-se ver que existem duas componentes distintas atuando, ou seja, um eixo correspondente à mortalidade propriamente dita e outro que corresponde às variáveis estruturais, número de moradias e moradias com esgoto. Nesses dois eixos as variáveis correspondentes a população mais jovem entram com sinal negativo.

O segundo fator pode ser também decomposto em um eixo relacionado com a mortalidade, propriamente dita, das pessoas menos idosas e um segundo eixo onde são contrapostas as variáveis correspondentes a uma situação de melhor desenvolvimento (alfabetização e saneamento) e variáveis correspondentes à mortalidade jovem.

O fator 4 contrapõe fortemente as variáveis correspondentes à mortalidade por doenças respiratórias àquelas que definem municípios com uma estrutura rural mais claramente definida.

Olhando o fator 5 com mais atenção nota-se que uma de suas dimensões é formada pelas variáveis que definem uma população com escolaridade mais avançada enquanto que outra é definida por variáveis correspondentes a uma estrutura urbana mais evoluída.

As observações acima podem ser verificadas aplicando-se novamente a análise fatorial em cada conjunto de variáveis correspondentes aos fatores definidos anteriormente. O Quadro 1 mostra um resumo dos resultados obtidos por essa estratégia.

Quadro 1

Resultados da aplicação da análise fatorial em cada um dos fatores definidos no item IV.2.1

Fatores	Variáveis	Cargas (%) dos Subfatores
1.1	Mortalidade acima de 65 anos	88
	Mort. doenças cardio-vasculares	82
	Mort. doenças neoplásicas	79
	Mortalidade de 50-65 anos	74
	População de 65 anos ou mais	73
	População de 50-65 anos	72
	População de menores de 1 ano	-62
	População de 1-4 anos	-61
1.2	Moradias com esgoto	89
	Total de moradias	78
	População de menores de 1 ano	-68
	População de 1-4 anos	-71
2.1	Mortalidade de menores de 1 ano	88
	Mortalidade geral	83
	Mort. doenças diarreicas 0-1 ano	78
	Mort. doenças diarreicas	77
	Mort. doenças respir. 1-4 anos	70
	Mortalidade de 1-4 anos	69
	Mort. doenças infec. parasitárias	60
2.2	Moradias com egoto	91
	Alfabet. com mais de 5 anos	88
	Mort. doenças diarreicas	-53
	Mortalidade de 1-4 anos	-61
	População de 5-19 anos	-94
4.2	Total de terra arável	86
	Total de terra cultivada	84
	Área do município	83
	Número de estabelecimentos rurais	72
	Culturas resp. 80% da produção	58
4.1	Mort. doenças respiratórias	90
	Mort. doenças respir. 0-1 ano	84
	Mort. doenças respir. 1-4 anos	81
5.1	Pessoas 11 amos ou mais de estudo	89
	Pessoas 8 anos ou mais de estudo	87
	Numero de leitos	85
	População de 65 anos ou mais	60
5.2	Moradias com água encanada	88
	Número de telefones	74
	Moradias com poço	-86

Com a análise fatorial é possível verificar como um conjunto grande de variáveis se interrelacionam formando agrupamentos de variáveis os quais podem ser entendidos como as diversas dimensões importantes de um determinado fenômeno de interesse. A

interpretabilidade dos fatores é de suma importância e serve, inclusive, para balizar a qualidade dos resultados da análise fatorial, que, nesse sentido, pode ser vista como uma técnica exploratória já que um resultado considerado extranho pelo especialista pode levar, através de uma análise mais detalhada, até a descoberta de algum problema com os dados.

A análise do comportamento das cidades em relação aos cinco primeiros fatores comuns, mostrou resultados bastante interessantes no sentido de agrupar os municípios de acordo com as características representadas pelas variáveis originais através dos fatores comuns. A propriedade de independência estatística dos fatore ortogonais facilita a análise já que se pode estudá-los separadamente.

Aparentemente o fator de número 3 é o que possui menos clareza de análise. Isso ocorre pelo fato do que vem a ser população rural, já que para sua definição não são levadas em conta as condições objetivas com relação ao que se refere às características rurais mas sim a legislação.[1] É, talvez, devido a isto que este fator é altamente correlacionado apenas com as variáveis relativas à distribuição etária da população rural, ou seja: se o município possui população rural os quantitativos populacionais por faixa etária são correlacionados entre si, porém, não são necessáriamente correlacionados com outras variáveis. Este aspecto mostra a relevância de uma criteriosa seleção das variáveis a serem utilizadas em um determinado estudo.

Este trabalho mostra a importância das técnicas de tratamento dos dados ausentes no sentido de se poder, mesmo a partir de uma matriz de dados incompleta, realizar análises que podem ser de suma utilidade para o estudo de fenômenos como o da saúde, por exemplo. Apesar de se ter consciência que se deve investir ao máximo no sentido de se obter

[1] A divisão da área de um município em sua parte rural e urbana é definida pelo poder legislativo municipal e tem como fator determinante a questão da arrecadação de impostos. Isso faz com que áreas com características fundamentalmente rurais sejam classificadas como urbanas e vice-versa. Por outro lado uma atividade agrícola (uma plantação por exemplo), mesmo realizada em estabelecimento localizado em área urbana, é sempre classificada como tal o que pode fazer com que municípios sem área rural (legalmente) tenham estabelecimentos agropecuários produtivos.

dados completos, sabe-se que isto nem sempre é possível de se concretizar, já que as falhas ocorrem independentemente da vontade do pesquisador.

Os resultados obtidos através da análise fatorial, pela sua razoável coerência, mostram que o tratamento aqui aplicado para estimar os dados ausentes pode ser considerado bastante satisfatório.

Pode-se apontar algumas linhas para a continuidade deste trabalho. Uma delas é na perspectiva da análise dos dados aqui trabalhados, no sentido de se explorar mais os resultados obtidos procurando através dos fatores extraídos, por exemplo, buscar grupos de municípios de comportamento semelhante para tentar entender um possível processo de regionalização do problema. Outra linha seria testar a qualidade dos resultados oferecidos pelas técnicas de preenchimento das lacunas da matriz de dados, por meio de um estudo de simulação.

CAPÍTULO VI

VI. REFERÊNCIAS BIBLIOGRÁFICAS

[1] BUSSAB, W. O. E HO, L. L., *Características regionais de saúde no estado de São Paulo - Análise estatística dos dados*, FUNDAP, São Paulo, 1983.

[2] PANERAI, R. B., *Multisectorial determinants of health in Brazil-Progress Report*, University of Virginia, Charlottesville, Va., 1985.

[3] LITTLE, R. J. A. e RUBIN, D. B., *Statistical analysis with missing data*, New York, John Wiley & Sons., 1987.

[4] WELLS, W. D. E SHETH, J. N., Factor analysis in marketing research, in *Handbook of market research*, Mc Graw-Hill, Inc., 1971.

[5] HARMAN, H. H., *Modern factorial analysis*, University of Chicago Press, 1975.

[6] JOHNSON, R. A. e WICHERN, D. W., *Applied multivariate statistical analysis*, Englewood Cliffs, NJ, Prentice Hall, 1982.

[7] SILVA, P. L. N., *Crítica e imputação de dados quantitativos utilizando o SAS*, série Informes de Matemática, IMPA, Rio de Janeiro, 1989.

[8] COCHRAN, W. G., *Sampling Techniques*, John Wiley & Sons. Inc., New York, 1977.

[9] HANSEN, M. H., HURWITZ, W. N. E MADOW, W.G., *Sample survey methods and theory*, John Wiley and Sons, Inc., New York, 1953.

[10] KIM, J. O. E CURRY, J., The treatment of missing values in multivariate analysis, *Social Methods Research*, **6**, pp. 215-240, 1977.

[11] AZEN, S. E VAN GUILDER, M., Conclusions regarding algorithms for handling incomplet data, *Proceedings of the Statistcal Computing Section, American Statistical Association*, pp 53-56, 1981.

[12] LITTLE, R. J. A. E SMITH, P. J., Edditing and imputation for quantitative survey data, *Journal of the American Statistical Association*, **82**, pp. 58-68, 1987.

[13] MCKENDRICK, A. G., Applications of mathematics to medical problems, *Proc. Edinburgh Math. Soc.*, **44**, pp 98-130, 1926.

[14] DEMPSTER, A. P., LAIRD, M. e RUBIN, D. B., Maximum likelihood from incomplete data via the EM algorithm, *Journal of The Royal Statistical Society*, **B**, **39**, pp. 1-38, 1977.

[15] SEARLE, R. R., *Linear models*, John Wiley & Sons, Inc., New York, 532 p., 1981

[16] BEATON, A. E., The use of special matrix operations in statistical calculus, *Research Bulletin*, **RB** 64-51., Princeton, 1964

[17] MOOD, A. M., GRAYBILL, F. A. E BOES, D. C., *Introduction to the theory of statistics*, Mc Graw-Hill, 1974.

[18] BEALE E. M. L. E LITTLE, R. J. A., Missing values in multivariate analysis, *Journal of the Royal Statististical Society*, **41**, pp 129-145, 1975.

[19] ANDERSON, T. W., *Introduction to statistical multivariate analysis*, John Wiley & Sons, Inc., New York, 1958.

[20] SOUZA, J., *Análise fatorial*, Editora Thesaurus, 1988.

[21] MARDIA, K. V., KENT, J. T. E BIBBY, J. M., *Multivariate analysis*, Academic Press, Inc., London, 1979.

[22] BARTLETT, M. S., A note on multiplyng factors for various Chi-squared approximations, *Journal of the Royal Statistical Society*, **16**, pp. 269-298, 1954.

[23] KAISER, H. F., The varimax criterion for analytic rotation in factor analysis, *Psychometrika*, **23**, pp. 187-200, 1958.

[24] KUBRUSLY, L. S., *O modelo de análise fatorial*, Tese de Mestrado, COPPE/UFRJ, Rio de Janeiro, 1981.

[25] ALMEIDA, R. M. V. R., *Estudo da correlação entre variáveis sócio-econômicas e indicadores de estado de saúde*, Tese de Mestrado, COPPE/UFRJ, Rio de Janeiro, 1987.

[26] SAS Institute Inc., *SAS User's Guide: Basics*, Version 5 Edition. Cary, N.C., 1290 p.0 p., 1985.

[27] DACHS, J. N. W., *Análise de dados e regressão*. São Paulo, UNICAMP, 1978.

[28] BOX, G. E. P. e COX, D. R., An analysis of transformations, *Journal of the Royal Statistical Society*, **B26**, pp. 211-252, 1964.

[29] SAS Institute Inc., *SAS User's Guide: Statistics*, 1982 Edition. Cary, N.C., 1290 p.0 p., 1982.

ANEXO A

Descrição e fontes de informação das variáveis

RELAÇÃO DAS VARIÁVEIS E FONTES DE COLETA

NASCIDOS VIVOS

Número total de registros no ano, inclui os nascimentos ocorridos no ano e os ocorridos em anos anteriores e registrados neste ano.

Fontes: -De 1981 a 1983 - SRIV (Servico de Recuperação de Informações por Variável) -IBGE

-De 1974 a 1980 - Estatística do registro civil -IBGE para todos municípios.

-1969 - Censo Demográfico de 1970 -IBGE -Outros anos p/capitais- Anuários estatísticos -IBGE

-Outros anos p/ municípios de MG e RGS, fornecidos pelas respectivas Secretarias de Saude.

-Outros anos, para municípios de SP, Servico Estadual de Analise Estatística - SEADE

NUMERO DE AUXILIO DOENÇA CONCEDIDOS

Número total de auxilio doença, ex-combatente, e plano basico,concedidos no ano pela Previdencia Social.

Fontes: -De 1980 a 1982, todos municípios, fornecido pelo DATAPREV.

APOSENTADOS POR INVALIDEZ

Total de auxilios concedidos a trabalhadores rurais, aposentadoria por invalidez, lei 1756/52, plano basico, e ex-combatentes, concedidos pela Previdencia Social.

Fontes: -De 1980 a 1982, todos municípios, fornecidos pelo DATAPREV.

MORTALIDADE GERAL

Soma de mortes por todas as causas em todas as idades , não inclui óbitos fetais.

Fontes: -De 1981 a 1983, todos municípios, SRIV -IBGE

-De 1977 a 1984, todos municípios, fornecidos pelo Ministério de Saude - FSESP.

-Anteriores a 1977, capitais, Anuários estatísticos -IBGE

-De 1950 a 1980, para municípios do RJ, Secretaria Estadual de Saude.

-De 1970 a 1982, para municípios de MG, Secretaria Estadual de Saude.

-De 1968 a 1982, para municípios do RGS, Sec. Estadual de Saude.

-De 1950 a 1982, para município de João Pessoa(30), Sec. Estadual de Saude.

-De 1960 a 1976, para municípios de SP, SEADE

MORTALIDADE POR FAIXA ETARIA

Fontes: -De 1977 a 1984, todos municípios, FSESP.

-Municipios do RGS,MG,RJ,PB, Secretarias Estaduais de Saude.

-De 1960 a 1976, municípios de SP, SEADE.

Faixas - De 1 a 4 anos

- De 5 a 19 anos

  OBS: mun. 30 de 69-82 5 a 24 anos

- De 20 a 49 anos

  OBS: mun 30 de 69-82 25 a 44 anos

- De 50 a 64 anos

  OBS: mun 30 de 69-82 45 a 64 anos
  mun 30 de 51-68 50 a 59 anos
  mun 03, 08, 39, 41, 45, 49, 56, 87, 89, 94, 96, 97, todos anos 50 a 69 anos

- Maiores de 65 anos

  OBS: mun 30 de 51-68 > de 60 anos
  mun 03, 08, 39, 41, 45, 49, 56, 87, 89, 94, 96, 97 todos anos > de 70 anos

MORTALIDADE DE MENORES DE 1 ANO DE IDADE

Total de mortes por todas as causas entre 0 e 1 anos de idade, não inclui natimortos.

Fontes: -De 1977 a 1984, todos municípios, FSESP.

-De 1970 a 1980, municípios de MG, Sec. Est. de Saude

-De 1968 a 1982,municípios do RGS, Sec. Est. de Saude

-De 1950 a 1982,município de João Pessoa(30), Sec. Est. de Saude.

-De 1981 a 1983, outros municípios, SRIV -IBGE

-De 1960 a 1976, municípios de SP, SEADE

-Anteriores a 1977,demais capitais, Anuários Estatísticos do IBGE.

NUMERO DE NATIMORTOS

Número total de óbitos fetais,e natimortos,com qualquer periodo de gestação. Por local de residência da mãe. São computados apenas os óbitos legalmente registrados.

Fontes: -De 1974 a 1980 ,todos municípios, Estatísticas de Registro Civil -IBGE

-Anteriores a 1974,capitais, Anuários Estatísticos IBGE

-De 1981 a 1983, todos municípios, SRIV -IBGE

-Municipios do RGS,MG,RJ,PB, Secretarias Estaduais de Saude.

-De 1960 a 1976, municípios de SP , SEADE.

MORTALIDADE POR DOENÇAS INFECCIOSAS E PARASITARIAS

Obtido pela soma de diversas causas de morte, conforme relacionado a seguir:

Soma de:	CID(9a)	CID-BR	LISTA A	LISTA B
Tuberculose resp.	010-012	02	A6	B5
Tb. outras formas	013-018	02	A7-A10	B6
Peste	020	03	A11	B7
Difteria	032	03	A15	B8
Coqueluche	033	03	A16	B9
Ang.estr. e escarl.	034	03	A17	B10
Inf. Meningococicas	036	03	A19	B11
Tetano	037	03	A20	
Poliomielite	045-049	04	A22-A23	
Febre amarela	060	04	A26	
Encefalites virais	062-065	04	A27	
Raiva	071	04		
Tifo e out.riqt.	080-083	05	A30	B15
Malaria	084	05	A31	B16
Sifilis	090-097	06	A34-A37	B17
Esquistossomose	120	07		
Outras doenças inf. e par. não relac.	Resto de 010-139	07	Resto de A6-A44	Resto de B5-B18

Fontes: -De 1977 a 1984, todos municípios, FSESP.
Municipios de RGS,RJ,MG,PB, Secretarias Estaduais de Saude.

-De 1960 a 1976, municípios de SP, SEADE.

-Capitais, outros anos, Anuários Estatísticos IBGE.

MORTALIDADE POR DOENÇAS CARDIOVASCULARES

Obtido pela soma de diversas causas de morte, conforme abaixo relacionado:

Soma de:	CID(9a)	CID-BR	LISTA A	LISTA B
Doença reum. ativa	390-392	25	A80	B25
Febre reum. ativa	393-398	25	A81	B26
Doença hipertensiva	400-404	26	A82	B27
Doença isquemica	410-414	27	A83	B28
Out. doenças do cor.	420-429	28	A84	B29
Doenças cerebrovasc.	430-438	29	A85	B30
Out doen.ap. circul.	440-459	30	A86-A88	B30

Fontes: -Mesmas da variável mortal. doenças infec. parasit.

MORTALIDADE POR DOENÇAS RESPIRATORIAS

Obtido pela soma de diversas causas de morte, conforme abaixo relacionado:

Soma de:	CID(9a)	CID-BR	LISTA A	LISTA B
Pneumonia e gripe	460-487	32	A89-A92	B31-B32
Bronquite,enfis.,asma	490-493	32	A93	B33
Pneumoconiose e outras causas ext.	500-508	32	A96	B33

Fontes: -Mesmas da variável mortal. doenças infec. parasit.

MORTALIDADE POR DOENÇA RESPIRATORIA, 0-1 ANO.

Mesmas causas da variável 140.

Fontes: -Mesmas da variável mortal. doenças infec. parasit.

MORTALIDADE POR DOENÇA RESPIRATORIA, 1-4 ANOS.

Mesmas causas da variável 140.

Fontes: -Mesmas da variável mortal. doenças infec. parasit.

MORTALIDADE POR DOENÇAS NEOPLASICAS.

Obtido pela soma de diversas causas de morte, conforme abaixo relacionado:

Soma de:	CID(9a)	CID-BR	LISTA A	LISTA B
Neoplasias malignas	140-209	08-14	A45-A60	B19
Neoplasias benignas	210-239	15-17	A61	B20

Fontes: -Mesmas da variável mortal. doenças infec. parasit.

MORTALIDADE POR ACIDENTES DE TRANSITO.

Obtido pela soma de diversas causas de morte, conforme abaixo relacionado:

Soma de:	CID(9a)	CID-BR	LISTA A	LISTA B
Ac. c/veic. aut.	E810-E823	E47	AE138	BE47
Out. ac. de trans.	E800-E848			

Fontes: -Mesmas da variável mortal. doenças infec. parasit.

MORTALIDADE POR EVENTOS VIOLENTOS

Obtido pela soma de diversas causas de morte, conforme relacionado a seguir:

Soma de:	CID(9a)	CID-BR	LISTA A	LISTA B
Quedas,fogo,etc.	E880-E899	E50-E51	AE141-AE142	
Suic. e auto inf.	E950-E959	E54	AE147	BE49
Hom. e out. viol.	E960-E999	E55-E56	AE148-AE150	BE50

Fontes: -Mesmas da variável mortal. doenças infec. parasit.

MORTALIDADE POR DOENÇAS DIARREICAS

Obtido pela soma de diversas causas de morte, conforme a seguir relacionado:

Soma de:	CID(9a)	CID-BR	LISTA A	LISTA B
Febre tifoide	001	01	A2	B2
Out. doenc. diarr.	008-009	01	A3-A4	B3
Enterites e out.	002-007	01	A5	B4

Fontes: -Mesmas da variável mortal. doenças infec. parasit.

MORTALIDADE POR DOENÇAS DIARREICAS MENORES DE 1 ANO

Mesmo conjunto de causas da variável 180.

Fontes: -Mesmas da variável mortal. doenças infec. parasit.

NUMERO DE LEITOS

Número total de leitos disponíveis no município, incluindo públicos e particulares, todas as especialidades

Fontes: -De 1981 a 1982,todos municpios, SRIV -IBGE

-De 1976 a 1979, Estatísticas de Saude -IBGE

-Municipios do RJ, 1974, Estatísticas de Saude do Estado da Guanabara

-Municipios de SP de 1977 a 1981 , -SEADE

-Municipios do RGS, Anuário Estatístico do RGS -FEE

-Municipio de João Pessoa(30), Sec. Estadual de Saude

-Municipios da BA,1968 a 1970, Anuário Estatístico da Bahia

-Municipio de Aracaju, Indicadores Sociais de Sergipe

NUMERO DE BAIXAS HOSPITALARES POR ANO

Número total de pacientes internados em hospitais no município, em hospitais públicos e particulares,todas as especialidades.

Fontes: -Mesmas da variável 200

NUMERO DE CONSULTAS AMBULATORIAIS POR ANO

Número total de pacientes atendidos em ambulatorios públicos e particulares.

Fontes: -Mesmas da variável 200

PRECIPITAÇÃO PLUVIOMETRICA ANUAL

Total de precipitação luviometrica no município no periodo de um ano.

TEMPERATURA MINIMA MEDIA MENSAL NO ANO

Menor temperatura minima (media mensal) registrada no ano,no município.

POPULAÇÃO TOTAL DO MUNICIPIO POR FAIXA ETARIA

Fontes: -Todos municípios, Censo Demográfico IBGE.

Faixas - Menores de 1 ano
- De 1 a 4 anos
- De 5 a 19 anos
- De 20 a 49 anos
- De 50 a 65 anos

OBS: Para o ano de 70-80,todos mun.: De 50 a 69 anos.

- Maiores de 65 anos

OBS: Para o ano de 70-80,todos mun.: Maiores de 70 anos.

POPULAÇÃO RURAL DO MUNICIPIO POR FAIXA ETARIA

Fontes: -Todos municípios, Censo Demográfico IBGE

Faixas - Menores de 1 ano
- De 1 a 4 anos
- De 5 a 19 anos
- De 20 a 49 anos
- De 50 a 65 anos

OBS: Para o ano de 70-80,todos mun.: De 50 a 69 anos.

- Maiores de 65 anos

OBS: Para o ano de 70-80,todos mun.: Maiores de 70 anos.

QUANTIDADE DE AGUA PRODUZIDA

Quantidade de agua potavel (tratada) produzida no município.

Fontes: -De 1974 a 1978, municípios de SP, Perfil Municipal, SEADE

-Municipios de MG,todos os anos,Sec. de Planejamento de Minas Gerais.

-Municipios do RGS,todos os anos, CORSAN

-Municipio de Porto Alegre, DMAE

EXTENSAO DE REDE DE ESGOTOS

Extensão total da rede de esgoto existente no município, excluindo-se a extensão de emissarios.

Fontes: -De 1974 a 1978, municípios de SP, Perfil Municipal, SEADE

-Municipios de MG,todos os anos,Sec. de Planejamento de Minas Gerais.

-Municipio de Porto Alegre(08), DMAE

NUMERO DE MORADIAS C/ALGUM SISTEMA DE ESGOTO

Número total de moradias que possuem algum sistema de esgoto, seja rede geral ou fossa septica.

Fontes: -Todos os municípios,(1950,60,70,80) Censos Demográficos do IBGE.

NUMERO TOTAL DE MORADIAS

Soma de moradias de todos os tipos, duraveis, rusticas, com e sem agua encanada.

Fontes: -50,60,70,80 todos municípios, Censo Demográfico -IBGE

NUMERO DE MORADIAS COM AGUA ENCANADA

Número total de moradias que dispoe de agua encanada, de rede geral ou não.

Fontes: -1960,70,80, todos municípios, Censo Demográfico -IBGE

-Mun. de POrto Alegre,todos anos, DMAE

-Mun. do RGS,todos anos, CORSAN

-De 1974 a 1978,Mun. de SP,Perfil Municipal, SEADE

NUMERO DE MORADIAS C/ POÇO OU NASCENTE

Número total de moradias abastecidas com poco ou nascente de agua.

Fontes: -Todos municípios,(60,70,80) Censo Demográfico do IBGE.

NUMERO DE VEICULOS AUTOMOTORES

Número total de veiculos automotores registrados no município, incluindo ônibus, caminhoes, utilitarios e motos.

Fontes: -De 1967 a 1972, todos municípios, Cadastro de Veiculos Automotores -IBGE

-De 1973 a 1980, mun. do RGS, - FEE (Fundação Estadual de Estatística).

-De 1976 a 1981, mun. de SP, Perfil Municipal -SEADE

NUMERO DE ONIBUS

Número de veiculos de transporte coletivo registrados no município incluindo lotações,e micro-ônibus. Observar que o veiculo registrado não presta obrigatoriamente serviço no mesmo município.

Fontes: -Mesmas da variável 570

NUMERO DE FERIDOS EM ACIDENTES DE TRANSITO

Número total de ferimentos devidos a acidentes de transporte tratados durante o ano,com e sem veiculos automotores.

Fontes: -De 1960 a 1979, municípios das capitais, Ministério dos Transportes,-DENATRAN

-Mun. de Porto Alegre(08), Pronto socorro municipal

-Mun. de Belo Horizonte(03), Pronto socorro

NUMERO DE APARELHOS TELEFONICOS

Número total de aparelhos telefônicos em funcionamento no município, independente do tipo e de ser ou não extensão.

Fontes: -Municipios das capitais, Anuários Estatísticos -IBGE

-João Pessoa(30) ,Anuário Estatístico da Paraiba

-De 1978 a 1981,Mun. de SP, Perfil Municipal SEADE

-Mun de ES, Cia Telefônica

-De 1967 as 1981, mun do RGS, - FEE.

-1955 , todos municpios, Enciclopedia dos Municipios IBGE

AREA DO MUNICIPIO

Area total do município, em km quadrado.

Fontes: -1950,1960,1970,1980, todos municípios, SRIV-IBGE

CONSUMO TOTAL DE ENERGIA ELETRICA

Total de energia elétrica consumida no município durante o ano, inclui consumo residencial, público, industrial e comercial.

Fontes: -Capitais, todos anos, Anuários Estatísticos do IBGE.
-Mun. do RGS, Anuários estatísticos - FEE.
-Mun da BA, Anuários estatísticos.

CONSUMO DE ENERGIA ELETRICA INDUSTRIAL

Energia elétrica consumida para uso industrial por ano.

Fontes: -Mesmas da variável 620.

PESSOAS C/MAIS DE 5 ANOS ALFABETIZADAS

Número total de pessoas com mais de cinco anos que sabem ler e escrever no município.

Fontes: -1950,1960,1970,1980,- Censo Demográfico

PESSOAS COM MAIS DE 8 ANOS DE ESTUDO

Número total de pessoas com 8 ou mais anos de estudo.

Fontes: -Mesmas da variável 700.

PESSOAS COM MAIS DE 11 ANOS DE ESTUDO

Número total de pessoas que estudaram 11 ou mais anos.

Fontes: -Mesmas da variável 700.

TOTAL DE TERRA ARAVEL

Total de terra utilizavel para produção agrícola. Esta variável e a soma de áreas de lavouras permanentes, temporarias, das terras em descanso e áreas produtivas não utilizadas. A área ocupada por matas e pastagens não esta incluida neste total.

Fontes: -1975 e 1980 Censo Agropecuário -IBGE

TOTAL DE AREA CULTIVADA

Total de área utilizada para obter a produção agrícola (variável 740) anual. Observar que uma mesma área pode ser utilizada para mais de uma cultura durante um ano.

Fontes: -De 1973 a 1980 ,todos municípios, Produção Agricola Municipal, Censo Agropecuário IBGE., SRIV.

-De 1968-1973, Mun da BA, Anuário Estatístico Da Bahia

-De 1974-1980, mun do RGS, Anuário Estatístico do RGS,FEE

VALOR DA PRODUÇÃO AGRICOLA

Valor total das principais culturas agrícolas do município durante o ano ,inclui as principais culturas permanentes e temporarias,bem como alguns produtos de origem animal e extrativo, conforme relação abaixo.

Abacaxi, Caqui, Manga, Abacate, Cebola, Marmelo, Algodão, Cera, Mel, Alho, Côco, Melancia, Amendoim, Feijão, Melão, Aveia, Fibra de sisal, Ovos, Azeitona, Fumo, Pera, Banana, Lã, Pessego, Batata-doce, Laranja, Soja, Batata-inglesa, Leite, Sorgo, Cafe, Limão, Tangerina, Caju, Maca, Tomate, Cana de açucar, Mandioca, Trigo e Uva

OBS:Valor da produção convertido para dolar em 1970.

Fontes: -De 1973 a 1980 ,todos municípios, Produção Agricola Municipal, Censo Agropecuário IBGE, e SRIV.

-De 1968-1973, Mun da BA, Anuário Estatístico Da Bahia

-De 1974-1980, mun do RGS, Anuário Estatístico do RGS,FEE

VALOR DO REBANHO

Valor total do rebanho do município,conforme relação abaixo.

Asininos Bovinos Muares
Caprinos Equinos Patos
Codornas Galinhas Perus
Coelhos Ovinos Suinos

OBS:Valor do rebanho convertido para dolar em 1970.

Fontes: -De 1973 a 1980 ,todos municípios, Produção Pecuaria Municipal, Censo Agropecuário IBGE, e SRIV.

-De 1968-1973, Mun da BA, Anuário Estatístico Da Bahia

-De 1974-1980, mun do RGS, Anuário Estatístico do RGS,FEE

NUMERO DE CULTURAS PARA 80% DA PRODUÇÃO

Número de culturas necessarias para se obter 80% do valor da produção agrícola.

Fontes: -Obtida a partir dos dados discriminados da variável valor da produção agrícola.

NUMERO DE ESTABELECIMENTOS AGRICOLAS

Número total de estabelecimentos dedicados a atividades agro-pecuarias, independente do tipo de proriedade.

Fontes: -1975 e 1980 Censo Agropecuário -IBGE

NUMERO DE EMPREGOS REGULARES

Número total de empregos regulares existentes no município, incluindo industria, comercio e profissoes liberais.

Fontes: -Todos municípios, 60,70, e 80, Censo Demográfico do IBGE.

POPULAÇÃO RURAL TOTAL

População rural do município, deve corresponder à soma das variáveis referentes a população rural por faixa etária.

Fontes: -Todos municípios, 60,70, 80, Censo Demográfico IBGE.

ANEXO B

Lista das variáveis e unidades de medida

*	V1	Nascidos vivos	Nasc./1000 Hab
*	V2	Trabalhadores em licenca	Lic/1000 >20a
*	V3	Aposentadorias prematuras	Apos/1000 >20a
*	V4	Mortalidade todas idades	Ob/1000 Hab
*	V5	Mortalidade 1 a 4 anos	Ob/10000 1-4a
*	V6	Mortalidade 5 a 19 anos	Ob/10000 5-19a
*	V7	Mortalidade 20 a 49 anos	Ob/10000 20-49
*	V8	Mortalidade 50 a 65 anos	Ob/10000 50-65
*	V9	Mortalidade > de 65 anos	Ob/10000 >65a
*	V10	Mortalidade < de 1 ano	Ob/10000 <1a
*	V11	Obitos fetais	Ob/10000 <1a
*	V12	Mort. d. infec. e paras.	Ob/10000 Hab
*	V13	Mort. d. cardio-vascular	Ob/10000 >20a
*	V14	Mort. d. respiratorias	Ob/10000 hab
*	V15	Mort. d. resp. 0-1 ano	Ob/10000 <1a
*	V16	Mort. d. resp. 1-4 anos	Ob/10000 1-4a
*	V17	Mort. d. neoplasicas	Ob/10000 >20a
*	V18	Mort. ac. trafego	Ob/10000 5-65a
*	V19	Mort. causas violentas	Ob/10000 Hab
*	V20	Mort. d. diarreicas	Ob/10000 0-4a
*	V21	Mort. d. diar. 0-1 ano	Ob/10000 <1a
*	V22	Num. de leitos	L./10000 Hab
	V23	Num. de internações	Int./10000 Hab
	V24	Num. de consultas	Cons./1000 Hab
	V25	Prec. pluvimétrica	mm/ano
	V26	Min. temp. media mensal	Grau Centigrado
*	V27	Pop. mun. menores 1 ano	% Pop. total
*	V28	Pop. mun. 1 a 4 anos	% Pop. total
*	V29	Pop. mun. 5 a 19 anos	% Pop. total
*	V30	Pop. mun. 20 a 49 anos	% Pop. total
*	V31	Pop. mun. 50 a 65 anos	% Pop. total
*	V32	Pop. mun. > de 65 anos	% Pop. total
*	V33	Pop. rur. < de 1 ano	/1000 pop. tot.
*	V34	Pop. rur. 1 a 4 anos	/1000 pop. tot.
*	V35	Pop. rur. 5 a 19 anos	/1000 pop. tot.
*	V36	Pop. rur. 20 a 49 anos	/1000 pop. tot
*	V37	Pop. rur. 50 a 65 anos	/1000 pop. tot.
*	V38	Pop. rur. acima 65 anos	/1000 pop. tot.
	V39	Volume de agua produzida	m3/10 Hab
	V40	Extens o rede de esgotos	m/1000 Hab
*	V41	Moradias c/ algum esgoto	Mor./1000 Hab
*	V42	Total de moradias	Mor./1000 Hab
*	V43	Moradias c/agua encanada	Mor./1000 Hab
*	V44	Moradias c/ poco	Mor./1000 Hab
	V45	Total de veiculos	Vei./1000 >19a
	V46	Total de coletivos	Vei./1000 Hab
	V47	Feridos por ac. transito	Ac./1000 vei
*	V48	Num. de telefones	Tel./1000 Hab
*	V49	Area do municipio	km2
	V50	Cons. energia eletrica	Mwh/1000 Hab
	V51	Cons. en. el. industrial	Mwh/10000 Hab
*	V52	Alfabetizados > 5 anos	Hab./1000 >5a
*	V53	Pes. c/+8 anos estudo	Hab./1000 >20a
*	V54	Pes. c/+11 anos estudo	hab./1000 >20a
*	V55	Total de terra aravel	Ha/1000 Hab.

(Continua)

(Continuação)

*	V56 Total terra cultivada	Ha/100000 Hab.
	V57 Val. da produção agr.	US$/Hab.
	V58 Val. do rebanho	US$/Hab.
*	V59 Cult.resp. 80% da prod. agríc.	Culturas
*	V60 Num. estab. rurais	E./1000 Hab. rur.
*	V61 Num. de empregos reg.	Emp./1000 19-64a
*	V62 População total do mun.	Hab.
	V63 População rural total	% pop. total

Obs.: As variáveis assinaladas com um asterisco são às que se referem a aplicação apresentada no Capítulo IV.

Anexo C

Lista dos Municípios

Lista dos Municípios

Cidade	Código- Nome	Estado
1	1- São Paulo	SP
2	2- Rio de Janeiro	RJ
3	3- Belo Horizonte	MG
4	4- Salvador	BA
5	5- Fortaleza	CE
6	6- Recife	PE
7	7- Brasilia	DF
8	8- Porto Alegre	RS
9	9- Nova Iguaçu	RJ
10	10- Curitiba	PR
11	11- Belem	PA
12	12- Goiânia	GO
13	13- Campinas	SP
14	14- Manaus	AM
15	15- São Gonçalo	RJ
16	16- Duque de Caxias	RJ
17	17- Santo André	SP
18	18- Guarulhos	SP
19	19- Osasco	SP
20	20- São Luiz	MA
21	21- São Bernardo do Campo	SP
22	22- Natal	RN
23	23- Santos	SP
24	24- Niteroi	RJ
25	25- Maceió	AL
26	26- São João de Meriti	RJ
27	27- Teresina	PI
28	29- Jaboatão	PE
29	30- João Pessoa	PA
30	31- Ribeirão Preto	SP
31	33- Londrina	PR
32	34- Aracaju	SE
33	35- Campo Grande	MS
34	36- Feira de Santana	BA
35	37- São José dos Campos	SP
36	38- Olinda	PE
37	39- Contagem	MG
38	41- Pelotas	RS
39	45- Uberlândia	MG
40	46- Joinville	SC
41	47- Diadema	SP

(continua)

Lista dos Municípios

(continuação)

Cidade	Código- Nome	Estado
42	49- Canoas	RS
43	50- Imperatriz	MA
44	52- Cuiaba	MT
45	53- Vitória	ES
46	54- Mauá	SP
47	55- Vila Velha	ES
48	56- Uberaba	MG
49	63- Florianópolis	SC
50	66- Carapicuíba	SP
51	76- Paulista	PE
52	77- Cascavel	PA
53	87- Ipatinga	MG
54	89- Rio Grande	RS
55	94- Novo Hamburgo	RS
56	95- Foz do Iguaçu	PR
57	96- Viamão	RS
58	97- Gravataí	RS
59	98- Sumaré	SP

ANEXO D

Prova dos resultados do item III.2.4.1

Prova do resultado apresentado no item II.2.4.1

Sejam ainda conhecidos o vetor de médias e a matriz das covariâncias, de todas as variáveis envolvidas numa regressão linear, particionados como abaixo:

$$\mu = \left[\mu_Y \,\vdots\, \mu_X\right] \quad \text{e} \quad \mathbb{V} = \left[\begin{array}{c|c} \mathbb{V}_{YY} & \mathbb{V}'_{XY} \\ \hline \mathbb{V}_{XY} & \mathbb{V}_{XX} \end{array}\right]$$

onde μ_Y é a média de Y
μ_X é o vetor de médias da matriz **X**
$\mathbb{V}_{YY}$ é a variância de Y
$\mathbb{V}_{XX}$ é a matriz de covariâncias de **X**
$\mathbb{V}_{XY}$ são as covariâncias de Y com **X**

Sejam b_0 e **b**, respectivamente, os estimadores de β_0 e β.

O preditor linear do modelo definido dessa maneira será calculado como:

$$\begin{aligned} \hat{Y} &= b_0 + b_1 X_1 + \ldots + b_p X_p \\ &= b_0 + \mathbf{Xb} \end{aligned}$$

O erro de predição pode ser calculado pela diferença abaixo:

$$Y - \hat{Y} = Y - b_0 - \mathbf{Xb}$$

Uma maneira de se calcular b_0 e **b** é determinar seus valores de maneira que minimizem o erro quadrático médio da predição, que definido como:

$$EQM = E[Y - b_0 - \mathbf{Xb}]^2$$

Resultado: os valores de b_0 e **b** que minimizam o erro quadrático médio são dados por:

$$\mathbf{b} = \mathbb{V}_{XX}^{-1}\mathbb{V}_{XY}$$
$$b_0 = \mu_Y - \mu_X\mathbf{b}$$

e o valor mínimo do erro quadrático médio é atingido quando:

$$EQM = \mathbb{V}_{YY} - \mathbb{V}'_{XY}\mathbb{V}^{-1}_{XX}\mathbb{V}_{XY}$$

Para verificar os resultados acima basta calcular o erro quadrático médio a partir da definição do modelo.

Prova: $\hat{Y} = b_0 + \mathbf{Xb}$

Somando e subtraindo o valor $(\mu_Y - \mu_X\mathbf{b})$, tem-se

$$\hat{Y} = b_0 + \mathbf{Xb} + (\mu_Y - \mu_X\mathbf{b}) - (\mu_Y - \mu_X\mathbf{b})$$

$$\hat{Y} = b_0 + \mathbf{Xb} + \mu_Y - \mu_X\mathbf{b} - \mu_Y + \mu_X\mathbf{b}$$

$$= \mu_Y + (\mathbf{X} - \mu_X)\mathbf{b} - (\mu_Y - b - \mu\, b_X)$$

Então o erro quadrático médio será:

$$EQM = E\left[(Y - \mu_Y) - (\mathbf{X} - \mu_X)\mathbf{b} + (\mu_Y - b_0 - \mu_X\mathbf{b})\right]^2$$

Expandindo o produto notável acima tem-se:

$$EQM = E\Big\{(Y - \mu_Y)^2 + \left[(\mathbf{X} - \mu_X)\mathbf{b}\right]^2 + (\mu_Y - b_0 - \mu_X\mathbf{b})^2 -$$
$$-2(Y - \mu_Y)(\mathbf{X} - \mu_X)\mathbf{b} + 2(Y - \mu)(\mu_Y - b_0 - \mu_X\mathbf{b}) -$$
$$-2(\mathbf{X} - \mu_X)\mathbf{b}(\mu_Y - b_0 - \mu_X\mathbf{b})$$

$$EQM = E(Y - \mu_Y)^2 + E\left[(\mathbf{X} - \mu_X)\mathbf{b}\right]^2 + E(\mu_Y - b_0 - \mu_X\mathbf{b})^2 -$$
$$-E\left[2(Y - \mu_Y)(\mathbf{X} - \mu_X)\mathbf{b}\right] + E\left[2(Y - \mu_Y)(\mu_Y - b_0 - \mu_X\mathbf{b})\right] -$$
$$-E\left[2(\mathbf{X} - \mu_X)\mathbf{b}(\mu_Y - b_0 - \mu_X\mathbf{b})\right]$$

Lembrando de algumas resultados da estatística matemática pode-se simplificar a expressão.

Sejam $\mathbf{Z}$ e $\mathbf{W}$ matrizes aleatórias com médias μ_Z e μ_W. Sejam, ainda, $\mathbf{G}$ e $\mathbf{H}$ matrizes constantes com dimensões compatíveis. Então:

1- $E(\mathbf{Z} - \mu_Z) = \mathbf{0}$

2- $E(\mathbf{Z} - \mu_Z)^2 = Cov(\mathbf{Z})$

3- $E[(\mathbf{Z} - \mu_Z)(\mathbf{W} - \mu_W)] = Cov(\mathbf{Z}, \mathbf{W})$

4- $E(\mathbf{GZH}) = \mathbf{G}E(\mathbf{Z})\mathbf{H} = \mathbf{G}\mu_Z\mathbf{H}$

5- $E(\mathbf{H}) = \mathbf{H}$

Portanto, aplicando as propriedades acima, tem-se:

$$EQM = \mathbb{V}_{YY} + \mathbf{b}'\mathbb{V}_{XX}\mathbf{b} + (\mu_Y - b_0 - \mu_X\mathbf{b})^2 - 2\mathbb{V}'_{XY}\mathbf{b}$$

Somando e subtraindo $\mathbb{V}'_{XY}\mathbb{V}^{-1}_{XX}\mathbb{V}_{XY}$ da expressão acima, resulta:

$$EQM = \mathbb{V}_{YY} - \mathbb{V}'_{XY}\mathbb{V}^{-1}_{XX}\mathbb{V}_{XY} + (\mu_Y - b_0 - \mu_X\mathbf{b})^2 + \left[\mathbf{b}'\mathbb{V}_{XX}\mathbf{b} + \mathbb{V}'_{XY}\mathbb{V}^{-1}_{XX}\mathbb{V}_{XY} - 2\mathbb{V}'_{XY}\mathbf{b}\right]$$

Assim, apenas a terceira e a quarta parcela dependem de b_0 e $\mathbf{b}$. Substituindo-os pelos valores propostos tem-se o erro quadrático médio minimizado, pois tais parcelas são zeradas.

$$EQM = \mathbb{V}_{YY} - \mathbb{V}'_{XY}\mathbb{V}^{-1}_{XX}\mathbb{V}_{XY} + (\mu_Y - \mu_Y + \mu_X\mathbf{b} - \mu_X\mathbf{b})^2 + \left[\mathbb{V}'_{XY}\mathbb{V}^{-1}_{XX}\mathbb{V}_{XX}\mathbb{V}^{-1}_{XX}\mathbb{V}'_{XY} + \mathbb{V}'_{XY}\mathbb{V}^{-1}_{XX}\mathbb{V}_{XY} - 2\mathbb{V}'_{XY}\mathbb{V}^{-1}_{XX}\mathbb{V}_{XY}\right]$$

Mas,

$$\mathbb{V}^{-1}_{XX}\mathbb{V}_{XX} = I$$

Então:

$$EQM = \mathbb{V}_{YY} - \mathbb{V}'_{XY}\mathbb{V}^{-1}_{XX}\mathbb{V}_{XY} + \left[2\mathbb{V}'_{XY}\mathbb{V}^{-1}_{XX}\mathbb{V}_{XY} - 2\mathbb{V}'_{XY}\mathbb{V}^{-1}_{XX}\mathbb{V}_{XY}\right]$$

$$EQM = \mathbb{V}_{YY} - \mathbb{V}'_{XY}\mathbb{V}^{-1}_{XX}\mathbb{V}_{XY} \qquad \blacksquare$$

Dessa maneira pode-se escrever o preditor dos valores de Y como função apenas dos valores de **X**, do vetor de médias e da matriz de covariâncias:

$$\hat{Y} = \mu_Y + (\mathbf{X} - \mu_X)\, \mathbb{V}_{XX}^{-1}\mathbb{V}_{XY}$$

ANEXO E

Comparação entre a aplicação do método EM e das médias

E.1 Desvios padrões e diferenças relativas

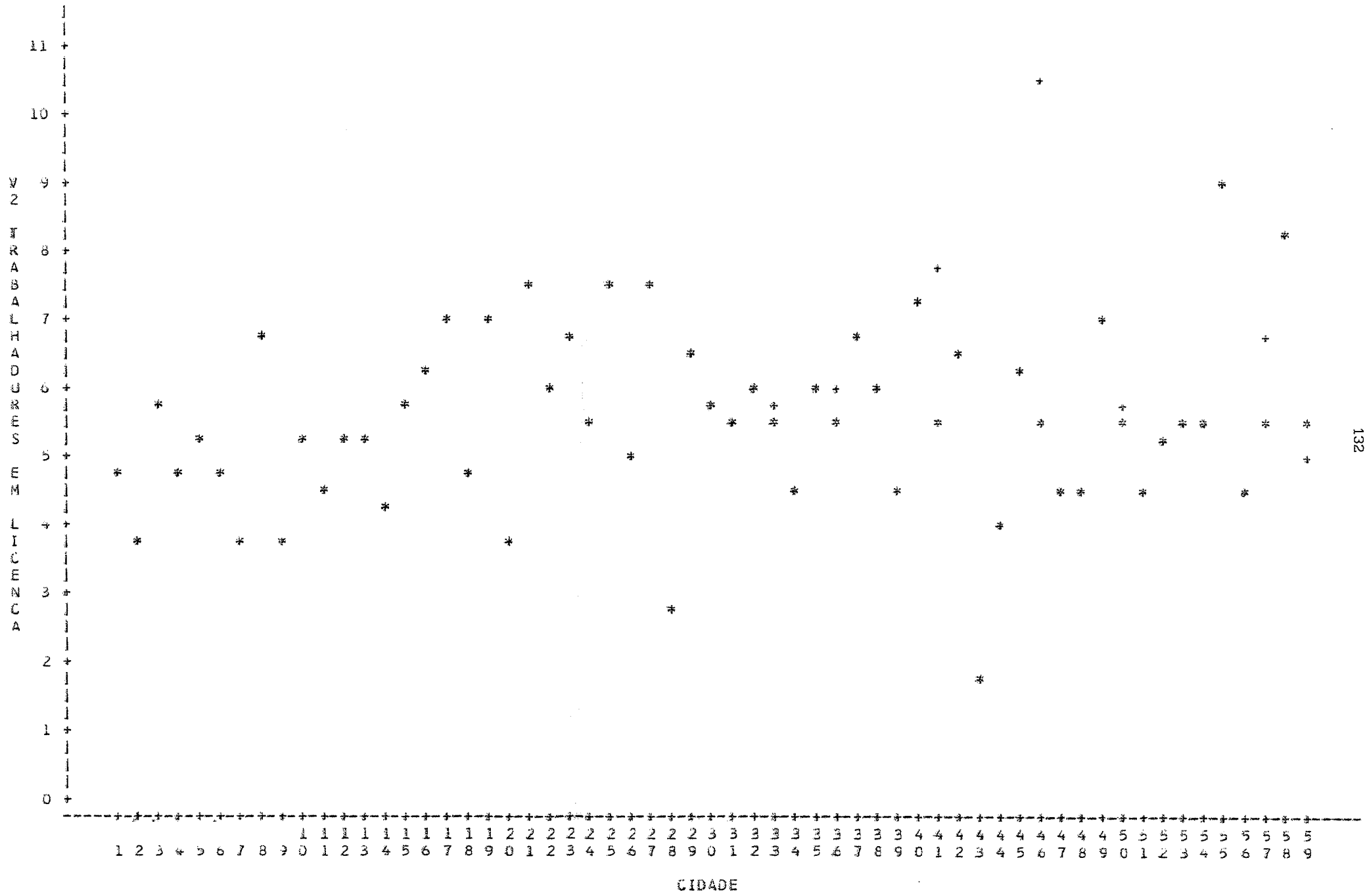
E.2 Gráficos comparativos

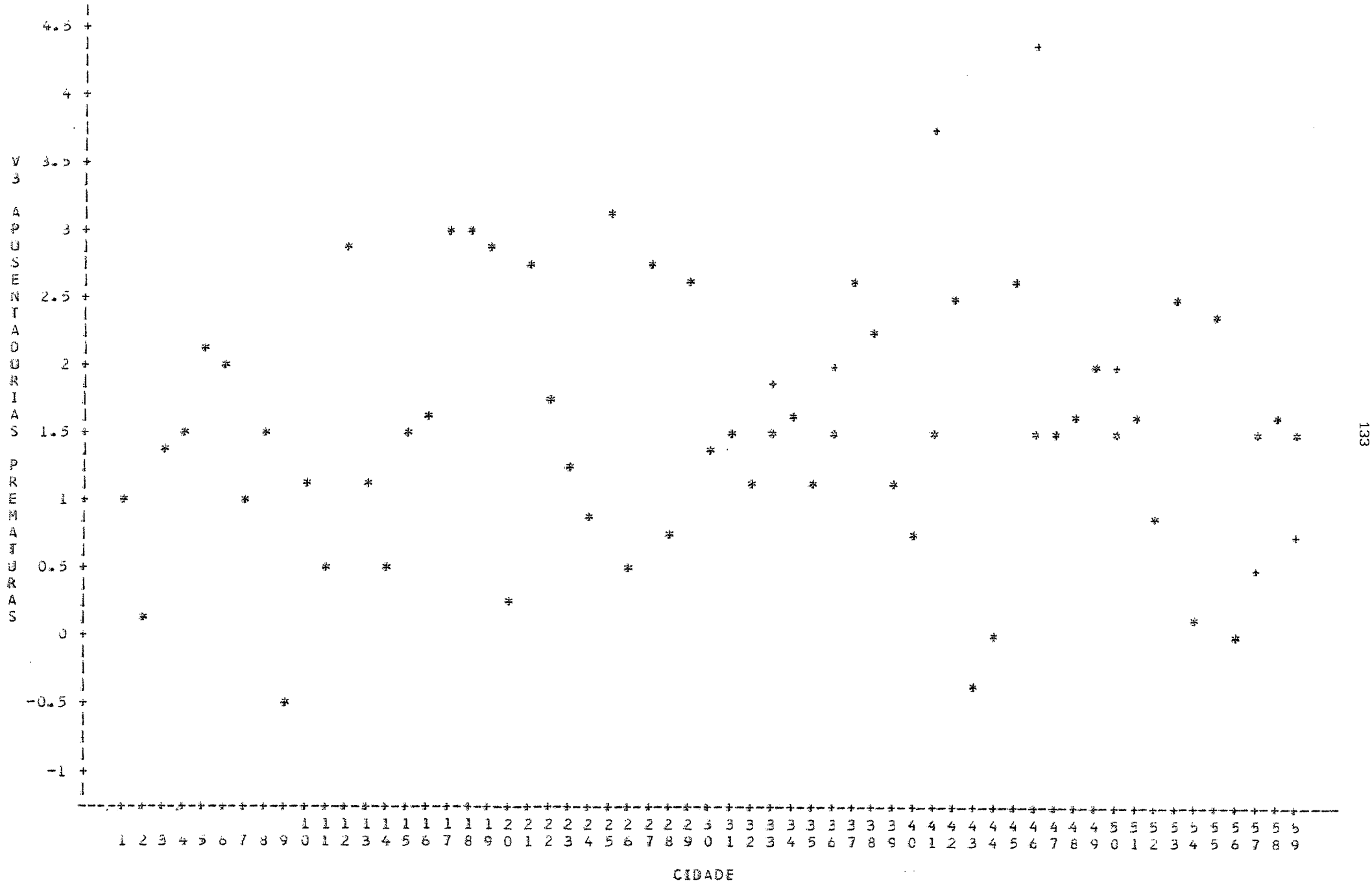
Comparação entre a aplicação do método EM e das médias

E.1 Desvios padrões e diferenças relativas

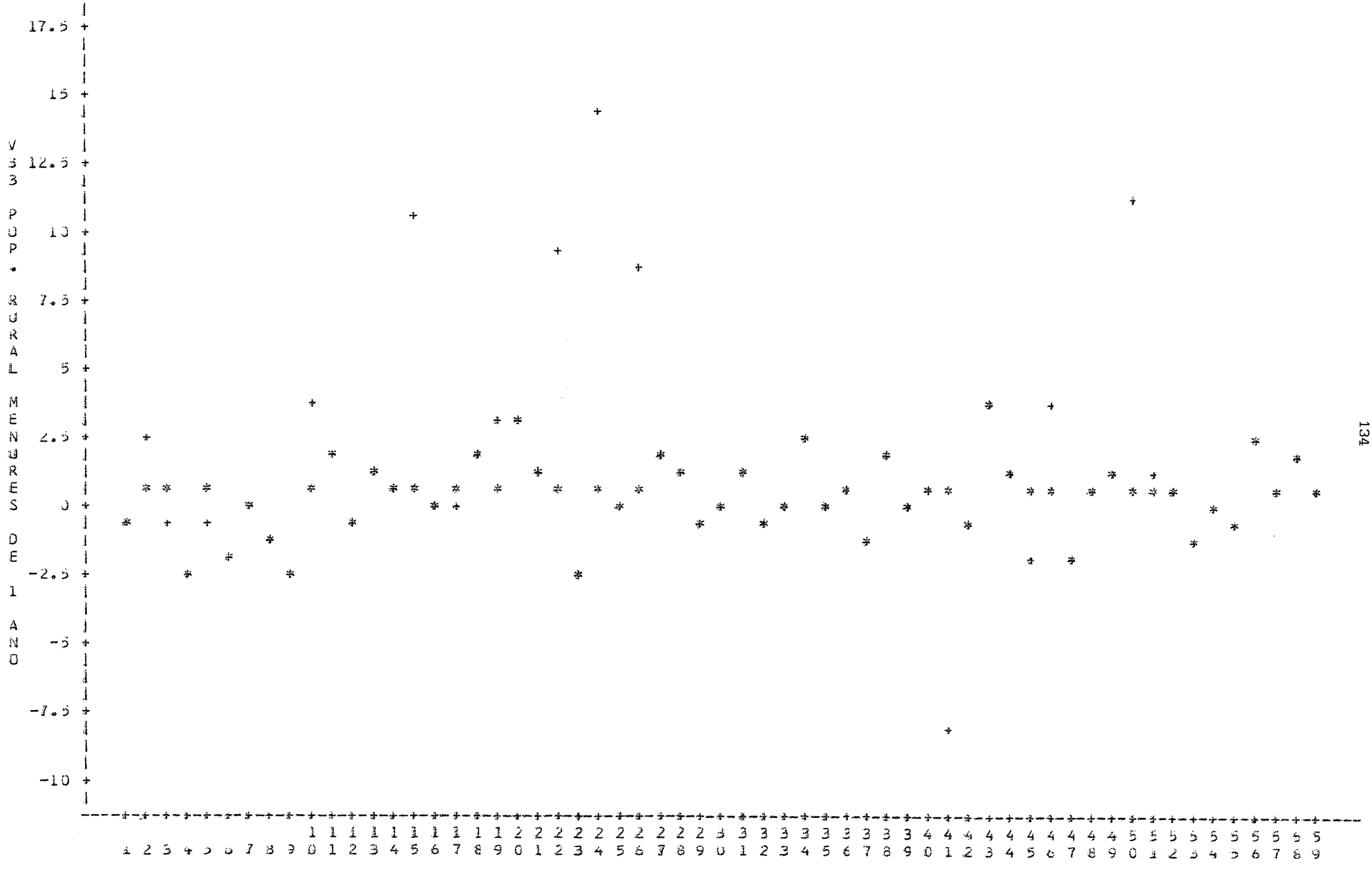
Variáveis	Au-sen-tes	Imputação			Dif. relat.	
		Sem	EM	Médias	EM	Médias
Trabalh. em licença	7	1,41	1,51	1,32	0,07	-0,06
Aposent. prematuras	7	0,95	1,02	0,89	0,08	-0,06
Pop. rur. < 1 ano	16	1,47	3,55	1,25	1,42	-0,15
Pop. rur. 1-4 anos	15	1,69	4,32	1,45	1,56	-0,14
Pop. rur. 5-19 anos	13	2,01	5,05	1,77	1,51	-0,12
Pop. rur. 20-49 anos	13	1,88	4,80	1,66	1,54	-0,12
Pop. rur. 50-65 anos	13	1,65	3,68	1,46	1,23	-0,12
Pop. rur.≥ 65 anos	13	1,40	3,14	1,23	1.25	-0,12
Moradias com esgoto	1	589665,91	58458,96	58455,37	-0,01	-0,01
Total de moradias	1	0,15	0,15	0,15	-0,01	-0,01
Moradias c/ água	2	237,22	235,24	233,09	-0,01	-0,02
Número de telefones	1	1,21	1,21	1,20	0,01	-0,01
Area do município	1	0,98	0,97	0,97	-0,01	-0,01
Alfabetizados >5 anos	1	5379,95	5371,10	5333,37	-0,00	-0,01
Total de terra arável	1	2,79	2,76	2,76	-0,01	-0,01
Total de terra cultiv.	5	2,81	3,44	2,69	0,22	-0,04
Estab. Rurais	10	3,98	4,66	3,62	0,17	-0,09

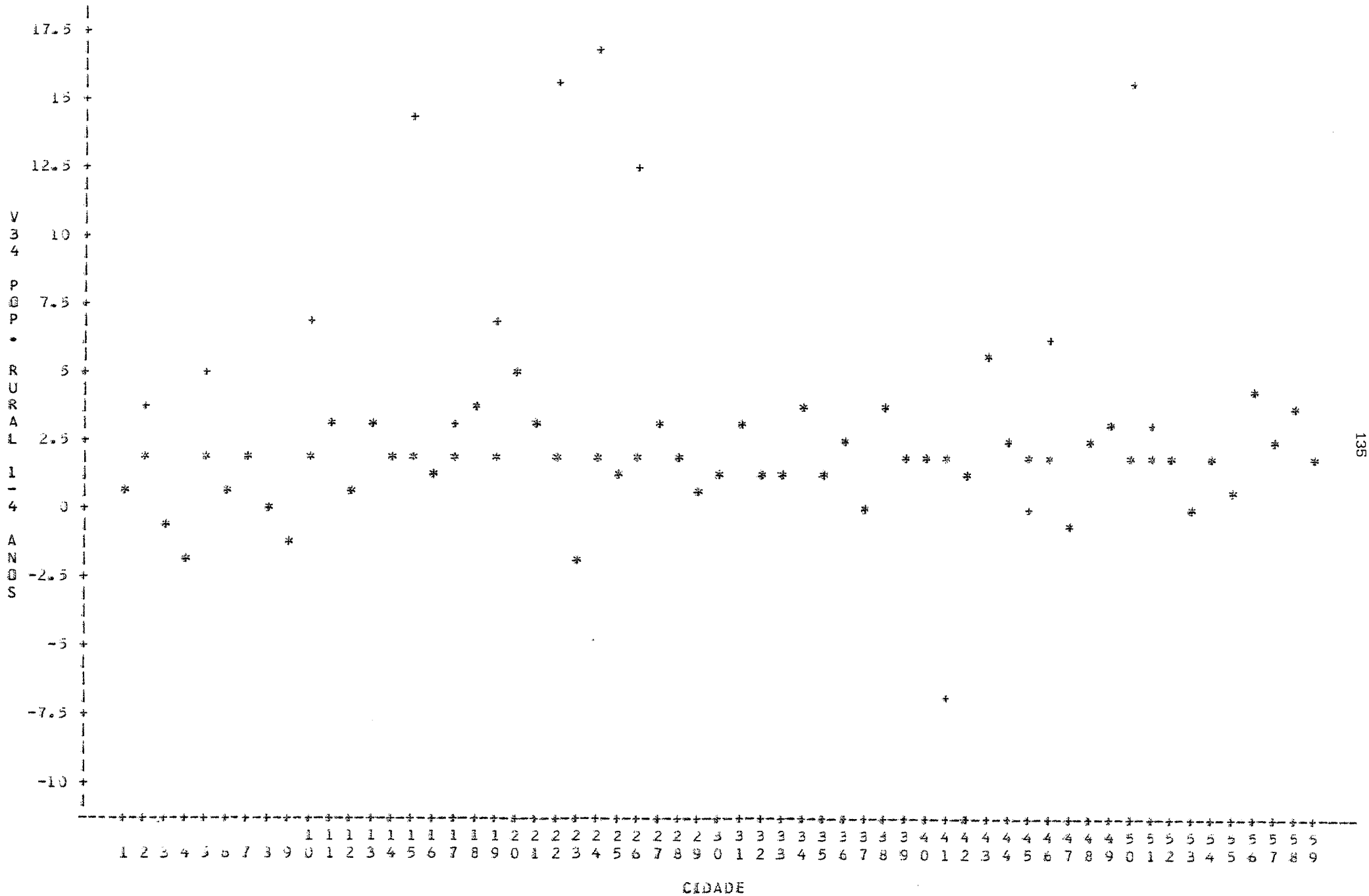
E.2 Gráficos comparativos dos resultados

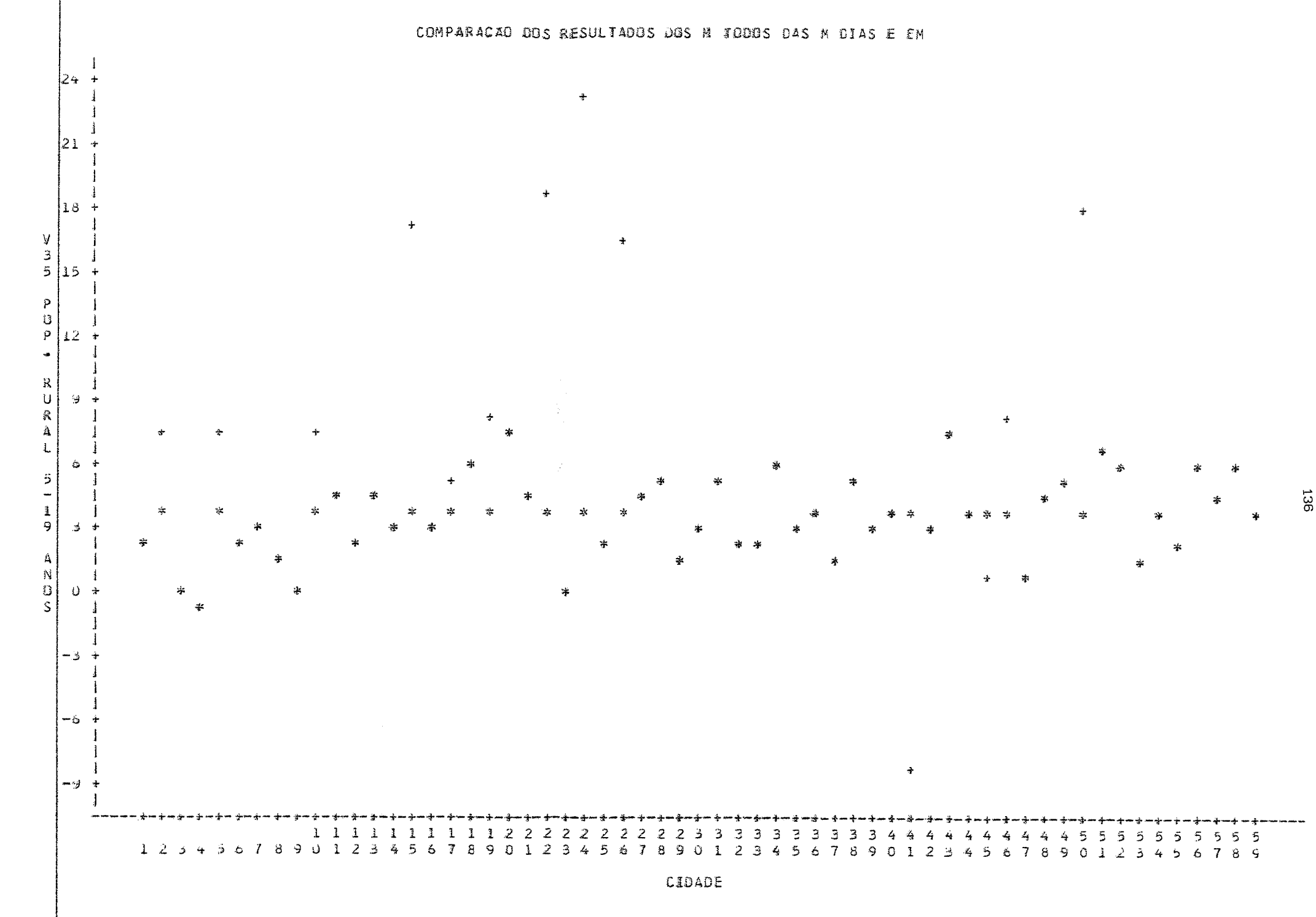

COMPARACAO DOS RESULTADOS DOS M TODOS DAS * DIAS E EM
V2 TRABALHADORES EM LICENCA
0 1 2 3 4 5 6 7 8 9 10 11
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59
CIDADE
132

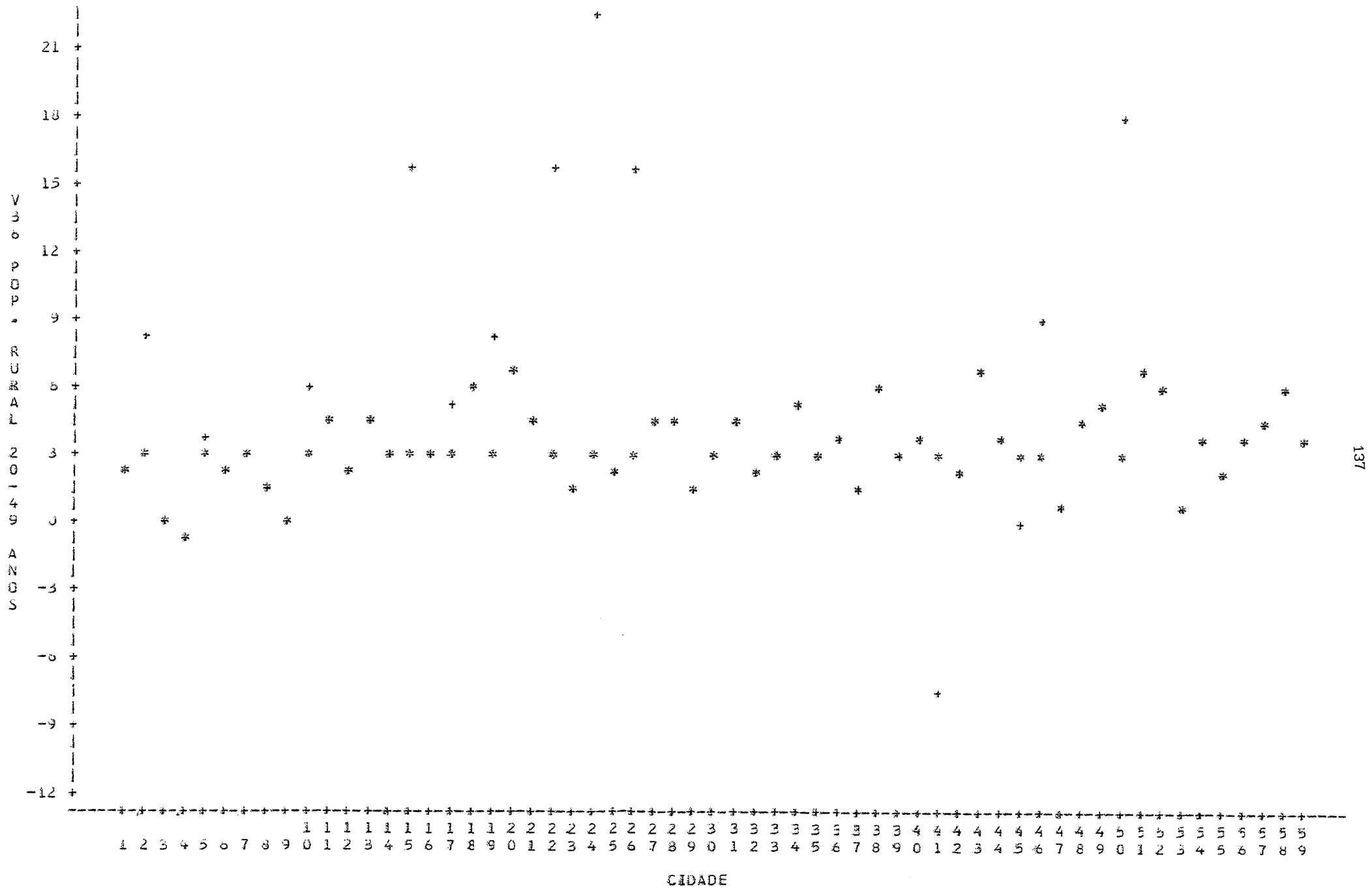

COMPARACAO DOS RESULTADOS DOS M TODOS DAS M DIAS E EM
VARIAS PREMATURAS APOSENTADORIAS
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
-0.5
-1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59
CIDADE

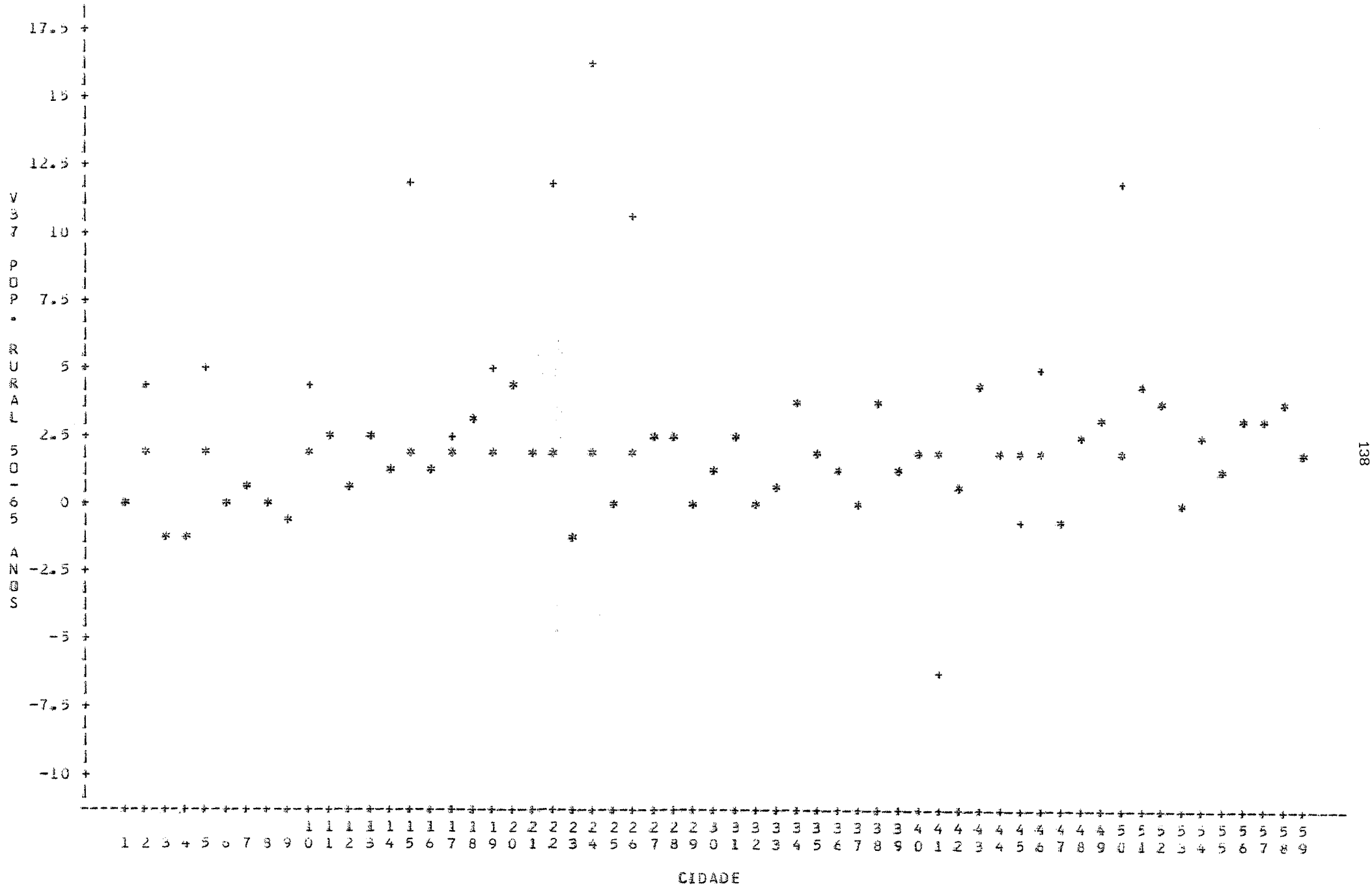
COMPARACAO DOS RESULTADOS DOS M TODOS DAS N DIAS E EM

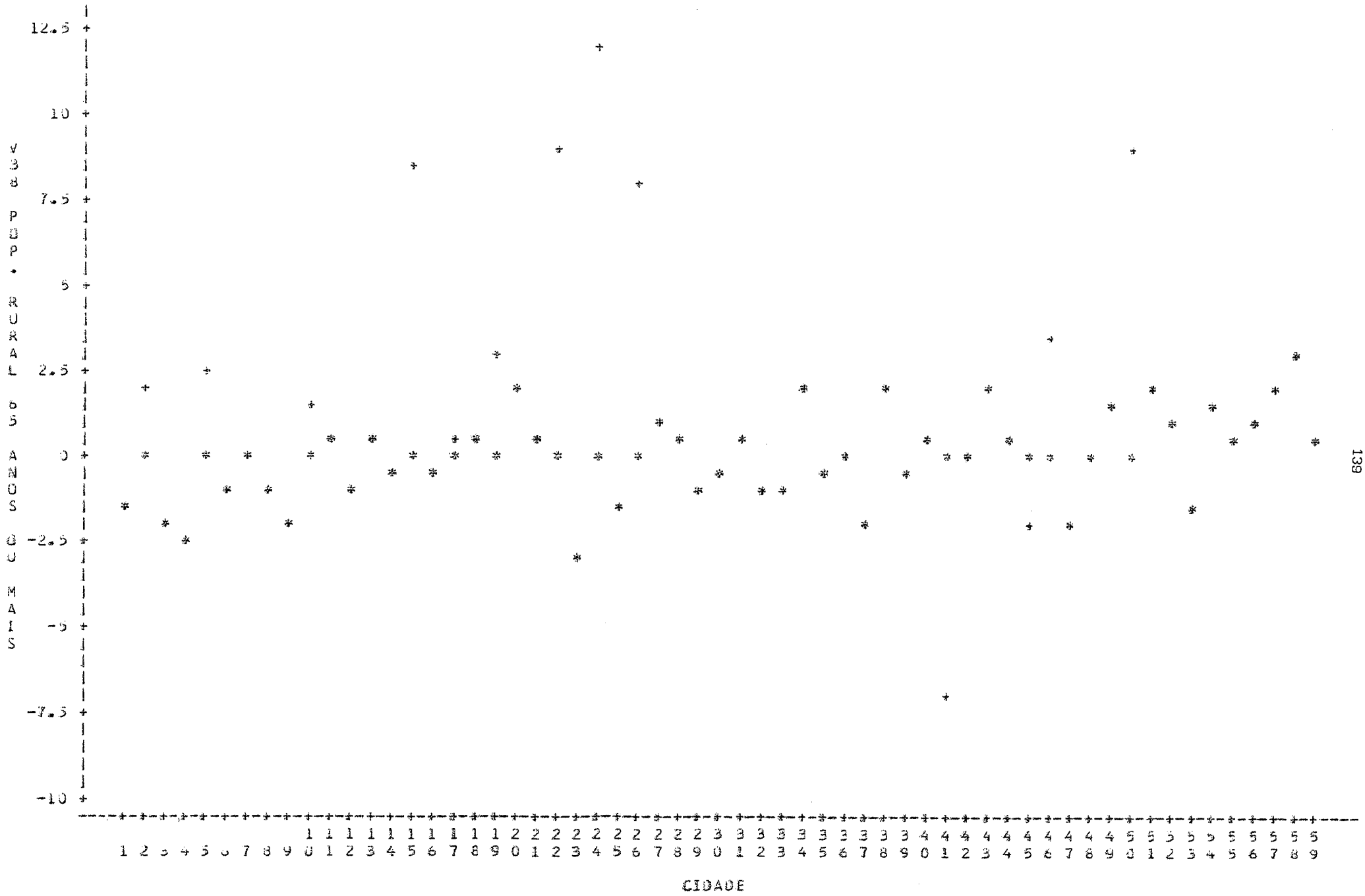


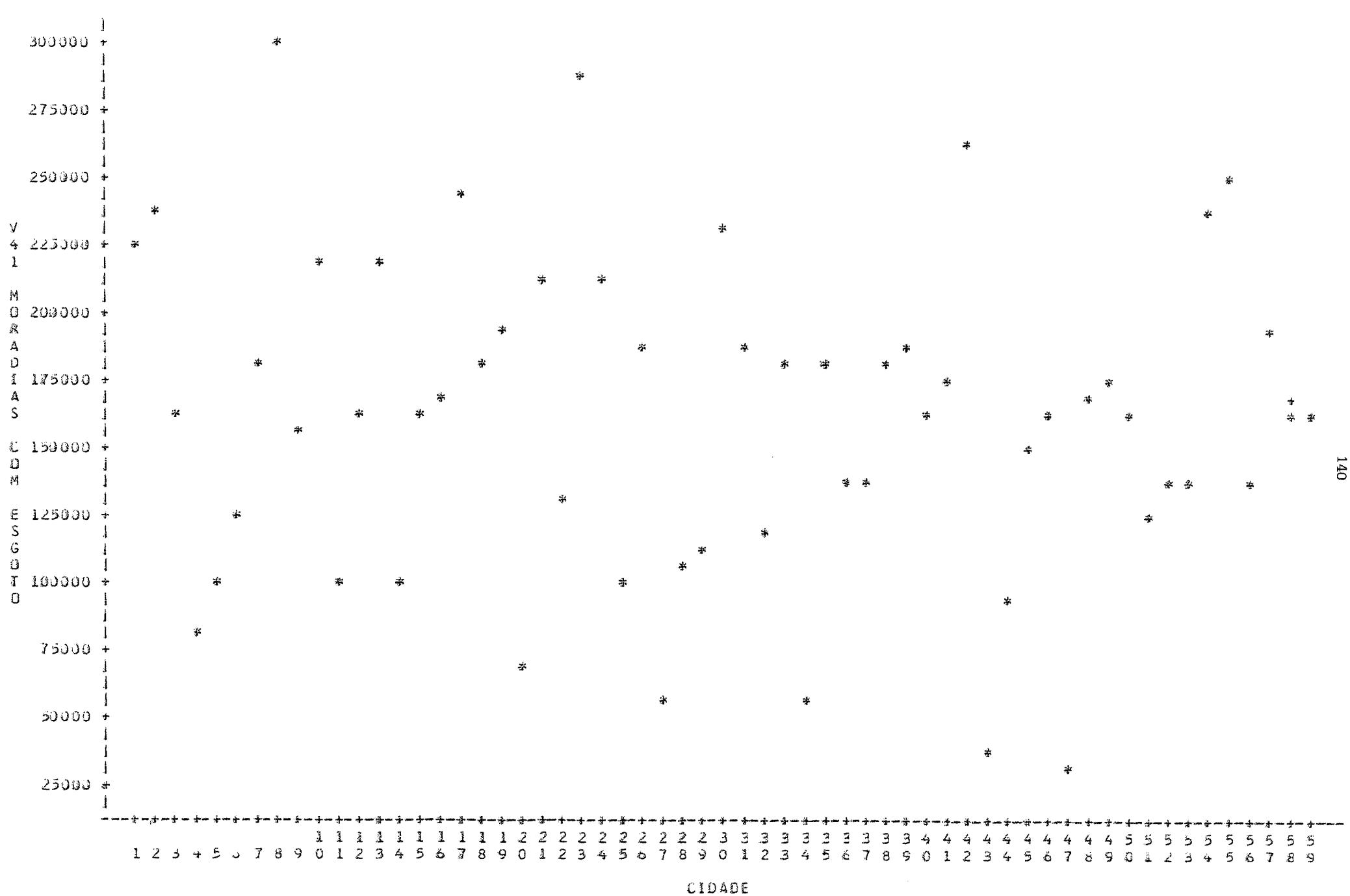

COMPARACAO DOS RESULTADOS DOS M TODOS DAS M DIAS E EM
V34 POP. RURAL 14 ANOS
17.5
15
12.5
10
7.5
5
2.5
0
-2.5
-5
-7.5
-10
CIDADE
135

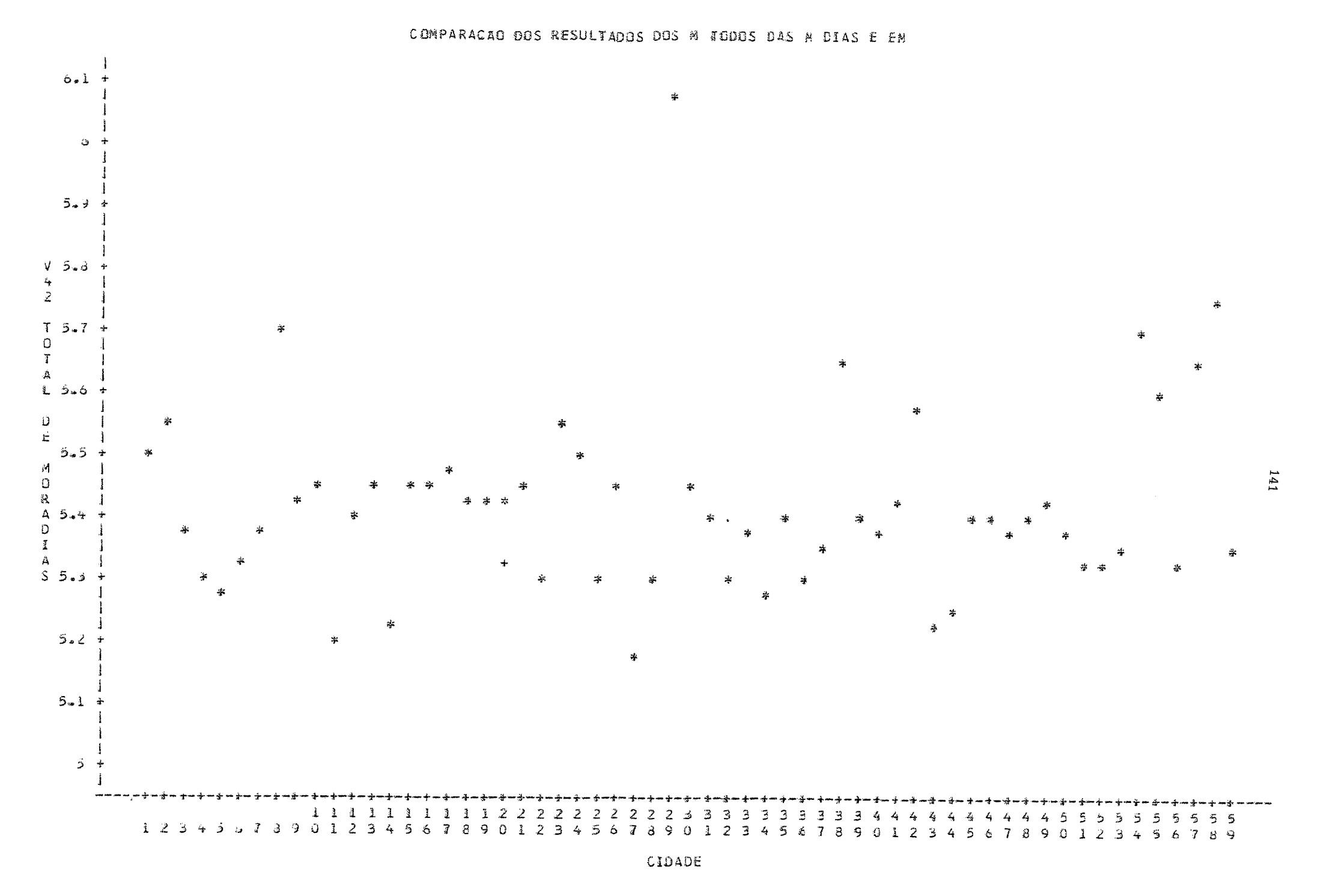

COMPARACAO DOS RESULTADOS DOS M TODOS DAS M DIAS E EM
V35 POP. RURAL 5-19 ANOS
24
21
18
15
12
9
6
3
0
-3
-6
-9
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59
CIDADE
136

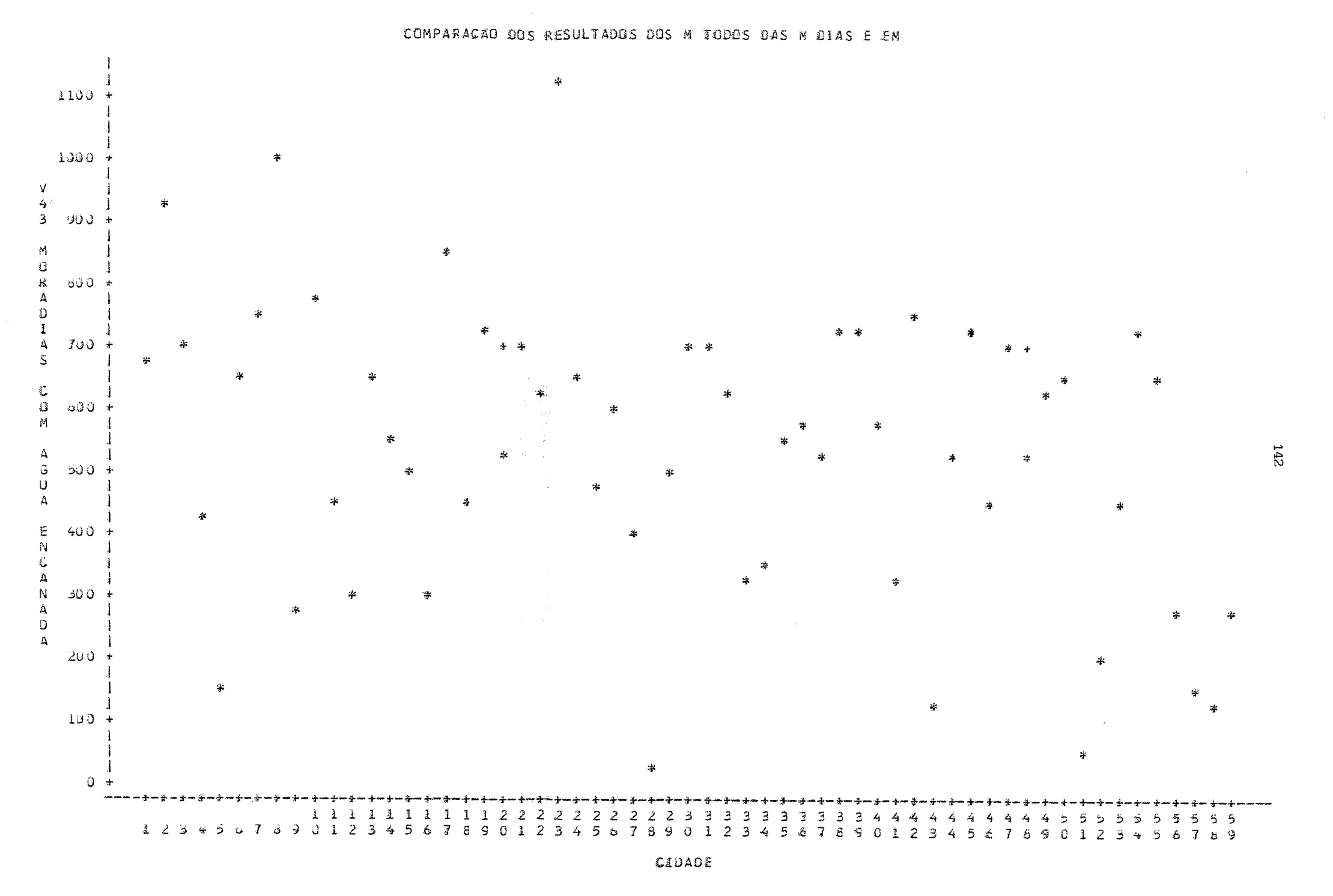

COMPARACAO DOS RESULTADOS DOS M TODOS DAS M DIAS E EM
V3o POP. RURAL 20-49 ANOS
21
18
15
12
9
6
3
0
-3
-6
-9
-12
137
CIDADE

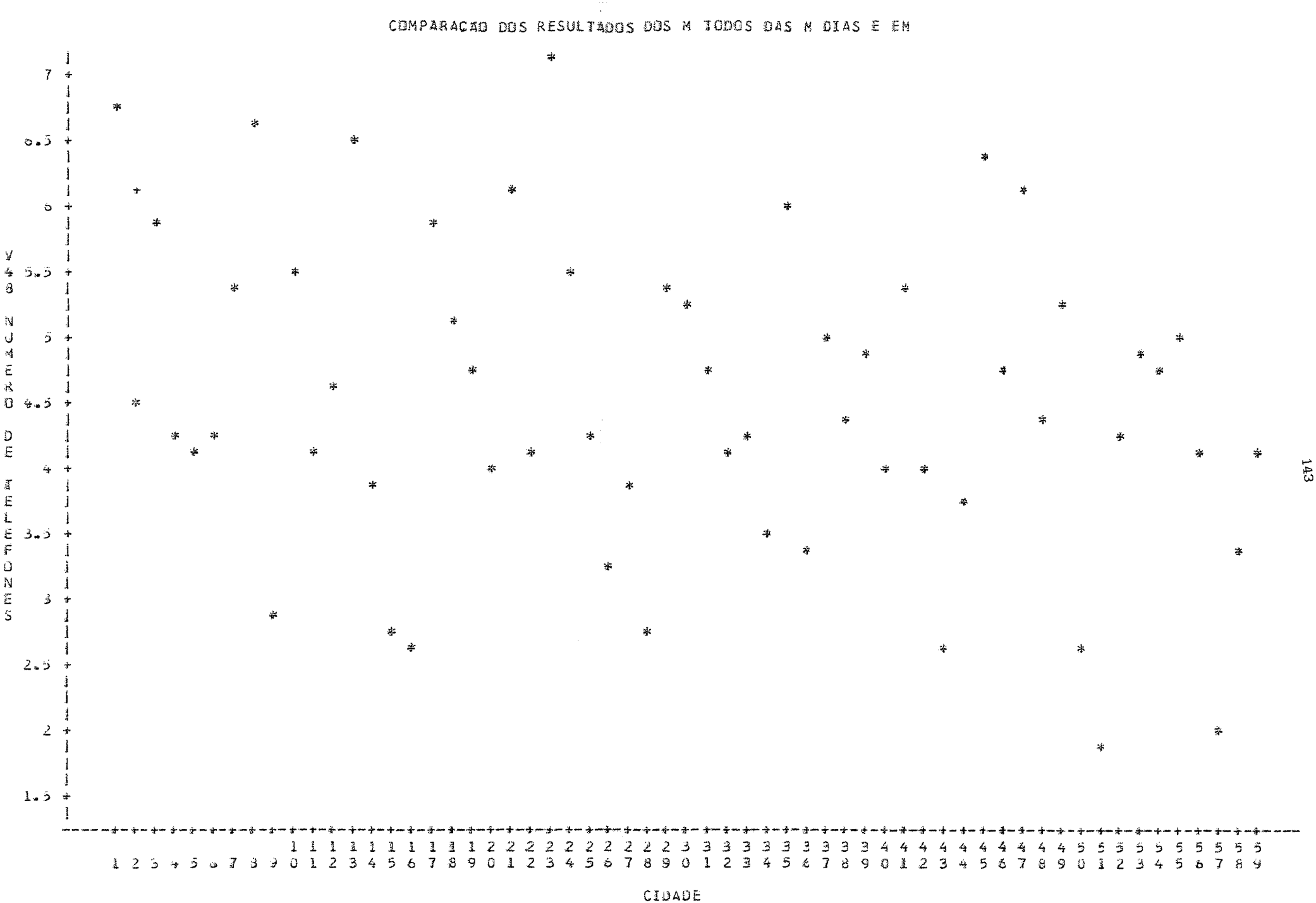

COMPARACAO DOS RESULTADOS DOS M TODOS DAS M DIAS E EM
17.5
15
12.5
10
7.5
5
2.5
0
-2.5
-5
-7.5
-10
V37 POP. RURAL 50-65 ANOS
CIDADE
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59

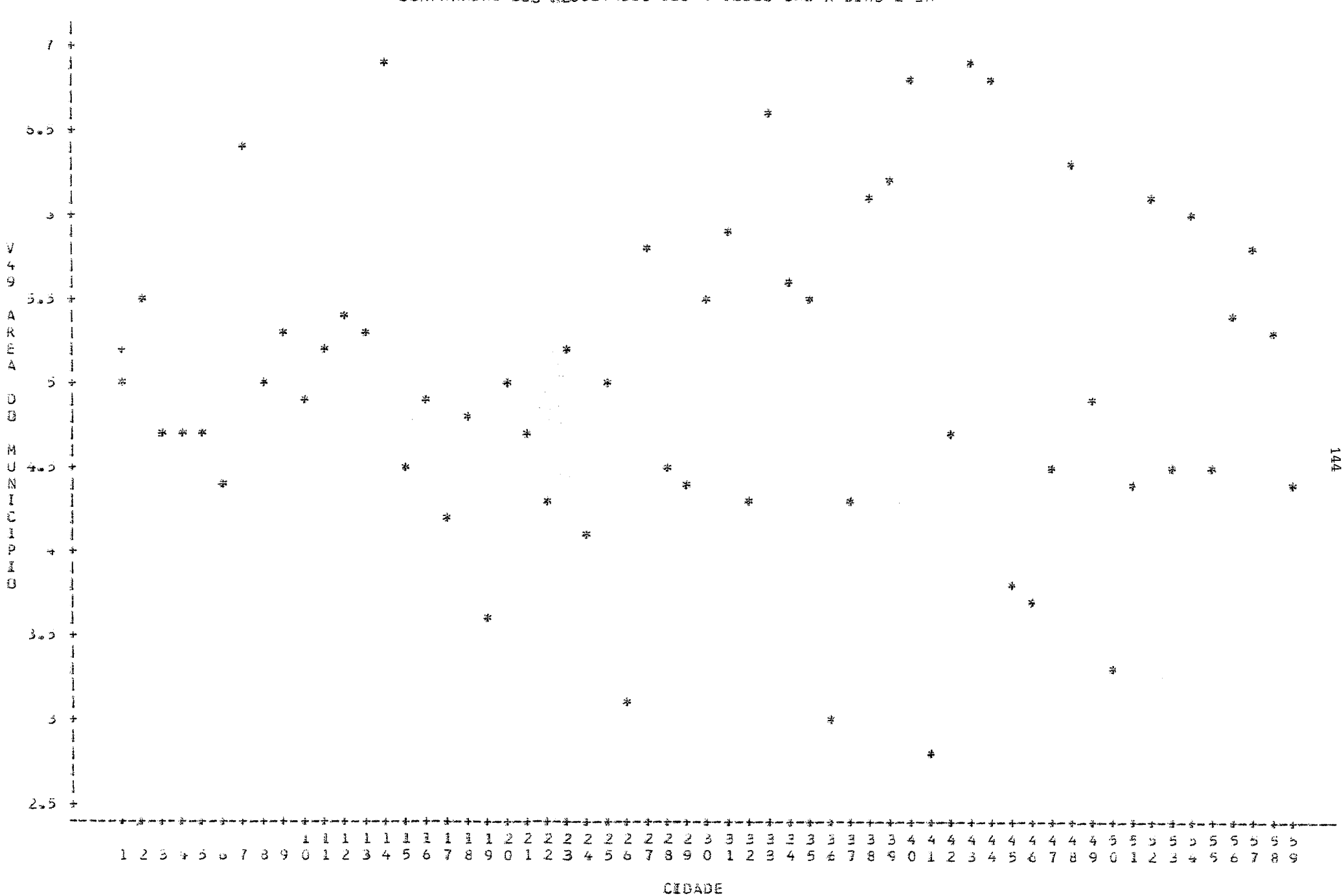

COMPARACAO DOS RESULTADOS DOS M TODOS DAS M DIAS E EM
12.5
10
7.5
5
2.5
0
-2.5
-5
-7.5
-10
VAR. POP. RURAL DOS ANOS DO MAIS
CIDADE
139

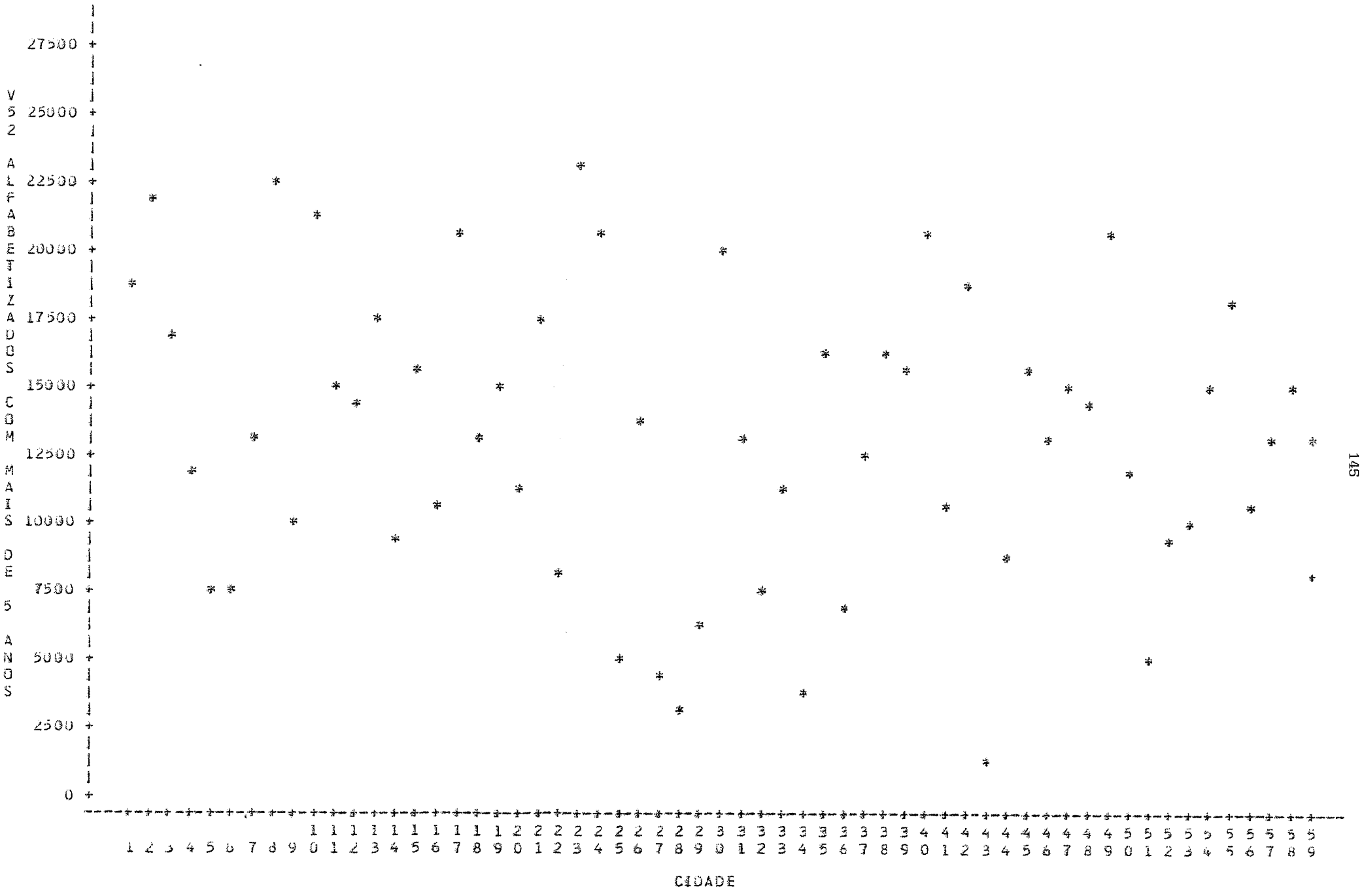

COMPARACAO DOS RESULTADOS DOS M TODOS DAS M DIAS E EM
VALMORADIAS COM ESGOTO
300000
275000
250000
225000
200000
175000
150000
125000
100000
75000
50000
25000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59
CIDADE
140

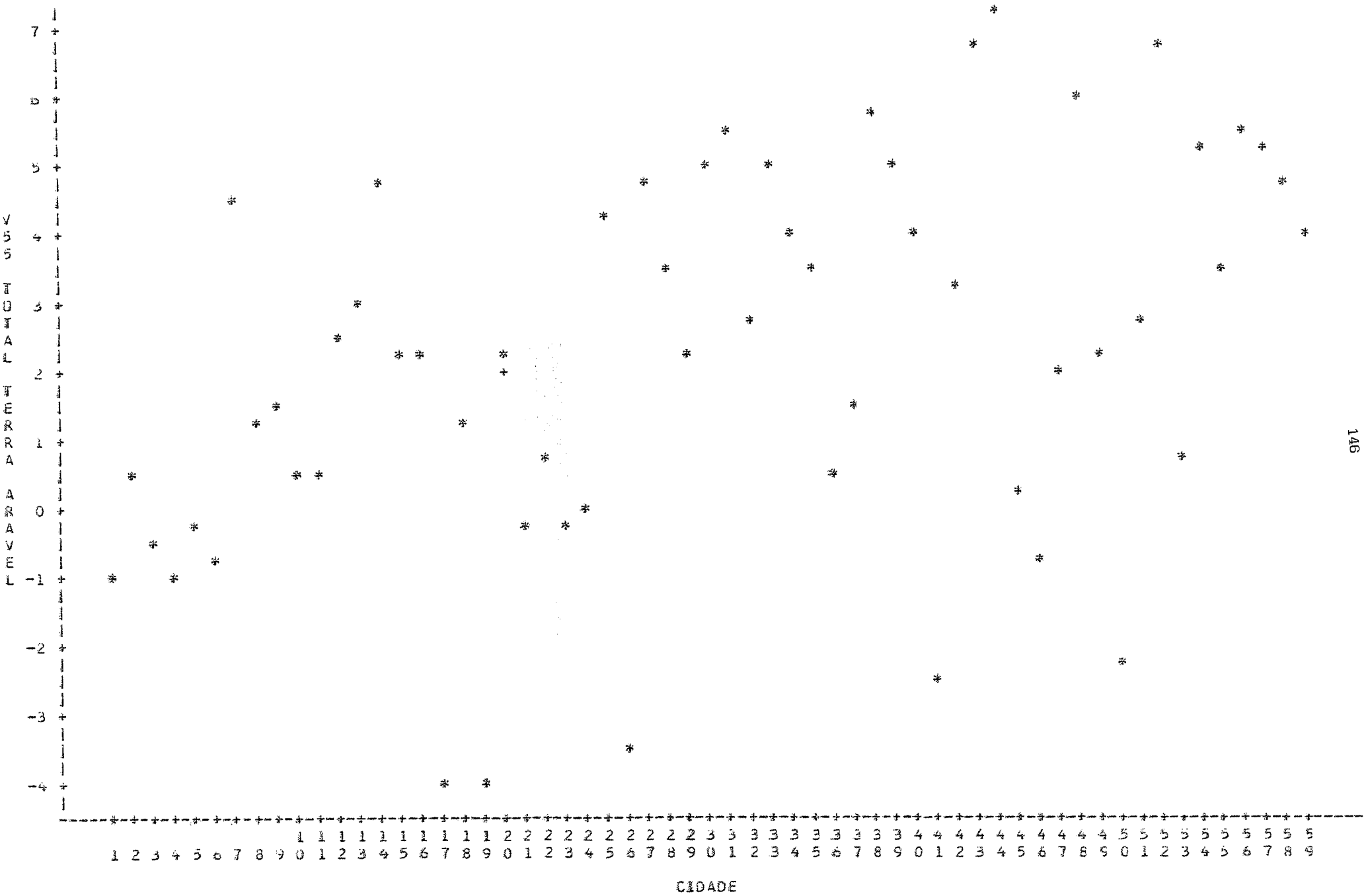

COMPARACAO DOS RESULTADOS DOS M TODOS DAS N DIAS E EM
V42 TOTAL DE MORADIAS
6.1
6
5.9
5.8
5.7
5.6
5.5
5.4
5.3
5.2
5.1
5
141
CIDADE

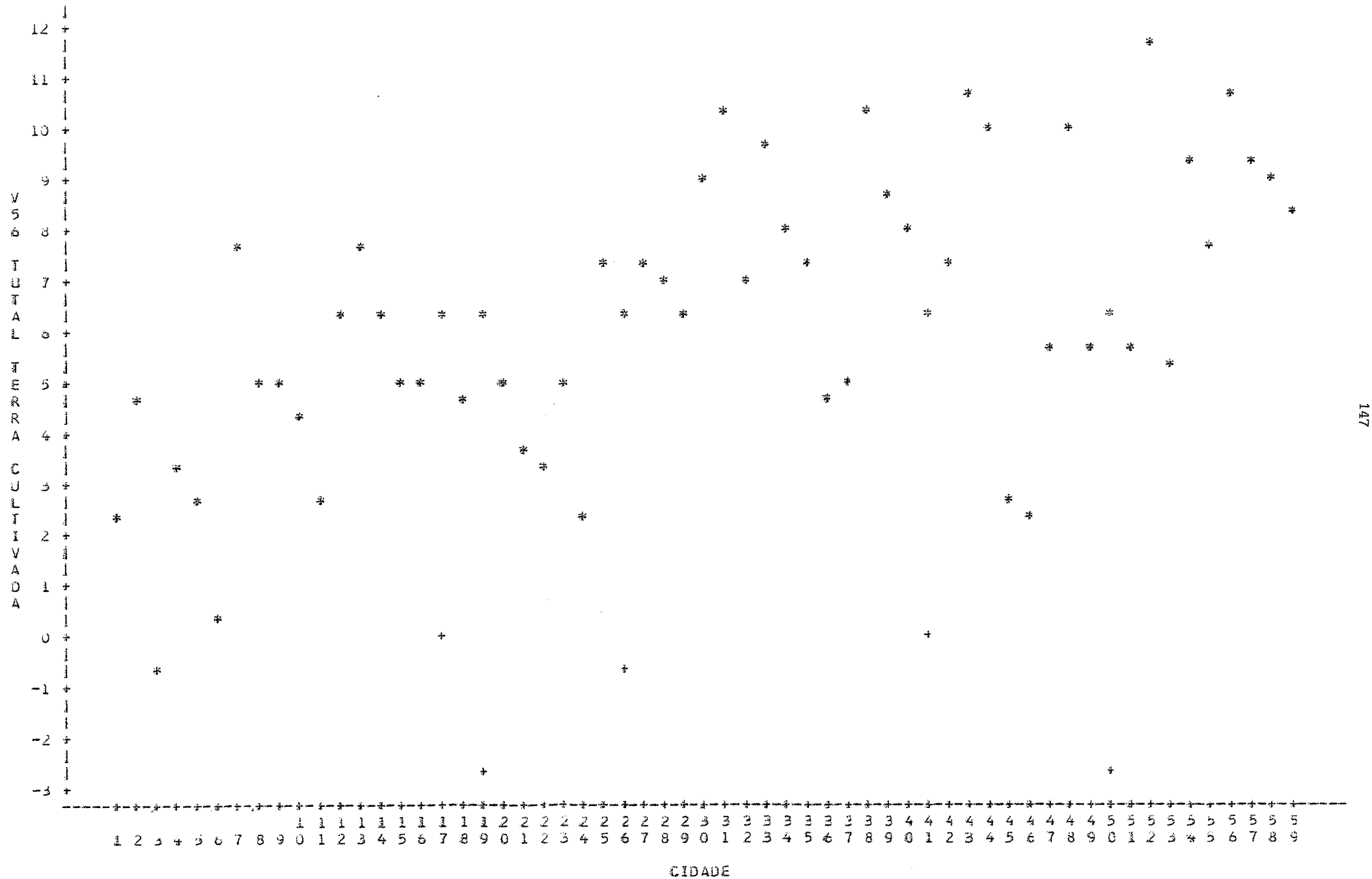

COMPARACAO DOS RESULTADOS DOS M TODOS DAS M DIAS E EM
V43 MGRADIAS CGM AGUA ENCANADA
1100
1000
900
800
700
600
500
400
300
200
100
0
CIDADE

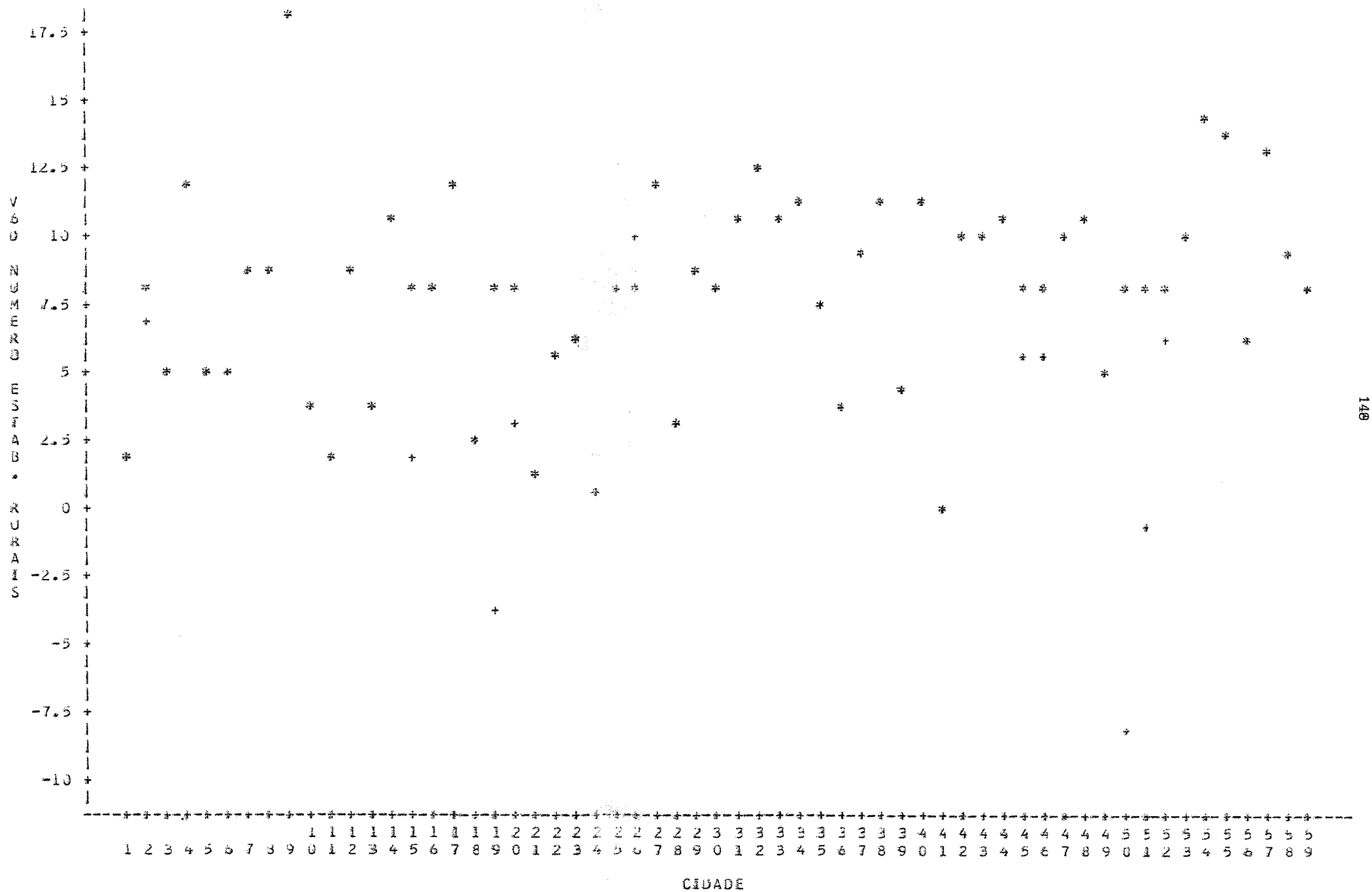

COMPARACAO DOS RESULTADOS DOS M TODOS DAS M DIAS E EM
VARIANCIA DE MESMO MEMBRO
7
6.5
6
5.5
5
4.5
4
3.5
3
2.5
2
1.5
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59
CIDADE


COMPARACAO DOS RESULTADOS DOS M TODOS DAS M DIAS E EM
V 4 9 A R E A D O M U N I C I P I O
7
6.5
6
5.5
5
4.5
4
3.5
3
2.5
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59
CIDADE
144


COMPARACAO DOS RESULTADOS DOS M TODOS DAS N DIAS E EM
V S 2 ALFABETIZADOS COM MAIS DE 5 ANOS
27500
25000
22500
20000
17500
15000
12500
10000
7500
5000
2500
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59
CIDADE


COMPARACAO DOS RESULTADOS DOS M TODOS DAS M DIAS E EM
VS TOTAL TERRA ARAVEL
7
6
5
4
3
2
1
0
-1
-2
-3
-4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59
CIDADE
146


COMPARACAO DOS RESULTADOS DOS M TODOS DAS M DIAS E EM
VSG TOTAL TERRA CULTIVADA
12
11
10
9
8
7
6
5
4
3
2
1
0
-1
-2
-3
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59
CIDADE


COMPARACAO DOS RESULTADOS DOS M TODOS DAS M DIAS E EM
VOLUMES ESTAB. RURAIS
NUMERO
17.5
15
12.5
10
7.5
5
2.5
0
-2.5
-5
-7.5
-10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59
CIDADE

ANEXO F

Exemplo de listagem da PROC UNIVARIATE do *SAS*

EXEMPLO DE SAIDA DA PROC UNIVARIATE DO SAS

UNIVARIATE

VARIABLE=V7 V7 MORTALIDADE 20-49 ANOS

MOMENTS

N	59	SUM WGTS	59
MEAN	38.9071	SUM	2295.52
STD DEV	8.96712	VARIANCE	80.4093
SKEWNESS	0.0724897	KURTOSIS	0.522277
USS	93975.8	CSS	4663.74
CV	23.0475	STD MEAN	1.16742
T:MEAN=0	33.3274	PROB>|T|	0.0001
SGN RANK	885	PROB>|S|	0.0001
NUM ¬= 0	59		
D:NORMAL	0.0778173	PROB>D	>.15

QUANTILES(DEF=4)

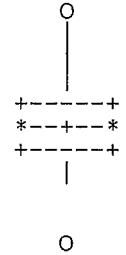
100% MAX	60.52	99%	60.52
75% Q3	44.4	95%	56.96
50% MED	38.24	90%	50.8
25% Q1	34.53	10%	27.27
0% MIN	13.27	5%	25.72
		1%	13.27
RANGE	47.25		
Q3-Q1	9.87		
MODE	47.73		

EXTREMES

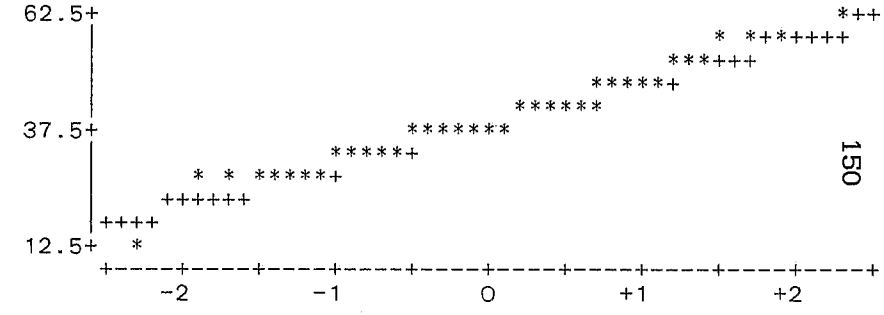
LOWEST	HIGHEST
13.27	51.25
25.22	56.06
25.72	56.96
25.73	57.42
26.95	60.52

```
STEM LEAF                          Õ
   6 1                             1
   5 677                           3
   5 111                           3
   4 5667888                       7
   4 00111122233334               13
   3 555556666666678899           18
   3 01133                         5
   2 56677889                      8
   2
   1
   1 3                             1
     ----+----+----+----+
  MULTIPLY STEM.LEAF BY 10**+01
```

BOXPLOT



NORMAL PROBABILITY PLOT



FREQUENCY TABLE

VALUE	COUNT	PERCENTS CELL	PERCENTS CUM
13.27	1	1.7	1.7
25.22	1	1.7	3.4
25.72	1	1.7	5.1
25.73	1	1.7	6.8
26.95	1	1.7	8.5
27.27	1	1.7	10.2
27.52	1	1.7	11.9
28.06	1	1.7	13.6
29.2	1	1.7	15.3
30.48	1	1.7	16.9
30.56	1	1.7	18.6
30.59	1	1.7	20.3
32.89	1	1.7	22.0
33.47	1	1.7	23.7
34.53	1	1.7	25.4
34.62	1	1.7	27.1
34.87	1	1.7	28.8
35.04	1	1.7	30.5
35.25	1	1.7	32.2
35.78	1	1.7	33.9
35.79	1	1.7	35.6
35.85	1	1.7	37.3
35.99	1	1.7	39.0
36.29	1	1.7	40.7
36.39	1	1.7	42.4
36.45	1	1.7	44.1
36.47	1	1.7	45.8
36.56	1	1.7	47.5
37.87	1	1.7	49.2
38.24	1	1.7	50.8
39.29	1	1.7	52.5
39.39	1	1.7	54.2
39.92	1	1.7	55.9
40.5	1	1.7	57.6
40.58	1	1.7	59.3
40.82	1	1.7	61.0
40.83	1	1.7	62.7
41.21	1	1.7	64.4
41.69	1	1.7	66.1
41.87	1	1.7	67.8
42.19	1	1.7	69.5
42.55	1	1.7	71.2
43.18	1	1.7	72.9
43.45	1	1.7	74.6
44.4	1	1.7	76.3
45.34	1	1.7	78.0
45.73	1	1.7	79.7
46.31	1	1.7	81.4
46.67	1	1.7	83.1
47.55	1	1.7	84.7
47.73	2	3.4	88.1
50.66	1	1.7	89.8
50.8	1	1.7	91.5
51.25	1	1.7	93.2
56.06	1	1.7	94.9
56.96	1	1.7	96.6
57.42	1	1.7	98.3
60.52	1	1.7	100.0

ANEXO G

Matrizes dos coeficientes fatoriais do Capítulo IV

G.1 Método das componentes principais - varimax -EM

G.2 Método das componentes principais - quartimax - EM

G.3 Método do fator principal - varimax - EM

G.4 Método do fator principal - quartimax - EM

G.5 Método das componentes principais - varimax - médias

G.5 Método das componentes principais - quartimax - médias

G.7 Método do fator principal - varimax - médias

G.8 Método do fator principal - quartimax - médias

ANALISE FATORIAL DOS DADOS PARA O ANO DE 1980
DADOS IMPUTADOS PELO METODO EM

ROTATION METHOD: VARIMAX

ROTATED FACTOR PATTERN

	FACTOR1	FACTOR2	FACTOR3	FACTOR4	FACTOR5	
V8	95 *	1	-5	2	-3	V8 MORTALIDADE 50-65 ANOS
V13	87 *	5	11	-25	15	V13 MORTAL. DOENCAS CARDIO-VASCULARES
V17	85 *	-14	-2	-2	33	V17 MORTAL. DOENCAS NEOPLASICAS
V31	83 *	-23	13	6	35	V31 POP. 50-65 ANOS
V32	71 *	-7	9	16	51 *	V32 POP. 65 ANOS OU MAIS
V7	71 *	-12	-7	-39	-4	V7 MORTALIDADE 20-49 ANOS
V42	56 *	-36	-2	11	-4	V42 TOTAL DE MORADIAS
V9	51 *	42	-2	0	45	V9 MORTALIDADE ACIMA DE 65 ANOS
V19	49	6	4	-33	-31	V19 MORTAL. CAUSAS VIOLENTAS
V28	-74 *	39	-8	3	-42	V28 POP. 1-4 ANOS
V27	-78 *	33	-13	-4	-33	V27 POP. MENORES DE 1 ANO
V5	-6	92 *	7	-7	3	V5 MORTALIDADE 1-4 ANOS
V20	-18	83 *	6	-38	5	V20 MORTAL. DOENCAS DIARREICAS
V21	-16	81 *	9	-42	4	V21 MORTAL. DOENCAS DIAR. 0-1 ANO
V29	-61 *	67 *	-3	22	-21	V29 POP. 5-19 ANOS
V10	2	67 *	-2	-46	17	V10 MORTALIDADE MENORES DE 1 ANO
V16	-2	63 *	22	-52 *	0	V16 MORTAL.DOENCAS RESP. 1-4 ANOS
V12	29	61 *	-10	-13	-3	V12 MORTAL. DOENCAS INFEC. E PARASIT.
V4	42	57 *	-15	-14	49	V4 MORTALIDADE GERAL
V61	-32	-50	-21	5	-2	V61 NUMERO EMPREGOS REGULARES
V41	50 *	-71 *	1	-17	8	V41 MORADIAS COM ESGOTO
V52	41	-76 *	5	-13	36	V52 ALFABETIZADOS COM MAIS DE 5 ANOS
V30	36	-80 *	-3	-33	11	V30 POP. 20-49 ANOS
V37	4	3	99 *	-5	-5	V37 POP. RURAL 50-65 ANOS
V35	2	5	99 *	-10	-1	V35 POP. RURAL 5-19 ANOS
V38	8	0	98 *	-6	-7	V38 POP. RURAL 65 ANOS OU MAIS
V36	7	0	98 *	-11	-2	V36 POP. RURAL 20-49 ANOS
V34	-1	5	98 *	-15	0	V34 POP. RURAL 1-4 ANOS
V33	2	4	98 *	-9	-2	V33 POP. RURAL MENORES DE 1 ANO
V56	-1	-5	-12	90 *	-19	V56 TOTAL TERRA CULTIVADA
V55	-5	4	-6	89 *	-20	V55 TOTAL TERRA ARAVEL
V49	-4	-10	-11	72 *	5	V49 AREA DO MUNICIPIO
V59	1	-29	-7	62 *	3	V59 CULTURAS RESP. 80% PROD.
V60	21	1	-36	57 *	-17	V60 NUMERO ESTAB. RURAIS
V14	43	10	10	-73 *	-20	V14 MORTAL.DOENCAS RESPIRAT.
V15	23	4	12	-81 *	-29	V15 MORTAL. DOENCAS RESP. 0-1 ANO
V54	17	6	0	8	93 *	V54 PESSOAS COM 11 OU MAIS ANOS ESTUDO
V53	23	5	3	4	90 *	V53 PESSOAS COM 8 OU MAIS ANOS ESTUDO
V22	33	16	4	14	67 *	V22 NUMERO DE LEITOS
V48	10	-49	-29	-16	64 *	V48 NUMERO DE TELEFONES
V43	32	-43	-2	-21	61 *	V43 MORADIAS COM AGUA ENCANADA
V62	13	13	-8	-42	52 *	V62 POPULACAO TOTAL
V44	-5	18	15	29	-63 *	V44 MORADIAS COM POCO
V3	-10	0	-16	-29	6	V3 APOSENTADORIAS PREMATURAS
V2	18	-26	-7	-15	-6	V2 TRABALHADORES EM LICENCA
V18	-17	9	-8	-1	16	V18 MORTAL. ACIDENTES DE TRAFEGO
V6	38	56 *	-9	2	11	V6 MORTALIDADE 5-19 ANOS
V1	16	2	-4	13	3	V1 NASCIDOS VIVOS
V11	22	11	-14	-15	38	V11 OBITOS FETAIS

ANALISE FATORIAL DOS DADOS PARA O ANO DE 1980
DADOS IMPUTADOS PELO METODO EM

ROTATION METHOD: VARIMAX

ROTATED FACTOR PATTERN

	FACTOR6	FACTOR7	FACTOR8	FACTOR9	
V8	4	0	2	0	V8 MORTALIDADE 50-65 ANOS
V13	-8	5	-6	-8	V13 MORTAL. DOENCAS CARDIO-VASCULARES
V17	2	12	10	-1	V17 MORTAL. DOENCAS NEOPLASICAS
V31	-5	-23	13	-2	V31 POP. 50-65 ANOS
V32	-8	-32	18	-7	V32 POP. 65 ANOS OU MAIS
V7	-13	34	-1	17	V7 MORTALIDADE 20-49 ANOS
V42	31	-10	33	9	V42 TOTAL DE MORADIAS
V9	3	14	2	-26	V9 MORTALIDADE ACIMA DE 65 ANOS
V19	-25	45	20	-6	V19 MORTAL. CAUSAS VIOLENTAS
V28	-13	11	7	-3	V28 POP. 1-4 ANOS
V27	-12	15	13	3	V27 POP. MENORES DE 1 ANO
V5	-10	14	-5	-3	V5 MORTALIDADE 1-4 ANOS
V20	4	9	13	8	V20 MORTAL. DOENCAS DIARREICAS
V21	6	6	12	9	V21 MORTAL. DOENCAS DIAR. 0-1 ANO
V29	-5	8	-15	-8	V29 POP. 5-19 ANOS
V10	38	12	3	2	V10 MORTALIDADE MENORES DE 1 ANO
V16	-21	8	-28	0	V16 MORTAL.DOENCAS RESP. 1-4 ANOS
V12	-24	7	-23	41	V12 MORTAL. DOENCAS INFEC. E PARASIT.
V4	22	14	10	-6	V4 MORTALIDADE GERAL
V61	14	39	-4	34	V61 NUMERO EMPREGOS REGULARES
V41	14	7	8	11	V41 MORADIAS COM ESGOTO
V52	1	16	-9	1	V52 ALFABETIZADOS COM MAIS DE 5 ANOS
V30	19	8	-3	13	V30 POP. 20-49 ANOS
V37	-3	-4	-1	-1	V37 POP. RURAL 50-65 ANOS
V35	-5	-2	-1	0	V35 POP. RURAL 5-19 ANOS
V38	0	-6	-1	-3	V38 POP. RURAL 65 ANOS OU MAIS
V36	-5	-5	-1	-1	V36 POP. RURAL 20-49 ANOS
V34	-2	-2	0	1	V34 POP. RURAL 1-4 ANOS
V33	-7	1	-2	1	V33 POP. RURAL MENORES DE 1 ANO
V56	-12	5	8	-2	V56 TOTAL TERRA CULTIVADA
V55	-15	12	3	0	V55 TOTAL TERRA ARAVEL
V49	-41	12	2	19	V49 AREA DO MUNICIPIO
V59	-4	-16	-17	35	V59 CULTURAS RESP. 80% PROD.
V60	-13	-11	-5	-3	V60 NUMERO ESTAB. RURAIS
V14	-10	11	-17	15	V14 MORTAL.DOENCAS RESPIRAT.
V15	-3	12	-15	10	V15 MORTAL. DOENCAS RESP. 0-1 ANO
V54	-2	5	-7	2	V54 PESSOAS COM 11 OU MAIS ANOS ESTUDO
V53	-3	11	-14	-7	V53 PESSOAS COM 8 OU MAIS ANOS ESTUDO
V22	13	22	19	27	V22 NUMERO DE LEITOS
V48	12	11	14	9	V48 NUMERO DE TELEFONES
V43	-6	-6	3	-7	V43 MORADIAS COM AGUA ENCANADA
V62	-34	-3	-11	38	V62 POPULACAO TOTAL
V44	2	26	-22	29	V44 MORADIAS COM POCO
V3	85 *	-9	-3	10	V3 APOSENTADORIAS PREMATURAS
V2	80 *	8	6	-17	V2 TRABALHADORES EM LICENCA
V18	1	87 *	7	-8	V18 MORTAL. ACIDENTES DE TRAFEGO
V6	-10	57 *	-14	16	V6 MORTALIDADE 5-19 ANOS
V1	2	5	81 *	-3	V1 NASCIDOS VIVOS
V11	23	34	-34	-44	V11 OBITOS FETAIS

ANALISE FATORIAL DOS DADOS PARA O ANO DE 1980
DADOS IMPUTADOS PELO METODO EM

ROTATION METHOD: QUARTIMAX

ROTATED FACTOR PATTERN

	FACTOR1	FACTOR2	FACTOR3	FACTOR4	FACTOR5	
V8	94 *	8	-6	4	-14	V8 MORTALIDADE 50-65 ANOS
V17	89 *	-6	-3	-1	23	V17 MORTAL. DOENCAS NEOPLASICAS
V31	88 *	-17	12	7	26	V31 POP. 50-65 ANOS
V13	88 *	13	10	-23	4	V13 MORTAL. DOENCAS CARDIO-VASCULARES
V32	76 *	-2	9	18	43	V32 POP. 65 ANOS OU MAIS
V7	72 *	-4	-8	-38	-13	V7 MORTALIDADE 20-49 ANOS
V42	59 *	-34	-3	10	-9	V42 TOTAL DE MORADIAS
V9	52 *	47	-3	2	38	V9 MORTALIDADE ACIMA DE 65 ANOS
V19	46	12	3	-31	-38	V19 MORTAL. CAUSAS VIOLENTAS
V29	-69 *	62 *	-2	22	-15	V29 POP. 5-19 ANOS
V28	-82 *	33	-7	3	-34	V28 POP. 1-4 ANOS
V27	-83 *	27	-13	-4	-25	V27 POP. MENORES DE 1 ANO
V5	-13	92 *	8	-5	2	V5 MORTALIDADE 1-4 ANOS
V20	-22	82 *	6	-36	5	V20 MORTAL. DOENCAS DIARREICAS
V21	-20	80 *	9	-40	4	V21 MORTAL. DOENCAS DIAR. 0-1 ANO
V10	0	68 *	-3	-46	15	V10 MORTALIDADE MENORES DE 1 ANO
V16	-6	65 *	22	-50 *	-1	V16 MORTAL.DOENCAS RESP. 1-4 ANOS
V12	24	65 *	-10	-11	-8	V12 MORTAL. DOENCAS INFEC. E PARASIT.
V6	35	63 *	-9	4	5	V6 MORTALIDADE 5-19 ANOS
V4	44	62 *	-16	-13	43	V4 MORTALIDADE GERAL
V61	-27	-50	-21	3	2	V61 NUMERO EMPREGOS REGULARES
V41	57 *	-66 *	0	-18	3	V41 MORADIAS COM ESGOTO
V52	52 *	-71 *	4	-14	32	V52 ALFABETIZADOS COM MAIS DE 5 ANOS
V30	45	-75 *	-3	-35	7	V30 POP. 20-49 ANOS
V37	4	3	99 *	-5	-5	V37 POP. RURAL 50-65 ANOS
V35	3	5	99 *	-10	-1	V35 POP. RURAL 5-19 ANOS
V38	8	0	98 *	-6	-8	V38 POP. RURAL 65 ANOS OU MAIS
V36	8	0	98 *	-11	-2	V36 POP. RURAL 20-49 ANOS
V34	-1	5	98 *	-15	0	V34 POP. RURAL 1-4 ANOS
V33	2	4	98 *	-9	-2	V33 POP. RURAL MENORES DE 1 ANO
V56	-5	-7	-12	90 *	-18	V56 TOTAL TERRA CULTIVADA
V55	-10	2	-5	89 *	-19	V55 TOTAL TERRA ARAVEL
V49	-4	-10	-10	73 *	5	V49 AREA DO MUNICIPIO
V59	2	-31	-7	62 *	4	V59 CULTURAS RESP. 80% PROD.
V60	17	1	-35	58 *	-18	V60 NUMERO ESTAB. RURAIS
V14	41	16	10	-71 *	-26	V14 MORTAL.DOENCAS RESPIRAT.
V15	21	8	12	-80 *	-32	V15 MORTAL. DOENCAS RESP. 0-1 ANO
V54	27	10	0	8	91 *	V54 PESSOAS COM 11 OU MAIS ANOS ESTUDO
V53	32	10	2	4	87 *	V53 PESSOAS COM 8 OU MAIS ANOS ESTUDO
V48	22	-46	-30	-17	63 *	V48 NUMERO DE TELEFONES
V22	40	21	3	15	62 *	V22 NUMERO DE LEITOS
V43	43	-39	-2	-21	58 *	V43 MORADIAS COM AGUA ENCANADA
V62	18	18	-8	-40	49	V62 POPULACAO TOTAL
V44	-14	17	15	29	-63 *	V44 MORADIAS COM POCO
V3	-8	-2	-17	-32	8	V3 APOSENTADORIAS PREMATURAS
V2	20	-25	-7	-18	-7	V2 TRABALHADORES EM LICENCA
V18	-14	13	-8	-1	17	V18 MORTAL. ACIDENTES DE TRAFEGO
V1	17	2	-5	14	1	V1 NASCIDOS VIVOS
V11	25	15	-15	-16	35	V11 OBITOS FETAIS

ANALISE FATORIAL DOS DADOS PARA O ANO DE 1980
DADOS IMPUTADOS PELO METODO EM

ROTATION METHOD: QUARTIMAX

ROTATED FACTOR PATTERN

	FACTOR6	FACTOR7	FACTOR8	FACTOR9	
V8	4	-2	0	-1	V8 MORTALIDADE 50-65 ANOS
V17	1	11	8	-2	V17 MORTAL. DOENCAS NEOPLASICAS
V31	-6	-22	11	-2	V31 POP. 50-65 ANOS
V13	-9	3	-7	-9	V13 MORTAL. DOENCAS CARDIO-VASCULARES
V32	-9	-32	16	-6	V32 POP. 65 ANOS OU MAIS
V7	-15	33	-2	15	V7 MORTALIDADE 20-49 ANOS
V42	30	-8	31	9	V42 TOTAL DE MORADIAS
V9	3	11	2	-27	V9 MORTALIDADE ACIMA DE 65 ANOS
V19	-26	43	19	-8	V19 MORTAL. CAUSAS VIOLENTAS
V29	-2	6	-13	-8	V29 POP. 5-19 ANOS
V28	-11	10	9	-3	V28 POP. 1-4 ANOS
V27	-11	14	15	3	V27 POP. MENORES DE 1 ANO
V5	-9	9	-3	-4	V5 MORTALIDADE 1-4 ANOS
V20	4	5	15	8	V20 MORTAL. DOENCAS DIARREICAS
V21	7	2	14	9	V21 MORTAL. DOENCAS DIAR. 0-1 ANO
V10	37	9	5	2	V10 MORTALIDADE MENORES DE 1 ANO
V16	-21	4	-26	-1	V16 MORTAL.DOENCAS RESP. 1-4 ANOS
V12	-24	4	-23	40	V12 MORTAL. DOENCAS INFEC. E PARASIT.
V6	-9	54 *	-14	14	V6 MORTALIDADE 5-19 ANOS
V4	22	11	10	-7	V4 MORTALIDADE GERAL
V61	13	43	-5	34	V61 NUMERO EMPREGOS REGULARES
V41	11	10	5	11	V41 MORADIAS COM ESGOTO
V52	-2	19	-11	1	V52 ALFABETIZADOS COM MAIS DE 5 ANOS
V30	16	11	-5	13	V30 POP. 20-49 ANOS
V37	-2	-3	-1	0	V37 POP. RURAL 50-65 ANOS
V35	-4	-2	-1	0	V35 POP. RURAL 5-19 ANOS
V38	1	-5	-1	-2	V38 POP. RURAL 65 ANOS OU MAIS
V36	-5	-4	-1	-1	V36 POP. RURAL 20-49 ANOS
V34	-2	-1	1	2	V34 POP. RURAL 1-4 ANOS
V33	-6	2	-2	1	V33 POP. RURAL MENORES DE 1 ANO
V56	-9	6	7	-3	V56 TOTAL TERRA CULTIVADA
V55	-12	12	2	-1	V55 TOTAL TERRA ARAVEL
V49	-39	13	1	18	V49 AREA DO MUNICIPIO
V59	-3	-13	-18	35	V59 CULTURAS RESP. 80% PROD.
V60	-11	-12	-6	-4	V60 NUMERO ESTAB. RURAIS
V14	-12	9	-17	14	V14 MORTAL.DOENCAS RESPIRAT.
V15	-6	11	-14	9	V15 MORTAL. DOENCAS RESP. 0-1 ANO
V54	-3	6	-8	3	V54 PESSOAS COM 11 OU MAIS ANOS ESTUDO
V53	-3	11	-15	-7	V53 PESSOAS COM 8 OU MAIS ANOS ESTUDO
V48	9	14	13	10	V48 NUMERO DE TELEFONES
V22	13	22	18	28	V22 NUMERO DE LEITOS
V43	-8	-4	2	-6	V43 MORADIAS COM AGUA ENCANADA
V62	-37	-3	-11	39	V62 POPULACAO TOTAL
V44	3	26	-22	27	V44 MORADIAS COM POCO
V3	84 *	-9	-3	11	V3 APOSENTADORIAS PREMATURAS
V2	79 *	9	6	-16	V2 TRABALHADORES EM LICENCA
V18	1	87 *	7	-9	V18 MORTAL. ACIDENTES DE TRAFEGO
V1	3	5	81 *	-2	V1 NASCIDOS VIVOS
V11	22	32	-34	-45	V11 OBITOS FETAIS

ANALISE FATORIAL DOS DADOS PARA O ANO DE 1980
DADOS IMPUTADOS PELO METODO EM

ROTATION METHOD: VARIMAX

ROTATED FACTOR PATTERN

	FACTOR1	FACTOR2	FACTOR3	FACTOR4	FACTOR5	
V8	95 *	1	-5	2	-2	V8 MORTALIDADE 50-65 ANOS
V13	86 *	5	11	-26	16	V13 MORTAL. DOENCAS CARDIO-VASCULARES
V17	84 *	-15	-2	-2	34	V17 MORTAL. DOENCAS NEOPLASICAS
V31	83 *	-22	13	6	35	V31 POP. 50-65 ANOS
V32	71 *	-6	9	16	52 *	V32 POP. 65 ANOS OU MAIS
V7	69 *	-14	-7	-37	-4	V7 MORTALIDADE 20-49 ANOS
V42	53 *	-36	-2	10	-3	V42 TOTAL DE MORADIAS
V9	49	40	-2	-2	46	V9 MORTALIDADE ACIMA DE 65 ANOS
V19	47	6	4	-31	-28	V19 MORTAL. CAUSAS VIOLENTAS
V28	-74 *	40	-8	2	-43	V28 POP. 1-4 ANOS
V27	-77 *	33	-13	-4	-34	V27 POP. MENORES DE 1 ANO
V5	-6	91 *	7	-9	5	V5 MORTALIDADE 1-4 ANOS
V20	-17	82 *	6	-39	4	V20 MORTAL. DOENCAS DIARREICAS
V21	-16	80 *	9	-43	3	V21 MORTAL. DOENCAS DIAR. 0-1 ANO
V29	-60 *	68 *	-3	21	-19	V29 POP. 5-19 ANOS
V10	2	64 *	-2	-48	17	V10 MORTALIDADE MENORES DE 1 ANO
V16	-1	61 *	22	-52 *	2	V16 MORTAL.DOENCAS RESP. 1-4 ANOS
V12	28	57 *	-10	-12	-2	V12 MORTAL. DOENCAS INFEC. E PARASIT.
V4	42	56 *	-15	-15	50	V4 MORTALIDADE GERAL
V61	-29	-47	-21	7	-4	V61 NUMERO EMPREGOS REGULARES
V41	49	-71 *	1	-15	7	V41 MORADIAS COM ESGOTO
V52	40	-78 *	5	-12	36	V52 ALFABETIZADOS COM MAIS DE 5 ANOS
V30	36	-82 *	-2	-32	10	V30 POP. 20-49 ANOS
V37	4	3	99 *	-6	-5	V37 POP. RURAL 50-65 ANOS
V35	2	5	99 *	-10	-1	V35 POP. RURAL 5-19 ANOS
V38	8	1	98 *	-6	-8	V38 POP. RURAL 65 ANOS OU MAIS
V36	7	0	98 *	-11	-2	V36 POP. RURAL 20-49 ANOS
V34	-1	5	98 *	-15	-1	V34 POP. RURAL 1-4 ANOS
V33	2	4	97 *	-10	-2	V33 POP. RURAL MENORES DE 1 ANO
V56	-1	-3	-12	90 *	-19	V56 TOTAL TERRA CULTIVADA
V55	-5	6	-6	89 *	-20	V55 TOTAL TERRA ARAVEL
V49	-4	-9	-11	72 *	4	V49 AREA DO MUNICIPIO
V59	0	-29	-8	59 *	2	V59 CULTURAS RESP. 80% PROD.
V60	18	1	-33	53 *	-13	V60 NUMERO ESTAB. RURAIS
V14	43	9	10	-71 *	-19	V14 MORTAL.DOENCAS RESPIRAT.
V15	23	3	12	-79 *	-28	V15 MORTAL. DOENCAS RESP. 0-1 ANO
V54	15	4	0	7	95 *	V54 PESSOAS COM 11 OU MAIS ANOS ESTUDO
V53	21	3	3	3	92 *	V53 PESSOAS COM 8 OU MAIS ANOS ESTUDO
V22	32	14	3	14	63 *	V22 NUMERO DE LEITOS
V48	9	-50	-28	-15	62 *	V48 NUMERO DE TELEFONES
V43	32	-43	-1	-20	58 *	V43 MORADIAS COM AGUA ENCANADA
V62	13	11	-7	-38	48	V62 POPULACAO TOTAL
V11	20	10	-13	-18	39	V11 OBITOS FETAIS
V44	-6	17	13	28	-59 *	V44 MORADIAS COM POCO
V3	-10	-2	-16	-30	6	V3 APOSENTADORIAS PREMATURAS
V2	17	-27	-7	-17	-5	V2 TRABALHADORES EM LICENCA
V18	-16	8	-8	-1	16	V18 MORTAL. ACIDENTES DE TRAFEGO
V6	38	54 *	-9	2	13	V6 MORTALIDADE 5-19 ANOS
V1	16	2	-4	14	2	V1 NASCIDOS VIVOS

ANALISE FATORIAL DOS DADOS PARA O ANO DE 1980
DADOS IMPUTADOS PELO METODO EM

ROTATION METHOD: VARIMAX

ROTATED FACTOR PATTERN

	FACTOR6	FACTOR7	FACTOR8	FACTOR9	
V8	4	1	2	-1	V8 MORTALIDADE 50-65 ANOS
V13	-8	6	-3	-6	V13 MORTAL. DOENCAS CARDIO-VASCULARES
V17	2	12	12	1	V17 MORTAL. DOENCAS NEOPLASICAS
V31	-6	-24	15	0	V31 POP. 50-65 ANOS
V32	-10	-34	22	-5	V32 POP. 65 ANOS OU MAIS
V7	-13	34	0	18	V7 MORTALIDADE 20-49 ANOS
V42	25	-8	31	5	V42 TOTAL DE MORADIAS
V9	4	15	3	-21	V9 MORTALIDADE ACIMA DE 65 ANOS
V19	-23	43	17	-2	V19 MORTAL. CAUSAS VIOLENTAS
V28	-13	13	7	-4	V28 POP. 1-4 ANOS
V27	-13	16	12	2	V27 POP. MENORES DE 1 ANO
V5	-10	16	-7	-3	V5 MORTALIDADE 1-4 ANOS
V20	3	11	16	14	V20 MORTAL. DOENCAS DIARREICAS
V21	5	8	15	15	V21 MORTAL. DOENCAS DIAR. 0-1 ANO
V29	-3	9	-19	-11	V29 POP. 5-19 ANOS
V10	36	13	5	9	V10 MORTALIDADE MENORES DE 1 ANO
V16	-19	8	-32	1	V16 MORTAL.DOENCAS RESP. 1-4 ANOS
V12	-23	8	-26	35	V12 MORTAL. DOENCAS INFEC. E PARASIT.
V4	22	15	12	-1	V4 MORTALIDADE GERAL
V61	11	30	-6	23	V61 NUMERO EMPREGOS REGULARES
V41	13	5	8	12	V41 MORADIAS COM ESGOTO
V52	0	15	-9	0	V52 ALFABETIZADOS COM MAIS DE 5 ANOS
V30	18	6	-1	15	V30 POP. 20-49 ANOS
V37	-3	-4	-2	-1	V37 POP. RURAL 50-65 ANOS
V35	-5	-2	-2	0	V35 POP. RURAL 5-19 ANOS
V38	0	-6	-1	-4	V38 POP. RURAL 65 ANOS OU MAIS
V36	-5	-5	-2	-1	V36 POP. RURAL 20-49 ANOS
V34	-2	-2	0	2	V34 POP. RURAL 1-4 ANOS
V33	-7	1	-3	1	V33 POP. RURAL MENORES DE 1 ANO
V56	-10	5	8	-4	V56 TOTAL TERRA CULTIVADA
V55	-13	11	3	-1	V55 TOTAL TERRA ARAVEL
V49	-37	10	1	18	V49 AREA DO MUNICIPIO
V59	-7	-15	-11	21	V59 CULTURAS RESP. 80% PROD.
V60	-11	-8	-2	-4	V60 NUMERO ESTAB. RURAIS
V14	-10	11	-23	16	V14 MORTAL.DOENCAS RESPIRAT.
V15	-3	11	-20	12	V15 MORTAL. DOENCAS RESP. 0-1 ANO
V54	-2	4	-7	4	V54 PESSOAS COM 11 OU MAIS ANOS ESTUDO
V53	-2	11	-15	-7	V53 PESSOAS COM 8 OU MAIS ANOS ESTUDO
V22	11	19	17	29	V22 NUMERO DE LEITOS
V48	10	10	20	13	V48 NUMERO DE TELEFONES
V43	-5	-6	9	-3	V43 MORADIAS COM AGUA ENCANADA
V62	-32	-2	-11	37	V62 POPULACAO TOTAL
V11	22	28	-21	-27	V11 OBITOS FETAIS
V44	1	23	-28	20	V44 MORADIAS COM POCO
V3	81 *	-9	0	11	V3 APOSENTADORIAS PREMATURAS
V2	73 *	8	10	-16	V2 TRABALHADORES EM LICENCA
V18	2	81 *	7	-4	V18 MORTAL. ACIDENTES DE TRAFEGO
V6	-8	58 *	-19	16	V6 MORTALIDADE 5-19 ANOS
V1	4	4	47	-1	V1 NASCIDOS VIVOS

ANALISE FATORIAL DOS DADOS PARA O ANO DE 1980
DADOS IMPUTADOS PELO METODO EM

ROTATION METHOD: QUARTIMAX

ROTATED FACTOR PATTERN

	FACTOR1	FACTOR2	FACTOR3	FACTOR4	FACTOR5	
V8	94 *	8	-6	4	-14	V8 MORTALIDADE 50-65 ANOS
V17	89 *	-6	-3	-1	24	V17 MORTAL. DOENCAS NEOPLASICAS
V31	88 *	-16	12	8	26	V31 POP. 50-65 ANOS
V13	87 *	13	10	-23	6	V13 MORTAL. DOENCAS CARDIO-VASCULARES
V32	77 *	-1	9	18	44	V32 POP. 65 ANOS OU MAIS
V7	71 *	-4	-7	-36	-12	V7 MORTALIDADE 20-49 ANOS
V42	57 *	-34	-3	10	-8	V42 TOTAL DE MORADIAS
V9	51 *	45	-3	-1	39	V9 MORTALIDADE ACIMA DE 65 ANOS
V19	45	12	4	-29	-34	V19 MORTAL. CAUSAS VIOLENTAS
V29	-69 *	63 *	-2	21	-13	V29 POP. 5-19 ANOS
V28	-81 *	33	-7	2	-34	V28 POP. 1-4 ANOS
V27	-83 *	26	-13	-4	-26	V27 POP. MENORES DE 1 ANO
V5	-13	92 *	7	-7	3	V5 MORTALIDADE 1-4 ANOS
V20	-22	82 *	6	-37	4	V20 MORTAL. DOENCAS DIARREICAS
V21	-20	79 *	9	-42	3	V21 MORTAL. DOENCAS DIAR. 0-1 ANO
V10	0	66 *	-3	-47	15	V10 MORTALIDADE MENORES DE 1 ANO
V16	-6	64 *	22	-51 *	0	V16 MORTAL.DOENCAS RESP. 1-4 ANOS
V12	22	62 *	-10	-10	-7	V12 MORTAL. DOENCAS INFEC. E PARASIT.
V6	34	61 *	-9	3	7	V6 MORTALIDADE 5-19 ANOS
V4	43	61 *	-16	-14	43	V4 MORTALIDADE GERAL
V61	-25	-48	-21	5	0	V61 NUMERO EMPREGOS REGULARES
V41	57 *	-66 *	0	-16	2	V41 MORADIAS COM ESGOTO
V52	51 *	-72 *	4	-13	32	V52 ALFABETIZADOS COM MAIS DE 5 ANOS
V30	45	-77 *	-3	-34	7	V30 POP. 20-49 ANOS
V37	4	3	99 *	-6	-5	V37 POP. RURAL 50-65 ANOS
V35	3	5	99 *	-10	-1	V35 POP. RURAL 5-19 ANOS
V38	8	1	98 *	-7	-8	V38 POP. RURAL 65 ANOS OU MAIS
V36	8	0	98 *	-11	-3	V36 POP. RURAL 20-49 ANOS
V34	-1	5	98 *	-16	0	V34 POP. RURAL 1-4 ANOS
V33	2	4	97 *	-10	-2	V33 POP. RURAL MENORES DE 1 ANO
V56	-5	-5	-11	91 *	-18	V56 TOTAL TERRA CULTIVADA
V55	-9	4	-5	90 *	-19	V55 TOTAL TERRA ARAVEL
V49	-4	-9	-10	73 *	4	V49 AREA DO MUNICIPIO
V59	1	-30	-7	58 *	3	V59 CULTURAS RESP. 80% PROD.
V60	14	0	-32	54 *	-15	V60 NUMERO ESTAB. RURAIS
V14	40	15	9	-70 *	-25	V14 MORTAL.DOENCAS RESPIRAT.
V15	21	8	12	-79 *	-32	V15 MORTAL. DOENCAS RESP. 0-1 ANO
V54	25	9	0	7	92 *	V54 PESSOAS COM 11 OU MAIS ANOS ESTUDO
V53	31	9	2	3	89 *	V53 PESSOAS COM 8 OU MAIS ANOS ESTUDO
V48	22	-46	-29	-16	61 *	V48 NUMERO DE TELEFONES
V22	39	19	2	14	59 *	V22 NUMERO DE LEITOS
V43	42	-38	-2	-20	54 *	V43 MORADIAS COM AGUA ENCANADA
V62	18	16	-7	-37	46	V62 POPULACAO TOTAL
V11	23	14	-14	-18	36	V11 OBITOS FETAIS
V44	-15	17	14	28	-59 *	V44 MORADIAS COM POCO
V3	-8	-4	-16	-33	7	V3 APOSENTADORIAS PREMATURAS
V2	19	-26	-8	-19	-6	V2 TRABALHADORES EM LICENCA
V18	-14	12	-8	-2	17	V18 MORTAL. ACIDENTES DE TRAFEGO
V1	17	2	-4	15	1	V1 NASCIDOS VIVOS

ANALISE FATORIAL DOS DADOS PARA O ANO DE 1980
DADOS IMPUTADOS PELO METODO EM

ROTATION METHOD: QUARTIMAX

ROTATED FACTOR PATTERN

	FACTOR6	FACTOR7	FACTOR8	FACTOR9	
V8	4	-1	-1	-3	V8 MORTALIDADE 50-65 ANOS
V17	1	11	9	-1	V17 MORTAL. DOENCAS NEOPLASICAS
V31	-6	-23	12	-1	V31 POP. 50-65 ANOS
V13	-9	4	-5	-9	V13 MORTAL. DOENCAS CARDIO-VASCULARES
V32	-10	-34	19	-5	V32 POP. 65 ANOS OU MAIS
V7	-15	34	-2	16	V7 MORTALIDADE 20-49 ANOS
V42	25	-7	28	6	V42 TOTAL DE MORADIAS
V9	4	11	2	-23	V9 MORTALIDADE ACIMA DE 65 ANOS
V19	-24	41	17	-5	V19 MORTAL. CAUSAS VIOLENTAS
V29	0	6	-16	-11	V29 POP. 5-19 ANOS
V28	-12	12	11	-4	V28 POP. 1-4 ANOS
V27	-12	15	16	3	V27 POP. MENORES DE 1 ANO
V5	-8	11	-4	-5	V5 MORTALIDADE 1-4 ANOS
V20	3	7	19	13	V20 MORTAL. DOENCAS DIARREICAS
V21	5	4	18	14	V21 MORTAL. DOENCAS DIAR. 0-1 ANO
V10	36	9	7	8	V10 MORTALIDADE MENORES DE 1 ANO
V16	-20	5	-30	-2	V16 MORTAL.DOENCAS RESP. 1-4 ANOS
V12	-23	5	-26	32	V12 MORTAL. DOENCAS INFEC. E PARASIT.
V6	-8	54 *	-19	12	V6 MORTALIDADE 5-19 ANOS
V4	22	11	12	-2	V4 MORTALIDADE GERAL
V61	10	34	-7	24	V61 NUMERO EMPREGOS REGULARES
V41	10	8	5	12	V41 MORADIAS COM ESGOTO
V52	-3	19	-12	0	V52 ALFABETIZADOS COM MAIS DE 5 ANOS
V30	15	10	-4	16	V30 POP. 20-49 ANOS
V37	-2	-3	-1	-1	V37 POP. RURAL 50-65 ANOS
V35	-4	-2	-1	1	V35 POP. RURAL 5-19 ANOS
V38	1	-5	-1	-3	V38 POP. RURAL 65 ANOS OU MAIS
V36	-5	-4	-1	-1	V36 POP. RURAL 20-49 ANOS
V34	-2	-1	1	3	V34 POP. RURAL 1-4 ANOS
V33	-7	2	-2	1	V33 POP. RURAL MENORES DE 1 ANO
V56	-7	6	7	-3	V56 TOTAL TERRA CULTIVADA
V55	-9	11	2	-1	V55 TOTAL TERRA ARAVEL
V49	-36	11	-1	18	V49 AREA DO MUNICIPIO
V59	-6	-12	-13	22	V59 CULTURAS RESP. 80% PROD.
V60	-10	-9	-3	-5	V60 NUMERO ESTAB. RURAIS
V14	-13	9	-22	14	V14 MORTAL.DOENCAS RESPIRAT.
V15	-6	10	-19	10	V15 MORTAL. DOENCAS RESP. 0-1 ANO
V54	-3	4	-8	4	V54 PESSOAS COM 11 OU MAIS ANOS ESTUDO
V53	-3	11	-16	-8	V53 PESSOAS COM 8 OU MAIS ANOS ESTUDO
V48	7	13	17	14	V48 NUMERO DE TELEFONES
V22	11	19	15	28	V22 NUMERO DE LEITOS
V43	-7	-4	7	-2	V43 MORADIAS COM AGUA ENCANADA
V62	-35	-3	-12	35	V62 POPULACAO TOTAL
V11	21	26	-21	-28	V11 OBITOS FETAIS
V44	2	22	-27	19	V44 MORADIAS COM POCO
V3	80 *	-9	0	13	V3 APOSENTADORIAS PREMATURAS
V2	72 *	9	9	-14	V2 TRABALHADORES EM LICENCA
V18	2	81 *	7	-5	V18 MORTAL. ACIDENTES DE TRAFEGO
V1	5	4	46	-1	V1 NASCIDOS VIVOS

ANALISE FATORIAL DOS DADOS PARA O ANO DE 1980
DADOS IMPUTADOS PELO METODO DAS MÉEDIAS

ROTATION METHOD: VARIMAX

ROTATED FACTOR PATTERN

	FACTOR1	FACTOR2	FACTOR3	FACTOR4	FACTOR5	
V8	93 *	-2	-12	1	-8	V8 MORTALIDADE 50-65 ANOS
V13	87 *	2	-12	-28	8	V13 MORTAL. DOENCAS CARDIO-VASCULARES
V31	87 *	-25	-6	3	28	V31 POP. 50-65 ANOS
V17	84 *	-19	-12	-4	29	V17 MORTAL. DOENCAS NEOPLASICAS
V32	78 *	-8	-5	14	44	V32 POP. 65 ANOS OU MAIS
V7	63 *	-18	-24	-37	-7	V7 MORTALIDADE 20-49 ANOS
V9	55 *	39	-3	-2	42	V9 MORTALIDADE ACIMA DE 65 ANOS
V42	55 *	-36	-5	10	-8	V42 TOTAL DE MORADIAS
V28	-75 *	43	13	5	-36	V28 POP. 1-4 ANOS
V27	-80 *	35	7	1	-27	V27 POP. MENORES DE 1 ANO
V5	-5	92 *	-4	-11	7	V5 MORTALIDADE 1-4 ANOS
V20	-18	80 *	-3	-42	10	V20 MORTAL. DOENCAS DIARREICAS
V21	-16	78 *	-2	-47	9	V21 MORTAL. DOENCAS DIAR. 0-1 ANO
V29	-59 *	70 *	14	20	-14	V29 POP. 5-19 ANOS
V16	-2	61 *	-9	-57 *	2	V16 MORTAL.DOENCAS RESP. 1-4 ANOS
V10	0	61 *	-7	-52 *	21	V10 MORTALIDADE MENORES DE 1 ANO
V12	24	57 *	-21	-14	2	V12 MORTAL. DOENCAS INFEC. E PARASIT.
V4	43	53 *	-17	-14	47	V4 MORTALIDADE GERAL
V61	-40	-53 *	-15	12	1	V61 NUMERO EMPREGOS REGULARES
V41	46	-75 *	-10	-17	3	V41 MORADIAS COM ESGOTO
V52	39	-81 *	-11	-10	27	V52 ALFABETIZADOS COM MAIS DE 5 ANOS
V30	31	-84 *	-17	-30	6	V30 POP. 20-49 ANOS
V36	-6	-1	97 *	9	-8	V36 POP. RURAL 20-49 ANOS
V35	-15	6	96 *	9	-9	V35 POP. RURAL 5-19 ANOS
V37	-7	1	96 *	17	-16	V37 POP. RURAL 50-65 ANOS
V34	-16	7	93 *	9	-2	V34 POP. RURAL 1-4 ANOS
V38	3	-4	92 *	17	-23	V38 POP. RURAL 65 ANOS OU MAIS
V33	-17	7	91 *	4	-1	V33 POP. RURAL MENORES DE 1 ANO
V55	-3	8	28	84 *	-15	V55 TOTAL TERRA ARAVEL
V49	-1	-7	14	73 *	8	V49 AREA DO MUNICIPIO
V56	-7	-13	31	69 *	-41	V56 TOTAL TERRA CULTIVADA
V59	1	-28	2	60 *	7	V59 CULTURAS RESP. 80% PROD.
V60	12	8	-28	59 *	-45	V60 NUMERO ESTAB. RURAIS
V14	36	6	-22	-75 *	-21	V14 MORTAL.DOENCAS RESPIRAT.
V15	17	0	-9	-85 *	-27	V15 MORTAL. DOENCAS RESP. 0-1 ANO
V54	22	2	-13	8	92 *	V54 PESSOAS COM 11 OU MAIS ANOS ESTUDO
V53	28	2	-14	5	86 *	V53 PESSOAS COM 8 OU MAIS ANOS ESTUDO
V22	36	11	-7	11	67 *	V22 NUMERO DE LEITOS
V48	1	-54 *	-29	-8	60 *	V48 NUMERO DE TELEFONES
V62	11	8	-31	-38	53 *	V62 POPULACAO TOTAL
V43	34	-47	-30	-11	47	V43 MORADIAS COM AGUA ENCANADA
V18	-20	6	5	-1	20	V18 MORTAL. ACIDENTES DE TRAFEGO
V6	34	52 *	-13	0	15	V6 MORTALIDADE 5-19 ANOS
V19	44	4	-6	-32	-34	V19 MORTAL. CAUSAS VIOLENTAS
V11	20	8	-14	-12	33	V11 OBITOS FETAIS
V3	-6	-3	-11	-14	11	V3 APOSENTADORIAS PREMATURAS
V2	20	-29	-13	-2	-3	V2 TRABALHADORES EM LICENCA
V1	17	3	5	13	2	V1 NASCIDOS VIVOS
V44	-9	17	29	14	-50	V44 MORADIAS COM POCO

ANALISE FATORIAL DOS DADOS PARA O ANO DE 1980
DADOS IMPUTADOS PELO METODO DAS MÉEDIAS

ROTATION METHOD: VARIMAX

ROTATED FACTOR PATTERN

	FACTOR6	FACTOR7	FACTOR8	FACTOR9	
V8	7	6	4	10	V8 MORTALIDADE 50-65 ANOS
V13	12	-8	-7	-8	V13 MORTAL. DOENCAS CARDIO-VASCULARES
V31	-20	-3	10	-6	V31 POP. 50-65 ANOS
V17	16	2	11	-1	V17 MORTAL. DOENCAS NEOPLASICAS
V32	-29	-6	14	-15	V32 POP. 65 ANOS OU MAIS
V7	38	-16	6	20	V7 MORTALIDADE 20-49 ANOS
V9	23	5	-1	-22	V9 MORTALIDADE ACIMA DE 65 ANOS
V42	-10	30	33	7	V42 TOTAL DE MORADIAS
V28	11	-14	7	-2	V28 POP. 1-4 ANOS
V27	14	-14	14	-1	V27 POP. MENORES DE 1 ANO
V5	16	-6	-5	5	V5 MORTALIDADE 1-4 ANOS
V20	10	-1	17	3	V20 MORTAL. DOENCAS DIARREICAS
V21	6	1	16	4	V21 MORTAL. DOENCAS DIAR. 0-1 ANO
V29	9	-2	-16	2	V29 POP. 5-19 ANOS
V16	8	-19	-29	1	V16 MORTAL.DOENCAS RESP. 1-4 ANOS
V10	13	36	7	5	V10 MORTALIDADE MENORES DE 1 ANO
V12	10	-21	-11	48	V12 MORTAL. DOENCAS INFEC. E PARASIT.
V4	21	25	12	-6	V4 MORTALIDADE GERAL
V61	34	9	5	30	V61 NUMERO EMPREGOS REGULARES
V41	5	11	10	7	V41 MORADIAS COM ESGOTO
V52	12	-1	-10	-5	V52 ALFABETIZADOS COM MAIS DE 5 ANOS
V30	5	15	1	8	V30 POP. 20-49 ANOS
V36	-9	-4	1	-1	V36 POP. RURAL 20-49 ANOS
V35	-1	-5	-1	0	V35 POP. RURAL 5-19 ANOS
V37	-5	-2	-2	0	V37 POP. RURAL 50-65 ANOS
V34	4	-10	6	2	V34 POP. RURAL 1-4 ANOS
V38	-9	6	-3	-2	V38 POP. RURAL 65 ANOS OU MAIS
V33	9	-12	7	6	V33 POP. RURAL MENORES DE 1 ANO
V55	12	-12	2	13	V55 TOTAL TERRA ARAVEL
V49	11	-40	3	18	V49 AREA DO MUNICIPIO
V56	8	-4	21	-1	V56 TOTAL TERRA CULTIVADA
V59	-21	-9	-10	36	V59 CULTURAS RESP. 80% PROD.
V60	-8	-1	-9	-17	V60 NUMERO ESTAB. RURAIS
V14	12	-14	-13	15	V14 MORTAL.DOENCAS RESPIRAT.
V15	11	-9	-11	11	V15 MORTAL. DOENCAS RESP. 0-1 ANO
V54	3	2	-8	-6	V54 PESSOAS COM 11 OU MAIS ANOS ESTUDO
V53	10	3	-18	-13	V53 PESSOAS COM 8 OU MAIS ANOS ESTUDO
V22	18	8	22	17	V22 NUMERO DE LEITOS
V48	10	16	22	-4	V48 NUMERO DE TELEFONES
V62	-4	-36	-4	20	V62 POPULACAO TOTAL
V43	1	-4	7	-31	V43 MORADIAS COM AGUA ENCANADA
V18	85 *	3	4	-1	V18 MORTAL. ACIDENTES DE TRAFEGO
V6	61 *	-7	-8	28	V6 MORTALIDADE 5-19 ANOS
V19	49	-20	18	1	V19 MORTAL. CAUSAS VIOLENTAS
V11	41	30	-37	-31	V11 OBITOS FETAIS
V3	-11	85 *	1	9	V3 APOSENTADORIAS PREMATURAS
V2	16	76 *	9	-18	V2 TRABALHADORES EM LICENCA
V1	6	7	78 *	-10	V1 NASCIDOS VIVOS
V44	17	1	-19	56 *	V44 MORADIAS COM POCO

ANALISE FATORIAL DOS DADOS PARA O ANO DE 1980
DADOS IMPUTADOS PELO METODO DAS MÉEDIAS

ROTATION METHOD: QUARTIMAX

ROTATED FACTOR PATTERN

	FACTOR1	FACTOR2	FACTOR3	FACTOR4	FACTOR5	
V8	92 *	3	-11	2	-17	V8 MORTALIDADE 50-65 ANOS
V31	89 *	-21	-5	4	21	V31 POP. 50-65 ANOS
V17	88 *	-13	-11	-3	21	V17 MORTAL. DOENCAS NEOPLASICAS
V13	88 *	8	-11	-26	0	V13 MORTAL. DOENCAS CARDIO-VASCULARES
V32	80 *	-5	-4	15	38	V32 POP. 65 ANOS OU MAIS
V7	65 *	-10	-23	-36	-14	V7 MORTALIDADE 20-49 ANOS
V9	57 *	43	-3	1	36	V9 MORTALIDADE ACIMA DE 65 ANOS
V42	56 *	-36	-4	10	-12	V42 TOTAL DE MORADIAS
V28	-80 *	39	12	5	-30	V28 POP. 1-4 ANOS
V27	-83 *	32	6	1	-21	V27 POP. MENORES DE 1 ANO
V5	-8	93 *	-4	-8	5	V5 MORTALIDADE 1-4 ANOS
V20	-19	81 *	-3	-39	9	V20 MORTAL. DOENCAS DIARREICAS
V21	-18	79 *	-2	-44	8	V21 MORTAL. DOENCAS DIAR. 0-1 ANO
V29	-64 *	67 *	13	21	-10	V29 POP. 5-19 ANOS
V16	-4	64 *	-9	-55 *	0	V16 MORTAL.DOENCAS RESP. 1-4 ANOS
V10	0	64 *	-7	-49	19	V10 MORTALIDADE MENORES DE 1 ANO
V12	22	61 *	-21	-10	-4	V12 MORTAL. DOENCAS INFEC. E PARASIT.
V6	34	58 *	-13	2	8	V6 MORTALIDADE 5-19 ANOS
V4	46	57 *	-17	-11	42	V4 MORTALIDADE GERAL
V61	-35	-52 *	-16	10	4	V61 NUMERO EMPREGOS REGULARES
V41	51 *	-71 *	-10	-18	1	V41 MORADIAS COM ESGOTO
V52	46	-77 *	-11	-12	25	V52 ALFABETIZADOS COM MAIS DE 5 ANOS
V30	36	-80 *	-17	-33	5	V30 POP. 20-49 ANOS
V36	-8	-2	97 *	8	-7	V36 POP. RURAL 20-49 ANOS
V35	-18	5	96 *	9	-7	V35 POP. RURAL 5-19 ANOS
V37	-10	-1	96 *	16	-15	V37 POP. RURAL 50-65 ANOS
V34	-18	6	93 *	8	0	V34 POP. RURAL 1-4 ANOS
V38	-1	-6	92 *	17	-22	V38 POP. RURAL 65 ANOS OU MAIS
V33	-19	7	91 *	3	0	V33 POP. RURAL MENORES DE 1 ANO
V55	-6	5	29	84 *	-15	V55 TOTAL TERRA ARAVEL
V49	-2	-8	15	73 *	7	V49 AREA DO MUNICIPIO
V56	-11	-17	32	68 *	-39	V56 TOTAL TERRA CULTIVADA
V59	1	-30	2	60 *	6	V59 CULTURAS RESP. 80% PROD.
V60	6	4	-27	59 *	-45	V60 NUMERO ESTAB. RURAIS
V14	35	11	-21	-73 *	-25	V14 MORTAL.DOENCAS RESPIRAT.
V15	16	4	-9	-84 *	-29	V15 MORTAL. DOENCAS RESP. 0-1 ANO
V54	30	6	-14	8	89 *	V54 PESSOAS COM 11 OU MAIS ANOS ESTUDO
V53	35	6	-14	6	84 *	V53 PESSOAS COM 8 OU MAIS ANOS ESTUDO
V22	42	15	-7	13	62 *	V22 NUMERO DE LEITOS
V48	10	-51 *	-30	-10	61 *	V48 NUMERO DE TELEFONES
V62	16	13	-31	-37	51 *	V62 POPULACAO TOTAL
V43	41	-44	-30	-12	46	V43 MORADIAS COM AGUA ENCANADA
V18	-15	10	4	-2	20	V18 MORTAL. ACIDENTES DE TRAFEGO
V19	43	9	-5	-31	-39	V19 MORTAL. CAUSAS VIOLENTAS
V11	24	12	-14	-13	31	V11 OBITOS FETAIS
V3	-3	-4	-11	-15	12	V3 APOSENTADORIAS PREMATURAS
V2	23	-29	-13	-3	-3	V2 TRABALHADORES EM LICENCA
V1	18	2	6	14	1	V1 NASCIDOS VIVOS
V44	-15	17	29	14	-52 *	V44 MORADIAS COM POCO

ANALISE FATORIAL DOS DADOS PARA O ANO DE 1980
DADOS IMPUTADOS PELO METODO DAS MÉEDIAS

ROTATION METHOD: QUARTIMAX

ROTATED FACTOR PATTERN

	FACTOR6	FACTOR7	FACTOR8	FACTOR9	
V8	4	5	2	8	V8 MORTALIDADE 50-65 ANOS
V31	-22	-4	8	-5	V31 POP. 50-65 ANOS
V17	14	0	8	-1	V17 MORTAL. DOENCAS NEOPLASICAS
V13	7	-9	-8	-9	V13 MORTAL. DOENCAS CARDIO-VASCULARES
V32	-31	-7	12	-12	V32 POP. 65 ANOS OU MAIS
V7	37	-18	4	18	V7 MORTALIDADE 20-49 ANOS
V9	18	5	-1	-23	V9 MORTALIDADE ACIMA DE 65 ANOS
V42	-10	29	31	9	V42 TOTAL DE MORADIAS
V28	11	-12	9	-5	V28 POP. 1-4 ANOS
V27	15	-13	16	-3	V27 POP. MENORES DE 1 ANO
V5	11	-5	-3	2	V5 MORTALIDADE 1-4 ANOS
V20	7	0	18	2	V20 MORTAL. DOENCAS DIARREICAS
V21	3	2	17	3	V21 MORTAL. DOENCAS DIAR. 0-1 ANO
V29	7	0	-14	-1	V29 POP. 5-19 ANOS
V16	4	-18	-28	-1	V16 MORTAL.DOENCAS RESP. 1-4 ANOS
V10	10	36	8	4	V10 MORTALIDADE MENORES DE 1 ANO
V12	7	-21	-12	45	V12 MORTAL. DOENCAS INFEC. E PARASIT.
V6	58 *	-6	-9	24	V6 MORTALIDADE 5-19 ANOS
V4	17	25	11	-7	V4 MORTALIDADE GERAL
V61	40	8	5	31	V61 NUMERO EMPREGOS REGULARES
V41	8	9	8	9	V41 MORADIAS COM ESGOTO
V52	15	-3	-11	-3	V52 ALFABETIZADOS COM MAIS DE 5 ANOS
V30	8	13	0	10	V30 POP. 20-49 ANOS
V36	-8	-3	1	-1	V36 POP. RURAL 20-49 ANOS
V35	0	-4	-1	0	V35 POP. RURAL 5-19 ANOS
V37	-5	-1	-2	0	V37 POP. RURAL 50-65 ANOS
V34	5	-9	6	2	V34 POP. RURAL 1-4 ANOS
V38	-8	6	-3	-3	V38 POP. RURAL 65 ANOS OU MAIS
V33	10	-11	7	6	V33 POP. RURAL MENORES DE 1 ANO
V55	13	-11	1	11	V55 TOTAL TERRA ARAVEL
V49	13	-39	2	17	V49 AREA DO MUNICIPIO
V56	10	-3	20	-3	V56 TOTAL TERRA CULTIVADA
V59	-17	-9	-12	37	V59 CULTURAS RESP. 80% PROD.
V60	-9	-1	-9	-20	V60 NUMERO ESTAB. RURAIS
V14	9	-15	-13	14	V14 MORTAL.DOENCAS RESPIRAT.
V15	10	-10	-11	10	V15 MORTAL. DOENCAS RESP. 0-1 ANO
V54	3	1	-9	-3	V54 PESSOAS COM 11 OU MAIS ANOS ESTUDO
V53	10	2	-18	-10	V53 PESSOAS COM 8 OU MAIS ANOS ESTUDO
V22	17	8	21	19	V22 NUMERO DE LEITOS
V48	14	14	21	0	V48 NUMERO DE TELEFONES
V62	-4	-37	-4	22	V62 POPULACAO TOTAL
V43	2	-6	6	-28	V43 MORADIAS COM AGUA ENCANADA
V18	85 *	3	4	-3	V18 MORTAL. ACIDENTES DE TRAFEGO
V19	47	-21	18	-2	V19 MORTAL. CAUSAS VIOLENTAS
V11	39	30	-37	-32	V11 OBITOS FETAIS
V3	-10	84 *	1	11	V3 APOSENTADORIAS PREMATURAS
V2	16	75 *	8	-17	V2 TRABALHADORES EM LICENCA
V1	5	8	78 *	-9	V1 NASCIDOS VIVOS
V44	18	2	-20	52 *	V44 MORADIAS COM POCO

ANALISE FATORIAL DOS DADOS PARA O ANO DE 1980
DADOS IMPUTADOS PELO METODO DAS MÉEDIAS

ROTATION METHOD: VARIMAX

ROTATED FACTOR PATTERN

	FACTOR1	FACTOR2	FACTOR3	FACTOR4	FACTOR5	
V8	93 *	-1	-12	-1	-7	V8 MORTALIDADE 50-65 ANOS
V31	87 *	-23	-6	-3	28	V31 POP. 50-65 ANOS
V13	86 *	2	-12	28	10	V13 MORTAL. DOENCAS CARDIO-VASCULARES
V17	83 *	-19	-12	4	29	V17 MORTAL. DOENCAS NEOPLASICAS
V32	78 *	-6	-5	-14	45	V32 POP. 65 ANOS OU MAIS
V7	62 *	-18	-24	36	-7	V7 MORTALIDADE 20-49 ANOS
V42	53 *	-35	-5	-11	-8	V42 TOTAL DE MORADIAS
V9	51 *	37	-4	4	43	V9 MORTALIDADE ACIMA DE 65 ANOS
V28	-75 *	43	13	-4	-37	V28 POP. 1-4 ANOS
V27	-79 *	35	7	0	-29	V27 POP. MENORES DE 1 ANO
V5	-6	91 *	-4	13	8	V5 MORTALIDADE 1-4 ANOS
V20	-18	79 *	-3	44	8	V20 MORTAL. DOENCAS DIARREICAS
V21	-17	76 *	-2	49	8	V21 MORTAL. DOENCAS DIAR. 0-1 ANO
V29	-59 *	71 *	14	-19	-13	V29 POP. 5-19 ANOS
V16	-2	59 *	-9	58 *	4	V16 MORTAL.DOENCAS RESP. 1-4 ANOS
V10	-1	59 *	-7	53 *	21	V10 MORTALIDADE MENORES DE 1 ANO
V12	23	55 *	-21	14	2	V12 MORTAL. DOENCAS INFEC. E PARASIT.
V4	41	51 *	-17	16	48	V4 MORTALIDADE GERAL
V43	33	-46	-29	11	46	V43 MORADIAS COM AGUA ENCANADA
V61	-35	-49	-14	-13	-3	V61 NUMERO EMPREGOS REGULARES
V41	46	-74 *	-11	15	2	V41 MORADIAS COM ESGOTO
V52	38	-82 *	-11	9	28	V52 ALFABETIZADOS COM MAIS DE 5 ANOS
V30	31	-85 *	-17	29	5	V30 POP. 20-49 ANOS
V35	-15	7	97 *	-9	-9	V35 POP. RURAL 5-19 ANOS
V36	-6	0	96 *	-9	-8	V36 POP. RURAL 20-49 ANOS
V37	-6	2	96 *	-17	-16	V37 POP. RURAL 50-65 ANOS
V34	-16	7	93 *	-8	-2	V34 POP. RURAL 1-4 ANOS
V38	3	-3	92 *	-18	-23	V38 POP. RURAL 65 ANOS OU MAIS
V33	-17	7	90 *	-4	-2	V33 POP. RURAL MENORES DE 1 ANO
V15	18	-1	-9	84 *	-27	V15 MORTAL. DOENCAS RESP. 0-1 ANO
V14	36	5	-22	73 *	-20	V14 MORTAL.DOENCAS RESPIRAT.
V60	10	8	-23	-53 *	-38	V60 NUMERO ESTAB. RURAIS
V59	1	-26	2	-57 *	4	V59 CULTURAS RESP. 80% PROD.
V56	-7	-11	31	-68 *	-40	V56 TOTAL TERRA CULTIVADA
V49	-2	-6	14	-71 *	7	V49 AREA DO MUNICIPIO
V55	-3	9	28	-83 *	-14	V55 TOTAL TERRA ARAVEL
V54	19	1	-13	-6	94 *	V54 PESSOAS COM 11 OU MAIS ANOS ESTUDO
V53	25	1	-14	-3	90 *	V53 PESSOAS COM 8 OU MAIS ANOS ESTUDO
V22	34	9	-8	-10	63 *	V22 NUMERO DE LEITOS
V48	0	-54 *	-29	8	57 *	V48 NUMERO DE TELEFONES
V62	11	7	-31	36	48	V62 POPULACAO TOTAL
V11	17	7	-14	16	36	V11 OBITOS FETAIS
V18	-20	5	5	2	20	V18 MORTAL. ACIDENTES DE TRAFEGO
V6	32	50 *	-13	1	17	V6 MORTALIDADE 5-19 ANOS
V19	43	4	-5	30	-31	V19 MORTAL. CAUSAS VIOLENTAS
V2	20	-29	-12	2	-1	V2 TRABALHADORES EM LICENCA
V3	-5	-4	-11	15	10	V3 APOSENTADORIAS PREMATURAS
V1	17	1	6	-14	1	V1 NASCIDOS VIVOS
V44	-9	17	28	-15	-48	V44 MORADIAS COM POCO

ANALISE FATORIAL DOS DADOS PARA O ANO DE 1980
DADOS IMPUTADOS PELO METODO DAS MÉEDIAS

ROTATION METHOD: VARIMAX

ROTATED FACTOR PATTERN

	FACTOR6	FACTOR7	FACTOR8	FACTOR9	
V8	9	6	3	7	V8 MORTALIDADE 50-65 ANOS
V31	-22	-3	10	-6	V31 POP. 50-65 ANOS
V13	13	-8	-6	-12	V13 MORTAL. DOENCAS CARDIO-VASCULARES
V17	17	2	13	-2	V17 MORTAL. DOENCAS NEOPLASICAS
V32	-31	-6	16	-13	V32 POP. 65 ANOS OU MAIS
V7	39	-17	7	16	V7 MORTALIDADE 20-49 ANOS
V42	-9	25	29	6	V42 TOTAL DE MORADIAS
V9	22	7	2	-20	V9 MORTALIDADE ACIMA DE 65 ANOS
V28	12	-15	6	-6	V28 POP. 1-4 ANOS
V27	15	-15	13	-5	V27 POP. MENORES DE 1 ANO
V5	18	-6	-6	4	V5 MORTALIDADE 1-4 ANOS
V20	11	-3	22	4	V20 MORTAL. DOENCAS DIARREICAS
V21	7	-2	21	6	V21 MORTAL. DOENCAS DIAR. 0-1 ANO
V29	10	-1	-19	1	V29 POP. 5-19 ANOS
V16	9	-18	-33	-1	V16 MORTAL.DOENCAS RESP. 1-4 ANOS
V10	12	34	14	13	V10 MORTALIDADE MENORES DE 1 ANO
V12	14	-25	-11	36	V12 MORTAL. DOENCAS INFEC. E PARASIT.
V4	20	25	18	-2	V4 MORTALIDADE GERAL
V43	0	-3	14	-30	V43 MORADIAS COM AGUA ENCANADA
V61	29	6	4	21	V61 NUMERO EMPREGOS REGULARES
V41	4	11	12	9	V41 MORADIAS COM ESGOTO
V52	12	0	-10	-7	V52 ALFABETIZADOS COM MAIS DE 5 ANOS
V30	4	14	4	10	V30 POP. 20-49 ANOS
V35	-1	-5	-1	1	V35 POP. RURAL 5-19 ANOS
V36	-10	-4	0	0	V36 POP. RURAL 20-49 ANOS
V37	-6	-1	-3	1	V37 POP. RURAL 50-65 ANOS
V34	5	-12	7	1	V34 POP. RURAL 1-4 ANOS
V38	-9	7	-4	-3	V38 POP. RURAL 65 ANOS OU MAIS
V33	9	-14	8	5	V33 POP. RURAL MENORES DE 1 ANO
V15	12	-9	-15	10	V15 MORTAL. DOENCAS RESP. 0-1 ANO
V14	14	-15	-17	12	V14 MORTAL.DOENCAS RESPIRAT.
V60	-7	1	-9	-16	V60 NUMERO ESTAB. RURAIS
V59	-16	-12	-7	24	V59 CULTURAS RESP. 80% PROD.
V56	7	-4	20	-3	V56 TOTAL TERRA CULTIVADA
V49	11	-38	2	13	V49 AREA DO MUNICIPIO
V55	12	-12	1	10	V55 TOTAL TERRA ARAVEL
V54	2	2	-5	-2	V54 PESSOAS COM 11 OU MAIS ANOS ESTUDO
V53	10	4	-17	-11	V53 PESSOAS COM 8 OU MAIS ANOS ESTUDO
V22	17	5	24	17	V22 NUMERO DE LEITOS
V48	8	14	29	0	V48 NUMERO DE TELEFONES
V62	-2	-36	-2	15	V62 POPULACAO TOTAL
V11	30	29	-19	-17	V11 OBITOS FETAIS
V18	76 *	5	6	-1	V18 MORTAL. ACIDENTES DE TRAFEGO
V6	65 *	-7	-10	23	V6 MORTALIDADE 5-19 ANOS
V19	45	-17	14	-1	V19 MORTAL. CAUSAS VIOLENTAS
V2	13	73 *	9	-15	V2 TRABALHADORES EM LICENCA
V3	-9	71 *	6	9	V3 APOSENTADORIAS PREMATURAS
V1	4	9	45	-6	V1 NASCIDOS VIVOS
V44	18	-1	-26	50	V44 MORADIAS COM POCO

ANALISE FATORIAL DOS DADOS PARA O ANO DE 1980
DADOS IMPUTADOS PELO METODO DAS MÉEDIAS

ROTATION METHOD: QUARTIMAX

ROTATED FACTOR PATTERN

	FACTOR1	FACTOR2	FACTOR3	FACTOR4	FACTOR5	
V8	92 *	4	-11	-3	-15	V8 MORTALIDADE 50-65 ANOS
V31	90 *	-20	-5	-5	22	V31 POP. 50-65 ANOS
V17	87 *	-13	-11	3	22	V17 MORTAL. DOENCAS NEOPLASICAS
V13	87 *	8	-11	26	2	V13 MORTAL. DOENCAS CARDIO-VASCULARES
V32	80 *	-4	-4	-16	40	V32 POP. 65 ANOS OU MAIS
V7	65 *	-11	-23	35	-13	V7 MORTALIDADE 20-49 ANOS
V42	55 *	-34	-4	-11	-11	V42 TOTAL DE MORADIAS
V9	54 *	41	-3	2	38	V9 MORTALIDADE ACIMA DE 65 ANOS
V19	42	9	-5	29	-35	V19 MORTAL. CAUSAS VIOLENTAS
V28	-79 *	39	12	-5	-31	V28 POP. 1-4 ANOS
V27	-82 *	31	6	0	-23	V27 POP. MENORES DE 1 ANO
V5	-9	92 *	-4	10	6	V5 MORTALIDADE 1-4 ANOS
V20	-20	80 *	-3	41	8	V20 MORTAL. DOENCAS DIARREICAS
V21	-18	78 *	-2	46	7	V21 MORTAL. DOENCAS DIAR. 0-1 ANO
V29	-64 *	67 *	13	-20	-9	V29 POP. 5-19 ANOS
V16	-5	63 *	-9	56 *	3	V16 MORTAL.DOENCAS RESP. 1-4 ANOS
V10	0	62 *	-7	51 *	19	V10 MORTALIDADE MENORES DE 1 ANO
V12	21	59 *	-20	11	-3	V12 MORTAL. DOENCAS INFEC. E PARASIT.
V4	44	55 *	-17	14	43	V4 MORTALIDADE GERAL
V61	-31	-49	-15	-11	0	V61 NUMERO EMPREGOS REGULARES
V41	51 *	-70 *	-10	16	0	V41 MORADIAS COM ESGOTO
V52	45	-78 *	-10	12	26	V52 ALFABETIZADOS COM MAIS DE 5 ANOS
V30	37	-82 *	-17	32	4	V30 POP. 20-49 ANOS
V35	-18	5	97 *	-9	-7	V35 POP. RURAL 5-19 ANOS
V36	-8	-2	96 *	-9	-7	V36 POP. RURAL 20-49 ANOS
V37	-10	0	96 *	-17	-14	V37 POP. RURAL 50-65 ANOS
V34	-18	6	92 *	-8	-1	V34 POP. RURAL 1-4 ANOS
V38	-1	-6	92 *	-17	-22	V38 POP. RURAL 65 ANOS OU MAIS
V33	-19	7	89 *	-4	-1	V33 POP. RURAL MENORES DE 1 ANO
V15	17	3	-9	83 *	-29	V15 MORTAL. DOENCAS RESP. 0-1 ANO
V14	36	10	-21	72 *	-24	V14 MORTAL.DOENCAS RESPIRAT.
V60	5	4	-23	-53 *	-38	V60 NUMERO ESTAB. RURAIS
V59	0	-28	3	-56 *	4	V59 CULTURAS RESP. 80% PROD.
V56	-10	-15	31	-68 *	-39	V56 TOTAL TERRA CULTIVADA
V49	-2	-7	14	-71 *	6	V49 AREA DO MUNICIPIO
V55	-6	7	28	-83 *	-14	V55 TOTAL TERRA ARAVEL
V54	27	4	-13	-7	92 *	V54 PESSOAS COM 11 OU MAIS ANOS ESTUDO
V53	33	5	-14	-4	87 *	V53 PESSOAS COM 8 OU MAIS ANOS ESTUDO
V22	40	14	-7	-11	59 *	V22 NUMERO DE LEITOS
V48	10	-52 *	-30	9	58 *	V48 NUMERO DE TELEFONES
V62	15	12	-31	35	46	V62 POPULACAO TOTAL
V43	40	-44	-29	12	45	V43 MORADIAS COM AGUA ENCANADA
V11	21	10	-14	16	34	V11 OBITOS FETAIS
V44	-14	18	28	-15	-50	V44 MORADIAS COM POCO
V18	-15	9	4	3	20	V18 MORTAL. ACIDENTES DE TRAFEGO
V6	34	57 *	-13	-1	11	V6 MORTALIDADE 5-19 ANOS
V2	23	-29	-12	4	-1	V2 TRABALHADORES EM LICENCA
V3	-3	-5	-11	16	11	V3 APOSENTADORIAS PREMATURAS
V1	18	1	6	-15	0	V1 NASCIDOS VIVOS

ANALISE FATORIAL DOS DADOS PARA O ANO DE 1980
DADOS IMPUTADOS PELO METODO DAS MÉEDIAS

ROTATION METHOD: QUARTIMAX

ROTATED FACTOR PATTERN

	FACTOR6	FACTOR7	FACTOR8	FACTOR9	
V8	5	5	0	6	V8 MORTALIDADE 50-65 ANOS
V31	-24	-5	7	-4	V31 POP. 50-65 ANOS
V17	14	0	9	-2	V17 MORTAL. DOENCAS NEOPLASICAS
V13	8	-9	-8	-13	V13 MORTAL. DOENCAS CARDIO-VASCULARES
V32	-34	-7	13	-11	V32 POP. 65 ANOS OU MAIS
V7	37	-19	5	15	V7 MORTALIDADE 20-49 ANOS
V42	-8	25	26	7	V42 TOTAL DE MORADIAS
V9	17	7	0	-21	V9 MORTALIDADE ACIMA DE 65 ANOS
V19	42	-17	14	-4	V19 MORTAL. CAUSAS VIOLENTAS
V28	13	-13	10	-8	V28 POP. 1-4 ANOS
V27	16	-13	17	-6	V27 POP. MENORES DE 1 ANO
V5	13	-4	-5	1	V5 MORTALIDADE 1-4 ANOS
V20	7	-2	24	3	V20 MORTAL. DOENCAS DIARREICAS
V21	3	0	23	5	V21 MORTAL. DOENCAS DIAR. 0-1 ANO
V29	8	2	-16	-2	V29 POP. 5-19 ANOS
V16	5	-18	-31	-4	V16 MORTAL.DOENCAS RESP. 1-4 ANOS
V10	8	35	14	12	V10 MORTALIDADE MENORES DE 1 ANO
V12	11	-24	-12	34	V12 MORTAL. DOENCAS INFEC. E PARASIT.
V4	16	25	16	-2	V4 MORTALIDADE GERAL
V61	34	5	4	22	V61 NUMERO EMPREGOS REGULARES
V41	6	8	9	11	V41 MORADIAS COM ESGOTO
V52	15	-3	-13	-5	V52 ALFABETIZADOS COM MAIS DE 5 ANOS
V30	8	12	2	13	V30 POP. 20-49 ANOS
V35	0	-4	-1	0	V35 POP. RURAL 5-19 ANOS
V36	-9	-3	0	0	V36 POP. RURAL 20-49 ANOS
V37	-5	0	-3	0	V37 POP. RURAL 50-65 ANOS
V34	6	-11	7	1	V34 POP. RURAL 1-4 ANOS
V38	-8	7	-5	-4	V38 POP. RURAL 65 ANOS OU MAIS
V33	10	-13	8	5	V33 POP. RURAL MENORES DE 1 ANO
V15	10	-10	-14	8	V15 MORTAL. DOENCAS RESP. 0-1 ANO
V14	11	-16	-16	10	V14 MORTAL.DOENCAS RESPIRAT.
V60	-8	1	-9	-18	V60 NUMERO ESTAB. RURAIS
V59	-13	-12	-9	24	V59 CULTURAS RESP. 80% PROD.
V56	9	-3	19	-4	V56 TOTAL TERRA CULTIVADA
V49	13	-38	1	12	V49 AREA DO MUNICIPIO
V55	13	-11	0	8	V55 TOTAL TERRA ARAVEL
V54	2	0	-8	1	V54 PESSOAS COM 11 OU MAIS ANOS ESTUDO
V53	9	3	-19	-9	V53 PESSOAS COM 8 OU MAIS ANOS ESTUDO
V22	16	5	21	19	V22 NUMERO DE LEITOS
V48	11	12	28	4	V48 NUMERO DE TELEFONES
V62	-2	-37	-2	17	V62 POPULACAO TOTAL
V43	0	-4	13	-26	V43 MORADIAS COM AGUA ENCANADA
V11	29	28	-20	-17	V11 OBITOS FETAIS
V44	19	-1	-27	46	V44 MORADIAS COM POCO
V18	76 *	5	6	-3	V18 MORTAL. ACIDENTES DE TRAFEGO
V6	61 *	-7	-12	20	V6 MORTALIDADE 5-19 ANOS
V2	14	72 *	8	-14	V2 TRABALHADORES EM LICENCA
V3	-8	71 *	5	11	V3 APOSENTADORIAS PREMATURAS
V1	4	9	44	-5	V1 NASCIDOS VIVOS

ANEXO H

H.1 Escores fatoriais para o ano de 1980

H.2 Gráficos dos escores fatoriais

ESCORES FATORIAIS PARA O ANO DE 1980
MÉTODO DAS COMPONENTES PRINCIPAIS

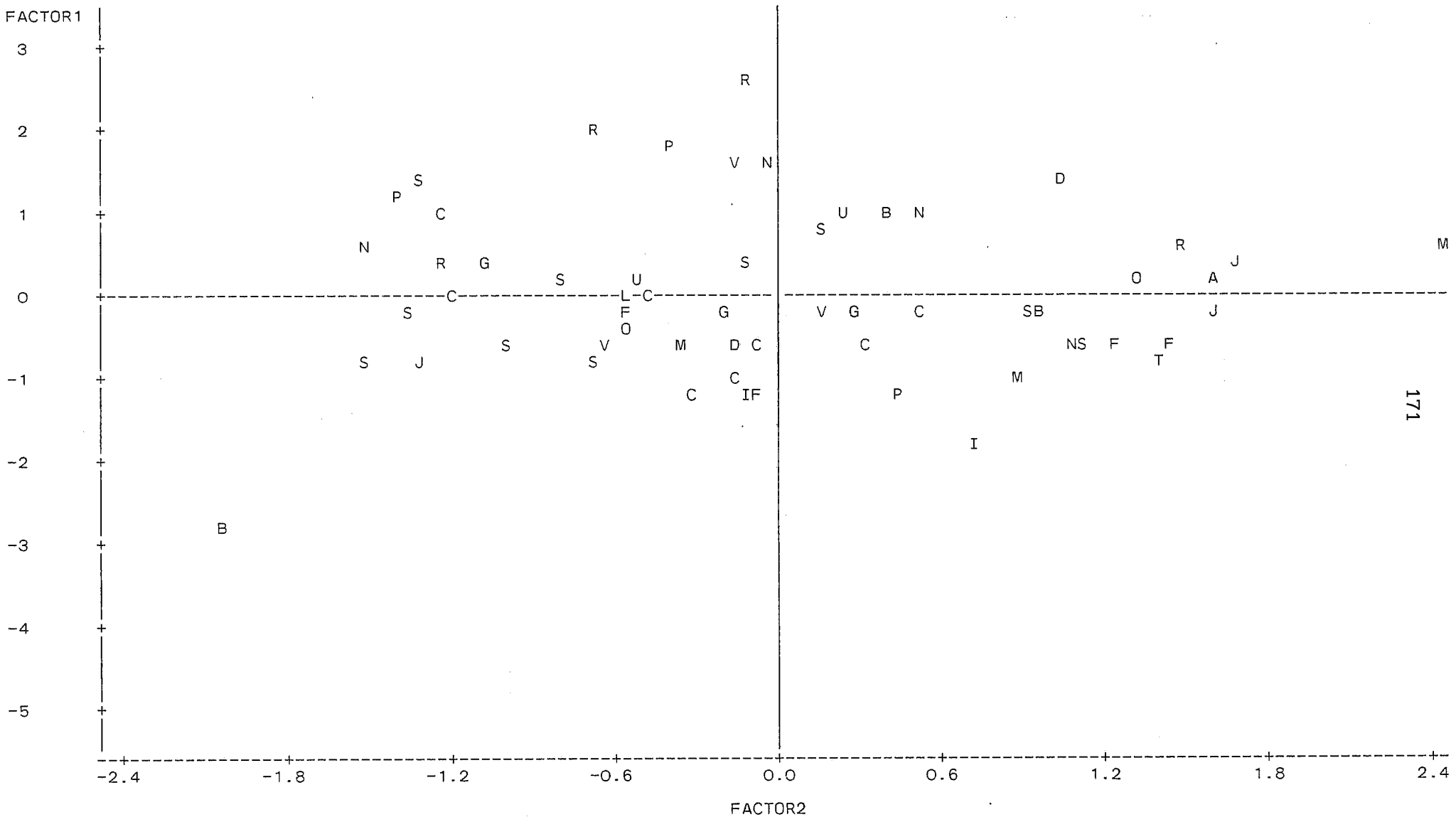
OBS	CIDADE	FACTOR1	FACTOR2	FACTOR3	FACTOR4	FACTOR5	FACTOR6	FACTOR7	FACTOR8	FACTOR9
1	SAO PAULO	0.2360	-0.8139	-0.7109	-1.4806	0.8923	-1.0242	-0.1836	-0.1388	1.1123
2	R DE JANEIRO	1.9089	-0.6680	0.3479	-0.3322	1.3272	-1.8697	0.6323	-0.2622	0.2562
3	B HORIZONTE	0.9411	0.3969	-1.0886	-1.2676	0.6035	-0.4280	1.1965	-0.4249	1.2084
4	SALVADOR	-0.2342	0.9336	-1.2477	-0.8998	0.4169	-1.1833	-1.3020	0.0282	1.6038
5	FORTALEZA	-0.5802	1.2567	0.3145	-0.2508	0.4228	0.5212	-0.1587	0.7699	1.0503
6	RECIFE	0.6350	1.4856	-0.7586	-0.9116	0.9460	-0.1343	-0.6508	-0.2006	-0.2475
7	BRASILIA	-2.8783	-2.0356	-0.2785	0.4301	1.0891	-0.8157	-1.9269	-0.7213	1.1115
8	P ALEGRE	1.1855	-1.3969	-0.6540	-0.0995	1.3106	0.0519	-0.3484	0.8853	0.7436
9	NOVA IGUACU	1.0849	0.5247	-1.2819	-0.5913	-1.7931	-2.0454	-0.5972	-2.0276	0.1939
10	CURITIBA	0.0421	-0.4840	0.5840	-0.1135	1.3598	-0.0229	0.7563	-0.4227	0.8288
11	BELEM	-0.2036	0.9560	-0.0299	-0.7736	0.6351	-1.3375	1.0264	-0.8995	-0.9734
12	GOIANIA	-0.1728	0.2648	-0.3713	0.9288	0.5173	1.0014	1.1135	-1.1808	2.6338
13	CAMPINAS	-0.0188	-1.2172	0.0549	0.1787	0.6294	-0.4584	-0.3498	0.6503	1.0065
14	MANAUS	-0.9169	0.8976	-0.4036	0.4513	0.3752	-1.4902	1.7317	0.5435	0.1165
15	S GONCALO	0.8059	0.1567	2.4233	-0.5761	-0.5675	-0.0305	0.3811	-0.8125	-0.7200
16	D DE CAXIAS	1.4897	1.0589	-0.5143	-0.6991	-1.6744	-0.1304	0.9829	-1.2091	0.5822
17	SANTO ANDRE	-0.1411	-1.3647	-0.2182	-1.1781	0.5868	0.6827	-1.1522	0.0756	-0.7703
18	GUARULHOS	-0.2647	-0.1912	0.0766	-1.0912	-0.6088	0.1527	0.7601	0.1758	1.2417
19	OSASCO	-0.3606	-0.5638	0.4437	-2.3987	-0.3101	0.3038	0.5419	0.8447	-0.3339
20	SAO LUIS	-0.6384	1.1138	0.5201	-0.0209	0.9198	-1.3027	-0.0414	-0.4568	-0.2817
21	S B DO CAMPO	-0.5303	-0.9904	-0.0599	-1.3153	0.2972	0.7768	0.1504	0.1468	0.0870
22	NATAL	-0.5287	1.0927	2.5123	-0.4374	0.8068	0.4093	-0.7745	0.3649	0.7295
23	SANTOS	1.3931	-1.3009	-1.1109	-0.8940	1.3849	-0.8810	-1.2416	2.4338	-1.6622
24	NITEROI	1.6729	-0.0258	3.6728	-0.1478	1.3869	-0.0011	0.6157	-0.1680	0.6924
25	MACEIO	0.5187	2.4260	-0.5995	0.0137	0.6538	1.3586	0.4673	0.9362	-0.6637
26	S J MERITI	0.4974	-0.1326	1.9256	-1.5177	-1.0998	-0.9430	-0.6479	0.0244	-0.4926
27	TERESINA	-0.7769	1.4023	-0.0317	0.6586	0.2259	0.8533	-0.2096	-1.3641	-0.4077
28	JABOATAO	-0.1831	1.6138	-0.1137	-0.0054	-0.8951	-0.6773	-0.9629	-0.7047	-1.0134
29	JOAO PESSOA	0.3366	1.6816	-0.5283	1.0791	1.0921	2.1699	-0.5106	2.5309	0.8572
30	RIB. PRETO	0.4369	-1.2299	-0.2689	0.9499	0.7349	-0.2717	-0.3958	0.3339	0.1911
31	LONDRINA	-0.0008	-0.5426	0.0957	1.2651	0.3439	0.1039	0.6419	-0.5644	0.0567
32	ARACAJU	0.2126	1.6002	-0.6990	-0.0673	0.4948	-0.1206	0.7088	1.4477	-2.3424
33	CAMPO GRANDE	-0.1489	0.5036	-0.2202	1.3767	0.1582	0.7595	0.8026	0.2798	1.2331
34	F DE SANTANA	-0.5881	1.4481	0.1736	1.0836	-0.3614	-0.3144	-2.0650	0.7353	-0.6226
35	S J CAMPOS	-0.8808	-1.5353	-0.2509	0.4635	0.4171	-0.4773	0.9401	0.5915	-1.4474
36	OLINDA	0.1653	1.3317	-0.3339	-0.3863	0.0284	0.4606	-1.5790	-1.0848	-0.8230
37	CONTAGEM	-0.6870	-0.0663	-0.8279	-0.4971	-0.8862	0.5923	0.3606	-0.6145	0.8419
38	PELOTAS	1.8568	-0.4154	0.2847	1.6811	-0.2582	0.6810	-1.7139	0.4966	0.7987
39	UBERLANDIA	0.1059	-0.5173	-0.3641	0.6107	0.0071	-0.6553	-0.0896	-0.4135	0.9391
40	JOINVILLE	-0.8297	-1.3022	0.0193	1.3741	0.0593	-0.0087	0.3289	-0.4743	-0.7665
41	DIADEMA	-0.6075	-0.1792	-2.6126	-2.2977	-1.2924	1.5308	0.8211	0.9385	0.6532
42	CANOAS	0.9047	-1.2578	-0.4543	-0.0683	-1.0439	0.0813	0.4033	-0.2429	-0.7758
43	IMPERATRIZ	-1.8143	0.7048	0.6008	1.0679	-1.3526	-2.3102	-0.7984	2.9199	1.4106
44	CUIABA	-0.5173	0.3033	-0.0973	1.5177	0.3039	-1.3797	0.7892	-0.9700	-0.5108
45	VITORIA	-0.2657	0.1717	-0.7506	-0.3397	1.9150	1.3554	0.7573	-0.1745	-1.3637
46	MAUA	-0.6898	-0.3746	0.6741	-0.9710	-0.8612	2.8279	-0.8018	-0.7434	-0.1537
47	VILA VELHA	-0.6182	-0.6418	-0.8980	0.2610	0.6004	-0.6923	-0.7704	-1.0740	-1.9109
48	UBERABA	1.0186	0.2278	-0.0994	1.3003	-0.0114	0.1979	-0.8779	-1.3117	1.4576
49	FLORIANOPOLS	-0.2129	-0.5570	0.4274	0.8973	1.8168	1.3896	0.6013	-1.5419	-1.8658
50	CARAPICUIBA	-1.1343	-0.3231	2.5580	-2.1832	-0.9908	0.4100	0.0393	0.4198	0.1502
51	PAULISTA	-1.1167	0.4458	0.4753	0.3089	-0.4759	0.3424	-1.8625	-0.4076	-0.7644
52	CASCAVEL	-1.0464	-0.1520	0.3530	1.3906	-0.8122	0.2105	2.0358	-0.2778	-0.2916
53	IPATINGA	-1.1232	-0.1125	-0.7455	-0.0074	-0.6323	0.5816	0.0689	-1.4452	-0.0028
54	RIO GRANDE	2.6358	-0.1250	-0.2298	0.8702	-0.9235	-0.3657	0.1481	0.8571	-0.6218

ESCORES FATORIAIS PARA O ANO DE 1980
MÉTODO DAS COMPONENTES PRINCIPAIS

OBS	CIDADE	FACTOR1	FACTOR2	FACTOR3	FACTOR4	FACTOR5	FACTOR6	FACTOR7	FACTOR8	FACTOR9
55	N HAMBURGO	0.6926	-1.5134	-0.36201	0.86906	-1.0299	1.58488	0.37490	0.07958	0.56400
56	F DO IGUACU	-1.2455	-0.0616	0.34163	0.50206	-1.0264	-0.83968	3.14089	1.58819	-0.50031
57	VIAMAO	1.6125	-0.1515	0.04721	0.95864	-2.2596	0.22060	-0.22639	-0.22035	-0.49890
58	GRAVATAI	0.3922	-1.0889	0.43782	1.06581	-1.9639	0.78323	-0.14739	1.42351	-0.73113
59	SUMARE	-0.8259	-0.6659	-0.14816	0.26572	-1.6285	-0.18363	-0.93288	0.03293	-0.79192

ANALISE FATORIAL DOS DADOS PARA O ANO DE 1980
GRAFICO DOS ESCORES FATORIAIS

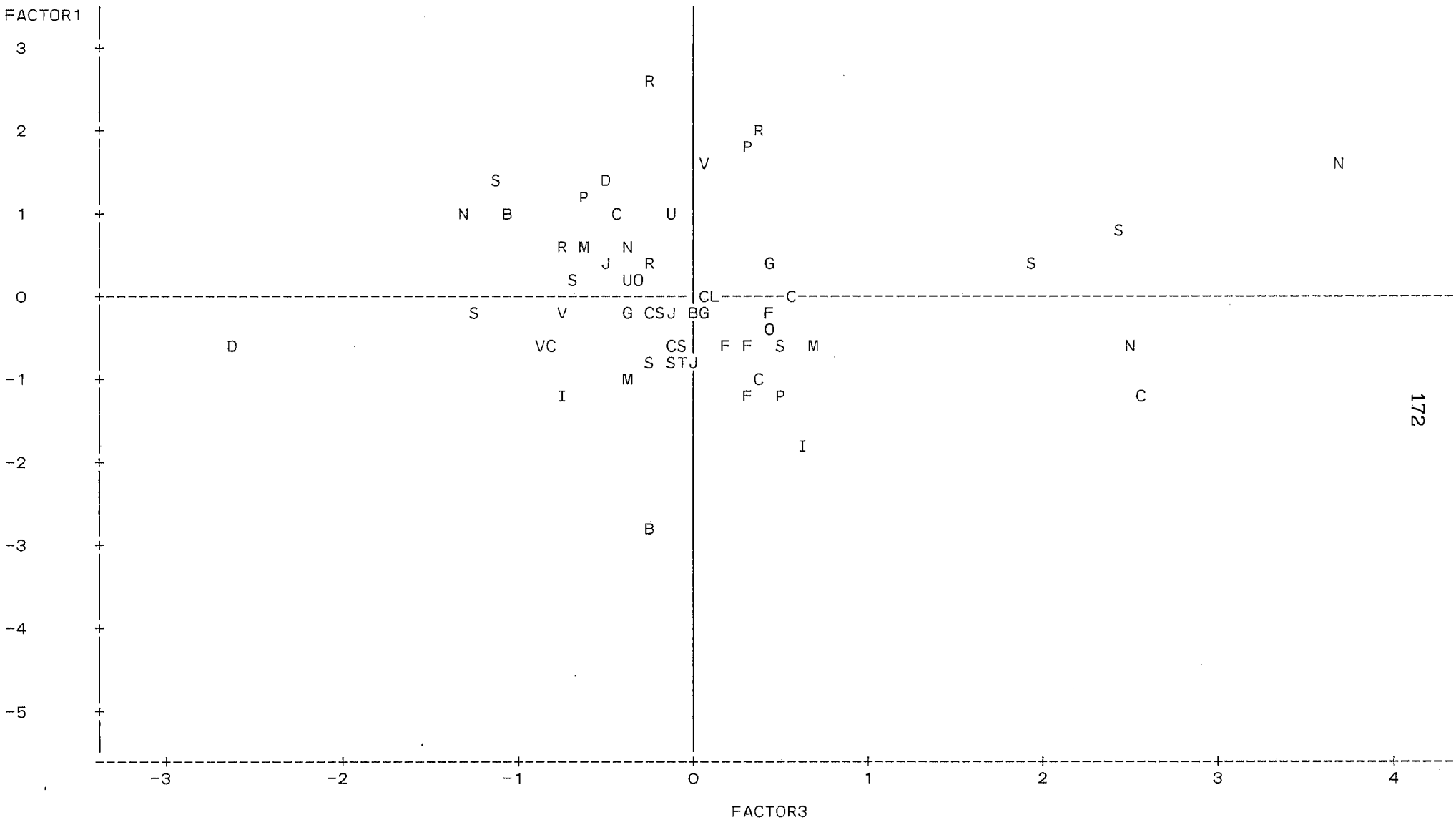
PLOT OF FACTOR1*FACTOR2 SYMBOL IS VALUE OF CIDADE



OBS: OS PONTOS SÃO MARCADOS PELA LETRA INICIAL DE CADA CIDADE

ANALISE FATORIAL DOS DADOS PARA O ANO DE 1980
GRAFICO DOS ESCORES FATORIAIS

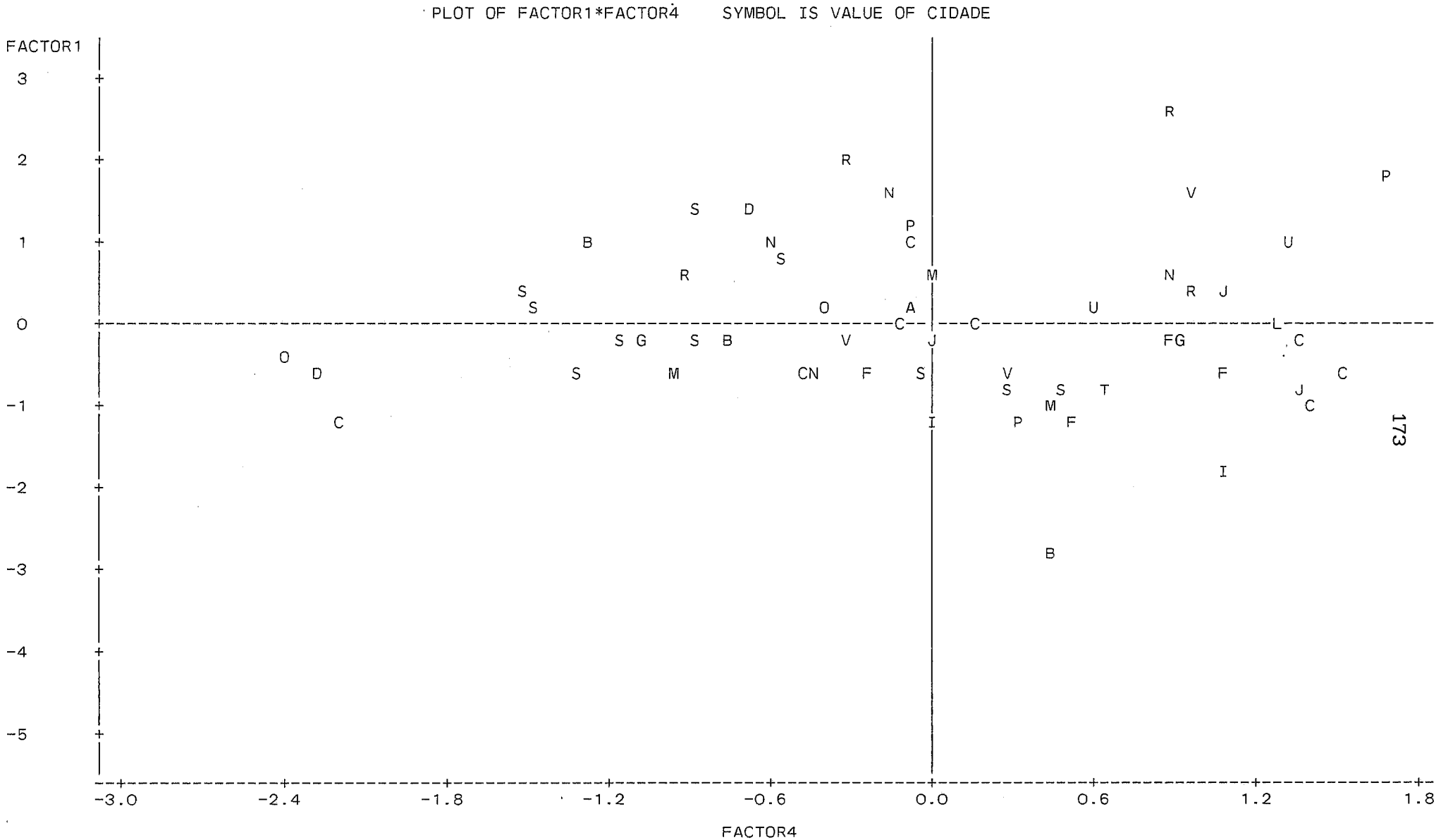
PLOT OF FACTOR1*FACTOR3 SYMBOL IS VALUE OF CIDADE



NOTE: 1 OBS HIDDEN

OBS: OS PONTOS SÃO MARCADOS PELA LETRA INICIAL DE CADA CIDADE

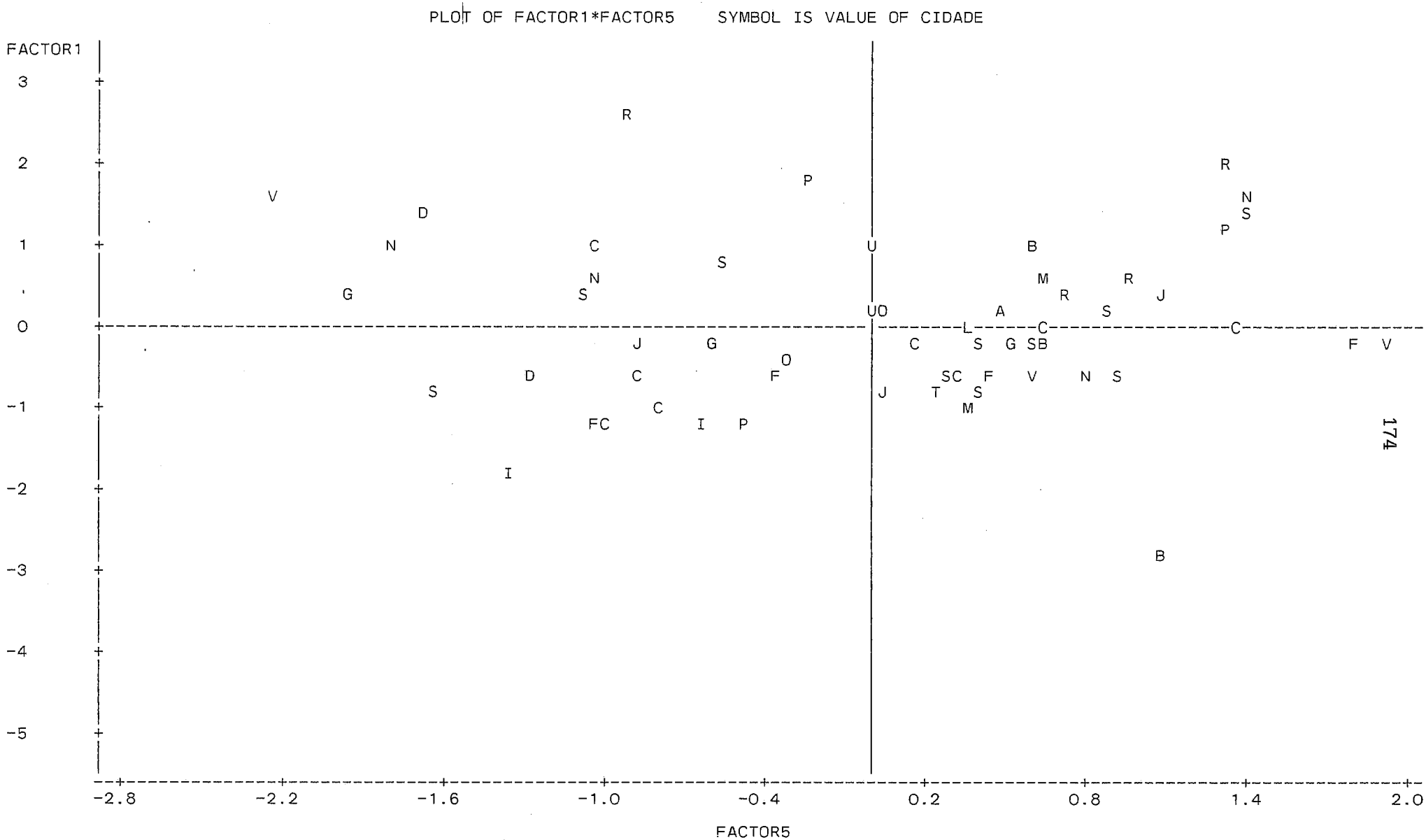
ANALISE FATORIAL DOS DADOS PARA O ANO DE 1980
GRAFICO DOS ESCORES FATORIAIS



NOTE: 1 OBS HIDDEN

OBS: OS PONTOS SÃO MARCADOS PELA LETRA INICIAL DE CADA CIDADE

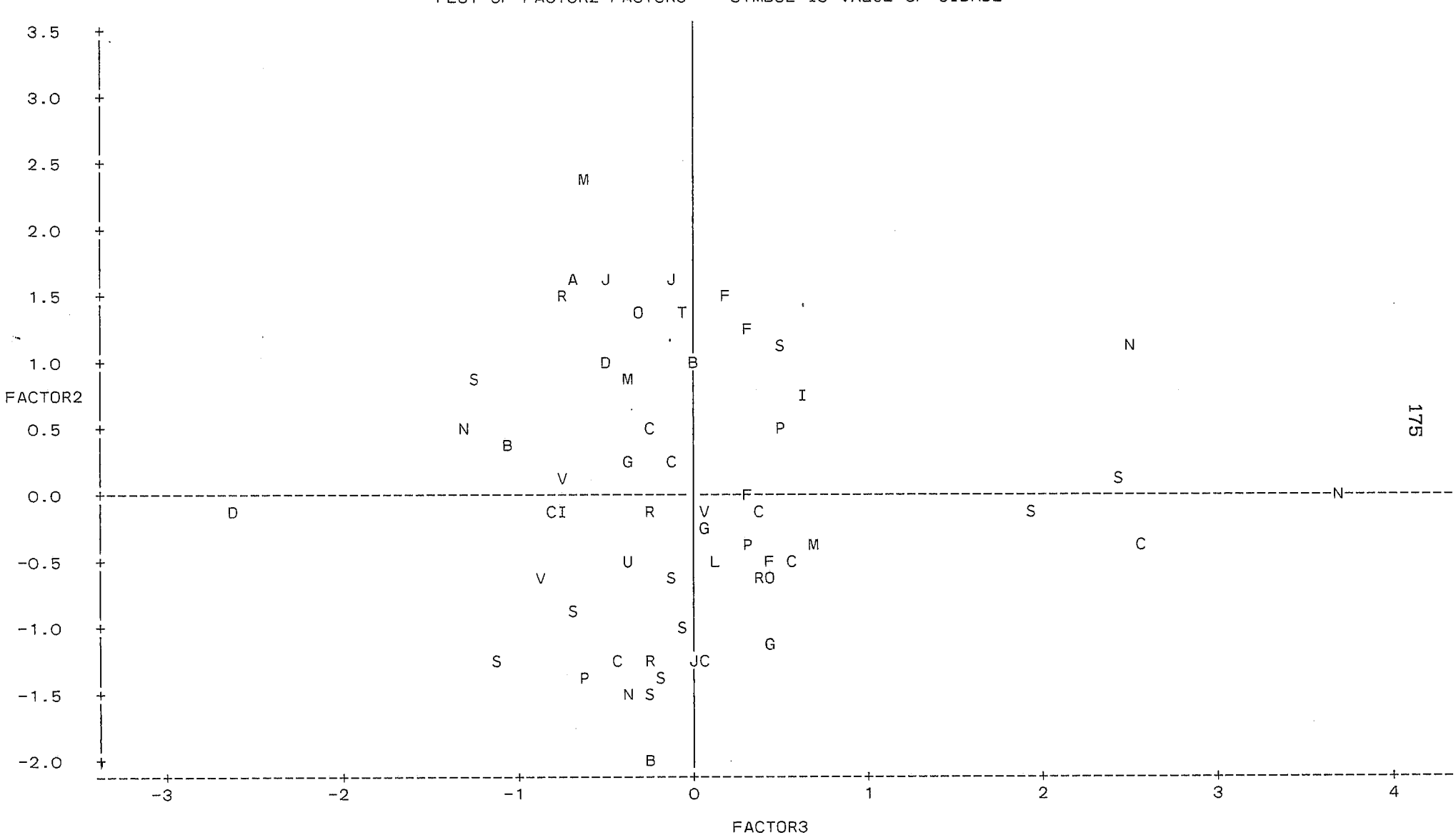
ANALISE FATORIAL DOS DADOS PARA O ANO DE 1980
GRAFICO DOS ESCORES FATORIAIS



NOTE: 1 OBS HIDDEN

OBS: OS PONTOS SÃO MARCADOS PELA LETRA INICIAL DE CADA CIDADE

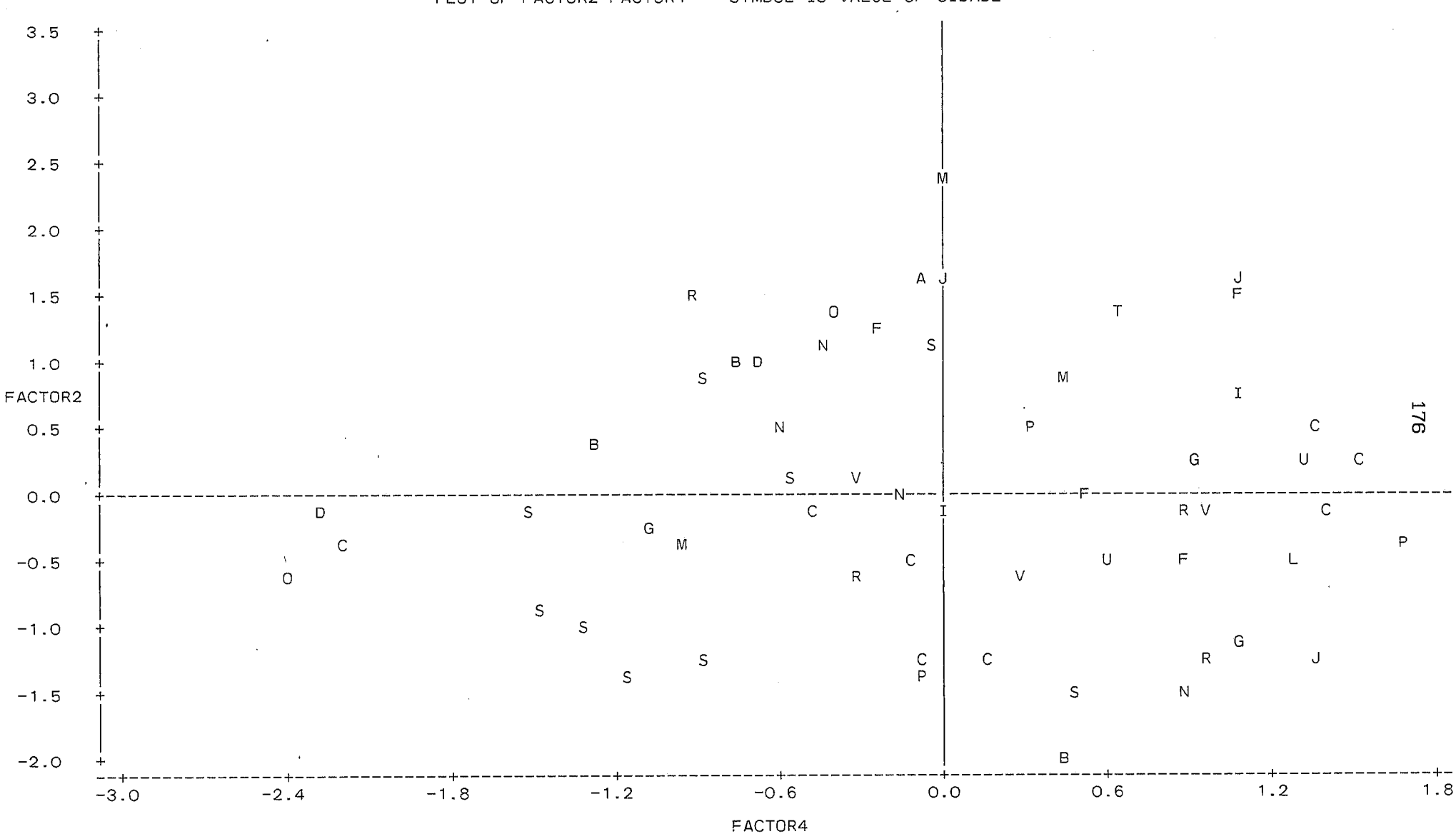
ANALISE FATORIAL DOS DADOS PARA O ANO DE 1980
GRAFICO DOS ESCORES FATORIAIS



NOTE: 1 OBS HIDDEN

OBS: OS PONTOS SÃO MARCADOS PELA LETRA INICIAL DE CADA CIDADE

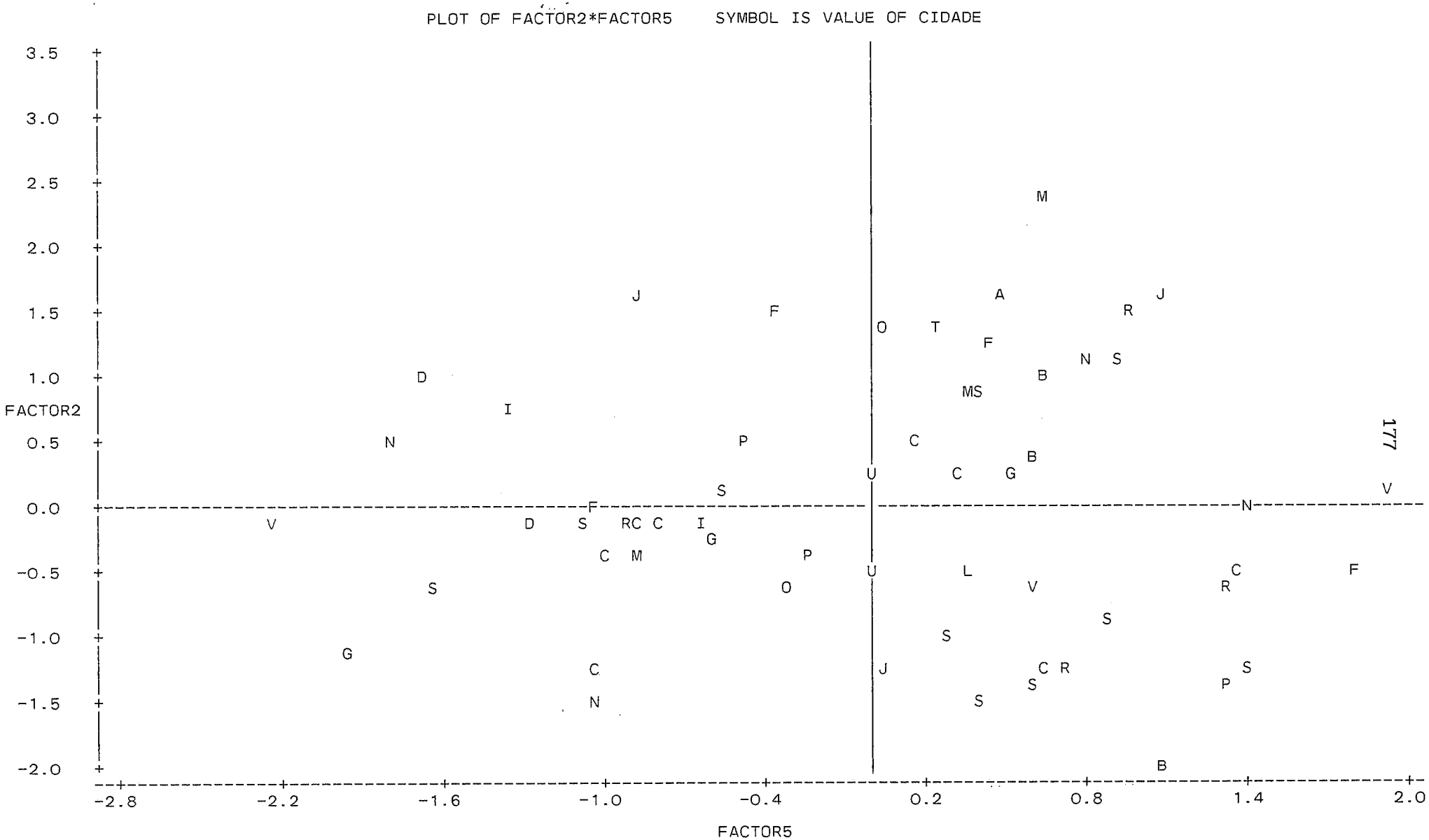
ANALISE FATORIAL DOS DADOS PARA O ANO DE 1980
GRAFICO DOS ESCORES FATORIAIS



NOTE: 1 OBS HIDDEN

OBS: OS PONTOS SÃO MARCADOS PELA LETRA INICIAL DE CADA CIDADE

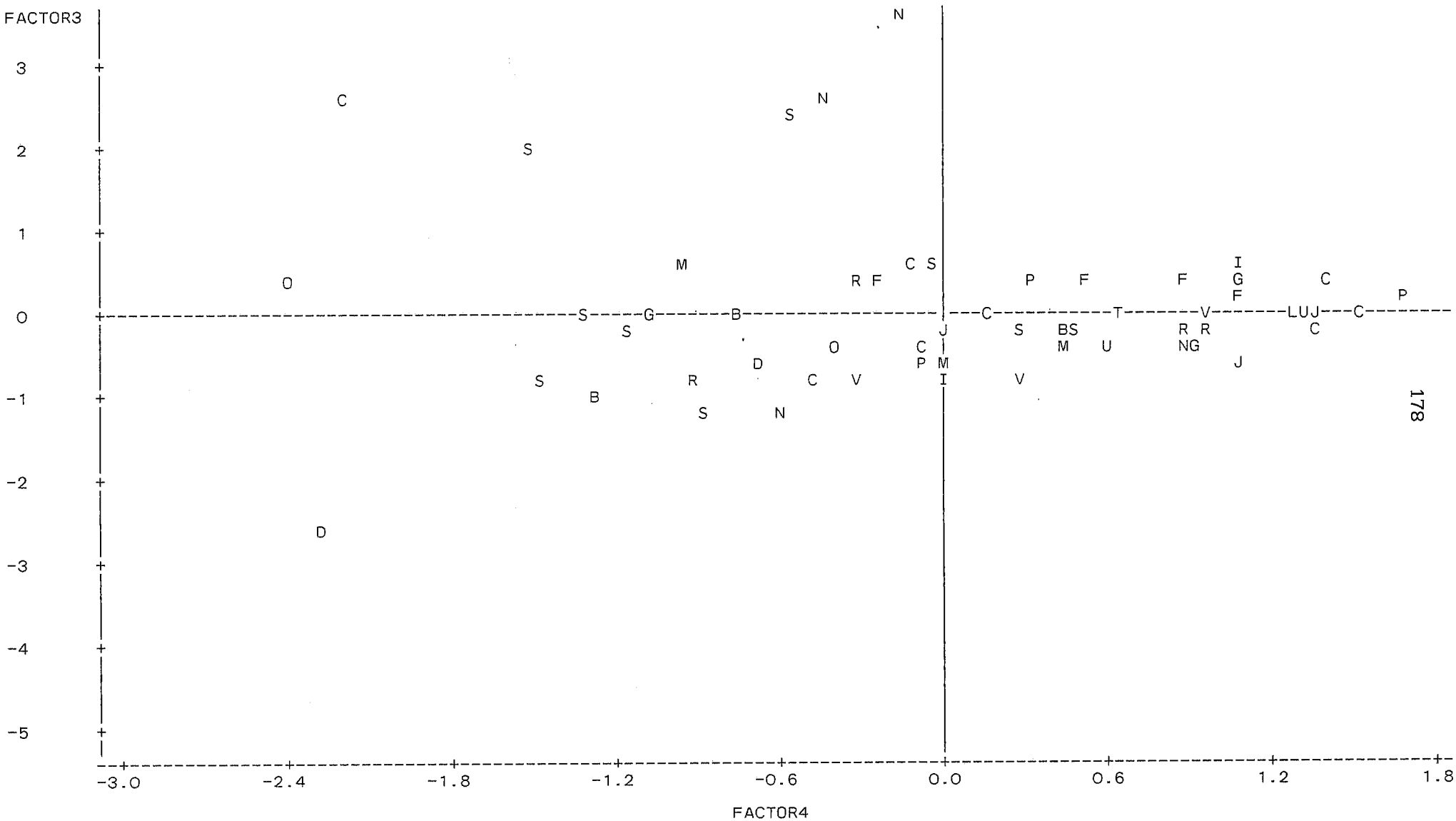
ANALISE FATORIAL DOS DADOS PARA O ANO DE 1980
GRAFICO DOS ESCORES FATORIAIS



OBS: OS PONTOS SÃO MARCADOS PELA LETRA INICIAL DE CADA CIDADE

ANALISE FATORIAL DOS DADOS PARA O ANO DE 1980
GRAFICO DOS ESCORES FATORIAIS

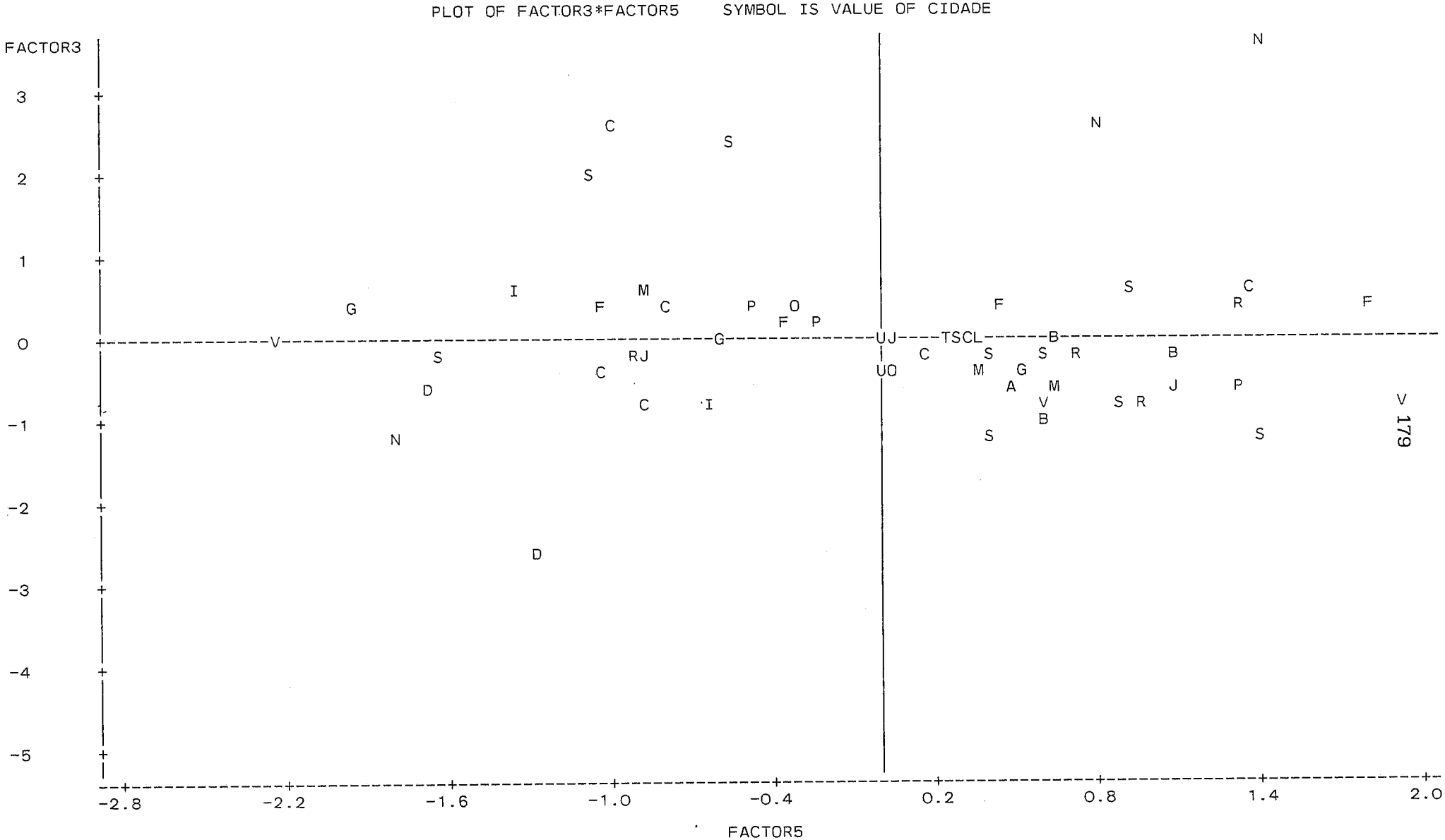
PLOT OF FACTOR3*FACTOR4 SYMBOL IS VALUE OF CIDADE



NOTE: 2 OBS HIDDEN

OBS: OS PONTOS SÃO MARCADOS PELA LETRA INICIAL DE CADA CIDADE

ANALISE FATORIAL DOS DADOS PARA O ANO DE 1980
GRAFICO DOS ESCORES FATORIAIS

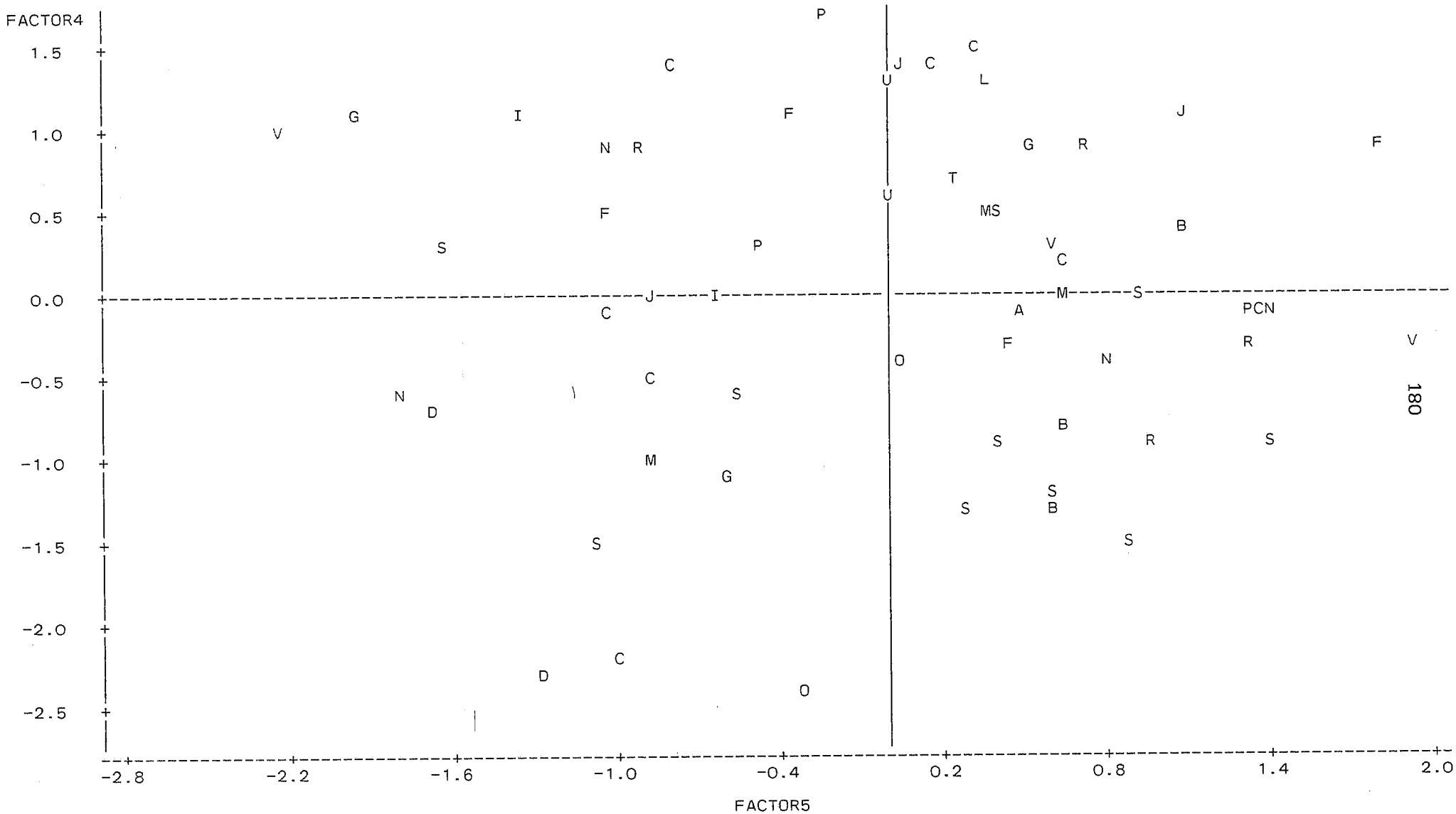


NOTE: 2 OBS HIDDEN

OBS: OS PONTOS SÃO MARCADOS PELA LETRA INICIAL DE CADA CIDADE

ANALISE FATORIAL DOS DADOS PARA O ANO DE 1980
GRAFICO DOS ESCORES FATORIAIS

PLOT OF FACTOR4*FACTOR5 SYMBOL IS VALUE OF CIDADE



OBS: OS PONTOS SÃO MARCADOS PELA LETRA INICIAL DE CADA CIDADE