UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMATICA
DEPARTAMENTO DE ESTATISTICA

# ESCALONAMENTO MULTIDIMENSIONAL

ADRIANA DEMARCHI LAUTERT

ORIENTAÇÃO : JANDYRA MARIA GUIMARÃES FACHEL

Monografia apresentada
para obtenção do Grau
de Bacharel em Estatística





Porto Alegre, julho de 1990.

## AGRADECIMENTOS

Agradeço, em especial, à Jandyra M. G. Fachel pela orientação dada, a Marcos L. Reina por sua dedicação e amizade, e à Jaqueline Chies pela cessão dos dados utilizados no trabalho.

Agradeço, também, a todos os amigos do Departamento de Estatística, em especial, a Ana Lúcia e Vânia, pela atenção dedicada.

ÍNDICE


| | | |
|---|---|---|
| 1 | CAPÍTULO 1 - INTRODUÇÃO ..... | 5 |
| 1.1 | Apresentação do Problema ..... | 5 |
| 1.2 | Nota Histórica ..... | 7 |
| 1.3 | Descrição do Trabalho ..... | 10 |
| 2 | CAPÍTULO 2 - MEDIDAS DE PROXIMIDADE ..... | 12 |
| 2.1 | Introdução ..... | 12 |
| 2.2 | Propriedades das Medidas de Proximidade ..... | 14 |
| 2.3 | Medidas de Proximidade para Dados Quantitativos ... | 16 |
| 2.4 | Medidas de Proximidade para Dados Qualitativos .... | 18 |
| 2.5 | Medidas de Proximidade para Dados Mistos ..... | 21 |
| 2.6 | Outros Procedimentos para Obtenção de Medidas de Proximidade ..... | 22 |
| 3 | CAPÍTULO 3 - PRINCIPAIS MÉTODOS DE ESCALONAMENTO MULTIDIMENSIONAL ..... | 26 |
| 3.1 | Introdução ..... | 26 |
| 3.2 | Método Métrico ..... | 27 |
| 3.2.1 | Solução Clássica ..... | 31 |
| 3.2.2 | Solução para uma Matriz de Dissimilaridade ou Similaridade ..... | 34 |
| 3.2.3 | Adequação de Ajuste ..... | 35 |
| 3.3 | Método Não-Métrico ..... | 36 |
| 3.3.1 | Adequação de Ajuste ..... | 39 |
| 3.3.2 | Obtenção da Solução ..... | 42 |
| 3.3.3 | Diagrama de Dispersão ..... | 45 |

| | | |
|---|---|---|
| 3.4 | Análise da Solução | 49 |
| 3.4.1 | Escolha do Número de Dimensões | 49 |
| 3.4.2 | Interpretação da Configuração | 52 |
| 3.5 | Comparação dos Métodos | 55 |
| 4 | CAPÍTULO 4 - APLICAÇÃO | 57 |
| 5 | CAPÍTULO 5 - NTSYS-pc | 77 |
| 5.1 | Introdução | 77 |
| 5.2 | Modos de Operação | 78 |
| 5.3 | Carregando o NTSYS-pc | 78 |
| 5.4 | Preparação do Arquivo de Dados | 79 |
| 5.5 | Descrição dos Programas | 81 |
| 6 | ANEXO | 91 |
| 7 | REFERÊNCIAS BIBLIOGRÁFICAS | 107 |
| 8 | BIBLIOGRAFIA | 111 |

# CAPITULO 1

# INTRODUÇÃO

## 1.1 - APRESENTAÇÃO DO PROBLEMA

Dadas as coordenadas de uma série de pontos (ou indivíduos ou objetos) é fácil calcular a distância entre cada par de pontos; mas dadas as distâncias entre cada par de pontos, achar as coordenadas dos pontos é um problema não muito fácil de se resolver.

Uma série de procedimentos, conhecidos como técnicas de Escalonamento Multidimensional, permitem a resolução desse tipo de problema. Mais especificamente, as técnicas de Escalonamento Multidimensional, tratam da representação de dados combinando uma série de pontos em um espaço geométrico e definindo algumas funções entre eles até refletir as relações que existem entre os dados ou que são supostas para gerar os dados.

Nas mais diversas áreas de pesquisa, como por exemplo, Psicologia, Sociologia, Antropologia, Educação, Economia, e outras, os dados a serem analisados podem consistir tanto de uma

matriz de dados (n x p) como de uma matriz (n x n) de medidas de "proximidade" entre pares de objetos ou entre pares de indivíduos. Usaremos o termo "proximidade" para denotar genericamente as medidas de distância, dissimilaridade ou similaridade.

Os procedimentos de Escalonamento Multidimensional, ou EMD, iniciam com uma matriz de proximidades ou com informações a respeito das medidas de proximidade, como por exemplo, a ordenação por postos destas medidas.

Quando os dados encontram-se sob a forma de uma matriz de proximidades esta pode originar-se ou diretamente, de experimentos nos quais indivíduos são questionados para avaliar a similaridade ou dissimilaridade de dois objetos, ou indiretamente, como uma medida derivada de uma matriz de dados (n x p) que compara cada par de objetos.

A partir da matriz de proximidades, ou das informações acerca destas medidas, o EMD tenta encontrar uma representação espacial de pontos, que possa refletir o modelo (ou a estrutura) presente na matriz de proximidades, com o objetivo de tornar os dados mais compreensíveis.

Essa representação espacial consiste de uma configuração geométrica de pontos (por configuração entende-se uma série de valores de coordenadas) em um espaço Euclidiano k-dimensional.

Para reproduzir as proximidades entre n pontos exatamente, pode-se precisar de até (n-1) dimensões, mas o propósito do uso do EMD é ver se existe uma configuração em um pequeno número de dimensões, a qual reproduza aproximadamente as proximidades observadas. O número de dimensões deve ser preferivelmente 2 ou 3. Configurações em 2 dimensões apresentam a vantagem de serem facilmente representadas graficamente e interpretadas. Configurações em 3 dimensões pode ser representadas graficamente, mas não tão facilmente interpretadas como as configurações em 2 dimensões.

Espera-se que haja, de alguma forma, máxima correspondência entre as medidas de proximidade observadas (ou informações a respeito destas) e as distâncias entre os pares de pontos da configuração obtida pelo EMD. Isto significa que a

maior das similaridades (ou a menor das dissimilaridades) entre dois objetos deve corresponder a menor distância entre dois pontos da configuração final, ou seja, os objetos que forem considerados muito similares (ou pouco dissimilares) estarão mais próximos um do outro, na configuração final, do que dos outros objetos.

Uma das formas mais importantes de análise dos resultados de EMD é feita observando simplesmente os pontos na configuração obtida, sendo que cada um deles deve indicar o objeto que ele representa.

## 1.2 - NOTA HISTÓRICA

As origens das aplicações do EMD encontram-se no campo da Psicologia, mas o EMD não se restringiu apenas às suas origens e atualmente é aplicado em várias áreas do conhecimento como Psicologia, Sociologia, Antropologia, Economia, Educação, e outras.

As bases teóricas do EMD foram lançadas por Young e Householder (1938) que demonstraram teoremas que fornecem as condições para se acomodar distâncias em um espaço Euclidiano e determinam a dimensão mínima desse espaço.

Richardson (1938) e Klingberg (1941) usaram esses conhecimentos nas primeiras aplicações conhecidas de EMD, no entanto a metodologia só começou a se desenvolver a partir dos anos 50 juntamente com o surgimento dos computadores que permitiram a Torgerson (1958), Messick e Abelson (1956), e outros, lidarem com um grande número de dados e com espaços de dimensão alta.

Foram desenvolvidos métodos de EMD a 'dois fatores' e a 'três fatores'. Os métodos de EMD a 'dois fatores' não consideravam as possíveis diferenças existentes entre os

julgamentos dos indivíduos que faziam parte do experimento, enquanto os métodos de EMD a 'três fatores' (ou Escalonamento de Diferenças Individuais) admitia a existência de diferenças entre os julgamentos dos indivíduos e as incorporava atribuindo pesos às dimensões.

Os métodos de EMD a 'dois fatores' foram classificados em métricos e não-métricos. Esta classificação foi introduzida por Coombs (1958), mas foi elaborada por Kruskal (1964) quando este chamou de EMD não-métrico o que Shepard (1962) havia chamado de 'Análise de Proximidades'.

O método de escalonamento métrico foi originalmente proposto por Torgerson (1958). Coombs (1958) desenvolveu seu método de EMD não-métrico baseado em dados ordinais de proximidades. Shepard (1962) desenvolveu sua técnica de EMD que consistia de um algoritmo para computador que permitia uma solução métrica a partir de dados não-métricos 'ordinais' e chamou de 'Análise de Proximidades'.

Kruskal (1964 a, b) construiu sua técnica de EMD não-métrico a partir da 'Análise de Proximidades' de Shepard. Ele aprimorou a técnica de Shepard centralizando-a em uma medida de 'Adequação de Ajuste' com a qual pretendia julgar qual configuração melhor se ajustava aos dados analisados. (MDSCALE, Kruskal 1964). O raciocínio básico para o algoritmo de EMD não-métrico dado por Kruskal (1964) tem formado a base para todos os trabalhos subsequentes nesta área. Kruskal (1964 b) sugere métodos para lidar com assimetrias, empates e observações perdidas, e também descreve os detalhes dos algoritmos desenvolvidos e implementados por ele para usar a técnica. Um exemplo de um modelo de EMD métrico (MDPREF) é dado por Carroll e Chang (1964).

Wagenaar e Padmos (1971) indicam que a interpretação da medida de 'Adequação de Ajuste' (STRESS) de Kruskal é fortemente dependente do número de objetos (ou indivíduos) envolvidos, e que uma simples interpretação em termos da avaliação verbal de Kruskal é muitas vezes não justificada. Spence (1970 e 1972) e Spence e Graef (1974) realizaram uma série de experimentos Monte Carlo e propuseram um método útil de obter uma melhor idéia da

verdadeira dimensionalidade de uma solução de EMD não-métrico, usando uma larga tabela de valores de "stress".

O primeiro método de EMD a 'três fatores' desenvolvido foi o EMD Tri-modal (Tucker 1964, 1972). O modelo Euclidiano ponderado, utilizado no EMD a 'três fatores', e os procedimentos associados a ele para ajustar modelos a dados empíricos foram propostos por várias pessoas aproximadamente ao mesmo tempo (Horam 1969, Bloxom 1968, Carroll e Chang 1970).

A complementação do EMD a 'três fatores' de maior sucesso é devido a Carroll e Chang (1970). Eles propõem um modelo para EMD, que admite diferenças individuais, conhecido como INDSCAL, o qual é aplicado quando existe no mínimo duas matrizes de proximidades, e que é, conceitualmente, simples generalização do EMD básico. Várias aplicações do modelo INDSCAL são apresentadas em 1970 e o número de aplicações publicadas cresce rapidamente. Vários métodos de EMD a 'três fatores' foram sugeridos por, entre outros, Tucker (1964, 1972), Harshman (1972), Tucker e Messick (1963), mas a mais bem sucedida aproximação continuou sendo o Modelo INDSCAL proposto por Carroll e Chang (1970). Segundo vários autores uma vantagem do EMD a 'três fatores' reside no fato das dimensões do espaço obtidas por esse modelo serem não rotacionáveis.

O programa ALSCAL (Escalonamento por Mínimos Quadrados Alternados) proposto por Takane e Young (1977) e a inter-relacionada série de programas que pertence ao pacote MULTISCALE proposto por Ramsay (1977) podem estimar modelos de escalonamento para diferenças proporcionais.

Ramsay (1977, 1980) e Takane (1981) sugerem aproximações para estimar modelos de EMD incluindo a aplicação de técnicas de verossimilhança para a estimação de coordenadas. Gower (1977) sugere um número de outros métodos os quais podem ser usados diretamente em matrizes assimétricas para obter uma representação espacial. Kruskal (1978) apresenta algumas outras formas de aplicação da própria técnica de EMD a 'três fatores' INDSCAL na análise de dados.

A eficiência e a robustez dos diversos métodos de EMD foi comprovada através de vários estudos usando técnicas de Monte

Carlo (Takane e Young (1977), Mac Callum e Cornélius (1977), Spence e Young (1978) , ou com os dados sintéticos (Kruskal, 1964) ou ainda usando dados dos quais se conhecia de antemão a configuração desejada (Torgerson, 1962 ) observando, então, que as técnicas de EMD realmente produzem configurações fiéis às medidas de proximidade das quais se parte. Mardia (1979) apresenta uma medida de 'Adequação de Ajuste' usada para comparar duas configurações.

Existem evidências de que o MULTISCALE ( Ramsay, 1977) é melhor do que o ALSCAL ( Takane, Young e De Leeuw, 1977), os quais, por sua vez, são melhores do que o INDSCAL ( Carroll e Chang, 1970).

Spence (1982),Null e Sarle (1982) e Heiser (1987) tem argumentado a necessidade de algoritmos de EMD que sejam resistentes ao efeito de observações estranhas ( 'outliers' ).

## 1.3 - DESCRIÇÃO DO TRABALHO

O principal objetivo deste trabalho é orientar quando se deve aplicar uma técnica de escalonamento multidimensional e como fazer para aplicar esta técnica, desde a coleta de dados até a interpretação dos resultados.

Foram considerados neste trabalho apenas os métodos de escalonamento a 'dois fatores', ou seja, os métodos que iniciam seu procedimento partindo de uma única matriz de proximidades.

Algumas medidas de proximidade que podem ser obtidas a partir de uma matriz de dados ,ou diretamente na coleta de dados, serão apresentadas no capítulo 2.

O capítulo 3 apresenta dois importantes métodos de escalonamento multidimensional a 'dois fatores', que são o método métrico clássico proposto por Torgerson e o método não-métrico proposto por Kruskal.

Com respeito ao método métrico, serão apresentados os procedimentos usados para obter soluções a partir de matrizes de distâncias, dissimilaridades e similaridades.

No método não-métrico um único procedimento iterativo é usado para obter soluções a partir de dissimilaridades ou similaridades e, portanto, este procedimento iterativo será apresentado juntamente com a medida de ajuste de configurações obtidas através do escalonamento não-métrico, conhecida como "stress" e introduzida por Kruskal.

O capítulo 3 ainda apresenta alguns procedimentos básicos utilizados para determinar o número de dimensões que deve ser usado para representar os dados e para interpretar a configuração obtida através do escalonamento, e uma comparação dos métodos métrico e não-métrico.

O capítulo 4 consiste de exemplos de aplicação dos métodos de escalonamento apresentados, feita na área de geologia. Uma comparação dos resultados obtidos com diferentes medidas de proximidade é apresentada, assim como, a comparação entre os resultados obtidos nas análises de escalonamento multidimensional e os resultados obtidos nas análises de agrupamento.

Finalmente, no capítulo 5, serão apresentados o programa estatístico NTSYS e as rotinas deste programa estatístico utilizadas na realização da aplicação de escalonamento multidimensional apresentada no capítulo anterior, juntamente com um guia de como usá-las. O programa estatístico, que será apresentado, inclue rotinas que podem ser úteis em aplicações de escalonamento métrico e de escalonamento não-métrico.

# CAPITULO 2

# MEDIDAS DE PROXIMIDADE

## 2.1 - INTRODUÇÃO

Quando uma análise de dados requer a aplicação de alguma técnica de EMD a primeira etapa a ser considerada é a definição da medida de proximidade a ser usada.

Na prática, os dados podem apresentar-se tanto sob a forma de uma matriz de dados, que apresenta observações de p objetos feitas por n indivíduos, como sob a forma de uma matriz de proximidades, que apresenta uma medida de proximidade para cada par de objetos.

Quando os dados são coletados na forma de medidas de proximidade a aplicação da técnica de EMD é imediata, mas quando os dados são coletados sob a forma de uma matriz de dados (n x p) é necessário que se obtenha medidas de proximidade derivadas desses dados, a fim de que se possa aplicar uma técnica de EMD.

Uma medida de proximidade, referida também como medida de similaridade, dissimilaridade ou distância, deve ser um valor

que indique o quanto dois objetos são semelhantes ( medida de similaridade ), ou o quanto dois objetos são diferentes ou estão distantes um do outro (medida de dissimilaridade ou distância), ou ainda um valor que indique o grau com que dois objetos são percebidos como sendo semelhantes ou diferentes por determinado indivíduo.

Existem vários procedimentos para se obter medidas de proximidade diretamente na coleta de dados . Um procedimento comum consiste em pedir a determinados indivíduos que julguem diretamente o grau de semelhança ( ou diferença ) entre dois objetos, numa escala de 1 a 9. Outros procedimentos serão descritos na seção 2.6.

Uma forma , também muito comum , de se obter medidas de proximidade é através de matrizes de dados que forneçam dados relativos aos objetos. Quando os dados são coletados sob a forma de uma matriz de dados, várias medidas de proximidade podem ser obtidas a partir desta matriz, uma vez que a matriz pode apresentar dados quantitativos, ou qualitativos, ou mistos, e existem medidas de proximidade apropriadas para cada um destes casos. Algumas destas medidas de proximidade serão apresentadas nas seções 2.3, 2.4 e 2.5.

Uma das formas mais comuns de se derivar uma medida de proximidade é calculando coeficientes de similaridade , coeficientes de dissimilaridade ou distâncias entre objetos.

Uma vez obtidas as medidas de proximidade elas são dispostas em matrizes, nas quais cada elemento $\delta_{ij}$ representa a proximidade entre o objeto i e o objeto j.

Medidas de similaridade e dissimilaridade são rigorosamente relacionadas de uma maneira inversa, ou seja, se $\delta$ é uma função definida sobre cada par de objetos, a qual mede a similaridade entre os objetos, então é fácil derivar uma medida correspondente de dissimilaridade assim como $\delta^* = (\text{constante} - \delta)$.

## 2.2 - PROPRIEDADES DAS MEDIDAS DE PROXIMIDADE

Definição 2.2.1 - Uma medida de similaridade entre objetos i e j, denotada por $\delta_{ij}$, deve satisfazer as seguintes condições:

(i) $\delta_{ij} = \delta_{ji}$

(ii) $\delta_{ij} > 0$

(iii) $\delta_{ij}$ cresce quando a similaridade entre os objetos i e j aumenta.

Definição 2.2.2 - Uma matriz $\Delta(n \times n)$ é chamada matriz de similaridades se $\delta_{ij} = \delta_{ji}$ e se $\delta_{ij} \leq \delta_{ii}$ para todo i, j.

Definição 2.2.3 - Uma medida de dissimilaridade entre objetos i e j, denotada por $\delta^*_{ij}$ e definida por $\delta^*_{ij} = \delta_{ii} - \delta_{ij}$, deve satisfazer as seguintes condições:

(i) $\delta^*_{ij} = \delta^*_{ji}$

(ii) $\delta^*_{ij} \geq 0$

(iii) $\delta^*_{ii} = 0$

(iv) $\delta^*_{ij}$ decresce quando a similaridade entre os objetos i e j aumenta.

Definição 2.2.4 - Uma matriz $\Delta^*(n \times n)$ é chamada uma matriz de dissimilaridades se $\delta^*_{ij} = \delta^*_{ji}$, $\delta^*_{ii} = 0$ e se $\delta^*_{ij} \geq \delta^*_{ii}$ para todo i, j.

Definição 2.2.5 - Uma medida de distância entre objetos i e j, denotada por $d_{ij}$, deve satisfazer as seguintes condições:

(i) $d_{ij} = d_{ji}$

(ii) $d_{ij} \geq 0$

(iii) $d_{ij} = 0$ se e somente se $i = j$

(iv) $d_{ij} \leq d_{ik} + d_{jk}$

Definição 2.2.6 - Uma matriz D(n x n) é chamada matriz de distâncias se $d_{ij} = d_{ji}$, $d_{ij} > d_{ii}$, $d_{ii}=0$ e $d_{ij} \leq d_{ik} + d_{jk}$ para todo i, j, k.

Definição 2.2.7 - Uma matriz de distâncias D(n x n) é chamada matriz de distâncias euclidianas se existir uma configuração de pontos em algum espaço euclidiano onde distâncias entre pontos são dadas por D, isto é, se para algum p, existir pontos $x_1, \ldots, x_n \in \mathbb{R}^p$, tal que

$$d_{ij}^2 = (x_i - x_j)'(x_i - x_j).$$

Para transformar medidas de dissimilaridade ($\delta_{ij}^*$) em distâncias métricas basta definir uma função monotônica como sendo a adição de uma constante, suficientemente grande, às medidas de dissimilaridade e deixar $\delta_{ii}^* = 0$. A menor constante que pode ser usada é:

$$C_{min} = \max_{i,j,k} (\delta_{ij}^* - \delta_{ik}^* - \delta_{jk}^*)$$

e $d_{ij}$ é definido como:

$$\begin{cases} d_{ij} = \delta_{ij}^* + C_{min} & \text{para } i \neq j \\ d_{ij} = 0 & \text{para } i = j \end{cases}$$

Alguma constante $C > C_{min}$ pode também transformar as medidas de dissimilaridade em distâncias métricas.

Definição 2.2.8 - A transformação padrão de uma matriz de similaridades $\Delta$ em uma matriz de distâncias D é definida por:

$$d_{ij} = (\delta_{ii} - 2\delta_{ij} + \delta_{jj})^{1/2}.$$

Note que sendo $\delta_{ij} \leq \delta_{ii}$ então $\delta_{ii} - 2\delta_{ij} + \delta_{jj} \geq 0$ e $d_{ii} = 0$. Portanto D é uma matriz de distâncias.

Teorema 2.2.1 - Se $\Delta \geq 0$, a matriz de distâncias D definida pela transformação padrão $d_{ij} = (\delta_{ii} - 2\delta_{ij} + \delta_{jj})^{1/2}$ é Euclidiana.

## 2.3 - MEDIDAS DE PROXIMIDADE PARA DADOS QUANTITATIVOS

Considere uma matriz de dados (n x p) apresentando n observações de p objetos.

A mais familiar medida de dissimilaridade entre objetos é a distância Euclidiana, tal que

$$d_{ij} = \text{distância entre os objetos } i \text{ e } j$$

$$= \left[ \sum_{k=1}^{n} \left( x_{ik} - x_{jk} \right)^2 \right]^{1/2}$$

A distância Euclidiana usada em dados brutos pode ser muito insatisfatória se as variáveis são medidas em diferentes unidades e tem variâncias diferentes, e também se as variáveis são correlacionadas, visto que a distância Euclidiana é muito afetada por mudança de escala, podendo mudar não só os valores da distância, como também os postos. Para evitar a mudança de postos a distância é usualmente calculada após as variáveis terem sido padronizadas.

A distância Euclidiana satisfaz a desigualdade métrica, é semi-definida positiva e é invariante sob transformações ortogonais dos x's.

Embora a distância Euclidiana seja uma métrica, pode-se facilmente achar uma função monótona dela a qual não é métrica, como por exemplo, a distância Euclidiana ao quadrado que é um coeficiente de dissimilaridade e não é métrica.

Contudo os postos produzidos pela distância Euclidiana e seu quadrado são os mesmos e são igualmente úteis para muitos propósitos. A distância Euclidiana ao quadrado é dada por

$$d_{ij}^{2} = \sum_{k=1}^{n} \left( x_{ik} - x_{jk} \right)^{2}.$$

Existem muitas outras possíveis métricas e uma a qual tem sido usada é a métrica de Minkowski definida como

$$M_{ij} = \left[ \sum_{k=1}^{n} | x_{ik} - x_{jk} |^{R} \right]^{1/R} , \quad R \geq 0$$

Quando R = 2, a métrica de Minkowski se reduz a distância Euclidiana e quando R = 1, tem-se a métrica

$$M_{ij} = \sum_{k=1}^{n} | x_{ik} - x_{jk} |$$

a qual é chamada métrica absoluta ou " city blok " ou distância de Manhattan.

A mais usual medida de similaridade entre duas variáveis (ou dois objetos) é o coeficiente de correlação de Pearson, também conhecido como correlação momento-produto, que é dado por

$$r = \frac{\Sigma \left[ X - \bar{X} \right] \left[ Y - \bar{Y} \right]}{\sqrt{\Sigma \left[ X - \bar{X} \right]^2 \Sigma \left[ Y - \bar{Y} \right]^2}}$$

Uma outra medida de similaridade que pode ser adotada é o cosseno do ângulo entre vetores dado por

$$\cos^2 \theta = \left[ X_i^T X_j \right] / \left[ \left[ X_i^T X_i \right] \cdot \left[ X_j^T X_j \right] \right]$$

$$= \sum_{k=1}^{n} \left[ x_{ik} x_{jk} \right]^2 / \left[ \sum_{k=1}^{n} x_{ik}^2 \sum_{k=1}^{n} x_{jk}^2 \right]$$

Existem outras medidas de distância, dissimilaridade e similaridade, que podem ser utilizadas, mas estas são as mais conhecidas e mais comumente usadas.

## 2.4 - MEDIDAS DE PROXIMIDADES PARA DADOS QUALITATIVOS

Para dados binários é comum definir um coeficiente de similaridade mais do que um coeficiente de dissimilaridade, uma vez que alguns coeficientes de similaridade são arranjados para ficarem no intervalo [0,1] e neste caso a correspondente medida de dissimilaridade pode ser obtida subtraindo de 1.

Considere que a presença ou ausência de p características sobre dois objetos X e Y seja denotada por $( x_1, \ldots, x_p )$ e $( y_1, \ldots, y_p )$, onde $x_i = 1$ ou 0 dependendo da i-ésima característica estar presente ou ausente para o objeto X, e considere que

$$a = \Sigma x_i y_i ,$$
$$b = \Sigma (1 - x_i) y_i ,$$

$$a = \sum x_i\, y_i ,$$
$$b = \sum (1 - x_i) y_i ,$$
$$c = \sum x_i\, (1 - y_i) ,$$
$$d = \sum (1 - x_i)\, (1 - y_i) ,$$

são as frequências das combinações $( x_i, y_i ) = (1,1), (0,1), (1,0)$ e $(0,0)$, respectivamente.

A mais simples medida de similaridade entre os objetos X e Y que pode ser obtida é

$$S_1 ( X,Y ) = a / p ,$$

onde a é o número de combinações positivas (1,1) e p é o número de características.

Mas neste caso $(1 - S_1)$ é tal que a dissimilaridade entre um indivíduo e ele mesmo não precisa ser zero e assim $d = (1 - S_1)$ não é um coeficiente de dissimilaridade.

O coeficiente de Jaccard é definido por

$$S_2 ( X,Y ) = a / ( a + b + c ).$$

Nota-se que $S_2 ( X,Y )$ exclue as combinações negativas (0,0) e dá igual peso para todas as outras combinações.

No caso do coeficiente de Jaccard $d = (1 - S_2)$ produz um coeficiente de dissimilaridade métrico.

O coeficiente de concordância simples é definido por

$$S_3 ( X,Y ) = ( a + d ) / p$$

e mede a proporção de atributos para os quais os indivíduos dão as mesmas respostas. O coeficiente de concordância simples inclue as combinações negativas (0,0) e dá pesos iguais para todas combinações. Obviamente $0 \leq S_3 ( X,Y ) \leq 1$, e a correspondente medida de dissimilaridade $d = (1 - S_3)$ satisfaz as condições para ser um coeficiente de dissimilaridade.

Vários outros coeficientes tem sido sugeridos, tais como

$$S_4 ( X,Y ) = 2a / ( 2a + b + c ),$$

$$S_5 ( X,Y ) = a / [ a + 2( b + c )]$$

e $$S_6 ( X,Y ) = 2( a + d ) / [ 2( a + d ) + ( b + c )].$$

Nota-se que os três coeficientes dão pesos diferentes para as combinações e que somente o coeficiente $S_6$ inclui as combinações negativas (0,0).

Kendal (1975) usa uma medida de similaridade definida por

$$K = a,$$

que é, simplesmente , o número de combinações positivas.

Diferentes coeficientes de similaridade podem ter diferentes valores para a mesma série de dados, mas este fato é relativamente sem importância se todos os coeficientes forem juntamente monotônicos, de modo que, se todos valores de pares de indivíduos em um coeficiente estão ordenados tal que eles formam uma série monotônica (que é uma série que cresce ou decresce sobre toda sua extensão) os valores para os pares de indivíduos calculados por outros coeficientes serão monotônicos.

Quando se tem uma matriz de dados contendo a informação presença / ausência sobre p atributos para n objetos e $x_{ij}$ = 1 ou 0, então uma matriz de similaridades, apresentando uma das medidas de similaridade apresentadas para cada par de objetos, pode ser obtida. Por exemplo,

$$S_1 = 1/p ( X X' ),$$

$$K = X X'$$

e $$S_3 = 1/p [ X X' + ( J - X )(J - X )'],$$

onde $J = 1\ 1'$.

Dados qualitativos para os quais os atributos tem mais do que duas categorias de classificação podem ser tratados de uma

maneira similar como dados binários com cada categoria sendo considerada como um atributo que pode estar presente ou ausente.

## 2.5 - MEDIDAS DE PROXIMIDADE PARA DADOS MISTOS

Existem casos em que uma matriz de dados pode apresentar tanto dados quantitativos como qualitativos. Em tais casos um coeficiente de similaridade proposto por Gower (1971) pode ser muito útil. Este coeficiente é definido como:

$$S_{ij} = \left[ \sum_{k=1}^{p} S_{ij,k} \right] / \left[ \sum_{k=1}^{p} W_{ij,k} \right]$$

onde o peso $W_{ij,k}$ é igual a 1 ou 0 dependendo se a comparação é considerada válida ou não para a variável $k$ . Os escores $S_{ij,k}$ podem ser obtidos como segue:

(A) Dados Qualitativos:

$S_{ij,k} = 0$, se os dados são diferentes.
$S_{ij,k} = 1$, se os dados são iguais.

(B) Dados Quantitativos:

$$S_{ij,k} = 1 - | X_{ik} - X_{jk} | / R_k \ , \ k = 1,2,\ldots,p \ ; \ i,j = 1,2,\ldots,n,$$

onde $R_k$ é a amplitude da variável $k$.

## 2.6 - OUTROS PROCEDIMENTOS PARA OBTENÇÃO DE MEDIDAS DE PROXIMIDADE

Existem vários procedimentos que são úteis para coletar dados diretamente sob a forma de medidas de proximidade e alguns destes procedimentos serão apresentados a seguir.

### 2.6.1 - Tríades

a) Método das Combinações Triádicas (Torgerson, 1962)

Dados n objetos, um total de $C_{n,3} = n(n-1)(n-2)/6$ tríades (grupos de três objetos) são apresentadas a cada indivíduo do experimento, o qual ao observar a tríade (i,j,k), por exemplo, deverá indicar quais são os dois objetos mais semelhantes e quais são os mais diferentes. Se o indivíduo indicar que i e j são mais semelhantes e i e k são mais diferentes, três julgamentos são inferidos:

(1) i é mais semelhante a j do que a k.
(2) j é mais semelhante a i do que a k.
(3) k é mais semelhante a j do que a i.

Apresentadas todas as tríades a todos os indivíduos, constrói-se para cada objeto k uma matriz $\left[{}_kp_{ij}\right]$ onde ${}_kp_{ij}$ é a proporção de vezes que o objeto k é indicado como mais semelhante a i do que a j , e os elementos desta matriz são então processados (Torgerson, 1962) até que sejam obtidas medidas de distância entre os objetos numa escala de razão.

b) Método Completo das Tríades

Dados n objetos, cada uma das $C_{n,3} = n(n-1)(n-2)/6$ tríades é apresentada três vezes a cada indivíduo, sendo que na

primeira vez o indivíduo deve julgar se i é mais semelhante a j ou a k; na segunda vez ele deve julgar se j é mais semelhante a i ou a k e na terceira vez se k é mais semelhante a i ou a j. Após todas as possíveis tríades terem sido apresentadas três vezes a cada um dos indivíduos do experimento, constrói-se as matrizes $\left[ {}_{k}p_{ij} \right]$ analogamente ao método das combinaçõs triadicas.

2.6.2 - Método dos Pares ( Wish e Carroll, 1973 )

Dados n objetos , um total de $C_{n,2} = n(n-1)/2$ pares são apresentados aos indivíduos e cada indivíduo deve atribuir uma nota de similaridade a cada par de objetos, numa escala que costuma ser de 9 a 10 pontos. A medida de similaridade $\delta_{ij}$ pode ser obtida através da média ou da mediana das notas atribuida por todos os sujeitos ao par de objetos (i,j).

2.6.3 - Método do Ponto de Âncora ( Green e Carmone, 1972 )

Os n objetos são apresentados n vezes a cada um dos indivíduos do experimento, sendo que a cada vez um dos n objetos é escolhido pelo experimentador como "ponto de âncora" e o indivíduo deve ordenar os demais (n-1) objetos de acordo com o seu grau de semelhança em relação ao objeto "ponto de âncora", até que todos os n objetos tenham desempenhado o papel de "ponto de âncora".

O programa para computador, Tricon, desenvolvido por Green, Carmone e Robinson (1968) obtém as medidas de similaridade para cada objeto.

## 2.6.4 - Agrupamento em Categorias

Os n objetos são apresentados a cada indivíduo e ele deve dividir estes objetos em categorias mutuamente exclusivas, de modo que os objetos pertencentes à cada categoria sejam similares entre si, podendo ser estabelecido limite máximo ou mínimo para o número de categorias. Este método é útil quando o experimentador tem um grande número de objetos.

As medidas de similaridade $(\delta_{ij})$ podem ser obtidas de diversas formas, tais como as seguintes:

(i) Green (1969) constrói uma matriz A(n x n) para cada indivíduo do experimento onde:

$$A_k = (a_{ij})_{nxn} \begin{cases} a_{ij} = 1 & \text{se os objetos } i \text{ e } j \text{ estiverem na mesma categoria} \\ a_{ij} = 0 & \text{caso contrário} \end{cases}$$

Em seguida efetua a soma de todas as matrizes $A_k$ obtendo assim a matriz de similaridade $\Delta = (\delta_{ij})_{nxn}$ onde $\delta_{ij}$ é a medida de similaridade entre os objetos i e j.

(ii) Wish e Carroll (1973) constroem a matriz $\Delta = (\delta_{ij})_{nxn}$ medindo a proporção de vezes que os objetos i e j se apresentam na mesma categoria, ou seja, $\delta_{ij} = S / R$ é a medida de similaridade entre os objetos i e j, onde R é o número total de indivíduos do experimento e S é o número de indivíduos que classificaram os objetos i e j na mesma categoria.

## 2.6.5 - Frequências de Semelhanças

Os n objetos são apresentados n vezes a cada indivíduo sendo que em cada vez um objeto i é escolhido pelo experimentador e o indivíduo deve identificar entre os (n-1) objetos restantes um objeto que mais se assemelhe ao objeto i segundo alguma

característica, até que todos os n objetos tenham ocupado o lugar do objeto i. A partir das respostas dos indivíduos constrói-se uma matriz $[\ f_{ij}\ ]$ de frequências onde cada elemento $f_{ij}$ corresponde à frequência com que o objeto i foi considerado similar ao objeto j.

As frequências absolutas $(f_{ij})$ e relativas $(f_{ij} / n)$ podem ser utilizadas como medida de similaridade (grosseira medida de similaridade), em que quanto mais alta a frequência, tanto maior a similaridade percebida.

Pode-se, também, a partir das frequências $f_{ij}$ obter-se um índice de similaridade

$$S_{ij} = (f_{ij} + f_{ji}) / (f_{i.} + f_{j.} + f_{.i} + f_{.j})$$

que ao serem normalizados resultam em medidas de similaridade $(\delta_{ij})$ (Mauser, 1972).

## 2.6.6. - Caracteres Concordantes

Os n objetos são estudados aos pares, segundo a presença ou ausência de alguns caracteres que interessam ao experimentador. Tendo-se os objetos i e j e os caracteres A , B , C , D , atribui-se o sinal (+) ao objeto cada vez que um caracter estiver presente e o sinal (-) quando estiver ausente. A medida de similaridade é, então, a proporção de sinais concordantes.

# CAPÍTULO 3

## PRINCIPAIS MÉTODOS DE ESCALONAMENTO MULTIDIMENSIONAL

### 3.1 - INTRODUÇÃO

Os dois tipos mais importantes de procedimento de escalonamento são denominados escalonamento métrico (também referido como escalonamento clássico ou Análise de Coordenadas Principais) e escalonamento não-métrico (também referido como escalonamento ordinal ou, simplesmente, escalonamento multidimensional).

Modelos clássicos de escalonamento multidimnesional foram designados para preservar informação métrica dos dados, assumindo que as proximidades eram uma função linear do modelo de distâncias, enquanto modelos de escalonamento multidimensional não-métricos procuravam somente preservar informação não-métrica dos dados . Em procedimentos de escalonamento multidimensional esta distinção tem sido implementada, basicamente, pela forma de regressão usada, sendo que, regressão linear de dados sobre distâncias tem sido usada no caso métrico e regressão monotônica

tem sido usada no caso não-métrico.

O escalonamento métrico usa as verdadeiras magnitudes das proximidades originais entre os objetos para obter uma representação geométrica destes objetos e, sendo assim, deve ser usado quando os dados a serem escalonados estiverem em escala de razão (portanto com propriedades de distâncias) ou na escala intervalar.

Por outro lado, o escalonamento não-métrico usa somente as propriedades ordinais das proximidades originais entre objetos e, sendo assim, deve ser usado quando os dados estiverem na escala ordinal.

O método de escalonamento métrico de Torgerson, também conhecido como método clássico, e o método não-métrico de Kruskal, serão descritos neste capítulo.

## 3.2 - MÉTODO MÉTRICO

O escalonamento métrico é um método de representação algébrica que tem como objetivo achar uma configuração de pontos a partir das dissimilaridades entre os objetos, o qual é apropriado particularmente quando as dissimilaridades são, exatamente ou aproximadamente, distâncias euclidianas.

Este método admite a existência de uma "verdadeira" configuração em k dimensões, cujas distâncias entre os pontos são $\delta_{ij}$ e tenta reconstruir esta configuração usando uma matriz de distâncias observadas D, cujos elementos são da forma de $d_{ij} = \delta_{ij} + e_{ij}$, onde os $e_{ij}$ são erros de medida acrescidos de erros de distorções causadas pelo fato das distâncias observadas não corresponderem exatamente às distâncias entre os pontos de uma configuração em $R^k$.

Deve-se notar que, dada uma série de distâncias euclidianas, não existe uma única representação dos pontos os

quais dão origem a estas distâncias. Isto significa que não se pode determinar a localização e a orientação da configuração. O problema de localização é usualmente superado centrando a configuração na origem, enquanto o problema de orientação pode ser superado submetendo a configuração obtida a uma transformação ortogonal, a qual deixa distâncias e ângulos inalterados. Em outras palavras, uma configuração obtida pelo escalonamento métrico é indeterminada com respeito a translações, rotações e reflexões.

Como exemplo, considere o caso em que se conhece as distâncias entre algumas capitais brasileiras, que são dadas na tabela 3.1. Com bases nestas informações nada se pode dizer a respeito da latitude e longitude de determinada capital e também não se pode dizer se determinada capital fica ao norte, sul, leste ou oeste, apesar de se poder dizer se esta capital é ou não um ponto extremo.

A reconstrução do mapa bidimensional do Brasil pode ser obtida através do escalonamento métrico, usando a distâncias entre algumas capitais brasileiras da tabela 3.1 , mas esta reconstrução pode ser feita de 'cabeça para baixo' ou de 'lado' e, quando isso acontece, é necessário então submeter a configuração obtida a uma rotação, ou reflexão, para obter o familiar mapa do Brasil.

O mapa bidimensional do Brasil obtido pelo método métrico de escalonamento multidimencional (Torgerson 1952, 1958) pode ser visto na figura 3.1.

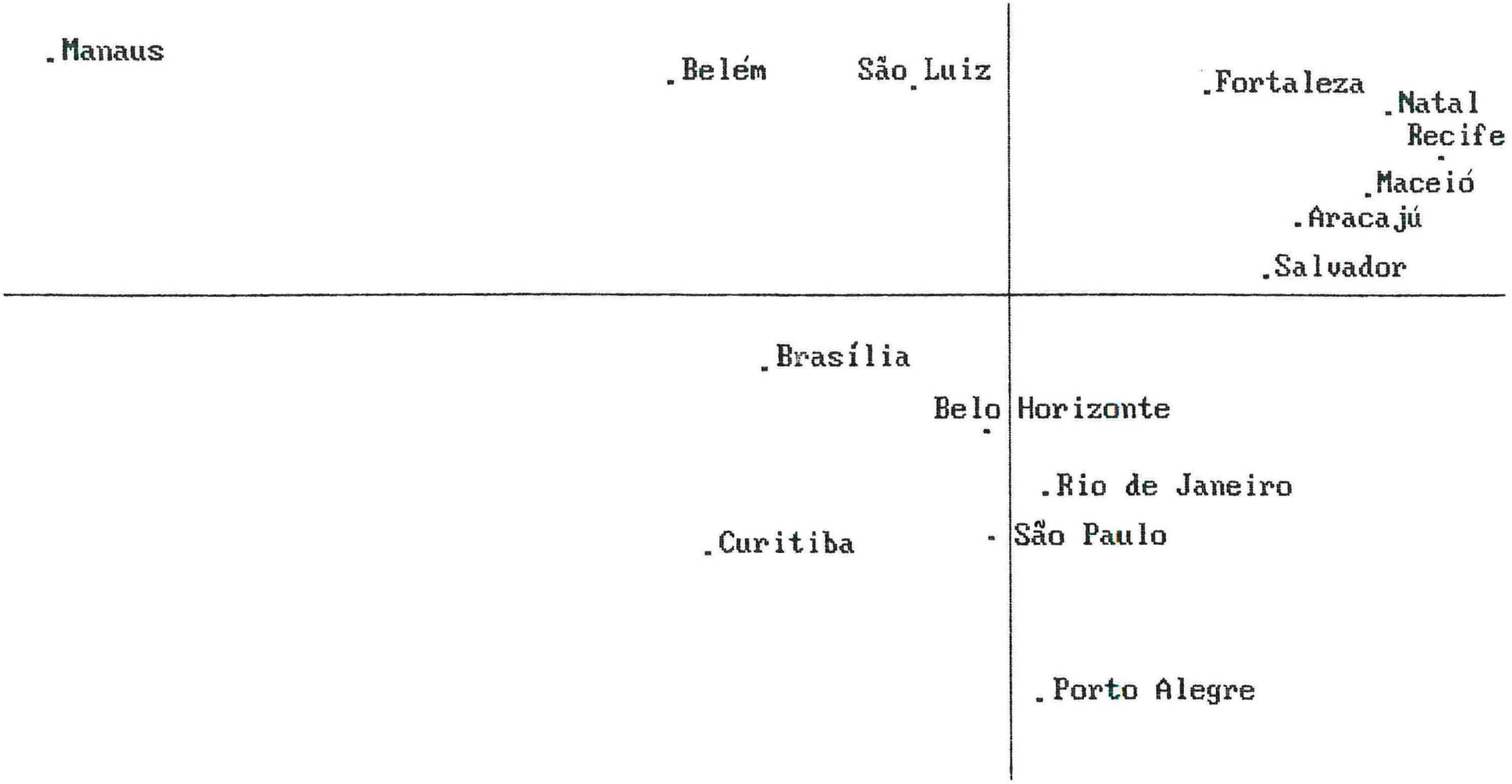
TABELA 3.1 - DISTÂNCIAS AÉREAS ENTRE ALGUMAS CAPITAIS BRASILEIRAS (KM)

| | ARACAJÚ | BELÉM | B.HORIZO | BRASÍLIA | CURITIBA | FORTALEZ | MACEIÓ |
|---|---|---|---|---|---|---|---|
| ARACAJÚ | 0 | | | | | | |
| BELÉM | 1590 | 0 | | | | | |
| B.HORIZO | 1350 | 2123 | 0 | | | | |
| BRASÍLIA | 1271 | 1627 | 589 | 0 | | | |
| CURITIBA | 2010 | 2574 | 823 | 1087 | 0 | | |
| FORTALEZ | 810 | 1138 | 1860 | 1682 | 2598 | 0 | |
| MACEIÓ | 202 | 1632 | 1416 | 1566 | 2205 | 717 | 0 |
| MANAUS | 2574 | 1288 | 2569 | 1967 | 2634 | 2295 | 2670 |
| NATAL | 606 | 1550 | 1800 | 1775 | 2580 | 444 | 435 |
| P.ALEGRE | 2520 | 3084 | 1370 | 1617 | 547 | 3126 | 2712 |
| RECIFE | 386 | 1680 | 1632 | 1632 | 2400 | 640 | 191 |
| R.JANEIR | 1485 | 2460 | 340 | 900 | 669 | 2190 | 1680 |
| S.LUIZ | 1237 | 493 | 1848 | 1530 | 2514 | 640 | 1200 |
| S.PAULO | 1740 | 2490 | 500 | 865 | 330 | 2238 | 1940 |
| SALVADOR | 267 | 1695 | 980 | 1053 | 1734 | 1018 | 464 |

| | MANAUS | NATAL | P.ALEGRE | RECIFE | R.JANEIR | S.LUIZ | S.PAULO |
|---|---|---|---|---|---|---|---|
| MANAUS | 0 | | | | | | |
| NATAL | 2658 | 0 | | | | | |
| P.ALEGRE | 3987 | 3069 | 0 | | | | |
| RECIFE | 2823 | 252 | 3083 | 0 | | | |
| R.JANEIR | 2854 | 2122 | 1133 | 1685 | 0 | | |
| S.LUIZ | 1752 | 1035 | 3042 | 1197 | 2271 | 0 | |
| S.PAULO | 3100 | 2486 | 844 | 2135 | 364 | 2360 | 0 |
| SALVADOR | 2617 | 870 | 2241 | 654 | 1220 | 1290 | 1486 |

| | SALVADOR |
|---|---|
| SALVADOR | 0 |

Figura 3.1 - Mapa do Brasil obtido pelo Método de Escalonamento Métrico



3.2.1 - Solução Clássica

Considere que se conhece a matriz de dados **X** de dimensões **n x p** e a matriz de 'soma de quadrados e produtos cruzados' de linhas

$$\mathbf{A} = \mathbf{XX'}$$

de dimensões **n x n**, na qual o **i,j**-ésimo termo é

$$a_{ij} = \sum_{k=1}^{p} x_{ik}\, x_{jk} .$$

Se $d_{ij}$ é a distância euclidiana entre os objetos **i** e **j**, então uma matriz de distâncias euclidianas ao quadrado $\mathbf{D^2}$ de dimensões **n x n**, usando os vetores linha **i** e **j** da matriz **X**, é tal que

$$d_{ij}^2 = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2$$

$$= \sum_{k=1}^{p} x_{ik}^2 + \sum_{k=1}^{p} x_{jk}^2 - 2 \sum_{k=1}^{p} x_{ik}\, x_{jk}$$

$$= a_{ii} + a_{jj} - 2a_{ij}$$

O procedimento métrico de escalonamento parte de uma matriz de distâncias **D** de dimensões **n x n** e tenta achar uma configuração de pontos em **k** dimensões, ou deja, tenta achar as coordenadas dos pontos.

Considere que se conhece a matriz de distâncias euclidianas e também a matriz $\mathbf{D^2}$ do quadrado das distâncias euclidianas . Então, as coordenadas dos pontos **i** e **j** podem ser obtidas em dois estágios, sendo que, o primeiro estágio consiste em encontrar os elementos $a_{ij}$ da matriz **A** em termos dos $d_{ij}^2$'s, e o segundo estágio consiste em fatorá-la na forma **A = XX'**.

Para obter os elementos $a_{ij}$, da matriz **A**, em termos dos $d^2_{ij}$'s é preciso inverter a equação

$$d^2_{ij} = a_{ii} + a_{jj} - 2a_{ij} ,$$

a qual é verdadeira para todo **i**, **j**.

Dado que não existe uma única solução, a não ser que se imponha uma dada localização, assume-se que $\bar{x} = 0$ de tal modo que $\sum_{i=1}^{n} x_{ik} = 0$, para todo **k**.

Usando a restrição $\sum_{i=1}^{n} x_{ik} = 0$ ( para todo **k** ) em

$$a_{ij} = \sum_{k=1}^{p} x_{ik}\, x_{jk} ,$$

a soma dos termos em qualquer linha de **A**, ou em qualquer coluna de **A**, torna-se zero. Consequentemente, somando-se

$$d^2_{ij} = a_{ii} + a_{jj} - 2a_{ij}$$

sobre **i**, sobre **j** e sobre **i** e **j**, obtem-se

$$\sum_{i=1}^{n} d^2_{ij} = T + na_{jj}$$

$$\sum_{j=1}^{n} d^2_{ij} = na_{ii} + T$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} d^2_{ij} = 2nT$$

onde $T = \sum_{i=1}^{n} a_{ii}$ é o traço da matriz **A**.

Resolvendo $d^2_{ij} = a_{ii} + a_{jj} - 2a_{ij}$ e suas somas em i, em j e em i e j, conjuntamente, tem-se

$$a_{ij} = -1/2 \left[ d^2_{ij} - d^2_{i.} - d^2_{.j} - d^2_{..} \right]$$

onde $d^2_{i.} = 1/n \sum_{i=1}^{n} d^2_{ij}$ (média dos termos da i-ésima linha de $D^2$);

$d^2_{.j} = 1/n \sum_{j=1}^{n} d^2_{ij}$ (média dos termos da j-ésima coluna de $D^2$) e

$d^2_{..} = 1/n^2 \sum_{i=1}^{n} \sum_{j=1}^{n} d^2_{ij}$ ( média geral dos termos de $D^2$).

Para obter os elementos da matriz A, à partir das distâncias quadradas, usa-se

$$a_{ij} = -1/2 \left[ d^2_{ij} - d^2_{i.} - d^2_{.j} - d^2_{..} \right].$$

Uma maneira fácil para obter os $a_{ij}$ é 'centrar duas vezes' $D^2$, primeiro subtraindo a média dos elementos em cada linha de todos os termos desta linha, e então fazendo a mesma coisa para as colunas, o que automaticamente torna o termo médio geral igual a zero, e o multiplicador ( 1/2 ) pode então ser aplicado.

Para obter a matriz de coordenadas X a partir da matriz A é preciso que se faça uma análise dos vetores característicos de A.

Se $D^2$ consiste do quadrado de distâncias euclidianas exato, a matriz A é simétrica e semi-definida positiva e, se A é de posto k, e $k \leq n$, então A tem k valores característicos positivos não nulos, os quais arranjados em ordem de magnitude são tais que

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k > 0.$$

Os corespondentes vetores característicos são denotados por $\{e_i\}$. Para que a soma de quadrados dos vetores característicos seja igual a $\lambda_i$, faz-se $f_i = \sqrt{(\lambda_i)}\,(e_i)$. Então, uma possível matriz de coordenadas **X**, tal que **A** = **XX'**, é tal que

$$X = \left[ f_1 , f_2 , \ldots , f_k \right]$$

onde **X** é de ordem (**n** x **k**) e se tem uma configuração em **k** dimensões, ou seja, as coordenadas do **i**-ésimo ponto são dadas pelas **i**-ésimas componentes de $\{f_i\}$.

### 3.2.2 - Solução para uma Matriz de Dissimilaridade ou Similaridade

Na prática, pode-se ter uma matriz de dissimilaridade, ao invés de uma matriz de distâncias Euclidiana, e, neste caso, se procede exatamente como se as distâncias fossem euclidianas.

Inicialmente, as 'distâncias' são elevadas ao quadrado e a matriz **A** é então formada, usando $a_{ij} = (-1/2)(d^2_{ij} - d^2_{i.} - d^2_{.j} + d^2_{..})$, e, finalmente, os vetores característicos de **A** são encontrados.

A matriz **A** será simétrica e terá zeros na diagonal, mas, como as distâncias não são euclidianas, não precisa ser semi-definida positiva. Consequentemente, a matriz **A** terá **n** valores característicos, sendo que, alguns deles podem ser negativos.

Se **A** tiver um número não muito grande de valores característicos negativos pequenos, uma satisfatória representação de coordenadas da matriz de dissimilaridade pode ser obtida pelos vetores característicos associados aos primeiros maiores valores característicos positivos. Então, usando

$$f_i = \sqrt{(\lambda_i)}\,(e_i) \ ,$$

obtém-se a matriz de coordenadas como $\mathbf{X} = [\, f_1, f_2, \ldots, f_s \,]$.

Todos valores característicos negativos são descartados, juntamente com qualquer pequeno valor característico positivo e, como regra geral, o maior valor característico negativo será menor do que o menor valor característico positivo, considerado como sendo grande.

Uma regra útil, denominada o 'critério do traço', diz que a soma dos valores característicos pequenos, tanto positivos como negativos, será aproximadamente zero, tal que a soma dos valores característicos positivos considerados como sendo grandes será aproximadamente igual ao posto de **XX'**.

Se **A** tem um número de grandes valores característicos negativos ou vários valores característicos de tamanho médio, tal que a solução requer um grande número de dimensões, o uso de escalonamento métrico pode não ser recomendável.

Quando se tem uma matriz de similaridades,ao invés de uma matriz de distâncias ou dissimilaridades, é necessário transformar esta matriz de similaridades em uma matriz de distâncias para poder obter a solução. Para transformar a matriz pode-se usar a transformação padrão descrita no capítulo 2.

### 3.2.3 - Adequação de Ajuste

Quando obtém-se a solução clássica, a partir de uma matriz de distâncias, uma possível medida da adequação de uma configuração em **k** dimensões pode ser obtida pela "proporção da matriz de distâncias explicada pela solução clássica **k**-dimensional", a qual é dada por

$$\alpha = \sum_{i=1}^{k} \lambda_i \; / \; \sum_{i=1}^{n} \lambda_i$$

onde $\lambda_i$ são os valores característicos de **A**.

Quando obtém-se a solução a partir de uma matriz de dissimilaridade, ou seja , distâncias não-euclidianas, a adequação da representação obtida usando-se os **k** primeiros vetores característicos pode ser obtida por

$$\alpha' = \sum_{i=1}^{k} \lambda_i \; / \; \sum_{i=1}^{n} |\lambda_i|$$

onde $\lambda_i$ são os valores característicos de **A**.

A diferença entre $\alpha$ e $\alpha'$ é a presença do módulo no denominador de $\alpha'$, a qual se deve ao fato de que alguns valores característicos de **A** podem ser negativos.

## 3.3 - MÉTODO NÃO-MÉTRICO

O escalonamento não-métrico, ou ordinal, assim como o escalonamento métrico, tem como propósito encontrar uma representação geométrica de pontos, cujas distâncias entre estes pontos combinem, de alguma maneira, com as similaridades (ou dissimilaridades) originais entre objetos (ou indivíduos).

A diferença básica entre os métodos é que o não-métrico usa somente a ordenação das similaridades para obter a solução, enquanto que o métrico usa informação métrica, ou seja, as verdadeiras magnitudes das similaridades, para obter a solução.

O escalonamento não-métrico, entretanto, toma como hipótese uma relação menos rígida, do que a assumida pelo métrico, entre $d_{ij}$ e $\delta_{ij}$ , como

$$d_{ij} = f ( \delta_{ij} ) + e_{ij}$$

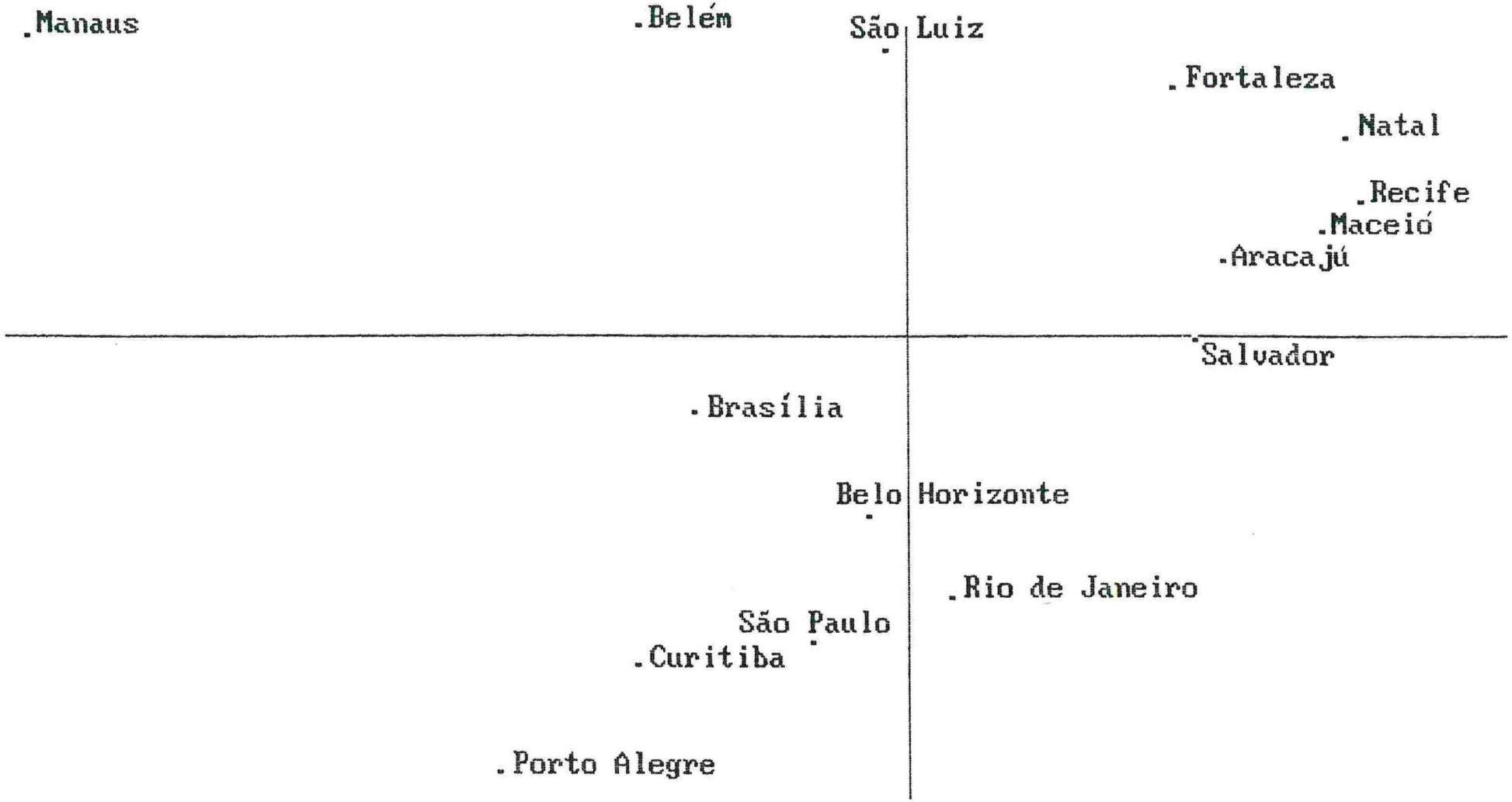
sendo f uma função monotônica desconhecida crescente ou decrescente. Em outras palavras, o escalonamento não-métrico relaxa a forte linearidade suposta pelo escalonamento métrico a respeito do tipo de função que liga as similaridades entre objetos às distâncias entre os pontos, que representam os objetos na configuração.

A grande vantagem do uso de escalonamento não-métrico é que o mesmo algoritmo básico é facilmente generalizado para diferentes tipos de dados e para diferentes modelos. É, portanto, aplicável a uma grande variedade de situações. Assim, frequências, probabilidades, postos, etc, são tão apropriados como medidas de proximidade, como são os coeficientes de correlação, associação, covariâncias e outros. Obviamente, quaisquer séries de medidas com a mesma ordenação de proximidades irão gerar a mesma solução métrica.

Como em escalonamento métrico, não se pode determinar uma única localização e uma única orientação da configuração obtida, isto é, a configuração é indeterminada com respeito a translação, rotação e reflexão. Além disso, a configuração obtida pelo método não-métrico de escalonamento é indeterminada com respeito a mudanças de escala, isto é, é indeterminada com respeito a expansão ou contração uniformes. Entretanto, o problema de não ter signicado a escala da configuração, pode ser superado fixando o centróide da configuração na origem, isto é, requerendo que a raiz quadrada da média das distâncias dos pontos a origem seja unitária.

Um exemplo de representação geométrica de pontos obtida pelo escalonamento não-métrico pode ser visto na figura 3.2, a qual mostra o mapa bidimensional do Brasil, obtido pelo procedimento não-métrico a partir das distâncias dadas na tabela 3.1.

Figura 3.2 - Mapa do Brasil obtido pelo Método de Escalonamento Não-Métrico



### 3.3.1 - Adequação de Ajuste

Como o método de escalonamento não-métrico, proposto por Kruskal, é centrado na medida de 'Adequação de Ajuste', denominada "Stress", considerou-se importante que esta medida fosse inicialmente apresentada.

O escalonamento não-métrico requer que as proximidades observadas entre objetos concordem, de alguma maneira, com as distâncias entre pontos da configuração obtida, ou seja, requer uma relação entre proximidades e distâncias do tipo

$$d_{ij} = f(\delta_{ij}) + e_{ij} ,$$

onde f é uma função monotônica qualquer .

Surge então a necessidade de definir uma função, a qual possa fornecer uma medida da bondade de ajuste (ou maldade de ajuste) das distâncias do espaço de configuração para com as proximidades observadas e, consequentemente, uma medida do ajuste da configuração obtida aos dados.

Esta função deve assumir o valor zero quando o modelo das distâncias do espaço de configuração ajustar-se perfeitamente ao modelo das proximidades e deve assumir valores maiores quando o ajuste vier a ser menos perfeito.

A função proposta, inicialmente, para medir o ajuste de qualquer configuração, foi a soma de quadrados de desvios, definida por

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} ( d_{ij} - f(\delta_{ij}))^2 .$$

Uma vez que, a monotonicidade nem sempre é obtida na configuração, define-se os números $\hat{d}_{ij} = f(\delta_{ij})$, os quais são monotonicamente relacionados com as proximidades observadas $\delta_{ij}$.

Então, tem-se

$$d_{ij} = \hat{d}_{ij} + e_{ij}$$

e a medida do ajuste vem a ser

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} ( d_{ij} - \hat{d}_{ij} )^2 .$$

Embora esta medida seja invariante com respeito a translações, reflexões e rotações, não é invariante com respeito a extensões e contrações uniformes.

Introduziu-se, então, um fator de escala, o qual não só normaliza a medida como também a torna invariante com respeito a mudanças de escala. Este fator de escala é definido por

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} ( d_{ij} )^2 .$$

A função que mede o ajuste de qualquer configuração pode ser obtida, então, por

$$\left[ \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} ( d_{ij} - \hat{d}_{ij} )^2 \; / \; \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} ( d_{ij} )^2 \right].$$

Kruskal propôs uma medida de adequação de ajuste definida por

$$S = \left[ \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} ( d_{ij} - \hat{d}_{ij} )^2 \; / \; \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} ( d_{ij} )^2 \right]^{1/2},$$

a qual ele chamou de "stress".

O "stress" é considerado uma função das ( **n** x **k** ) coordenadas dos pontos da configuração obtida e, sendo assim, o valor do "stress" depende da configuração obtida. Sabe-se que o valor do "stress" aumenta quando **n** aumenta e/ou quando **k** diminui.

Pode-se dizer, ainda, que o "stress" representa a extensão para a qual os postos de ordenação dos $d_{ij}$'s não combinam com os postos de ordenação dos $\delta_{ij}$'s e, sendo assim, quando os postos dos $d_{ij}$'s e dos $\delta_{ij}$'s combinam perfeitamente, o valor do "stress" é zero.

Para certos tipos de dados, o "stress" é obtido pela expressão:

$$S' = \left[ \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} ( d_{ij} - \hat{d}_{ij} )^2 \; / \; \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} ( d_{ij} - \bar{d} )^2 \right]^{1/2},$$

onde $\bar{d}$ é a média aritmética dos $d_{ij}$'s .

Kruskal sugere que o valor do "stress" obtido por S, cujo denominador é $\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} ( d_{ij} )^2$, seja informalmente interpretado de acordo com o seguinte padrão:

| Valor do "Stress" | Adequação de Ajuste |
|---|---|
| 0.20 | Ruim |
| 0.10 | Regular |
| 0.05 | Bom |
| 0.025 | Excelente |
| 0.00 | Perfeito |

Como o valor do "stress" obtido por S', cujo denominador é $\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} ( d_{ij} - \bar{d} )^2$, é usualmente duas vezes o tamanho daquele obtido por S, para mesma configuração, Kruskal sugere que esse seja interpretado como segue:

| Valor do "Stress" | Adequação de Ajuste |
|---|---|
| 0.40 | Ruim |
| 0.20 | Regular |
| 0.10 | Bom |
| 0.05 | Excelente |
| 0.00 | Perfeito |

### 3.3.2 - Obtenção da Solução

Em escalonamento multidimensional não-métrico nenhuma solução analítica é possível. O procedimento usado para obter a solução é um procedimento iterativo, o qual é aplicável tanto a medidas de similaridade quanto a medidas de dissimilaridade.

Procedimento Iterativo:

Para n objetos obtém-se a informação inicial, a qual é o posto de ordenação dos $m = n(n-1)/2$ valores de similaridades, $\delta_{ij}$'s, a respeito de todos os pares de objetos. Supondo que não existe empates, os m valores de similaridades são arranjados em uma ordenação estritamente ascendente como

$$\delta_{i_1j_1} < \delta_{i_2j_2} < \ldots < \delta_{i_mj_m}$$

onde $\delta_{i_1j_1}$ é a menor das similaridades e o subscrito $i_1j_1$ indica o par de objetos que são menos similares, isto é, o par de objetos com posto 1 na ordenação das similaridades.

Somente a ordenação destas medidas de similaridade é preservada na solução métrica obtida pelo escalonamento não-métrico.

Os objetos são então representados por pontos em um espaço euclidiano k-dimensional e $X_i$ denota o vetor de coordenadas do ponto correspondente ao i-ésimo objeto, sendo $i = 1,\ldots,n$. Esta representação por pontos em um espaço euclidiano k-dimensional é denominada configuração inicial, a qual pode ser uma configuração aleatória de pontos ou uma configuração obtida a partir de uma análise de coordenadas principais.

A configuração inicial deve obedecer duas condições:

(i) $p_i \neq p_j$ , $\forall$ $i \neq j$, sendo $p_i$ e $p_j$ pontos da configuração.

(ii) X não pode estar contida em $\mathbb{R}^t$, tal que $t < k$ sendo k a dimensão do espaço da configuração procurada.

Qualquer que seja a configuração inicial, esta deve ser normalizada, isto é, seu centro de gravidade deve ser localizado na origem do sistema e a soma das distâncias de seus pontos à origem do sistema deve ser igual a 1, ou seja,

$$\sum_{i=1}^{n} \sum_{t=1}^{k} X_{it} = 1$$

onde $X_{it}$ é a coordenada de $p_i$ na dimensão $t$. Uma função de distância, $d(x_i, x_j) = d_{ij}$, é definida neste espaço euclidiano $k$-dimensional. Por simplicidade a conhecida distância euclidiana não-ponderada,

$$d_{ij} = \left[ (X_i - X_j)'(X_i - X_j) \right]^{1/2}$$

é assumida. Calcula-se então a distância euclidiana entre todos os pares de pontos da configuração inicial.

O propósito de qualquer procedimento de escalonamento não-métrico é encontrar uma série de pontos em um espaço de dimensão mínima, tal que as (dis)similaridades dadas entre os objetos são uma função monótona das distâncias euclidianas entre os pontos que representam os objetos.

Se as similaridades estiverem na ordem $\delta_{i1j1} < \delta_{i2j2} < \ldots < \delta_{imjm}$, então a monotonicidade é dita como sendo perfeita se as correspondentes distâncias euclidianas na solução estiverem na ordem $d_{i1j1} \geq d_{i2j2} \geq \ldots \geq d_{imjm}$.

Quando estivermos trabalhando com dissimilaridades, ao invés de similaridades, e estas estiverem ordenadas como $\delta^*_{i1j1} < \delta^*_{i2j2} < \ldots < \delta^*_{imjm}$, onde $\delta^*_{i1j1}$ é a menor das dissimilaridades, então a monotonicidade será perfeita se as correspondentes distâncias euclidianas estiverem na ordem $d_{i1j1} \geq d_{i2j2} \geq \ldots \geq d_{imjm}$.

Em outras palavras, a ordenação ascendente das similaridades deve implicar a ordenação descendente das

distâncias euclidianas, que deve ser exatamente análoga a ordenação descendente das dissimilaridades.

A completa conformidade para a monotonicidade pode ser adquirida com uma representação em um espaço de dimensionalidade alta, como $k \geq n-1$, sendo que se tem $n$ objetos no total. Mas o propósito do escalonamento multidimensional é encontrar uma representação de baixa dimensão, na qual a conformidade para a monotonicidade possa ser pelo menos razoável, se não puder ser perfeita.

A partir das medidas de $d_{ij}$ estima-se os valores ajustados $\hat{d}_{ij}$, que estão monotonicamente relacionadas com as medidas de similaridade. Os $\hat{d}_{ij}$ são apenas números reais (também referidos como pseudo-distâncias ou disparidades ou discrepâncias) e, sendo assim, não existe uma configuração de pontos cujas distâncias entre os pontos sejam $\hat{d}_{ij}$. Os $\hat{d}_{ij}$ são definidos como mínimos quadrados ajustados às distâncias euclidianas $d_{ij}$ e são obtidos por um método de regressão designados para produzir distâncias monotônicas ajustadas, conhecido como regressão monotônica ou isotônica.

Os valores ajustados $\hat{d}_{ij}$ devem ser monotonicamente relacionados aos $\delta_{ij}$, ou seja, se $\delta_{i1j1} < \delta_{i2j2} < \ldots < \delta_{imjm}$, então, $\hat{d}_{i1j1} \geq \hat{d}_{i2j2} \geq \ldots \geq \hat{d}_{imjm}$.

Obtidas as distâncias euclidianas, $d_{ij}$, entre os pontos da configuração e os valores ajustados $\hat{d}_{ij}$, pode-se calcular o valor do stress, que é a medida de adequação de ajuste proposta por Kruskal para medir o ajuste da configuração, definida por

$$S = \left[ \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} ( d_{ij} - \hat{d}_{ij} )^2 \; / \; \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij}^2 \right]^{1/2}$$

O stress pode então ser usado como base para um método sistemático de obtenção dos valores $\hat{d}_{ij}$ que minimizam o stress sujeito a condição de serem monótonos em relação aos valores das similaridades observadas $\delta_{ij}$'s.

O valor do stress depende da configuração e

sendo assim pode-se escrever $S(X_1, X_2, \ldots, X_n)$, onde $X_i$ é o vetor de coordenadas do ponto correspondente ao i-ésimo objeto.
Se o stress não for suficientemente pequeno então a posição dos pontos no espaço da configuração devem ser alterados a fim de reduzir o valor do stress.
Calcula-se então um fator de correção para mover a configuração na direção do menor stress.
Mover a configuração na direção do menor stress consiste em mover os pontos na direção dada pelo vetor das primeiras derivadas parciais do stress com respeito a configuração **X**, isto é, na direção dada pelo vetor gradiente. O gradiente consiste das derivadas parciais do stress com respeito a configuração **X**, e é obtida por:

$$g_{ip} = \partial S / \partial X_{ip} = (1/NS) \sum_{j=1}^{n} [ (1 - d_{ij} / \hat{d}_{ij}) - S^2] (x_{ip} - x_{jp})$$

onde $N = \sum d_{ij}$ é o fator que normaliza a expressão, e **p** é o número de dimensões da configuração. Se o gradiente for zero, então um possível mínimo local foi alcançado.

A configuração é reconstruída de forma a minimizar o stress, e desta forma, iterativamente novas configurações são obtidas, das quais novos $d_{ij}$ e $\hat{d}_{ij}$ originam-se, o stress é recalculado e este processo é repetido até que um suficientemente bom ajuste é obtido ou até que seja obtida uma configuração que não possa ser melhorada, a qual é chamada configuração de stress mínimo. O método utilizado para minizar o stress mínimo, geralmente, é o 'steepest descent'.

### 3.3.3 - Diagrama de Dispersão

O diagrama de dispersão é um gráfico construído, primeiramente, para verificar se a relação exigida entre as distâncias $d_{ij}$ entre pares de pontos da configuração e as medidas

de proximidades observadas entre os pares de objetos é obedecida na solução obtida e, secundariamente, para obter uma visualização do ajuste da configuração aos dados, uma vez que o valor numérico do ajuste foi obtido através do stress.

Um diagrama de dispersão pode ser obtido através da construção de um gráfico cujas coordenadas são $d_{ij}$ (distâncias entre os pares de pontos da configuração) e $\delta_{ij}$ (proximidades entre os pares de objetos). O diagrama de dispersão deve apresentar um ponto para cada par de objetos possível.

No escalonamento não-métrico supõe-se que $d_{ij} = f(\delta_{ij})$ onde f é uma função monótona crescente ou decrescente.

Qualquer que seja a suposição feita com relação a f, o diagrama de dispersão pode mostrar que os dados sugerem outro tipo de função, que não a suposta.

Se os pontos do diagrama de dispersão mostrarem claramente uma curva diferente daquela referente a função suposta, o valor do stress poderá estar indevidamente aumentado pela suposição inadequada com relação a f e neste caso, é necessário analisar os dados novamente fazendo uma suposição mais adequada com relação a função f.

O diagrama de dispersão pode mostrar se existe 'degeneração', ou seja, se os pontos da configuração estão fortemente agrupados, ou se a maioria dos pontos estão sobre ou próximos de um número muito pequeno de posições.

A degeneração pode ocorrer pelo fato dos objetos terem um agrupamento natural em três ou menos grupos, ou pelo fato de as dissimilaridades entre os objetos dos diferentes grupos serem todas ou quase todas maiores do que as dissimilaridades entre os objetos de um mesmo grupo. Neste caso todos ou quase todos os pontos que representam os objetos de um mesmo grupo convergem para a mesma posição e o stress converge para valores próximos de zero, apesar de não atingir o zero na prática.

Em particular o stress muito pequeno não deve ser tomado como indicador de um bom ajuste.

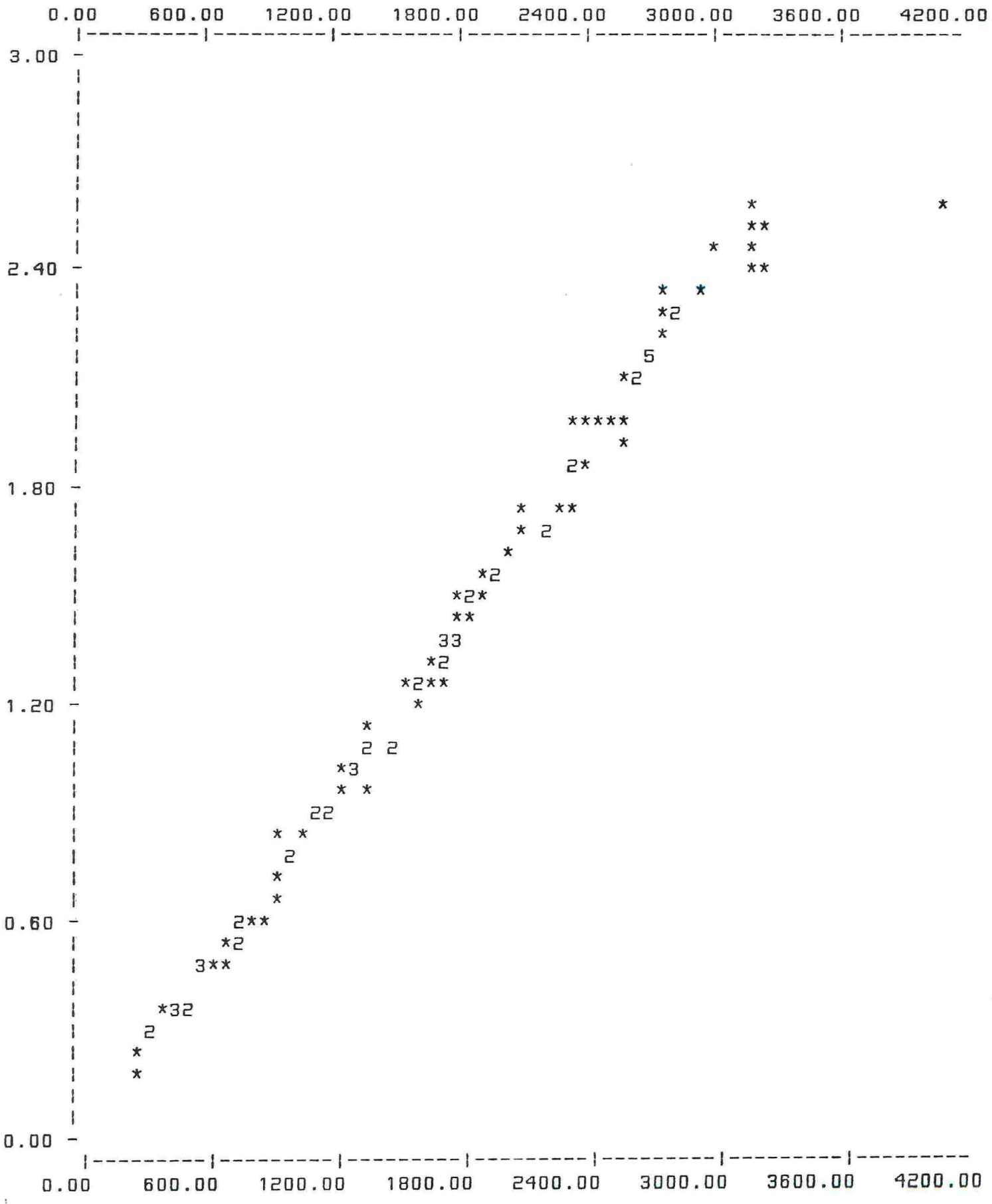
Nos casos em que ocorrer degeneração uma nova análise, para cada grupo separadamente deve ser feita ou uma análise pelo

EMD métrico.

Considere como exemplo o diagrama de dispersão construído para a configuração bidimensional obtida pelo escalonamento não-métrico para os dados da tabela 3.1, apresentado na figura 3.3.

Observando o diagrama de dispersão da figura 3.3, nota-se que este sugere a existência de uma relação monotônica crescente entre as dissimilaridades originais e as distâncias do espaço de configuração. Sendo assim, pode-se concluir que a relação monotônica entre $\delta_{ij}$ e $\mathbf{d}_{ij}$ assumida pelo escalonamento não-métrico é obedecida na solução obtida e que a configuração se ajusta perfeitamente aos dados.

FIGURA 3.3 - DIAGRAMA DE DISPERSÃO DA SOLUÇÃO OBTIDA PELO ESCALONAMENTO NÃO-MÉTRICO PARA OS DADOS DA TABELA 3.1



## 3.4 - ANÁLISE DA SOLUÇÃO

Analisar os resultados obtidos através de uma análise de escalonamento multidimensional é uma tarefa não muito simples.

Poder interpretar os resultados de uma análise de escalonamento implica em escolher um número de dimensões apropriado para representar os dados. Neste caso serão apresentados inicialmente alguns procedimentos os quais podem orientar a escolha do número de dimensões apropriado.

Finalmente serão apresentados alguns procedimentos os quais podem ser utilizados para obter uma melhor interpretação da configuração no número de dimensões escolhido.

### 3.4.1 - Escolha do Número de Dimensões

O principal objetivo de quem utiliza uma técnica de escalonamento multidimensional, geralmente, é encontrar a configuração que melhor representa os dados em estudo e que permite uma melhor compreensão dos dados.

Na prática, a verdadeira dimensão do espaço é desconhecida e parte-se então em procura de um número de dimensões, que seja apropriado para representar os dados.

Sabe-se que quanto maior o número de dimensões tanto menor é o valor do stress e que, tendo-se **n** objetos, uma configuração com $p \geq n-1$ dimensões terá stress zero. Entretanto, isto não quer dizer que a configuração que melhor representa os dados seja aquela cujo valor do stress é menor do que de qualquer outra. O que se procura é uma configuração com dimensionalidade apropriada para representar os dados e não uma configuração que tenha o menor valor do stress.

Obviamente, o melhor seria que duas dimensões fossem apropriadas para representar os dados, mas nem sempre duas

dimensões são suficientes e uma configuração em maior número de dimensões é então requerida.

Uma vez que o escalonamento multidimensional é quase sempre usado como um modelo descritivo para interpretar e representar os dados, várias considerações devem ser levadas em conta quando se está decidindo acerca da dimensionalidade do espaço a ser usado, como por exemplo, a facilidade de interpretação da solução, o ajuste de modelo aos dados, etc. Qualquer informação que se tenha a respeito dos dados deve ser levada em consideração.

Existem alguns procedimentos intuitivos, os quais fornecem uma orientação sobre o número de dimensões que deve ser usado para melhor representação e interpretação dos dados.

Um dos procedimentos usa o valor do "stress" como guia. Este consiste em realizar análises de escalonamento usando diferentes números de dimensões; Construir um gráfico cujas coordenadas são **k** (número de dimensões usada) e S (valor do "stress" para a configuração **k**-dimensional obtida); observar em que ponto do gráfico se forma um 'cotovelo' ou 'ângulo' e procurar as coordenadas correspondentes a este ponto, as quais são o número de dimensões apropriado para representar os dados e o valor do "stress" para a configuração que tem este número de dimensões.

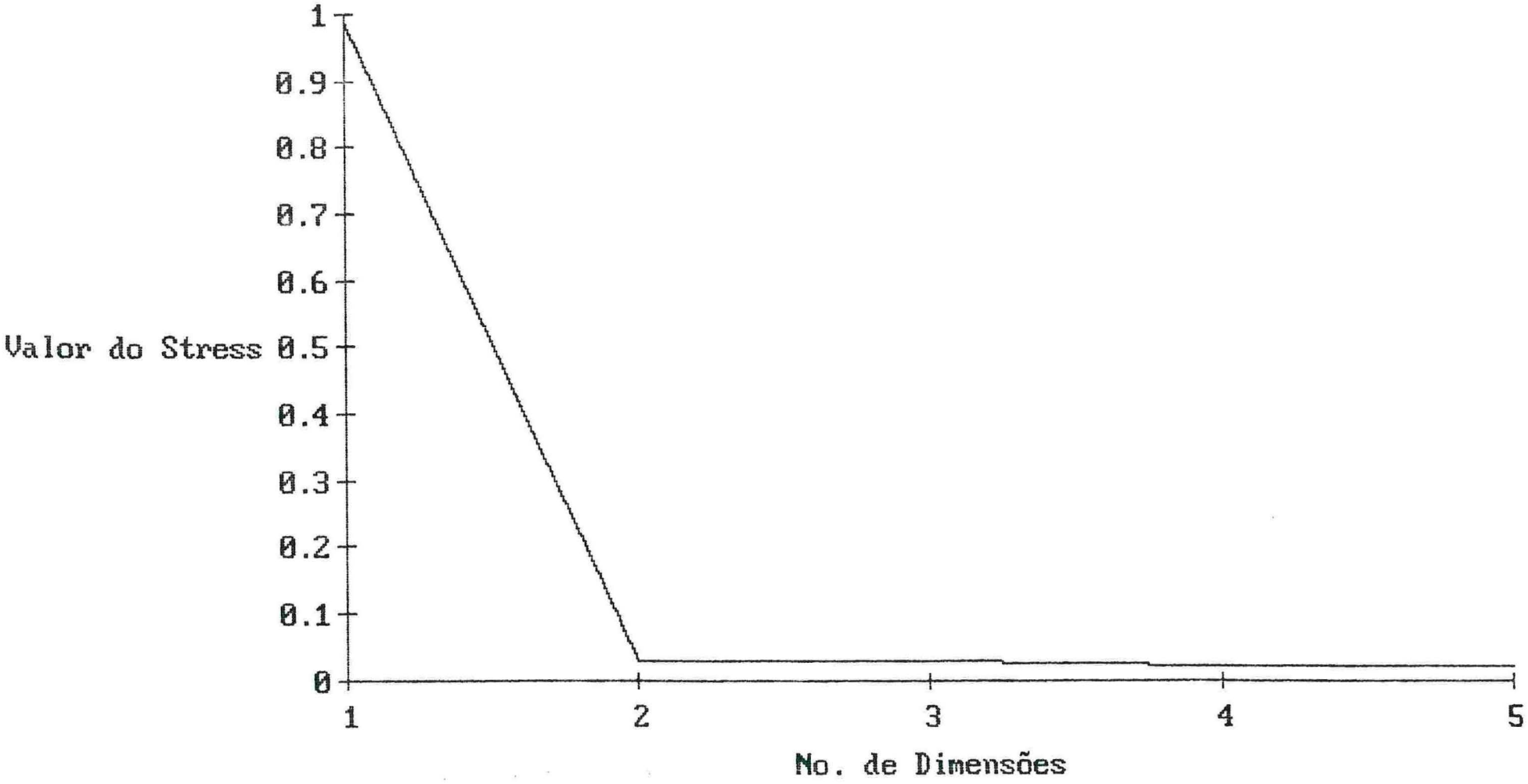
Como exemplo, considere que foram realizados análise de escalonamento multidimensional não-métrico de 1 a 5 dimensões usando as distâncias da tabela 3.1. O gráfico do número de dimensões, **k**, pelo valor so stress, S, é construido para este exemplo e é apresentado na figura 3.4.

Observando o gráfico pode-se dizer que forma-se um 'cotovelo' no ponto cujas coordenadas são **k** = 2 e S = 0.03.

Nota-se que o valor do stress diminui rapidamente de **k** = 1 para **k** = 2 e que para **k** > 2 diminui muito lentamente, e é exatamente isto que faz com que se forme um 'cotovelo' ou um 'ângulo' no gráfico.

Conclui-se então que uma configuração de 2 dimensões é apropriada para representar estes dados e que o stress desta configuração é 0.03.

Figura 3.4 - Gráfico do Stress em função da Dimensão do Espaço para os Dados da Tabela 3.1



Outro procedimento intuitivo, que pode e deve ser usado para escolher o número de dimensões apropriado, usa como guia de interpretabilidade da solução. Este procedimento leva em consideração a facilidade com que os eixos podem ser interpretados. Se uma solução em **p** dimensões permite uma satisfatória interpretação dos eixos e uma solução em **p** + 1 dimensões não permite uma interpretação dos eixos ou não fornece nenhuma informação adicional a respeito dos dados, escolhe-se uma solução em **p** dimensões.

Por outro lado, pode-se ter uma configuração em **p** dimensões sem interpretação, mas em **p** + 1 dimensões teríamos uma interpretação total, sendo portanto esta a melhor solução.

Soluções em apenas 2 dimensões são, geralmente, preferíveis por serem facilmente representadas graficamente e por serem facilmente interpretadas, mas nem sempre revelam a 'estrutura' escondida nos dados e nem sempre permitem uma interpretação satisfatória.

Ainda outro procedimento que pode ser usado o qual leva em consideração a estabilidade da solução consiste em fazer novas análises dos dados, ou seja, divide-se os sujeitos aleatoriamente em sub-grupos e aplica-se o EMD a cada sub-grupo em separado.

Os procedimentos apresentados são os mais simples e que pode ser facilmente realizados, mas existem outros procedimentos mais complexos para determinar o número de dimensões apropriado para representar os dados (ver Gnanadesikan, 1977).

### 3.4.2 - Interpretação da Configuração

Após ter-se determinado o número de dimensões apropriado para representar os dados e obtido a configuração com a dimensionalidade apropriada, parte-se para a interpretação desta configuração, que é uma das etapas mais importantes da análise de escalonamento.

Um dos métodos que pode ser usado para interpretar uma solução de escalonamento multidimensional é identificar, ou

procurar um significado particular para as dimensões (ou eixos) da configuração.

Para algumas configurações a interpretação das dimensões pode ser facilmente obtida simplesmente observando estas configurações, mas nem sempre isto é possível e, nestes casos, existem alguns recursos que podem ser utilizados a fim de facilitar a interpretação da solução. A determinação do recurso que deve ser usado dependerá do tipo de dados a serem analisados e do tipo de configuração obtida.

Configurações obtidas pelos métodos de escalonamento multidimensional a 'dois fatores', apresentados neste capítulo, podem ser submetidas a translações, rotações ou reflexões, e usando estes recursos é possível que se encontre novos e mais significativos eixos, os quais permitam uma melhor interpretação da solução obtida.

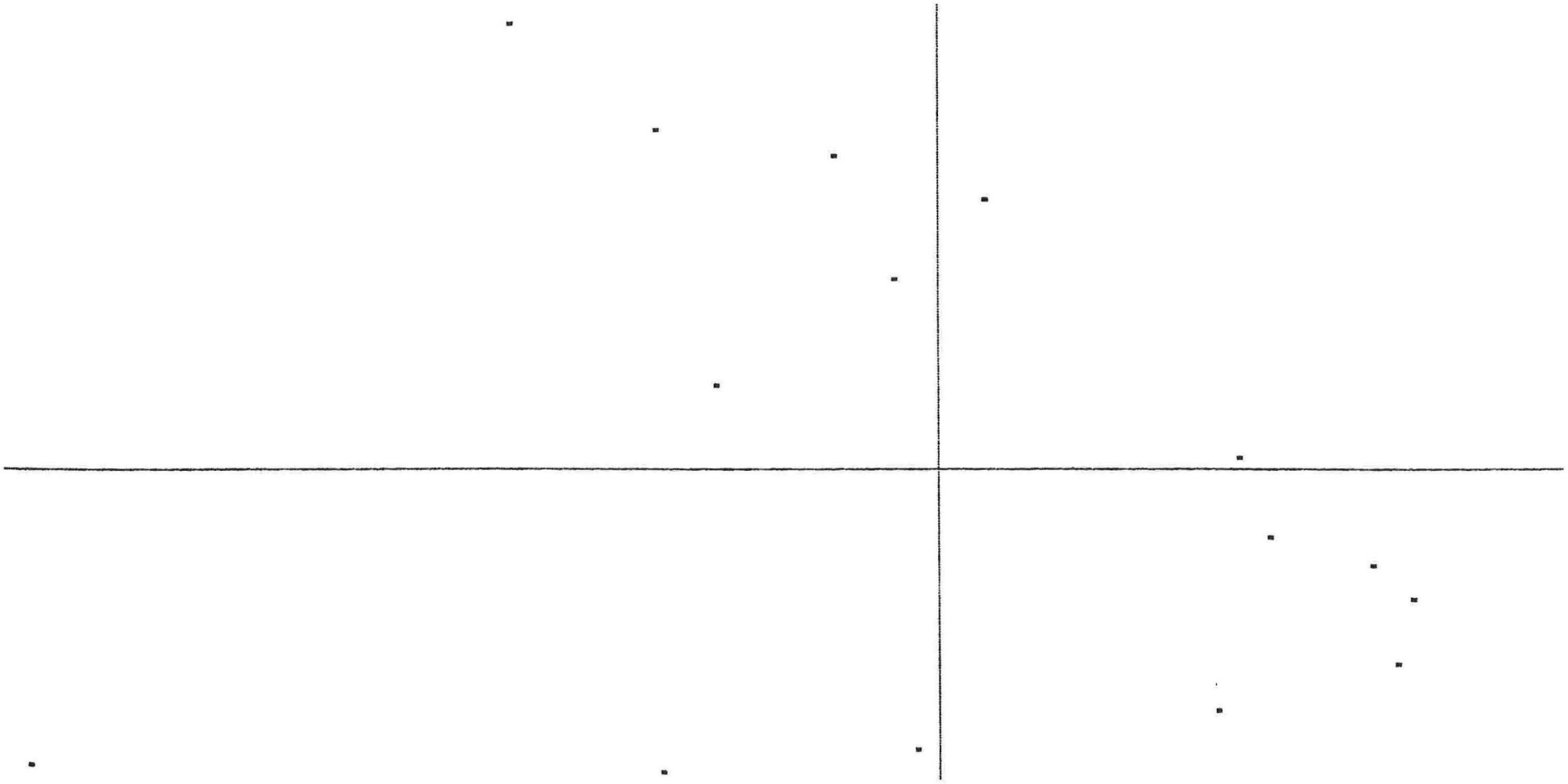
Como exemplo, considere a figura 3.5, a qual apresenta a configuração de pontos obtida pelo método de escalonamento não-métrico para os dados da tabela 3.1. Não se pode dizer que ela representa uma reconstrução do mapa do Brasil, mas se submetermos esta configuração a uma reflexão obtém-se a configuração apresentada na figura 3.2, a qual representa uma perfeita reconstrução do mapa do Brasil.

Contudo, é difícil observar a configuração em todas as direções e não só naquelas definidas pelos eixos coordenados, e numa única configuração pode haver mais direções interpretáveis do que aquelas definidas pelos eixos.

Quando se tem uma configuração em três ou mais dimensões, por exemplo, a interpretação da configuração através da observação visual torna-se difícil. Para suprimir as falhas da observação pode-se usar um procedimento estatístico de regressão linear, quando forem conhecidas algumas das características do conjunto de objetos. Uma interpretação melhorada das dimensões pode, ainda, ser obtida usando as características (variáveis) que melhor distinguem os objetos que se localizam nos extremos opostos das várias direções dos planos da configuração obtida.

Outro método que pode ser utilizado para interpretar a solução obtida, o qual é geralmente preferido, consiste em

Figura 3.5 - Solução obtida pelo Escalonamento Não-Métrico para os Dados da Tabela 3.1



identificar grupos de objetos e evitar falar de dimensões.

A diferença básica entre interpretação das dimensões e interpretação por vizinhanças, que consiste em identificar grupos de objetos que aparecem na configuração, é que a interpretação das dimensões se baseia fundamentalmente em grandes distâncias, enquanto a interpretação por vizinhanças se baseia em pequenas distâncias ou grandes similaridades.

Finalmente, interpretação é muitas vezes auxiliada por comparações de uma configuração com outras configurações obtidas pelo escalonamento multidimensional para a mesma série de dados.

## 3.5 - COMPARAÇÃO DOS MÉTODOS

É interessante fazer uma comparação entre os métodos de escalonamento, apresentados nas seções anteriores deste capítulo (método clássico e método não-métrico), sob alguns aspectos, como por exemplo, facilidade dos cálculos, robustez, qualidade dos resultados, etc.

Pode-se dizer, obviamente, que a obtenção da solução é mais simples, no método clássico do que no método não-métrico. Isto se deve à complicada natureza do procedimento iterativo utilizado para obter a solução no escalonamento não-métrico o que exige uma quantidade de cálculos muito maior do que os envolvidos na obtenção da solução clássica.

Não é conhecido quanto robusto é o método clássico para transformações monótonas de funções de distâncias, entretanto, sabe-se que ambos os métodos dão resultados similares quando os dados consistem de distâncias aproximadamente ou exatamente euclidianas, e que o escalonamento não-métrico produz resultados satisfatórios, enquanto o escalonamento clássico produz resultados não satisfatórios, quando distâncias não euclidianas são usadas. Segundo Chatfield (1980) estudos foram realizados na

Universidade de Bath para comparar os dois métodos e concluiram que o escalonamento não-métrico é geralmente melhor que escalonamento métrico.

Como exemplo de comparação dos métodos, considere os mapas do Brasil reconstruídos pelo escalonamento clássico (Figura 3.1) e pelo escalonamento não-métrico (Figura 3.2), a partir das distâncias entre algumas capitais brasileiras (dadas na tabela 3.1).

Observando as posições de cada uma das capitais brasileiras, em cada uma das reconstruções do mapa do Brasil, e comparando com as posições das mesmas no mapa do Brasil, pode-se dizer que a reconstrução do mapa obtido pelo escalonamento não-métrico é mais perfeita do que a obtida pelo escalonamento clássico.

Este fato confirma o que foi dito anteriormente a respeito do escalonamento não-métrico obter melhores resultados do que o escalonamento clássico quando distâncias não euclidianas são usadas, uma vez que neste exemplo da reconstrução do mapa do Brasil foram usadas distâncias aéreas entre algumas capitais brasileiras, as quais são distâncias não-euclidianas.

# CAPÍTULO 4

## APLICAÇÃO

Neste capítulo apresentamos uma aplicação de Escalonamento Multidimensional na área de Geologia, na qual foram usados dados originados de análises químicas vulcânicas da região ocidental do Rio Grande do Sul e Santa Catarina, cedidos por Jaqueline Chies, do Instituto de Geologia da UFRGS.

Um dos objetivos desta aplicação é demonstrar com dados reais que o método não-métrico produz soluções equivalentes para medidas de proximidade que tenham a mesma ordenação. Um outro objetivo é verificar se o escalonamento não-métrico é capaz de produzir uma configuração que evidencie a estrutura dos dados, já conhecida de trabalhos anteriores feitos nesta área, ou ainda se a configuração obtida sugere alguma outra estrutura que possa estar presente nos dados.

Os dados consistem de análises feitas sobre 116 amostras de rochas de locais diferentes (todas as amostras tem o mesmo peso) em termos de elementos químicos que estavam presentes na constituição destas rochas. Estes elementos químicos dividem-se em elementos maiores, os quais são medidos em percentagem (%) do peso em termos de óxido, e elementos traço, os

quais são medidos em partes por milhão (ppm), visto que estão presentes em menos de um por cento (1%) na composição das rochas.

Foram feitas medições de 20 elementos químicos, mas somente 15 elementos foram considerados na presente aplicação, uma vez que 5 dos 20 elementos químicos apresentavam observações perdidas. Desses 15 elementos considerados, 10 são medidos em percentagem do peso em termos de óxido ($SiO_2$, $TiO_2$, $Al_2O_3$, FeO, MnO, MgO, CaO, $Na_2O$, $K_2O$ e $P_2O_5$ ) e 5 são medidos em partes por milhão (Ni, Ba, Rb, Sr e Zr). Sendo assim, tem-se uma matriz de dados de dimensões (116 x 15), que apresenta as observações de 15 elementos químicos feitas para 116 amostras de rochas de locais diferentes.

Como os dados referentes aos elementos maiores e aos elementos traço apresentam-se em unidades de medida diferentes, as análises foram feitas com dados padronizados.

Os dados são quantitativos e, portanto, optou-se por usar, exploratoriamente, como medidas de proximidade a distância euclidiana, a distância euclidiana ao quadrado, a distância taxonômica média, a distância média de Manhattan, a correlação momento-produto e variância e covariância, as quais são apropriadas para este tipo de dados.

Mostraremos que o método não-métrico de escalonamento multidimensional produz resultados similares aos resultados obtidos pelo método clássico e ainda, que o método não-métrico produz resultados melhores do que o clássico, nos casos em que são usadas distâncias, as quais são não-euclidianas.

Como na presente aplicação serão usadas tanto medidas de distâncias euclidianas como não-euclidianas e, também, similaridades (o método métrico requer que similaridades sejam transformadas em distâncias), optou-se por aplicar o método não-métrico, mediante uso da rotina MDSCALE do programa NTSYS ( ver capítulo 5 ) que realiza o procedimento iterativo do método não-métrico de Kruskal, descrito no capítulo 3.

Sabe-se que as rochas se dividem em dois grupos devido ao teor de óxido de silício ($SiO_2$) presente na sua composição . Um dos grupos é formado por rochas básicas, o qual será referido como grupo de rochas do tipo 1, e o outro é formado

por rochas ácidas e será referido como grupo de rochas do tipo 2. As rochas do tipo 1 tem menos de 65% de SiO na sua composição, enquanto as rochas do tipo 2 tem mais do que 65% de SiO.

Observando a figura 4.1, a qual apresenta a quantidade de SiO presente na composição das amostras de rochas, pode-se identificar quais as amostras de rochas que pertencem ao grupo de rochas do tipo 1 e quais pertencem ao grupo de rochas do tipo 2. As amostras de rochas de números 31, 33, 43, 44, 47, 48, 58, 60, 64, 65, 66, 67, 94, 110, 111, 112, 113, 114, 115 e 116 são amostras de rochas do tipo 2 e as restantes são amostras de rochas do tipo 1.

Sabe-se também, que cada um destes grupos de rochas divide-se em dois novos grupos devido aos teores de óxido de titânio (TiO), óxido de fósforo ($P_2O_5$), estrôncio (Sr), zircônio (Zr) e bário (Ba) presentes na sua composição. Para saber a quantidade de cada um destes elementos presente na composição de cada uma das amostras de rochas, pode-se observar as figuras 4.2, 4.3, 4.4.

Há ainda outros gráficos no anexo que apresentam nitidamente a diferença das amostras de rochas do tipo 1 e 2, e é importante que esses sejam observados com atenção a fim de se identificar quais os elementos que discriminam melhor os grupos de rochas.

Serão apresentados a seguir os resultados obtidos através das análises feitas com matrizes de proximidades derivadas da matriz de dados padronizados.

Inicialmente, foram feitas análises de escalonamento usando as matrizes de distâncias euclidianas, distâncias euclidianas ao quadrado e distâncias taxonômicas médias. Obteve-se três configurações iguais para estas três matrizes de proximidades, como era de se esperar, visto que as medidas de proximidade usadas tinham a mesma ordenação. As três configurações tem um "stress" de 0.127, podendo ser consideradas como boas configurações, segundo o critério de avaliação de Kruskal. Pode-se observar na figura 4.5, que as amostras de rochas do tipo 1 estão todas próximas umas das outras, assim como as amostras de rochas do tipo 2, e que o grupo de amostras de

Figura 4.1 - Quantidade de SiO presente na composição das Amostras de Rochas

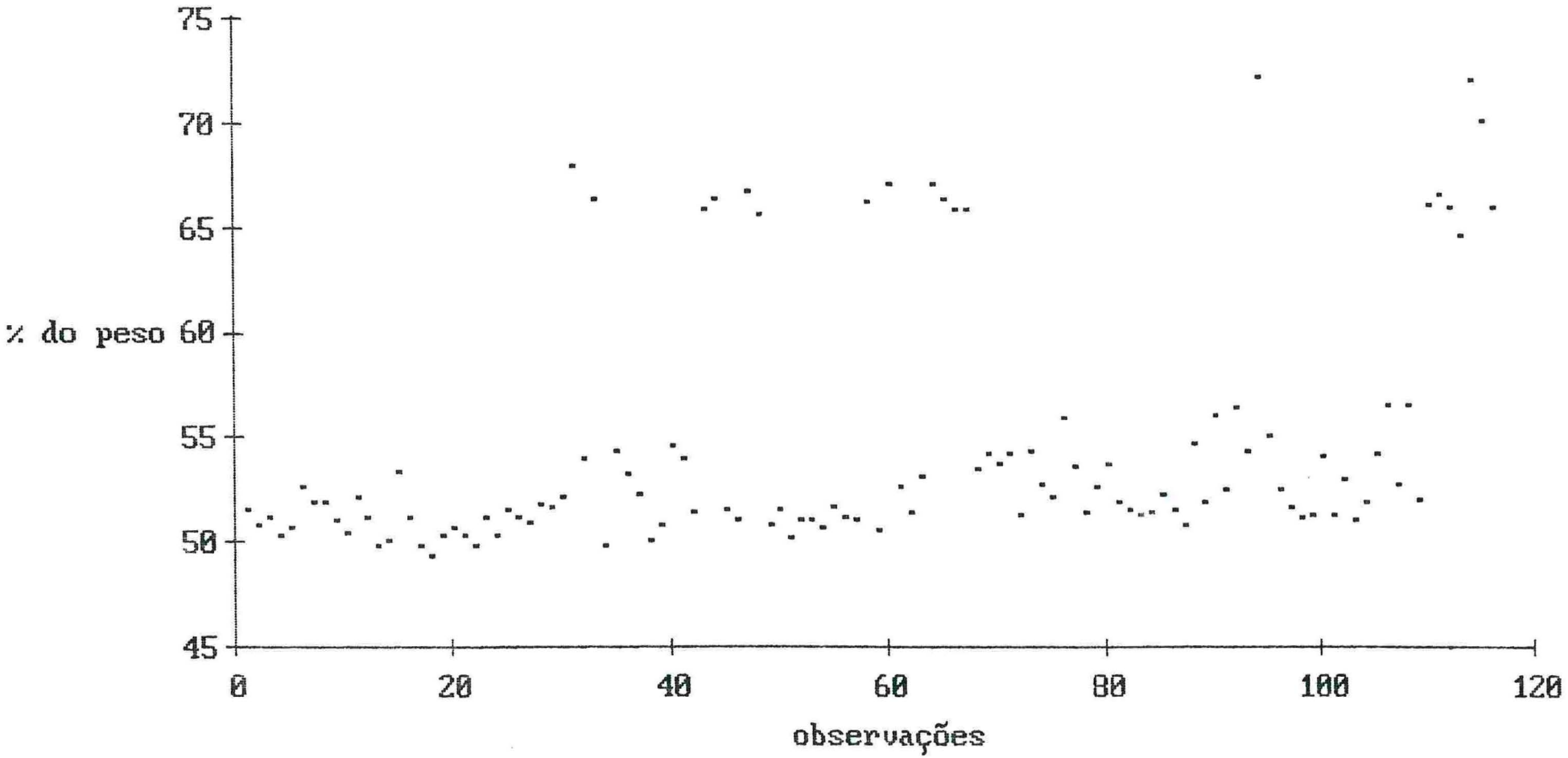


Figura 4.2 - Quantidade de TiO presente na composição das Amostras de Rochas

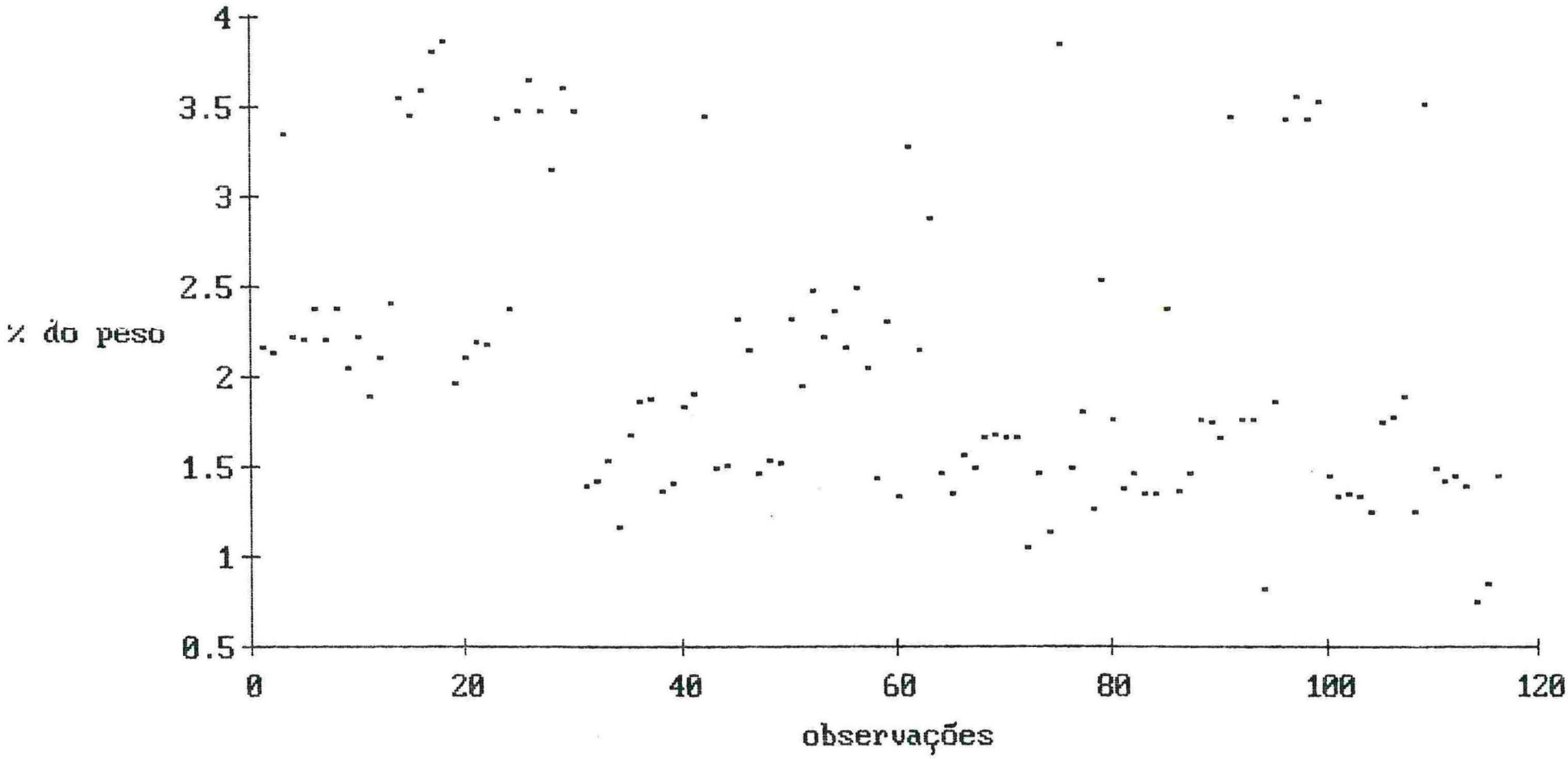


Figura 4.3 - Quantidade de Ba presente na composição das Amostras de Rochas

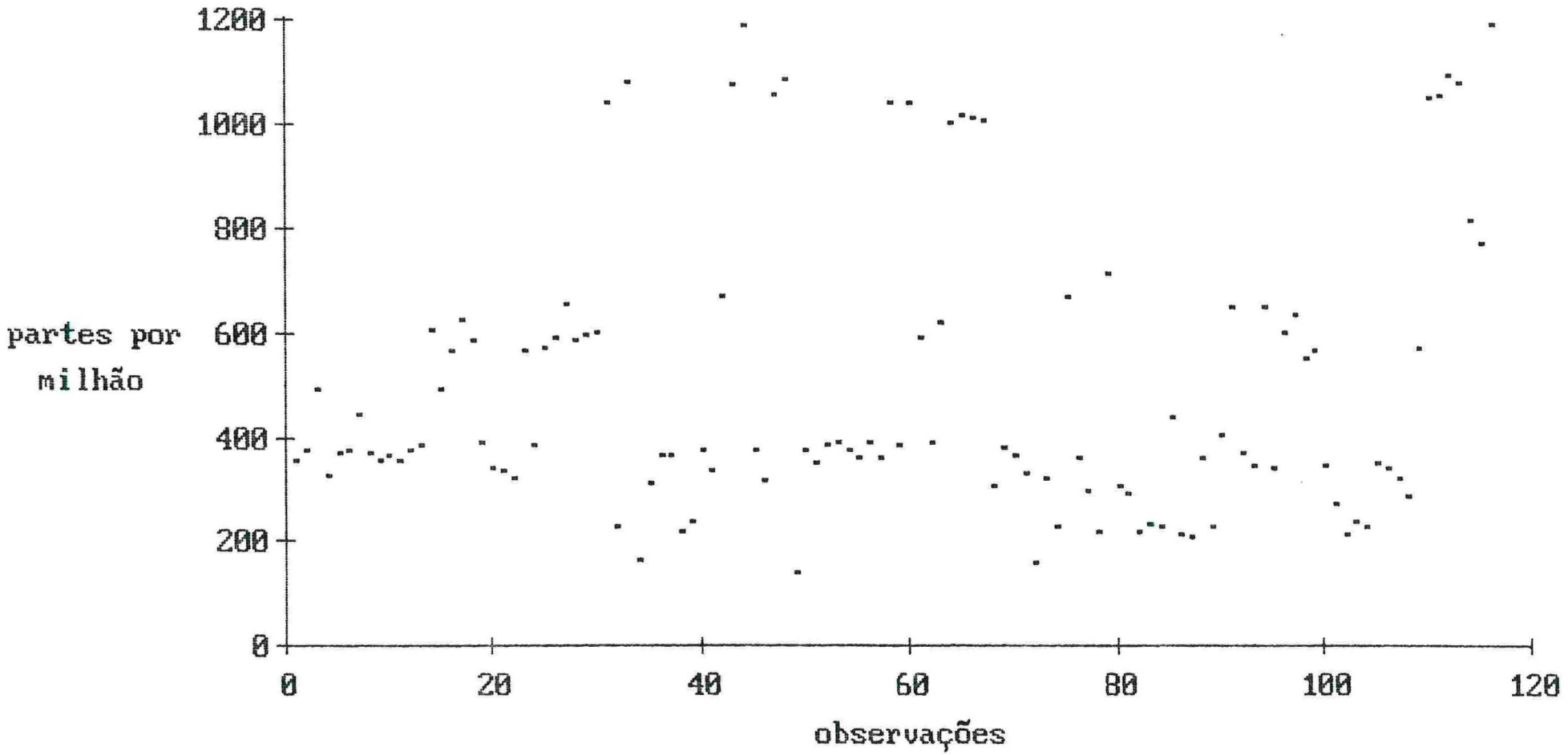


Figura 4.4 - Quantidade de Zr presente na composição das Amostras de Rochas

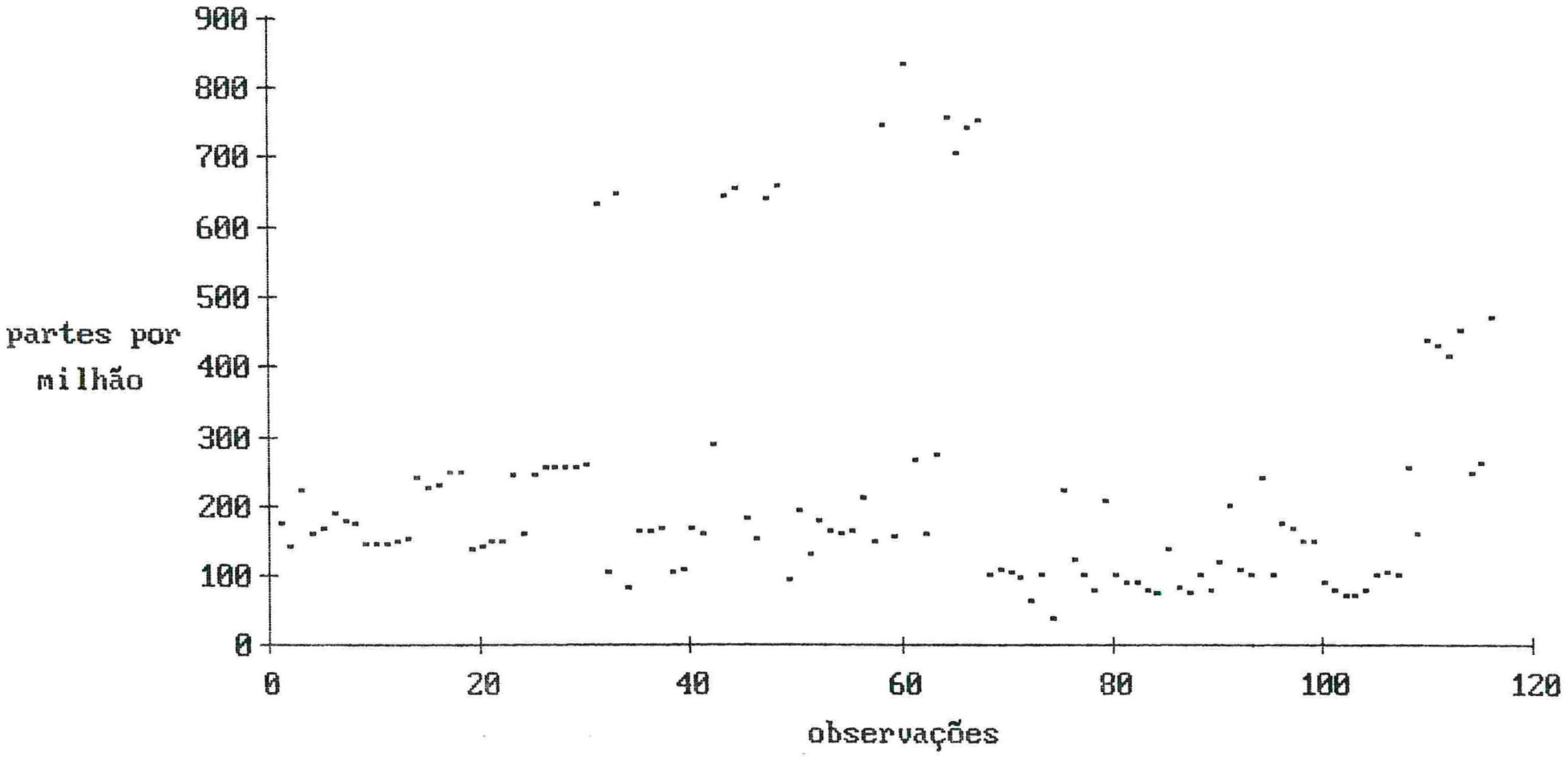
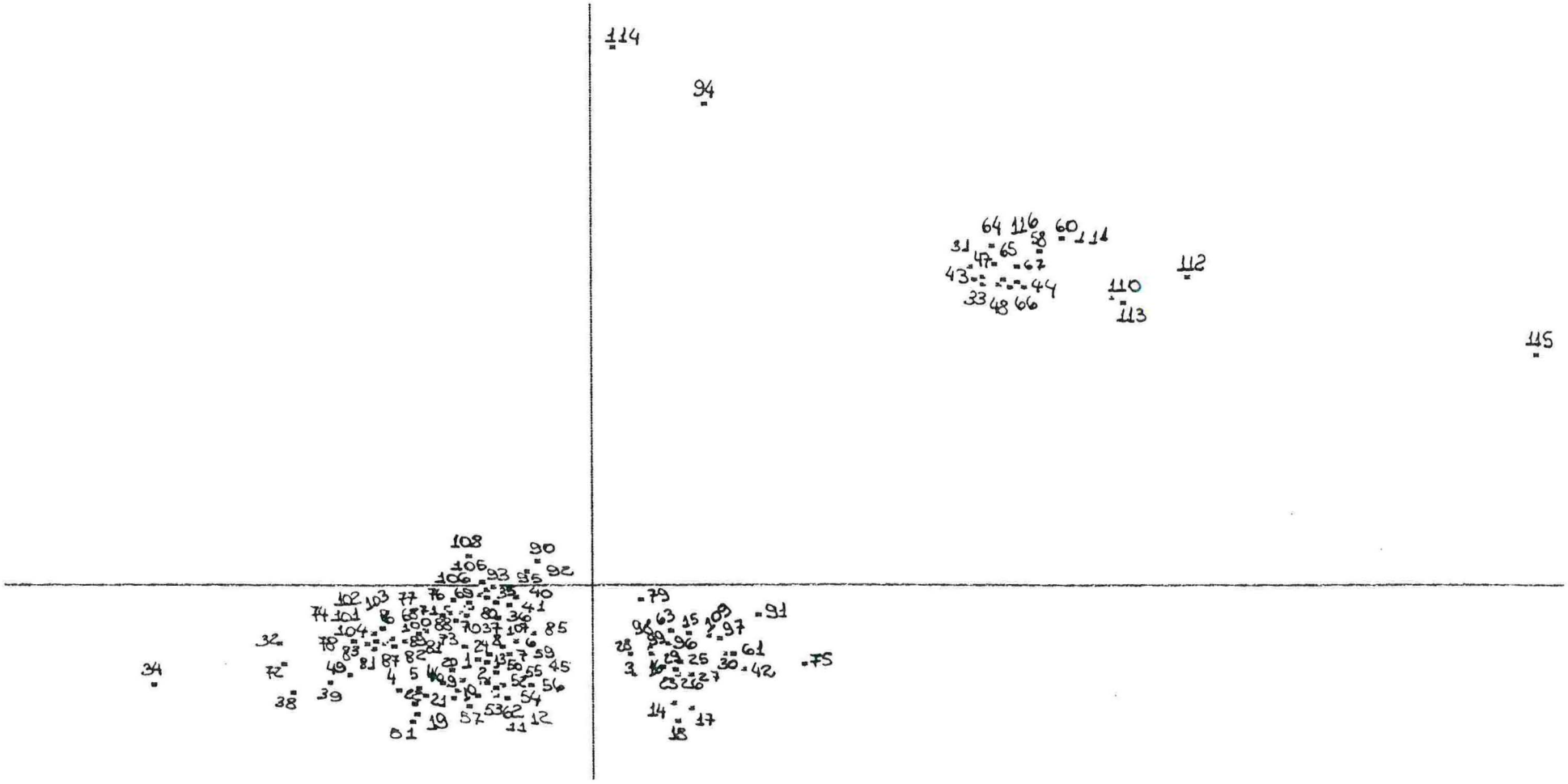


Figura 4.5 - Solução obtida pelo Escalonamento Não-Métrico
para a Matriz de Distâncias Euclidianas



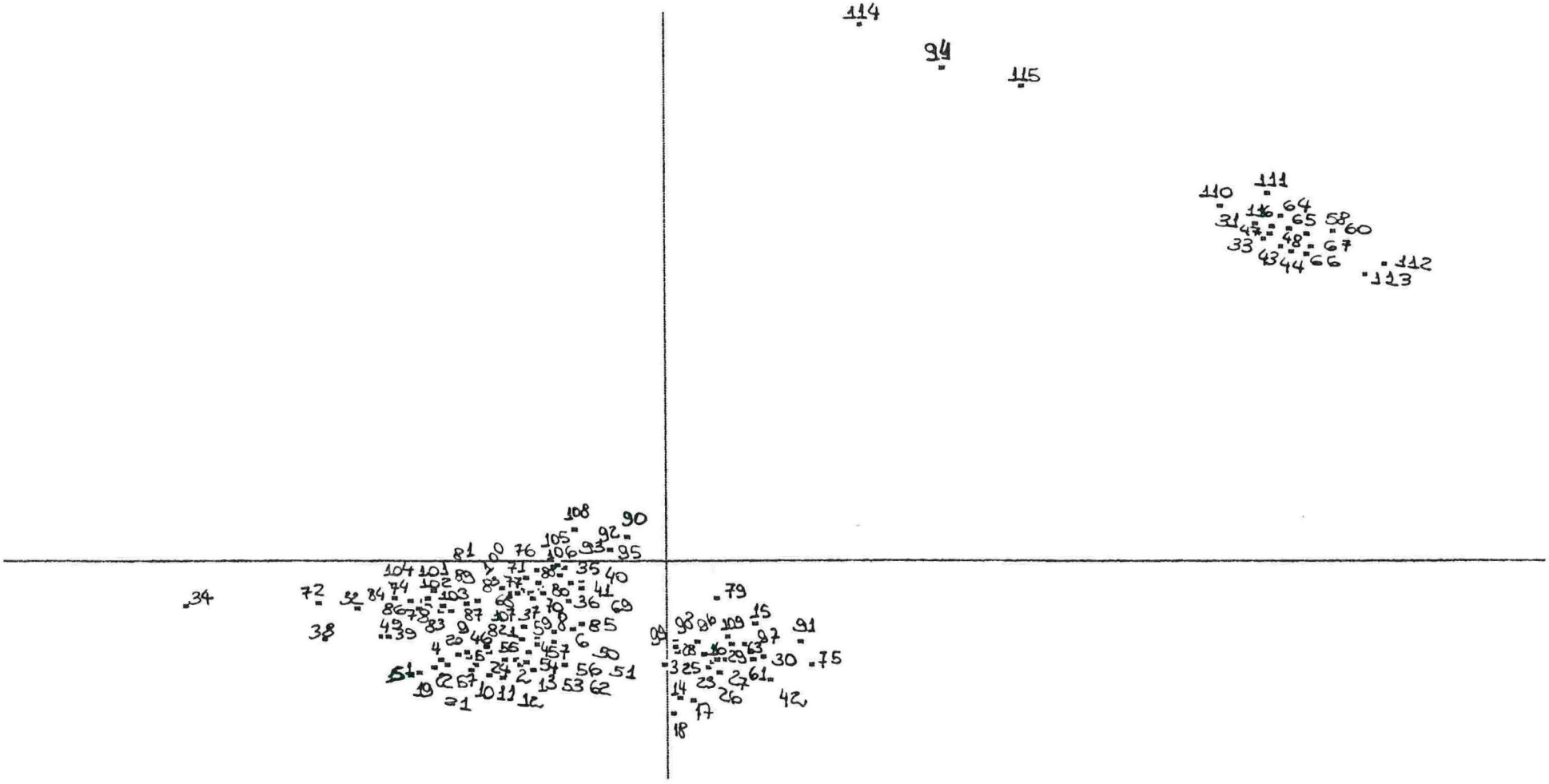
rochas do tipo 1 está distante do grupo de amostras de rochas do tipo 2. Nota-se que as amostras de rochas do tipo 1 podem ser divididas em dois novos grupos, como era esperado. Pode-se notar, também, que as amostras de rochas do tipo 2 de números 94 e 114 ficaram próximas uma da outra, mas isoladas, e que a amostra de rocha de número 115 também aparece distante de todas as outras amostras de rochas do tipo 2. Esperava-se que estas três amostras de rochas formassem um segundo grupo de amostras de rochas do tipo 2, mas como se tratam de apenas três amostras de rochas que se diferenciam das restantes , considerou-se que todas as amostras de rochas do tipo 2 formam um único grupo. Pode-se dizer, então, que as configurações obtidas a partir destas três matrizes de proximidades representam satifatoriamente a estrutura presente nos dados, mas não perfeitamente.

Utilizando-se a matriz de distâncias médias de Manhattan, para os dados padronizados, obteve-se uma configuração com um "stress" de 0.086 ( excelente ), que ao ser observada pode também ser considerada como uma excelente configuração ( ver figura 4.6 ). A configuração obtida aqui é muito semelhante à configuração obtida para a matriz de distâncias euclidianas, mas sugere que as amostras de rochas de números 94, 114 e 115 formam um segundo grupo de amostras de rochas do tipo 2, como era esperado, visto que estas três amostras de rochas do tipo 2 se diferenciam das outras amostras de rochas do tipo 2. Em razão disto, a configuração obtida usando a matriz de distâncias médias de Manhattan foi considerada como a configuração que melhor representa a estrutura dos dados analisados.

Já era esperado que a matriz de distâncias médias de Manhattan fornecesse uma configuração melhor do que a obtida com a matriz de distâncias euclidianas , visto que, a distância média de Manhattan é uma distância não-euclidiana ,e que o método de escalonamento não-métrico fornece resultados melhores quando distâncias não-euclidianas são usadas.

As análises feitas usando a matriz de correlação e a matriz de variância e covariância, para dados padronizados, produziram configurações com "stress" de 0.265 e 0.267, respectivamente, sendo consideradas como tendo um ajuste regular

Figura 4.6 - Solução obtida pelo Escalonamento Não-Métrico
para a Matriz de Distâncias Médias de Manhattan



aos dados, segundo o critério de avaliação de Kruskal.

As configurações obtidas para essas duas matrizes podem ser vistas nas figuras 4.7 e 4.8. Pode-se notar que estas configurações sugerem que o grupo de rochas do tipo 1 pode ser dividido em 3 novos grupos, sendo que as amostras de rochas do tipo 2 aparecem todas muito próximas formando um único grupo.

Considerou-se as configurações obtidas utilizando as matrizes de distâncias euclidianas e de distâncias médias de Manhattan, como sendo satisfatórias por representarem, a estrutura presente nos dados.

Para verificar se a solução considerada como sendo a melhor obedece a relação de monotonicidade entre as proximidades originais e as distâncias entre os pontos do espaço de configuração, construiu-se um diagrama de dispersão, o qual tem como coordenadas as distâncias médias de Manhattan e as distâncias euclidianas estimadas entre os pontos da configuração. Este diagrama de dispersão é apresentado na figura 4.9. Nota-se que o diagrama de dispersão sugere uma relação monotônica entre as distâncias observadas e as distâncias do espaço de configuração e, portanto, conclui-se que a configuração obtida é bem ajustada aos dados.

Com o objetivo de verificar se uma configuração em apenas duas dimensões é realmente suficiente para representar os dados, realizou-se análises de escalonamento multidimensional para 1, 2, 3, 4 e 5 dimensões, usando a matriz de distâncias médias de Manhattan. Conhecidos os valores do "stress" para os diferentes números de dimensões, construiu-se um gráfico do valor do "stress" em função do número de dimensões, o qual é apresentado na figura 4.10. Pode-se notar a formação de um leve 'cotovelo' no ponto em que o número de dimensões é igual a dois e que o valor do "stress" decresce rapidamente quando se aumenta o número de dimensões de 1 para 2 e que a partir deste, começa a diminuir mais lentamente. Em razão disto, conclui-se que duas dimensões são suficientes para representar os dados.

Após ter-se confirmado o número de dimensões apropriado para representar os dados, parte-se para a interpretação da solução. Como a configuraço obtida permite que sejam feitas,

Figura 4.7 - Solução obtida pelo Escalonamento Não-Métrico para a Matriz de Correlação Momento-Produto

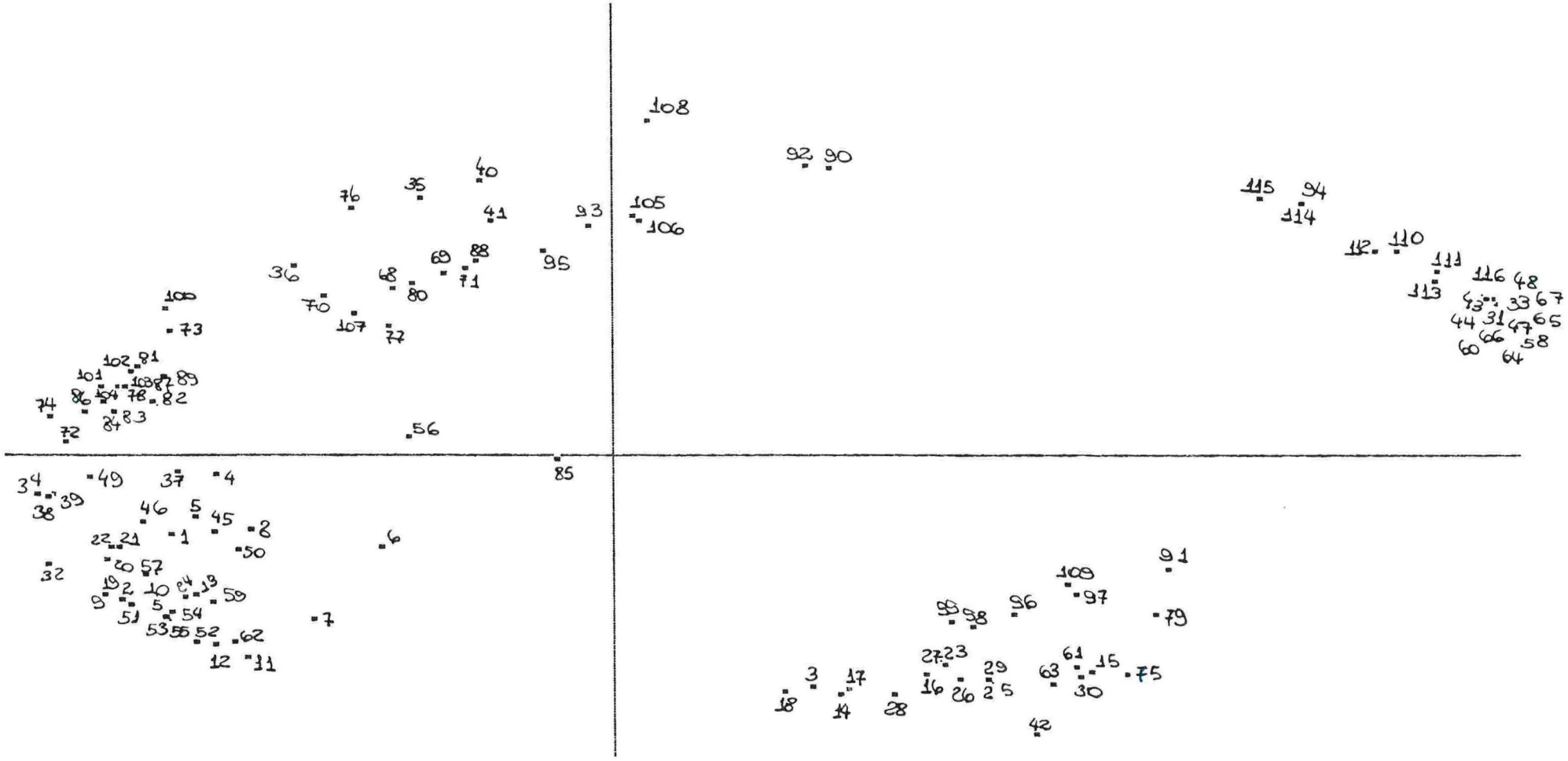


Figura 4.8 - Solução obtida pelo Escalonamento Não-Métrico para a Matriz de Variância e Covariância

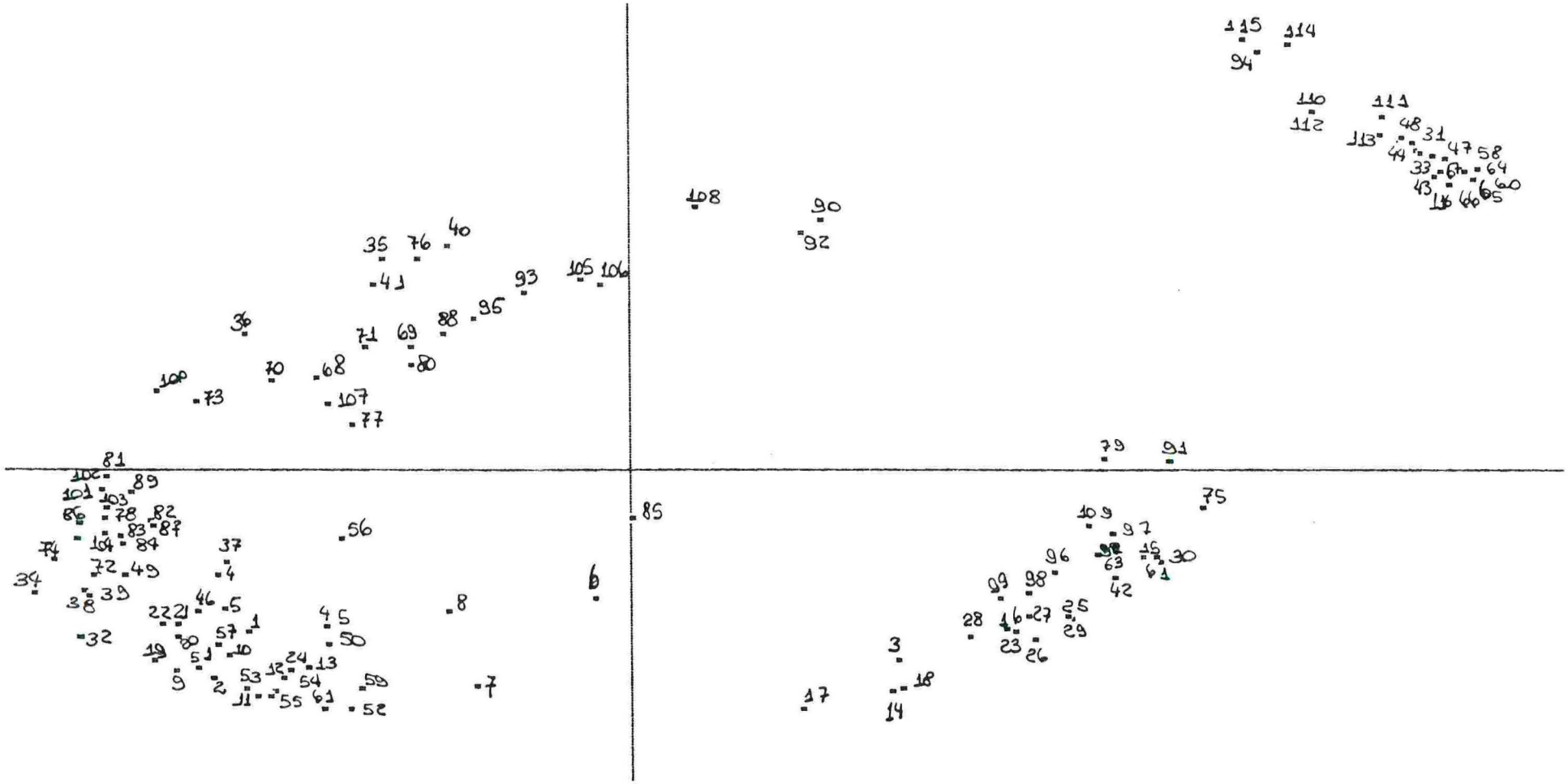


FIGURA 4.9 - DIAGRAMA DE DISPERSÃO

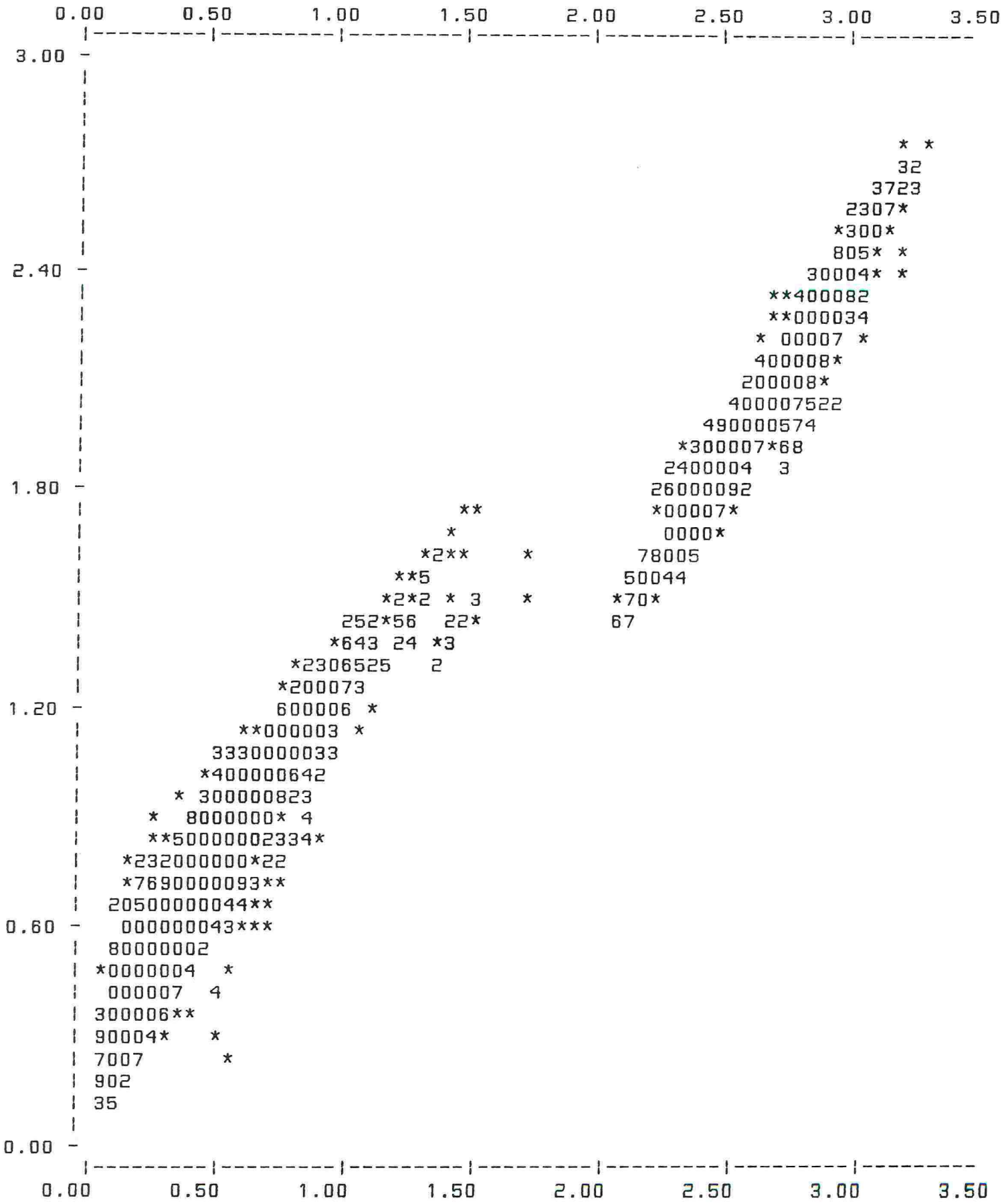
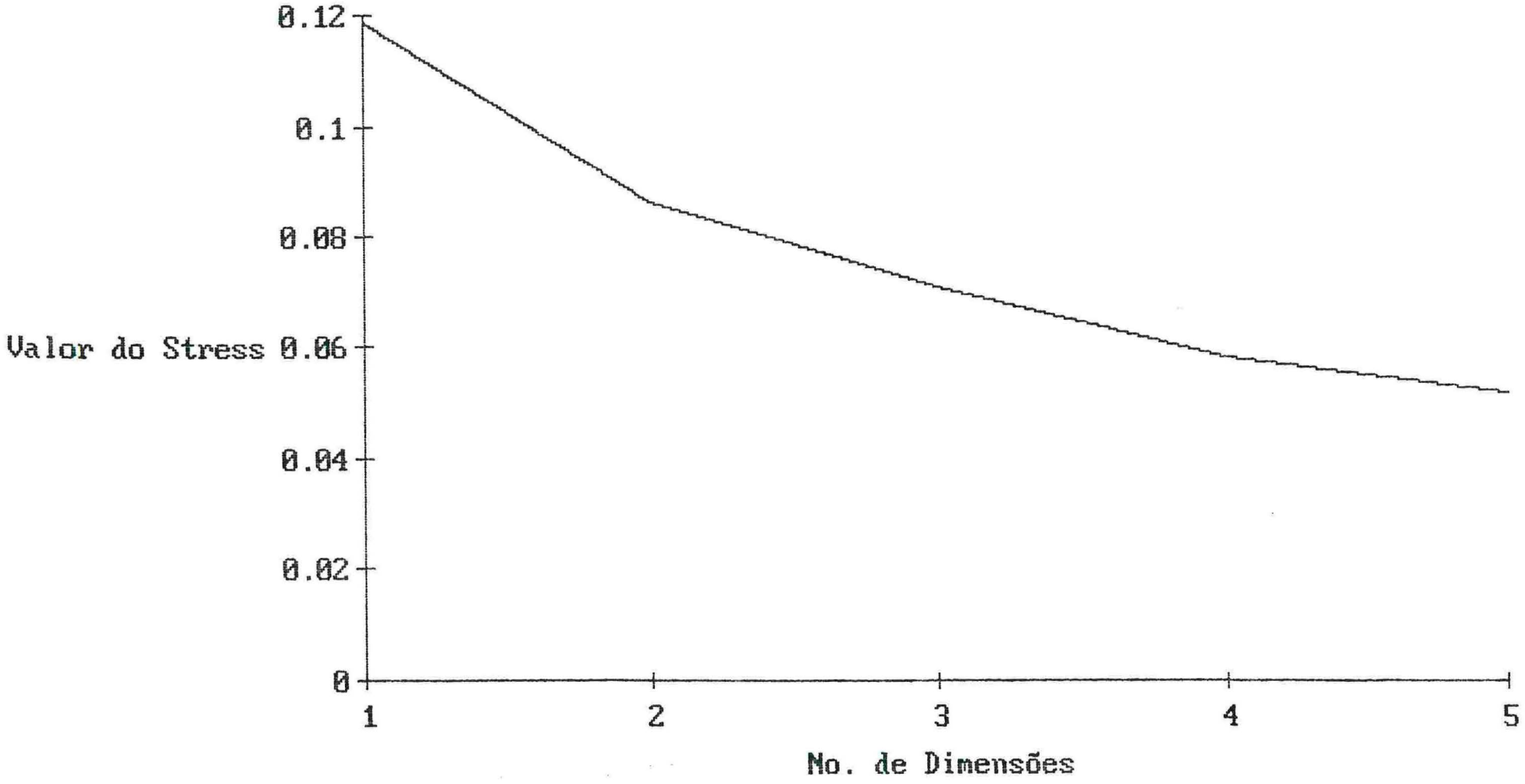


Figura 4.10 - Gráfico do Stress em função da Dimensão do Espaço



tanto a interpretação das dimensões como a interpretação por vizinhanças, apresentaremos inicialmente a interpretação das dimensões e, posteriormente, a interpretação por vizinhanças.

A dimensão 1 só pode ser interpretada como o teor de $P_2O_5$, Ba, Sr e Zr, sendo que as amostras de rochas com baixos teores destes elementos estão mais para o lado esquerdo da dimensão 1 e as amostras de rochas com altos teores destes elementos estão mais para o lado direito da dimensão 1.

A dimensão 2 da configuração obtida pode ser interpretada como o tipo de rocha, uma vez que os grupos de amostras de rochas do tipo 1 e 2 estão nos extremos opostos desta dimensão. Isto pode ser visto, observando-se a figura 4.11, que apresenta a interpretação das dimensões da configuração. Pode-se dizer que as amostras de rochas do tipo 1 estão mais próximas da base da dimensão 2 e que as amostras de rochas do tipo 2 estão mais próximas do topo da dimensão 2.

A dimensão 2 pode, também, se interpretada como o teor de SiO, $K_2O$, e Rb, sendo que, as amostras de rochas com baixos teores de SiO, $K_2O$ e Rb estão mais próximas da base e as amostras de rochas com altos teores destes elementos estão mais próximas do topo da dimensão 2. E ainda, pode ser interpretada como o teor de FeO, MgO, CaO e Ni,sendo que, as amostras de rochas que tem altos teores destes elementos estão mais próximas da base da dimensão 2 e as amostras de rochas que tem baixos teores destes elementos estão mais próximas do topo da dimensão 2.

A interpretação por vizinhanças consiste em esquecer as dimensões (eixos) da configuração e identificar os grupos de amostras de rochas que são sugeridos pela configuração.

Observando a figura 4.12, que apresenta a interpretação por vizinhanças para a configuração obtida a partir da matriz de distâncias médias de Manhattan, pode-se dizer que na parte superior direita da configuração estão as amostras de rochas do tipo 2 e que na parte inferior da configuração estão as amostras de rochas do tipo 1. E ainda, pode-se dizer que as amostras de rochas do tipo 1 se dividem em 2 novos grupos, sendo que um destes é formado pelas amostras de rochas de números 94, 114 e

Figura 4.11 - Interpretação das Dimensões da Configuração obtida usando a Matriz de Distâncias Médias de Manhattan

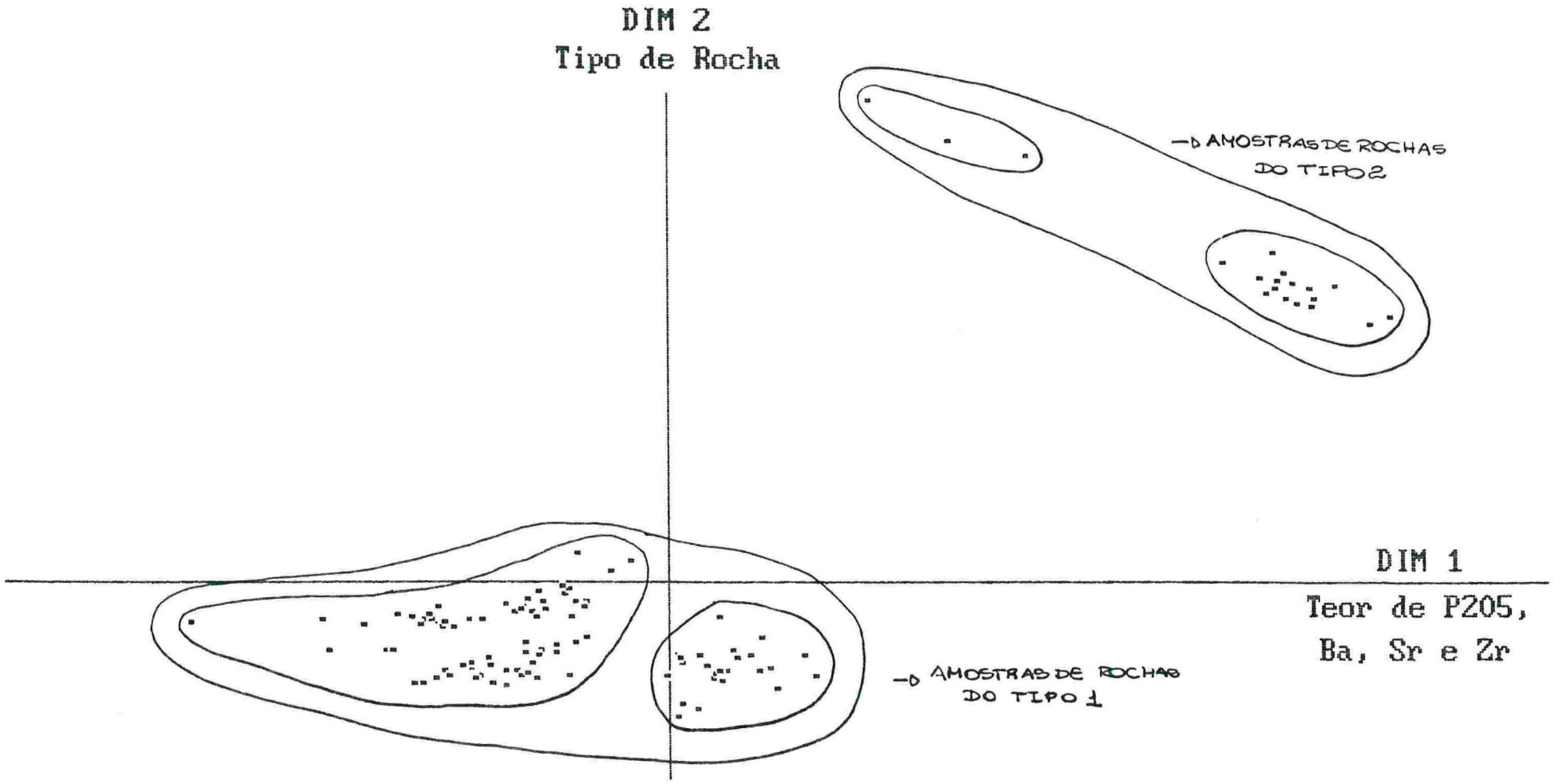
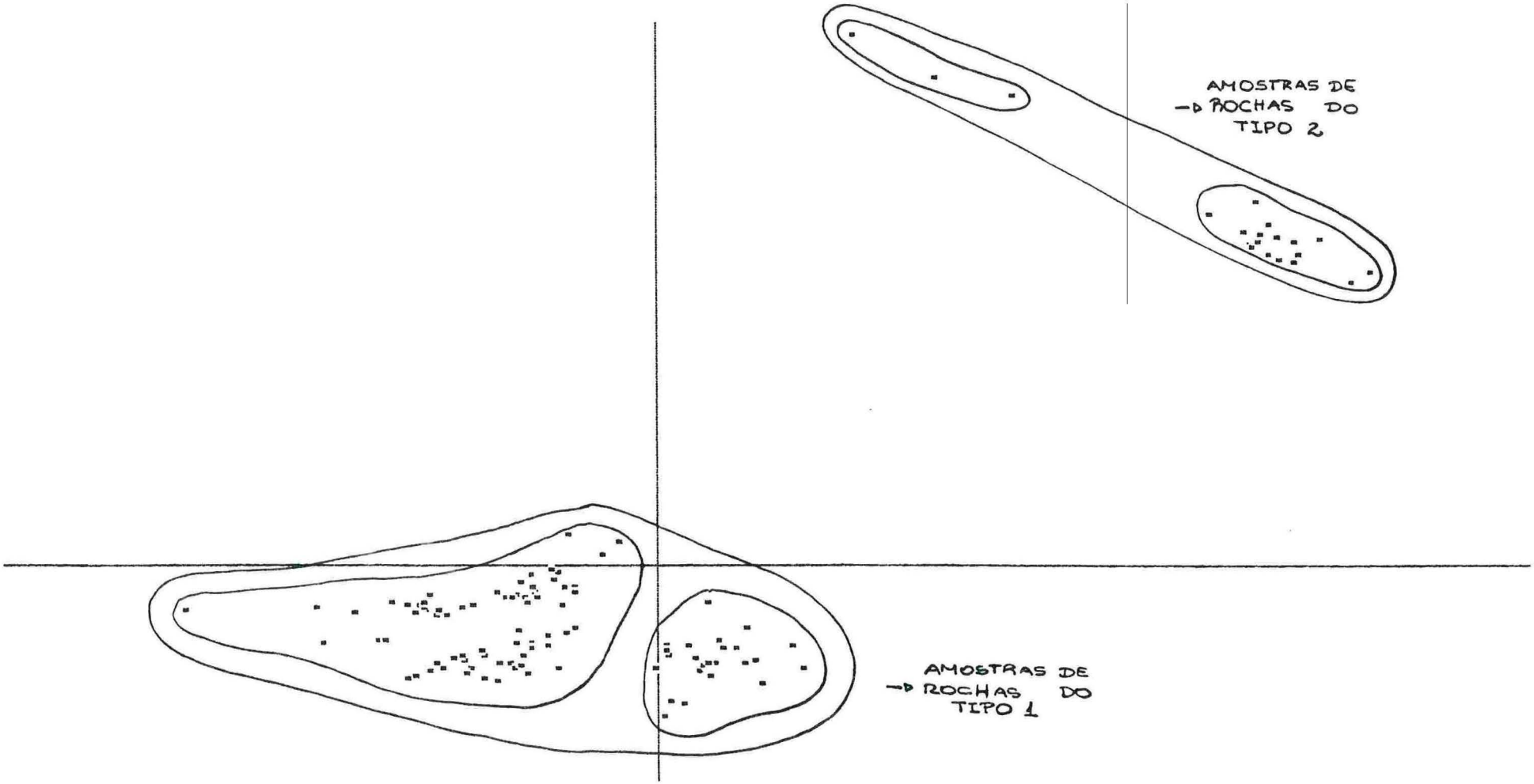


Figura 4.12 - Interpretação por Vizinhanças da Configuração obtida usando a Matriz de Distâncias Médias de Manhattan



115 e o outro é formado pelas restantes. Pode-se dizer, ainda, que as amostras de rochas do tipo 1 também se dividem em dois novos grupos, como está indicado na figura 4.12.

Após ter-se identificado os grupos de amostras de rochas que se formaram, pode-se identificar quais as características que estão associadas às amostras de rochas de cada grupo, observando os gráficos que estão no anexo.

Uma das formas de interpretar vizinhanças numa configuração envolve a aplicação de técnicas de Análise de Agrupamento ("Cluster Analysis"). Em razão disso, considerou-se importante apresentar os resultados de uma análise de agrupamento feita usando a matriz de distâncias euclidianas ao quadrado, obtida a partir da matriz de dados padronizados, como matriz de proximidades.

Os grupos de amostras de rochas sugeridos pela configuração obtida a partir da matriz de distâncias médias de Manhattan, e da matriz de distâncias euclidianas ao quadrado, coincidem com os grupos formados pela análise de agrupamento. Isto pode ser verificado, observando as figuras 4.5, 4.6 e a figura 4.13, a qual apresenta os resultados da análise de agrupamento.

Conclui-se, então, que os resultados obtidos com a análise de escalonamento multidimensional e com a análise de agrupamento, resultam nos mesmos grupos de amostras de rochas.

# CAPITULO 5

## NTSYS-pc

### 5.1 - INTRODUÇÃO

NTSYS-pc ( Numerical Taxonomy SYStem ) é um sistema de programas que executa vários tipos de operações, que são úteis para achar e expor modelos e estruturas que possam estar presentes em dados multivariados. Os programas foram originalmente planejados para serem utilizados na área de Biologia, mas são de grande aplicabilidade.

Os programas do NTSYS-pc podem ser usados para obter matrizes de proximidades entre objetos a partir de uma matriz de dados e, também, para sumarizar informações em termos de grupos de objetos similares (análise de agrupamento) ou em termos de um arranjo espacial sobre um ou mais eixos de coordenadas (escalonamento multidimensional ou análise de ordenação ).

NTSYS-pc representa uma "micro" versão do NTSYS, que é um grande sistema de programas escritos em FORTRAN para computadores de grande porte. O sistema consiste de um número de módulos separados coordenados pelo programa principal NTSYS. Muitos módulos executam algumas operações distintas sobre uma

matriz de dados e produzem uma matriz resultante que pode ser armazenada para ser processada por outro programa deste sistema, ou por outros programas, visto que o formato dos arquivos é muito simples. Em um microcomputador com 640K de memória RAM é possível para muitos dos programas processar matrizes com até 400 objetos ou variáveis, sendo que a principal limitação no uso dos programas será o tempo de cálculo e espaço no disco para armazenar matrizes de dimensões (400x400).

Em adição, também existe um suplemento NTSYS-pc que contém programas adicionais que são de interesse mais especializado, como o MDSCALE, que executa análise de escalonamento multidimensional linear e não-métrico, e o CORRESP, que executa análise de correspondência.

## 5.2 - MODOS DE OPERAÇÃO

O NTSYS-pc pode ser usado de duas maneiras, sendo que, uma consiste em usar os programas separadamente e a outra consiste em montar um arquivo de comandos de linha em que cada linha inicia com o nome do programa que se quer usar seguido dos parâmetros do programa definidos. Quando se sabe todos os programas que vão ser usados e a ordem com que eles devem ser usados, costuma-se construir um arquivo de comandos de linha , ao invés de usar os programas separadamente.

## 5.3 - CARREGANDO O NTSYS-pc

Para acessar o NTSYS-pc, usando-se os programas separadamente, digita-se NTSYS-pc e tecla-se 'ENTER'. Aparecerá

# A N E X O

Gráfico 1.1 - Quantidade de SiO presente na composição das Amostras de Rochas

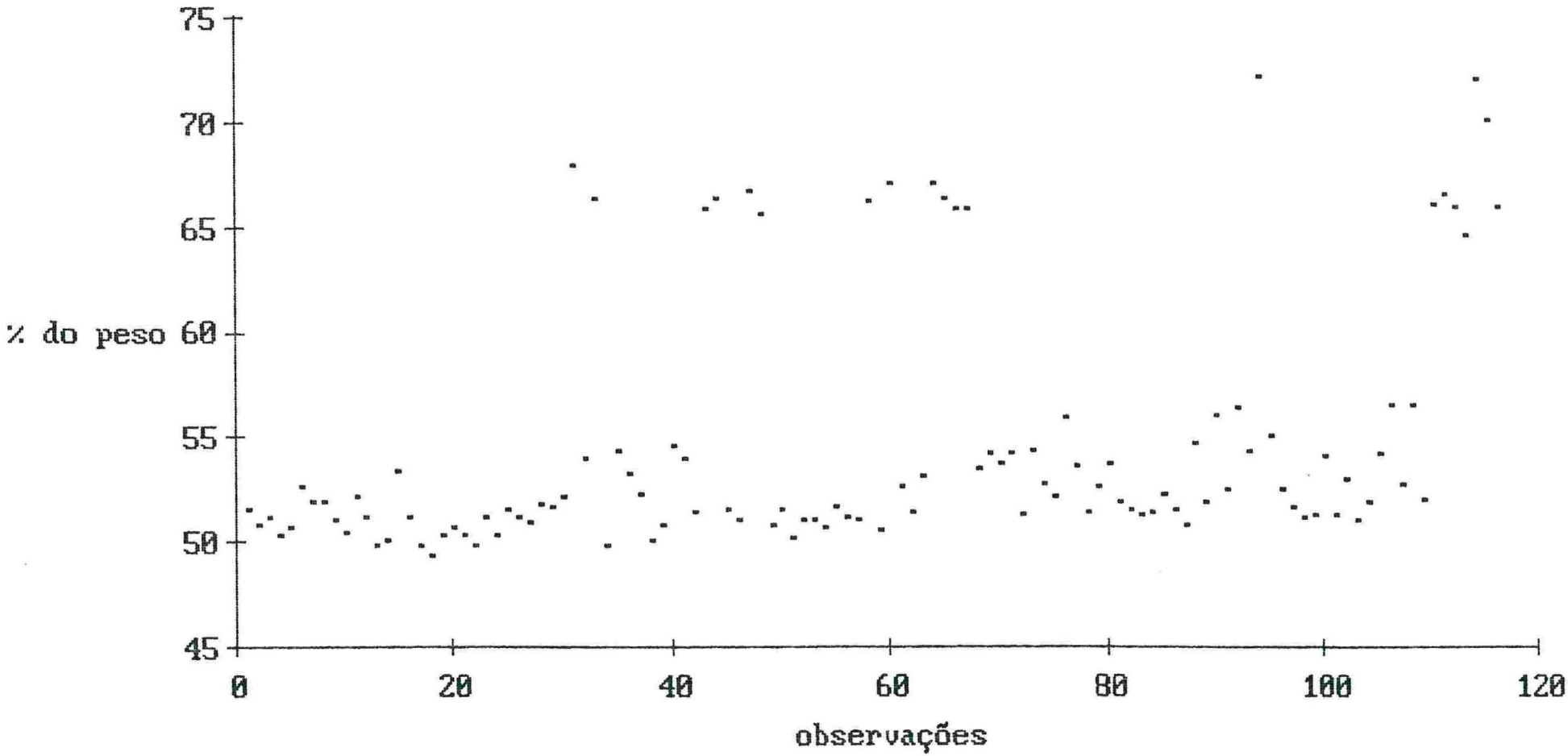


Gráfico 1.2 - Quantidade de TiO presente na composição das Amostras de Rochas

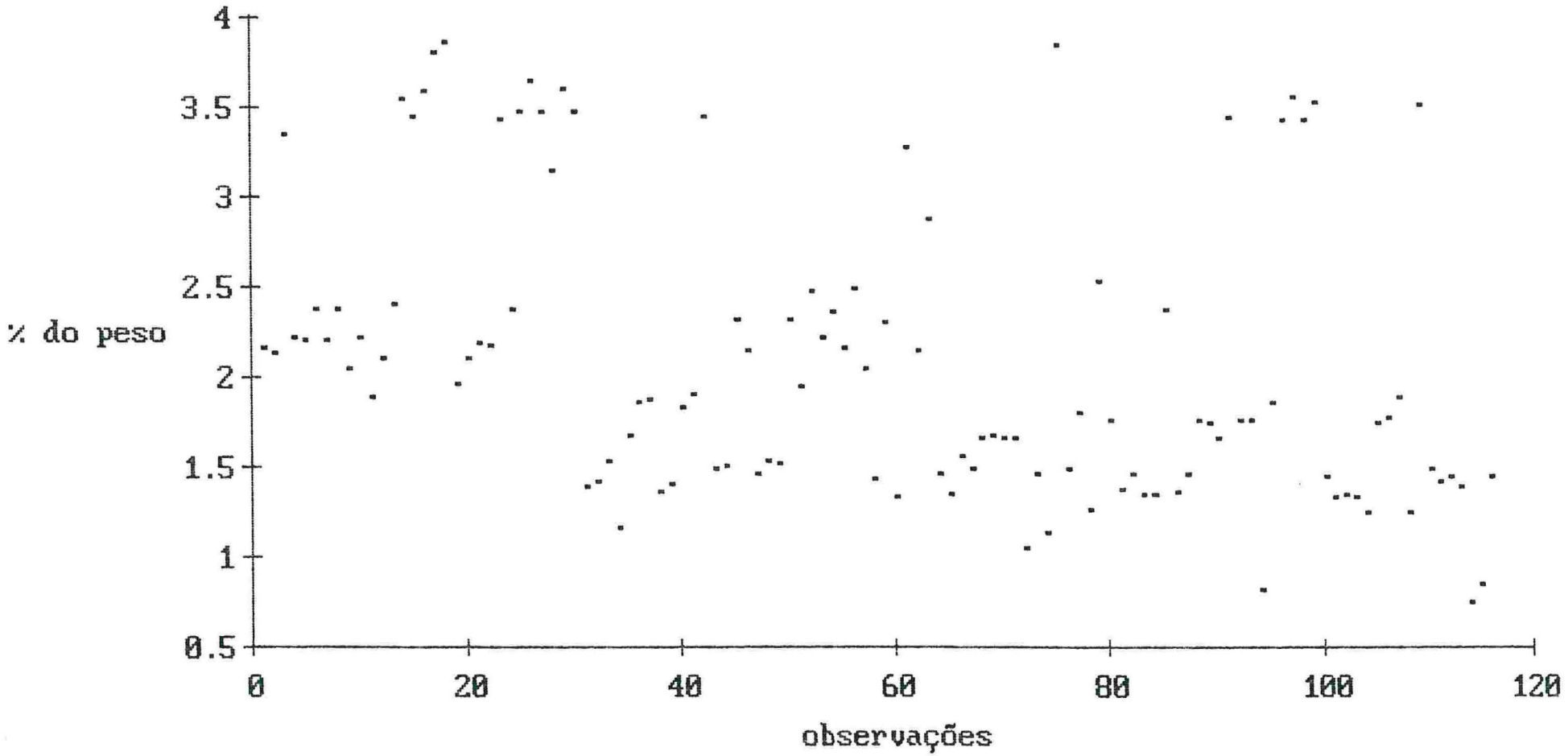


Gráfico 1.3 - Quantidade de Al2O3 presente na composição das Amostras de Rochas

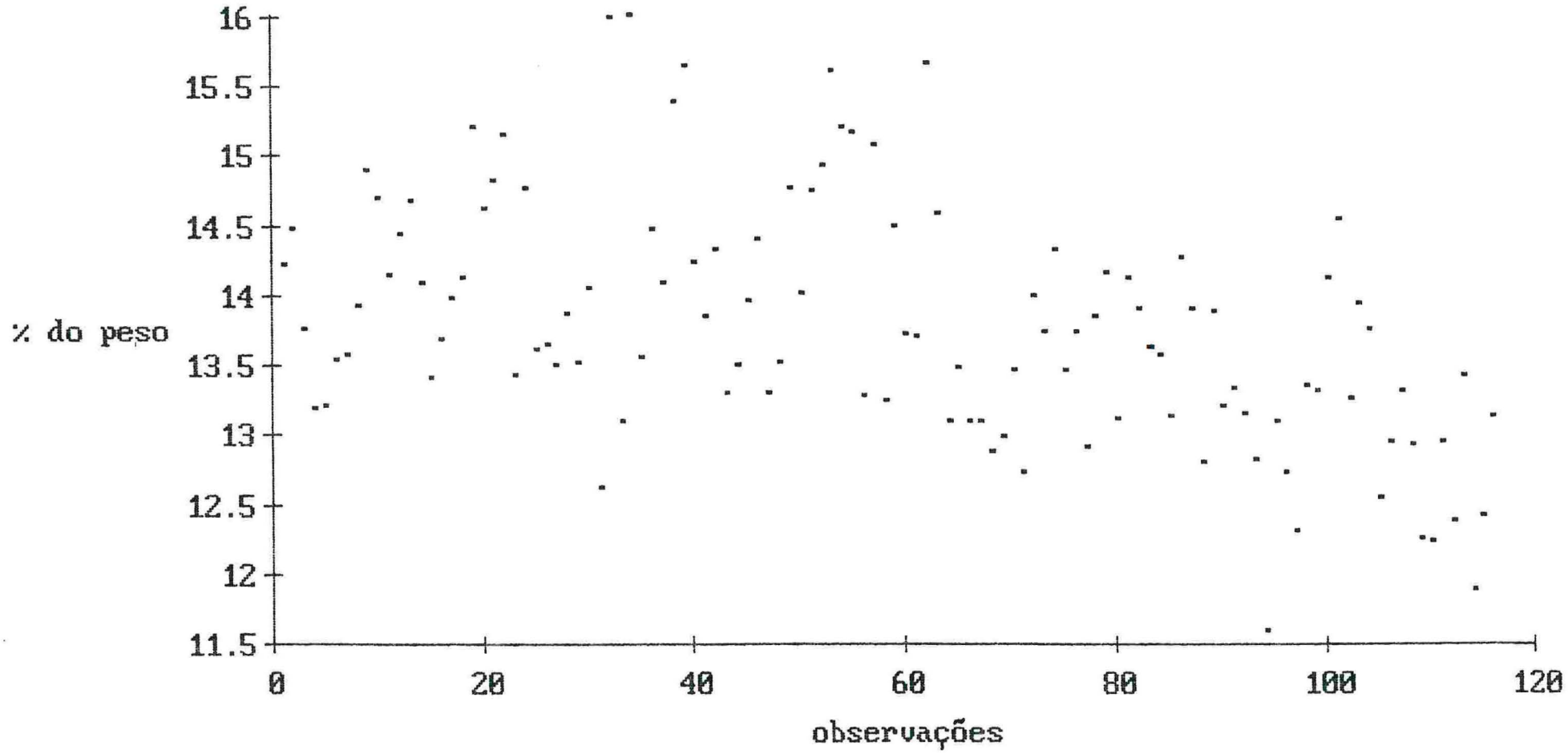


Gráfico 1.4 - Quantidade de FeO presente na composição das Amostras de Rochas

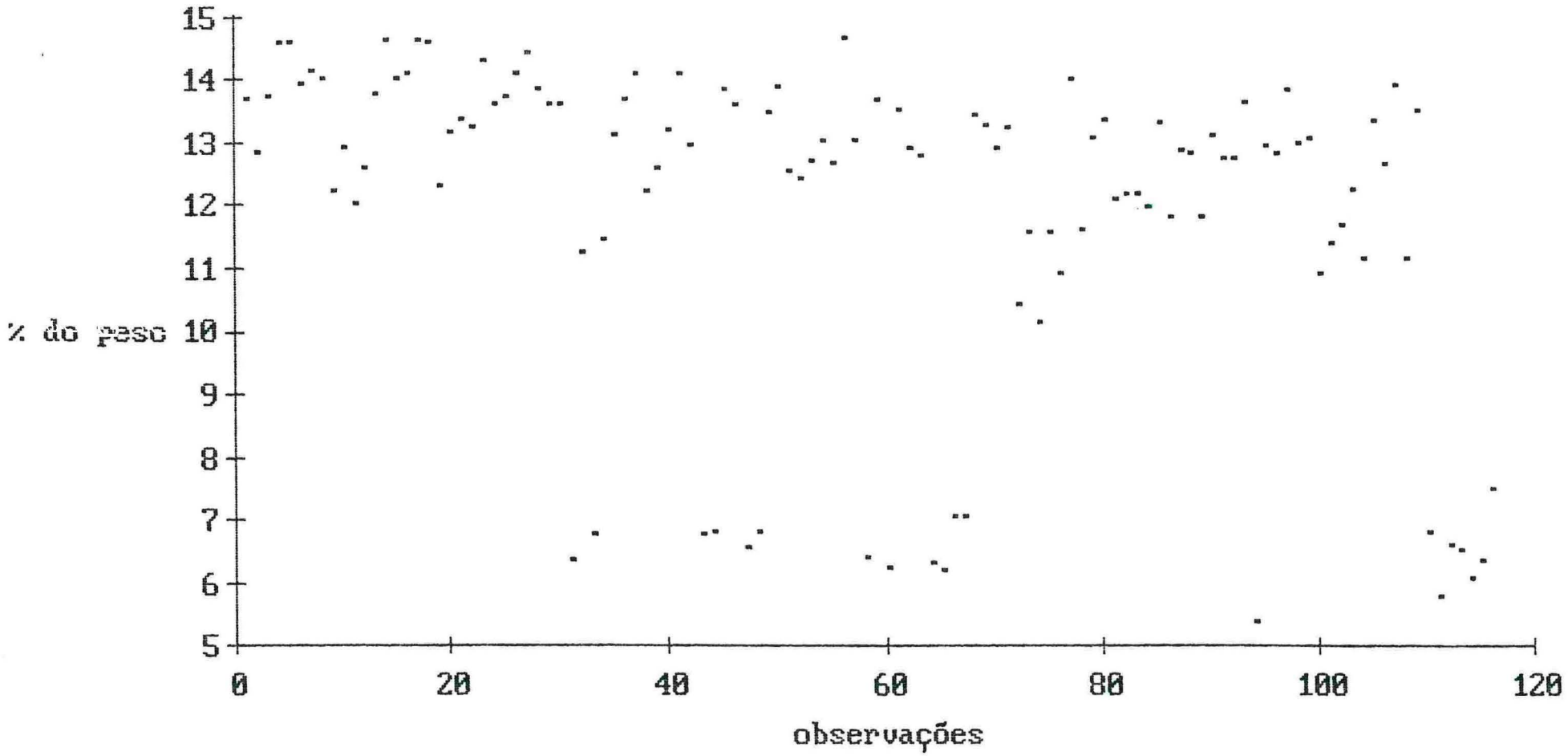


Gráfico 1.5 - Quantidade de MnO presente na composição das Amostras de Rochas

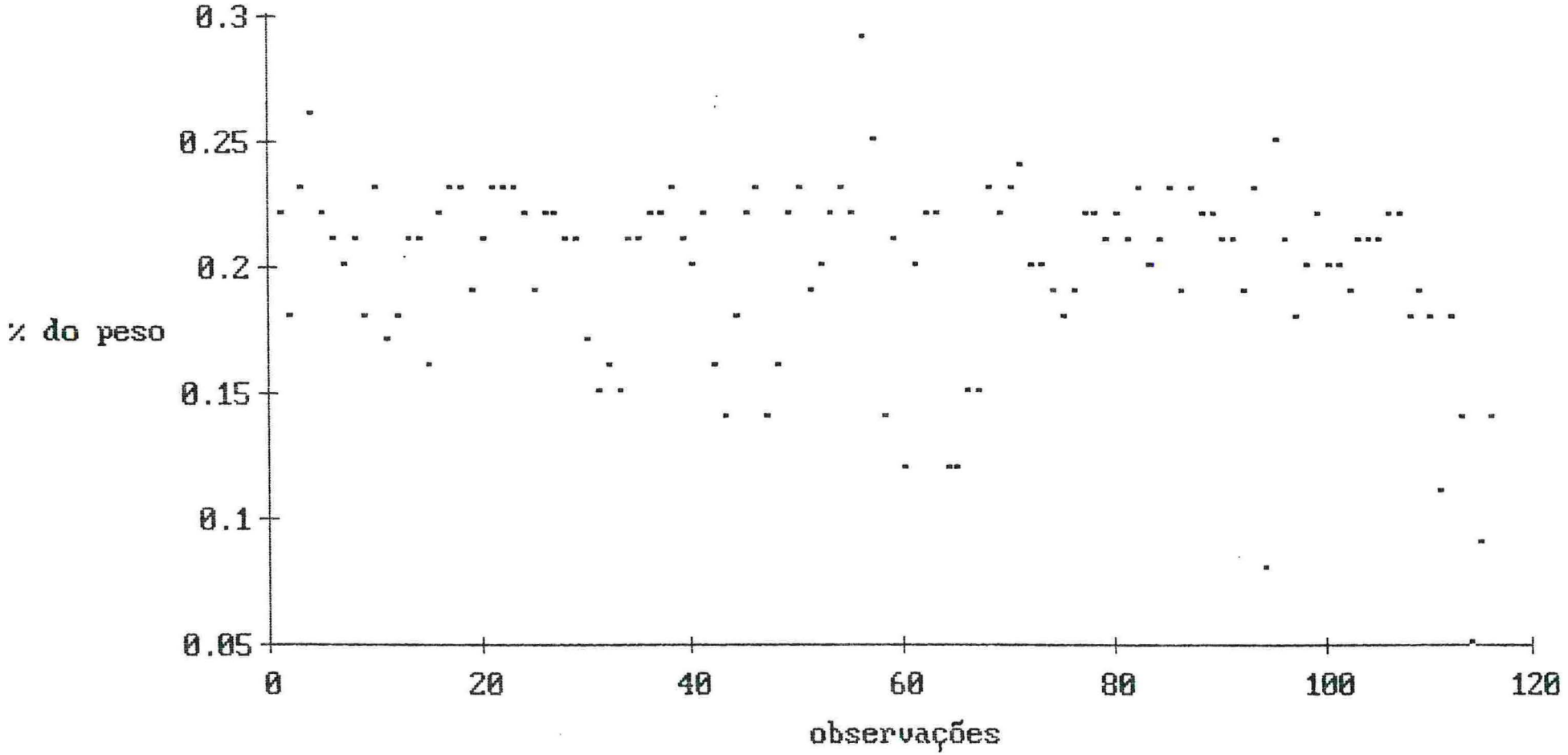


Gráfico 1.6 - Quantidade de MgO presente na composição das Amostras de Rochas

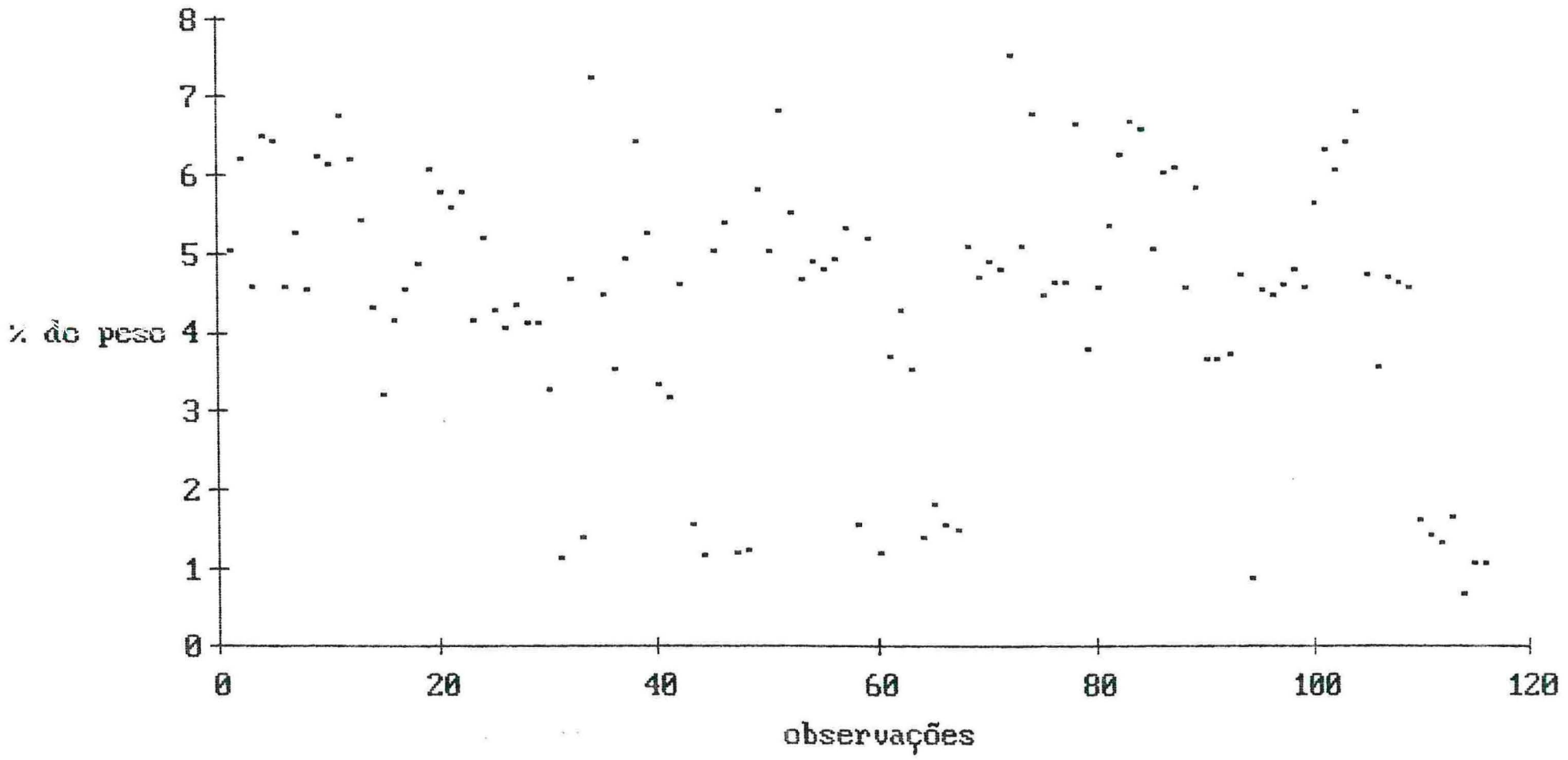


Gráfico 1.7 - Quantidade de CaO presente na composição das Amostras de Fachas

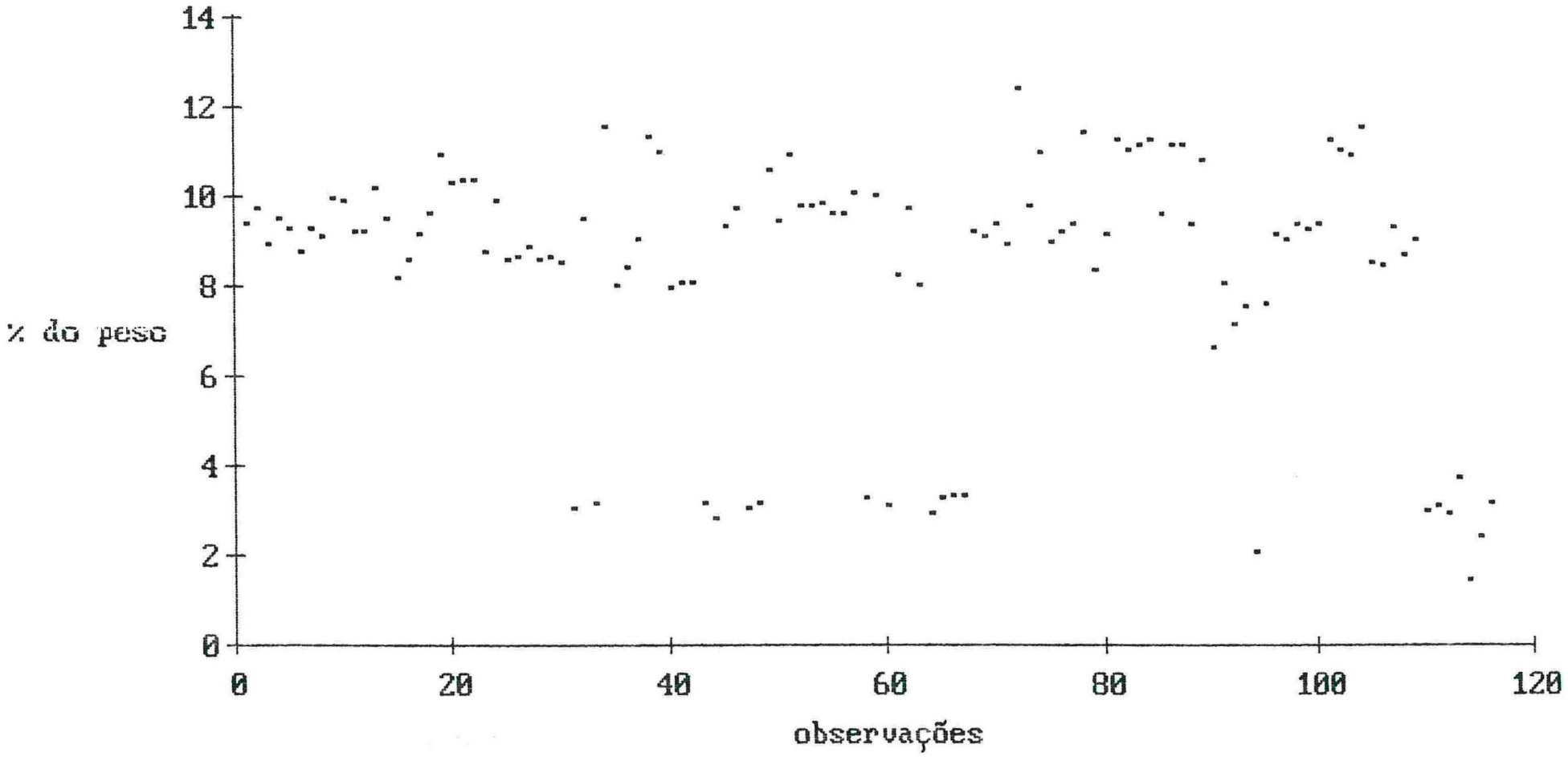


Gráfico 1.8 - Quantidade de Na20 presente na composição das Amostras de Rochas

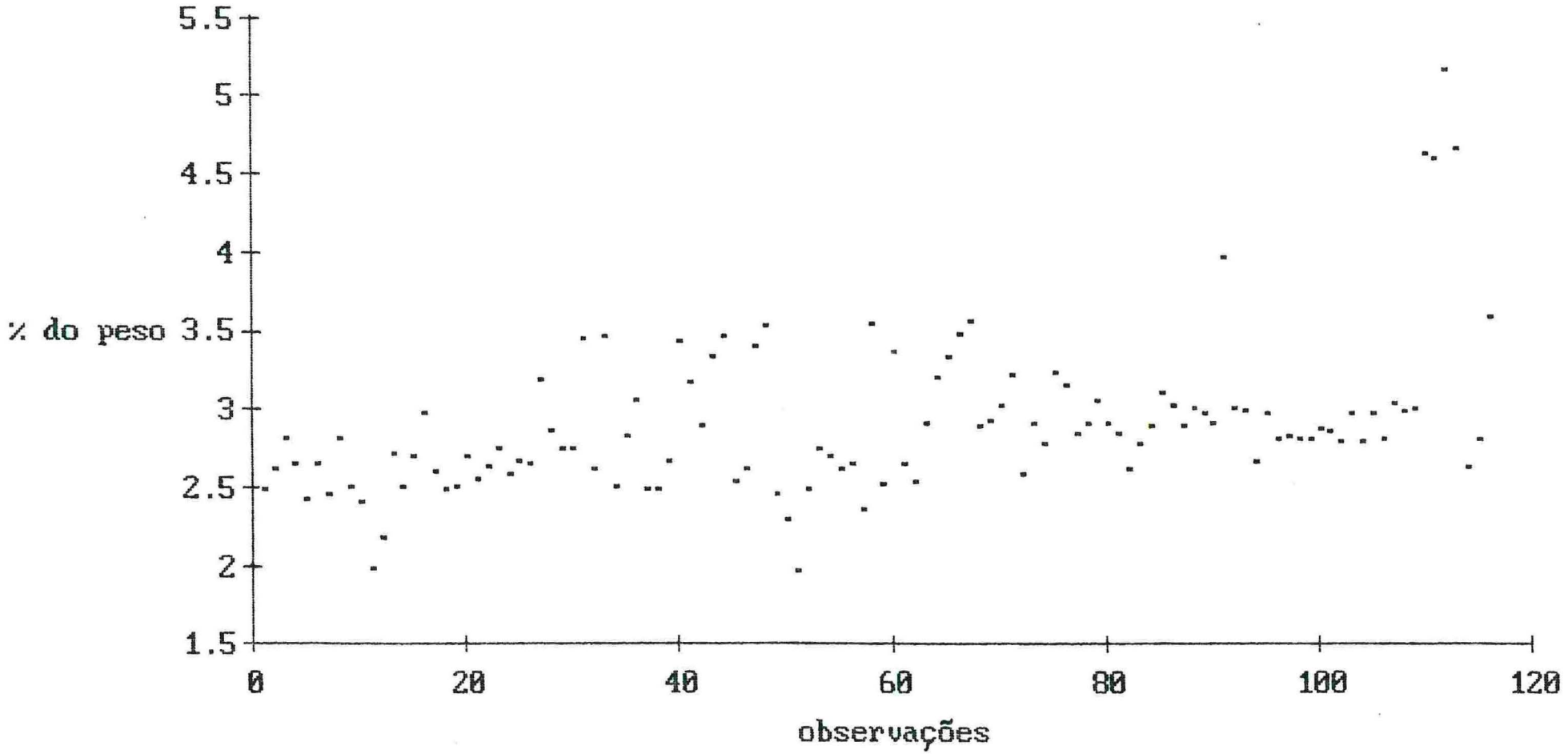


Gráfico 1.9 - Quantidade de K20 presente na composição das Amostras de Rochas

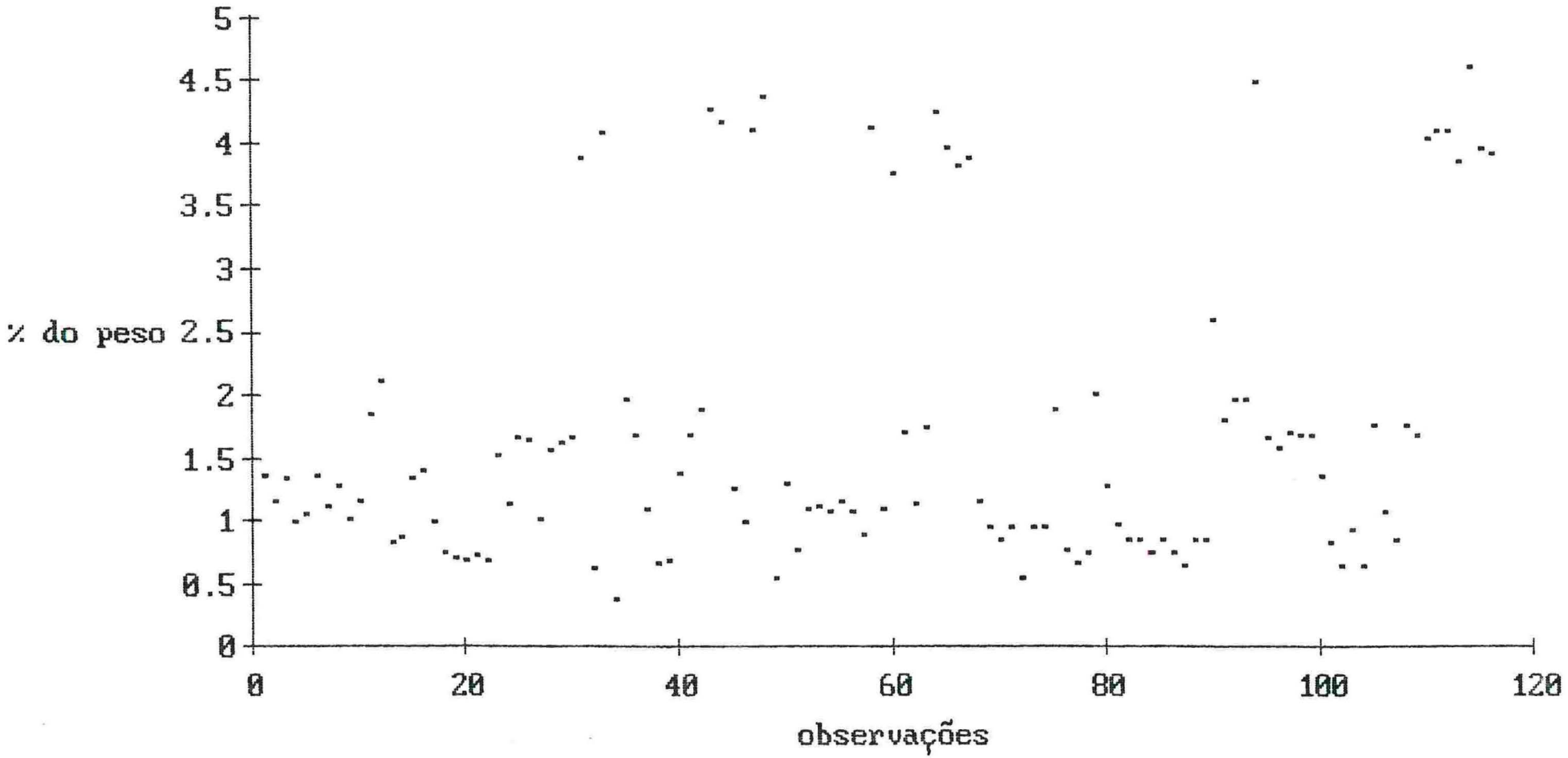
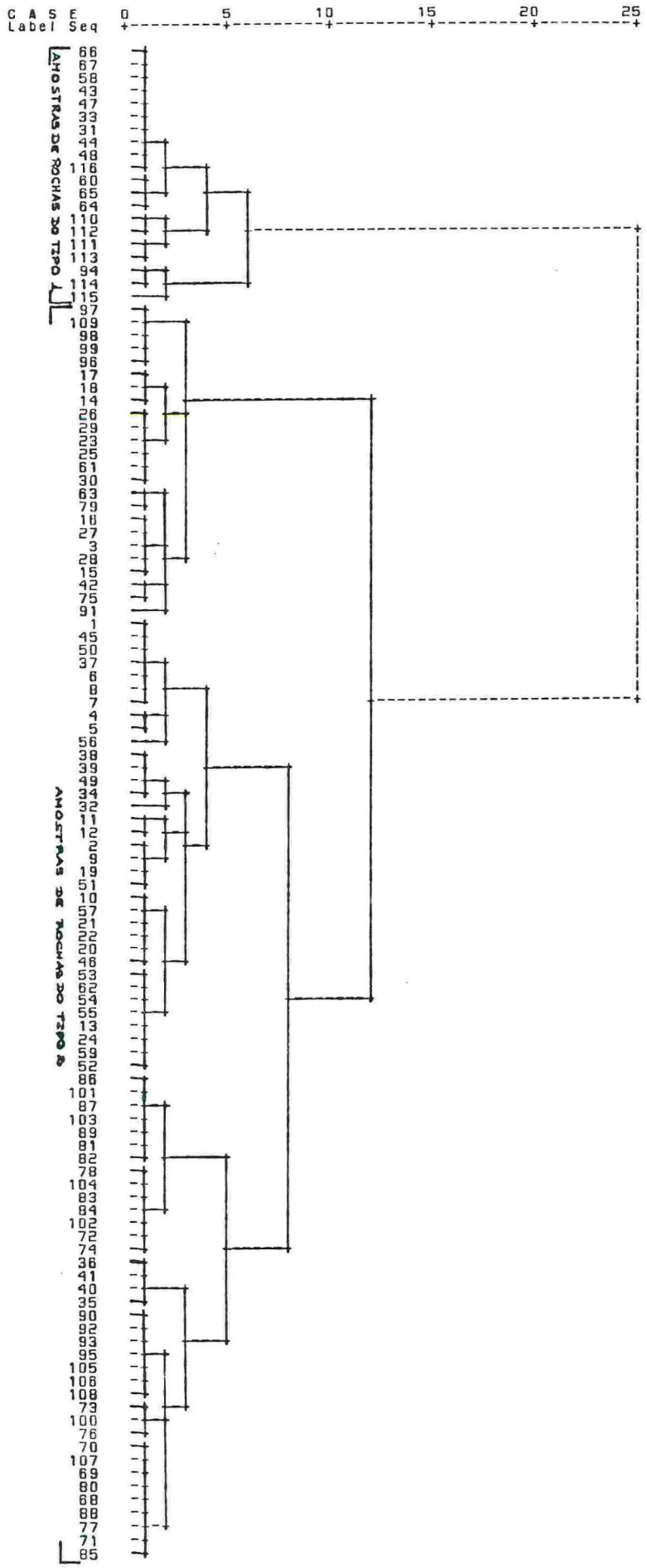


FIGURA 4.13 - RESULTADO DA ANÁLISE DE AGRUPAMENTO

Dendograma obtido usando o Método Ward

Rescaled Distance Cluster Combine



na tela um menú que apresenta o nome de todos os programas do NTSYS, que podem ser selecionados posicionando o cursor sobre o nome do programa escolhido e pressionando a tecla de função F2.

Aparecerá na parte inferior da tela as seguintes funções: F1 → para acessar um arquivo de ajuda, F2 → para carregar o programa selecionado e para executar o programa, F3 → para sair do NTSYS, e F4 → para selecionar um arquivo de dados do diretório do NTSYS.

Menú Principal:

```
                NTSYS, Version 1.30 - (C) F. James Rohlf, 1987

ê             Serial number  =  270
COPH            PROJ                  RAM available  = 377712 bytes
CORRESP         SAHN                  Listing device = CON:
DCENTER         SIMGEND
EIGEN           SIMINT                No graphics adaptor
MDSCALE         SIMQUAL
MOD3DG          STAND
MST             TRANSF
MXCOMP          TREE
MXCOMPG         TREEG
MXPLOT
MXPLOTG
OUTPUT
é

 Use cursor keys to select; F1=help, F2=run, F3=exit, F4=dir
```

## 5.4 - PREPARAÇÃO DO ARQUIVO DE DADOS

Os arquivos de dados que podem ser lidos pelo NTSYS são arquivos em ASCII ( não arquivos binários ). Um arquivo para uma matriz de dados inicial pode ser preparado com um editor que

tenha um modo de caracter ASCII puro.

Cada arquivo pode conter 4 linhas de comentários, sendo que algumas são opcionais. A primeira linha do arquivo é opcional e o primeiro caracter deve ser (") ou ('). A informação nestas linhas podem ser comentários ou títulos referentes a matriz de dados.

A segunda linha deve conter 4 números inteiros e possivelmente um número real, separados por um espaço em branco.
- O primeiro número é um código para indicar o tipo de matriz que será usada: 1=matriz de dados retangular, 2=matriz de dissimilaridade simétrica, 3=matriz de similaridade simétrica, 4=matriz diagonal, 5=matriz árvore para dados de dissimilaridade, 6=matriz árvore para dados de similaridade, 7= matriz diagrama para dados de dissimilaridade, e 8=matriz diagrama para dados de similaridade.
- O segundo e o terceiro número são os números de linhas e colunas da matriz de dados , sendo que quando cada linha tem um nome coloca-se a letra L depois do segundo número e quando as colunas tem nomes coloca-se um L depois do terceiro número (sem espaço entre o número e a letra L).
- O quarto número é zero se não existem observações perdidas na matriz de dados e é 1 se houver observações perdidas, sendo que o número 1 deve aparecer seguido do código numérico usado para denotar as observações perdidas.

A terceira linha do arquivo de dados deve apresentar os nomes das linhas,ou das colunas, se for o caso. Se a letra L não foi colocado após o segundo e/ou o terceiro número da segunda linha do arquivo de dados, a terceira linha não é necessária.

A partir da quarta linha do arquivo de dados pode ser digitada a matriz de dados propriamente dita, sendo que, se a matriz de dados for simétrica as linhas começarão na primeira coluna e terminarão no elemento da diagonal principal da matriz.

## 5.5 - DESCRIÇÃO DOS PROGRAMAS

Serão descritos a seguir os programas do NTSYS-pc e serão apresentados exemplos de definições de parâmetros que os programas exigem que sejam definidos.

COPH - Este programa lê uma 'matriz árvore', que pode ter sido produzida pelo programa SAHN, e produz uma matriz de valores de distâncias ultramétricas que representam a similaridade ou dissimilaridade entre pares de objetos.

Parâmetros do Programa:

COPH

| | |
|---|---|
| Name of tree matrix | A:MATRIZ.TRE |
| Name for cophenetic matrix | B:COPHEN.MAT |
| Listing file | PRN |

DCENTER - Transforma uma matriz simétrica para a forma de produto escalar tal que esta pode ser fatorada, usando o programa EIGEN, resultando em uma análise de coordenadas principais.

Parâmetros do Programa:

DCENTER

| | |
|---|---|
| Name of input matrix | A:MATRIZ.DIS |
| Name for result matrix | B:RESULT.MAT |
| Square distances? | Y |
| Listing file | PRN |

EIGEN - Calcula matrizes de valores e vetores característicos para uma matriz simétrica de proximidades. O número de dimensões a ser guardado pode ser escolhido. O tipo de normalização do comprimento dos vetores pode ser escolhido para ser igual ao valor característico (código a ser usado :LAMBDA), para ser igual a raiz quadrada do valor característico (código :SQRT(LAMBDA)), para ser igual a 1 (código :1), ou para ser igual a reciproca do valor característico (código :1/LAMBDA).

Parâmetros do Programa:

EIGEN

| | |
|---|---|
| Name of input matrix | A:MATRIZ.SIM |
| Number of dimensions | 2 |
| Name for eigenvector matrix | B:VETCAR.MAT |
| Name for eigenvalue matrix | B:VALCAR.MAT |
| Length of vectors? | SQRT(LAMBDA) |
| Listing file | PRN |

MST - Produz uma árvore a partir de uma matriz de similaridade ou dissimilaridade.

Parâmetros do Programa:

MST

| | |
|---|---|
| Name of input matrix | A:MATRIZ.DAD |
| Name for output graph matrix | B:GRAFIC.MAT |
| Listing file | PRN |

MXCOMP - O programa lê duas matrizes simétricas de similaridade ou dissimilaridade, calcula sua correlação , e então constrói um

gráfico de uma matriz versus a outra.

Parâmetros do Programa:

MXCOMP

| | |
|---|---|
| Name of X input matrix | A:DISTAN.MAT |
| Name of Y input matrix | A:MATRIZ.DIS |
| Listing file | PRN |

MXPLOT - Representa graficamente pares de linhas ou colunas de uma matriz, uma contra a outra.

Parâmetros do Programa:

MXPLOT

| | |
|---|---|
| Name of input matrix | A:MATRIZ.GRA |
| Direction to plot by | R |
| Variable (I) for X-axis | 1 |
| Variable (J) for Y-axis | 2 |
| X min | -1.0 |
| X max | 1.0 |
| No. intervals for X-axis | 4 |
| Decimal places for X-axis | 2 |
| Y min | -0.5 |
| Y max | 2.8 |
| No. intervals for Y-axis | 4 |
| Decimal places for Y-axis | 2 |
| X label | I |
| Y label | II |
| Identify the points? | Y |
| Listing file | PRN |

OUTPUT - Este programa imprime, ou grava num disco, uma matriz formatada em linhas e colunas . É usado, geralmente, para imprimir matrizes produzidas por outros programas do sistema NTSYS-pc.

Parâmetros do Programa:

OUTPUT

| | |
|---|---|
| Name of matrix | A:RESULT.MAT |
| Field width | 9 |
| Number of decimal places | 3 |
| Page width | 79 |
| Listing file | PRN |

PROJ - Projeta objetos de uma matriz de dados sobre um ou mais eixos. Os eixos são usualmente vetores característicos de uma matriz de correlação entre variáveis.

Parâmetros dos Programas:

PROJ

| | |
|---|---|
| Name of data matrix | A:MATRIZ.DAD |
| OTUs = rows or cols? | C |
| Name of factor matrix | B:FATOR.MAT |
| Name for projection matrix | B:PROJE.MAT |
| Listing file | PRN |

SAHN - Realiza agrupamento com os vários algoritmos que Sneath e Sokal, em 1973, referem como :Sequencial, Aglomerativo e Hierárquico.

Parâmetros do Programa:

SAHN

| | |
|---|---|
| Name of input matrix | A:MATRIZ.DAD |
| Method | UPGMA |
| Name for output matrix | A:RESULT.MAT |
| beta | -0.25000 |
| Listing file | PRN |

Códigos dos Métodos de Agrupamento:

- COMPL - Método de Ligação Completa.
- FLEXI - Agrupamento Flexível.
- SINGL - Método de Ligação Simples.
- UPGMA - Método Não-Ponderado, que usa médias aritméticas.
- WPGMA - Método Ponderado, que usa médias aritméticas.
- WPGMC - Método Ponderado, que usa médias centróide.
- WPGMS - Método Ponderado, que usa médias de Spearman.

SIMGEND - Calcula vários coeficientes de distância genética para dados de frequências.

Parâmetros do Programa:

SIMGEND

| | |
|---|---|
| Name of input matrix | A:MATRIZ.DAD |
| Coefficient | NEI |
| Name for output matrix | B:RESULT.MAT |
| By rows or cols? | R |
| No. loci | 0 |
| Listing file | PRN |

STAND - Realiza uma variedade de transformações linares das variáveis da matriz de dados.

Parâmetros do Programa:

```
                                    STAND

Name of input matrix              A:MATRIZ.DAD
Direction of standardization      R
Name for output matrix            B:RESULT.MAT
Subtract                          YBAR
Divide                            STD
Constant                          0.00000
Listing file                      PRN
```

Opções de Subtração:

- MIN - O valor mínimo de cada variável é subtraído.
- YBAR - A média de cada variável é subtraida.

Opções de Divisão:

- MAX - Divide pelo valor máximo de cada variável.
- RANGE - Divide pela amplitude de cada variável.
- STD - Divide pelo desvio padrão de cada variável.
- SQRT(SSD) - Divide pela raiz quadrada da soma de quadrados.
- SQRT(SY) - Divide pela raiz quadrada da soma dos valores de cada variável.
- SQRT(SY2) - Divide pela raiz quadrada da soma dos valores de cada variável ao quadrado.
- SUMY - Divide pela soma dos valores de cada variável.

As médias, variâncias, tamanhos de amostras, mínimo e máximo para cada variável são apresentados, assim como os dados transformados.

SIMINT - Calcula uma variedade de coeficientes de similaridade e

dissimilaridade para dados quantitativos.

Parâmetros do Programas:

SIMINT

| | |
|---|---|
| Name of input matrix | A:MATRIZ.DAD |
| Coefficient | EUCLID |
| Name for output matrix | B:RESULT.MAT |
| By rows or cols? | R |
| Listing file | PRN |

Códigos dos Coeficientes:

- CORR - Correlação Momento-Produto.
- COSINE - Cosseno do Ângulo entre Vetores.
- DIST - Distância Taxonômica Média.
- DISTSQ - Distância Média ao Quadrado.
- EUCLID - Distância Euclidiana.
- MANHAT - Distância Média de Manhattan.
- N - Tamanho da Amostra para cada Coeficiente ( isto é de interesse quando existem valores perdidos na matriz de dados).
- VARCOV - Variâncias e Covariâncias.
- XXT - Matriz de Dados pela sua Transposta.

SIMQUAL - Calcula uma variedade de coeficientes de similaridade e dissimilaridade para dados qualitativos (nominais).

Parâmetros do Programa:

SIMQUAL

| | |
|---|---|
| Name of input matrix | A:MATRIZ.DAD |
| Coefficient | SM |
| Name for output matrix | B:RESULT.MAT |
| By rows or cols? | C |
| Positive code | 1 |
| Negative code | 0 |
| Listing file | PRN |

TRANSF - Realiza uma série de transformações lineares e não-lineares de todos os elementos de uma matriz, ou de linhas, ou de colunas selecionadas.

Parâmetros do Programa:

TRANSF

| | |
|---|---|
| Name of matrix | A:MATRIZ.DAD |
| Transf. codes | LOG[3-5] |
| Direction | C |
| Result | B:RESULT.MAT |
| Listing file | PRN |

MDSCALE - O programa MDSCALE foi usado para realizar as análises de escalonamento multidimensional não-métrico, apresentadas no capítulo 4..

Parâmetros do Programa:

MDSCALE

| | |
|---|---|
| Name of input matrix | A:MATRIZ.DIS |
| Number of dimensions | 4 |
| Name of initial matrix | A:MTRIZ.INI |
| Direction of vectors | C |
| Name of final matrix | B:FINAL.MAT |
| Regression type | MONO |
| Maximum iterations | 50 |
| Minimum stress | 0.00100 |
| Maximum stress ratio | 0.99900 |
| Name of dhat matrix | B:DISCRE.MAT |
| Name of dconfig matrix | B:DISTAN.MAT |
| Minimum gradient | 0.00000 |
| Listing file | PRN |

CORRESP - Este programa realiza uma análise de correspondência que inicia com uma tabela de contingência retangular.

Parâmetros do Programa:

CORRESP

| | |
|---|---|
| Name of input matrix | A:TABELA.FRE |
| Number of dimensions | 3 |
| Name for row factor matrix | B:FATLIN.MAT |
| Name for col factor matrix | B:FATCOL.MAT |
| Name for eigenvalue matrix | B:VALCAR.MAT |
| Name for row abs contr. mx. | B:ABSLIN.MAT |
| Name for column abs contr. mx | B:ABSCOL.MAT |
| Name for row corr. sq. mx. | B:CORRSQ.LIN |
| Name for column corr. sq. mx. | B:CORRSQ.COL |
| Listing file | PRN |

O NTSYS-pc ainda apresenta outros programas e qualquer informação a respeito destes outros programas e dos programas apresentados pode ser obtida consultando o manual do NTSYS-pc.

Gráfico 1.10 - Quantidade de P205 presente na composição das Amostras de Rochas

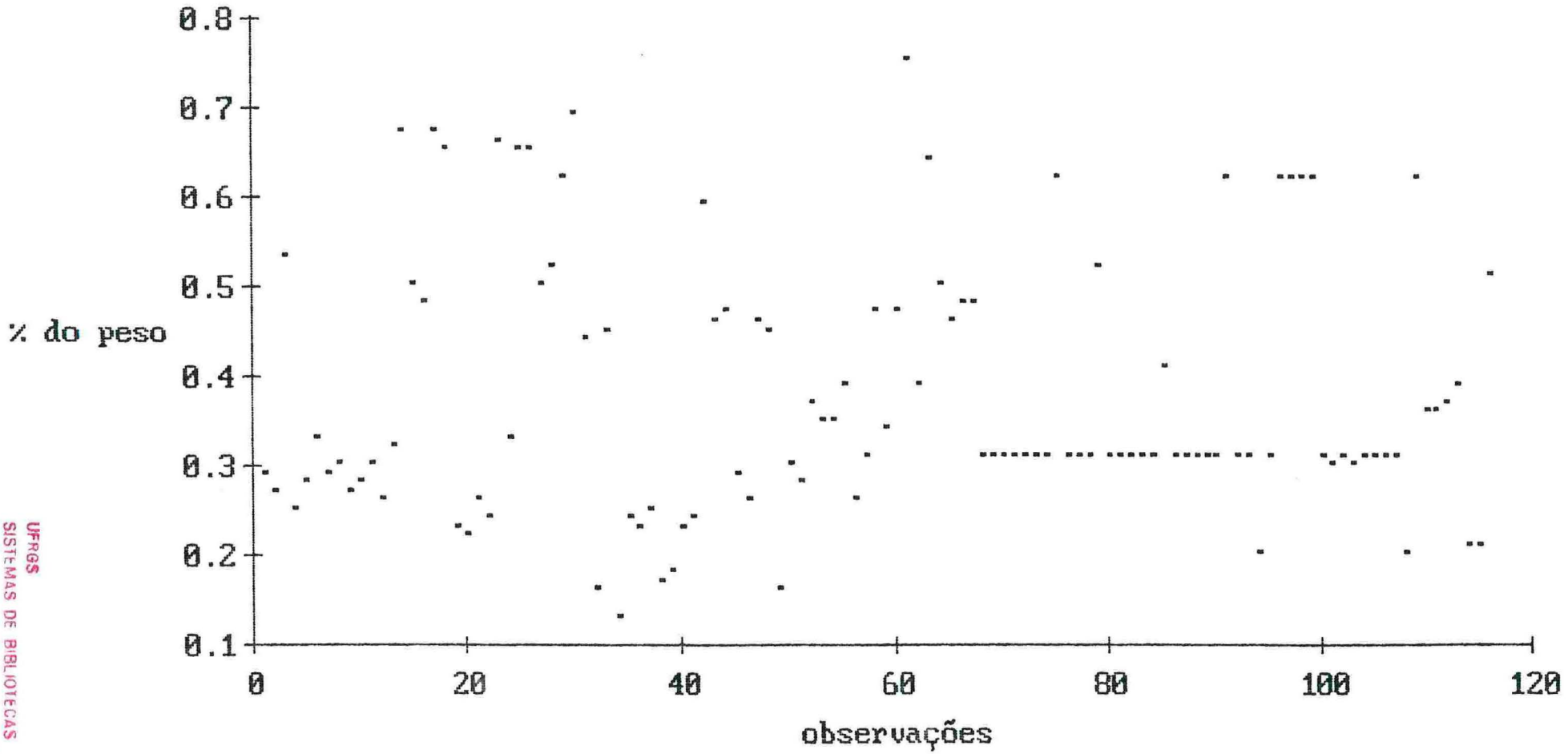


Gráfico 1.11 - Quantidade de Ni presente na composição das Amostras de Rochas

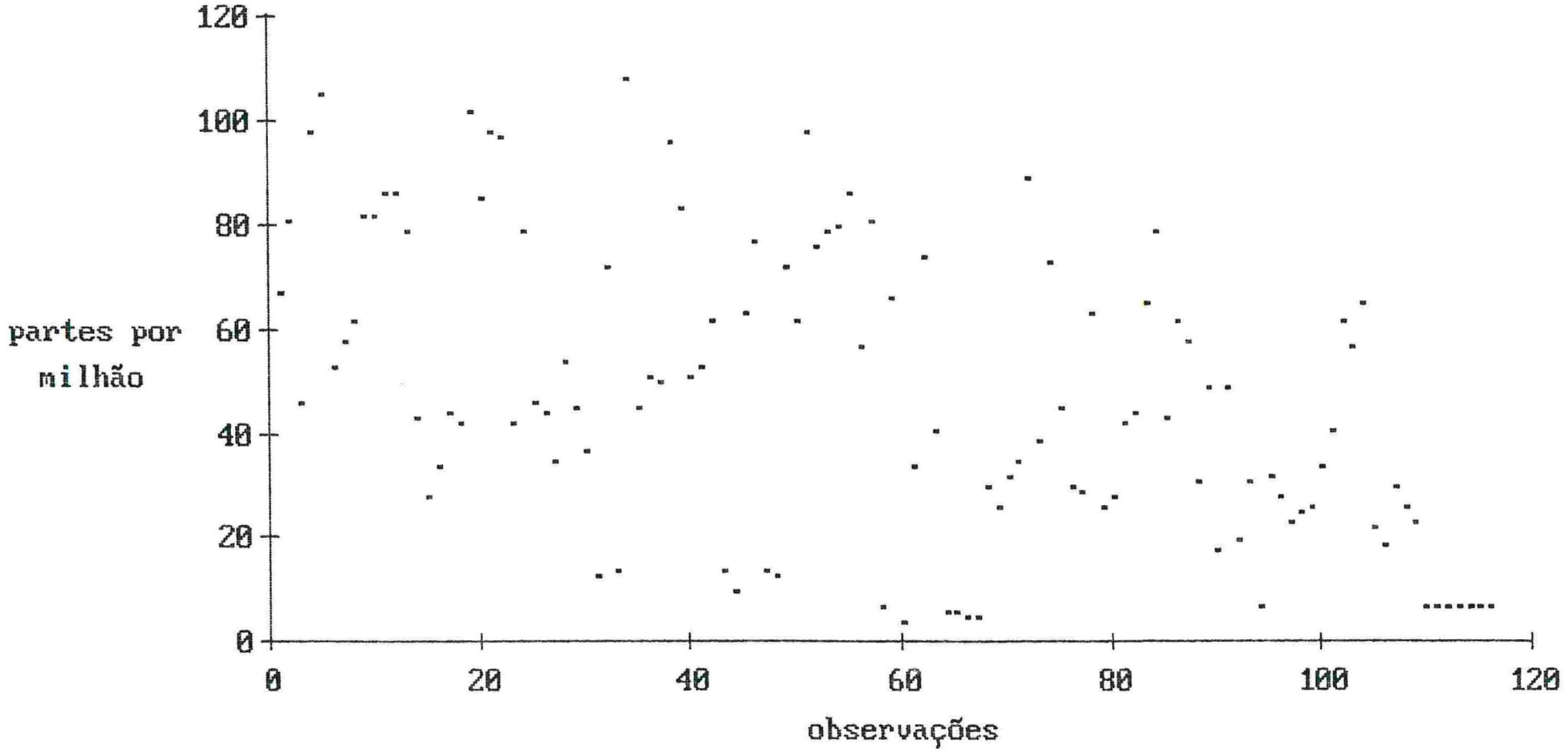


Gráfico 1.12 - Quantidade de Ba presente na composição das Amostras de Rochas

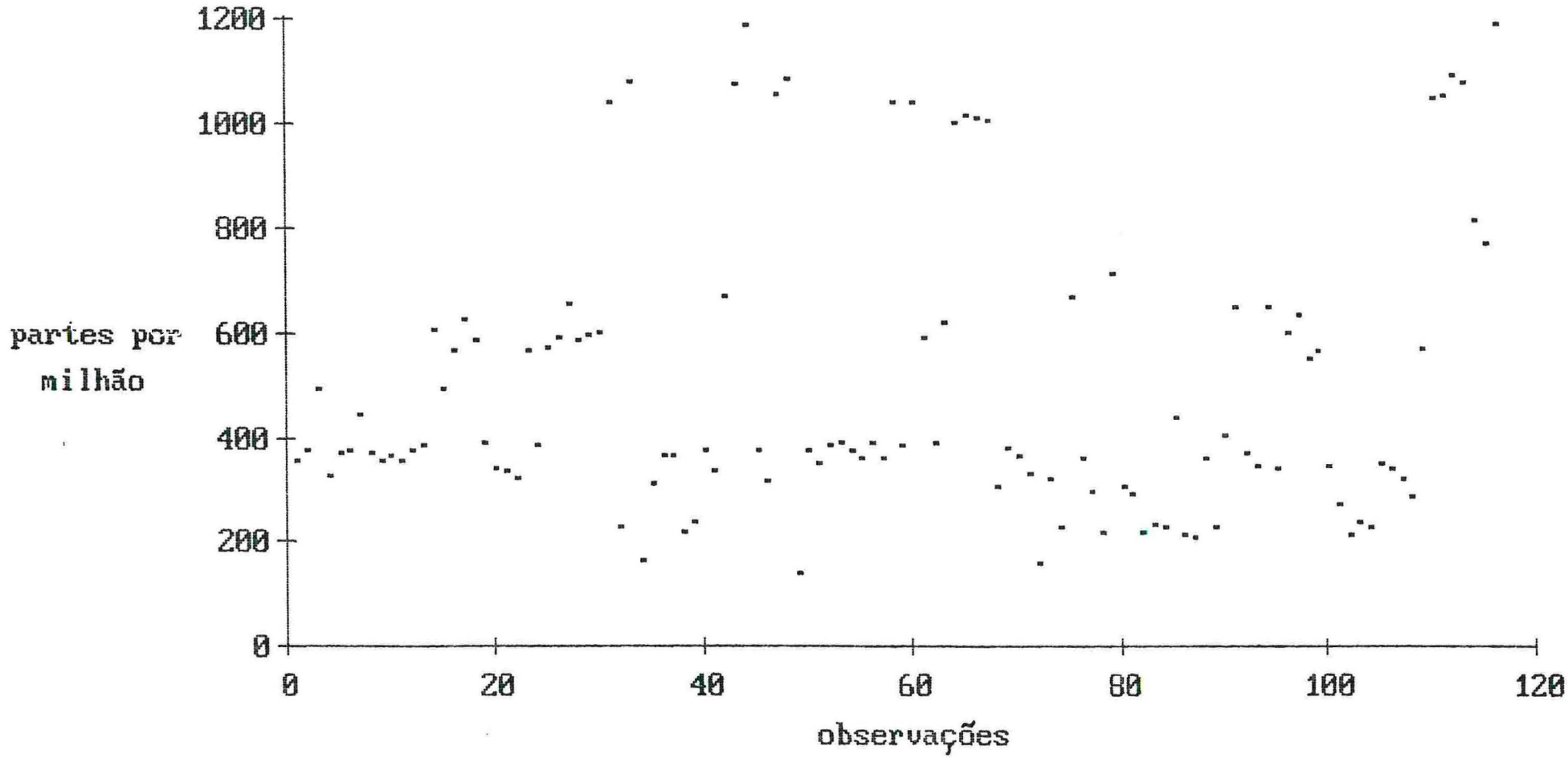


Gráfico 1.13 - Quantidade de Rb presente na composição das Amostras de Rochas

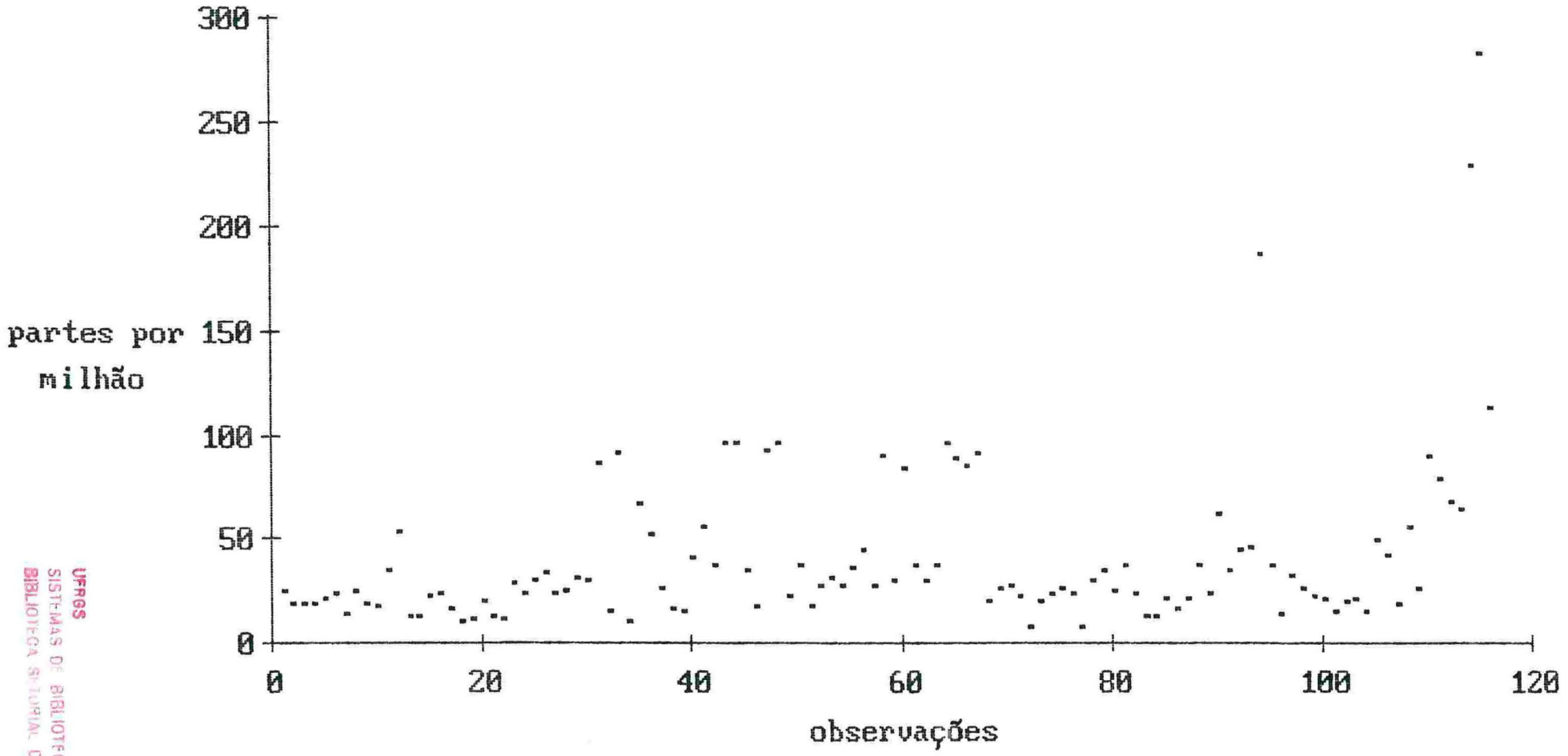


Gráfico 1.14 - Quantidade de Sr presente na composição das amostras de Rochas

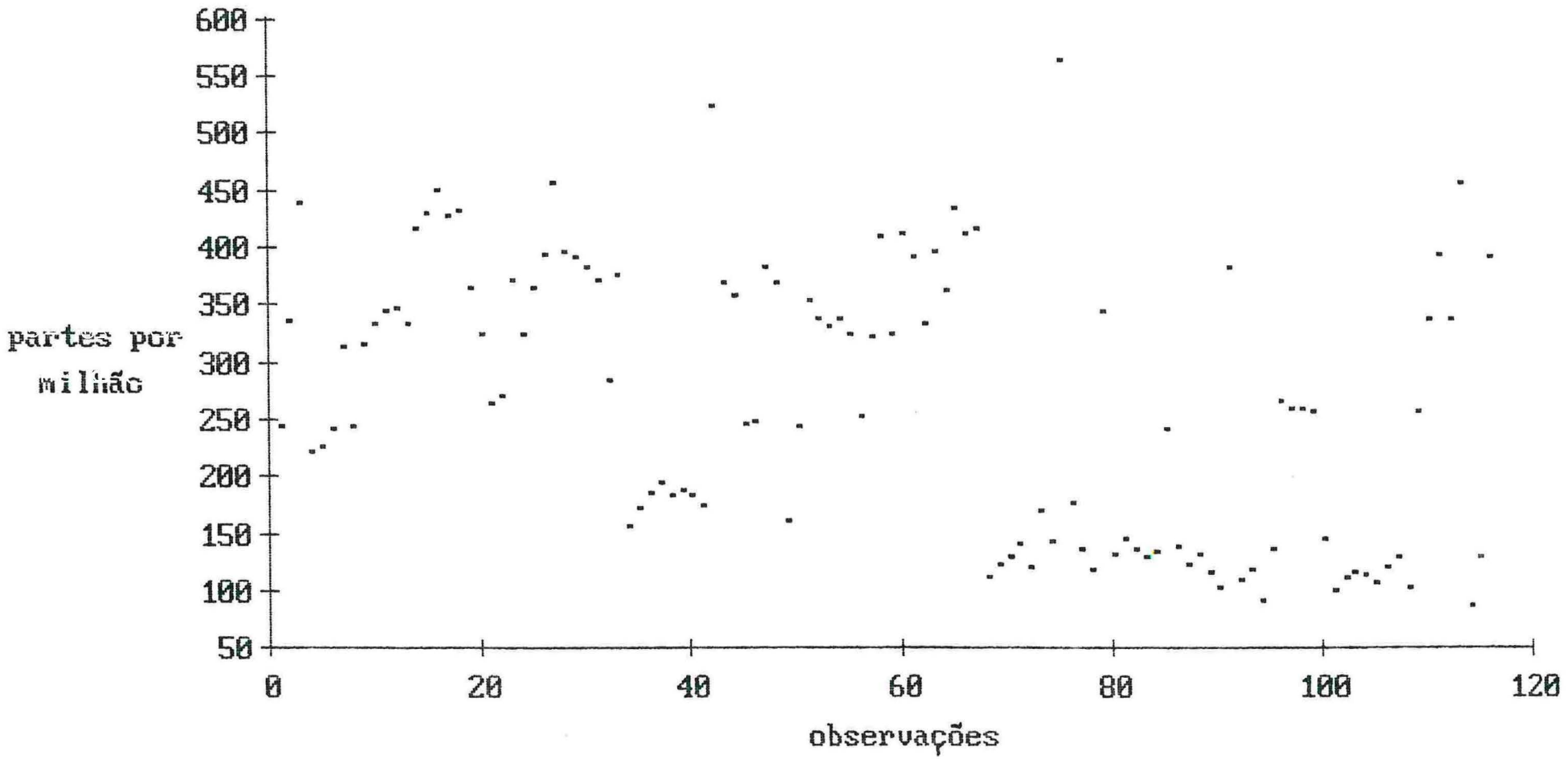
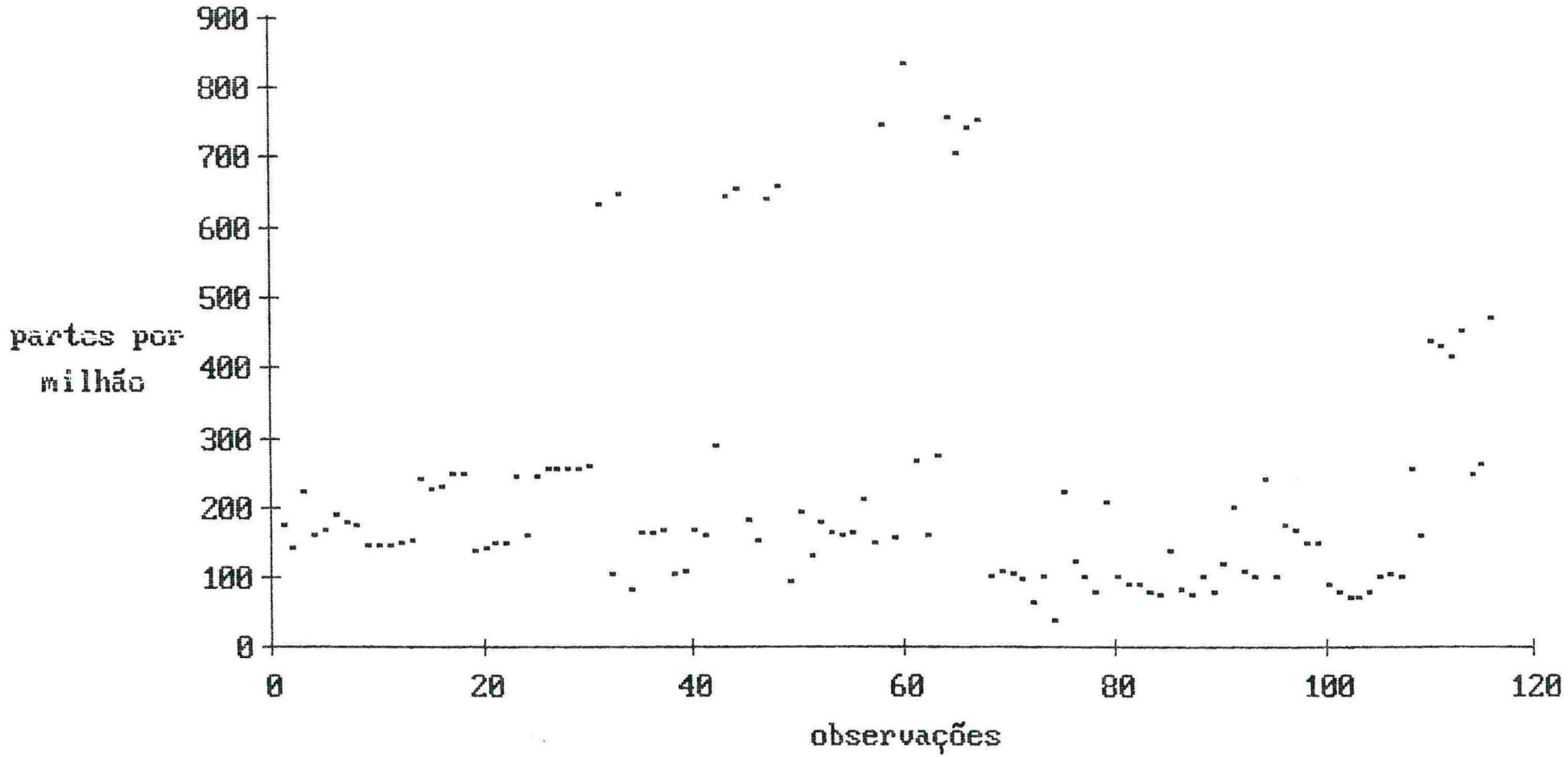


Gráfico 1.15 - Quantidade de Zr presente na composição das amostras de Rochas



## REFERÊNCIAS BIBLIOGRÁFICAS


Bloxom, B. (1968) - Individual Differences in Multidimensional Scaling , *Research Bulletin* 45-68 , Princeton, N.J., Educational Testing Service.

Carroll, J.D. and Chang, J.J. (1964) - A General Index of Nonlinear Correlation and Its Application to the Interpretation of Multidimensional Scaling Solutions , *American Psychologist*, **19**.

Carroll, J.D. and Chang, J.J. (1970) - Analysis of Individual Differences in Multidimensional Scaling via an N-way Generalization of the Eckart-Young Decomposition , *Psychometrika*, 35, 283-319.

Coombs, C.H. (1958) - An Application of a Nonmetric Model for Multidimensional Analysis of Similarities , *Psychological Reports*, **4**, 511-518.

Gower, J.C. (1977) - The Analysis of Asymmetry and Orthogonality, *in Recent Developments in Statistics*, ed. J. Barra, Amsterdam: North-Holland, 109-123.

Harshman, R.A. (1972) - PARAFAC2 : Mathematical and Technical Notes, *in Working Papers in Phonetics*, 22, Los Angeles : University of California.

Heiser, W.J. (1987) - Multidimensional Scaling with Least Absolute Residuals , *Presented at the First Conference of the International Federation of Classification Societies* , Aachen,

Federal Republic of Germany.

Horan, C.B. (1969) - Multidimensional Scaling : Combining Observations when Individuals have Different Perceptual Structures , *Psychometrika*, **34**, 139-165.

Klingberg, F.L. (1941) - Studies in Measurement of the Relations among Sovereign States , *Psychometrika*, 6, 335-352.

Kruskal, J.B. (1964a) - Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypotesis , *Psychometrika*, 29, 1-27.

Kruskal, J.B. (1964b) - Nonmetric Multidimensional Scaling : A Numerical Method , *Psychometrika*, 29, 115-129.

Kruskal, J.B. and Wish, M. (1978) - *Multidimensional Scaling* , Beverly Hills, Sage.

Mac Callum, R.C. and Cornelius III, E.T. (1977) - A Monte Carlo Investigation of Recovery of Structure by ALSCAL , *Psychometrika*, **42**, 401-428.

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979) - *Multivariate Analysis* , London , Academic Press.

Messick, S.J. and Abelson, R.P. (1956) - The Additive Constant Problem in Multidimensional Scaling , *Psychometrika* , **21**, 1-15.

Null, C.H. and Sarle, W. (1982) - Robust Multidimensional Scaling , *Paper Presented at the Joint Meeting of the Psychometric end Classification Societies*, Montreal, Canada.

Ramsay, J.O. (1977) - Maximum Likelihood Estimation in Multidimensional Scaling , *Psychometrika*, **42**, 241-266.

Ramsay, J.O. (1980) - Some Small Sample Results for Maximum Likelihood Estimation in Multidimensional Scaling , *Psychometrika*, **45**, 139-144.

Richardson, M.W. (1938) - Multidimensional Psychophysics , *Psychological Bulletin*, 35, 659-660.

Shepard, R.N. (1962a) - Analysis of Proximities : Multidimensional Scaling with an Unknown Distance Function - I, *Psychometrika*, 27, 125-140.

Shepard, R.N. (1962b) - Analysis of Proximities : Multidimensional Scaling with an Unknown Distance Function - II , *Psychometrika*, 27, 219-246.

Spence, I (1970) - *Multidimensional Scaling : An Empirical and Theoritical Investigation*, Ph.D. Thesis, University of Toronto.

Spence, I. (1972) - A Monte Carlo Evaluation of Three Nonmetric Multidimensional Scaling Algorithms , *Psychometrika*, 37, 461-486.

Spence, I. (1982) - Robust Multidimensional Scaling - *Paper Presented at the Joint Meeting of the Psychometric and Classification Societies*, Montreal, Canada.

Spence, I.A. and Graef, J. (1974) - The Determination of the Underlying Dimensionality of an Empirically Obtained Matrix of Proximities , *Multivariate Behavioral Research*, 9, 331-342.

Spence, I. and Young, F.W. (1978) - Monte Carlo Studies in Nonmetric Scaling , *Psychometrika*, **43**, 1, 115-117.

Takane, Y. (1981) - Multidimensional Successive-Categories Scaling : A Maximum Likelihood Method , *Psychometrika*, **46**, 9-28.

Takane, Y., Young, F.W. and De Leeuw, J. (1977) - Nonmetric Individual Differences Multidimensional Scaling : An Alternating Least Squares Method with Optimal Scaling Features, *Psychometrika*, **42**, 7-67.

Torgerson, W.S. (1962) - *Theory and Methods of Scaling* (2nd ed.), New York, Wiley.

Tucker, L.R. (1964) - The Extension of Factor Analysis to Three-Dimensional Matrices , in N. Frederiksen and H. Gulliksen (eds), *Contibuition to Mathematical Psychology*, New York, Holt, Rinehart and Winston.

Tucker, L.R. (1972) - Relations between Multidimensional Scaling and Three-Mode Factor Analysis, *Psychometrika*, 37, 3-28.

Tucker, L.R. and Messick, S. (1963) - An Individual Differences Model for Multidimensional Scaling, *Psychometrika*, **28**, 333-367.

Wagenaar, W.A. and Padmos, P. (1971) - Quantitative Interpretation of Stress in Kruskal's Multidimensional Scaling

Technique , *Brit. J. Math. and Statist. Psychol.*, 24, 101-110.

Young, G. and Householder, A.S. (1938) - Discussion of a Set of Points in Terms of their Mutual Distances , *Psychometrika*, 3, 19-22.

# BIBLIOGRAFIA


Chatfield, C. and Collins, A.J. (1980) - *Introduction to Multivariate Analysis*, Chapman and Hall, London.

Everitt, B.S. (1980) - *Cluster Analysis* (2nd ed.), Heinemann Educational Books, London.

Everitt, B.S. and Dunn, G. (1983) - *Advanced Methods of Data Exploration and Modelling*, Heinemann Educational Books, New Hampshire.

Gnanadesikan, R. (1977) - *Methods for Statistical Data Analysis of Multivariate Observations*, John Wiley and Sons, New York.

Hair, J.F.Jr., Anderson, R.E. and Tatham, R.L. (1987) - *Multivariate Data Analysis - With Readings* (2nd ed.), Macmillan Publishing Company, New York.

Johnson, R.A. and Wichern, D.W. (1988) - *Applied Multivariate Statistical Analysis* (2nd ed.), Prentice Hall, New Jersey.

Mackay, D., Schofield, N. and Whiteley, P. (1983) - *Data Analysis and the Social Sciences*, Frances Printer, London.

Malzyner, M.S.L. (1981) - *Escalonamento Multidimensional*, Universidade de São Paulo, São Paulo.

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979) - *Multivariate Analysis*, Academic Press, New York.