UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE GEOCIÊNCIAS
ÁREA DE GEOLOGIA DE PETRÓLEO

Este exemplar corresponde a redação fina da tese defendida por Rosane Trajano de Faria e aprovad pela comissão julgadora em 15/01/1993

ORIENTADOR

Dissertação apresentada ao Instituto de Geociências
como requisito parcial à obtenção do título de
Mestre em Geoengenharia de Reservatório

TRATAMENTO DE DADOS MULTIVARIADOS ATRAVÉS DA ANÁLISE DE CORRESPONDÊNCIA EM ROCHAS CARBONÁTICAS

Autora: ROSANE TRAJANO DE FARIA
Orientador: JOSÉ FERREIRA DE CARVALHO †
Co-orientador: PAULO TIBANA †

CAMPINAS
Estado de São Paulo - Brasil
Janeiro, 1993

F 225 t

FARIA, Rosane Trajano de
Tratamento de dados multivariados através da análise de correspondência em rochas carbonáticas. Campinas: Universidade Estadual de Campinas - UNICAMP. Instituto de Geociências. Área de Geologia de Petróleo, 1993.
138 p. (Dissertação de mestrado)
Inclui bibliografia

1. Métodos Estatísticos. 2. Rochas Carbonáticas.

CDD - 552.58 072

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE GEOCIÊNCIAS
ÁREA DE GEOLOGIA DE PETRÓLEO

A dissertação "Tratamento de dados multivariados através da análise de correspondência em rochas carbonáticas", elaborada por Rosane Trajano de Faria e aprovada por todos os membros da Banca Examinadora foi aceita pela Subcomissão de Pós-graduação em Geoengenharia de Petróleo como requisito parcial à obtenção do Título de Mestre em Geoengenharia de Reservatório.

Campinas, 15 de janeiro de 1993.

Banca Examinadora:

José Ferreira de Carvalho, Dr.
(Orientador)

Cláudio Bettini, Dr.
(Examinador)

Joel Carneiro de Castro, Dr.
(Examinador)

Aos meus pais

Rosa e Wilson

ÍNDICE

Pág.

AGRADECIMENTOS i

RESUMO ii

ABSTRACT iii

LISTA DE FIGURAS iv

LISTA DE TABELAS vii

LISTA DE SIGLAS, ABREVIATURAS E SÍMBOLOS ix

CAPÍTULO 1. INTRODUÇÃO 01
1.1 Introdução 01
1.2 Objetivos 05
1.3 Histórico 05
1.4 Trabalhos Anteriores 07

CAPÍTULO 2. DECOMPOSIÇÃO DE MATRIZES EM VALORES SINGULARES 11
2.1 Introdução 11
2.2 Decomposição Espectral de Matrizes Simétricas 13
2.3 Decomposição Ordinária de Matrizes Retangulares 20
2.4 Aproximação de Matrizes por Matrizes de Posto Menor 27
2.5 Decomposição Generalizada de Matrizes Retangulares 28

CAPÍTULO 3. TEORIA DA ANÁLISE DE CORRESPONDÊNCIA 31
3.1 Desenvolvimento Teórico da Análise de Correspondência 31
3.1.1 Construção das Nuvens de Pontos 34
3.1.2 Definição de Distância 38
3.1.3 Ajustamento das Nuvens de Pontos 39
3.1.4 A Dualidade do Método 43
3.1.5 Princípio da Equivalência Distributiva 48
3.2 Interpretação dos Resultados 50
3.2.1 Inércia 50
3.2.2 Contribuição Absoluta 52

3.2.3 Contribuição Relativa 53
3.3 Perfil Suplementar 55
3.4 Análise de Correspondência Múltipla 55
3.4.1 Geometria das Colunas de Z 58
3.4.2 Geometria das Linhas de Z 59
3.4.3 Tabela de Burt 60

CAPÍTULO 4. APLICAÇÃO DA ANÁLISE DE CORRESPONDÊNCIA 63
4.1 Os Dados 63
4.2 Apresentação dos Resultados 75
4.2.1 Análise de Correspondência Simples Variável x Variável 75
4.2.2 Análise de Correspondência Simples Poço x Variável 85
4.2.3 Análise de Correspondência Múltipla 91
4.2.4 Análise de Correspondência Simples das Fácies Sedimentares 93
4.2.5 Análise de Correspondência Simples das Fácies por Cronozona . . 93

CAPÍTULO 5. CONCLUSÕES E DISCUSSÃO 101
5.1 Conclusões 101
5.2 Discussão 103

CAPÍTULO 6. REFERÊNCIAS BIBLIOGRÁFICAS 105

APÊNDICE 113
A.1 Conceitos Geométricos no Espaço Multidimensional 113
A.1.1 Ângulo, Distância e Produto Escalar 113
A.1.2 Espaço Euclidiano Ponderado 116
A.1.3 Associando Massa aos Vetores 118
A.1.4 Identificando Subespaços Ótimos 121

PROGRAMAS 123
P.1 Programa de Categorização dos Dados da Fm Lagoa Feia 123
P.2 Programa de Definição de Fácies Sedimentares 126
P.3 Programa de Fácies Sedimentares simplificadas 130
P.4 Programa de Realização da Análise de Correspondência no SAS 131
P.5 Programa de Análise de Correspondência (SAS/IML) 132

ANEXOS 134
AX.1 Arquivo de Saída do PROC CORRESP do SAS 134

AGRADECIMENTOS

Aos meus orientadores, JOSÉ FERREIRA DE CARVALHO e PAULO TIBANA, da Universidade Estadual de Campinas, pela dedicação, estímulo e paciência durante a realização desta tese.

A UYARA MUNDIM PRAÇA e MARIA DOLORES CARVALHO, da PETROBRÁS, pela cessão dos dados utilizados neste trabalho e discussão dos resultados.

A MOACIR AMÉRICO CORNETTI pela confecção dos desenhos e pela ajuda na edição do texto.

Aos amigos e colegas de curso pela convivência agradável durante este período, em especial a: Valcir, Flávio, Senira, Paulo e Maria Baldissera, Eduardo e Tânia Edelwein, Edna e Cris.

À PETROBRÁS pela oportunidade que me foi dada em participar deste curso de mestrado.

Ao CONVÊNIO UNICAMP/PETROBRÁS pelo apoio técnico-científico.

TRATAMENTO DE DADOS MULTIVARIADOS ATRAVÉS DA ANÁLISE DE CORRESPONDÊNCIA EM ROCHAS CARBONÁTICAS

(TESE DE MESTRADO: Dez./92)

ROSANE TRAJANO DE FARIA

Orientador: Dr. José Ferreira de Carvalho

Co-orientador: Prof. Paulo Tibana

Instituto de Geociências - Curso de Geoengenharia de Reservatórios

Convênio UNICAMP/PETROBRÁS

RESUMO

A análise de correspondência é uma técnica exploratória de grandes matrizes de dados. Sua aplicação resulta em representações gráficas simultâneas das amostras e das variáveis de uma matriz de dados no mesmo plano fatorial, possibilitando uma interpretação fácil das relações entre elas.

A análise de correspondência é uma técnica, mais geométrica do que estatística, que utiliza distâncias Euclidianas ponderadas para analisar as similaridades entre os pontos no espaço. É mais indicada para variáveis discretas, mas pode ser aplicada a dados contínuos desde que estes sejam adequadamente discretizados.

O objetivo deste trabalho é divulgar a técnica através de sua aplicação a um conjunto de dados geológicos referente à descrição macroscópica de rochas carbonáticas da Formação Lagoa Feia na Bacia de Campos.

Através dos resultados obtidos, pode-se agrupar poços com características semelhantes em relação às diversas variáveis e fazer o mapeamento da distribuição destes grupos.

MULTIVARIATE DATA ANALYSIS - AN APPLICATION OF CORRESPONDENCE ANALYSIS TO CARBONATE ROCKS

(MASTER THESIS: Dec./92)

ROSANE TRAJANO DE FARIA

Adviser: Dr. José Ferreira de Carvalho

Co-adviser: Prof. Paulo Tibana

Instituto de Geociências - Curso de Geoengenharia de Reservatórios

Convênio UNICAMP/PETROBRÁS

ABSTRACT

Correspondence analysis is an exploratory tool for the analyses of large data matrices. Its application results in simultaneous graphical displays of objects and variables of data matrices on the same coordinate plane, providing easy interpretation of the relations among them.

The great advantage of graphical representation of a data matrix is to provide easier interpretation rather than numerical descriptions. Besides, it is a good method to summarize, simplify and allow global understanding of matrix information.

Correspondence analysis is more suitable for discrete variables, but it can be applied to continuous variables as well, after adequate transformations.

This technique was applied to a set of data comprising macroscopic core description of a sequence of carbonate rocks from Lagoa Feia Formation, Campos basin.

From the results of this analysis, it was possible to identify groups of wells having similar behavior with respect to several variables and to map the geographic distribution of those groups.

LISTA DE FIGURAS

Pág.

1.1: Representação das amostras no espaço das variáveis; representação das variáveis no espaço das amostras 2

2.1: Representação aproximada da relação entre os indivíduos de **A** ou análise de componentes principais 16

2.2: Representação bidimensional da relação entre as variáveis de **A** ou análise de componentes principais das colunas de **A** 19

2.3: Representação bidimensional das linhas de **X** 24

2.4: Representação bidimensional das colunas de **X** 24

2.5: Biplot ou análise RQ-modal das linhas e colunas de **X** 25

2.6: Fluxograma da relação entre R, Q e RQ-modal 26

3.1: Histogramas das categorias da variável tamanho de conchas em relação à variável matriz terrígena 35

3.2: Disposição dos pontos-linha (categorias da variável tamanho de conchas) no espaço tridimensional das colunas (variável matriz terrígena) 36

3.3: Configuração dos 8 pontos-linha no espaço Euclidiano ponderado das colunas 39

3.4: Análise de correspondência simples dos pontos-linha (tamanho de conchas) 42

3.5: Análise de correspondência dos pontos-coluna (matriz terrígena) 46

3.6: Análise de correspondência entre as linhas e colunas da tabela de contingência **N** (tamanho de conchas x matriz terrígena) 49

3.7: Representação unidimensional das linhas e colunas da tabela 3.1 51

3.8: Coordenada do i-ésimo perfil-linha relativa ao k-ésimo eixo principal o qual está a uma distância d_i do centróide c. 54

3.9: Fluxograma do resumo da análise de correspondência 56

3.10: Análise de correspondência múltipla das variáveis tamanho de conchas e matriz terrígena 59

4.1: Localização da Bacia de Campos e dos campos produtores na Formação Lagoa Feia (Pampo, Badejo, Linguado e Trilha). 64

4.2: Seqüências Deposicionais da Formação Lagoa Feia, seção tipo 66

4.3: Modelo deposicional esquemático da Seqüência das Coquinas 67

4.4: Localização dos poços amostrados na Bacia de Campos 68

4.5: Análise de correspondência simples das variáveis granulometria e quantidade de matriz carbonática 76

4.6: Análise de correspondência simples das variáveis granulometria e quantidade de matriz terrígena 77

4.7: Análise de correspondência simples das variáveis granulometria e quantidade de concha aberta e fechada 78

4.8: Análise de correspondência simples das variáveis granulometria e tamanho de concha 79

4.9: Análise de correspondência simples das variáveis granulometria e espessura de concha 80

4.10: Análise de correspondência simples das variáveis granulometria e empacotamento 81

4.11: Análise de correspondência simples das variáveis quantidade de matriz carbonática e matriz terrígena 82

4.12: Análise de correspondência simples das variáveis quantidade de matriz carbonática e empacotamento 83

4.13: Análise de correspondência simples das variáveis quantidade de matriz terrígena e empacotamento 84

4.14: Análise de correspondência simples dos poços com a variável tamanho de conchas, sem ponderar pela espessura 86

4.15: Análise de correspondência simples dos poços com a variável tamanho de conchas, ponderada pela espessura 87

4.16: Mapeamento da variável tamanho de conchas com base nos agrupamentos obtidos da análise de correspondência 88

4.17: Análise de correspondência simples dos poços com a variável matriz terrígena, ponderada pela espessura 89

4.18: Mapeamento dos poços com relação à variável quantidade de matriz terrígena, ponderada pela espessura 90

4.19: Análise de correspondência múltipla de todas as variáveis 92

4.20: Análise de correspondência simples dos poços com as fácies sedimentares, ponderada pela espessura 94

4.21: Mapeamento da distribuição das fácies sedimentares 95

4.22: Análise de correspondência simples dos poços com as fácies simplificadas, da cronozona CB 96

4.23: Análise de correspondência simples dos poços com as fácies simplificadas, da cronozona CC 97

4.24: Análise de correspondência simples dos poços com as fácies simplificadas, da cronozona CD 98

4.25: Análise de correspondência simples dos poços com as fácies simplificadas, da cronozona CE 99

4.26: Análise de correspondência simples dos poços com as fácies simplificadas, da cronozona CF 100

5.1: Fluxograma da análise fatorial 103

A.1: Distância e ângulo entre dois vetores **a** e **b** 114

A.2: Pontos **a** e **b** no espaço bidimensional, vetor resultante da diferença e distância entre os pontos 115

A.3: Pontos no espaço multidimensional e suas projeções em um subespaço, representado por um plano 122

LISTA DE TABELAS

Pág.

2.1: Medidas hipotéticas feitas em quatro indivíduos 14

2.2: Comparação entre a decomposição ordinária, espectral e generalizada . 29

3.1: Tabela de contingência **N**, das amostras da Formação Lagoa Feia 32

3.2: Matriz de correspondência **P** 33

3.3: Matriz dos perfis-linha **R**, obtida pela divisão de cada elemento de **N**, pelo total de sua respectiva linha 36

3.4: Tabela de contingência $\mathbf{N}^T$, das amostras da Formação Lagoa Feia 44

3.5: Matriz dos perfis-coluna **C**, obtida pela divisão de cada elemento de $\mathbf{N}^T$ pelo total de sua respectiva coluna 44

3.6: Inércia dos pontos com relação ao primeiro eixo principal e suas contribuições 53

3.7: Matriz indicadora **Z** 57

3.8: Tabela de Burt - análise de correspondência múltipla 61

4.1: Parte da matriz original de dados da Formação Lagoa Feia 69

4.2: Categorias da variável granulometria 70

4.3: Categorias da variável energia deposicional 71

4.4: Categorias da variável matriz terrígena 72

4.5: Categorias da variável forma de conchas 72

4.6: Categorias da variável tamanho de conchas 73

4.7: Categorias da variável tipo de rocha 73

4.8: Categorias da variável espessura de conchas 74

4.9: Categorias da variável empacotamento 74

LISTA DE SIGLAS, ABREVIATURAS E SÍMBOLOS

Termos granulométricos

GRAN	: granulometria
SXO	: seixo
GNL	: grânulo
MGO	: muito grosseiro
GRO	: grosseiro
MED	: médio
FNO	: fino
MFN	: muito fino
SLT	: silte
ARG	: argila

Termos Texturais

SCO	: sem conchas
ARSO	: carbonatos arenosos
MUIA	: carbonatos muito arenosos
PURO	: carbonatos puros
G/P	: grainstones a packstones
WST	: wackestones
MST	: mudstones
EMP	: empacotamento
FRX	: frouxo
NOR	: normal
DEN	: denso
MAT	: quantidade de matriz
MTE	: quantidade de matriz terrígena
MCA	: quantidade de matriz carbonática
CON	: quantidade de conchas
ABE	: quantidade de conchas abertas
FEC	: quantidade de conchas fechadas
QUE	: quantidade de conchas quebradas
INT	: quantidade de conchas inteiras
TIPO	: tipo de rocha
NTAM	: tamanho de conchas (AA,...GG)
NESPC	: espessura de conchas (A,...G)

Tipos de Rocha

CLU	: calcilutito
CRE	: calcarenito
C	: calcirrudito
CMC	: calcirrudito muito calcarenítico
CC	: calcirrudito calcarenítico
SLX	: silexito
REC	: carbonatos recristalizados
GST	: gastrópodes
BIO	: carbonatos bioclásticos
AREN	: terrígenos

Termos Estatísticos

SVD	: decomposição de uma matriz em seus valores e vetores singulares
R-modal	: análise das variáveis de uma tabela de contingência
Q-modal	: análise dos objetos de uma tabela de contingência
RQ-modal	: análise simultânea das variáveis e objetos de uma tabela de contingência

Termos Matemáticos

Z	: matriz indicadora
N	: tabela de contingência
P	: matriz de correspondência
R	: matriz dos perfis-linha
C	: matriz dos perfis-coluna
D	: matriz diagonal
1	: matriz quadrada de uns
r	: vetor das massas das linhas
c	: vetor das massas das colunas
a e **b**	: vetor ponto
i	: número de linhas de uma matriz
j	: número de colunas de uma matriz
n	: total de elementos de uma tabela de contingência
n_{ij}	: elemento de uma tabela de contingência
n_{i+}	: total de elementos das linhas de uma tabela de contingência
n_{+j}	: total de elementos das colunas de uma tabela de contingência
r	: posto de uma matriz
$\mathbf{X}^T$	: transposta da matriz **X**
$\mathbf{D}_r^{-1}$	: inversa da matriz diagonal dos elementos de **r**
α, ϕ, ω	: valores singulares de uma matriz
λ	: autovalores de uma matriz

PETROBRÁS	: Petróleo Brasileiro S/A
CENPES	: Centro de Pesquisas e Desenvolvimento Leopoldo A. Miguez de Mello
DEPEX	: Departamento de Exploração/Petrobrás

1 - INTRODUÇÃO

1.1 INTRODUÇÃO

No estudo de um determinado fenômeno, procura-se medir todos os aspectos considerados relevantes para o seu entendimento; posteriormente, emprega-se a estatística para resumir, simplificar e eventualmente explicar esse conjunto de dados. O entendimento das estruturas ou relações identificadas no fenômeno permite conhecê-lo melhor e fazer previsões sobre o seu comportamento.

Na geologia, é muito comum trabalhar-se com grandes quantidades de dados de natureza multivariada. Por exemplo, quando se descreve uma determinada seção de rocha carbonática com o objetivo de determinar o seu ambiente de deposição, procura-se quantificar parâmetros considerados importantes tais como: quantidade e tipos de grãos, matriz, cimento e estruturas sedimentares. Geralmente estes dados estão arranjados na forma de tabelas ou matrizes, onde as linhas representam as amostras e as colunas representam as variáveis que foram quantificadas.

Desta massa de dados, quer-se extrair informações que permitam identificar grupos de amostras ou reconhecer comportamentos similares entre as variáveis. Tendo-se reconhecido variabilidade nos dados, pode-se identificar a razão desta variabilidade e mostrar as possíveis causas - fatores principais - responsáveis pela estrutura dos dados. Estes fatores podem ser, por exemplo, as diferentes condições do ambiente de deposição, tais como: nível de energia, profundidade, luminosidade, temperatura e salinidade da água, influxo de terrígenos, dentre outros.

As representações gráficas, tais como diagramas de correlação, histogramas, diagramas triangulares e tetraédricos, ou mesmo os gráficos provenientes da **análise fatorial**, são de grande utilidade no reconhecimento das

estruturas de um conjunto de dados. O ponto comum entre estas representações é que as amostras podem ser referidas ao espaço p-dimensional das p variáveis e as variáveis podem ser referidas ao espaço n-dimensional das n amostras (figura 1.1).

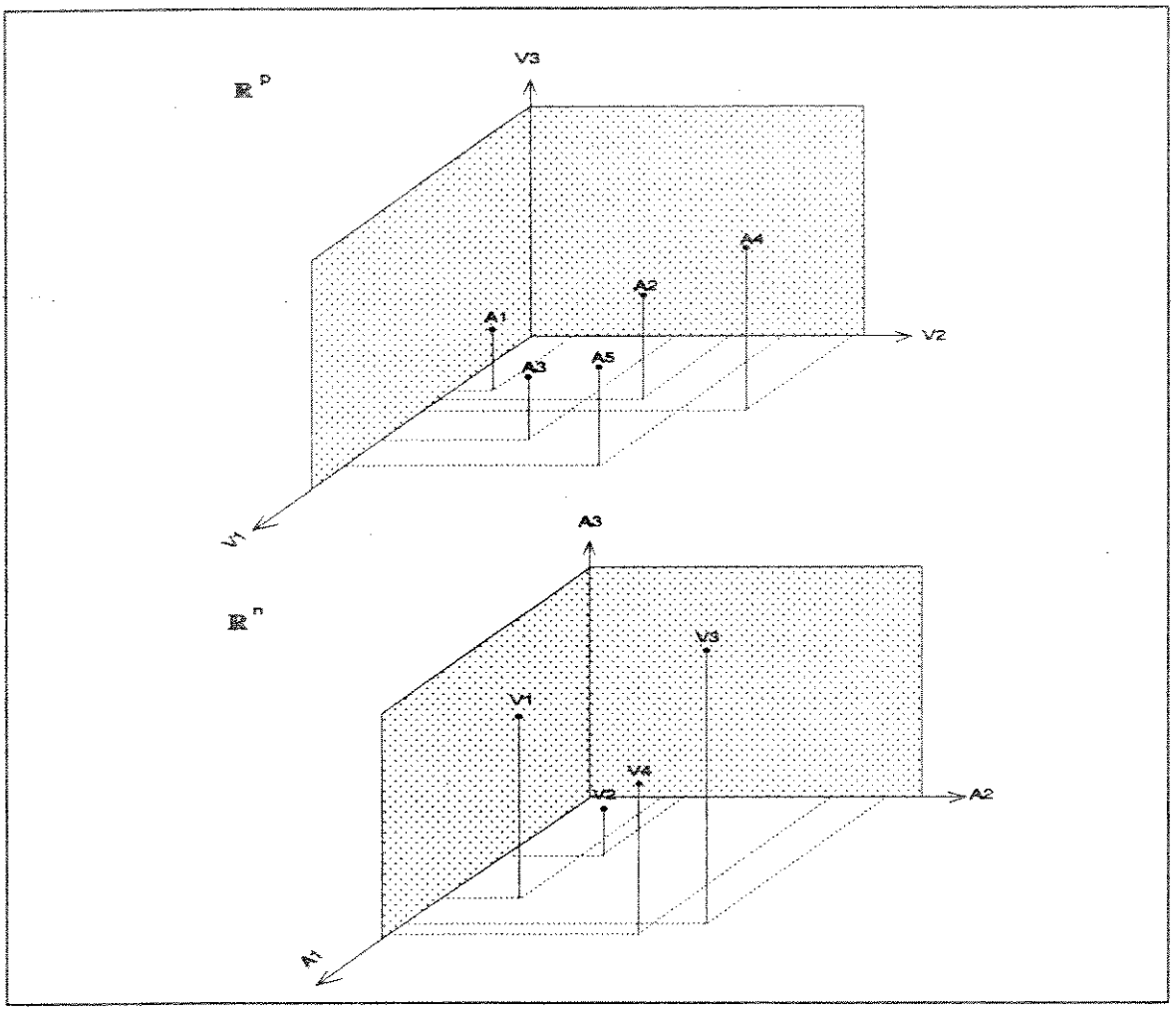

Figura 1.1: a) Representação das amostras no espaço das variáveis ($\mathbb{R}^p$); b) representação das variáveis no espaço das amostras ($\mathbb{R}^n$).

Nota-se que a estrutura das duas representações é a mesma, sendo os pontos ora representados pelas amostras ora pelas variáveis, o mesmo acontecendo com os espaços. Os dois espaços nos quais se representa o conjunto de dados satisfazem uma relação dual. Esta dualidade significa que se pode estudar o problema ora em um espaço ora no outro.

Neste tipo de representação obtém-se uma nuvem de pontos referentes às amostras ou às variáveis. A forma da nuvem ou o aparecimento de grupos isolados podem indicar as relações entre os pontos, levando a uma melhor intrepretação geológica dos dados. Estas relações são medidas através das distâncias entre os pontos que representam, por exemplo, as variáveis que se deseja analisar. Logo, é essencial a compreensão do modelo utilizado na determinação destas distâncias.

No entanto, dados multidimensionais são impossíveis de serem visualizados. Um dos objetivos da **análise fatorial** é reduzir o número de dimensões do espaço com o qual se está trabalhando. O que se deseja é recuperar um certo número de feições consideradas essenciais, em um espaço de dimensão menor. Ou ainda, deseja-se a melhor representação possível do espaço das amostras e do espaço das variáveis neste espaço de dimensão reduzida. Para isto, projeta-se a nuvem de pontos sobre um plano, no caso de representações bidimensionais. Estes planos são selecionados pela sua capacidade de preservar ao máximo as distâncias entre os pontos, refletindo, o melhor possível, as relações entre eles. A representação gráfica obtida a partir da projeção dos pontos no plano permite visualizar a distribuição das variáveis e das amostras, bem como as relações entre elas.

Selecionar estes planos, matematicamente, significa decompor a matriz de distâncias entre os pontos em seus valores e vetores singulares. Cada eixo do plano selecionado responde por uma porcentagem da variância total dos dados e permite avaliar a capacidade que cada eixo tem de representar a nuvem de pontos. Quanto maior o percentual, melhor a representação dos pontos no espaço e, conseqüentemente, das relações existentes entre os pontos.

Existe uma grande variedade de técnicas de análise de dados multivariados, cujo princípio fundamental é a decomposição de matrizes em valores singulares (geralmente referida como SVD - *Singular Value Decomposition*). Essa decomposição visa reduzir a dimensionalidade da matriz de dados, através do cálculo dos seus valores e vetores singulares, para que ela possa ser representada graficamente, facilitando, conseqüentemente, sua interpretação.

O arcabouço da SVD engloba uma grande variedade de técnicas

multidimensionais, unificando o que aparentemente parece ser diferente. A SVD é a base da análise de componentes principais, do biplot, da análise de correlação canônica, da análise canônica e da análise de correspondência. Essas técnicas são todas variações sobre o mesmo tema: o tema é a álgebra e a geometria da SVD (Greenacre, 1984).

A **análise de correspondência**, objeto de estudo deste trabalho, é um método de análise fatorial que considera simultaneamente as relações entre as linhas e as colunas de uma tabela de contingência. Ela é especialmente indicada para descrever matrizes com grande volume de dados discretos ou categóricos e sem uma estrutura claramente definida *a priori*. Ela permite visualizar as relações mais importantes de um grande conjunto de variáveis entre si ou as relações entre as amostras e as variáveis. Os resultados são apresentados sob a forma de gráficos onde estão representadas as variáveis e/ou amostras; a medida da distância entre os pontos no gráfico permite estabelecer as similaridades e diferenças entre eles. Essa técnica pode também ser aplicada a variáveis contínuas, desde que estas estejam devidamente classificadas ou categorizadas.

A forma básica da análise de correspondência consiste na sua aplicação a tabelas de contingência de dupla entrada, quando então é conhecida como **análise de correspondência simples**. No entanto, pode também ser aplicada a tabelas de contingência de múltiplas entradas - **análise de correspondência múltipla**.

A análise de correspondência pode ser vista como uma variação da análise de componentes principais e da análise fatorial, diferenciando-se destas, entre outros aspectos, por permitir a inclusão de variáveis categóricas. Segundo Greenacre (1984), a geometria da análise de correspondência simples fornece as regras básicas para a sua interpretação. Todas as outras formas de análise de correspondência são aplicações do mesmo algoritmo a outros tipos de matrizes de dados, com adaptações na interpretação.

Ainda de acordo com Greenacre (op. cit.), a análise de correspondência é uma técnica de exploração de dados multivariados, razoavelmente simples do ponto de vista matemático e computacional. É uma técnica mais geométrica do que estatística, que converte uma matriz de dados não negativos em um tipo particular

de gráfico bidimensional, onde as linhas e as colunas são representadas por pontos.

A análise de correspondência é também um método de análise fatorial RQ-modal que utiliza um escalonamento no espaço Euclidiano ponderado.

1.2 OBJETIVOS

O objetivo deste trabalho é estudar a **análise de correspondência** e aplicá-la a um conjunto de dados geológicos. Os dados escolhidos referem-se à descrição macroscópica de rochas carbonáticas da Formação Lagoa Feia, Bacia de Campos. Esse conjunto de dados é composto pela descrição das amostras coletadas ao longo dos testemunhos de 28 poços distribuídos na bacia, sendo que cada amostra é caracterizada por diversas variáveis.

A grande quantidade de informação contida nesse conjunto de dados, uma matriz com 1415 linhas por 25 colunas, faz dele um objeto adequado para aplicação de técnicas de análise exploratória de dados.

Como a análise de correspondência permite projetar amostras e variáveis em um mesmo gráfico, considerando informações quantitativas e qualitativas, ela torna-se uma ferramenta mais poderosa do que as demais. Através de sua aplicação, pode-se analisar as relações existentes entre os poços e as variáveis, o que permite agrupar poços com características semelhantes. O estudo das relações entre as variáveis entre si permite identificar quais delas são mais importantes, por exemplo, na definição de fácies sedimentares. A análise das relações entre as fácies e os poços permite mapear sua distribuição espacial ao longo da bacia.

1.3 HISTÓRICO

A origem da análise de correspondência é algo confuso na literatura estatística. Isto se deve ao fato de ela ter sido desenvolvida independentemente por vários autores, em contextos diferentes e sob diversas denominações.

No entanto, existe um consenso em atribuir a Hartley (1935) a origem da análise de correspondência. No artigo publicado em seu nome original alemão (Hirschfeld), ele apresentou uma formulação algébrica da correlação entre linhas e colunas de uma tabela de contingência. Quase na mesma época, Richardson e Kuder (1933) e Horst (1935), independentemente, tinham sugerido idéias semelhantes.

Fisher (1940) desenvolveu a mesma teoria na forma de análise discriminante, sobre uma tabela de contingência, em um estudo clássico sobre a relação entre as cores de olhos e cabelos de um grupo de crianças.

Pouco depois, Guttman (1941) desenvolveu a mesma teoria, em outro contexto, tratando o caso geral de mais de duas variáveis qualitativas, o que pode ser hoje relacionado à análise de correspondência múltipla. Fisher e Guttman apresentaram essencialmente a mesma teoria em contextos diferentes (biométrica e psicometria, respectivamente). Estas duas escolas dividiam a mesma teoria matemática e os mesmos procedimentos computacionais da análise de correspondência, mas geravam resultados numéricos e não resultados gráficos.

Nos anos 40 e 50, novas abordagens matemáticas foram desenvolvidas, principalmente por Guttman e seus colaboradores. Hayashi (1952) introduziu a técnica no Japão, sob a denominação de "quantificação de dados qualitativos".

A utilização da análise de correspondência sob denominações diversas ocasionou uma confusão na literatura, que só foi esclarecida na metade dos anos 60, através de Jean-Paul Benzécri, que apresentou uma abordagem geométrica do método. O termo francês *correspondence* foi usado para expressar um "sistema de associação entre elementos de dois conjuntos", no caso, linhas e colunas. Na verdade, o termo representa uma entidade matemática específica denominada de tabela de contingência.

Essa abordagem foi rapidamente incorporada pelos estatísticos franceses. O grupo liderado por Benzécri adquiriu experiência prática em análise de correspondência e vários trabalhos estão publicados no "*Les Cahiers de l'Analyse des Données*". Entretanto, o estilo matemático, dotado de uma notação algébrica extremamente rigorosa, dificultou sua divulgação em outros países.

Mais recentemente, a abordagem francesa da análise de correspondência

tem recebido maior atenção. Ela foi discutida muito ampla e competentemente na literatura inglesa por Greenacre (1984). O livro de Lebart, Morineau e Warwick (1984) foi traduzido para o inglês e existem inúmeros artigos em jornais estatísticos que comparam a abordagem francesa com a anglo-americana. A contribuição japonesa foi revista por Nishisato (1980).

Atualmente, essa técnica é bem conhecida e pode ser facilmente encontrada na literatura. Uma revisão histórica completa pode ser encontrada em Rijckevorsel e De Leeuw (1988).

1.4 TRABALHOS ANTERIORES

Várias técnicas de análise de dados multivariados têm sido aplicadas na geologia desde o início dos anos 60. Dentre as mais utilizadas, destacam-se a análise de componentes principais, análise de agrupamento (*cluster analysis*), análise discriminante e análise de regressão. Todas elas encontram aplicações na geoquímica, paleontologia, sedimentologia e, mais recentemente, em geologia de petróleo, com alguns trabalhos realizados no âmbito da PETROBRÁS. Dentre os últimos cabe ressaltar os trabalhos de Souza Jr. (1988, 1991) e Bucheb (1991).

Dentre as inúmeras aplicações de análise de dados multivariados na geologia, cabe destacar-se o clássico trabalho de Imbrie e Purdy (1962), utilizando pela primeira vez a análise fatorial Q-modal para classificar sedimentos carbonáticos do Recente das Bahamas. Em Davis (1986), encontram-se várias referências sobre os artigos publicados nesta área.

Só muito recentemente, a análise de correspondência tornou-se viável computacionalmente, de modo que suas aplicações na geologia são relativamente poucas e recentes.

A análise de correspondência foi introduzida na geologia através de David, Campiglio e Darling, em 1974. Eles apresentam o método, o programa e uma aplicação a dados geoquímicos coletados em um corpo intrusivo posteriormente metamorfizado, o batólito Bourlamaque, no Canadá. A aplicação é usada para testar

o método e mostrar sua capacidade de separar e identificar processos geológicos tais como magmatismo e metamorfismo.

Teil (1975) apresenta uma abordagem teórica da análise de correspondência e algumas considerações sobre a interpretação dos resultados gerados por ela. Também compara a análise de correspondência com a análise de componentes principais, e faz comentários sobre os tipos de dados adequados para a aplicação dessa técnica.

Teil e Cheminee (1975) mostram um exemplo de aplicação da análise de correspondência a dados geoquímicos de uma série vulcânica na Etiópia. Os dados utilizados consistem em análises químicas de óxidos e elementos traços das amostras coletadas nesta área. Primeiro eles analisam conjuntamente as relações entre as variáveis (óxidos e elementos traços) e as amostras e, depois, as relações entre as amostras e entre as variáveis separadamente. Os resultados obtidos da análise de correspondência permitiram agrupar as amostras de forma a obter um boa indicação do grau de diferenciação da série vulcânica.

David, Dagbert e Beauchemin (1977) fazem inicialmente uma abordagem teórica da análise de correspondência (utilizando também o termo análise fatorial não-paramétrica), a qual permite tirar vantagem da dualidade entre as análises R e Q-modais e de usar procedimentos de ponderação que reduzem os problemas de escala. Depois apresentam alguns exemplos de aplicação dessa técnica a dados geoquímicos e a dados sobre a espessura de diferentes unidades estratigráficas.

Zhou, Chang e Davis (1983) discutem os princípios gerais da análise RQ-modal, que não é exclusividade da análise de correspondência; fazem uma revisão das condições sob as quais os vários tipos de procedimentos são apropriados e uma crítica dos resultados obtidos da análise de correspondência feita por David et alii (1977), sugerindo considerar os componentes principais como procedimento da análise RQ-modal.

Davis (1986) apresenta uma abordagem dos métodos de análise de dados multivariados, incluindo a análise de correspondência, com aplicações a dados hipotéticos.

Bonham-Carter et alii (1986) utilizam a análise de correspondência para

determinar a distribuição dos foraminíferos do Cenozóico, na margem noroeste do Atlântico (Canadá). Para isso, eles constroem um zoneamento bioestratigráfico, usando um método denominado de RASC e a análise de correspondência para verificar se existe uma tendência faunística sistemática, e se essa tendência varia geograficamente com o tempo.

Pereira et alii (1990) apresentam um trabalho na linha de geoengenharia de reservatórios, combinando técnicas de estatística multivariada, no caso a análise de correspondência, geoestatística (*krigagem*) e modelos geológicos para fornecer uma melhoria nos processos de modelagem e na estratégia de desenvolvimento de um campo de petróleo, através do refinamento da descrição do reservatório. Através da análise de correspondência, foi possível selecionar um pequeno subconjunto de variáveis para definir grupos de poços que apresentavam características semelhantes. Esse subconjunto de variáveis reproduz o padrão global dos dados, no que diz respeito à distribuição da qualidade de óleo e das características geológicas e petrofísicas do reservatório.

Fabian-Goyheneche et alii (1991) propõem a utilização de métodos de classificação, da análise de variância e da análise de correspondência para caracterizar os parâmetros geológicos mais importantes, relacionados às propriedades petrofísicas da rocha, visando a definição de "fácies petrofísicas".

A análise de correspondência tem sido aplicada a muitas áreas do conhecimento científico, como técnica exploratória, podendo-se destacar dentre outras psicologia, biologia, saúde, educação, economia e pesquisa de mercado.

Em estatística, a análise de correspondência é abordada por Greenacre (1984), Greenacre e Hastie (1987), Lebart, Morineau e Warwick (1984) e Hill (1974). Em português, destaca-se a tese de mestrado de Cardoso (1991).

2 - *DECOMPOSIÇÃO DE MATRIZES EM VALORES SINGULARES*

2.1 INTRODUÇÃO

Segundo Davis (1986), existe uma grande variedade de procedimentos computacionais englobados pelo nome genérico de **análise fatorial**, cujo objetivo é tentar revelar estruturas existentes dentro de um conjunto de observações multivariadas, através de representações gráficas.

O ponto comum entre os métodos de análise fatorial é a **decomposição de matrizes em valores singulares**, geralmente referida como SVD (*Singular Value Decomposition*). A decomposição de uma matriz em seus valores e vetores singulares (SVD) é uma ferramenta matemática que foi desenvolvida por matemáticos franceses e italianos no final do século passado. Ela é utilizada para selecionar os subespaços que melhor se ajustam a uma nuvem de pontos pelo método dos mínimos quadrados. Eckart e Young (1936), foram os primeiros a aplicá-la, daí o fato de ela também ser conhecida como "decomposição de Eckart-Young". O termo tem ainda outros sinônimos: "estrutura básica", "decomposição singular", "forma canônica", e "redução de tensor".

Segundo Cardoso (1991), não são muitos os livros textos de estatística que abordam esta ferramenta. Literatura relevante sobre SVD em estatística deve-se a Good (1969), Chambers (1977), Gabriel (1978), Rao (1980), Mandel (1982), Searle (1982) e Greenacre e Underhill (1982); em matemática destaca-se Ben-Israel e Greville (1974).

Quando se utilizam gráficos estatísticos para representar as relações entre as variáveis que se deseja analisar, parte-se de sua representação no espaço das amostras, ou o contrário, no caso de analisarem-se as amostras. Estas variáveis são representadas por pontos no espaço Euclidiano, gerando uma nuvem cuja forma será

analisada. No entanto, quando se trabalha com matrizes de dimensões muito grandes, não é possível analisar estes gráficos multidimensionais, daí a necessidade de se reduzirem as suas dimensões. Isto é feito ajustando um subespaço (no caso bidimensional) a esta nuvem de pontos. A projeção dos pontos sobre este espaço bidimensional ótimo favorece a interpretação dos dados através das relações entre eles; se dois pontos estão próximos, diz-se que eles são correlacionados e, se estão muito afastados, que eles não têm correlação.

Ajustar subespaços de menor dimensão a uma nuvem de pontos significa rotacionar os eixos de modo que seja possível ver a nuvem de um ângulo diferente. Os novos eixos que definem o subespaço de menor dimensão indicam as direções de maior variância dos dados, e são perpendiculares entre si.

A definição de uma medida da distância ou **métrica** entre os pontos determina o tipo de análise empregada. Quando estas medidas são tomadas no espaço Euclidiano ordinário, têm-se os procedimentos de componentes principais, análise fatorial R-modal, Q-modal e RQ-modal. Quando as distâncias são medidas no espaço Euclidiano ponderado, tem-se a análise de correspondência.

Antes de abordar-se a decomposição de matrizes, define-se a notação utilizada neste trabalho. Simbolicamente, uma matriz de dimensão nxp é representada por $\mathbf{X}$ (letra maiúscula em negrito) e os seus elementos por x_{ij}, com i variando de 1 até n (número de linhas) e com j variando de 1 até p (número de colunas). Um conjunto de números reais $x_1 \ldots x_p$ é representado através do vetor $\mathbf{x}$ (letra minúscula em negrito), e convenciona-se escrevê-lo como um vetor coluna:

$$\boldsymbol{x} = \begin{vmatrix} x_1 \\ x_2 \\ \cdot \\ x_p \end{vmatrix} \qquad (1)$$

O índice p representa a ordem do vetor. Para escrever vetores como vetores linha, utiliza-se a notação $\mathbf{x}^T$ (transposta de $\mathbf{x}$): $\mathbf{x}^T = [x_1 \ldots x_p]$ ou $\mathbf{x} = [x_1 \ldots x_p]^T$. Embora a convenção seja escrever vetores como vetores coluna, utiliza-se a sua forma

transposta, (1), por uma questão de espaço. Os vetores podem ser representados graficamente de tal forma que cada vetor é desenhado como um ponto no espaço, em função de seus valores ou coordenadas ($x_1 ... x_p$).

2.2 DECOMPOSIÇÃO ESPECTRAL DE MATRIZES SIMÉTRICAS

Decompor uma matriz quadrada e simétrica em seus valores e vetores singulares significa expressá-la como o produto de três outras matrizes de forma e interpretação geométrica particularmente simples. As novas matrizes contêm as coordenadas dos eixos que definem o subespaço ótimo e as coordenadas das projeções da nuvem de pontos sobre este subespaço. A representação dos pontos no subespaço de menor dimensão fornece uma visualização das relações entre os pontos.

O caso mais simples de SVD é a chamada **decomposição espectral** de matrizes em seus autovalores e autovetores, ou ainda, *eigendecomposition* de uma matriz, a qual existe para qualquer matriz quadrada e simétrica **B** (nxn), de posto[1] $r \leq n$:

$$\boldsymbol{B}_{nxn} = \boldsymbol{V}_{nxr}\, \boldsymbol{D}_{\lambda rxr}\, \boldsymbol{V}^{T}_{rxn} = \sum_{r}^{R} \lambda_r v_r v_r^{T} \qquad (2)$$

com: $\mathbf{V}^T\mathbf{V} = \mathbf{I}$; isto significa que os vetores colunas de **V** são ortonormais (ortogonais e de comprimento unitário); $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_r > 0$ e $\mathbf{D}_\lambda$ é a matriz que contém os autovalores na sua diagonal principal e zeros nas outras posições.

Neste caso, os vetores singulares direitos e esquerdos, dados pelas colunas de **V** e pelas linhas de $\mathbf{V}^T$, são idênticos e são comumente conhecidos como autovetores de **B**, enquanto os valores singulares, λ, são chamados de autovalores (valor latente, valor característico, raiz latente caraterística).

As coordenadas dos novos eixos são dadas pelas colunas da matriz **V** ou pelas linhas de $\mathbf{V}^T$; as coordenadas das projeções dos pontos sobre estes novos eixos são dadas pelas colunas da matriz resultante do produto de **B** por **V**.

---

[1] posto de uma matriz é igual ao número de autovalores não-nulos.

A decomposição espectral de matrizes tem aplicação na análise de componentes principais (ACP) e na análise fatorial R-modal e Q-modal. Um exemplo de aplicação pode ilustrar melhor o que foi introduzido. Suponha-se que foram medidas três variáveis em quatro indivíduos de uma espécie determinada, resultando em uma matriz **A** de dimensão 4x3, conforme a tabela 2.1.

Tabela 2.1: Medidas feitas em quatro indivíduos (dados hipotéticos).

	V_1	V_2	V_3	Média das linhas
Indivíduo A	3	10	8	7
Indivíduo B	4	9	9	7,33
Indivíduo C	8	12	13	11
Indivíduo D	7	11	10	9,33
Média das colunas	5,5	10,5	10	8,66

Primeiramente, pode-se analisar as relações entre os indivíduos no espaço das variáveis, ou análise das linhas de **A**. Para simplificar os cálculos, pode-se subtrair de cada observação a sua média. Esta operação significa que a origem dos eixos foi transladada para o ponto médio da nuvem de pontos. A matriz resultante é:

$$\boldsymbol{X} = \begin{vmatrix} -2,5 & -0,5 & -2 \\ -1,5 & -1,5 & -1 \\ 2,5 & 1,5 & 3 \\ 1,5 & 0,5 & 0 \end{vmatrix} \tag{3}$$

Constrói-se uma matriz quadrada e simétrica $\mathbf{R} = \mathbf{X}^T\mathbf{X}$, ou matriz de variância-covariância das variáveis (espaço).

$$\boldsymbol{R} = \begin{vmatrix} 17 & 8 & 14 \\ 8 & 5 & 7 \\ 14 & 7 & 14 \end{vmatrix} \tag{4}$$

Em seguida, calcula-se a decomposição espectral de $\mathbf{R} = \mathbf{V}\mathbf{D}_\lambda\mathbf{V}^T$, através da subrotina

"*eigen*" do SAS/IML.

$$V = \begin{vmatrix} 0,698 & 0,559 & -0,447 \\ 0,349 & 0,279 & 0,894 \\ 0,625 & -0,780 & 0,000 \end{vmatrix} \qquad D_\lambda = \begin{vmatrix} 33,539 & 0 & 0 \\ 0 & 1,461 & 0 \\ 0 & 0 & 1 \end{vmatrix} \tag{5}$$

As colunas de **V** fornecem as coordenadas dos novos eixos da nuvem de pontos-linha (indivíduos), e são também os eixos de máxima variância dos dados. Define-se a variância total como a soma das variâncias individuais que estão localizadas na diagonal principal da matriz de variância-covariância **R**, o que é denominado de **traço** da matriz. Neste caso, a variância total é 17 + 5 + 14 = 36. Observa-se que a soma dos autovalores é igual ao traço da matriz: 33,54 + 1,46 + 1 = 36. Como os autovalores representam os comprimentos dos semi-eixos principais, os eixos também representam a variância dos dados, cada um deles considerando uma porção da variância, igual ao autovalor dividido pelo traço. Desta maneira, o primeiro eixo considera (33,54/36) x 100 = 93,16% da variância total; o segundo eixo considera (1,46/36) x 100 = 4,05% e o terceiro (1/36) x 100 = 2,8%.

Como se deseja uma representação bidimensional dos dados, despreza-se o último autovalor e o último autovetor de **X** (última coluna de **V**). Desta maneira está-se representando os dados com 93,16 + 4,05 = 97,21% da variância total, o que é uma representação quase exata dos mesmos.

As duas primeiras colunas da matriz **F**, formada pelo produto de **X** pelas duas primeiras colunas de **V**, fornecem as coordenadas da projeção dos pontos-linha sobre os dois novos eixos.

$$F = XV = \begin{vmatrix} -3,170 & 0,023 \\ -2,196 & -0,477 \\ 4,144 & -0,524 \\ 1,222 & 0,979 \end{vmatrix} \tag{6}$$

Pode-se plotar as coordenadas das duas primeiras colunas de **F** em um gráfico como o da figura 2.1. Observa-se que os indivíduos A e B são semelhantes entre si e diferentes de C e D em relação às variáveis medidas.

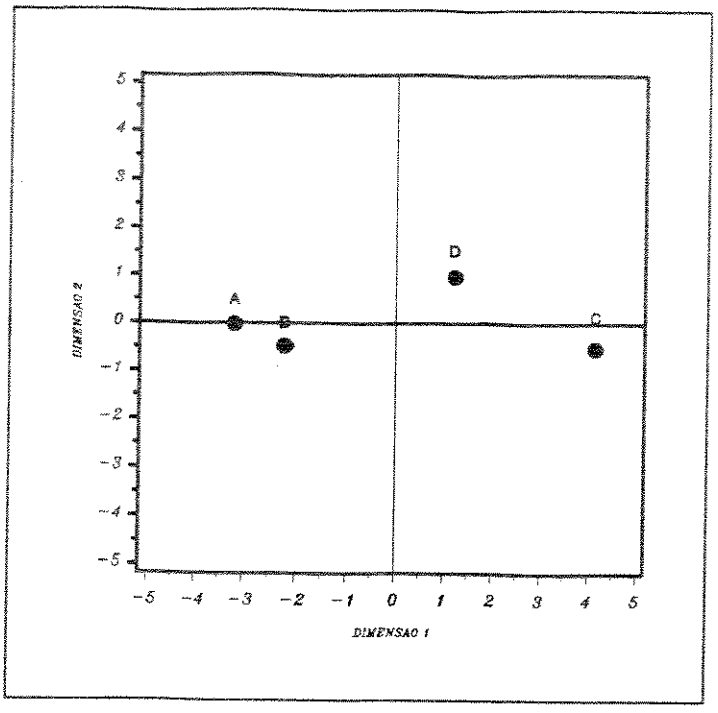

Figura 2.1: Representação aproximada da relação entre os indivíduos de A ou análise de componentes principais.

Desta forma, reduziu-se a dimensão original de **R**, que de 3 passou a 2. O primeiro eixo posiciona-se na direção de maior variabilidade dos dados, enquanto o segundo, perpendicular ao primeiro, posiciona-se na direção de segunda maior variabilidade. Estes novos eixos são chamados de componentes principais e esta análise é também conhecida como **análise de componentes principais** dos indivíduos de **A**.

Na análise de componentes principais quer-se somente encontrar a direção destes novos eixos (**V**) e medir suas magnitudes, λ (comprimento dos semi-eixos). Os autovetores têm comprimentos proporcionais à variação que eles representam. A análise de componentes principais consiste na transformação linear das p variáveis originais em p novas variáveis, onde cada nova variável é uma combinação linear das originais. Este processo é realizado de modo que cada nova variável considere, sucessivamente, a máxima variância possível dos dados. Quando as p novas variáveis forem calculadas toda a variância dos dados terá sido considerada.

No contexto da **análise fatorial**, o vetor formado pela multiplicação do autovetor pelo correspondente valor singular α (raiz quadrada do autovalor) é referido como **fator**, no caso as colunas de $\mathbf{VD}_\alpha$. Lembre-se que os autovetores têm comprimento unitário, ou seja, a soma dos seus elementos ao quadrado é igual a 1. Se eles são multiplicados pelo correspondente valor singular, eles são escalonados de modo que seus comprimentos sejam proporcionais à magnitude de seus valores singulares. Os elementos do fator são denominados de **ponderadores** e representam a proporção ou peso que se deve associar a cada variável para que se possa projetar os indivíduos sobre os fatores e obter seus escores.

Os escores dos indivíduos sobre os fatores são encontrados multiplicando-se $\mathbf{X}$ pelos fatores ponderados $\mathbf{VD}_\alpha$. A disposição dos pontos-linha ou indivíduos no diagrama bidimensional, neste caso, será a mesma da figura 2.1, com a diferença do fator $\mathbf{D}_\alpha$, e este procedimento é conhecido como **análise R-modal**.

Na análise fatorial, as relações dentro do conjunto das p variáveis representam as correlações de cada uma das variáveis com os r fatores mutuamente não-correlacionados; geralmente, assume-se que $r < p$. As variâncias das p variáveis são, portanto, derivadas das variâncias dos r fatores. A análise fatorial pode ser abordada em termos dos componentes principais. Neste contexto, extraem-se os autovalores e autovetores da matriz de variância-covariância ou de correlação. Isto assegura que todas as variáveis são igualmente ponderadas, permitindo converter os componentes principais em fatores. Se os autovetores são calculados na forma padronizada, eles são transformados de modo a definir vetores cujos comprimentos sejam proporcionais à variação que eles representam. Desta forma, a transformação dos autovetores padronizados ou unitários em fatores não afeta as direções dos vetores, somente os seus comprimentos.

Da mesma forma, pode-se também analisar as relações entre as variáveis no espaço dos indivíduos. Para simplificar os cálculos pode-se subtrair de cada observação a sua média, no caso a média das linhas, resultando na matriz $\mathbf{X}$:

$$X = \begin{vmatrix} -4 & 3 & 1 \\ -3{,}33 & 1{,}66 & 1{,}66 \\ -3 & 1 & 2 \\ -2{,}33 & 1{,}66 & 0{,}66 \end{vmatrix} \quad (7)$$

Em seguida, constrói-se uma matriz quadrada e simétrica $\mathbf{Q} = \mathbf{X}\mathbf{X}^T$, ou matriz de variância-covariância entre os indivíduos.

$$Q = \begin{vmatrix} 26 & 20 & 17 & 15 \\ 20 & 16{,}67 & 15 & 11{,}67 \\ 17 & 15 & 14 & 10 \\ 15 & 11{,}67 & 10 & 8{,}67 \end{vmatrix} \quad (8)$$

Calcula-se a decomposição espectral de $\mathbf{Q} = \mathbf{U}\mathbf{D}_\lambda\mathbf{U}^T$. As colunas de **U** fornecem as coordenadas dos eixos da nuvem de pontos-coluna (variáveis). As colunas da matriz **G**, formada pelo produto de $\mathbf{X}^T$ por **U**, fornecem as coordenadas das projeções dos pontos-coluna sobre os novos eixos. Estes eixos são chamados de componentes principais e esta análise é também conhecida como **análise de componentes principais** das colunas de **A**.

$$U = \begin{vmatrix} 0{,}633 & -0{,}584 & -0{,}504 & 0{,}731 \\ 0{,}511 & 0{,}282 & 0{,}202 & -0{,}786 \\ 0{,}451 & 0{,}714 & -0{,}189 & 0{,}501 \\ 0{,}367 & -0{,}264 & 0{,}818 & 0{,}354 \end{vmatrix} \qquad D_\lambda = \begin{vmatrix} 63{,}00 & 0 & 0 & 0 \\ 0 & 2{,}33 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{vmatrix} \quad (9)$$

$$G = X^T U = \begin{vmatrix} -6{,}45 & -0{,}13 & 0 & 0 \\ 3{,}81 & -1{,}01 & 0 & 0 \\ 2{,}63 & 1{,}14 & 0 & 0 \end{vmatrix} \quad (10)$$

Neste caso, como os dois últimos autovalores são iguais a zero, a dimensão da matriz original foi reduzida e a representação bidimensional é exata, considerando 100% da variância dos dados. No caso de se ponderar os eixos pelos seus respectivos valores singulares, realiza-se a **análise Q-modal**, ou análise das colunas de **A**.

A figura 2.2 mostra a representação gráfica bidimensional das relações

entre as variáveis V_1, V_2 e V_3. As variáveis V_2 e V_3 opõem-se a V_1, ou seja V_2 e V_3 são mais parecidas em relação aos indivíduos.

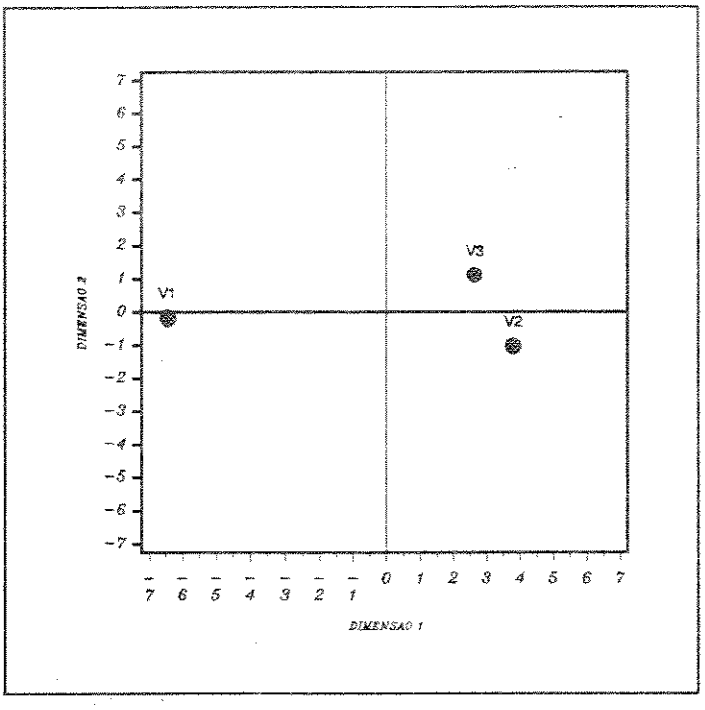

Figura 2.2: Representação bidimensional da relação entre as variáveis ou análise de componentes principais das colunas de A.

Resumindo, a decomposição espectral de matrizes simétricas é a própria análise de componentes principais das linhas ou colunas de **A**. As análises fatoriais R e Q-modal, também se utilizam desta decomposição, escalonando os novos eixos encontrados pelo fator $\mathbf{D}_\alpha = \mathbf{D}_\lambda^{1/2}$.

Apesar de ser possível construir matrizes quadradas e simétricas a partir da matriz que está sendo analisada, pós ou pré-multiplicando sua transposta por ela mesma, este cálculo é computacionalmente caro, a depender do tamanho da matriz de dados, e os resultados obtidos perdem precisão. Assim, quando se trabalha com matrizes retangulares, utiliza-se a decomposição ordinária (subrotina "*svd*" no SAS/IML), descrita a seguir.

2.3 DECOMPOSIÇÃO ORDINÁRIA DE MATRIZES RETANGULARES

O teorema fundamental da decomposição de matrizes diz que qualquer matriz real **X** (nxp), retangular, de posto r, pode ser expressa da seguinte forma:

$$\boldsymbol{X}_{nxp} = \boldsymbol{U}_{nxr}\, \boldsymbol{D}_{\alpha rxr}\, \boldsymbol{V}^{T}_{rxp} \tag{11}$$

onde: $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$ e $\alpha_1 \geq \ldots \alpha_r > 0$.

Outra forma de escrever a equação acima é:

$$\boldsymbol{X} = \sum_{r=1}^{R} \alpha_r \boldsymbol{u}_r \boldsymbol{v}_r^{T} \tag{12}$$

onde: $\mathbf{u}_1 \ldots \mathbf{u}_r$ e $\mathbf{v}_1 \ldots \mathbf{v}_r$ são as colunas de **U** e **V**. Os valores $\alpha_1 \ldots \alpha_r$, com r=1...R, são chamados de valores singulares de **X**, enquanto que os R vetores ortonormais $\mathbf{u}_r$ e $\mathbf{v}_r$, com r=1...R, são chamados de vetores singulares esquerdos e direitos, respectivamente. Os vetores singulares esquerdos $\mathbf{u}_r$, formam uma base ortonormal para as colunas de **X** no espaço n-dimensional e são os autovetores de $\mathbf{X}\mathbf{X}^T$ associados aos autovalores $\alpha_1^2 \ldots \alpha_r^2$. Os R vetores singulares direitos $\mathbf{v}_r$, formam uma base ortonormal para as linhas de **X** no espaço p-dimensional e são os autovetores de $\mathbf{X}^T\mathbf{X}$, com os mesmos autovalores. Os elementos $\alpha_1 \ldots \alpha_r$ da diagonal da matriz $\mathbf{D}_\alpha$ são os valores singulares de **X**.

A SVD na forma da equação 12, pode ser interpretada como a soma das matrizes $\mathbf{u}_r\mathbf{v}^T_r$, padronizadas de posto 1, com r = 1...R, e com os valores singulares indicando a magnitude de cada uma das R-dimensões da matriz.

Note-se que a SVD ordinária consiste em matrizes reais e existe para qualquer matriz retangular, enquanto que a decomposição espectral de uma matriz quadrada geralmente envolve elementos complexos, se a matriz for não-simétrica.

A existência da SVD de **X** pode ser provada a partir da existência de uma matriz **B**, quadrada e simétrica obtida da multiplicação da transposta de **X** por ela mesma:

$$\begin{aligned} B &= X^T X = V D_\lambda V^T \\ &= (U D_\alpha V^T)^T U D_\alpha V^T \\ &= V D_\alpha^2 V^T = V D_\lambda V^T \end{aligned} \tag{13}$$

Segundo Davis (1986), sob certas circunstâncias, a relação recíproca entre **R** e **Q** é satisfeita, mas isto depende da natureza da transformação ou escalonamento feito na matriz original. No exemplo mostrado, trabalhou-se com as matrizes de variância-covariância das linhas e das colunas de **A**, separadamente. Isto implica que os espaços dos indivíduos e das variáveis têm escalas diferentes (os autovalores são diferentes).

As diferenças de escala distorcem as soluções no modo R e Q, não sendo possível plotar as duas soluções no mesmo gráfico. Existem várias maneiras de evitar esta distorção. Uma delas é omitir o escalonamento, o que faz com que os autovetores e autovalores de $\mathbf{X}^T\mathbf{X}$ sejam iguais aos de $\mathbf{X}\mathbf{X}^T$. Infelizmente, a desvantagem deste procedimento é que a análise Q-modal fica sensível à escolha das unidades de medidas e os resultados podem refletir as magnitudes médias das variáveis ao invés das suas variâncias e covariâncias, não sendo utilizado na prática.

Outra alternativa é buscar uma forma de escalonamento das linhas que produza uma medida significativa de suas relações e que, ao mesmo tempo, resulte em uma medida significativa das relações entre as colunas, o que é conhecido como **análise RQ-modal** ou **biplot**. Ou ainda, escalonar igualmente as linhas e as colunas de uma tabela de contingência, como faz a **análise de correspondência**.

Suponha-se, por exemplo, que se faça um escalonamento de **A** da seguinte forma: $\mathbf{X} = a_{ij} - a_{i+} - a_{+j} + a_{++}$, o que equivale a subtrair de cada observação a média das linhas e a média das colunas e somar a média total.

$$X = \begin{vmatrix} -0{,}84 & 1{,}16 & 0{,}34 \\ -0{,}17 & -0{,}17 & 0{,}33 \\ 0{,}16 & -0{,}84 & 0{,}66 \\ 0{,}83 & 1{,}17 & 0{,}67 \end{vmatrix} \tag{14}$$

Fazendo-se a decomposição espectral da matriz $\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}_\lambda\mathbf{V}^T$ obtêm-se os seguintes autovetores e autovalores:

$$V = \begin{vmatrix} -0,110 & 0,765 & -0,634 \\ 0,984 & 0,169 & 0,033 \\ -0,133 & 0,621 & 0,773 \end{vmatrix} \qquad D_\lambda = \begin{vmatrix} 3,491 & 0 & 0 \\ 0 & 2,148 & 0 \\ 0 & 0 & 0,368 \end{vmatrix} \tag{15}$$

Fazendo-se a decomposição espectral da matriz $\mathbf{XX}^T = \mathbf{UD}_\lambda\mathbf{U}^T$ obtêm-se os seguintes autovetores e autovalores:

$$U = \begin{vmatrix} 0,685 & -0,449 & 0,508 & 0,267 \\ -0,103 & 0,031 & 0,588 & -0,801 \\ -0,499 & 0,266 & 0,627 & 0,535 \\ 0,520 & 0,853 & 0,500 & 0,003 \end{vmatrix} \qquad D_\lambda = \begin{vmatrix} 3,491 & 0 & 0 & 0 \\ 0 & 2,148 & 0 & 0 \\ 0 & 0 & 0,368 & 0 \end{vmatrix} \tag{16}$$

Fazendo-se a decomposição ordinária da matriz $\mathbf{X} = \mathbf{UD}_\lambda\mathbf{V}^T$ obtêm-se os seguintes vetores e valores singulares:

$$U = \begin{vmatrix} 0,685 & -0,449 & 0,508 \\ -0,103 & 0,031 & 0,588 \\ -0,499 & 0,266 & 0,627 \\ 0,520 & 0,853 & 0,500 \end{vmatrix} \qquad D_\alpha = \begin{vmatrix} 1,868 & 0 & 0 \\ 0 & 1,465 & 0 \\ 0 & 0 & 0,607 \end{vmatrix} \tag{17}$$

$$V = \begin{vmatrix} -0,110 & 0,765 & -0,634 \\ 0,984 & 0,169 & 0,033 \\ -0,133 & 0,621 & 0,773 \end{vmatrix} \tag{18}$$

Neste caso, a decomposição espectral de $\mathbf{X}^T\mathbf{X}$ e de $\mathbf{XX}^T$ tem os mesmos autovalores e os mesmos autovetores de $\mathbf{X}$, ou seja, a decomposição espectral está contida na decomposição ordinária de $\mathbf{X}$. Há muito que se sabia da existência da relação entre os autovetores de $\mathbf{R}$ e os autovetores de $\mathbf{Q}$, já que ambas são originadas da mesma matriz original. Este fato permite a caracterização de grupos de amostras em relação às variáveis e reduz consideravelmente o tempo de computação uma vez que, ao invés de se trabalhar com uma matriz nxn ($\mathbf{XX}^T$), tudo é feito com a matriz pxp ($\mathbf{X}^T\mathbf{X}$), que na maioria das vezes é muito menor. Por exemplo, na maioria dos problemas é comum se ter mais amostras do que variáveis: 1000 amostras e 20 variáveis. A matriz $\mathbf{X}^T\mathbf{X}$ = 20 x 20 = 400 enquanto que $\mathbf{XX}^T$ = 1000 x 1000 = 1000000.

Pode-se provar que as duas matrizes têm os mesmos autovalores. Seja

$\mathbf{v}_k$ o k-ésimo autovetor de $\mathbf{X}\mathbf{X}^T$ e $\mathbf{u}_k$ o k-ésimo autovetor de $\mathbf{X}^T\mathbf{X}$ e λ_k o k-ésimo autovalor de $\mathbf{X}\mathbf{X}^T$. Por definição:

$$XX^TV_k = \lambda_k V_k \tag{19}$$

Se r é o posto da matriz $\mathbf{X}\mathbf{X}^T$, então existe um máximo de r autovalores não-nulos. Multiplicando-se a equação 19 por $\mathbf{X}^T$ tem-se:

$$X^TX(X^TV_k) = \lambda_k(X^TV_k) \tag{20}$$

Isto significa que para cada k ($k \leq r$), $\mathbf{X}^T\mathbf{V}_k$ é um autovetor de $\mathbf{X}^T\mathbf{X}$ e λ_k é seu autovalor associado. Além disto, os autovalores não-nulos de $\mathbf{X}^T\mathbf{X}$ e $\mathbf{X}\mathbf{X}^T$ são os mesmos. Lembrando-se que estes autovalores divididos pela sua soma são interpretados como a porcentagem da variância explicada em cada eixo, observa-se que no primeiro eixo de $\mathbb{R}^p$ tem-se a mesma variância do primeiro eixo de $\mathbb{R}^n$, e assim para o segundo, terceiro e para o r-ésimo eixo.

Como o vetor $\mathbf{V}_k$ é unitário, não se pode ter ao mesmo tempo $\mathbf{U}_k = \mathbf{X}^T\mathbf{V}_k$ também unitário. Como os vetores são definidos somente dentro de uma constante multiplicativa pode-se obter dois conjuntos de vetores unitários usando-se a equação 19. Em um espaço $\mathbf{U}_k = (1/\sqrt{\lambda})\mathbf{X}^T\mathbf{V}_k$ e no outro $\mathbf{V}_k = (1/\sqrt{\lambda})\mathbf{X}\mathbf{U}_k$ com k variando de 1 a r.

As propriedades duais das duas análises devem ser sempre lembradas e não há razão para realizá-las separadamente, pois assim fazendo, perde-se uma quantidade importante de informação visual, uma das grandes vantagens da análise fatorial (David et alii, 1974).

Neste caso, as matrizes **F** e **G**, definidas em (6) e (10), são também as mesmas na decomposição espectral e na decomposição ordinária. Como a transformação é o mesmo para as linhas e colunas, pode-se representá-los no mesmo espaço. As figuras 2.3 e 2.4 mostram a disposição dos pontos-linha e dos pontos-coluna, em seus respectivos espaços, e a figura 2.5 mostra a sobreposição dos dois gráficos, permitindo interpretar simultaneamente as relações entre as linhas e as

colunas da tabela **A**.

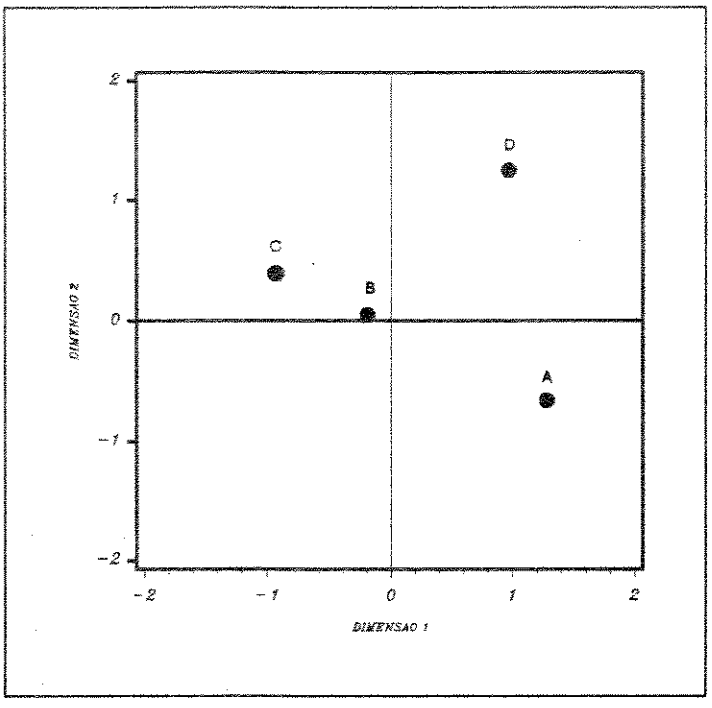

Figura 2.3: Representação bidimensional das linhas de **X**.

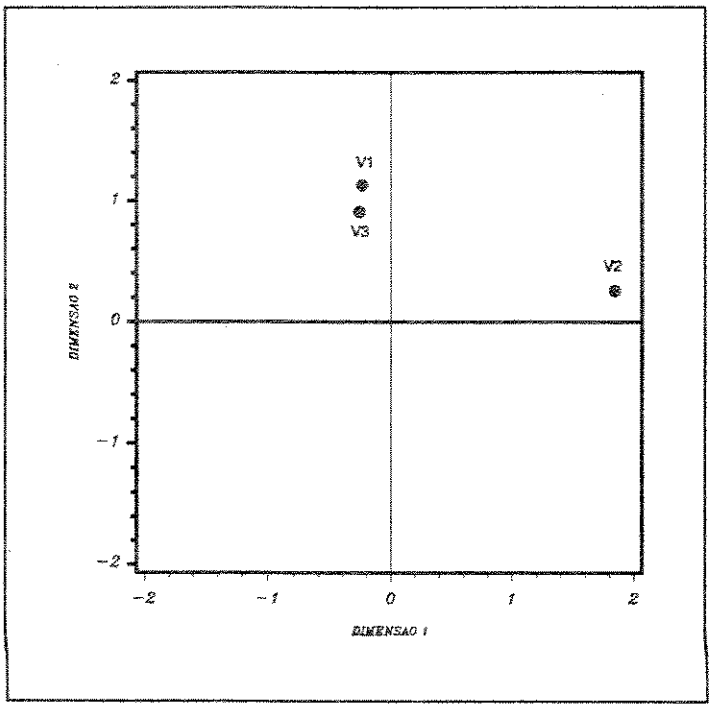

Figura 2.4: Representação bidimensional das colunas de **X**.

Observa-se que, na figura 2.5, a disposição dos pontos-linha e dos pontos-coluna difere da disposição apresentada nas figuras 2.1 e 2.2, pois o escalonamento foi diferente. Neste tipo de análise, os indivíduos C e B estão relacionados e diferenciam-se dos indivíduos A e D. As variáveis V_1 e V_3 diferem de V_2. A análise conjunta dos dois gráficos mostra uma associação dos indivíduos B e C com as variáveis V_1 e V_3, enquanto os indivíduos A e D associam-se moderadamente com a variável V_2.

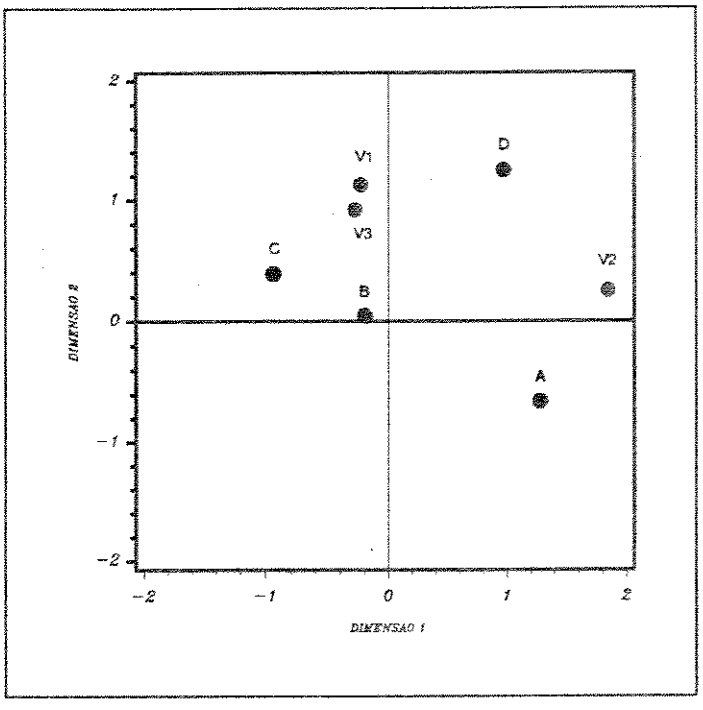

Figura 2.5: Biplot ou análise RQ-modal das linhas e colunas de X.

Conclui-se, que as várias versões da análise fatorial diferem basicamente na forma pela qual os dados são escalonados antes da decomposição da matriz. O escalonamento determina a medida de similaridade e, conseqüentemente, a natureza da solução fatorial (Zhou et alii, 1983).

O fluxograma da figura 2.6 mostra as relações entre as decomposições espectral e ordinária nas análises ACP, R-modal, Q-modal e RQ-modal.

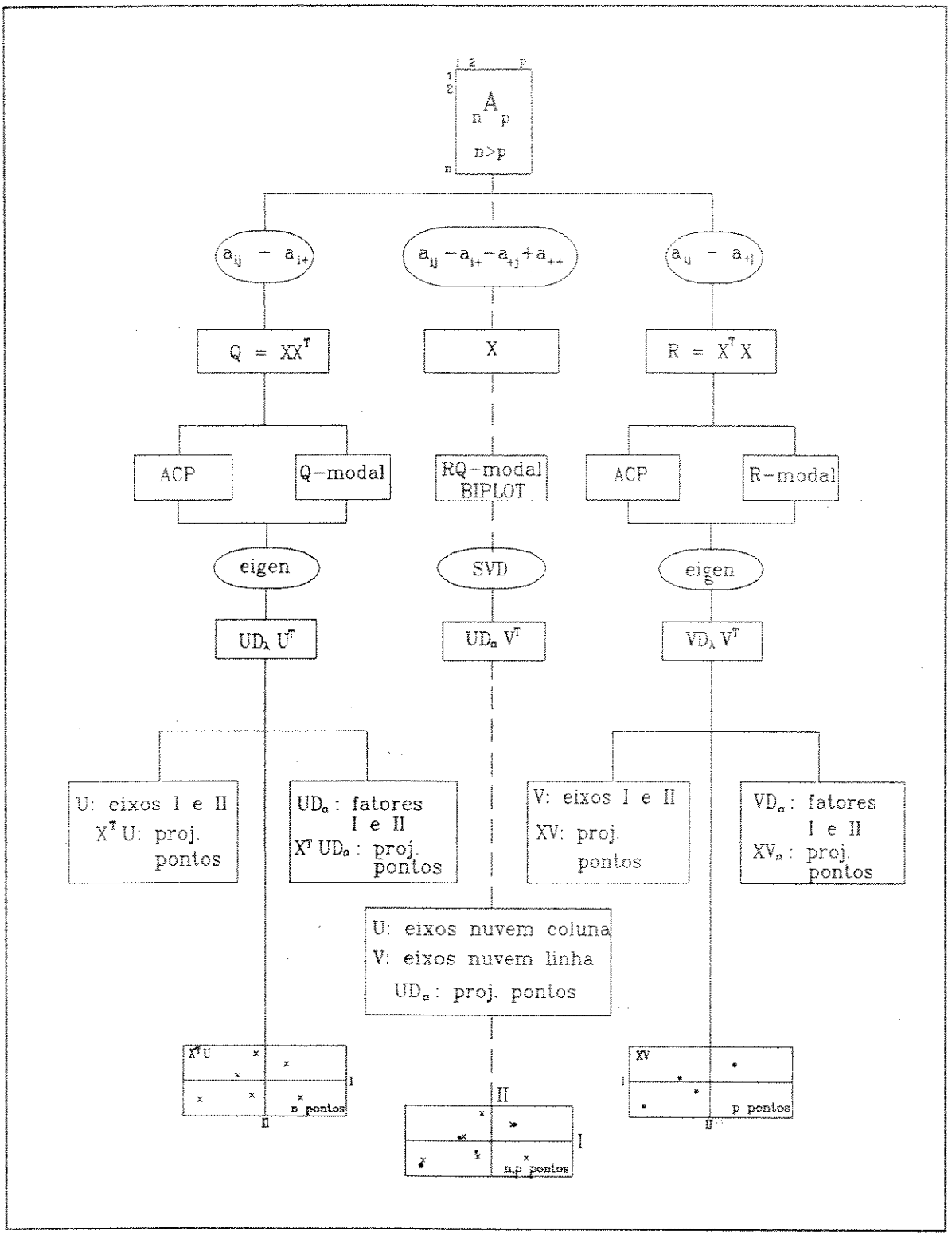

Figura 2.6: Fluxograma da relação entre R, Q e RQ-modal.

2.4 APROXIMAÇÃO DE MATRIZES POR MATRIZES DE POSTO MENOR

Uma das grandes utilidades da SVD é que ela permite desprezar os últimos termos da equação 12, correspondentes aos menores valores singulares, como foi feito no primeiro exemplo. Isto resulta em uma aproximação da matriz **X** pela matriz $\mathbf{X}_{R^*}$:

$$X_{R^*} = \sum_r^{R^*} \alpha_r u_r v_r^T \tag{21}$$

de posto $R^* < r$, que minimiza a distância entre os pontos de **X** e os pontos de **A**, no sentido dos mínimos quadrados:

$$\|X - A\|^2 = \sum_i \sum_j (x_{ij} - a_{ij})^2 \tag{22}$$

para todas as matrizes **A** de posto R^* (ou menor). $\mathbf{X}_{R^*}$ é chamada de matriz de aproximação de **X**, de posto R^*, no sentido dos mínimos quadrados, e pode ser escrita na forma decomposta como:

$$X_{R^*} = U_{R^*} D_{\alpha R^*} V_{R^*}^T \tag{23}$$

onde $\mathbf{U}_{R^*}$, $\mathbf{V}_{R^*}$ e $\mathbf{D}_{\alpha R^*}$ são as submatrizes relevantes de **U**, **V** e $\mathbf{D}_\alpha$.

Da equação 23, as linhas e as colunas de $\mathbf{X}_{R^*}$ são equivalentemente os pontos nos respectivos subespaços de dimensão R^* que melhor se ajustam às linhas e colunas de **X**, no sentido da menor soma do quadrado das distâncias Euclidianas. Isto efetivamente soluciona o problema de minimizar distâncias entre pontos e subespaços, na ausência de massa para os pontos e peso para os eixos, que é também conhecido como análise de componentes principais ordinária.

2.5 DECOMPOSIÇÃO GENERALIZADA DE UMA MATRIZ RETANGULAR

Até agora foram analisados casos em que se trabalhava no espaço Euclidiano ordinário. No entanto, muitas vezes, a medida de distância neste espaço não é adequada, principalmente lidando-se com vetores de freqüência, como é o caso da análise de correspondência. Isto porque a distância entre 0 e 5% é intuitivamente muito maior do que a distância entre 50 e 55%. Neste caso, é mais adequado utilizar uma distância ponderada pela freqüência.

Voltando-se para o problema de selecionar o subespaço que mais se aproxima da nuvem de pontos no espaço Euclidiano ponderado, introduz-se uma generalização da definição de SVD. Se $\Omega_{(nxn)}$ e $\Phi_{(pxp)}$ são duas matrizes simétricas positivas definidas[2], então qualquer matriz real $\mathbf{X}_{(nxp)}$ de posto R pode ser expressa da seguinte forma:

$$X_{nxp} = N_{nxr}\, D_{\alpha rxr}\, M^{T}_{rxp} = \sum_{r}^{R} \alpha_r n_r m_r^{T} \qquad (24)$$

onde as colunas de **N** e **M** são ortonormalizadas com relação a Ω e Φ respectivamente:

$$N^{T}\Omega N = M^{T}\Phi M = I \qquad (25)$$

As colunas de **N** e **M** podem ser chamadas de vetores singulares generalizados, esquerdo e direito respectivamente. Eles são ainda as bases ortonormais para as colunas e linhas de **X**, onde as métricas impostas aos espaços n e p-dimensional não são mais as Euclidianas ordinárias, mas as métricas generalizadas ou ponderadas, definidas por Ω e Φ, respectivamente. Da mesma forma, os elementos da diagonal da matriz diagonal, $\mathbf{D}_{\alpha}$, podem ser chamados de valores singulares generalizados, ordenados do maior para o menor.

A SVD generalizada pode ser facilmente provada, a partir da SVD ordinária de **B**:

---

[2] uma matriz positiva definida é aquela que tem todos seus autovalores positivos.

$$\begin{aligned} \boldsymbol{B} &= \boldsymbol{\Omega}^{1/2}\boldsymbol{X}\boldsymbol{\Phi}^{1/2} \\ &= \boldsymbol{\Omega}^{1/2}\boldsymbol{U}\boldsymbol{D}_{\alpha}\boldsymbol{\Phi}^{1/2}\boldsymbol{V}^{T} \end{aligned} \qquad (26)$$

Fazendo $\mathbf{N} = \Omega^{1/2}\mathbf{U}$ e $\mathbf{M} = \Phi^{1/2}\mathbf{V}$, têm-se as equações 24 e 25. Deve ser lembrado que, se Ω tem decomposição espectral $\Omega = \mathbf{W}\mathbf{D}_{\mu}\mathbf{W}^{T}$, então $\Omega^{1/2} = \mathbf{W}\mathbf{D}_{\mu}{}^{1/2}\mathbf{W}^{T}$).

Em resumo, a SVD é a base da análise fatorial, da qual a análise de correspondência é um tipo de procedimento. A tabela 2.2 sintetiza o que foi discutido. O capítulo seguinte apresenta o desenvolvimento do algoritmo utilizado na análise de correspondência.

Tabela 2.2: Comparação entre a decomposição ordinária, espectral e generalizada. (De Greenacre, 1984).

SVD ORDINÁRIA (definida para geometria Euclidiana ordinária) $\mathbf{X}_{(n\times p)} = \mathbf{U}_{(n\times r)}\,\mathbf{D}_{\alpha(r\times r)}\,\mathbf{V}^{T}_{(r\times p)}$ $\mathbf{X} = \sum \alpha\, \mathbf{u}_r\, \mathbf{v}_r^{T}$ onde: $\mathbf{U}^{T}\mathbf{U} = \mathbf{V}^{T}\mathbf{V} = \mathbf{I}$ (isto implica que as colunas de U e V são ortonormais) r = posto de $\mathbf{X}$ $(r \le \text{mín}\,\{n,p\})$
ESPECTRAL (caso particular da SVD, para matrizes quadradas) $\mathbf{B}_{(n\times n)} = \mathbf{X}^{T}\mathbf{X} = \mathbf{V}_{(n\times r)}\,\mathbf{D}_{\lambda(r\times r)}\,\mathbf{V}^{T}_{(r\times n)}$ $\mathbf{B} = \sum \lambda\, \mathbf{v}_r\, \mathbf{v}_r^{T}$ com $\lambda = \alpha^{2}$
SVD GENERALIZADA (definida para geometria Euclidiana ponderada) $\mathbf{X}_{(n\times p)} = \mathbf{N}_{(n\times r)}\,\mathbf{D}_{\alpha(r\times r)}\,\mathbf{M}^{T}_{(r\times p)}$ onde $\mathbf{N}^{T}\Omega\mathbf{N} = \mathbf{M}^{T}\Phi\mathbf{M} = \mathbf{I}$

3 - TEORIA DA ANÁLISE DE CORRESPONDÊNCIA

3.1 DESENVOLVIMENTO TEÓRICO

A abordagem teórica da análise de correspondência apresentada a seguir é a abordagem geométrica de Greenacre (1984). Para desenvolver a teoria da análise de correspondência simples, utilizou-se um subconjunto de dados da Formação Lagoa Feia. A matriz de dados em questão refere-se à descrição macroscópica de testemunhos, na qual foram medidas diversas variáveis para cada amostra: tamanho das conchas, espessura das conchas, quantidade de matriz terrígena, quantidade de matriz carbonática, dentre outras, como será visto com mais detalhe no capítulo seguinte.

Suponha-se, por exemplo, que se deseja estudar as relações existentes entre duas dessas variáveis: tamanho de conchas e quantidade de matriz terrígena. Para isso foi construída a tabela de contingência, denotada por **N** (tabela 3.1). Essa tabela foi obtida através da transformação da matriz original de dados contínuos em uma matriz de dados discretizados segundo critérios geológicos (Programa P.1). A variável tamanho de conchas é categorizada em intervalos de tamanho, do maior para o menor: AA, BB, CC, DD, EE, FF, GG e SCO. As categorias da variável quantidade de matriz terrígena são: AREN (terrígenos), ARSO (carbonatos arenosos), MUIA (carbonatos muito arenosos) e PURO (carbonatos puros).

Na tabela 3.1, **N** é uma matriz 8 x 4, de números não-negativos, onde se denota cada elemento de **N** por n_{ij}, o total das linhas de **N** por n_{i+} (i= 1,...,I), o total das colunas de **N** por n_{j+} (j= 1,...,J) e o total geral simplesmente por n. Nela observa-se que 2 amostras caem simultaneamente nas categorias AREN e AA, 1 amostra nas categorias ARSO e AA, e assim por diante, até que todas as 1280 amostras tenham sido classificadas dessa forma.

Tabela 3.1: Tabela de contingência, N, das amostras da Formação Lagoa Feia.

Tamanho de conchas	Quantidade de matriz terrígena				Totais (n_{i+})
	AREN	ARSO	MUIA	PURO	
AA	2	1	5	0	8
BB	0	4	1	1	6
CC	0	9	0	1	10
DD	14	20	27	15	76
EE	16	40	30	22	108
FF	37	101	43	84	265
GG	10	92	42	85	229
SCO	426	6	42	104	578
Totais (n_{+j})	505	273	190	312	1280

Seja **P** a matriz de correspondência obtida pela divisão de cada elemento de **N** pelo total geral dos elementos, n. Na tabela 3.2, a soma das linhas é denotada por $\mathbf{r}_i = n_{i+}/n$ e a soma das colunas por $\mathbf{c}_i = n_{+j}/n$. $\mathbf{D}_r$ é a matriz diagonal formada pelos elementos de **r**, na sua diagonal principal e $\mathbf{D}_c$ é a matriz cuja diagonal principal é formada pelos elementos de **c**. Em notação vetorial, $\mathbf{P} = (1/n)\mathbf{N}$, onde:

$$\mathbf{N} = [n_{ij}], \text{ com } n_{ij} \geq 0$$
$$n = \mathbf{1}^T \mathbf{N}\, \mathbf{1} \text{ com } \mathbf{1} = [1...1]^T$$
$$\mathbf{r} = \mathbf{P1} \text{ e } \mathbf{c} = \mathbf{P}^T\mathbf{1}$$
$$\mathbf{D}_r = \text{diag}(\mathbf{r}) \text{ e } \mathbf{D}_c = \text{diag}(\mathbf{c})$$

Cabe observar que **P** nada mais é do que a matriz de freqüências relativas; os elementos de **r** são as probabilidades marginais das linhas e, mais adiante, serão usados como massa das linhas e centróide das colunas; os elementos de **c** são as probabilidades marginais das colunas e serão usados como massa das

colunas e centróide das linhas.

Tabela 3.2 - Matriz de correspondência P.

	c_1	c_2	c_3	c_4	$r_i=n_{i.}/n$
r_1	0,00156	0,00078	0,00390	0,00000	0,00625
r_2	0,00000	0,00312	0,00078	0,00078	0,00468
r_3	0,00000	0,00703	0,00000	0,00078	0,00781
r_4	0,01093	0,01562	0,02109	0,01172	0,05937
r_5	0,01250	0,03125	0,02343	0,01719	0,08437
r_6	0,02890	0,07890	0,03359	0,06563	0,20703
r_7	0,00781	0,07187	0,03281	0,06641	0,17890
r_8	0,33280	0,00469	0,03281	0,08125	0,45156
$c_j= n_{+j}/n$	0,39453	0,21328	0,14843	0,24375	1

$$D_r = \begin{vmatrix} 0,00625 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0,00468 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0,00781 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0,05937 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0,08437 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0,20703 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0,17890 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0,45156 \end{vmatrix}$$

$$D_c = \begin{vmatrix} 0,39453 & 0 & 0 & 0 \\ 0 & 0,21328 & 0 & 0 \\ 0 & 0 & 0,14843 & 0 \\ 0 & 0 & 0 & 0,24375 \end{vmatrix}$$

Massa das linhas $(r)^T$ = | 0,00625 0,00468 0,00781 0,05937 0,08437 0,20703 0,17890 0,45156 |

Massa das colunas (c) = | 0,39453 0,21328 0,14843 0,24375 $|^T$

Notacionalmente é mais fácil trabalhar com **P** do que com **N**, uma vez que a análise de correspondência lida com as freqüências relativas dos dados e, assim, torna-se invariante em relação ao total de observações.

3.1.1 Construção das Nuvens de Pontos

Para comparar mais facilmente as oito categorias da variável tamanho de conchas, construiu-se a tabela 3.3, denotada por **R**, a matriz dos perfis das linhas. Essa tabela de freqüências relativas das linhas é obtida dividindo-se cada elemento de **N** pelo total de sua respectiva linha; estes valores podem ser expressos como porcentagem, se forem multiplicados por 100. Para facilitar mais a interpretação, pode-se desenhar histogramas para cada uma dessas linhas, conforme a figura 3.1. Nela, observa-se que os perfis das categorias FF e GG são os mais semelhantes com relação às categorias da variável quantidade de matriz terrígena.

*Tabela 3.3: Matriz dos Perfis-linha, **R**, obtida pela divisão de cada elemento de **N**, pelo total de sua respetiva linha.*

Matriz dos Perfis-linha, **R**.					
a_1	0,250000	0,125000	0,625000	0,000000	1
a_2	0,000000	0,666667	0,166667	0,166667	1
a_3	0,000000	0,900000	0,000000	0,100000	1
a_4	0,184211	0,263158	0,355263	0,197368	1
a_5	0,148148	0,370370	0,277778	0,203704	1
a_6	0,139623	0,381132	0,162264	0,316981	1
a_7	0,043668	0,401747	0,183406	0,371179	1
a_8	0,737024	0,010381	0,072664	0,179931	1
	1,502674	3,118455	1,843042	1,53583	1

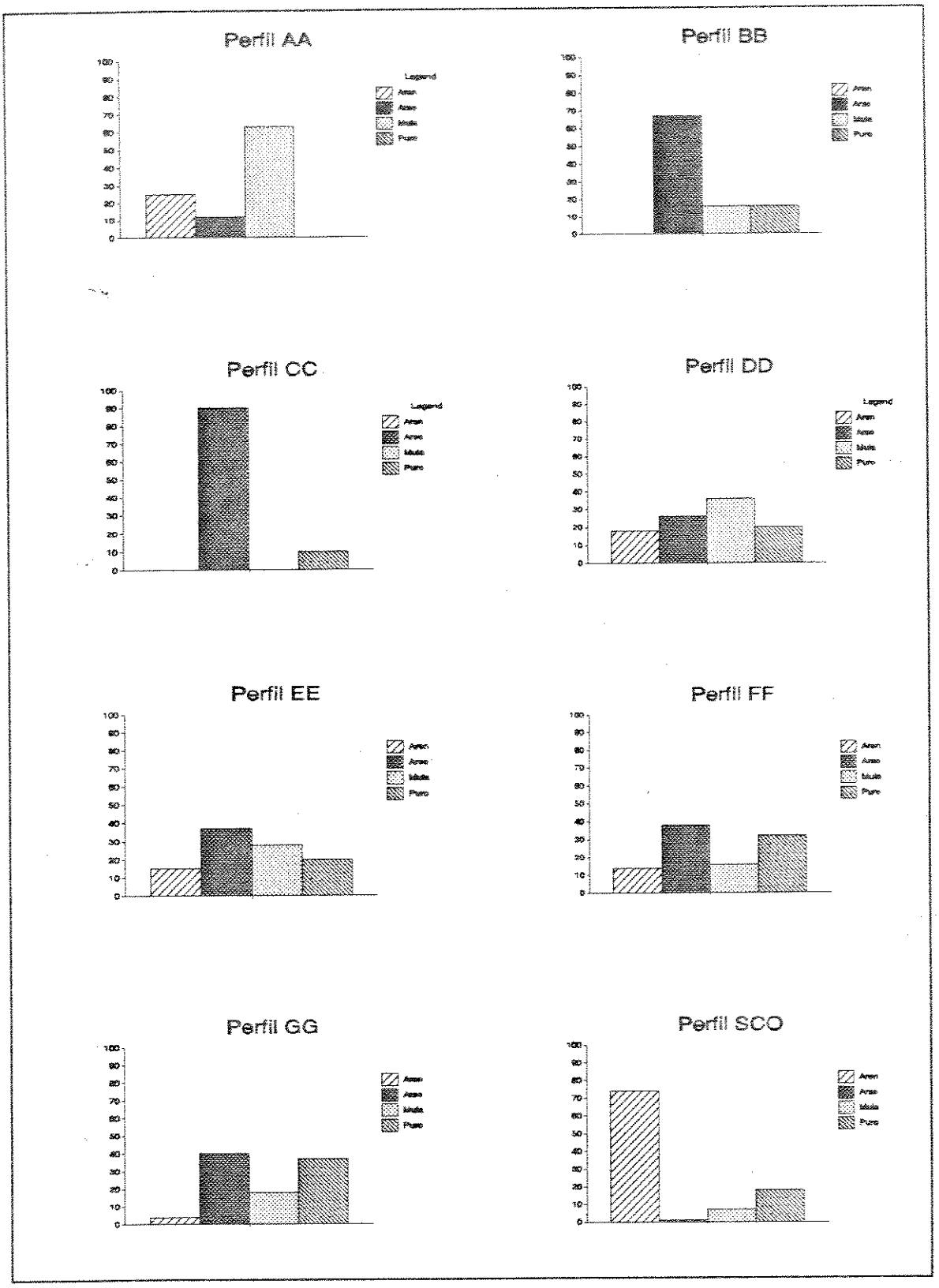

Figura 3.1: Histogramas das categorias da variável tamanho de conchas em relação à variável quantidade de matriz terrígena.

Na análise de correspondência, cada linha de **R** é denominada de **perfil distância** e, representa um conjunto de freqüências relativas cuja soma é igual a 1. Esse conjunto de freqüências relativas, na terminologia estatística convencional, é conhecido como densidade de probabilidade amostral. No entanto, na análise de correspondência, utiliza-se o termo **perfil distância** porque em alguns contextos a interpretação probabilística não é aplicável.

A compreensão das relações entre essas duas variáveis pode ser facilitada através de uma interpretação geométrica. Cada coluna da tabela 3.3 define as coordenadas de cada categoria sobre um eixo de um espaço geométrico. Por exemplo, as colunas AREN, ARSO e MUIA definem um espaço de três dimensões como o da figura 3.2. A representação gráfica desse espaço está restrita a três dimensões, mas ele pode ser analiticamente estendido a n dimensões com ângulos retos uns com os outros, fornecendo, nesse caso, uma representação de todas as quatro colunas da tabela.

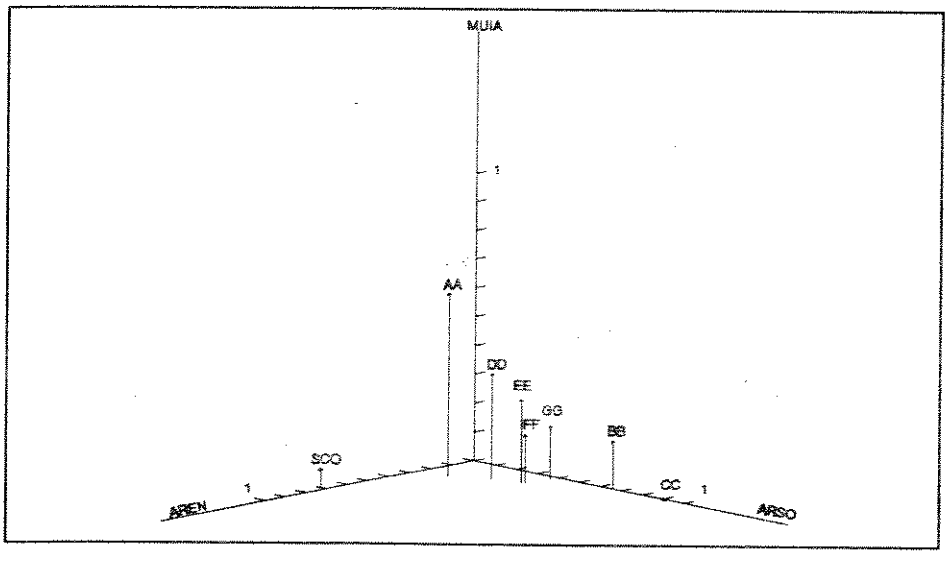

Figura 3.2: Disposição dos pontos-linha (categorias da variável tamanho de conchas) no espaço tridimensional das colunas (variável matriz terrígena).

Dentro do espaço 4-D (espaço das categorias de matriz terrígena), cada linha (categoria de tamanho de conchas) pode ser considerada como um ponto localizado de acordo com seus valores. Se, para cada um desses pontos, desenhar-se

uma linha reta, da origem até o ponto, tem-se uma representação vetorial dos pontos. Resumindo, as oito categorias de tamanho de conchas são plotadas como vetores no espaço imaginário das quatro categorias (dimensão) de matriz terrígena, que representa um espaço vetorial. De forma semelhante, as quatro categorias de matriz terrígena podem ser plotadas no espaço imaginário, 8-D, das categorias de tamanho de conchas.

A medida da relação entre as características dos pontos (extremidades de vetores) pode ser, por exemplo, o ângulo entre os vetores; quanto menor o ângulo definido por dois pontos e a origem, em $\mathbb{R}^n$, maior será a relação entre eles e, quando dois vetores coincidem, diz-ze que eles são perfeitamente correlacionados. Outra medida da relação entre dois pontos pode ser a distância entre eles; quanto mais próximo um ponto estiver do outro, maior a relação entre eles. A disposição dos pontos no espaço reflete as inter-relações entre eles. Do ponto de vista geométrico, as técnicas de análise fatorial definem agrupamentos de pontos quando o número de dimensões é maior do que três.

Se os componentes dos vetores no espaço J-dimensional forem os próprios valores n_{ij}, as proximidades entre os elementos podem ficar deturpadas pela falta de padronização dos dados. Na análise de correspondência, utilizam-se os perfis distâncias, que são dados pelas probabilidades condicionais de a observação aparecer na coluna j, dado que pertence à linha i, ou seja, n_{ij} deve ser dividido pelo total da linha. De forma semelhante, as probabilidades condicionais de a observação aparecer na linha i dado que pertence à coluna j é o mesmo que dividir n_{ij} pelo total da coluna j.

Cada linha de **R** é denotada por $\mathbf{a}_i = [n_{ij}/n_{i+}, \ldots, n_{ij}/n_{i+}]^T$, por exemplo, $\mathbf{a}_2 = [0{,}0000\ 0{,}666667\ 0{,}166667\ 0{,}166667]^T$, e pode ser considerada como as coordenadas de um vetor no espaço Euclidiano, representado no espaço das colunas como um ponto. Assim, tem-se I (8) pontos-linha no espaço J (4)-dimensional. Como existe uma dependência linear entre as coordenadas dos vetores-perfil (cada perfil-linha tem soma igual a 1), isto significa, geometricamente, que os oito pontos-linha estão contidos exatamente em simplexo de comprimento igual a 1. Os perfis das linhas $\mathbf{a}_i$ (com i=1,...,I) são escritos nas linhas de **R**.

A escolha dos perfis das linhas ou das colunas como coordenadas dos pontos no espaço $\mathbb{R}^J$ e $\mathbb{R}^I$ considera que todos os pontos-linha e pontos-coluna têm a mesma importância. Quando se estuda a geometria de um conjunto de pontos, a atribuição de diferentes massas aos vetores determina diferentes graus de importância aos pontos no espaço. Logo, cada ponto tem um peso proporcional à sua freqüência. Atribui-se aos pontos-linha massa igual aos elementos de **r** e, aos pontos-coluna, massa igual aos elementos de **c**. Em notação vetorial, $\mathbf{R} = \mathbf{D}_r^{-1}\mathbf{P} = [a_1 \ldots a_J]^T$.

3.1.2 Definição de Distância

Uma vez definida a nuvem de pontos-linha, faz-se necessário definir a função distância, ou métrica, entre esses pontos, de forma que ela não dependa da escala de medida utilizada. Como já foi mencionado, já que a análise de correspondência lida com vetores de freqüências relativas, a medida de distância no espaço Euclidiano não é uma medida adequada. Isto porque a distância entre 0 e 5% é intuitivamente muito maior do que a distância entre 50 e 55%. Conseqüentemente, a distância escolhida é a **distância Euclidiana ponderada**, conhecida também como **distância quiquadrado**.

No exemplo dado, os perfis-linha são 8 vetores dentro de um tetraedro com os vértices nos valores unitários nas 4 dimensões. Este tetraedro pode ser visto como se tivesse sido esticado diferencialmente ao longo de cada dimensão para fornecer as posições dos pontos que se quer investigar. Esticar os eixos equivale a ponderá-los pelo inverso da raiz quadrada dos centróides: $1/(0{,}39453)^{1/2} = 1{,}59$; $1/(0{,}21328)^{1/2} = 2{,}16$; $1/(0{,}14843)^{1/2} = 2{,}60$ e $1/(0{,}24375)^{1/2} = 2{,}02$. O lado mais esticado corresponde à categoria menos freqüente de matriz terrígena, MUIA, e o lado menos esticado corresponde à categoria mais freqüente, AREN, conforme a figura 3.3. Os pontos são situados neste novo sistema de coordenadas, da mesma forma que no caso dos eixos de comprimentos unitários.

As distâncias nesta nova configuração são as distâncias quiquadradas utilizadas na análise de correspondência. Assim, esta análise pode ser vista como o

seguinte problema: como encontrar os eixos principais dos pontos, com massas associadas, neste novo sistema de coordenadas? De forma geral, se uma matriz de dados tem I linhas e J colunas, e supondo-se J ≤ I, então o espaço dos perfis-linha é um poliedro (J - 1) dimensional com J vértices.

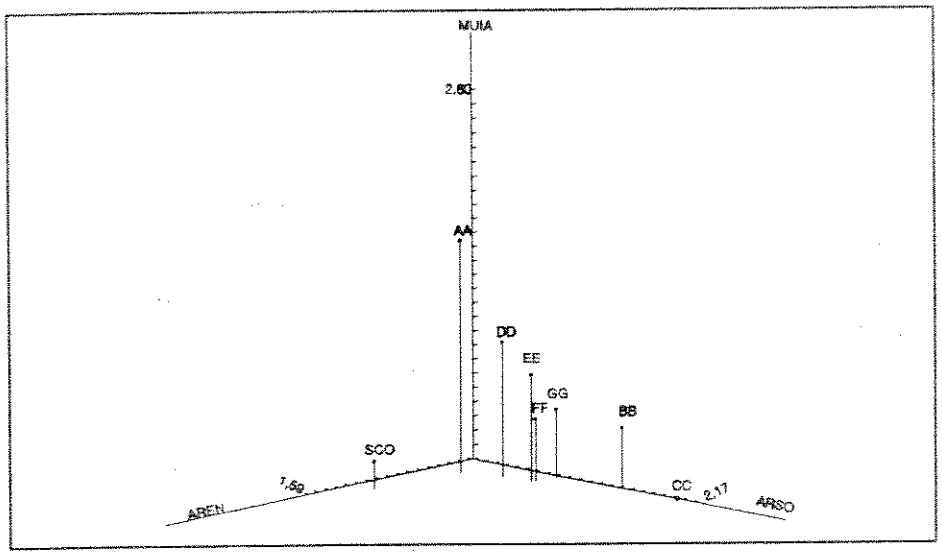

Figura 3.3: Configuração dos 8 pontos-linha no espaço Euclidiano ponderado das colunas.

Portanto, a distância Euclidiana ponderada entre dois pontos-linha, i e i', é definida como:

$$d^2\ (i\ ,\ i') = (\boldsymbol{a}_i - \boldsymbol{a}_{i'})^T\ \boldsymbol{D}_c^{-1}(\boldsymbol{a}_i - \boldsymbol{a}_{i'}) \qquad (1)$$

3.1.3 Ajustamento das Nuvens de Pontos

Uma vez definida a nuvem de pontos-linha no espaço Euclidiano ponderado J-dimensional, deseja-se encontrar o subespaço de dimensão k, com k ≤ 4, que mais se aproxime de todos os pontos da nuvem. Deve-se considerar que estes têm massas diferentes e que a distância entre eles é a distância Euclidiana ponderada. Logo, o subespaço ajustado deve estar mais próximo dos pontos de maiores massas.

Define-se também o centro de gravidade (ou centróide ou ponto médio) da nuvem de pontos-linha, que é o vetor $\mathbf{c} = [c_1 \ldots c_J]^T$, e representa, nesse caso, as

proporções de todas as amostras nas categorias da variável matriz terrígena. Ele é também conhecido como perfil médio das linhas, por ser a média ponderada dos perfis-linha: $c_j = \sum_i r_i(p_{ij}/r_i)/\sum_i r_i = \sum_i p_{ij}$, pois $\sum_i r_i = 1$. É importante observar que o centróide dos perfis-linha, com massas definidas pelos elementos de **r**, é o perfil dos totais das colunas, ou seja, o vetor das massas dos pontos-coluna. Em notação vetorial $\mathbf{c} = \mathbf{R}^T\mathbf{r}$.

A variação espacial total da nuvem de pontos-linha em relação ao seu centróide é quantificada por sua **inércia total**, que é a soma ponderada das distâncias ao quadrado de um ponto ao seu respectivo centróide, considerando as massas dos pontos e a métrica definida, no caso $\mathbf{D}_c^{-1}$. Em notação vetorial, a inércia é o traço da matriz:

$$\textit{Inércia}\ (\boldsymbol{R}) = \textit{traço}\ [\boldsymbol{D}_r\ (\boldsymbol{R} - \mathbf{1}\boldsymbol{c}^T)^T\ \boldsymbol{D}_c^{-1}(\boldsymbol{R} - \mathbf{1}\boldsymbol{c}^T)] \qquad (2)$$

Uma vez calculada a matriz de distâncias dos perfis-linha ao seu centróide, deseja-se encontrar o subespaço de menor dimensão que melhor se ajusta a esta nuvem de pontos-linha. A projeção da nuvem neste subespaço permite analisar as similaridades entre os pontos-linha. Dessa forma, a geometria dos 8 pontos-linha fica completamente especificada pelo cálculo da decomposição generalizada em valores singulares da matriz **R** ou da matriz centrada no seu ponto médio ou centróide, $\mathbf{R} - \mathbf{1}\mathbf{c}^T$.

O subespaço de dimensão k (com $k \leq \min\{I,J\}$), que melhor se ajusta à nuvem de pontos-linha, é definido pelos k vetores singulares generalizados (direitos e esquerdos, respectivamente), de $\mathbf{R} - \mathbf{1}\mathbf{c}^T$, correspondentes aos k maiores valores singulares. Este problema é análogo a encontrar os maiores componentes principais de uma matriz de I observações e J variáveis, considerando-se a massa das linhas ($\mathbf{D}_r$) e a métrica quiquadrado ($\mathbf{D}_c^{-1}$).

A escolha natural para esse ajuste é o método dos mínimos quadrados ponderados pela massa das linhas, r_i. Com essa escolha da função objetivo o subespaço de melhor ajuste passa pelo centróide **c**. Se a origem do gráfico for deslocada para **c**, o subespaço que melhor se ajusta à nuvem de pontos-linha pode ser visto como o gerado pelos autovetores correspondentes aos k-maiores autovalores da

matriz não-simétrica $\mathbf{R} - \mathbf{1c}^T$. Essa decomposição fornece a solução teórica completa do problema de minimizar a distância entre dois pontos no espaço Euclidiano ponderado, utilizando-se o método dos mínimos quadrados.

Os eixos principais e as coordenadas dos perfis-linha com relação a esses eixos são obtidos da decomposição generalizada em valores singulares de $\mathbf{R} - \mathbf{1c}^T$, de modo que os vetores singulares direitos e esquerdos são ortonormalizados em relação a $\mathbf{D}_r$ e $\mathbf{D}_c^{-1}$, respectivamente, conforme a equação:

$$\boldsymbol{R} - \mathbf{1}\boldsymbol{c}^T = \boldsymbol{N}\,\boldsymbol{D}_\phi\,\boldsymbol{M}^T \tag{3}$$

com:

$$\boldsymbol{N}^T\,\boldsymbol{D}_r\,\boldsymbol{N} = \boldsymbol{I} \qquad e \qquad \boldsymbol{M}^T\,\boldsymbol{D}_c^{-1}\,\boldsymbol{M} = \boldsymbol{I} \tag{4}$$

As colunas de **M** definem os eixos principais e as colunas de $\mathbf{ND}_\phi$ definem as coordenadas dos pontos-linha sobre estes novos eixos.

Como a representação gráfica desejada é bidimensional, as duas primeiras colunas de **M** definem as coordenadas dos dois eixos principais da nuvem linha no espaço 4-D; as duas primeiras colunas de $\mathbf{ND}_\phi$ definem as coordenadas dos pontos-linha com relação a esses novos eixos. Seja $\mathbf{F} = \mathbf{ND}_\phi$; então $F_{(2)}$ é:

$$F_{(2)} = \begin{vmatrix} 0{,}25857 & 1{,}33826 \\ 1{,}04152 & -0{,}16107 \\ 1{,}22088 & -0{,}64244 \\ 0{,}42291 & 0{,}49179 \\ 0{,}56056 & 0{,}24398 \\ 0{,}55481 & -0{,}11032 \\ 0{,}72729 & -0{,}10783 \\ -0{,}73837 & -0{,}02269 \end{vmatrix} \tag{5}$$

Os valores singulares ϕ, da matriz centrada são: 0,67737 0,19686 e 0,11062. As coordenadas dos eixos principais, $\mathbf{M}_{(2)}$, são:

$$M_{(2)} = \begin{vmatrix} -0{,}462012 & -0{,}027414 \\ 0{,}280621 & -0{,}134897 \\ 0{,}091879 & 0{,}340738 \\ 0{,}089512 & -0{,}178427 \end{vmatrix} \tag{6}$$

Pode-se plotar as linhas de $\mathbf{F}_{(2)}$ como pontos no sistema de coordenadas retangular como o da figura 3.4, que é a análise de correspondência dos perfis-linha da matriz original de dados **N**.

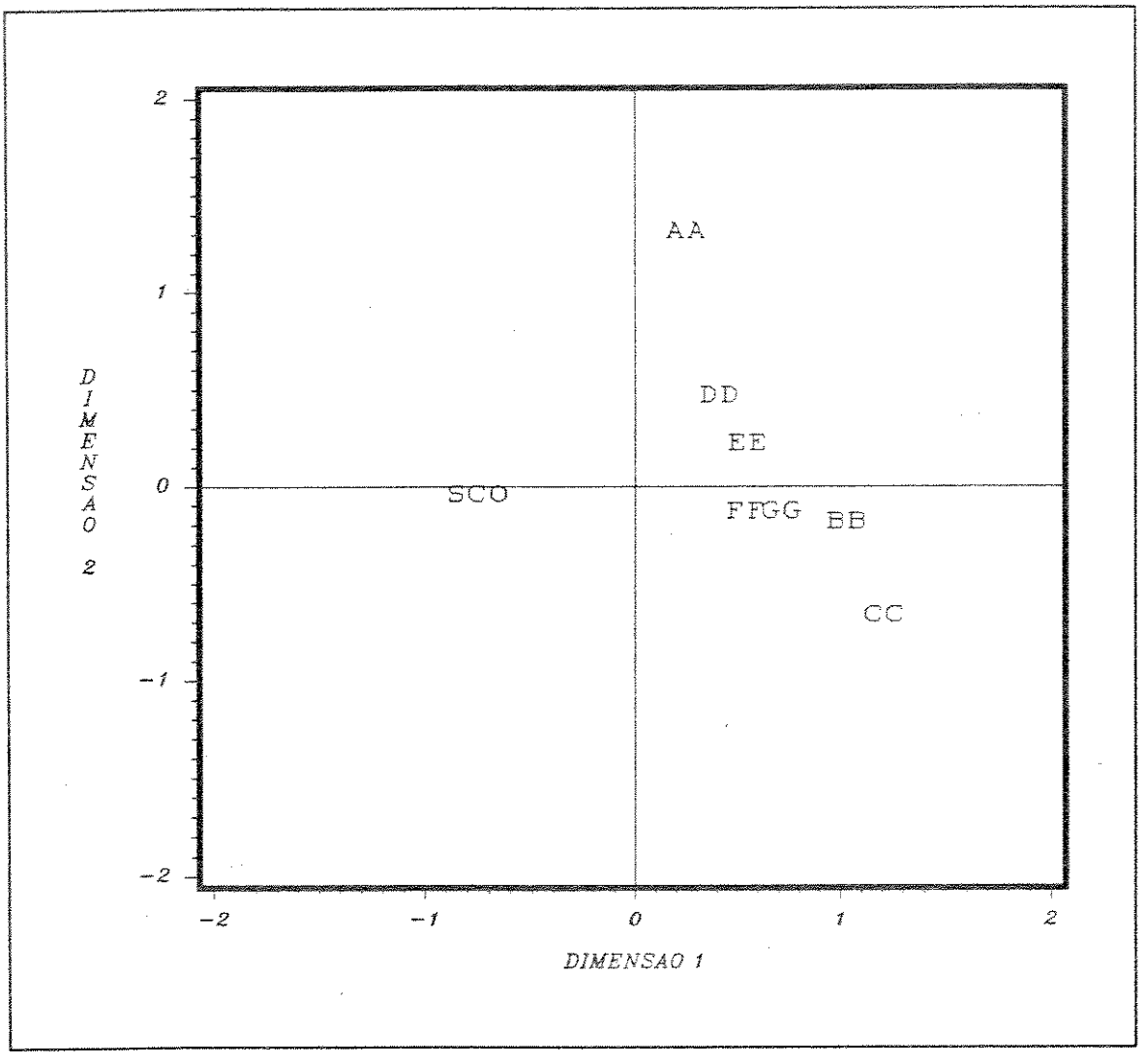

Figura 3.4: Análise de correspondência simples dos pontos-linha (tamanho de conchas) dimensão 1 (90%), dimensão 2 (7.6%).

A qualidade dessa representação bidimensional de pontos, calculada em termos do quadrado da soma dos valores singulares, é 97,6%. A inércia total é igual à soma dos valores singulares ao quadrado: $(0,67737)^2 + (0,19686)^2 + (0,11062)^2 = 0,50982$. Por convenção, escreve-se a porcentagem da inércia em cada eixo. Por

exemplo, para o eixo 1, $\phi_1^2 = (0,67737)^2 = 0,45883$ que corresponde a 90,00% da inércia total; para o eixo 2, $\phi_2^2 = (0,19686)^2 = 0,03875$, que corresponde a 7,60% da inércia total e para o eixo 3, $\phi_3^2 = (0,11062)^2 = 0,01224$, que corresponde a 2,40% da inércia total. Assim, pode-se interpretar a figura 3.4 como uma representação quase exata dos pontos, uma vez que os dois eixos consideram 97,6% da inércia total dos pontos.

Na figura 3.4, as amostras com conchas tamanho DD, EE, GG, FF e BB são relativamente similares em termos da quantidade de matriz terrígena. As amostras sem concha (SCO) e com conchas de tamanho AA e CC são bastante diferentes com relação à quantidade de matriz terrígena. Dessa forma, as linhas da matriz de dados foram mapeadas como pontos no plano e o exame de suas posições relativas sugere similaridades e diferenças entre os tamanhos de conchas em relação ã quantidade de matriz terrígena.

3.1.4 A Dualidade do Método

Acabou-se de examinar a geometria dos pontos-linha da tabela de contingência **N**, o que equivale a uma análise de componentes principais ponderada ou ainda, a uma análise R-modal. Da mesma forma, pode-se investigar a geometria dos pontos-coluna da mesma tabela. A geometria dos perfis-coluna está diretamente relacionada com a geometria dos perfis-linha, justificando o termo análise de correspondência.

Considere agora o conjunto das colunas da tabela **N**. Uma maneira conveniente de abordar a geometria dos perfis-coluna é aplicar a mesma metodologia à matriz transposta $\mathbf{N}^T$ (tabela 3.4).

Dividindo-se cada linha de $\mathbf{N}^T$ por seu total, tem-se a matriz **C** (JxI) dos perfis-coluna. Neste caso, examinam-se as relações entre as categorias da variável matriz terrígena no espaço das categorias da variável tamanho de conchas (tabela 3.5). Os quatro perfis das categorias de matriz terrígena, $\mathbf{b}_i$ (i=1,...,J) são escritos nas linhas de **C** e definem 4 pontos no espaço 8-D.

Tabela 3.4: Tabela de Contingência, N^T, das amostras da Formação Lagoa Feia.

Matriz Terrígena	Tamanho de conchas								Totais
	AA	BB	CC	DD	EE	FF	GG	SCO	
AREN	2	0	0	14	16	37	10	426	505
ARSO	1	4	9	20	40	101	92	6	273
MUIA	5	1	0	27	30	43	42	42	190
PURO	0	1	1	15	22	84	85	104	312
Totais	8	6	10	76	108	265	229	578	1280

Esses perfis são ponderados pelas massas (dadas pelos elementos de **c**), $\mathbf{D}_c^{-1}$, no espaço ponderado pela métrica quiquadrado definida por $\mathbf{D}_r^{-1}$. Em notação vetorial $\mathbf{C} = \mathbf{D}_c^{-1}\mathbf{P}^T = [b_1 ... b_J]^T$. Assim, os elementos de **r** e **c** desempenham papel duplo: por um lado, ponderando os perfis e por outro, ponderando ou reescalonando os eixos ou dimensões.

Tabela 3.5: Matriz dos Perfis-coluna, C, obtida pela divisão de cada elemento de N^T, pelo total de sua respectiva coluna.

Matriz dos Perfis-coluna, C.									
b_1	0,00396	0,00000	0,00000	0,02772	0,03168	0,07327	0,01980	0,84356	1
b_2	0,00366	0,01465	0,03297	0,07326	0,14652	0,36996	0,33699	0,02198	1
b_3	0,02632	0,00526	0,00000	0,14210	0,15789	0,22632	0,22105	0,22105	1
b_4	0,00000	0,00320	0,00320	0,04808	0,07051	0,26923	0,27243	0,33333	1

As coordenadas dos eixos principais da nuvem-coluna e as coordenadas dos pontos-coluna com relação ao subespaço bidimensional ótimo são fornecidas pela decomposição generalizada de $\mathbf{C} - \mathbf{1r}^T$:

$$C - \mathbf{1}r^T = WD_\omega Z^T \tag{7}$$

com:

$$W^T D_c W = Z^T D_r^{-1} Z = I \tag{8}$$

onde: as duas primeiras colunas de **Z** definem as coordenadas dos eixos principais da nuvem de pontos-coluna no espaço das linhas e as duas primeiras colunas de $\mathbf{WD}_\omega$ definem as coordenadas principais dos pontos-coluna com relação a esses novos eixos. Dessa forma, se $\mathbf{G} = \mathbf{WD}_\omega$,

$$G_{(2)} = \begin{vmatrix} -0,793227 & -0,013679 \\ 0,8912356 & -0,124511 \\ 0,4192748 & 0,4518939 \\ 0,2487495 & -0,144104 \end{vmatrix} \tag{9}$$

Os valores singulares ω são: 0,67737 0,19686 e 0,11062, idênticos aos obtidos da análise das linhas. As coordenadas dos eixos principais, $\mathbf{Z}_{(2)}$ são:

$$Z_{(2)} = \begin{vmatrix} -0,002388 & 0,042487 \\ 0,007207 & -0,003835 \\ 0,014081 & -0,025495 \\ 0,014081 & -0,025495 \\ 0,037070 & 0,148327 \\ 0,069825 & 0,104569 \\ 0,169571 & -0,116017 \\ 0,192090 & -0,097996 \\ -0,492231 & -0,052042 \end{vmatrix} \tag{10}$$

Pode-se plotar as linhas de $\mathbf{G}_{(2)}$ como pontos no sistema de coordenadas retangular como o da figura 3.5, conhecido como análise de correspondência dos perfis-coluna da matriz original de dados **N**, ou ainda análise Q-modal. A qualidade dessa representação de pontos é semelhante à dos pontos-linha.

Na verdade, não é necessário recalcular a solução dual, uma vez que ela pode ser obtida do primeiro problema. As matrizes **F** e **M**, da solução R-modal, estão relacionadas com as matrizes **G** e **Z**, da solução Q-modal, da seguinte forma:

$$F = D_r^{-1} W D_\omega \qquad e \qquad G = D_c^{-1} M D_\phi \tag{11}$$

Por exemplo, o elemento m_{31} de **M** é 0,0918792. O correspondente elemento de **G** é: $g_{31} = m_{31}\,\phi_1/c_3 = (0{,}0918792)(0{,}6773686)/(0{,}148375) = 0{,}419274$. Note-se que os sinais das colunas de **G** e **M** devem concordar com os sinais de **F** e **W**. Quando se calculam as soluções dos dois problemas separadamente, pode acontecer que os sinais sejam diferentes. Contudo, tendo visto que os dois problemas estão relacionados, não é necessário calculá-los separadamente. A solução de um problema é suficiente para obter a solução do outro, graças à equação 11, na qual a concordância de sinais está implícita.

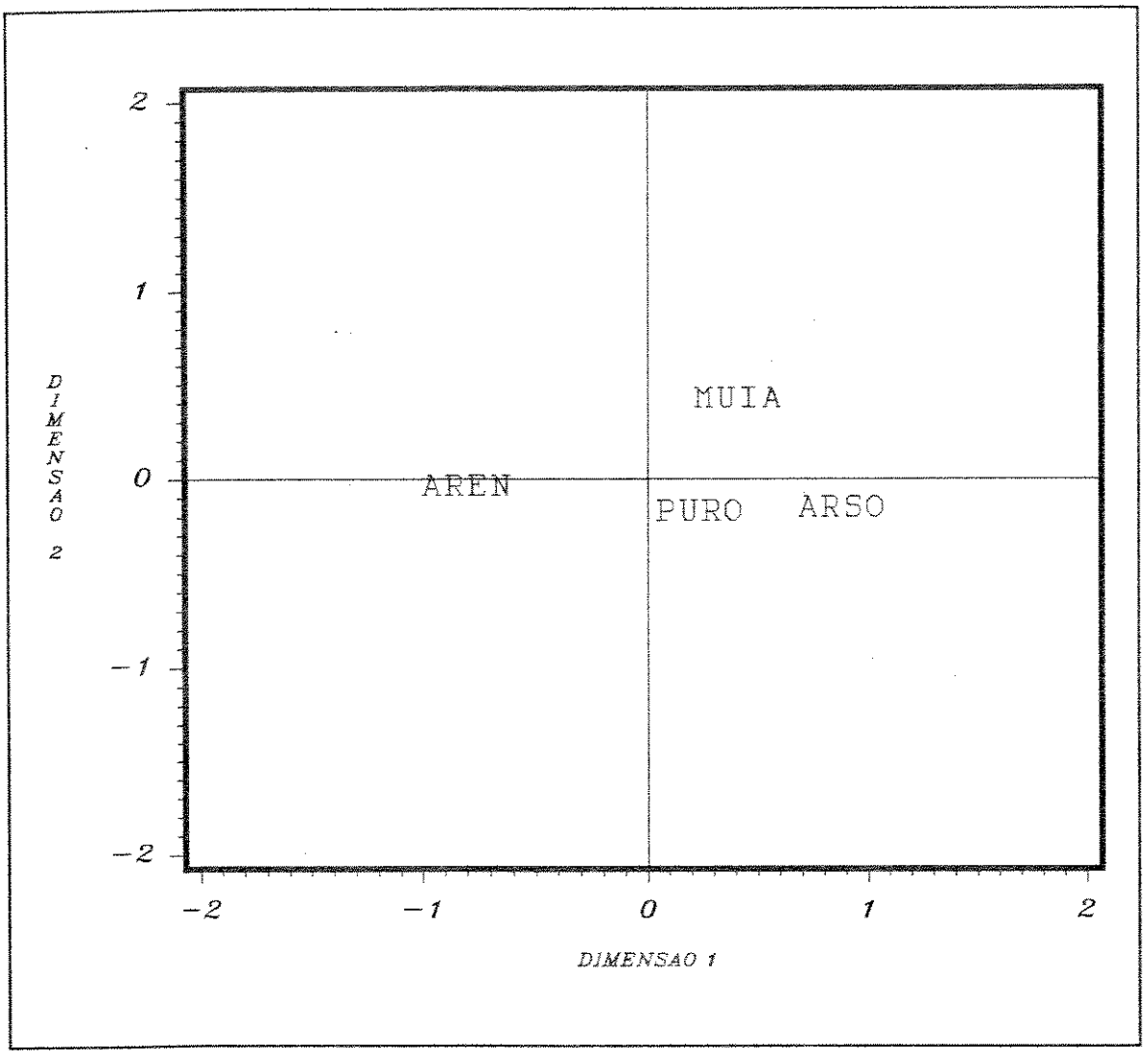

Figura 3.5: Análise de correspodência dos pontos-coluna (matriz terrígena); dimensão 1 (90%) e dimensão 2 (7,6%).

Nota-se que a inércia total e sua decomposição em inércias principais são exatamente as mesmas nos dois casos ($\phi = \omega$). A equação 11 mostra que as coordenadas dos pontos perfis com relação aos eixos principais, da análise R-modal, estão relacionadas aos eixos originais dos pontos perfis da análise Q-modal e vice-versa, pela pré ou pós-multiplicação das matrizes diagonais. Essa simetria dos dois problemas, mais o fato de os valores singulares e seus quadrados (as inércias principais) serem os mesmos nos dois casos, são a essência da dualidade do método.

Considerando esta dualidade, é suficiente calcular a matriz de correspondência **P**, centrá-la no centróide das linhas e colunas ($\mathbf{P} - \mathbf{rc}^T$) e decompô-la para ajustar os subespaços k-dimensionais relativos às linhas e colunas, de modo que esses subespaços sejam os mais próximos dos pontos em termos da soma ponderada das distâncias ao quadrado. Eles serão definidos pelos k vetores singulares generalizados (esquerdos e direitos, respectivamente) de $\mathbf{P} - \mathbf{rc}^T$, nas métricas $\mathbf{D}_c^{-1}$ e $\mathbf{D}_r^{-1}$, correspondendo aos k maiores valores singulares.

Em outras palavras, os autovetores direitos e esquerdos definem os eixos principais das nuvens linhas e colunas, respectivamente. Portanto, em símbolos:

$$P - rc^T = AD_\alpha B^T \tag{12}$$

com:

$$A^T D_r^{-1} A = I \qquad e \qquad B^T D_c^{-1} B = I \tag{13}$$

sendo $\alpha_1 > \ldots \geq \alpha_k > 0$, os valores singulares de $\mathbf{P} - \mathbf{rc}^T$. Note que $\alpha = \Phi = \omega$.

As colunas de **A** definem os eixos principais da nuvem linha e as colunas de **B** definem os eixos principais da nuvem coluna. A decomposição em valores singulares de $\mathbf{P} - \mathbf{rc}^T$, desprezando o último valor singular (que é zero) e o último vetor singular direito e esquerdo, é exatamente a decomposição em valores singulares de **P**, desprezando o primeiro valor singular (que é igual a um), o primeiro vetor singular esquerdo, **r**, e o primeiro vetor singular direito **c**.

A inércia total é a mesma em ambas as nuvens de pontos (linhas e colunas). É a média ao quadrado do coeficiente de contingência calculado sobre N,

e também a estatística quiquadrada dividida pelo total geral, n. Um valor significativo da estatística quiquadrado significa, geometricamente, que ocorre um desvio significativo dos perfis-linha em relação ao seu centróide.

Na prática, não há interesse nas matrizes **M** e **Z**, que definem as coordenadas dos eixos principais das nuvens linha e coluna. A relação entre os novos eixos com os eixos originais é de importância secundária. Há interesse pelas posições relativas dos pontos em relação aos novos eixos.

As coordenadas principais dos perfis-linha com relação aos seus eixos principais estão relacionadas com os eixos principais da outra nuvem de pontos dos perfis-coluna por simples reescalonamento. O inverso também é verdadeiro. Considere os perfis-linha com relação aos seus eixos principais **B** (na métrica quiquadrado $\mathbf{D}_c^{-1}$). Note-se que, como os eixos principais são ortonormais ($\mathbf{B}^T\mathbf{D}_c^{-1}\mathbf{B} = \mathbf{I}$), essas coordenadas são dadas pelo produto escalar dos perfis centrados ($\mathbf{R} - \mathbf{1c}^T$) com **B**. Conseqüentemente, pode-se escrever:

$$\boldsymbol{F} = \boldsymbol{D}_r^{-1}\boldsymbol{A}\boldsymbol{D}_\alpha \qquad e \qquad \boldsymbol{G} = \boldsymbol{D}_c^{-1}\boldsymbol{B}\boldsymbol{D}_\alpha \tag{14}$$

A figura 3.6 mostra a representação simultânea das linhas e colunas da tabela de contingência **N**, no mesmo plano fatorial ou análise de correspondência simples de **N**.

3.1.5 Princípio da Equivalência Distributiva

Uma das vantagens de se usar a distância quiquadrado e da dualidade resultante das duas nuvens de pontos é o chamado de Princípio da Equivalência Distributiva. Ele estabelece que, se dois perfis, por exemplo perfis-linha, são idênticos (ou equivalentes distributivamente), então essas duas linhas da tabela de contingência original podem ser somadas para gerar uma única linha sem afetar a geometria dos perfis-coluna. O mesmo resultado ocorre com perfis-coluna idênticos.

Geometricamente, isso significa que se pode unir dois pontos da nuvem

que caiam na mesma posição, cuja massa será a soma das massas de cada ponto, sem afetar a geometria dos pontos da outra nuvem. Esse resultado é peculiar à geometria da análise de correspondência.

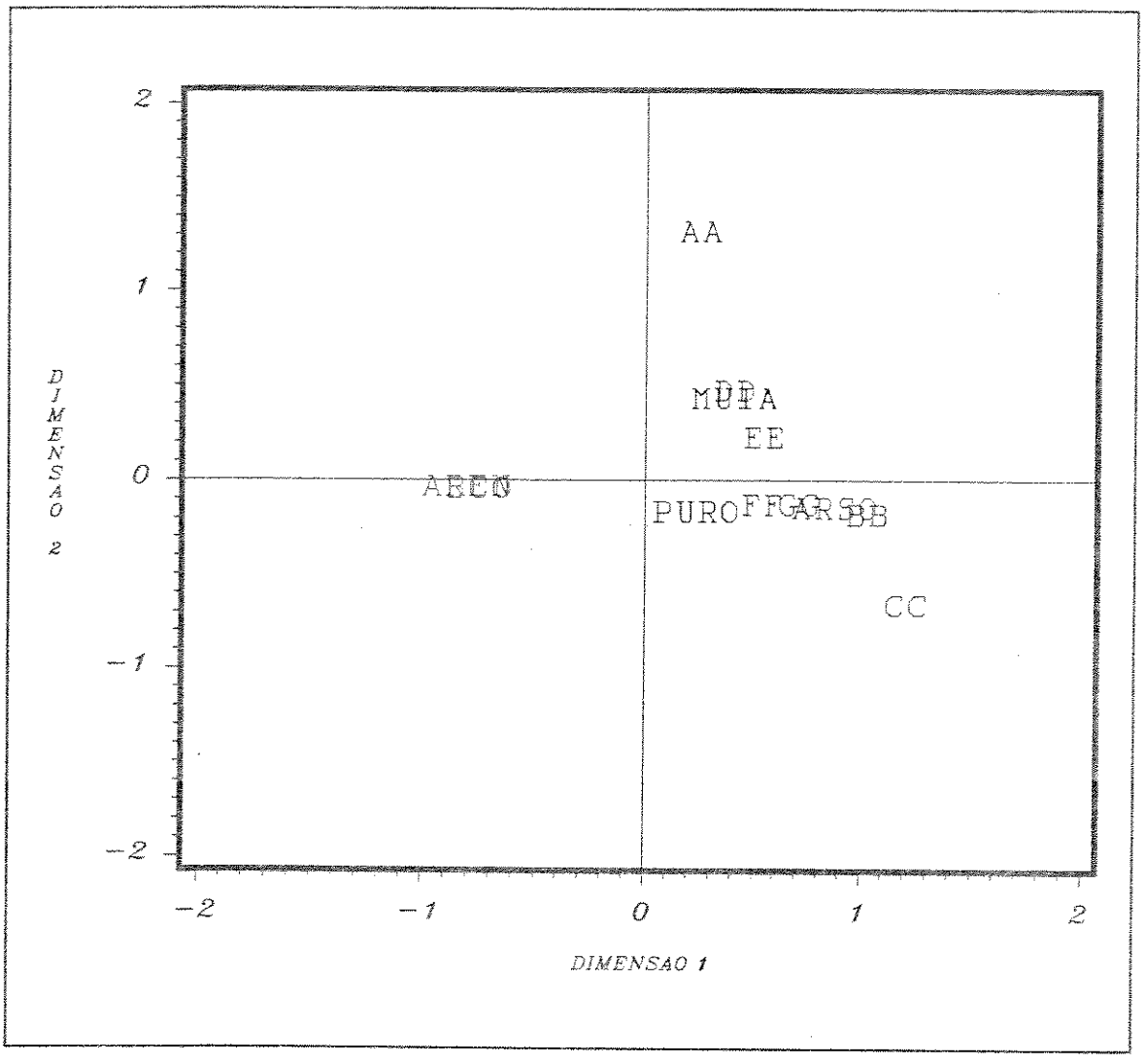

Figura 3.6: Análise de correspondência simples entre as linhas e colunas da tabela de contingência N, (tamanho de conchas x matriz terrígena). Dim. 1 (90%), Dim. 2 (7,6%).

3.2 INTERPRETAÇÃO DOS RESULTADOS

Quando se aplica a análise de correspondência, é comum a representação gráfica bidimensional dos pontos-linha e dos pontos-coluna, obtida através dos primeiros eixos principais tomados dois a dois. A interpretação dos gráficos representando as nuvens de pontos nos planos de projeção é o mais difícil objeto da análise. Uma interpretação geométrica da análise de correspondência é apresentada em Greenacre e Hastie (1987).

Deve-se verificar quais são os eixos mais representativos da análise e interpretar a proximidade entre os pontos de uma mesma nuvem. Não é possível interpretar a proximidade entre pontos de nuvens diferentes, uma vez que nenhuma medida de distância entre nuvens diferentes foi estabelecida. Apenas a distância entre pontos de uma mesma nuvem foi explicitamente definida. A análise de pontos pertencentes a nuvens diferentes se dá através de ângulos entre os vetores que vão da origem do gráfico até cada um dos pontos.

Como a representação bidimensional mostra a projeção dos pontos originais sobre esse plano e não mostra quais os pontos que estão mais próximos e quais estão mais distantes, informações adicionais devem ser consideradas para se interpretar corretamente o gráfico. Deseja-se, na verdade, entender a geometria de um conjunto de pontos em um espaço multidimensional, através de um gráfico aproximado, de menor dimensão, e quer-se saber onde esse gráfico é preciso e onde não o é. Isso é análogo a construir um modelo para os dados e estudar tanto o modelo como a qualidade do ajuste do modelo aos dados, ou seja, onde o modelo se ajusta bem aos dados e onde não se ajusta.

3.2.1 Inércia

Como já foi dito, a inércia quantifica a variação espacial de cada nuvem de pontos, ou ainda, é a soma ponderada das distâncias ao quadrado dos pontos aos seus centróides. Em notação matricial:

$$inércia(r) = traço[D_r(R - 1c^T)\ D_c^{-1}\ (R - 1c^T)^T] \quad (15)$$

A inércia total é igual em ambas as nuvens e pode-se verificar que:

$$inércia(r) = inércia(total) = traço[D_r^{-1}(P - rc^T)\ D_c^{-1}\ (P - rc^T)^T] \quad (16)$$

No exemplo em questão, obteve-se a representação quase exata dos pontos, porque somente 2,4% da inércia total não está representada neste espaço bidimensional. Na verdade, a melhor representação seria em três dimensões, onde seria possível considerar também a inércia do terceiro eixo. Na prática, contudo, quando se lida com grandes matrizes de dados, é raro obter uma representação bidimensional tão boa. Se grande porcentagem da inércia total é explicada ao longo de outros eixos principais, significa que os pontos não são bem representados pelos dois primeiros eixos.

Para ilustrar melhor os princípios envolvidos, escolheu-se a análise de correspondência unidimensional dos referidos dados, dada na figura 3.7. Ela representa uma visão aproximada dos dados e sabe-se que é uma boa visão pois 90% da inércia é representada por esse eixo. Informalmente, pode-se interpretar essa dimensão como separando as amostras com conchas, à direita, das amostras sem conchas, à esquerda, ou ainda como separando os terrígenos dos carbonatos.

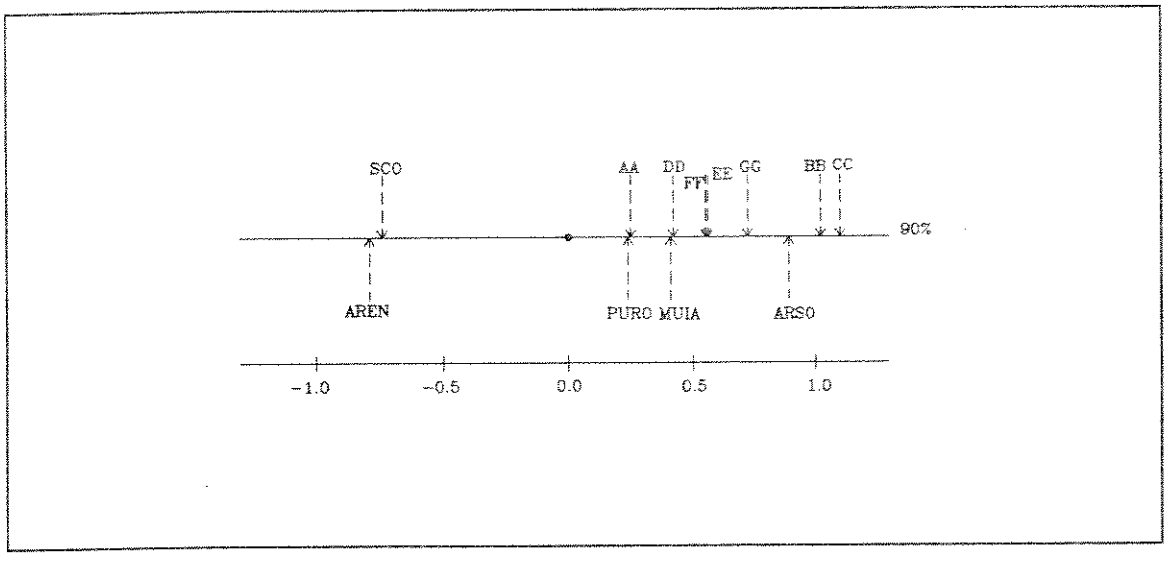

Figura 3.7: Representação unidimensional das linhas e colunas da tabela 3.1.

3.2.2 Contribuição Absoluta

Mais formalmente, portanto, pode-se quantificar quanto cada ponto participa no estabelecimento do primeiro eixo. A inércia ao longo dele, 0,45883 = $(0,67737)^2$, é igual à soma ponderada, pelas massas dos pontos, das distâncias ao quadrado de cada perfil-linha em relação à origem. Cada termo dessa soma pode ser expresso como a porcentagem da inércia principal e, por isso, é chamada de **contribuição absoluta** dos pontos à inércia principal ou ao eixo principal. Por exemplo, o ponto AA tem massa igual a 0,00625 e a distância em relação ao centróide é 0,25857. A sua contribuição absoluta à inércia do eixo principal é então 0,00625 x $(0,25857)^2$ = 0,000417, que é 0,09% de 0,45883 (inércia do primeiro eixo).

Neste exemplo vê-se, pela tabela 3.6, que os pontos que representam o tamanho de conchas GG e SCO contribuem com mais de 70% da inércia principal dos perfis-linha, enquanto que, entre os perfis-coluna, só o ponto AREN contribui com 54,1%. Imaginando esses pontos nas suas posições de origem, nos dois espaços correspondentes, exercendo forças de atração para o eixo principal em virtude de suas posições e de suas massas, então eles são os pontos com alta contribuição que estabelecem a orientação final dos eixos principais. Os pontos com maior massa contribuem mais para o primeiro eixo. Assim, o eixo principal tende mais para a direção dos pontos de maior massa. No entanto, existem casos em que pontos com massas relativamente pequenas têm alta contribuição à inércia dos eixos, devido à sua grande distância do centróide.

Com base nas posições dos pontos no primeiro eixo e no conhecimento dos pontos que mais contribuem para ele, pode-se associar algum nome descritivo ao eixo, para guiar a interpretação. Nesse exemplo, o primeiro eixo alinha grupos em termos de presença ou ausência de conchas, ou em termos de presença ou ausência de matriz terrígena.

Tabela 3.6 : Inércias dos pontos com relação ao primeiro eixo principal e suas contribuições.

	Inércia do ponto no primeiro eixo	Contribuição do ponto para a inércia principal (%)	Contribuição relativa: $\cos^2\theta$	Ângulo entre o ponto e o primeiro eixo
Linhas				
AA	0,25857	0,9	0,035882	79,1
BB	1,04152	1,1	0,783262	332,2
CC	1,22088	2,5	0,525059	43,6
DD	0,42291	2,3	0,425032	49,3
EE	0,56056	5,8	0,808115	26,0
FF	0,55481	13,9	0,961653	348,7
GG	0,72729	20,6	0,955772	347,9
SCO	-0,73837	53,6	0,999017	181,8
Colunas				
AREN	-0,793227	54,1	0,996632	183,3
ARSO	0,891236	36,9	0,958302	348,2
MUIA	0,419275	5,7	0,461356	47,2
PURO	0,248750	3,3	0,550495	42,1

3.2.3 Contribuição Relativa

Tendo-se interpretado a primeira dimensão, deseja-se agora saber quão perto cada ponto está desse subespaço unidimensional. Por exemplo, pode-se medir o ângulo θ entre um ponto observado e o primeiro eixo. É conveniente examinar o co-seno ao quadrado desse ângulo porque, para cada ponto, os co-senos ao quadrado desses ângulos com o conjunto completo de eixos principais ortogonais somam 1. Uma abordagem equivalente vem do fato de que a inércia de um ponto $r_i\, d_i^2$ (ou seja, o i-ésimo perfil-linha com massa r_i e distância d_i do centróide) é decomposta ao longo dos eixos principais. A parte dessa inércia ao longo do primeiro eixo é $r_i f_{i1}^2$, onde f_{i1}^2 é a coordenada do ponto sobre este eixo (figura 3.8).

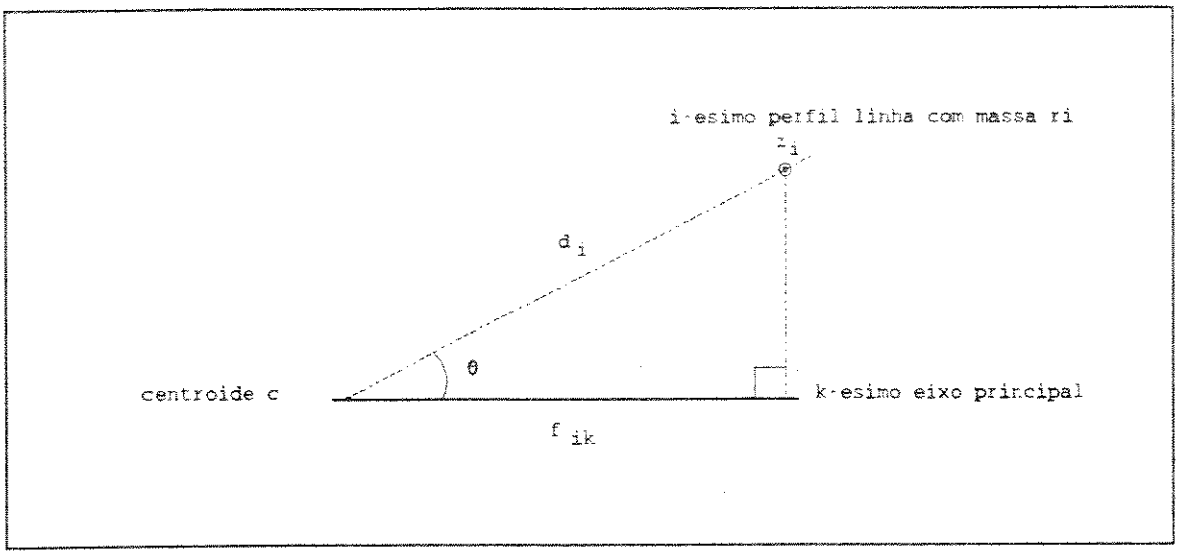

Figura 3.8: Coordenada do i-ésimo perfil-linha relativa ao k-ésimo eixo principal, o qual está a uma distância d_i do centróide c. $Cos^2\theta = (f_{ik}/d_i)^2$ é denominado de contribuição relativa do eixo para o k-ésimo ponto.

Expressando em termos de proporção da inércia total dos pontos tem-se:

$$\frac{r_i f_i^2}{r_i d_i^2} = \left(\frac{f_{ij}}{d_i}\right)^2 = \cos^2\theta \qquad (17)$$

O $\cos^2\theta$ é designado de contribuição do eixo para a inércia do ponto. Se $\cos^2\theta$ é um valor alto, então o eixo explica a inércia do ponto muito bem; equivalentemente, se θ é um ângulo pequeno, o vetor perfil é dito estar na direção do eixo, ou estar relacionado com o eixo. Os valores de $\cos^2\theta$ são também chamados de **contribuições relativas**, pois são independentes da massa dos pontos.

Geralmente, uma alta contribuição do ponto para a inércia do eixo implica alta contribuição relativa do eixo para a inércia do ponto, mas o inverso não é necessariamente verdadeiro. A contribuição relativa dá uma indicação da qualidade da representação de cada ponto individualmente.

Todo esse desenvolvimento teórico pode ser acompanhado através do programa P.5, realizado no módulo IML (*Interactive Matrix Language*) do SAS. No entanto, não é necessário utilizar esse programa, uma vez que o procedimento PROC

CORRESP, do módulo STAT, realiza a análise de correspondência, gerando todos esses resultados: inércia, contribuições, coordenadas dos pontos, etc, de forma bastante simples. No anexo AX.1 encontra-se o resultado da análise de correspondência do exemplo dado.

O fluxograma da figura 3.9 mostra o resumo esquemático das operações realizadas na análise de correspondência.

3.3 PERFIL SUPLEMENTAR

Uma vez que os eixos principais das nuvens de pontos foram estabelecidos, é possível representar pontos adicionais no espaço definido pelos perfis. O conceito de **perfil suplementar** é muito importante na análise de correspondência, bem como em qualquer representação gráfica baseada na SVD. Esses pontos adicionais podem ser vistos como pontos com massa igual a zero, uma vez que eles não participam da definição da distância quiquadrado, nem na determinação dos eixos principais.

A vantagem de se mostrar tais pontos é que eles podem melhorar a interpretação dos eixos principais e dos padrões ou estruturas observadas nos perfis.

3.4 ANÁLISE DE CORRESPONDÊNCIA MÚLTIPLA

Esse tipo de análise é utilizada para a representação gráfica de mais de duas variáveis discretas. A análise de correspondência múltipla é, essencialmente, a aplicação do mesmo algoritmo descrito na análise de correspondência simples, aplicada a uma matriz indicadora **Z** (tabela 3.7) e não mais à tabela de contingência **N**.

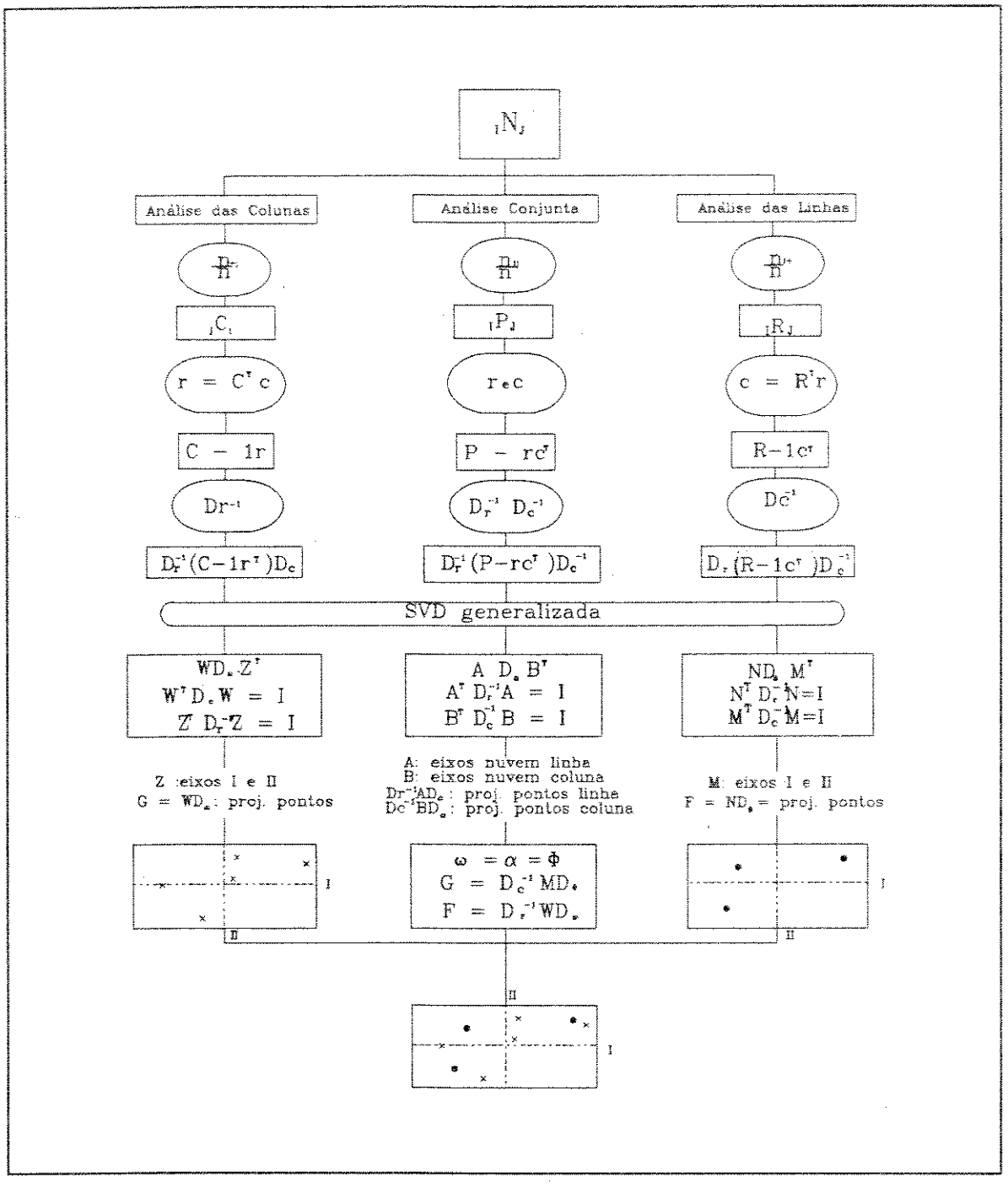

Figura 3.9: Fluxograma do resumo da análise de correspondência.

Tabela 3.7: Matriz Indicadora Z.

	AA	BB	CC	DD	EE	FF	GG	SCO	AREN	ARSO	MUIA	PURO
1	1	0	0	0	0	0	0	0	1	0	0	0
2	1	0	0	0	0	0	0	0	1	0	0	0
3	1	0	0	0	0	0	0	0	0	1	0	0
4	1	0	0	0	0	0	0	0	0	0	1	0
5	1	0	0	0	0	0	0	0	0	0	1	0
6	1	0	0	0	0	0	0	0	0	0	1	0
7	1	0	0	0	0	0	0	0	0	0	1	0
8	1	0	0	0	0	0	0	0	0	0	1	0
1280	0	0	0	0	0	0	0	1	0	0	0	1

A tabela de contingência **N** pode ser considerada como uma matriz condensada originada da matriz indicadora **Z** (1280 x 12). As colunas da matriz indicadora referem-se às 8 categorias da variável tamanho de conchas e às 4 categorias da variável matriz terrígena; cada linha de **Z** consiste de 10 zeros e 2 uns, de modo que os uns indicam quais as categorias que cada amostra pertence. Por exemplo, na tabela 3.1, observa-se que 2 amostras caem na categoria AREN (da variável matriz terrígena) e na categoria AA (da variável tamanho de conchas); assim, existem 2 linhas de **Z** cujos elementos são [1 0 0 0 0 0 0 0 1 0 0 0]. Como existem somente duas variáveis discretas, a única informação perdida quando se condensou a matriz indicadora na forma da tabela de contingência **N** foi a identificação de cada amostra.

O que se quer agora é fazer a análise de correspondência da matriz indicadora e ver sua relação com a análise anterior da tabela de contingência. A figura 3.10 mostra a análise de correspondência da matriz indicadora. A despeito da grande mudança de escala, a configuração dos pontos permanece a mesma dos pontos-linha e pontos-coluna da tabela de contingência, **N**, já citada. O que pode ser ressaltado é que o gráfico foi esticado ao longo dos seus eixos. Além disso, ver-se-á mais adiante,

que as posições relativas ao longo desses eixos permanecem idênticas, apesar das diferenças nas escalas dos eixos. A inércia total e a inércia principal são bem maiores e as diferenças nas sucessivas porcentagens da inércia são menos drásticas. Por exemplo, a primeira e a segunda inércia são 0,83868 e 0,59843, respectivamente (16,7% e 11,9% da inércia total, respectivamente), enquanto que esses mesmos valores caem para 0,45883 e 0,03875 (90% e 7,6%, respectivamente), na análise anterior. De fato, obtem-se dez eixos principais da matriz indicadora, enquanto que a análise da tabela de contingência gerou somente três eixos.

3.4.1 Geometria das Colunas de Z

Denotam-se os números de linhas e de colunas da tabela de contingência, **N**, por J_1 e J_2, respectivamente. A matriz indicadora associada é denominada **Z**, com I linhas e $(J_1 + J_2)$ colunas, e é particionada em $\mathbf{Z} = [\mathbf{Z}_1\ \mathbf{Z}_2]$, com $\mathbf{Z}_1$ = [AA BB CC DD EE FF GG SCO] e $\mathbf{Z}_2$ = [AREN ARSO MUIA PURO]. A tabela de contingência das duas variáveis cruzadas é $\mathbf{N} = \mathbf{Z}_1^T\mathbf{Z}_2$, ou ainda, $\mathbf{Z}^T_{1\,(8 \times 1280)} \times \mathbf{Z}_{2\,(1280 \times 4)} = \mathbf{N}_{(8 \times 4)}$.

Como já foi observado, não existe qualquer diferença entre o gráfico das coordenadas das linhas e colunas de **N** e o gráfico das colunas de **Z**, desprezando-se todas as dimensões de **Z** com inércias menores ou iguais a 1/2. Contudo, existe uma diferença substancial nas inércias principais, que afetarão o gráfico nas coordenadas principais. Primeiro, as porcentagens das inércias na análise de **Z** são muito menores e, segundo, como já foi dito, a diferença entre os seus valores é menos drástica, o que indica que os perfis-colunas de **Z** estão dispersos mais "esfericamente" do que os perfis-linha e coluna de **N**. Como as inércias principais mais interessantes da análise de **Z** são maiores do que 1/2, as porcentagens deveriam ser calculadas sobre as quantidades α_k - 1/2, com k = 1,...,J_2 - 1, que, nesse exemplo, são as porcentagens referentes aos valores da raiz quadrada das inércias principais da análise de **N**.

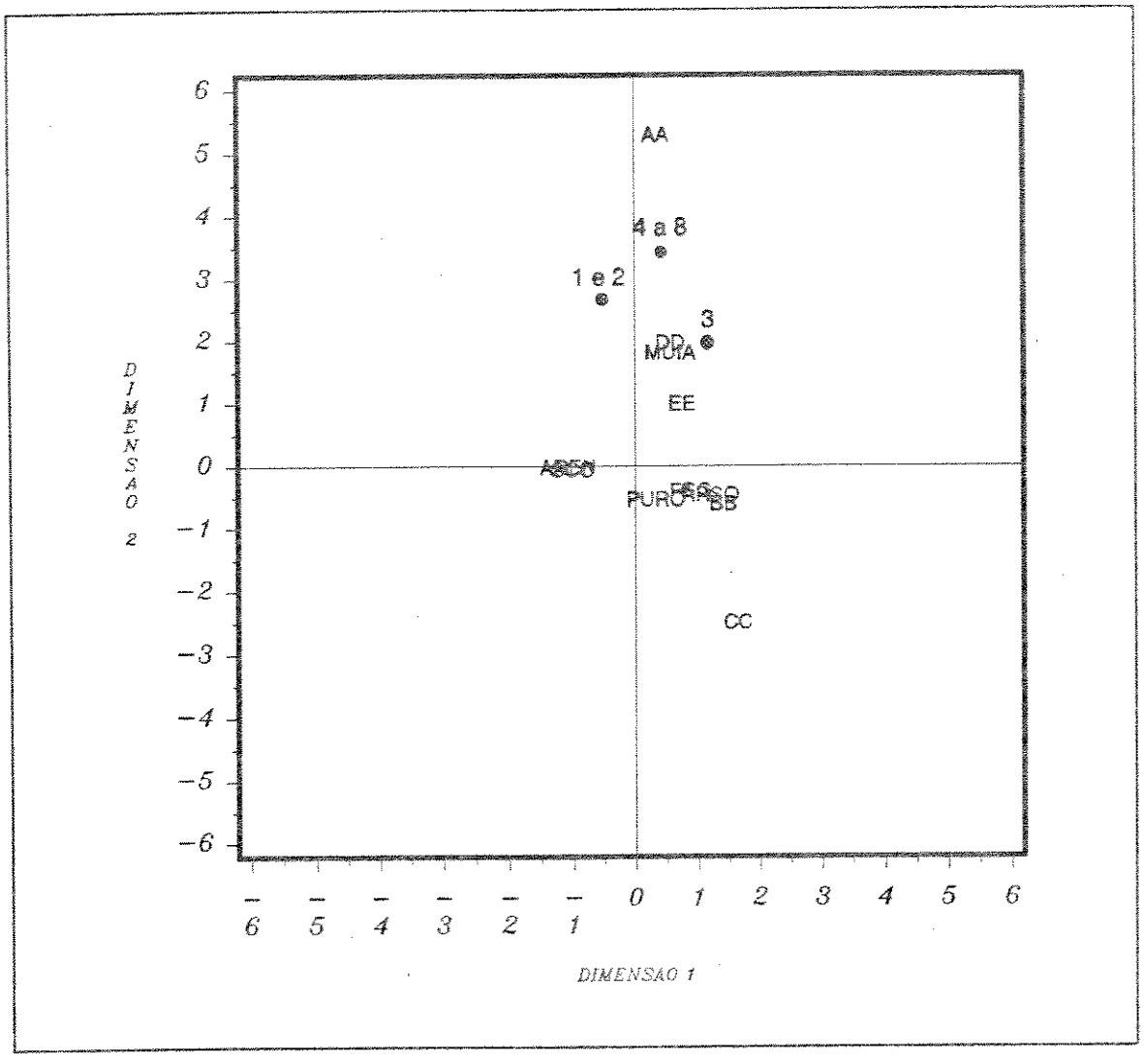

Figura 3.10: Análise de correspondência múltipla das variáveis tamanho de conchas e matriz terrígena. Dimensão 1 (16,77%), dimensão 2 (11,97%).

3.4.2 Geometria das Linhas de Z

Os perfis-linha são, agora, de forma muito especial. Cada linha de Z consiste de zeros, exceto para as duas categorias nas quais eles caem, de modo que cada perfil-linha tem valores de 1/2 nessas mesmas posições e zero nas demais. Nessa situação, a relação de baricentro indica que cada ponto-linha cairá a meio caminho entre os pontos de suas respectivas categorias.

3.4.3 Tabela de Burt

É interessante comparar a análise de $\mathbf{Z}$ com a matriz simétrica $\mathbf{Z}^T\mathbf{Z}$, (JxJ), com $J = J_1 + J_2$ a qual é denominada de "Tabela de Burt", em homenagem a Burt (1950). A tabela 3.8 mostra a tabela de Burt, cuja estrutura é a seguinte: os elementos fora da diagonal principal correspondem à tabela de contingência ($\mathbf{N}^T$ na posição inferior e $\mathbf{N}$ na posição superior), que condensa a associação entre as variáveis para todas as amostras I. Os elementos da diagonal principal da tabela de Burt definem também duas submatrizes: a diagonal dessas submatrizes correspondem ao somatório das linhas e das colunas de $\mathbf{N}$, respectivamente.

Como a matriz de Burt é simétrica positiva definida, é claro que sua análise de correspondência produz dois conjuntos idênticos de coordenadas para as linhas e colunas. De novo, a única diferença está nos valores das inércias principais, que afetarão as escalas das coordenadas principais. Com relação a isso, também pode ser mostrado que as inércias principais na análise da matriz de Burt, α^B são as raízes quadradas das inércias da matriz indicadora $\mathbf{Z}$, α^Z.

Segundo Greenacre e Hastie (1987), a geometria da matriz indicadora na análise de correspondência múltipla é reconhecidamente menos convincente e, conseqüentemente com menor apelo visual.

Tabela 3.8: Tabela de Burt - Análise de correspondência múltipla.

	AA	BB	CC	DD	EE	FF	GG	SCO	AREN	ARSO	MULA	PURO	Σ
AA	8	0	0	0	0	0	0	0	2	1	5	0	16
BB	0	6	0	0	0	0	0	0	0	4	1	1	12
CC	0	0	10	0	0	0	0	0	0	9	0	1	20
DD	0	0	0	76	0	0	0	0	14	20	27	15	152
EE	0	0	0	0	108	0	0	0	16	40	30	22	216
FF	0	0	0	0	0	265	0	0	37	101	43	84	530
GG	0	0	0	0	0	0	229	0	10	92	42	85	458
SCO	0	0	0	0	0	0	0	578	426	6	42	104	1156
AREN	2	0	0	14	16	37	10	426	505	0	0	0	1010
ARSO	1	4	9	20	40	101	92	6	0	273	0	0	546
MULA	5	1	0	27	30	43	42	42	0	0	190	0	380
PURO	0	1	1	15	22	84	85	104	0	0	0	312	624
Σ	16	12	20	152	216	530	458	1156	1010	546	380	624	5120

4 - APLICAÇAO DA ANÁLISE DE CORRESPONDÊNCIA

4.1 OS DADOS

Como exemplo de aplicação da análise de correspondência, escolheu-se um conjunto de dados referente à Formação Lagoa Feia na Bacia de Campos. Esse conjunto de dados consiste na descrição macroscópica de testemunhos da seqüência deposicional denominada "Coquinas". Foi coletado na fase inicial do projeto **Análise Regional da Seqüência das Coquinas. Formação Lagoa Feia - Bacia de Campos,** realizado pelo CENPES/DEPEX e gentilmente cedido pela PETROBRÁS. O objetivo do referido projeto é detalhar o modelo deposicional das coquinas, que constituem o principal reservatório de hidrocarbonetos da formação, para localizar novos objetivos para a exploração (Carvalho et alii, 1992 - no prelo).

A Bacia de Campos situa-se na costa leste brasileira, com aproximadamente 98% de sua área submersa na Plataforma Continental do Estado do Rio de Janeiro. Cobre uma área de 31.000 quilômetros quadrados até a cota batimétrica de 300 metros, a leste do meridiano de 42 graus W. A bacia é limitada ao sul pelo Arco de Cabo Frio e ao norte pelo Arco de Vitória (figura 4.1).

Os sedimentos da Formação Lagoa Feia foram depositados durante os estágios *rift valley* e proto-oceânico da Bacia de Campos, nos andares Buracica, Jiquiá e Alagoas. Encontram-se sobrepostos discordantemente sobre rochas basálticas extrusivas da "Seqüência Vulcânica Sedimentar" (Formação Cabiúnas). Consistem em conglomerados e arenitos líticos de planície aluvial, folhelhos lacustrinos, siltitos e carbonatos flúvio-lacustrinos (Carvalho et alii, 1984), recobertos por camadas de halita e anidrita; estas últimas equivalem às primeiras incursões marinhas na bacia. Atribui-se aos sedimentos da Formação Lagoa Feia uma origem não-marinha, uma vez que, até o momento, não foram encontrados fósseis marinhos.

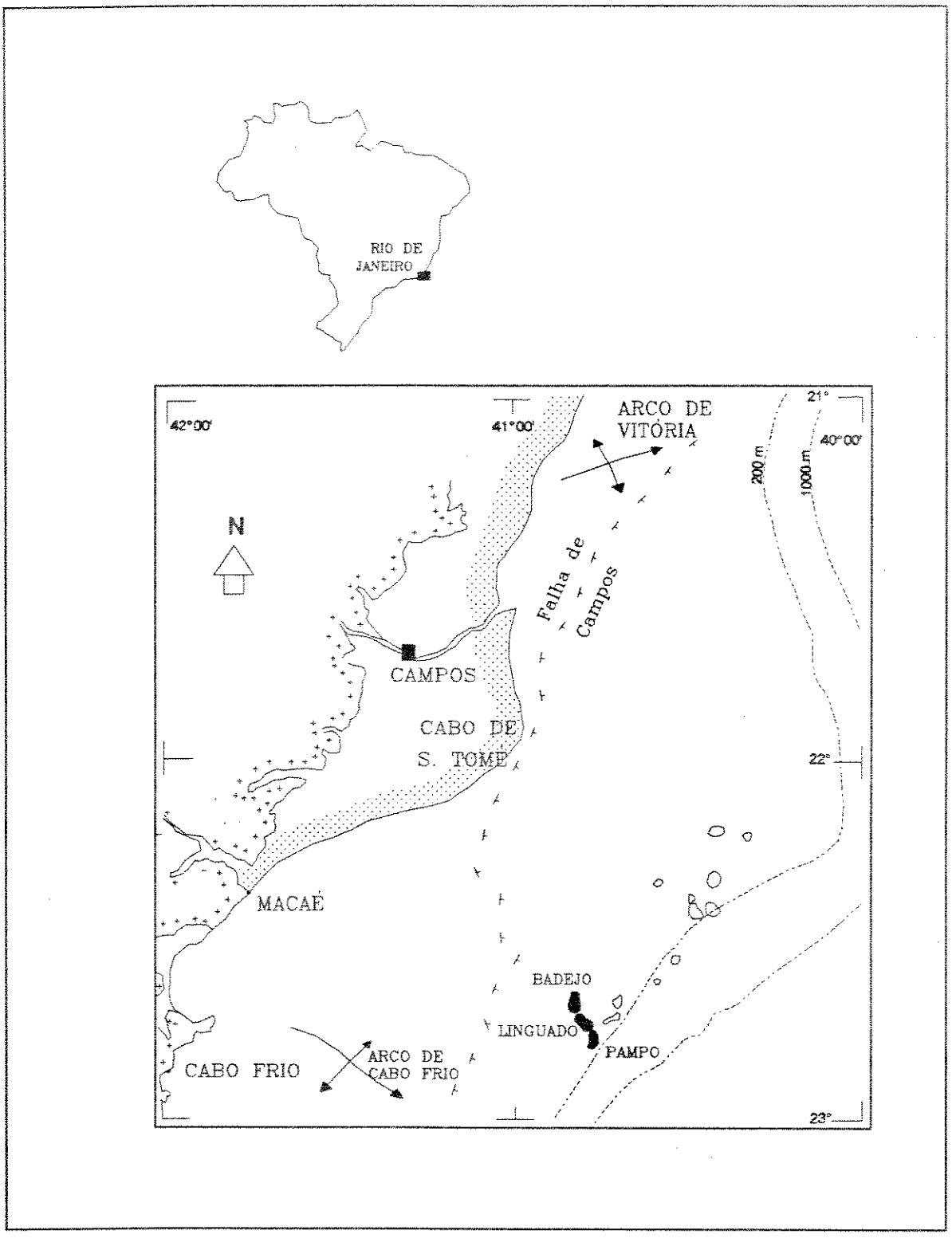

Figura 4.1: Localização da Bacia de Campos e dos campos produtores na Formação Lagoa Feia.

Dentre os inúmeros trabalhos realizados na Formação Lagoa Feia, destacam-se os de Castro e Azambuja (1981), Carvalho et alii (1984), Bertani e Carozzi (1984), Dias et alii (1987) e Carvalho et alii (1992 - no prelo).

A ocorrência da Formação Lagoa Feia está limitada a oeste pela falha-charneira de Campos, apresentando continuidade tanto para a Bacia de Santos quanto para a Bacia do Espírito Santo. Para leste avança em direção a águas profundas até à Província de Domos Salinos identificada em linhas sísmicas (Szatmari et alii, 1983 e Lobo & Ferradaes, 1983 apud Dias et alii, 1987).

A Formação Lagoa Feia é dividida em quatro seqüências deposicionais (Dias et alii, op. cit.): 1) "Sequência Clástica Basal", 2) "Sequência Talco-estevensítica", 3) "Sequência das Coquinas" e 4) "Sequência Clasto-evaporítica". A seqüência das coquinas é a mais importante da formação pois nela se encontram as rochas-reservatórios e o principal gerador de petróleo da Bacia de Campos (figura 4.2).

A principal característica desta seqüência é a ocorrência de expressiva deposição carbonática com ampla distribuição na bacia. Carvalho et alii (1984) definiram várias fácies sedimentares nestes carbonatos: calcilutitos; calcarenitos bioclásticos, peloidais e oolíticos; calcirruditos de pelecípodes, resultado do retrabalhamento dos bioacumulados de pelecípodes, originalmente depositados *in situ*. Nas porções mais distais (lacustre profundo), encontram-se margas e principalmente folhelhos escuros, carbonosos, com alto conteúdo de matéria orgânica, que são responsáveis pela geração de petróleo na bacia. A figura 4.3 mostra o modelo deposicional esquemático da seqüência das coquinas.

São individualizados dois corpos de coquinas no andar Jiquiá Superior: Coquina superior e Coquina inferior. As coquinas são rochas constituídas de conchas retrabalhadas de pelecípodes e ostracodes formando espessas camadas. Geralmente constituem seqüências cíclicas, nas quais os sedimentos grosseiros (calcirruditos) sobrepõem-se aos mais finos (calcarenitos e calcilutitos). Carapaças de gastrópodes são os principais constituintes no corpo de coquinas superior. A porosidade é secundária, predominantemente vugular e interpartícula. Subordinadamente ocorrem porosidades móldica e intercristalina.

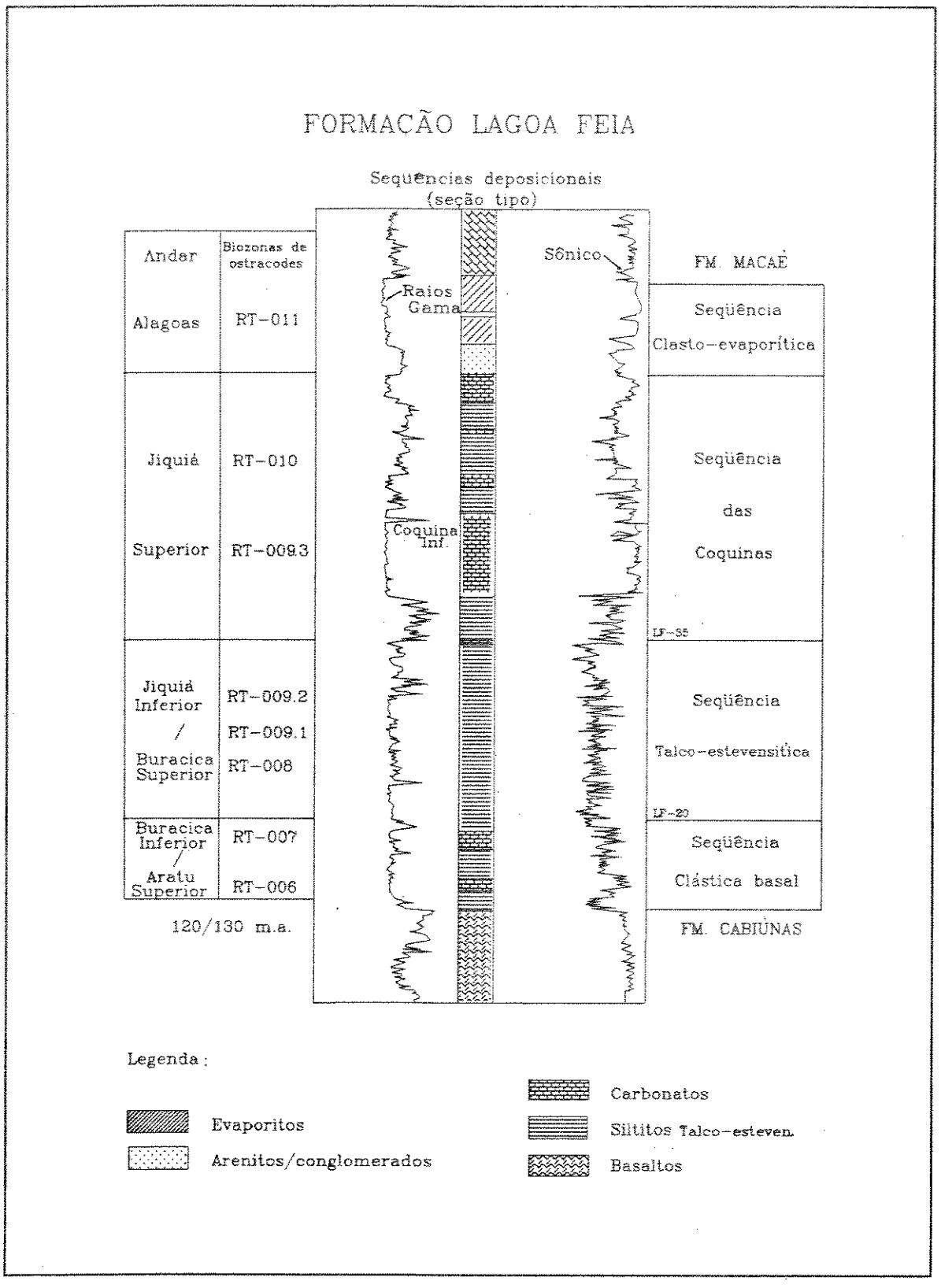

Figura 4.2: Seqüências deposicionais da Formação Lagoa Feia, seção tipo.

(Adaptado de Dias et alii, 1987)

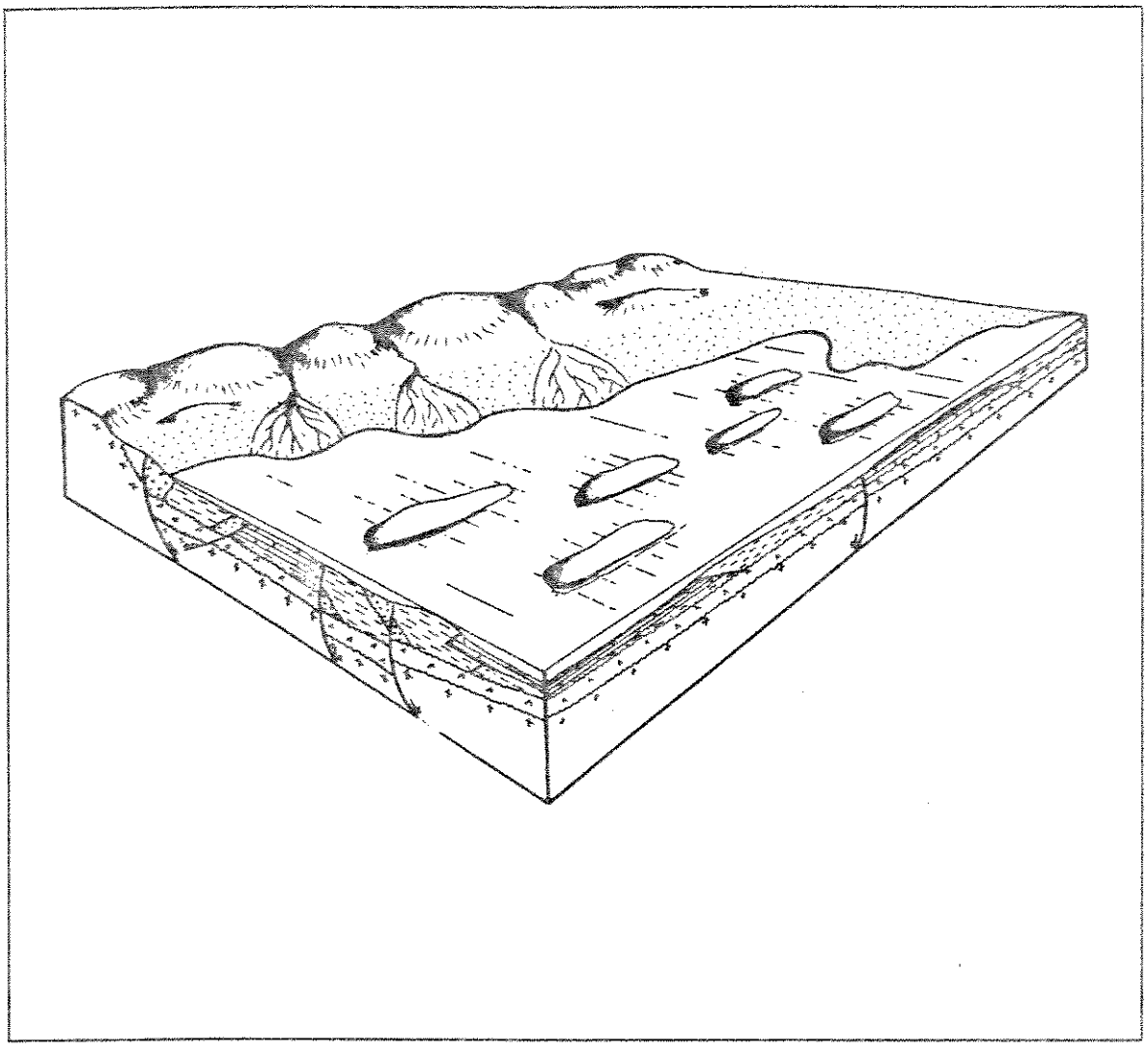

Figura 4.3: Modelo deposicional esquemático da seqüência das coquinas (modificado de Dias et alii, 1987).

O conjunto de dados no qual se aplicou a análise de correspondência consiste na descrição macroscópica dos testemunhos da Formação Lagoa Feia. Esta descrição enfatizou os aspectos texturais, com o objetivo de definir fácies sedimentares; ênfase é dada ao tipo de ocorrência das conchas, no que diz respeito principalmente à sua forma (se estão abertas ou fechadas, inteiras ou quebradas), à sua espessura e tamanho (Carvalho et alii, 1992 - no prelo). A figura 4.4 mostra a distribuição dos poços amostrados na bacia.

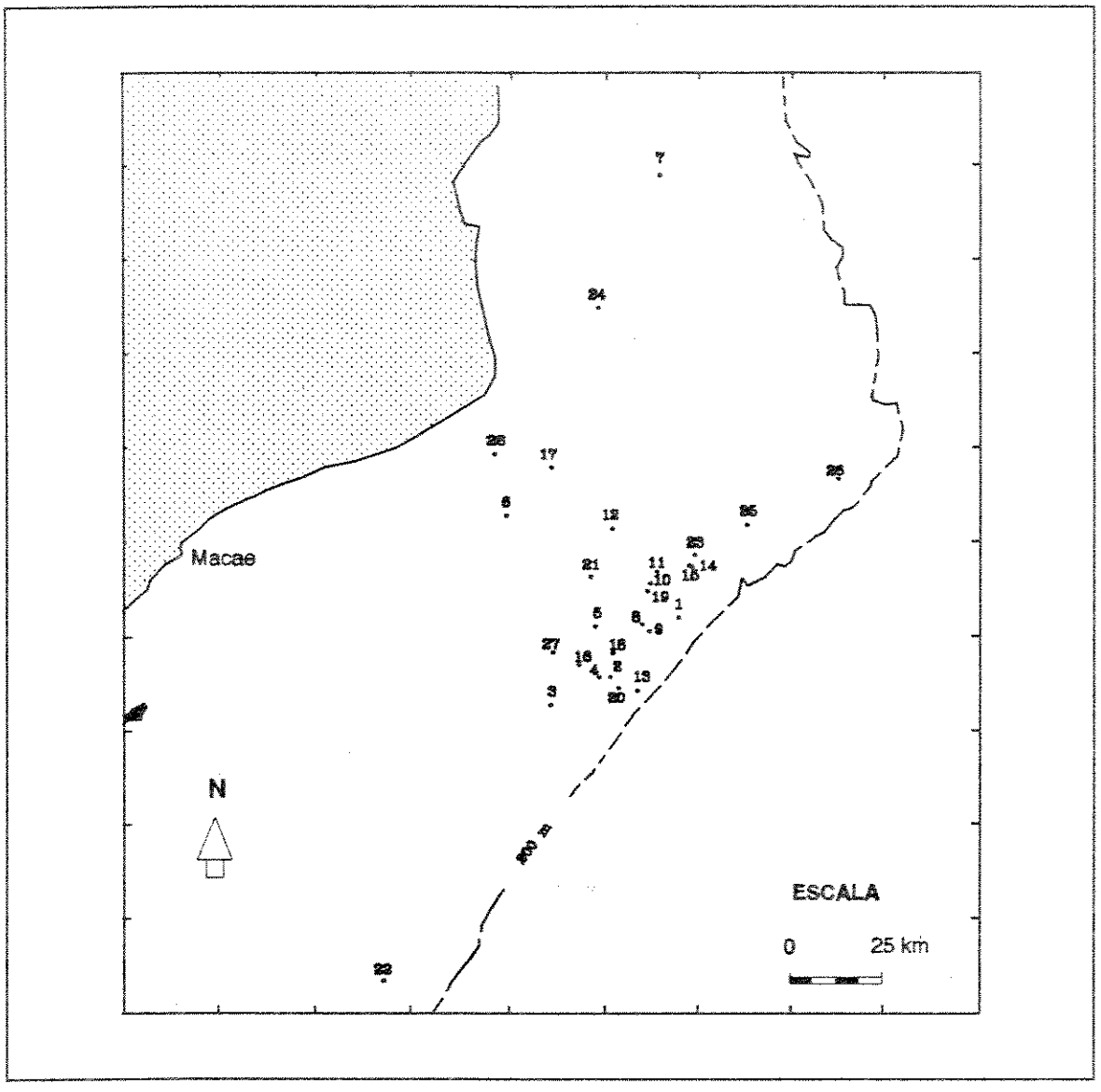

Figura 4.4: Localização dos poços amostrados na Bacia de Campos

A matriz original de dados pode ser vista parcialmente na tabela 4.1, sendo formada por 1415 linhas e 25 colunas. As linhas referem-se aos poços onde foram medidas as variáveis: profundidade, espessura, granulometria, quantidade de matriz e quantidade de conchas. A variável quantidade de matriz é posteriormente categorizada em matriz terrígena e matriz carbonática. A variável quantidade de conchas é categorizada em forma, tamanho e espessura. Quanto à forma, podem ser abertas, fechadas, inteiras e quebradas. Quanto ao tamanho e à espessura, foram categorizadas segundo critérios que serão vistos adiante.

Tabela 4.1: Parte da matriz original de dados da Formação Lagoa Feia.

POÇO	PROF	ESP	GRAN	MAT	CON	MATRIZ		CONCHAS																	
						MCA	MTE	FORMA				TAMANHO							ESPESSURA						
								ABE	FEC	INT	QUE	AA	BB	CC	DD	EE	FF	GG	A	B	C	D	E	F	G
1	4299,00	30	5	50	50	95	5	100	0	0	100	0	0	0	0	0	5	95	0	0	0	0	0	0	100
1	4299,30	5	8	60	40	100	0	100	0	10	90	0	0	0	0	30	20	50	0	0	0	0	0	0	100
1	4299,35	65	5	50	50	95	5	100	0	0	100	0	0	0	0	5	5	90	0	0	0	0	0	0	100
1	4300,00	70	5	50	50	95	5	95	5	5	95	0	0	0	0	5	10	85	0	0	0	0	0	10	90
1	4300,70	10	0	100	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	4300,80	50	5	50	50	95	5	100	0	0	100	0	0	0	0	5	5	90	0	0	0	0	0	0	100
1	4301,30	10	5	100	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	4301,40	45	5	50	50	95	5	100	0	0	100	0	0	0	0	5	5	90	0	0	0	0	0	0	100
1	4301,85	5	8	40	60	100	0	100	0	0	100	0	0	0	5	30	30	35	0	0	0	0	0	10	90
1	4301,90	25	5	50	50	95	5	100	0	0	100	0	0	0	5	5	10	80	0	0	0	0	0	0	100
1	4302,15	5	8	40	60	100	0	100	0	0	100	0	0	0	5	30	30	35	0	0	0	0	0	10	90
1	4302,20	30	5	50	50	95	5	100	0	0	100	0	0	0	0	5	5	90	0	0	0	0	0	0	100
1	4302,50	20	8	40	60	100	0	100	0	0	100	0	0	0	5	30	30	35	0	0	0	0	0	10	90
1	4302,70	130	5	50	50	95	5	100	0	0	100	0	0	0	5	5	10	80	0	0	0	0	0	0	100
1	4304,00	20	6	60	40	100	0	100	0	0	100	0	0	0	0	5	10	85	0	0	0	0	0	0	100
1	4304,20	30	3	70	30	90	10	100	0	0	100	0	0	0	0	0	5	95	0	0	0	0	0	0	100
1	4304,50	30	6	50	50	100	0	100	0	0	100	0	0	0	5	0	15	80	0	0	0	0	0	0	100
1	4304,80	5	7	50	50	95	5	90	10	20	80	0	0	0	0	5	30	65	0	0	0	0	0	0	100
1	4304,85	15	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
1	4305,00	10	6	50	50	80	20	95	5	5	95	0	0	0	0	0	10	90	0	0	0	0	0	0	100
2	3540,00	20	0	90	10	100	0	80	20	50	50	0	0	20	10	60	10	0	0	0	0	0	0	30	0	70
2	3540,20	40	0	100	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Cada uma destas variáveis foi medida em um intervalo de espessura considerado homogêneo, portanto cada uma delas reflete a média para o intervalo. Foram também coletados alguns dados de petrofísica (porosidade e permeabilidade), que não foram incluídos neste trabalho, por serem medidas muito localizadas e ausentes na maioria das amostras. A análise de correspondência não considera valores ausentes para uma determinada variável, eliminando toda a linha que representa a amostra se alguma variável está ausente.

Para aplicar a análise de correspondência neste conjunto de dados, foi necessário, inicialmente, transformar a matriz de dados originais (constituída de valores numéricos) em uma matriz composta de valores discretos. Os critérios geológicos para essa discretização foram definidos por Carvalho et alii, 1992 (no prelo), e podem também ser vistos no programa P.1.

Dessa forma, os valores de granulometria, que já estavam discretizados entre 0 e 8, foram transformados nos caracteres correspondentes conforme a tabela 4.2.

Tabela 4.2: Categorias da variável granulometria

	Caracter	Diâmetro (mm)	Classe granulométrica
0	ARG	< 0,004	argila
1	SLT	0,004 - 0,062	silte
2	MFN	0,062 - 0,125	areia muito fina
3	FNO	0,125 - 0,250	areia fina
4	MED	0,250 - 0,500	areia média
5	GRO	0,500 - 1,000	areia grossa
6	MGO	1,000 - 2,000	areia muito grossa
7	GNL	2,000 - 4,000	grânulo
8	SXO	>4,000	seixo

Os valores referentes à quantidade de conchas (ou o seu complemento - matriz -) foram arranjados de forma que permitissem identificar três categorias indicadoras da energia deposicional do ambiente: *Grainstones/Packstones*, *Wackestones* e *Mudstones* (tabela 4.3). Cabe observar que, na descrição macroscópica dos testemunhos, não se separou cimento de matriz, pois, muitas vezes, eles apresentavam-se bastante recristalizados, dificultando a identificação dos constituintes originais. Desta forma, a variável MCA (matriz carbonática) engloba tanto o que poderia ser matriz quanto cimento carbonático. Portanto, na interpretação desta variável, isto deve ser considerado.

Tabela 4.3: Categorias da matriz carbonática.

	Caracter	Intervalo
Sem conchas	SCO	con = 0%
Mudstones	MST	0 < con < 10%
Wackestones	WST	10 ≤ con < 30%
Grainstones/Packstones	G/P	con ≥ 30%

Os valores da variável quantidade de matriz terrígena foram arranjados de modo a distinguir terrígenos de carbonatos e os vários graus de contaminação detrítica dentro desses carbonatos - tabela 4.4. Grande parte dos testemunhos consiste somente de rochas terrígenas, mas foram também incluídos na análise de correspondência, uma vez que fazem parte da matriz de dados.

Tabela 4.4: Categorias de matriz terrígena.

	Caracter	Intervalo
Carbonato Puro	PURO	mte = 0%
Carbonato Arenoso	ARSO	0 < mte ≤ 10%
Carbonato Muito Arenoso	MUIA	10 < mte ≤ 50%
Terrígenos	AREN	mte > 50%

A tabela 4.5 mostra a discretização da variável forma de conchas. As tabelas 4.6 e 4.7 mostram os critérios de discretização da variável tamanho de conchas. A primeira mostra os intervalos de tamanho de conchas que foram utilizados no exemplo do capítulo anterior, e refere-se somente à leitura da matriz original, considerando 10 % como valor mínimo. Já a segunda mostra os critérios utilizados na definição dos tipos de rocha com base nesta variável. As tabelas 4.8 e 4.9 mostram os critérios definidos para a variável espessura de conchas e empacotamento da rocha. Observe-se que só foram detalhados os tipos "calcirrudíticos", não tendo sido definidos, neste trabalho, critérios para os calcarenitos, calcilutitos ou mesmo para os terrígenos.

Tabela 4.5: Categorias da forma das conchas.

	Caracter	Intervalo
Sem conchas	SCO	abe = 0%
Conchas fechadas	FEC	0 < abe < 50%
Conchas abertas	ABE	abe ≥ 50%
Sem conchas	SCO	que = 0%
Conchas inteiras	INT	0 < que < 50%
Conchas quebradas	QUE	que ≥ 50%

Tabela 4.6: Categorias do tamanho de concha.

	Caracter	Intervalo
Sem conchas	SCO	con = 0%
con ≤ 0,2 cm	GG	GG > 10%
0,2 < con ≤ 0,6 cm	FF	FF > 10%
0,6 < con ≤ 1 cm	EE	EE > 10%
1 < con ≤ 2 cm	DD	DD > 10%
2 < con ≤ 3 cm	CC	CC > 10%
3 < con ≤ 4 cm	BB	BB > 10%
con < 4 cm	AA	AA > 105

Tabela 4.7: Categorias do Tipo de Rocha.

	Calcirrudito	Calcirrudito Calcarenítico	Calcirrudito Muito Calcarenítico	Calcarenito/ Calcilutito
	C	CC	CMC	CRE
	0 ≤ GG < 20%	20 ≤ GG < 50%	50 ≤ GG < 70%	GG ≥ 70%
FF ≥ 30%	Cmf	CCmf	CMCmf	CRE
EE ≥ 30%	Cf	CCf	CMCf	CRE
DD ≥ 30%	Cm	CCm	CMCm	CRE
CC ≥ 30%	Cg	CCg	CMCg	CRE
BB ≥ 30%	Cmg	CCmg	CCmg	CRE
AA ≥ 30%	Cmg	CCmg	CMCmg	CRE

Tabela 4.8: Categorias da espessura de conchas.

	Caracter	Intervalo
Conchas muito grossas	A (con > 0,5 cm)	A ≥ 10%
Conchas grossas	B (0,4 < con ≤ 0,5 cm)	B ≥ 10%
Conchas grossas	C (0,3 < con ≤ 0,4 cm)	C ≥ 10%
Conchas médias	D (0,2 < con ≤ 0,3 cm)	D ≥ 10%
Conchas finas	E (0,1 < con ≤ 0,2 cm)	E ≥ 10%
Conchas muito finas	F (con = 0,1 cm)	F ≥ 10%
Conchas G	G (con < 0,1 cm)	G ≥ 10%

Tabela 4.9: Categorias de empacotamento.

	Caracter	Intervalo
Frouxo	FRX	30 ≤ con < 50%
Normal	NOR	50 ≤ con < 70%
Denso	DEN	con ≥ 70%

Feita a discretização da matriz de dados, procedeu-se à análise de correspondência, utilizando o *software* SAS, através do procedimento PROC CORRESP, para a análise de correspondência simples e PROC CORRESP MCA para a análise de correspondência múltipla. Este procedimento calcula as coordenadas dos eixos do plano de melhor ajuste aos dados, as coordenadas das projeções dos pontos no referido plano, a inércia dos eixos, a contribuição relativa e absoluta dos pontos, a massa dos pontos e a qualidade da representação.

No anexo AX.1, apresenta-se o arquivo de saída do PROC CORRESP entre as categorias da variável quantidade de matriz terrígena e as categorias da variável tamanho de concha, que foi apresentado no capítulo anterior. Observe-se que

este procedimento não desenha os gráficos; portanto, para isto é necessário utilizar o procedimento PROC GPLOT para plotar o arquivo de saída da análise de correspondência (veja programa P.4).

4.2 APRESENTAÇÃO DOS RESULTADOS

4.2.1 Análise de Correspondência Simples Variável x Variável

Como a análise de correspondência é mais indicada para tabelas de contingência de dupla entrada, ela foi inicialmente aplicada ao estudo das relações existentes entre as variáveis, tomadas duas a duas.

A figura 4.5 mostra a análise de correspondência simples entre as categorias da variável granulometria e as categorias da variável quantidade de matriz carbonática. Há uma forte associação dos *Grainstones/Packstones* (G/P) com as granulometrias mais grosseiras (SXO, GNL, MGO e GRO). As amostras sem conchas (SCO) associam-se fortemente com as granulometrias mais finas (ARG, SLT, MFN), enquanto os *Wackestones* (WST) e os *Mudstones* (MST) associam-se às granulometrias MED e FNO. Esse gráfico é a representação quase exata dos pontos, uma vez que 99,02% da inércia é explicada pelos dois primeiros eixos.

Os pontos de maior massa ou que estão mais distantes são os que mais contribuem para a inércia; para a inércia do primeiro eixo, estes pontos são SCO, ARG e, G/P; para o segundo eixo, GNL, MFN e MST. Desta representação conclui-se que as amostras sem conchas, provavelmente os sedimentos terrígenos, têm granulometria mais fina, certamente correspondendo a folhelhos, margas e siltitos. Os sedimentos carbonáticos G/P apresentam maior quantidade de conchas e têm conseqüentemente granulometrias mais grosseiras. Já os MST e WST, como têm quantidades menores de conchas, têm também granulometrias intermediárias. Logo, a presença de conchas está diretamente relacionada à granulometria: quanto maior a quantidade de conchas mais grosseiros são os sedimentos.

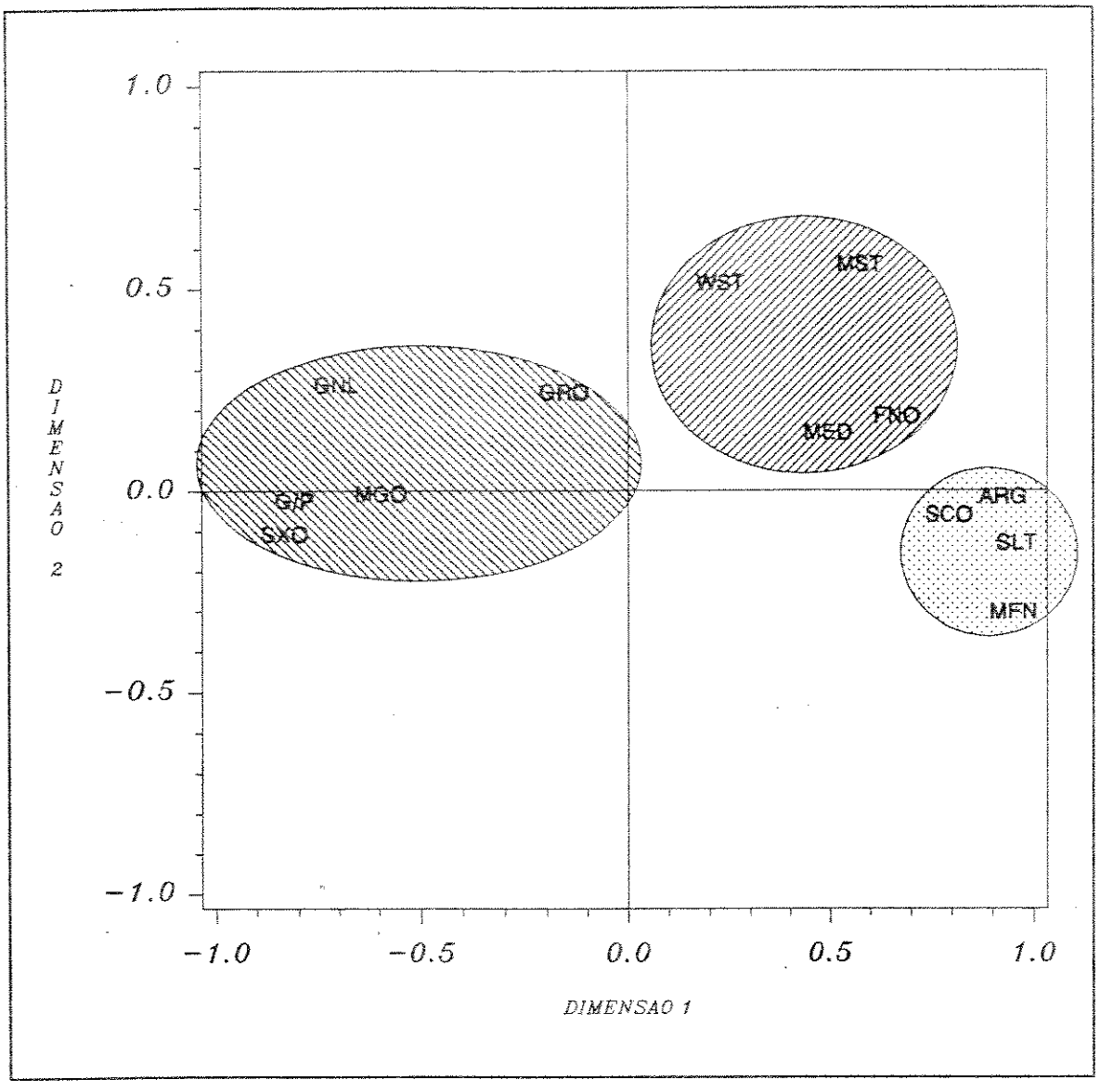

Figura 4.5: Análise de correspondência simples das variáveis granulometria e quantidade de matriz carbonática. Dimensão 1 (95,15%) e dimensão 2 (3,87%).

A figura 4.6 mostra as relações entre as categorias da variável granulometria e as categorias da variável quantidade de matriz terrígena. As amostras de terrígenos (AREN) associam-se com as classes granulométricas mais finas (SLT, MFN, FNO, MED), as amostras de carbonatos arenosos (ARSO) associam-se com granulometria GNL, SXO, GRO, e as amostras de carbonatos muito arenosos (MUIA) e de carbonatos puros (PURO) associam-se com granulometria MGO.

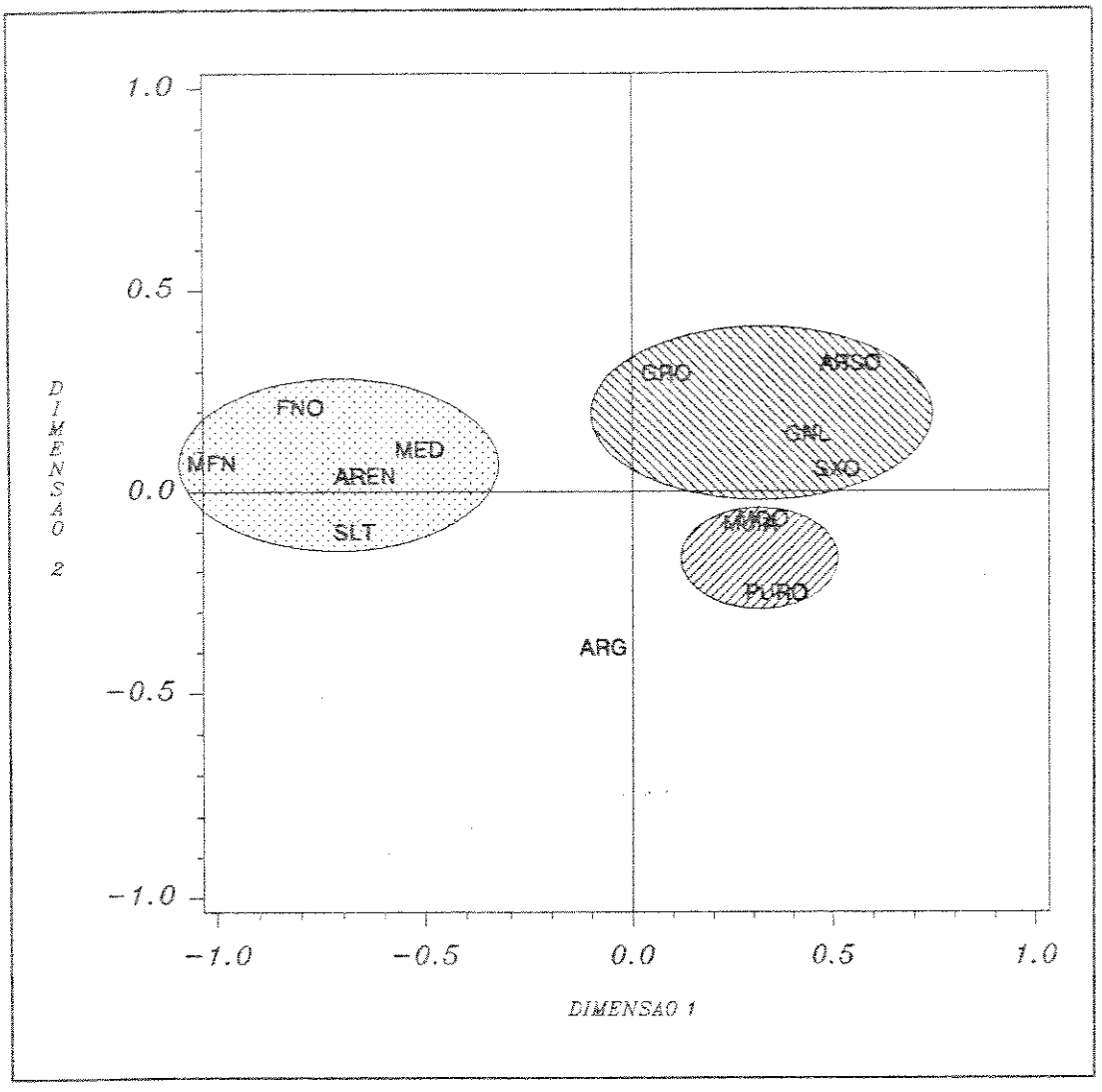

Figura 4.6: Análise de correspondência simples das variáveis granulometria e quantidade de matriz terrígena. Dimensão 1 (85,51%) e dimensão 2 (12,62%).

É interessante notar que as amostras de granulometria argila (ARG) situam-se entre os AREN e PURO, indicando que tanto os terrígenos (folhelhos) quanto os carbonatos puros (calcilutitos) têm granulometria argila. As amostras ARSO têm granulometria mais grosseira, devido à presença de conchas e até mesmo grãos de areia grosseira. É comum encontrarem-se conchas entre terrígenos nas áreas mais proximais. Esta representação bidimensional explica 98,13% da inércia total dos dados.

A representação gráfica da análise de correspondência entre a granulometria e a forma das conchas (ABE, FEC) considera 100% da inércia dos

pontos (figura 4.7). Observa-se um agrupamento de pontos à esquerda, que corresponde a uma íntima associação entre conchas ABE e granulometrias GNL, SXO, MGO e GRO, e um outro agrupamento à direita, correspondendo a uma forte relação entre amostras sem conchas SCO e as granulometrias MED, FNO, MFN, SLT, ARG.

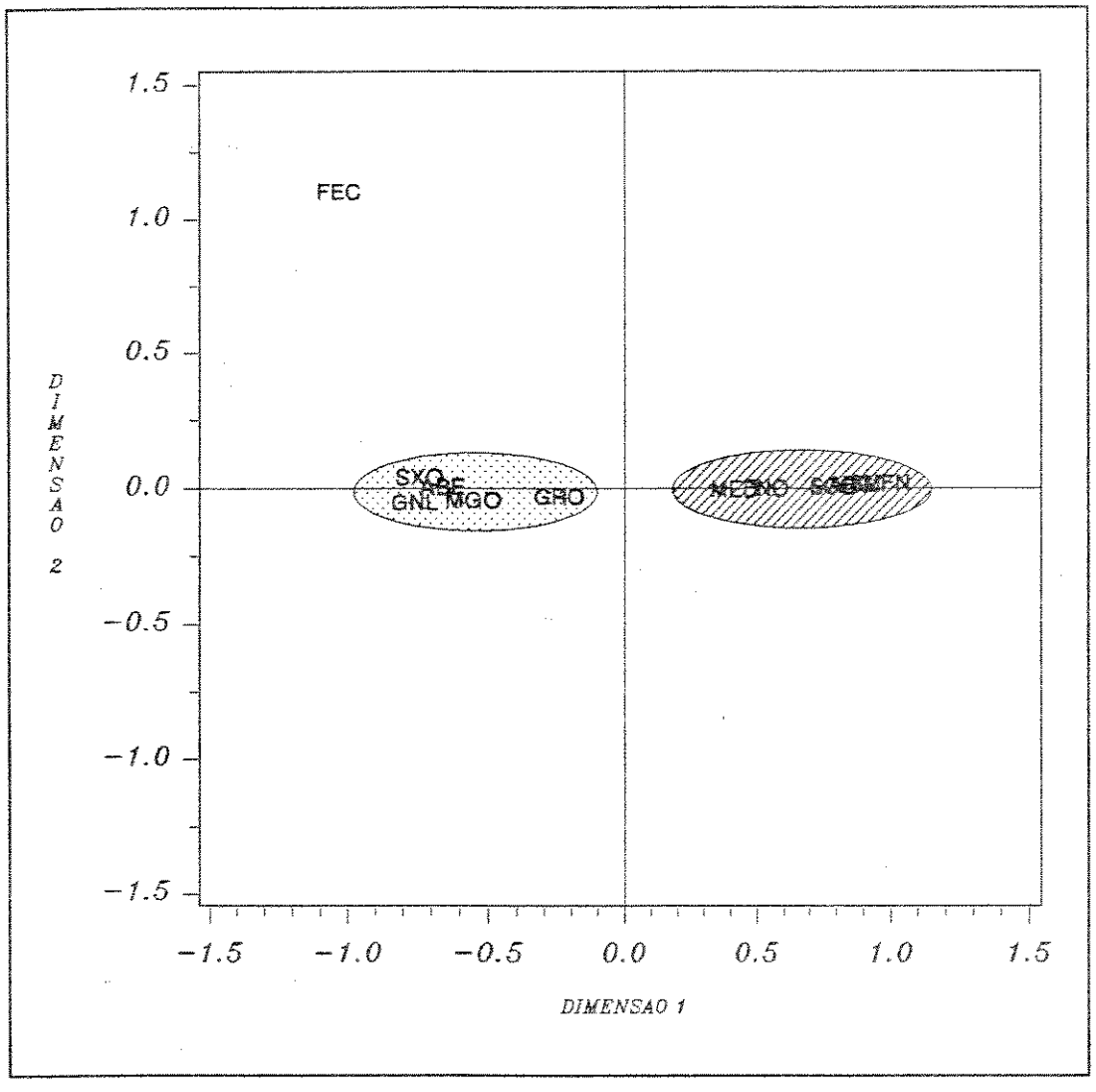

Figura 4.7: Análise de correspondência simples das variáveis granulometria e quantidade de concha aberta e fechada. Dimensão 1 (99,82%) e dimensão 2 (0,18%).

As amostras com conchas fechadas, FEC, não se associam fortemente à classe granulométrica, mas têm uma tendência às granulometrias mais grosseiras, como era de se esperar; de fato, a porcentagem de conchas fechadas é muito pequena. A mesma disposição gráfica é obtida na análise entre a granulometria e a forma de

conchas categorizadas em INT e QUE. As conchas quebradas situam-se na mesma posição que as conchas abertas e as inteiras situam-se na mesma posição que as conchas fechadas. Basicamente, as conchas abertas estão quebradas e as conchas fechadas estão inteiras.

A figura 4.8 mostra a relação entre as categorias da variável granulometria e as categorias da variável tipo de rocha.

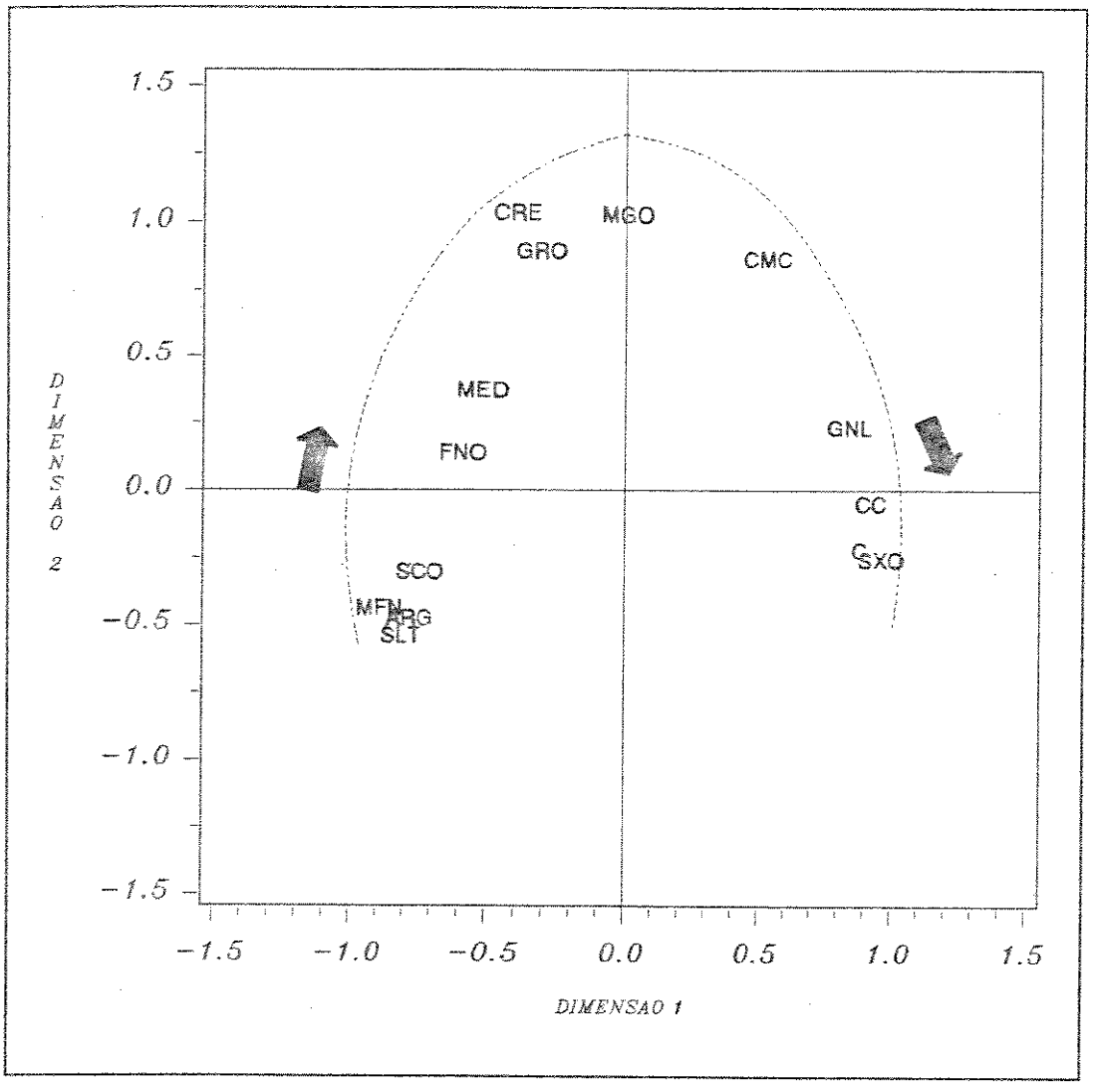

Figura 4.8: Análise de correspondência simples das variáveis granulometria e tipo de rocha. Dimensão 1 (65,63%) e dimensão 2 (27,93%).

Observa-se uma feição curva na disposição dos pontos: da direita para a esquerda, há um aumento da granulometria; no mesmo sentido, passa-se de amostras SCO, para CRE, CMC, CC e C. Resumindo, o primeiro eixo reflete um

aumento de granulometria e, conseqüentemente, um aumento no tamanho das conchas, neste sentido.

A figura 4.9 mostra a análise entre as categorias de granulometria e as categorias de espessura de conchas, considerando 97,30% da inércia total dos pontos.

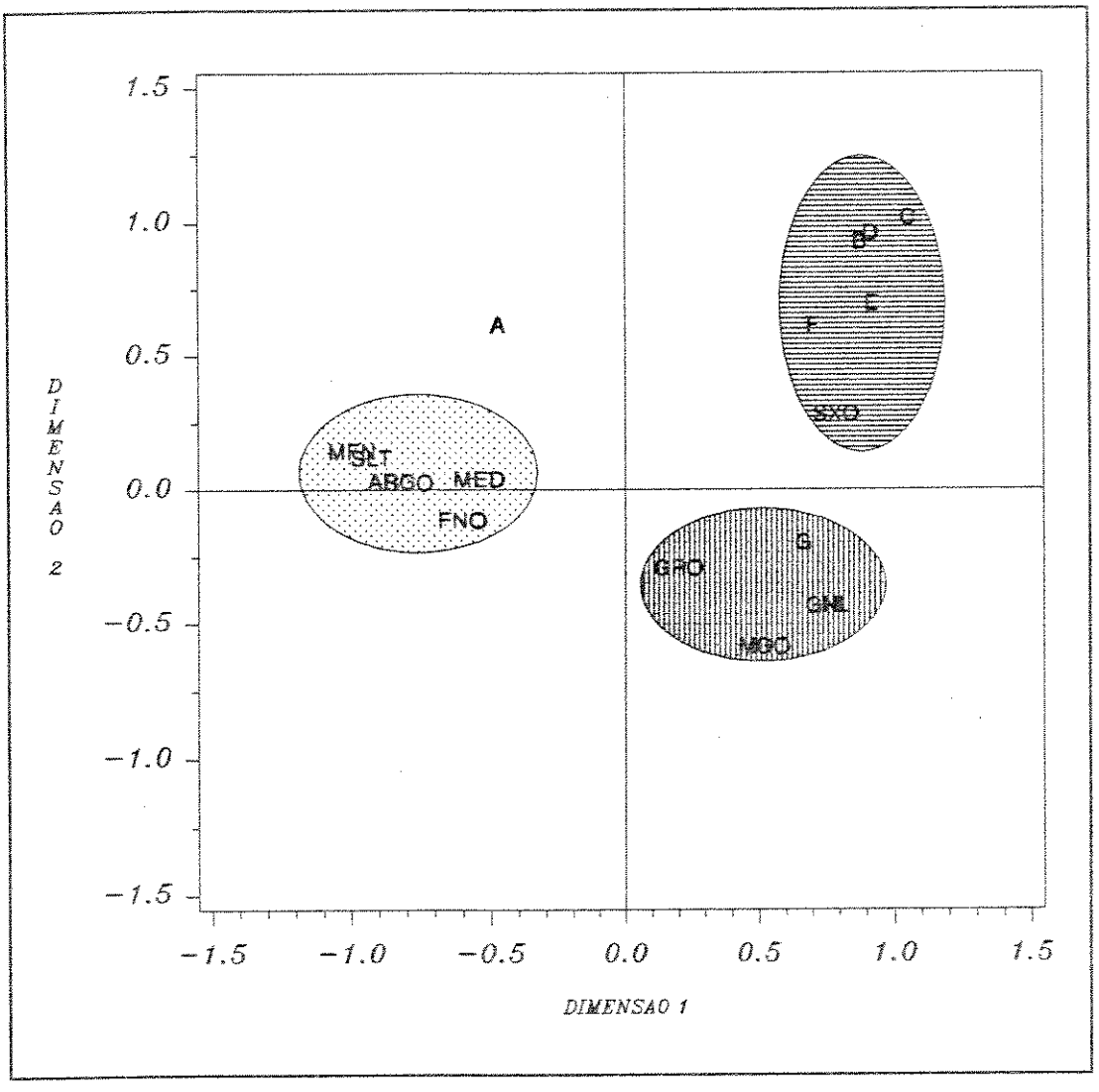

Figura 4.9: Análise de correspondência simples das variáveis granulometria e espessura de concha. Dimensão 1 (85,91%) e dimensão 2 (11,39%).

Observa-se que as amostras sem conchas (SCO) associam-se novamente às classes granulométricas ARG, SLT, MFN, FNO e MED. As amostras com as menores espessuras de conchas (G) relacionam-se com as granulometrias GNL, MGO e GRO, enquanto as amostras com espessura de conchas F, E, D, C e B associam-se

à granulometria SXO. Observa-se que o primeiro eixo mostra, da esquerda para a direita, um aumento da espessura das conchas. A exceção é a categoria A, que mostra alguma associação com as granulometrias mais finas, possivelmente porque este tipo de concha deve estar presente em calcilutitos, ou mesmo em siltitos, e associada a terrígenos.

As relações entre a granulometria e o empacotamento são mostradas na figura 4.10, que é considerada uma ótima representação dos pontos. As três categorias de empacotamento, frouxo, normal e denso, relacionam-se com as classes granulométricas SXO, GNL, MGO e GRO. As amostras SCO, de novo, associam-se a granulometrias mais finas.

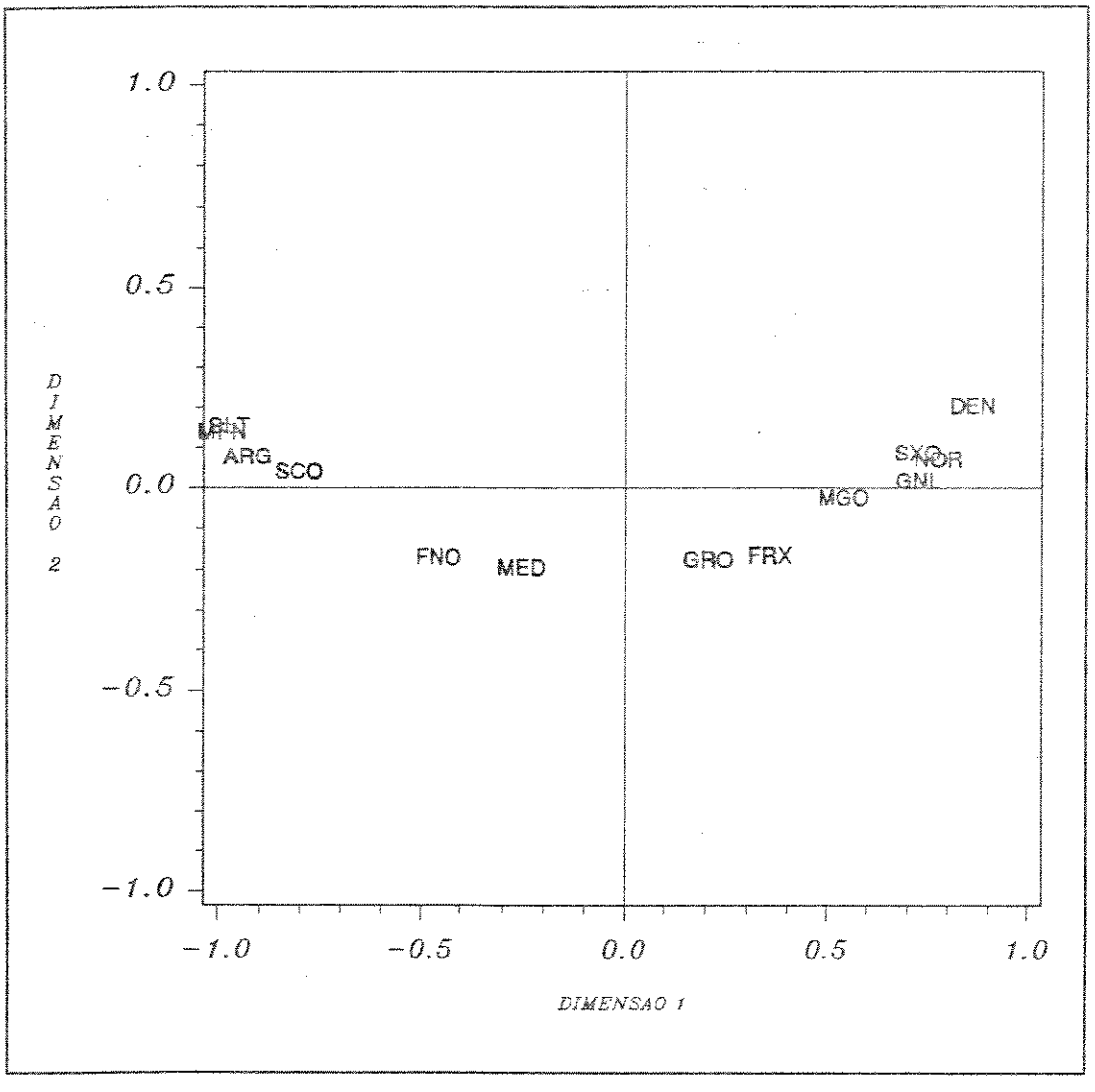

Figura 4.10: Análise de correspondência simples das variáveis granulometria e empacotamento. Dimensão 1 (96,95%) e dimensão 2 (2,90%).

A figura 4.11 mostra a relação entre a quantidade de matriz carbonática e a quantidade de matriz terrígena. Os *Grainstones/Packstones* associam-se predominantemente a carbonatos arenosos, puros e muito arenosos, no lado direito do gráfico, em oposição ao lado esquerdo, que mostra uma estreita associação entre terrígenos (AREN) e amostras sem conchas (SCO). Os *Mudstones* mostram maior associação com terrígenos, provavelmente, porque estão associados a margas e siltitos.

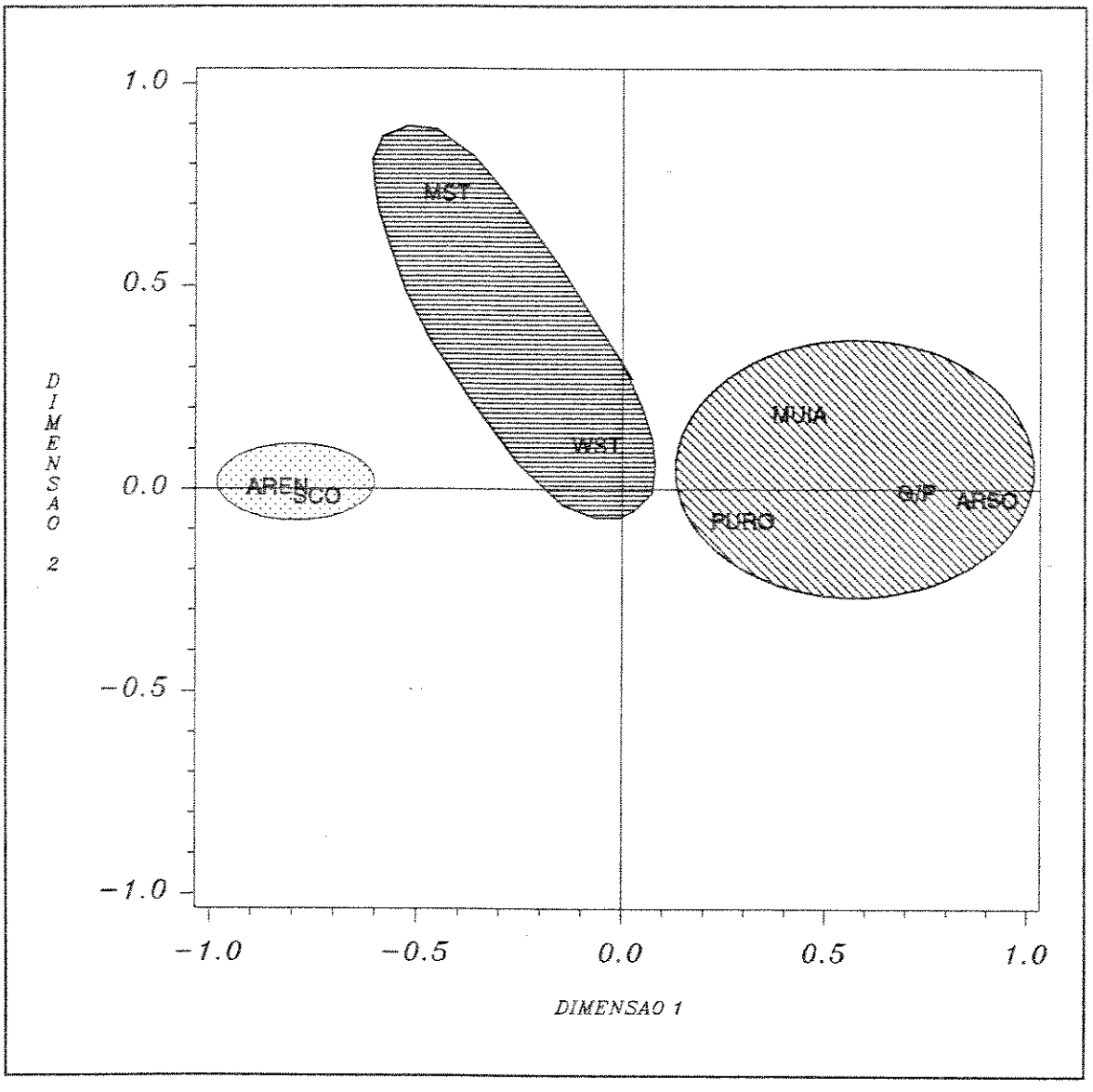

Figura 4.11: Análise de correspondência simples das variáveis quantidade de matriz carbonática e matriz terrígena. Dimensão 1 (98,36%) e dimensão 2 (1,34%).

Outro resultado interessante é a relação entre as categorias de matriz carbonática e as categorias do empacotamento. As amostras G/P têm empacotamento normal e denso, pois têm maior quantidade de conchas, enquanto as amostras MST e WST têm empacotamento FRX, pois têm menor quantidade de conchas (figura 4.12).

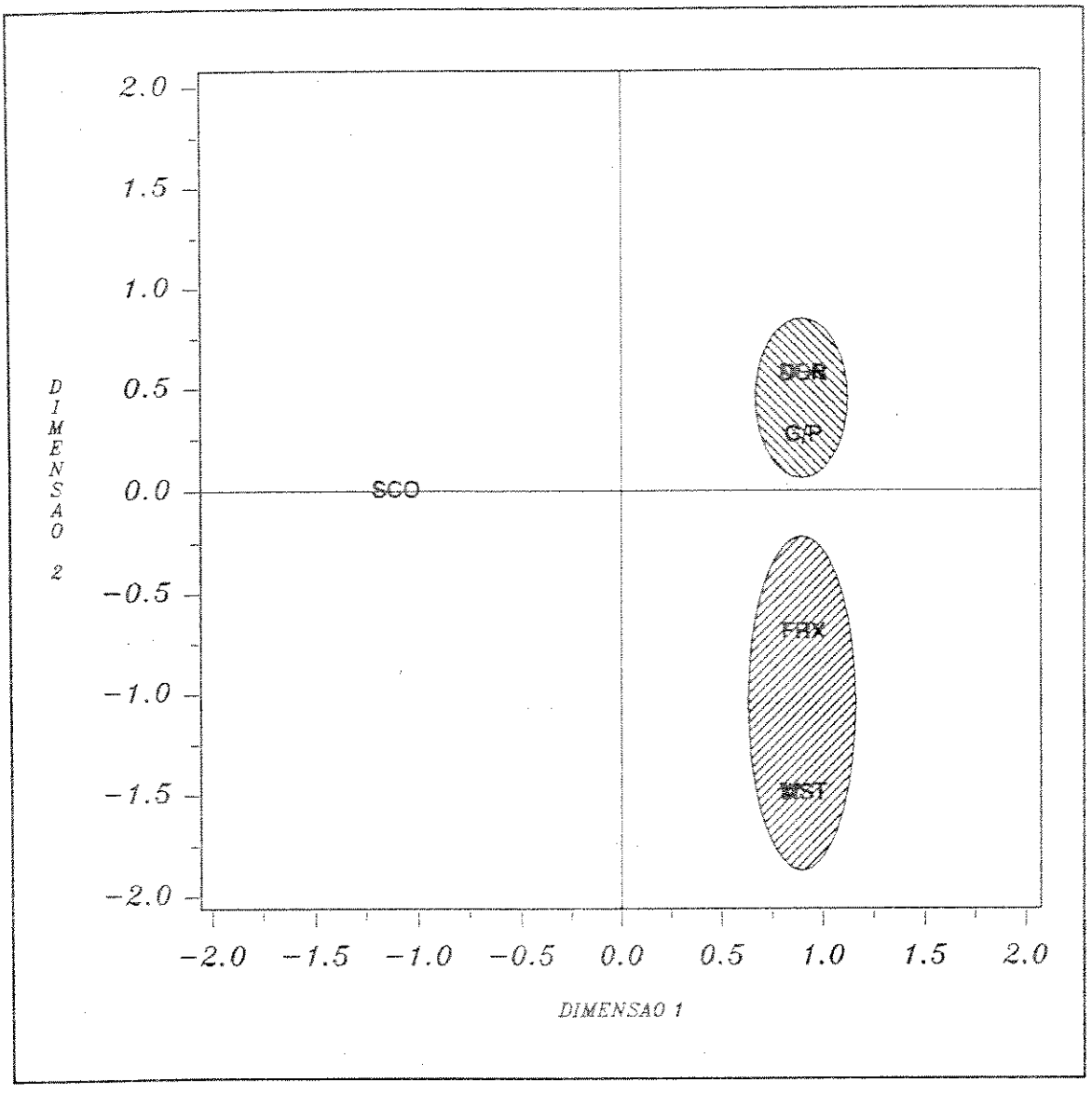

Figura 4.12: Análise de correspondência simples das variáveis quantidade de matriz carbonática e empacotamento. Dimensão 1 (82,11%) e dimensão 2 (17,89%).

A figura 4.13 mostra a relação entre as categorias da variável quantidade de matriz terrígena e o empacotamento. As amostras MUIA têm empacotamento FRX, as amostras ARSO têm empacotamento NOR e as amostras PURO têm empacotamento DEN. Conclui-se que as amostras com maior quantidade de conchas, DEN, associam-se a carbonatos mais puros.

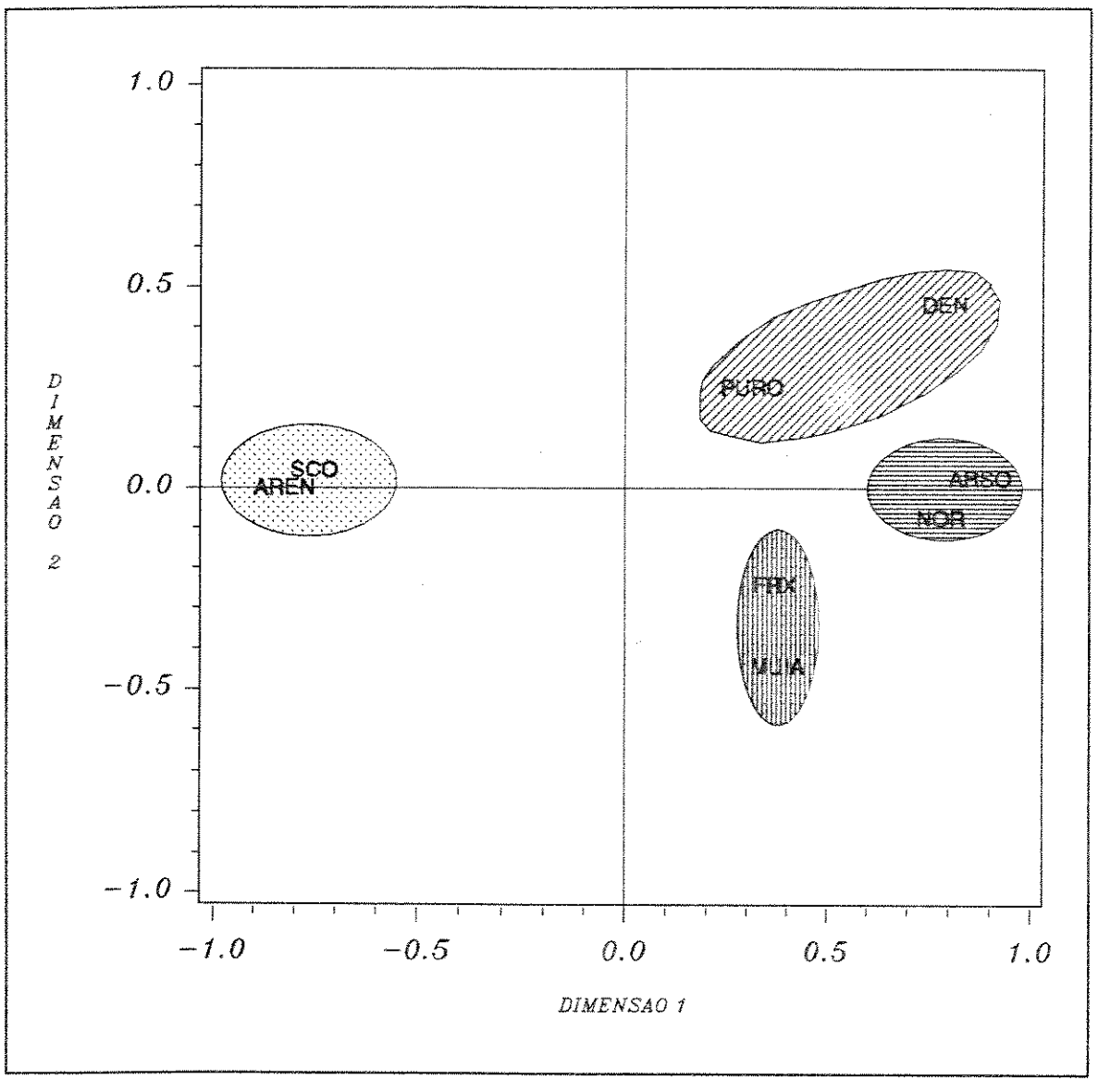

Figura 4.13: Análise de correspondência simples das variáveis quantidade de matriz terrígena e empacotamento. Dimensão 1 (91,12%) e dimensão 2 (8,68%).

Dessa forma, pode-se avaliar qualquer relação entre variáveis, duas a duas. Note-se que a representação bidimensional é sempre muito boa para os casos descritos acima.

As relações encontradas nas análises descritas são decorrentes dos critérios geológicos definidos na discretização das variáveis. Pode-se, por exemplo, através da análise de correspondência, detectar critérios erroneamente definidos, quando a interpretação das relações entre os dados analisados não tiver sentido geológico. Pode-se também simplificar determinadas categorias através de seu agrupamento; por exemplo, a espessura de conchas poderia ser discretizada em um número menor de categorias, uma vez que B, C, D, E e F não se distinguem, em termos da variável granulometria (todas elas associam-se à granulometria seixo - figura 4.9).

Desta forma, além de analisar as relações entre as variáveis, pode-se também avaliar os critérios definidos.

4.2.2 Análise de Correspondência Simples Poço x Variável

Pode-se ainda estudar as relações entre os poços e uma determinada variável, o que permite agrupar poços com características semelhantes. A figura 4.14 apresenta o resultado da análise de correspondência entre os poços e a variável tamanho de conchas. Observa-se a ocorrência de calcarenitos nos poços 1, 10 e 18, de calcirruditos muito calcareníticos nos poços 2, 4, 14, 15, 20, 22, 23, 25 e 26; de calcirruditos calcareníticos nos poços 3, 6, 9, 17, e 21, e de calcirruditos nos poços 8, 13, 16 e 27. Já os poços 5, 7, 11, 12, 19, 24 e 28 apresentam-se sem conchas.

Na figura 4.15, pode-se avaliar o efeito da ponderação da variável tamanho de conchas pela espessura que ocorre em cada poço. A ponderação resulta em uma melhor definição do agrupamento dos poços segundo essa variável. Assim, por exemplo, o poço 1 tem menor espessura de calcarenitos do que o poço 10. Os poços 4 e 17 ficam mais definidos como sem conchas e os poços 2, 3, 15 e 20 associados a calcirrudito calcarenítico. Esse tipo de resultado permite mapear a

distribuição de grupos de poços com mesmo tipo de rocha (figura 4.16). Observa-se que os poços sem conchas situam-se na borda oeste da bacia; mais a leste situam-se os carbonatos, com uma concentração de calcarenitos na parte central enquanto os calcirruditos mais grosseiros concentram-se nas bordas, ao sul da área estudada.

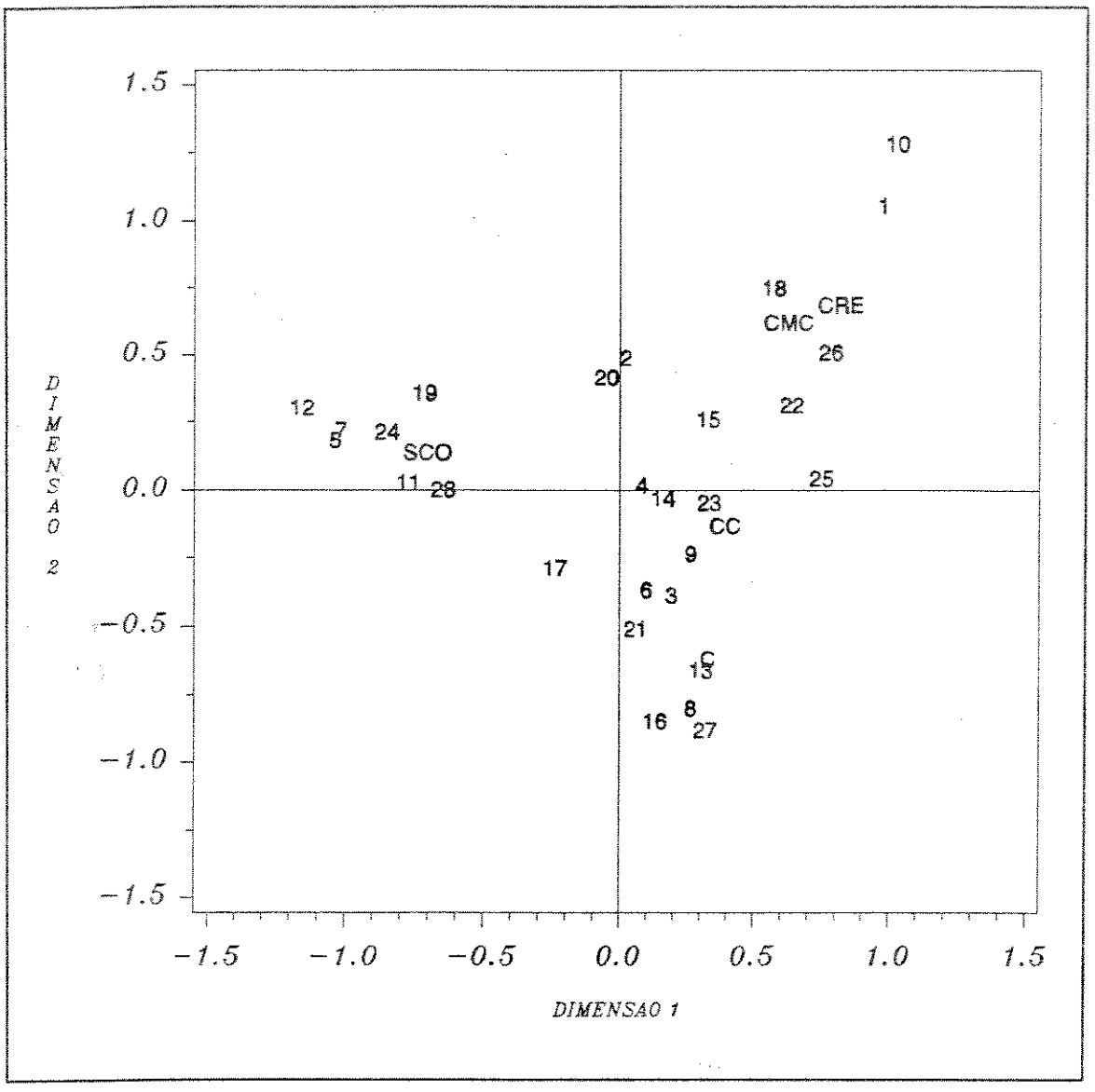

Figura 4.14: Análise de correspondência simples dos poços com a variável tamanho de conchas, sem ponderar pela espessura. Dimensão 1 (56,70%) e dimensão 2 (29,79%).

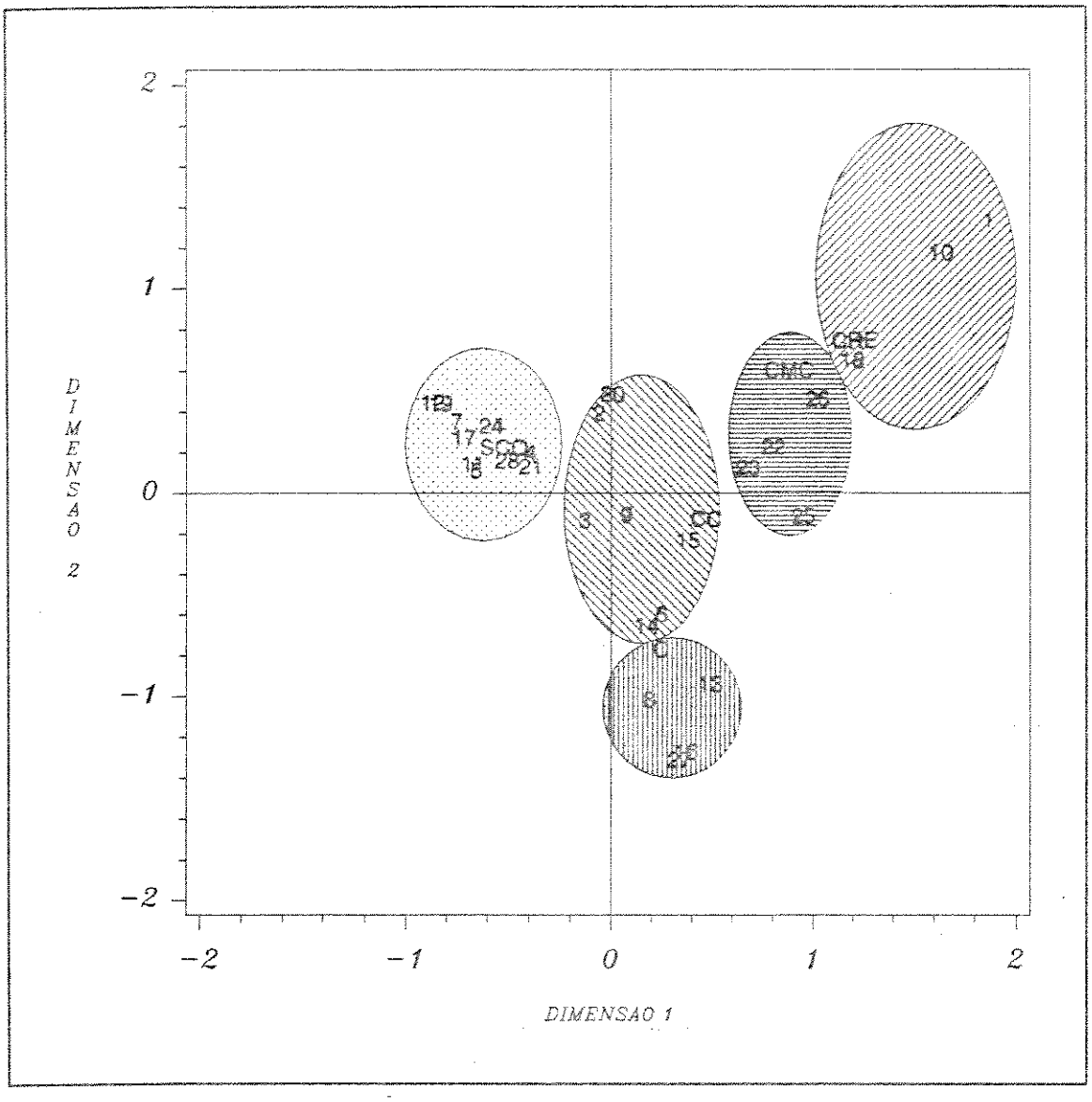

Figura 4.15: Análise de correspondência simples dos poços com a variável tamanho de conchas, ponderada pela espessura. Dimensão 1 (49,18%) e dimensão 2 (34,21%).

Outra variável interessante de se mapear é a quantidade de matriz terrígena. A figura 4.17 mostra o resultado da análise de correspondência e a figura 4.18 mostra o mapeamento da variável na bacia. Colocando-se no mapa esses resultados, pode-se observar uma tendência de os poços com terrígenos se distribuírem na área mais proximal da bacia, justamente onde se concentram os poços sem conchas; os carbonatos puros concentram-se em uma faixa de direção NE-SW,

mais ou menos paralela à costa, sendo seguidos por carbonatos arenosos e muito arenosos.

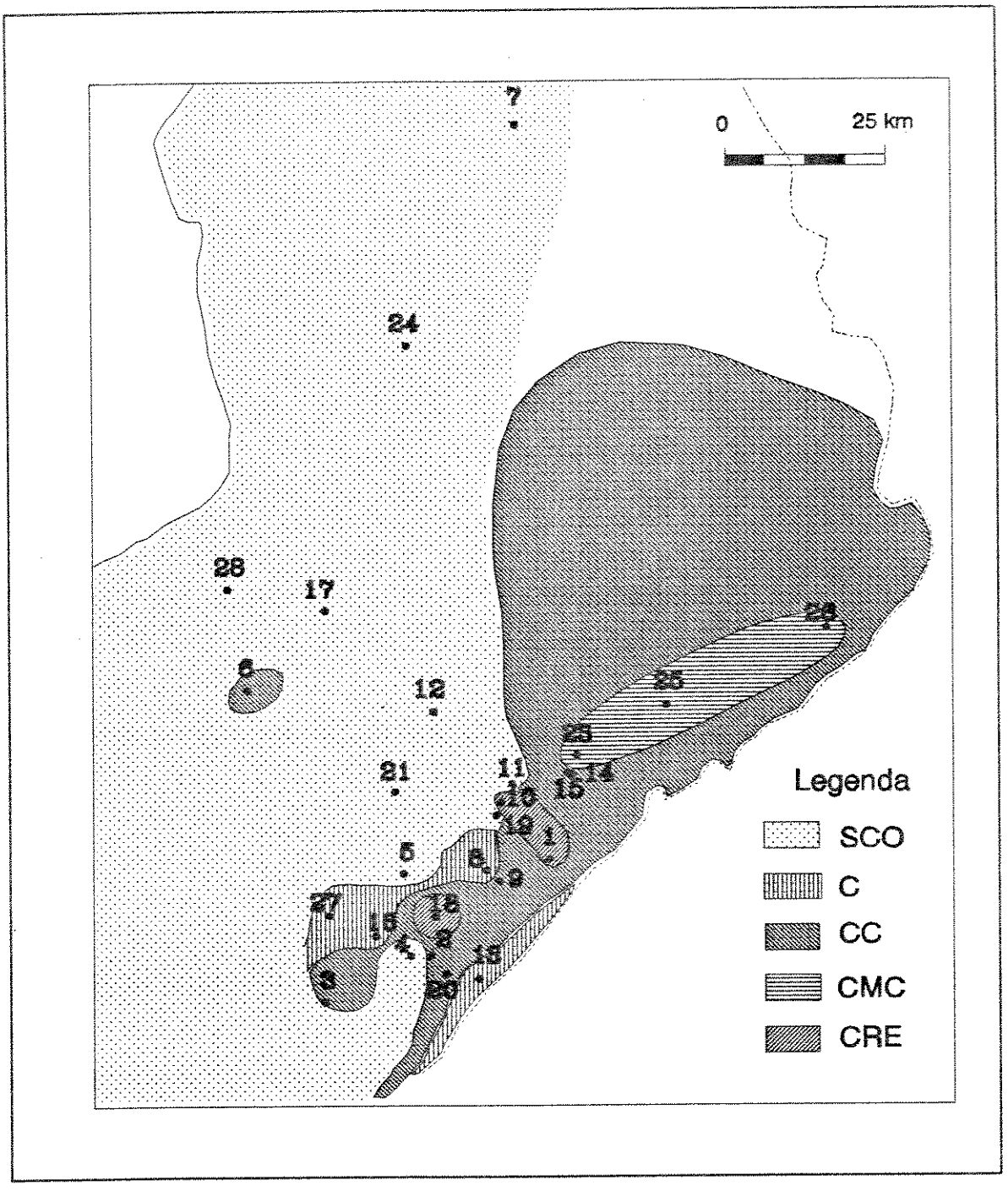

Figura 4.16: Mapeamento da variável tipo de rochas, com base nos agrupamentos obtidos da análise de correspondência.

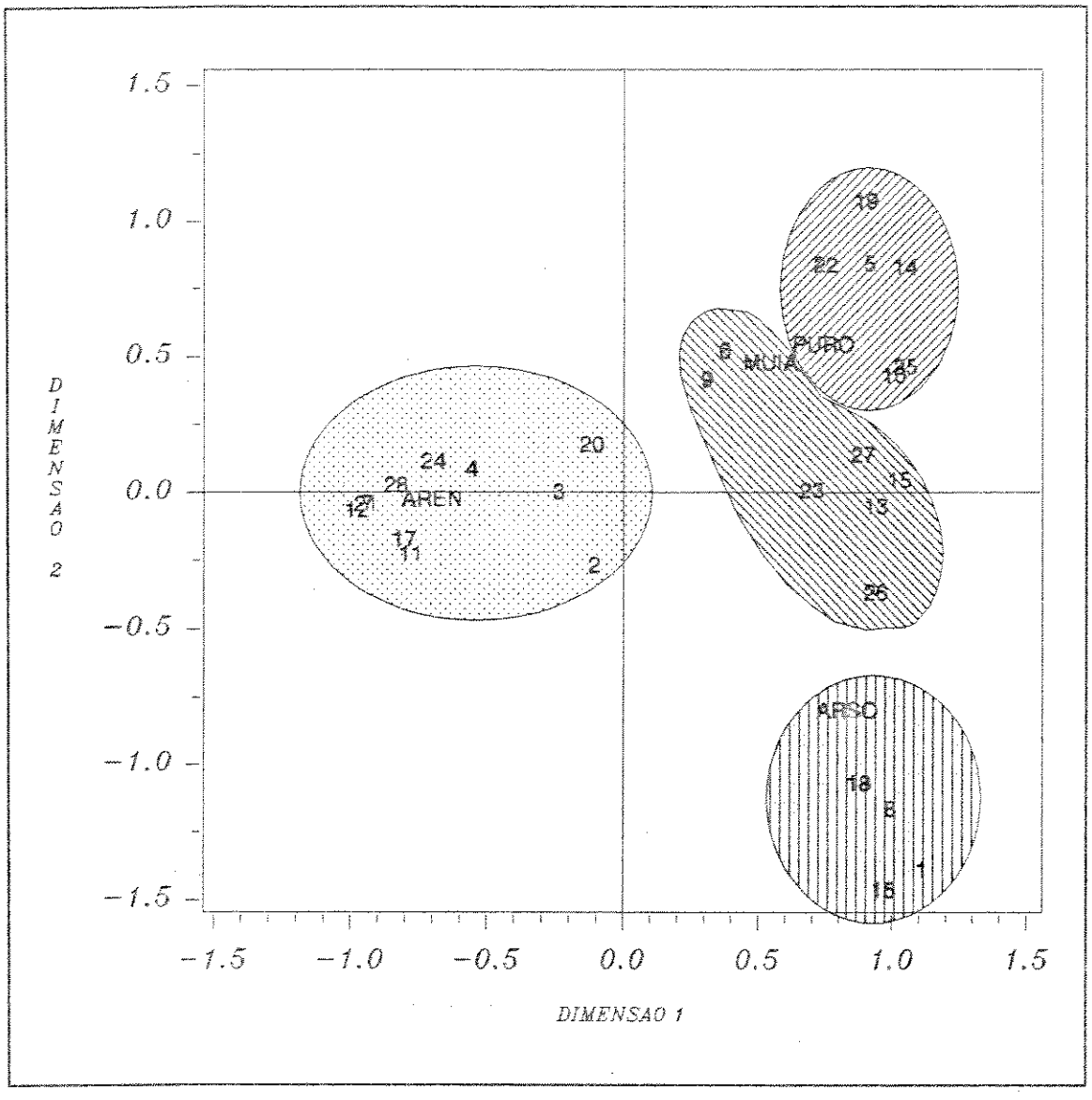

Figura 4.17: Análise de correspondência simples dos poços com a variável quantidade de matriz terrígena, ponderada pela espessura. Dimensão 1 (61,75%) e dimensão 2 (23,73%).

Observe-se que a representação bidimensional da análise de correspondência simples dos poços com as variáveis é também considerada boa.

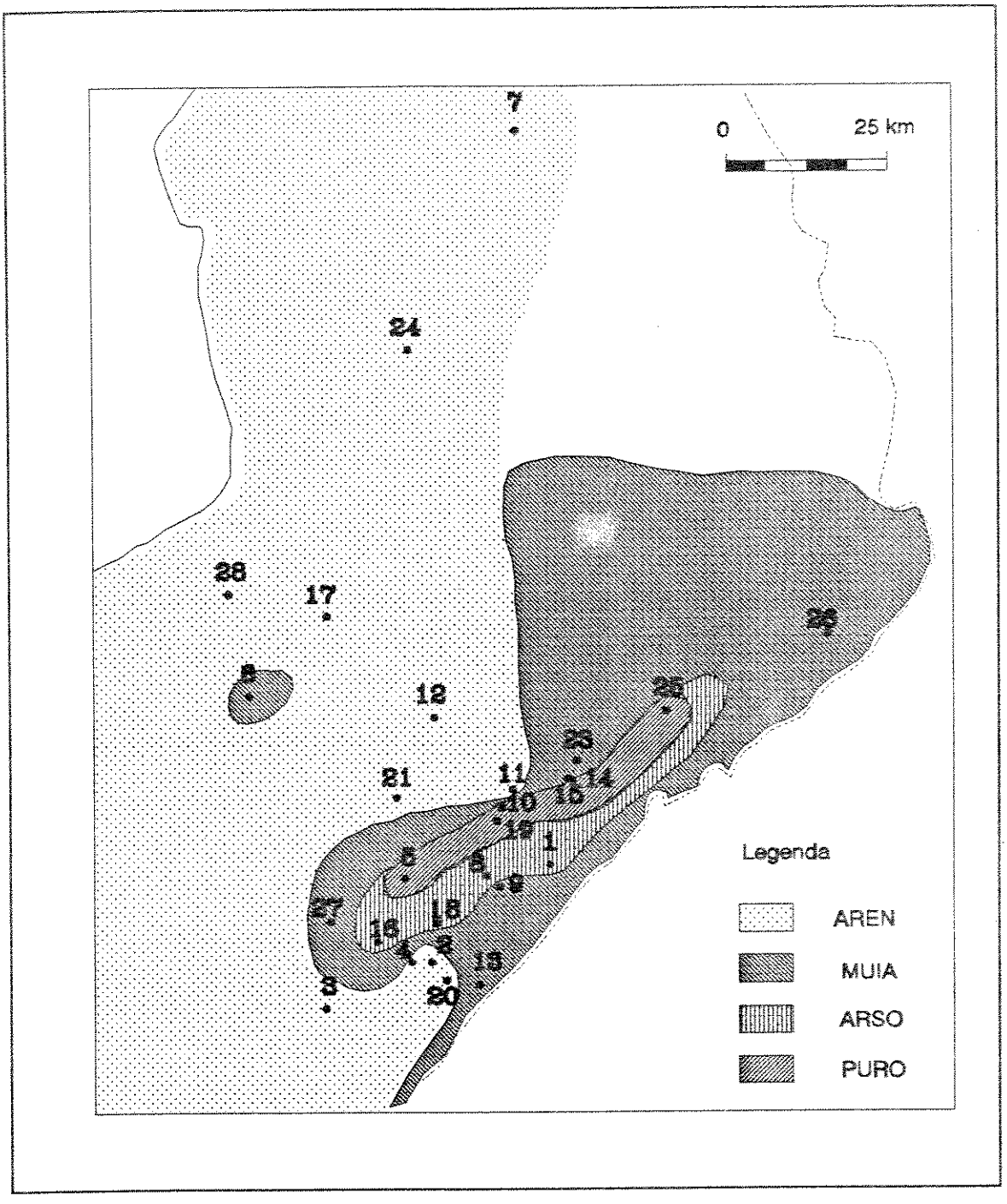

Figura 4.18: Distribuição dos poços em relação à variável matriz terrígena, ponderada pela espessura.

4.2.3 Análise de Correspondência Múltipla

Foi também realizada a análise de correspondência múltipla entre todas as variáveis, para entenderem-se as relações existentes entre elas. Como foi citado no capítulo anterior, esse tipo de análise nem sempre resulta em uma representação bidimensional boa. A figura 4.19 mostra este resultado: observam-se cinco agrupamentos no plano fatorial formado pelos dois primeiros eixos. No entanto, a inércia de cada um deles é pequena (22,14 + 5,80% = 27,84%), ou seja, a aproximação bidimensional não é boa e seriam necessárias mais de três dimensões para uma boa representação dos pontos.

Observa-se uma separação nítida ao longo da dimensão 1, entre terrígenos (I), à esquerda e carbonatos, à direita. Os terrígenos, ou grupo I, associam-se à ausência de conchas, com a granulometria variando de argila a areia média. Os carbonatos separam-se em 4 grupos com relação à dimensão 2. O grupo II é formado por calcirruditos muito calcareníticos e calcarenitos de granulometria grosseira. O grupo III é formado por calcirruditos calcareníticos, dominantemente *Grainstones/Packstones*, arenosos a puros, empacotamento normal a denso, com conchas abertas e quebradas e granulometria grosseira e granulosa. O grupo IV é formado de calcirruditos dominantemente *Wackestones* e *Mudstones*, muito arenosos, de empacotamento frouxo com conchas inteiras e de espessura E. Por fim, o grupo V reúne as categorias conchas fechadas e espessuras A a D, que se situam mais longe da origem, indicando que são categorias pouco expressivas nos dados, não se associando fortemente a qualquer dos tipos definidos de rochas. A dimensão 2 indica uma diminuição da granulometria dos carbonatos, do topo para a base.

Esses quatro agrupamentos podem ser grosseiramente relacionados a fácies sedimentares. Como abordagem exploratória de grande massa de dados, essa primeira simplificação dos dados pode ser útil na indicação de procedimentos posteriores.

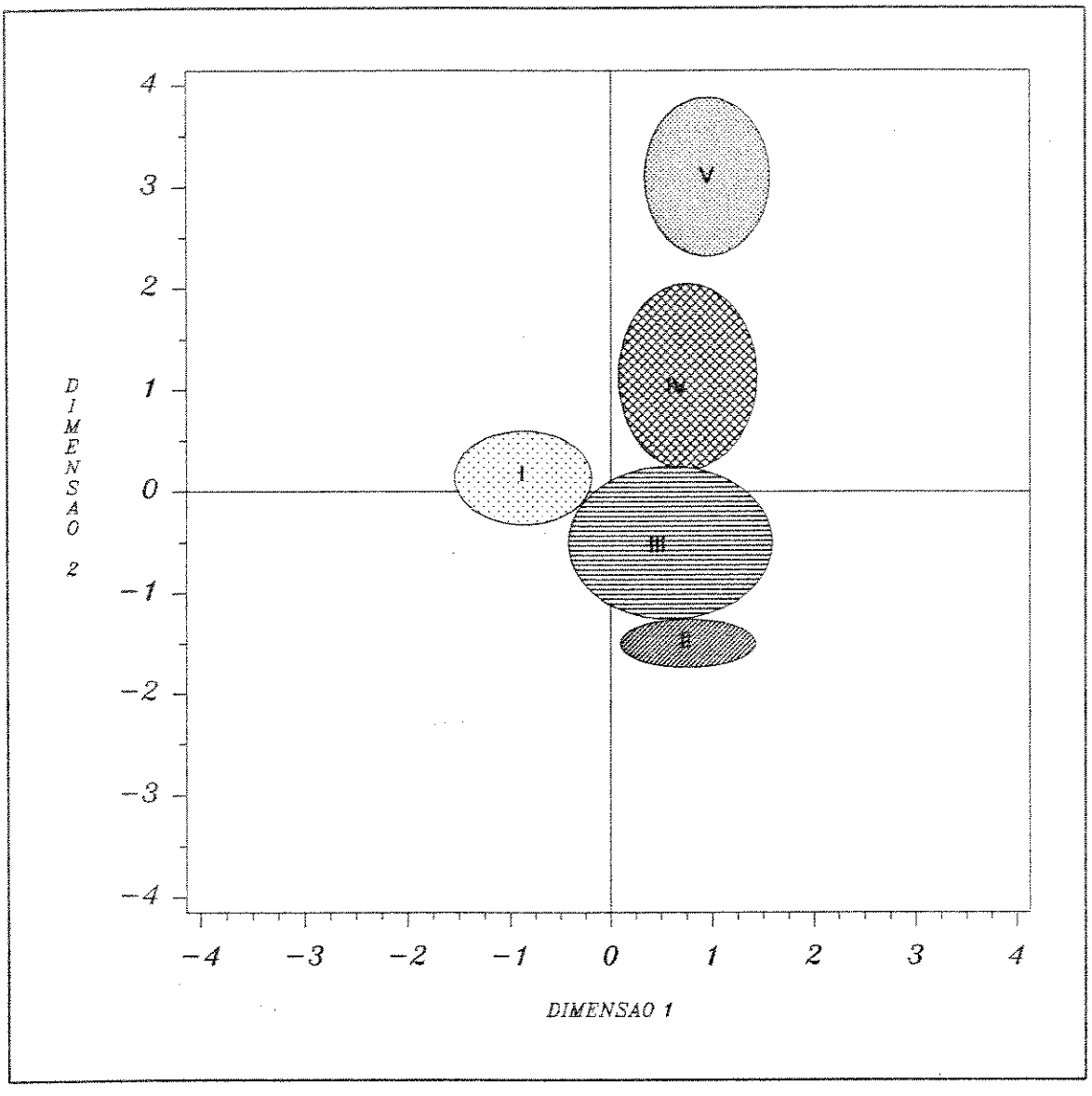

Figura 4.19: Análise de correspondência múltipla de todas as variáveis. Dimensão 1 (22,14%) e dimensão 2 (5,80%).

4.2.4 Análise de Correspondência Simples das Fácies Sedimentares

Através da análise de correspondência, pode-se também estudar a distribuição faciológica dos poços ao longo da bacia. Para isso, foi necessário transformar a matriz original em uma matriz de fácies. Os critérios utilizados nesta transformação foram os mesmos definidos por Carvalho et alii, 1992, e podem ser vistos no programa P.2. Na figura 4.20, observa-se o resultado da análise de correspondência entre os poços e as fácies que ocorrem em cada um deles, ponderada pela espessura. A partir da identificação da fácies predominante em um determinado poço, pode-se mapear a distribuição destas fácies como mostra a figura 4.21. Apesar de a amostragem não ser abrangente, ela pode dar uma idéia da distribuição faciológica na bacia.

4.2.5 Análise de Correspondência Simples das Fácies Sedimentares por Cronozona

Outro procedimento utilizado foi o mapeamento das fácies sedimentares por cronozonas. A seqüência das coquinas foi dividida em seis cronozonas (Carvalho et alli, 1992 - no prelo), com base na ocorrência de ostracodes. São elas, da base para o topo: CA, CB, CC, CD, CE e CF. O mapeamento das fácies sedimentares, por cronozona, fornece uma visão mais realista de sua distribuição na bacia, uma vez que não mistura sedimentos de idades diferentes, mas não foi feito pois o número de poços em cada uma delas é muito pequeno. As figuras 4.22 a 4.26 mostram a análise de correspondência entre as fácies e os poços por cronozona.

Finalmente, pode-se concluir que essa técnica permite uma visualização global fácil das relações entre as variáveis e dos poços com as variáveis, de forma a ajudar na interpretação da distribuição das fácies sedimentares e mesmo no agrupamento das variáveis para levar a classificação das amostras. Para isto é necessário que os valores numéricos sejam categorizados segundo critérios com significado geológico.

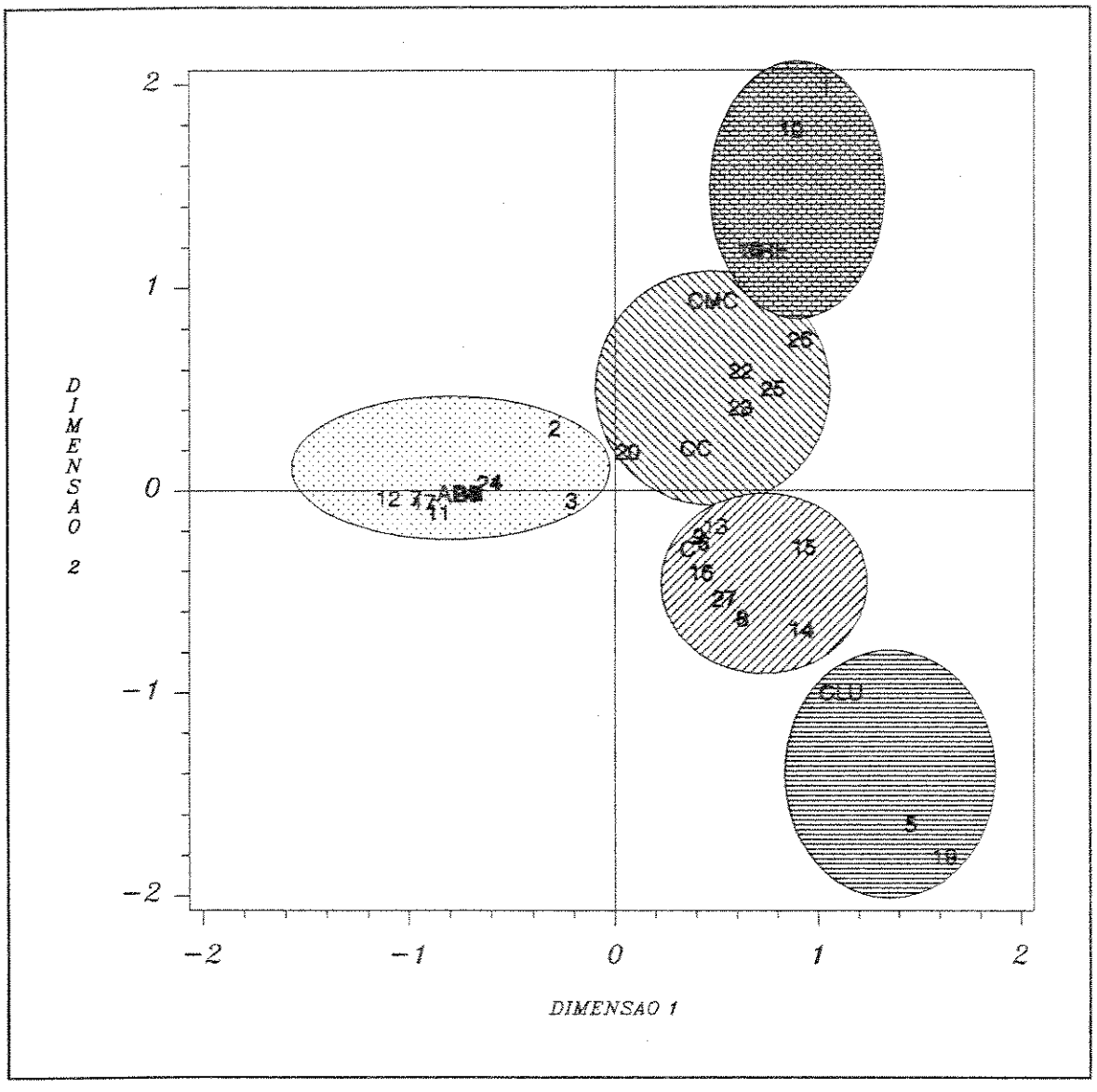

Figura 4.20: Análise de correspondência simples dos poços com as fácies, ponderada pela espessura.
Dimensão 1 (42,12%) e dimensão 2 (26,63%).

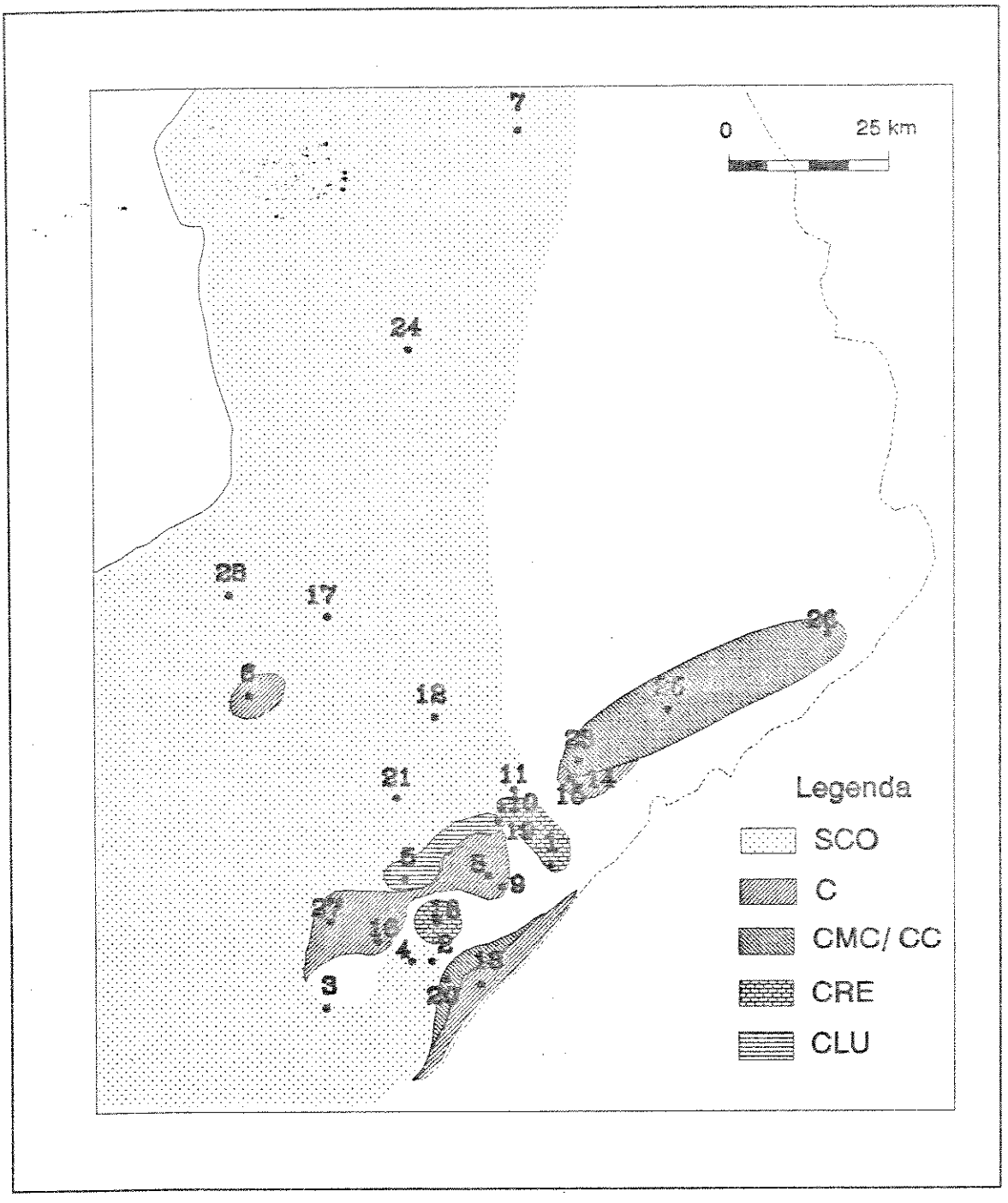

Figura 4.21: Mapeamento da distribuição das fácies sedimentares na bacia.

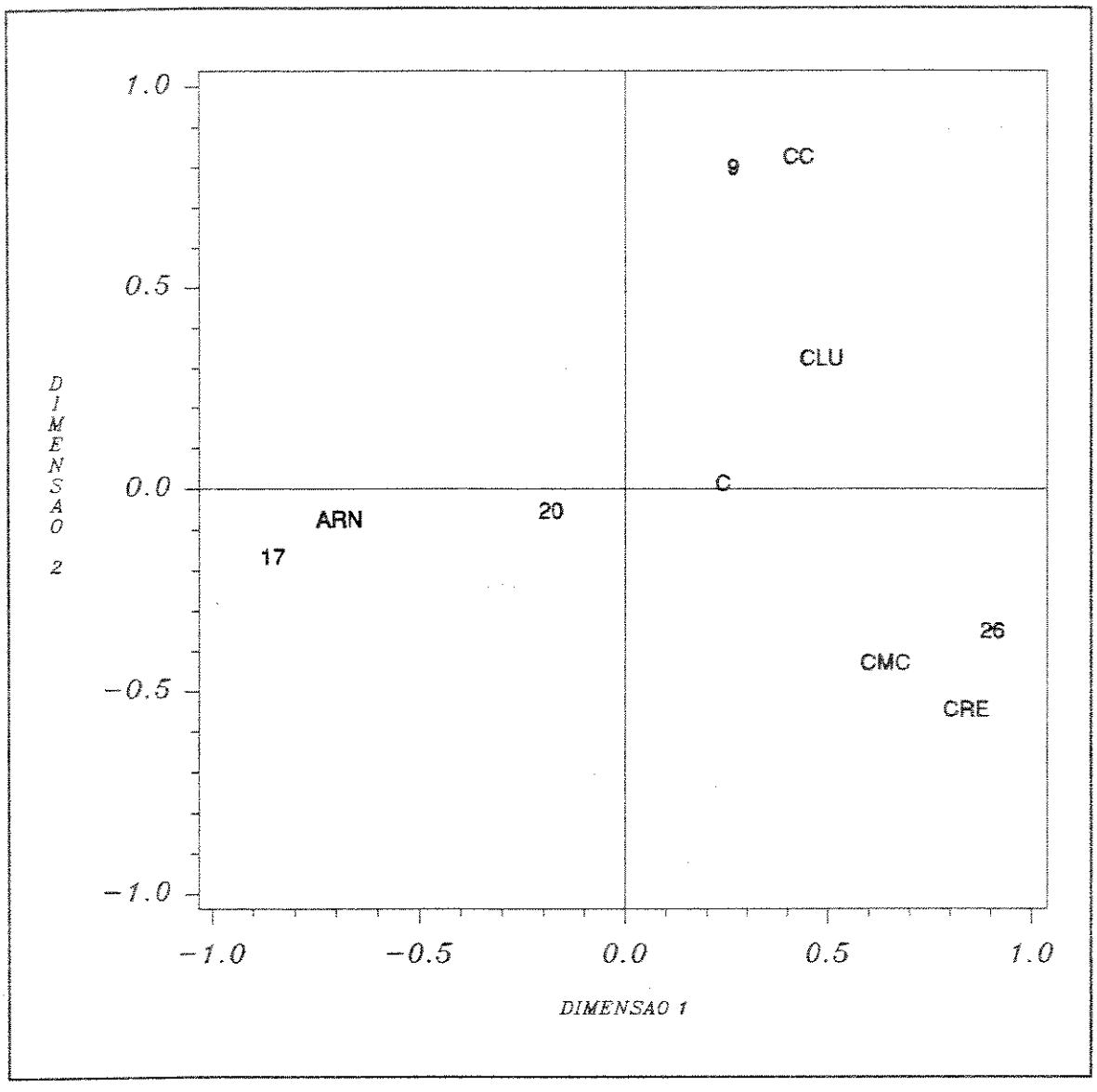

Figura 4.22: Análise de correspondência entre os poços e as fácies simplificadas, ponderadas pela espessura, na cronozona CB. Dimensão 1 (71,94%), dimensão 2 (23,77%).

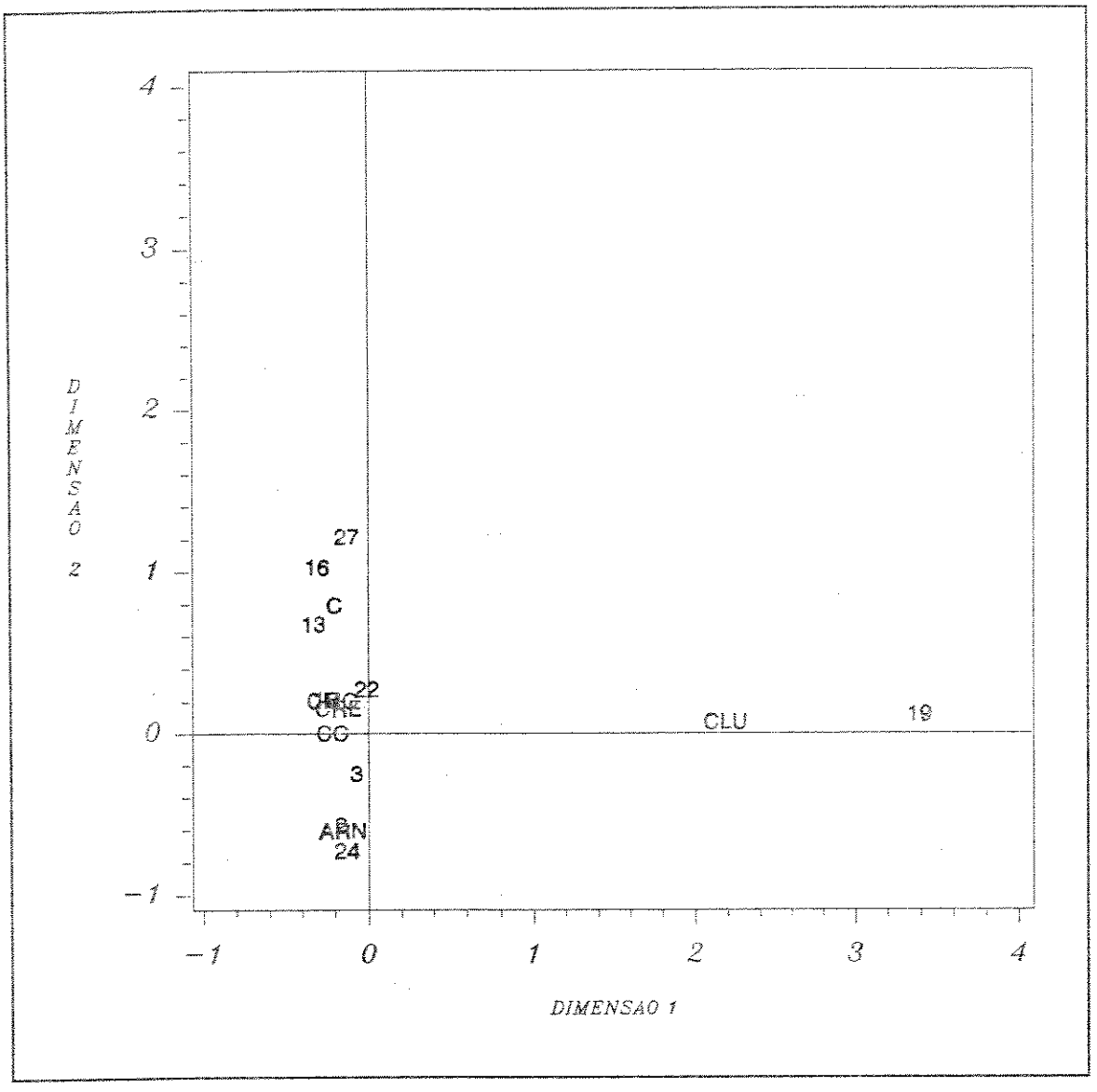

Figura 4.23: Análise de correspondência entre os poços e as fácies simplificadas, ponderadas pela espessura, na cronozona CC. Dimensão 1 (36,85%), dimensão 2 (33,97%).

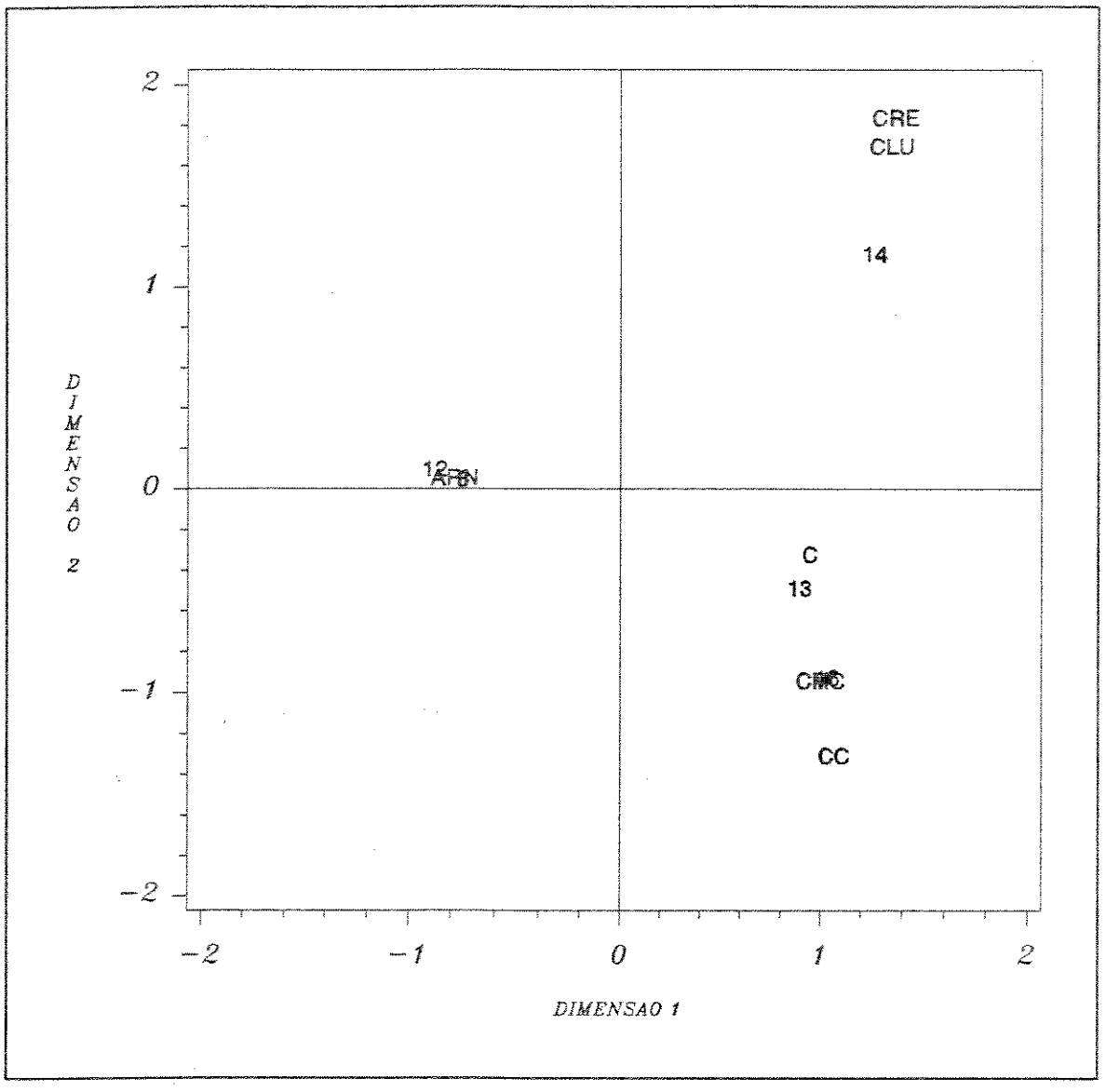

Figura 4.24: Análise de correspondência entre os poços e as fácies simplificadas, ponderadas pela espessura, na cronozona CD. Dimensão 1 (53,08%), dimensão 2 (46,62%).

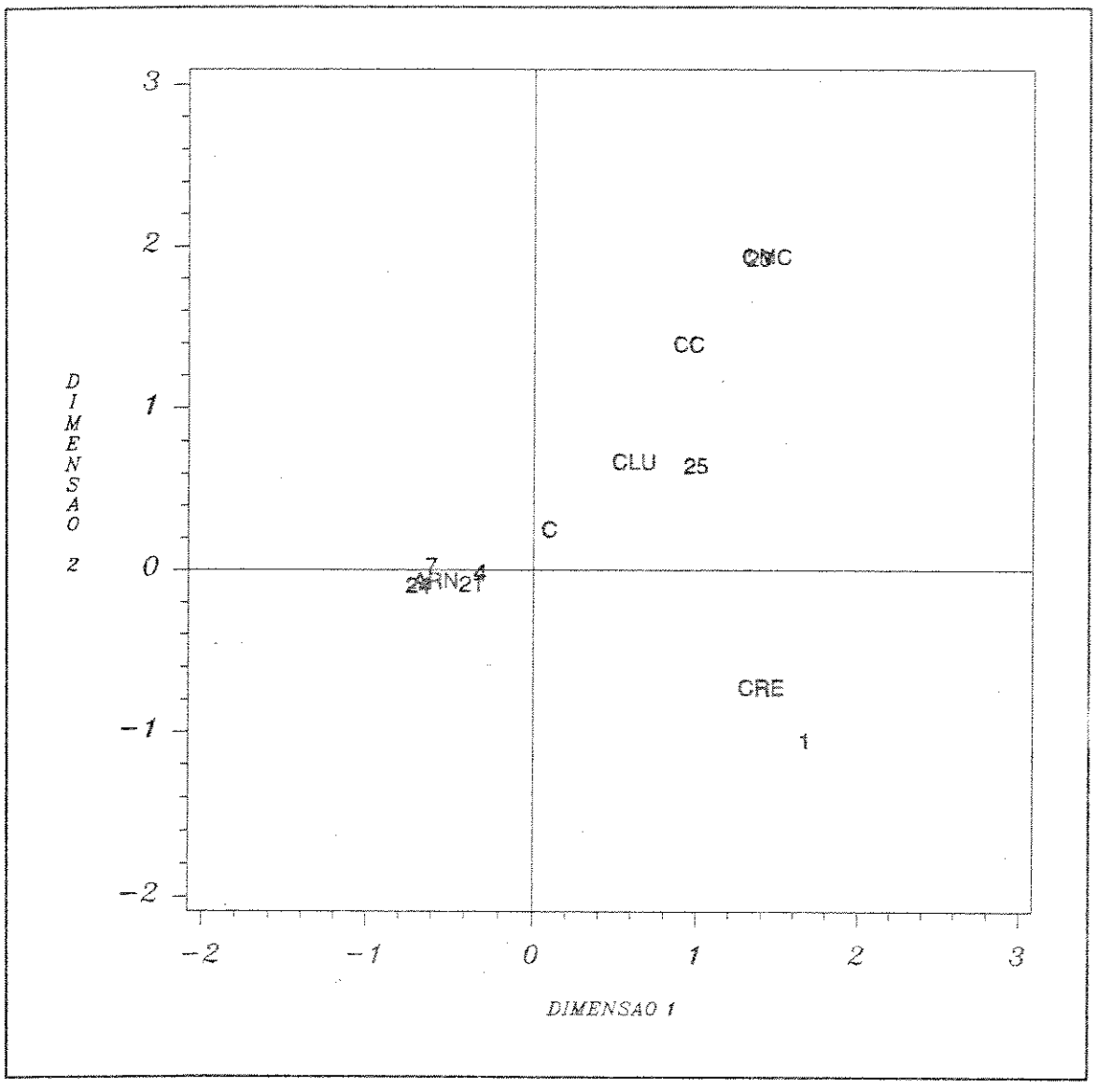

Figura 4.25: Análise de correspondência entre os poços e as fácies simplificadas, ponderadas pela espessura, na cronozona CE Dimensão 1 (58,73%), dimensão 2 (31,44%).

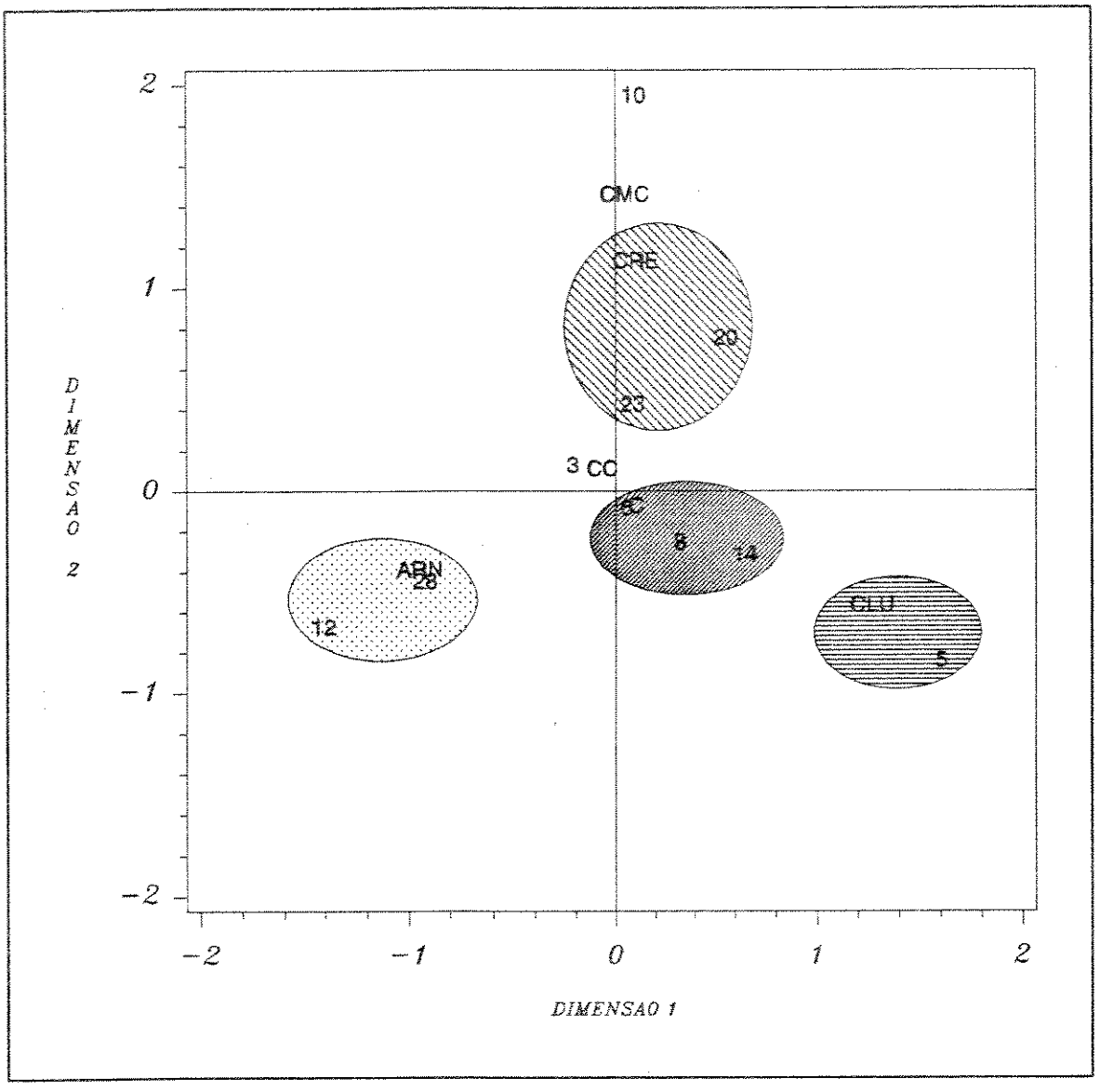

Figura 4.26: Análise de correspondência entre os poços e as fácies simplificadas, ponderadas pela espessura, na cronozona CF Dimensão 1 (38,26%), dimensão 2 (27,50%).

5 - CONCLUSÕES E DISCUSSÃO

5.1 CONCLUSÕES

Os métodos estatísticos descritivos têm como compromisso facilitar a interpretação dos dados com um mínimo de perda de informação. Alguns deles representam exatamente os dados, mas podem ser difíceis de serem totalmente assimilados visualmente. Outros sacrificam alguma informação (geralmente o mínimo possível), para fornecer representações de interpretação mais fácil. A vantagem da análise de correspondência é que o ganho na facilidade de interpretação excede a perda de informações.

A análise de correspondência é uma ferramenta poderosa na análise exploratória de dados, pois permite visualizar ao mesmo tempo as relações entre as linhas e colunas de uma tabela de contingência: no caso analisado, entre os poços e as variáveis e entre as diversas variáveis entre si. Sua aplicação através do SAS é fácil e transparente para o usuário, que necessita somente saber interpretar os resultados da análise. No anexo AX.1 encontra-se um exemplo de saída da análise de correspondência.

A análise de correspondência tem a vantagem de considerar conjuntamente informações qualitativas e quantitativas, fazendo uma ponte entre as abordagens puramente descritivas e os modelos numéricos. No exemplo apresentado, foram consideradas informações descritivas (fácies sedimentares) e informações quantitativas categorizadas.

Embora a literatura geológica registre uma diminuição do número de trabalhos publicados na área de análise de dados multivariados, em comparação com um aumento de publicações na área de geoestatística (Agterberg e Griffiths, 1991), estas técnicas podem ser muito úteis na fase inicial de projetos envolvendo grandes

volumes de dados. O trabalho de Pereira et alii (1990) exemplifica bem este fato.

Neste trabalho, a aplicação desta técnica ficou restrita ao conjunto de dados disponíveis. Embora os resultados possam parecer um tanto previsíveis, deve-se ressaltar que esta metodologia resulta em interpretações rápidas e fáceis das relações entre os dados, podendo ser aplicada a dados nos quais, *a priori*, não se conhecem as relações existentes. Na área de geoengenharia de reservatórios, onde geralmente se trabalha com grandes quantidades de informações, quase impossíveis de serem analisadas manualmente, estas técnicas de análise de dados multivariados fornecem as relações mais importantes entre os dados, simplificando sobremaneira o entendimento dos mesmos.

Através da análise de correspondência, poder-se-ia, por exemplo, analisar as relações existentes entre dados petrográficos, fácies sedimentares, dados petrofísicos, dados de perfis e dados de produção. Esta análise permitiria caracterizar os reservatórios em termos destas informações, através da definição de fácies petrofísicas, fácies perfis e unidades de fluxo.

Alguns aspectos da análise de correspondência devem ser ressaltados: 1) deve-se lembrar que esta técnica não considera a distribuição dos dados, ou seja, é não-paramétrica, conseqüentemente, não existe teste de confiança ou regra para aceitar ou rejeitar a significância dos fatores ou eixos; 2) o gráfico nada mais é do que a representação simultânea das análises R e Q-modal com escala apropriada: a proximidade entre os pontos amostrais caracteriza membros de grupos similares, enquanto que a proximidade entre pontos de variáveis implica em comportamento similar; 3) estes gráficos representam somente as projeções dos pontos sobre o plano fatorial formado pelos dois primeiros eixos, de modo que, quando dois pontos estão sobrepostos, no plano formado pelos dois primeiros fatores, podem estar na verdade distantes no espaço original. A projeção destes pontos sobre o plano formado pelo primeiro e terceiro fator pode resolver este tipo de dúvida; 4) a proximidade de um grupo de pontos de amostras e de variáveis é realmente o que ajuda a identificar grupos e caracterizá-los.

5.2 DISCUSSÃO

Segundo Zhou et alii (1983) as várias versões da análise fatorial diferem basicamente na forma pela qual os dados são escalonados. O escalonamento determina a medida de similaridade e, conseqüentemente, a natureza da solução fatorial. Por exemplo, o co-seno θ é a medida de similaridade entre amostras, utilizada na análise Q-modal de Imbrie (1963); o coeficiente de correlação ou covariância são as medidas de similaridade utilizadas na análise R-modal; o perfil-distância é a medida de similaridade utilizada na análise de correspondência, e assim por diante. Depois de escalonados os dados, fatora-se a matriz através da rotação dos eixos de referência, sob certas condições, tais como considerar a máxima variância dos dados. Depois da fatoração, a configuração dos pontos analisados ainda permanece a mesma e pode ser vista mais claramente em uma dimensão reduzida. A redução da dimensão é dada pela decomposição da matriz de dados em seus valores e vetores singulares (figura 5.1).

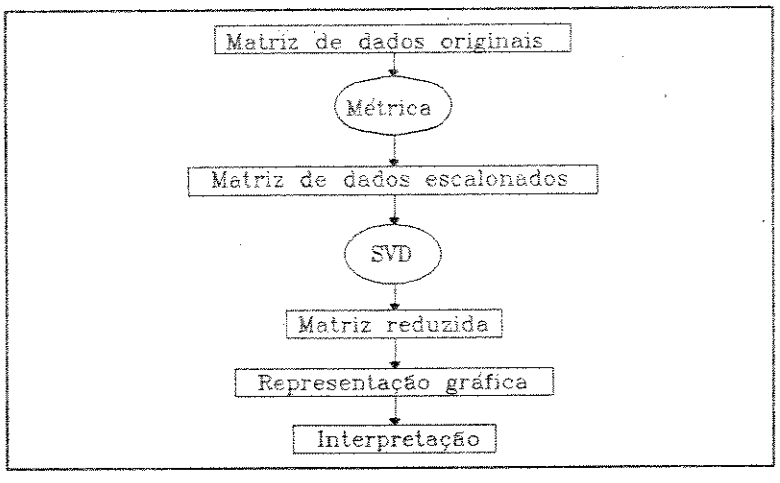

Figura 5.1: Fluxograma da análise fatorial.

Em muitas aplicações geológicas, o coeficiente de correlação e, mais raramente a covariância, são medidas apropriadas de similaridades entre variáveis; logo, a análise de componentes principais é um procedimento R-modal satisfatório. Já na análise Q-modal o problema é mais complicado. Nela, geralmente são necessárias medidas de similaridade proporcionais, porque as proporções, mais do

que a magnitude dos constituintes indicam a fonte de uma observação. Nestas situações, o co-seno θ fornece uma boa medida e a análise Q-modal de Imbrie é realizada. Contudo, em outras situações, a magnitude dos constituintes é mais importante. Por exemplo, suponha-se que duas amostras foram analisadas em relação à presença de elementos traços; a primeira contém 100 ppm de Cu e 500 ppm de Pb e a segunda contém 10 ppm de Cu e 50 ppm de Pb. Note-se que, em termos da importância como guia de mineralização, as duas amostras são bem diferentes, apesar de terem a mesma proporção de Cu e Pb. Neste caso, é importante utilizar uma medida de similaridade que seja sensível à diferença de magnitude, como por exemplo a distância Euclidiana, e a análise de coordenadas principais é o procedimento Q-modal satisfatório.

Quando o mesmo tipo de escalonamento para as variáveis e para as amostras resulta em uma medida apropriada de similaridade para o problema em questão, a análise RQ-modal pode ser aplicada. Na análise de correspondência, utiliza-se o perfil-distância como medida de similaridade, tanto para as linhas quanto para as colunas de uma tabela de contingência. Este tipo de medida é sensível às proporções, mas não é sensível às diferenças de magnitude. Na aplicação estudada, quando se utilizou a análise de correspondência simples entre duas variáveis, ou a análise de correspondência simples entre poços e variáveis, ou ainda, a análise de correspondência múltipla entre todas as variáveis, o perfil-distância é adequado, uma vez que se está interessado em proporções.

Geralmente, a medida de similaridade entre amostras é mais relevantemente expressa pela distância Euclidiana do que por uma medida de similaridade proporcional. Neste caso, o biplot é um procedimento adequado, uma vez que utiliza, como medida de similaridade entre amostras, a distância Euclidiana, e como medida de similaridade entre as variáveis, o coeficiente de correlação. Este procedimento pode ser aplicado aos dados, e uma comparação dos resultados da análise de correspondência e do biplot pode gerar conclusões importantes.

6 - REFERÊNCIAS BIBLIOGRÁFICAS

ABLER, R.; ADAMS, J. S.; GOULD, P. - 1971 - *Spatial organization - the geographer's view of the world*. New Jersey: Prentice-Hall, Inc., Englewood Cliffs. 587p.

AGTERBERG, F. P. & GRIFFITHS, C. M. - 1991 - Computer applications in stratigraphy 1989/1990: a review. *Computers & Geosciences, 17* (8):1105-1118.

BAUMGARTEN, C. S.; DULTRA, A. J. C.; CARVALHO, M. D. - 1983 - *Zoneamento do intervalo de coquinas Lagoa Feia em Pampo, Linguado e Badejo*. PETROBRÁS/DEPEX/DIRSUL/SEDESU. Relatório Interno.

BAUMGARTEN, C. S. - 1985 - Evolução estrutural de Pampo, Badejo e Linguado durante a deposição da Fm. Lagoa Feia. *Boletim Técnico da PETROBRÁS, 28* (2):91-101.

BAUMGARTEN, C. S. et alli - 1988 - Coquinas da Formação Lagoa Feia, Bacia de Campos: Evolução da geologia de desenvolvimento. *Boletim de Geociências da PETROBRÁS*, 2 (1):27-36.

BEN-ISRAEL, A. & GREVILLE, T. N. E. - 1974 - *Generalized inverses: theory and applications*. New York: John Wiley & Sons.

BENZÉCRI, J. P. - 1973 - *L'Analyse des données*, Tome II. L'Analyse des correspondences. Paris: Dunod. 619p.

BERTANI, R. T. & CAROZZI, A. V. - 1984 - Microfacies, depositional models and diagenesis of Lagoa Feia Formation (Lower Cretaceous) Campos basin, offshore Brazil. PETROBRÁS/CENPES, *Ciência-Técnica-Petróleo, 14*:104.

BONHAM-CARTER, G. F. - 1965 - A numerical method of classification using qualitative and semi-quantitative data, as applied to the facies analysis of limestones. *Bulletin of Canadian Petroleum Geology, 13* (4):482-502.

BONHAM-CARTER, G. F.; GRADSTEIN, F. M.; D'IORIO, M. A. - 1986 - Distribution of Cenozoic foraminifera from the northwestern Atlantic margin analyzed by correspondence analysis. *Computers & Geosciences*, 12 (4B):621-635.

BOUROCHE, J. M. & SAPORTA, G. - 1980 - *L'Analyse des données*. Paris, Presses Universitaires de France. 127p.

BUCHEB, J. A. - 1991 - *Aplicação de tratamento estatístico multivariante em dados de perfis de poços da Bacia de Sergipe-Alagoas*. Tese de mestrado. Universidade Federal do Pará. 136p.

BUZAS, M. A. - 1979 - Quantitative Biofacies Analysis. In: Foraminiferal Ecology and Paleoecology, Houston, Texas. *SEPM - Short Course* n. 6, p.11-20.

CARDOSO, R. H. A. - 1990 - *Análise de dados multivariados através de técnicas baseadas na decomposição em valores singulares*. Tese de mestrado. Universidade Estadual de Campinas. 118p.

CARR, J. R. - 1990 - Corsponde: a portable fortran-77 program for correspondence analysis. *Computers & Geosciences, 16* (3):289-307.

CARVALHO, M.D; MONTEIRO, M.C.; PIMENTEL, A. M.; REHIM, H. A. A. A. - 1984 - *Microfácies, diagênese e petrofísica das coquinas da Formação Lagoa Feia em Badejo, Linguado e Pampo - Bacia de campos*. PETROBRÁS/CENPES. Relatório Interno n.672-4001, 130p.

CARVALHO, M.D et alii - 1992 - *Análise regional da seqüência das coquinas. Formação Lagoa Feia - Bacia de Campos*. PETROBRÁS/CENPES/DEPEX. Relatório Interno, no prelo.

CARVALHO, M. S. - 1992 - *Análise de correspondência - uma aplicação do método em avaliação de serviços*. Escola Nacional de Saúde Pública, Secretaria de Des. Educacional. Rio de Janeiro.

CASTRO, J. & AZAMBUJA FILHO, N. C. - 1981 - *Fácies, análise estratigráfica e reservatórios da Formação Lagoa Feia. Cretáceo Inferior da Bacia de Campos*. Rio de Janeiro, PETROBRÁS, 110 p. Relatório Interno.

CHAMBERS, J. M. - 1977 - *Computacional methods for data analysis*. New York: John Wiley & Sons.

CHAYES, F. - 1971 - *Ratio correlation - a manual for students of petrology and geochemistry*. Chicago: The University of Chicago Press. 99 p.

DAVID, M.; DAGBERT, M.; BEAUCHEMIN, Y. - 1977 - Statistical analysis in geology: correspondence analysis method. *Quaterly of the Colorado School of Mines, 72* (1):1-60.

DAVID, M.; CAMPIGLIO, C.; DARLING, R. - 1974 - Progress in R- and Q-Mode analysis: Correspondence analysis and its application to the study of geological processes. *Can. J. Earth Sci., 11*:131-146.

DAVIS, J. C. - 1986 - *Statistics and data analysis in geology*. New York: John Wiley & Sons. 646 p.

DIAS, J. L. et alii - 1987 - *Estudo regional da Formação Lagoa Feia*. PETROBRÁS/DEPEX/CENPES. Relatório Interno, v.1, 143p.

DIAS, J. L.; OLIVEIRA, J. Q.; VIEIRA, J. C. - 1988 - Sedimentological and stratigraphic analysis of the Lagoa Feia Formation, rift phase of Campos Basin, offshore Brazil. *Revista Brasileira de Geociências, 18* (03):252-260.

ECKART, C. & YOUNG, B. - 1936 - The approximation of one matrix by another of lower rank. *Psychometrika, 1* (03):221-218.

EVERITT, B. S. - 1977 - *The analysis of contingency tables*. New York, John Wiley & Sons, 95p.

FABIAN-GOYHENECHE, C.; PELLERIN, F. M.; GLOTIN, G.; PFLUGFELDER, B. - 1990 - Contribution of the statistical tool for the synthesis of geological and petrophysical data: application to a complex carbonate reservoir. *BCREDP*, Soc. Nat. Elf-Acquitaine, *15*:308-322.

FISHER, R. A. - 1940 - The precision of discriminant functions. *Ann. Eugen., 10*:422-429.

FORSYTHE, G.; MALCOLM, M.A.; MOLER, C. B. - 1977 - *Computer methods for mathematical computations*. New Jersey: Prentice-Hall, Inc, 259p.

GABRIEL, K. R. - 1971 - The biplot graphic display of matrices with application to principal component analysis. *Biometrika, 58* (3):453-467.

GABRIEL, K. R. - 1978 - Least squares aproximation of matrices by additive and multiplicative models. *J. Am. Stat. Soc., 40* (2):186-196.

GOOD, I. J. - 1969 - Some applications of the singular decomposition of a matrix. *Technometrics, 11*:823-831.

GREENACRE, M. J. - 1984 - *Theory and applications of correspondence analysis*. London: Academic Press, 364p.

GREENACRE, M. J. & HASTIE, T. - 1987 - The geometric interpretation of correspondence analysis. *JASA, 82*:437-447.

GREENACRE, M. J. & UNDERHILL, L. G. - 1982 - Scaling a data matrix in low-dimensional Euclidean space. IN: HAWKINS, D. M. ed., *Topics in applied multivariate analysis*. Cambridge: Cambridge University Press, p.183-268.

GUTTMAN, L. - 1941 - The quantification of a class of attributes: A theory and method of scale construction. IN: HORST, P., *The prediction of personal adjustment*. New York: Social Science Research Council, p.319-348.

HALMOS, P. R. - 1974 - *Finite-dimensional vector spaces*. New York: Springer-Verlag, 200p.

HARBAUGH, J. W. & MERRIAM, D. F. - 1968 - Computer applications in stratigraphic analysis. New York: John Wiley & Sons, Inc.

HAYASHI, C. - 1950 - On the quantification of qualitative data from the mathematico-statistical point of view. *Ann. Inst. Statist. Math.*, 2:35-47.

HILL, M. O. - 1974 - Correspondence analysis - a neglected multivariate method. *Jour. Royal Stat. Soc. Ser. C, Appl. Stat.*, 23 (03):340-354.

HIRSCHFELD, H. O. - 1935 - A connection between correlation and contingency. *Cambridge Philosophical Soc. Proc. (Math. Proc.)*, 31:520-524.

HORSCHUTZ, P. M. C. & SCUTA, M. S. - 1992 - *Fácies-perfis e mapeamento de qualidade do reservatório de coquinas da Formação Lagoa Feia do Campo de Pampo.* PETROBRÁS/DEPEX/DIGED/DIGEO. Relatório Interno.

HORST, P. - 1935 - Measuring complex attitudes. *J. Social Psychol.*, 6:369-374.

IMBRIE, J. & PURDY, E. G. - 1962 - Classification of modern Bahamian carbonate sediments. In: Classification of Carbonate Rocks, a Symposium: *AAPG*, Memoir 1, p.253-272.

ISNARD, P.; MALLET, J. L.; CAZES, P.; SATTRAN, V. - 1972 - Correlations géologiques de traitement des données. IN: LAFITTE, P. , *Traité d'information géologique*. Paris: Masson et Cie Éditeurs, cap. 9, p.379-533.

LEBART, L. MORINEAU, A.; WARWICK, K. M. - 1984 - *Multivariate descriptive statistical analysis*. New York: John Wiley & Sons, 231p.

LECLERC, A. - 1975 - L'analyse des correspondences sur juxtaposition de tableaux de contingence. *Revue de Statistique Appliquée*, 23 (3):5-16.

LOBO, A. P. & FERRADAES, J. O. - 1983 - *Reconhecimento preliminar do talude e sopé continentais da bacia de Campos*. PETROBRÁS. Relatório Interno, 27p.

MANDEL, J. - 1982 - Use of the singular value decomposition in regression analysis. *Am. Statitician*, 36:15-24.

NISHISATO, S. - 1980 - *Analysis of categorical data: dual scaling and its applications*. Toronto: University of Toronto Press.

PEREIRA, H. G.; SILVA, A. C.; SOARES, A.; RIBEIRO, L.; CARVALHO, J. - 1990 -

Improving reservoir description by using geostatistical and multivariate data analysis techniques. *Mathematical Geology*, 22 (8):879-913.

RAO, C. R. - 1980 - Matrix approximations and redution of dimensionality in multivariate statistical analysis. IN: KRISHNAIAH, P.R.,ed., *Multivariate analysis*. Amsterdam: North Holland, vol. 5.

RICHARDSON, M. & KUDER, G. F. - 1933 - Making a rating scale that measures. *Personnel J., 12*:36-40.

RIJCKEVORSEL, J. L. A. VAN & DE LEEUW, J. - 1988 - *Component and correspondence analysis*. New York: John Wiley & Sons Ltd, 146p.

SAS Institute Inc. - 1988 - *SAS language guide*. Release 6.03 Edition. Cary, NC: SAS Institute Inc.

SAS Institute Inc. - 1988 - *SAS/IML user's guide*. Release 6.03 Edition. Cary, NC: SAS Institute Inc.

SAS Institute Inc. - 1987 - *SAS/STAT guide for personal computers*. Version 6 edition. Cary, NC: SAS Institute Inc.

SAS Institute Inc. - 1978 - *SAS/GRAPH guide for personal computers*. Version 6 edition. Cary, NC: SAS Institute Inc.

SCHALLER, H. - 1973 - Estratigrafia da Bacia de Campos. IN: CONGRESSO BRASILEIRO DE GEOLOGIA, 27, Aracaju, 1973. *Anais*... São Paulo, SBG, v.3, p.247-258.

SCHALLER, H.; TERRA, G.S.; SOUZA CRUZ, C. E.; SPADINI, A. R. - 1981 - *Estudo preliminar dos reservatórios da Fm. Lagoa Feia, área de Badejo/Pampo, Bacia de Campos*. Rio de Janeiro, PETROBRÁS/CENPES/LABOR. Relatório Interno, 17p.

SEARLE, S. R. - 1982 - *Matrix algebra useful for statistics*. New York: John Wiley & Sons.

SHILOV, G. E. - 1961 - *An introduction to the theory of linear spaces*. New Jersey: Prentice-Hall, Inc, 310 p. Traduzido do russo por Silverman R. A..

SIZE, W. B. - 1987 - *Use and abuse of statistical methods in the earth science*. Oxford University Press, 169p.

SOUZA JR., O. G. - 1988 - *Simulação condicional de unidades de fluxo na área II do projeto piloto de injeção de vapor na Formação Açu (Ksup), Campo de Estreito, Bacia Potiguar, Brasil*. Tese de mestrado. Universidade Federal de Ouro Preto, 168p.

SOUZA JR., O. G. - 1991 - Análise de dados multivariados, uma eficiente ferramenta para descrição e caracterização de reservatórios. IN: STOG, 3, Cabo Frio, 1991. *Anais*... Rio de Janeiro, STOG, v.1, p.121-130.

SZATMARI, P.; LOBO, A. P. et alii - 1983 - *Arcabouço tectônico da bacia de Campos e áreas adjacentes*. PETROBRÁS. Relatório Interno, 35p.

TEIL, H. - 1975 - Correspondence factor analysis: an outline of its method. *Mathematical Geology*, 7 (1):3-13.

TEIL, H. & CHEMINEE, J. L. - 1075 - Application of correspondence factor analysis to the study of major and trace elements in the Erta Ale Chain (Afar, Ethiopia). *Mathematical Geology*, 7 (1):13-31.

ZHOU, D.; CHANG, T.; DAVIS, J. C. - 1983 - Dual extraction of R-Mode and Q-Mode factor solutions. *Mathematical Geology*, 15 (5):581-606.

APÊNDICE

A.1 CONCEITOS GEOMÉTRICOS NO ESPAÇO MULTIDIMENSIONAL

Apresentam-se alguns conceitos geométricos no espaço multidimensional, que ajudarão no entendimento da análise de correspondência: a definição de distância, ângulo e produto escalar entre pontos no espaço multidimensional; a atribuição de pesos às dimensões do espaço e de massa aos pontos individuais; e a identificação de subespaços de menor dimensionalidade, que melhor se aproximem de um conjunto de pontos dados (Cardoso, 1990 apud Greenacre, 1984).

A.1.1 Distância, Ângulo e Produto Escalar

A expressão **espaço Euclidiano** é uma maneira formal de denominar o espaço físico com o qual estamos acostumados e que adotamos em todas as representações de pontos no espaço. Assim, uma linha reta é um espaço unidimensional, um plano é um espaço bidimensional e o espaço tridimensional é o que vemos à nossa volta.

Como já foi mencionado, a definição de uma medida de distância, também chamada de **métrica**, entre pontos em um espaço multidimensional de dados é muito importante na aplicação das técnicas de análise de dados multivariados. A distância e o ângulo são ambos quantidades escalares (números reais), definidos em termos de 2 pontos. A **distância** é o valor que quantifica a proximidade de um ponto (ou vetor) **a** a um ponto **b**, e o **ângulo** é o valor que quantifica quão rapidamente 2 vetores divergem a partir de uma origem comum. A representação gráfica destes

conceitos, para quaisquer dois pontos **a** e **b** é mostrada na figura A.1.

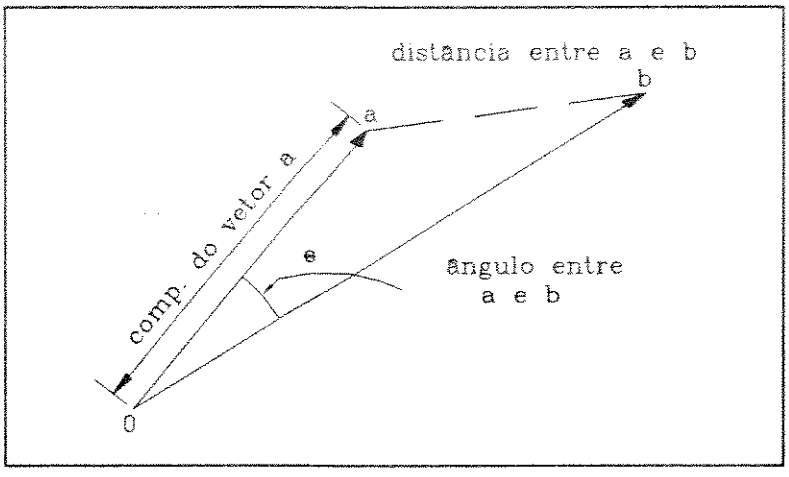

*Figura A.1: Distância e ângulo entre dois vetores **a** e **b**.*
(Adaptado de Greenacre, 1984).

Conhecendo-se as distâncias de **a** e de **b** até a origem (comprimento ou norma dos vetores **a** e **b**, respectivamente), e o ângulo entre eles, pode-se então encontrar a distância de **a** até **b**.

Ambos os conceitos de distância e ângulo estão englobados em um único conceito que é fundamental quando se trabalha no espaço multidimensional, conhecido como **produto escalar** ou **produto interno**.

Sejam **a** e **b**, dois pontos no espaço Euclidiano bidimensional, como mostra a figura A.2.

Os comprimentos ou normas dos vetores **a** e **b** são dados por:

$$\|a\| = (a_1^2 + a_2^2)^{1/2} \qquad e \qquad \|b\| = (b_1^2 + b_2^2)^{1/2} \tag{1}$$

A distância entre os pontos **a** e **b**, denotada por $d(\mathbf{a},\mathbf{b})$ é definida como:

$$d\,(a,b) = [(a_1-b_1)^2 + (a_2-b_2)^2]^{1/2} \tag{2}$$

O ângulo θ entre **a** e **b** tem coseno igual a:

$$\cos\theta = \frac{(a_1b_1 + a_2b_2)}{[(a_1^2 + a_2^2)(b_1^2 + b_2^2)]^{1/2}} \qquad (3)$$

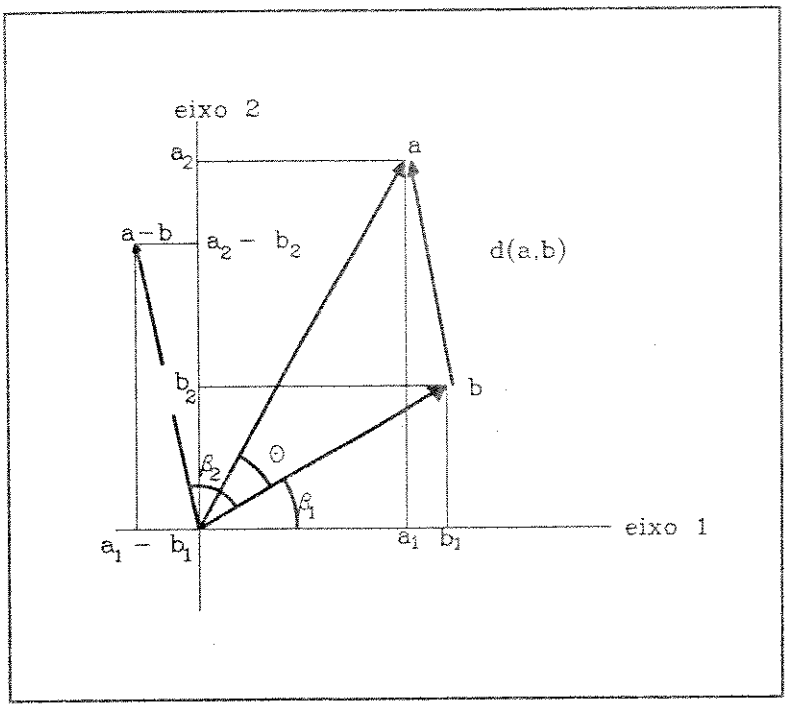

*Figura A.2: Pontos **a** e **b** no espaço bidimensional, vetor resultante da diferença e distância entre os pontos. (Adaptado de Greenacre, 1984).*

Todas as equações acima podem ser expressas em termos do produto escalar de **a** e **b**, denotado por <**a**,**b**>: <**b**,**a**> = $a_1b_1 + a_2b_2$.

Em notação vetorial a expressão $a_1b_1 + a_2b_2$ é simplesmente $\mathbf{a}^T\mathbf{b}$, a transposta de **a** multiplicada por **b**. As equações acima podem ser escritas na forma de produto escalar e em notação vetorial:

$$\|\boldsymbol{a}\| = \langle \boldsymbol{a},\boldsymbol{a} \rangle^{1/2} = (\boldsymbol{a}^T\boldsymbol{a})^{1/2} \qquad (4)$$

$$d(\boldsymbol{a},\boldsymbol{b}) = (\langle \boldsymbol{a} - \boldsymbol{b},\ \boldsymbol{a} - \boldsymbol{b} \rangle)^{1/2} = ((\boldsymbol{a} - \boldsymbol{b})^T(\boldsymbol{a} - \boldsymbol{b}))^{1/2} \qquad (5)$$

$$\cos\theta = \frac{\langle a,b\rangle}{(\langle a,a\rangle \langle b,b\rangle)^{1/2}} = \frac{a^T b}{a^T a b^T b} \tag{6}$$

A distância $d(\mathbf{a},\mathbf{b})$ pode ser avaliada em termos dos comprimentos de **a** e **b** e do cos θ da forma conhecida como regra do coseno:

$$d^2(a,b) = (a - b)^T (a - b)$$

$$d^2(a,b) = a^T a + b^T b - 2a^T b$$

$$d^2(a,b) = \|a\| + \|b\| - 2\|a\| \cdot \|b\| \cdot \cos\theta \tag{7}$$

É importante ressaltar que os resultados acima dependem inteiramente da perpendicularidade do sistema de coordenadas, ou seja, os vetores bases devem ser ortogonais. Dois vetores são **ortogonais** se o produto escalar entre eles é zero, ou seja, eles não têm componentes na direção um do outro. Se, além disto, os vetores têm comprimentos unitários, diz-se que eles são **ortonormais**. Então, eles constituem uma base ortonormal para o espaço Euclidiano. Todas as definições acima serão válidas se os vetores forem expressos em um sistema de coordenadas ortonormais.

Estendendo-se as definições acima para o espaço Euclidiano J-dimensional pode-se definir o produto escalar de quaisquer dois vetores $\mathbf{a} = [a_1 ... a_j]^T$ e $\mathbf{b} = [b_1 ... b_j]^T$ como sendo:

$$\langle a,b \rangle = \sum_j a_j b_j = a^T b \tag{8}$$

A.1.2 Espaço Euclidiano Ponderado

Na **análise de correspondência**, não se mede a distância entre dois pontos no espaço vetorial através da distância Euclidiana usual, mas através de uma métrica específica denominada de **distância Euclidiana ponderada**. Isto porque a distância entre dois pontos pode ser influenciada diretamente pelas unidades de

medida das variáveis. Uma maneira de evitar esse problema é dividir cada medida por seu respectivo desvio padrão, antes de calcular a distância Euclidiana. Essa forma padronizada de medida permanece a mesma para qualquer unidade escolhida originalmente.

Outra forma equivalente de contornar o problema é ponderar os eixos de coordenadas: os vetores originais são mantidos, mas, a definição do produto escalar contém um fator ponderador em cada termo, de modo que a distância entre dois pontos é dada por:

$$d^2\ (\boldsymbol{a}\ ,\ \boldsymbol{b}) = \frac{a_1 b_1}{s_1^2} + \frac{a_2 b_2}{s_2^2} = \boldsymbol{a}^T\ \boldsymbol{D}_{s^2}^{-1}\ \boldsymbol{b} \qquad (9)$$

onde: $\mathbf{D}_{s^2}^{-1}$ é a matriz diagonal que contém na sua diagonal principal o inverso das variâncias.

Geometricamente, os vetores **a** e **b** são desenhados nas suas unidades originais mas o produto escalar e, conseqüentemente, as distâncias e os comprimentos nesse espaço são calculados usando a equação 9, onde as unidades de **a** estão ponderadas com relação às unidades de **b**. A este espaço chama-se de **Euclidiano ponderado**, com os ponderadores sendo, nesse caso, o inverso das variâncias. Pode-se imaginar esse espaço como tendo sido esticado ao longo de seus eixos. Outra forma semelhante de ver o mesmo problema é imaginar que as unidades de medida foram mudadas em cada eixo, com as distâncias unitárias sobre cada um deles sendo inversamente proporcionais ao seu respectivo ponderador.

Em geral, o espaço Euclidiano ponderado é definido pelo produto escalar:

$$\boldsymbol{a}^T \boldsymbol{D}_q \boldsymbol{b} = \sum_j q_j a_j b_j \qquad (10)$$

onde: $q_1 ... q_j$ são números reais positivos que definem os ponderadores relativos associados às J dimensões. A distância ao quadrado entre dois pontos **a** e **b** neste espaço é, então, a soma **ponderada** das diferenças das coordenadas ao quadrado (equação 11). Esse tipo de função distância é geralmente denominada de **métrica diagonal**.

$$d^2(a,b) = (a-b)^T D_q (a-b) = \sum_j q_j (a_j - b_j)^2 \quad (11)$$

Algumas técnicas de análise de dados multidimensionais, como a análise de correspondência, lidam com vetores de freqüências relativas **u** como pontos em um espaço multidimensional. Tais vetores são conhecidos como **perfis**. Um dos exemplos mais comuns de distância Euclidiana ponderada entre vetores de freqüência é a **estatística quiquadrado** χ^2:

$$\chi^2 = (o - e)^T D_e^{-1} (o - e) \quad (12)$$

onde: **o** e **e** são freqüências observadas e esperadas, respectivamente, e D_e^{-1} é a matriz diagonal dos inversos das freqüências esperadas.

Seja **u** = (1/n)**o** e $\bar{\mathbf{u}}$ = (1/n)**e**, as freqüências relativas observadas e esperadas, respectivamente, onde n é a freqüência observada total, então a estatística χ^2 acima é dada por:

$$\chi^2 = n(u - \bar{u})^T D_{\bar{u}}^{-1} (u - \bar{u}) = \frac{n\sum_j (u_j - \bar{u}_j)^2}{\bar{u}_j} \quad (13)$$

A distância ao quadrado entre **u** e $\bar{\mathbf{u}}$ é dada por:

$$d^2(u, \bar{u}) = (u - \bar{u})^T D_{\bar{u}}^{-1} (u - \bar{u}) \quad (14)$$

A equação 14 é uma distância Euclidiana ponderada, onde os pesos são os inversos das freqüências relativas esperadas. Por causa da proporcionalidade dessa distância à estatística de χ^2, essa função distância é denominada de **distância quiquadrado**. O fator de proporcionalidade é o tamanho da amostra.

A.1.3 Associando Massa aos Vetores

Muitos métodos estatísticos comumente ponderam suas observações

por alguma razão. Para distinguir a ponderação dos eixos, citada acima, da ponderação dos pontos, utiliza-se o termo **massa** quando se referir à quantidade que pondera os pontos, e **peso** quando se referir à quantidade que pondera os eixos. No estudo da geometria de um conjunto de vetores, associar massas diferentes a cada um deles significa dar diferentes graus de importância às suas posições no espaço.

Quando se deseja identificar o subespaço de menor dimensão que melhor se ajusta a todos os pontos, e quando esses pontos têm massas diferentes, é claro que pontos com massa maior deverão estar mais próximos desse subespaço do que os pontos com massa menor.

O **centróide** ou **centro de gravidade** de um conjunto de pontos $\mathbf{a}_1,...\mathbf{a}_I$ com diferentes massas $w_1,...w_I$, é a média ponderada dos pontos:

$$\bar{\boldsymbol{a}} = \frac{\sum_i w_i \boldsymbol{a}_i}{\sum_i w_i} \tag{15}$$

Assim, $\bar{\mathbf{a}}$ tende para os pontos de maior massa.

A estatística χ^2 apresentada acima descreve a distância ao quadrado entre o vetor $\mathbf{u}$ de freqüências relativas observadas e o vetor $\bar{\mathbf{u}}$ de freqüências relativas esperadas, multiplicada pela freqüência observada total, n.

Suponha que existam *s* subpopulações de tamanho n_i, com i=1,...,s com perfil $\mathbf{u}_i = [u_{i1}\ u_{i2}\ ...\ u_{ik}]^T$. Então para cada subpopulação pode-se calcular a estatística χ_i^2:

$$\chi_i^2 = n_i\ (u_i - \bar{u})^T\ D_{\bar{u}}^{-1}\ (u_i - \bar{u}) \tag{16}$$

e para a soma das subpopulações tem-se:

$$\chi^2 = \sum_{i=1}^{s} \chi_i^2 \tag{17}$$

A estatística χ^2 acima pode ser descrita como a soma ponderada das distâncias ao quadrado entre $\mathbf{u}_i$ e $\bar{\mathbf{u}}$, com:

$$\bar{\boldsymbol{u}} = \frac{\sum_i n_i u_i}{\sum_i n_i} \tag{18}$$

Comparando a equação 18 com a equação 14, verifica-se que o vetor $\bar{\mathbf{u}}$ é o centróide dos vetores de freqüências relativas das subpopulações, sendo que cada um é ponderado pelo tamanho da subpopulação. O termo **perfil médio** é empregado para denominar $\bar{\mathbf{u}}$.

Introduzem-se agora as seguintes definições:

$$n = \sum_i n_i \tag{19}$$

$$w_i = \frac{n_i}{n} \tag{20}$$

$$d_i^2 = (\boldsymbol{u} - \bar{\boldsymbol{u}})^T \boldsymbol{D}_{\bar{\boldsymbol{u}}}^{-1} (\boldsymbol{u} - \bar{\boldsymbol{u}}) \tag{21}$$

A equação (21) define a distância ao quadrado entre $\bar{\mathbf{u}}_i$ e $\bar{\mathbf{u}}_i$, na métrica definida por $\mathbf{D}_{\bar{\mathbf{u}}}^{-1}$.

$$\textit{Inércia } (I) = \frac{\chi^2}{n} \tag{22}$$

A equação (22) define a inércia total do conjunto de I vetores perfis. O centróide $\bar{\mathbf{u}}$ e a inércia in(I) podem ser expressos como médias ponderadas:

$$\bar{\boldsymbol{u}} = \Sigma\, w_i\, u_i \tag{23}$$

$$\textit{Inércia } (I) = \Sigma\, w_i\, d_i^2 \tag{24}$$

Assim, o perfil médio $\bar{\mathbf{u}}$ é um vetor que representa o centróide dos perfis individuais, enquanto que a inércia é uma medida de quanto os perfis individuais estão espalhados ao redor do centróide.

A.1.4 Identificando Subespaços Ótimos

O objetivo, aqui, é encontrar o subespaço de menor dimensão que melhor contenha o conjunto de pontos, ou seja, o que mais se aproxime do conjunto de pontos.

Primeiramente, é necessário definir a proximidade de um conjunto de pontos a um subespaço dado. Intuitivamente, a distância entre um ponto e um subespaço dado é a menor distância entre esse ponto e todos os pontos contidos no subespaço. Assim, a proximidade de um conjunto de pontos a um subespaço pode ser definida como a média, ou a média ponderada, do correspondente conjunto das menores distâncias. Por razões de simplicidade algébrica, bem como por conveniências geométricas, a medida de proximidade é baseada nas distâncias ao quadrado.

A figura A.3 mostra uma nuvem de pontos em um espaço Euclidiano ponderado J-dimensional, com um subespaço S, de menor dimensão k, desenhado esquematicamente como um plano cortando o espaço. Para um ponto típico $\mathbf{u}_i$, $\hat{\mathbf{u}}_i$ representa o ponto no subespaço que está mais próximo de $\mathbf{u}_i$; a distância mínima entre eles é igual a d_i. Se $\mathbf{u}_i$ é ponderado pela massa w_i (i=1,...,I), então a definição de proximidade do conjunto inteiro de pontos ao subespaço S é dado por:

$$\Psi\ (S\ ;\ u_1 \ldots u_I) = \Sigma_i\ w_i\ d_i^2 \qquad (25)$$

onde:

$$d_i^2 = \|u_i - \hat{u}\|^2\ D_q = (u_i - \hat{u}_i)^T\ D_q\ (u_i - \hat{u}_i) \qquad (26)$$

e $\mathbf{D}_q$ é uma matriz diagonal de pesos (da dimensão) positivos.

A distância ao quadrado, d_i^2, depende do subespaço S, e o objetivo é encontrar o subespaço S^* que minimiza a função ψ na equação acima.

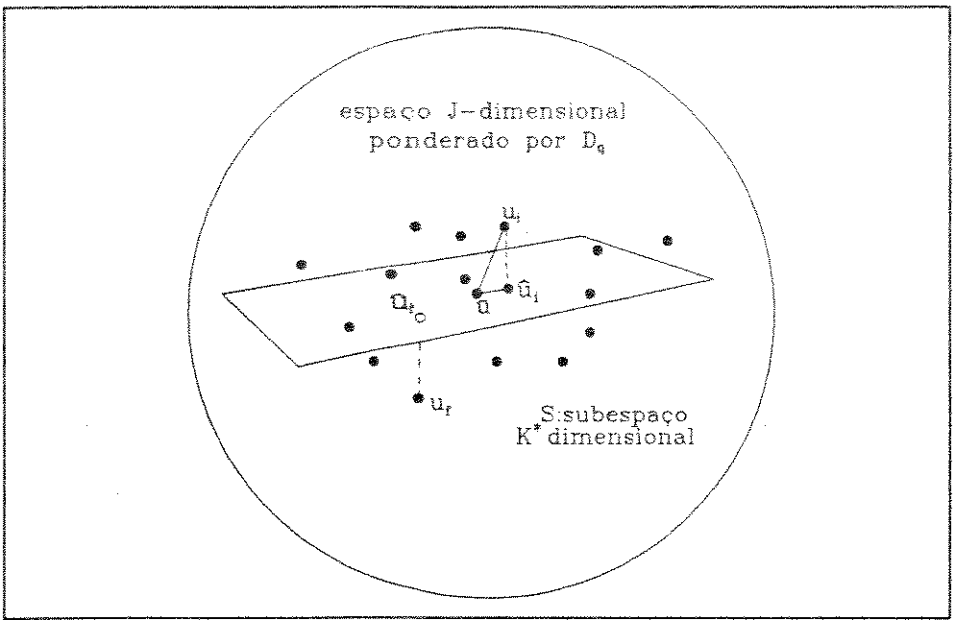

Figura A.3: Pontos em um espaço multidimensional e suas projeções em um subespaço, representado por um plano. (Greenacre, 1984).

Imagine agora um ponto s como um subespaço de dimensão zero. A função ψ torna-se:

$$\boldsymbol{\psi}\ (s\ ;\ u_1 \ldots u_I) = \Sigma\ w_i\ (u_i - s)^T\ \boldsymbol{D}_q\ (u_i - s) \qquad (27)$$

já que $\mathbf{u}_i$ é igual a s para todo i.

O centróide $\bar{\mathbf{u}}$ é o ponto que minimiza essa função (fazendo a derivada dessa função em relação aos elementos de s igual a zero). Assim, o centróide é, nesse sentido, o ponto mais próximo de todos os pontos $\mathbf{u}_1 \ldots \mathbf{u}_I$ dados. Pode-se mostrar que, na busca por um subespaço k-dimensional ótimo, necessita-se apenas considerar os subespaços S que contenham $\bar{\mathbf{u}}$ e, conseqüentemente, pode-se desenhar $\bar{\mathbf{u}}$ no subespaço candidato da figura A.3. Portanto, qualquer subespaço S que é ótimo no sentido de minimizar a função ψ, deve incluir o centróide.

A solução teórica completa para o problema de minimizar a função ψ para uma dimensão k, está embutida nos conceitos de decomposição em valores singulares e aproximação de uma matriz por matriz de posto menor, apresentada no capítulo 2.

PROGRAMA P.1

Categorização dos dados da descrição macroscópica da Fm Lagoa Feia

```
DATA TRAB1;                    /* ARQUIVO DE DADOS CATEGORIZADOS */
        SET IN.LAGFEIA;        /* ARQUIVO ORIGINAL DE DADOS */

IF      GRAN = 8 THEN NGRAN = 'SXO';
ELSE IF GRAN = 7 THEN NGRAN = 'GNL';
ELSE IF GRAN = 6 THEN NGRAN = 'MGO';
ELSE IF GRAN = 5 THEN NGRAN = 'GRO';
ELSE IF GRAN = 4 THEN NGRAN = 'MED';
ELSE IF GRAN = 3 THEN NGRAN = 'FNO';
ELSE IF GRAN = 2 THEN NGRAN = 'MFN';
ELSE IF GRAN = 1 THEN NGRAN = 'SLT';
ELSE IF GRAN = 0 THEN NGRAN = 'ARG';

IF      MTE EQ 00                       THEN TE = 'PURO';
ELSE IF MTE GT 00 AND MTE LE 10         THEN TE = 'ARSO';
ELSE IF MTE GT 10 AND MTE LE 50         THEN TE = 'MUIA';
ELSE IF MTE GT 50                       THEN TE = 'AREN';

IF CON EQ 0 THEN DO;
      AB      =  'SCO';
      QU      =  'SCO';
      NCON    =  'SCO';
      EMP     =  'SCO';
      NTAM    =  'SCO';
      TIPO    =  'SCO';
      NESPC   =  'SCO';
      ADJ     =  'SCO';
GOTO FIM;
END;

ELSE IF CON GT 00 AND CON LT 10 THEN NCON = 'MST';
ELSE IF CON GE 10 AND CON LT 30 THEN NCON = 'WST';
ELSE IF CON GE 30                THEN NCON = 'G/P';

IF      CON LT 50                       THEN EMP = 'FRX';
ELSE IF CON GE 50 AND CON LT 70 THEN EMP = 'NOR';
ELSE IF CON GE 70                       THEN EMP = 'DEN';

IF      ABE GE  0 AND ABE LT 50    THEN AB = 'FEC';
ELSE IF ABE GE 50                  THEN AB = 'ABE';

IF      QUE GE  0 AND QUE LT 50    THEN QU = 'INT';
ELSE IF QUE GE 50                  THEN QU = 'QUE';

TAM = MAX (AA, BB, CC, DD, EE, FF, GG);
    IF TAM = GG THEN DO;
          IF GG GT 00 AND GG LT 10 THEN NTAM = 'CRUGG';
          ELSE IF GG GE 10         THEN NTAM = 'GG';
    END;
    IF TAM = FF THEN DO;
          IF FF GT 00 AND FF LT 10   THEN NTAM = 'CRUFF';
          ELSE IF FF GE 10           THEN NTAM = 'FF';
    END;
    IF TAM = EE THEN DO;
          IF EE GT 00 AND EE LT 10   THEN NTAM = 'CRUEE';
```

```
            ELSE IF EE GE 10              THEN NTAM = 'EE';
    END;
    IF TAM = DD THEN DO;
            IF DD GT 00 AND DD LT 10 THEN NTAM = 'CRUDD';
            ELSE IF DD GE 10              THEN NTAM = 'DD';
    END;
    IF TAM = CC THEN DO;
            IF CC GT 00 AND CC LT 10  THEN NTAM = 'CRUCC';
            ELSE IF CC GE 10              THEN NTAM = 'CC';
    END;
    IF TAM = BB THEN DO;
            IF BB GT 00 AND BB LT 10   THEN NTAM = 'CRUBB';
            ELSE IF BB GE 10              THEN NTAM = 'BB';
    END;
    IF TAM = AA THEN DO;
            IF AA GT 00 AND AA LT 10 THEN NTAM = 'CRUAA';
            ELSE IF AA GE 10              THEN NTAM = 'AA';
    END;

IF      GG LT 20                          THEN TIPO = 'C  ';
ELSE IF GG GE 20 AND GG LT 50     THEN TIPO = 'CC ';
ELSE IF GG GE 50 AND GG LT 70     THEN TIPO = 'CMC';
ELSE IF GG GE 70                          THEN TIPO = 'CRE';
IF GG EQ 100 THEN GOTO FIM;

TAM = MAX (AA, BB, CC, DD, EE, FF);
IF TAM = FF THEN DO;
    Fcor = FF/(100 - GG) * 100;
    IF      Fcor GE 30 THEN ADJ = 'mf';
    ELSE IF            ADJ = '  ';
END;
IF TAM = EE THEN DO;
    Ecor = EE/(100 - GG) * 100;
    IF     Ecor GE 30 THEN ADJ = 'f ';
    ELSE IF            ADJ = '  ';
END;
IF TAM = DD THEN DO;
    Dcor = DD/(100 - GG) * 100;
    IF     Dcor GE 30 THEN ADJ = 'm ';
    ELSE IF            ADJ = '  ';
END;
IF TAM = CC THEN DO;
    Ccor = CC/(100 - GG) * 100;
    IF     Ccor GE 30 THEN ADJ = 'g ';
    ELSE IF            ADJ = '  ';
END;
IF TAM =  BB THEN DO;
    Bcor = BB/(100 - GG) * 100;
    IF Bcor GE 30 THEN ADJ = 'g ';
    ELSE IF          ADJ = '  ';
END;
IF TAM =  AA THEN DO;
    Acor = AA/(100 - GG) * 100;
    IF Acor GE 30 THEN ADJ = 'mg';
    ELSE IF          ADJ = '  ';
END;

ESPC = MAX (A, B, C, D, E, F, G);
    IF ESPC = G THEN DO;
        IF       G GE 30 THEN NESPC = 'G ';
        ELSE IF  NESPC = '  ';
    END;
    IF ESPC = F THEN DO;
        IF       F GE 30 THEN NESPC = 'F ';
        ELSE IF  NESPC = '  ';
```

```
    END;
    IF ESPC = E THEN DO;
        IF      E GE 30 THEN NESPC = 'E ';
        ELSE IF  NESPC = '  ';
    END;
    IF ESPC = D THEN DO;
        IF      D GE 30 THEN NESPC = 'D ';
        ELSE IF  NESPC = '  ';
    END;
    IF ESPC = C THEN DO;
        IF      C GE 30 THEN NESPC = 'C ';
        ELSE IF  NESPC = '  ';
    END;
    IF ESPC = B THEN DO;
        IF      B GE 30 THEN NESPC = 'B ';
        ELSE IF  NESPC = '  ';
    END;
    IF ESPC = A THEN DO;
        IF      A GE 30 THEN NESPC = 'A ';
        ELSE IF  NESPC = '  ';
    END;
FIM;

IF CON EQ 70 AND MTE EQ 0 AND ABE EQ '.' AND QUE EQ '.' THEN PF = 'GST';
IF CON EQ '.' AND MTE EQ '.' AND ABE EQ '.' AND QUE EQ '.' THEN PF = 'REC';
IF CON EQ 0 AND MAT EQ 100 AND MTE EQ '.' AND ABE EQ '.' AND QUE EQ '.' THEN PF = 'SLX';
IF MAT EQ '.' AND CON EQ '.' AND ABE EQ '.' AND QUE EQ '.' THEN PF = 'BIO';

KEEP POCO COD PROF NGRAN ESP NCON TE AB QU NTAM TIPO ADJ NESPC EMP PF;
PROC PRINT;
RUN;
```

PROGRAMA P.2

Definição de Fácies Sedimentares

```
DATA FAC;
   SET IN.TRAB1;

IF NTAM = 'C  ' THEN DO;

   IF ADJ EQ '  ' THEN DO;
      IF       TE = 'PURO' AND EMP = 'FRX' THEN FAC = 'CPF    ';
      ELSE IF TE = 'PURO' AND EMP = 'NOR' THEN FAC = 'CPN    ';
      ELSE IF TE = 'PURO' AND EMP = 'DEN' THEN FAC = 'CPD    ';
      ELSE IF TE = 'ARSO' AND EMP = 'FRX' THEN FAC = 'CAF    ';
      ELSE IF TE = 'ARSO' AND EMP = 'NOR' THEN FAC = 'CAN    ';
      ELSE IF TE = 'ARSO' AND EMP = 'DEN' THEN FAC = 'CAD    ';
      ELSE IF TE = 'MUIA' AND EMP = 'FRX' THEN FAC = 'CMAF   ';
      ELSE IF TE = 'MUIA' AND EMP = 'NOR' THEN FAC = 'CMAN   ';
      ELSE IF TE = 'MUIA' AND EMP = 'DEN' THEN FAC = 'CMAD   ';
   END;

   ELSE IF ADJ = 'mf'  THEN DO;
      IF       TE = 'PURO' AND EMP = 'FRX' THEN FAC = 'CPFmf  ';
      ELSE IF TE = 'PURO' AND EMP = 'NOR' THEN FAC = 'CPNmf  ';
      ELSE IF TE = 'PURO' AND EMP = 'DEN' THEN FAC = 'CPDmf  ';
      ELSE IF TE = 'ARSO' AND EMP = 'FRX' THEN FAC = 'CAFmf  ';
      ELSE IF TE = 'ARSO' AND EMP = 'NOR' THEN FAC = 'CANmf  ';
      ELSE IF TE = 'ARSO' AND EMP = 'DEN' THEN FAC = 'CADmf  ';
      ELSE IF TE = 'MUIA' AND EMP = 'FRX' THEN FAC = 'CMAFmf ';
      ELSE IF TE = 'MUIA' AND EMP = 'NOR' THEN FAC = 'CMANmf ';
      ELSE IF TE = 'MUIA' AND EMP = 'DEN' THEN FAC = 'CMADmf ';
   END;

   ELSE IF ADJ = 'f ' THEN DO;
      IF       TE = 'PURO' AND EMP = 'FRX' THEN FAC = 'CPFf   ';
      ELSE IF TE = 'PURO' AND EMP = 'NOR' THEN FAC = 'CPNf   ';
      ELSE IF TE = 'PURO' AND EMP = 'DEN' THEN FAC = 'CPDf   ';
      ELSE IF TE = 'ARSO' AND EMP = 'FRX' THEN FAC = 'CAFf   ';
      ELSE IF TE = 'ARSO' AND EMP = 'NOR' THEN FAC = 'CANf   ';
      ELSE IF TE = 'ARSO' AND EMP = 'DEN' THEN FAC = 'CADf   ';
      ELSE IF TE = 'MUIA' AND EMP = 'FRX' THEN FAC = 'CMAFf  ';
      ELSE IF TE = 'MUIA' AND EMP = 'NOR' THEN FAC = 'CMANf  ';
      ELSE IF TE = 'MUIA' AND EMP = 'DEN' THEN FAC = 'CMADf  ';
   END;

   ELSE IF ADJ = 'm ' THEN DO;
      IF       TE = 'PURO' AND EMP = 'FRX' THEN FAC = 'CPFm   ';
      ELSE IF TE = 'PURO' AND EMP = 'NOR' THEN FAC = 'CPNm   ';
      ELSE IF TE = 'PURO' AND EMP = 'DEN' THEN FAC = 'CPDm   ';
      ELSE IF TE = 'ARSO' AND EMP = 'FRX' THEN FAC = 'CAFm   ';
      ELSE IF TE = 'ARSO' AND EMP = 'NOR' THEN FAC = 'CANm   ';
      ELSE IF TE = 'ARSO' AND EMP = 'DEN' THEN FAC = 'CADm   ';
      ELSE IF TE = 'MUIA' AND EMP = 'FRX' THEN FAC = 'CMAFm  ';
      ELSE IF TE = 'MUIA' AND EMP = 'NOR' THEN FAC = 'CMANm  ';
      ELSE IF TE = 'MUIA' AND EMP = 'DEN' THEN FAC = 'CMADm  ';
   END;

   ELSE IF ADJ = 'g ' THEN DO;
      IF       TE = 'PURO' AND EMP = 'FRX' THEN FAC = 'CPFg   ';
      ELSE IF TE = 'PURO' AND EMP = 'NOR' THEN FAC = 'CPNg   ';
```

```
        ELSE IF TE = 'PURO' AND EMP = 'DEN' THEN FAC = 'CPDg    ';
        ELSE IF TE = 'ARSO' AND EMP = 'FRX' THEN FAC = 'CAFg    ';
        ELSE IF TE = 'ARSO' AND EMP = 'NOR' THEN FAC = 'CANg    ';
        ELSE IF TE = 'ARSO' AND EMP = 'DEN' THEN FAC = 'CADg    ';
        ELSE IF TE = 'MUIA' AND EMP = 'FRX' THEN FAC = 'CMAFg   ';
        ELSE IF TE = 'MUIA' AND EMP = 'NOR' THEN FAC = 'CMANg   ';
        ELSE IF TE = 'MUIA' AND EMP = 'DEN' THEN FAC = 'CMADg   ';
    END;

    ELSE IF ADJ = 'mg' THEN DO;
        IF      TE = 'PURO' AND EMP = 'FRX' THEN FAC = 'CPFmg   ';
        ELSE IF TE = 'PURO' AND EMP = 'NOR' THEN FAC = 'CPNmg   ';
        ELSE IF TE = 'PURO' AND EMP = 'DEN' THEN FAC = 'CPDmg   ';
        ELSE IF TE = 'ARSO' AND EMP = 'FRX' THEN FAC = 'CAFmg   ';
        ELSE IF TE = 'ARSO' AND EMP = 'NOR' THEN FAC = 'CANmg   ';
        ELSE IF TE = 'ARSO' AND EMP = 'DEN' THEN FAC = 'CADmg   ';
        ELSE IF TE = 'MUIA' AND EMP = 'FRX' THEN FAC = 'CMAFmg  ';
        ELSE IF TE = 'MUIA' AND EMP = 'NOR' THEN FAC = 'CMANmg  ';
        ELSE IF TE = 'MUIA' AND EMP = 'DEN' THEN FAC = 'CMADmg  ';
    END;
END;

IF NTAM = 'CC' THEN DO;

    IF ADJ = '  ' THEN DO;
        IF      TE = 'PURO' AND EMP = 'FRX' THEN FAC = 'CCPF    ';
        ELSE IF TE = 'PURO' AND EMP = 'NOR' THEN FAC = 'CCPN    ';
        ELSE IF TE = 'PURO' AND EMP = 'DEN' THEN FAC = 'CCPD    ';
        ELSE IF TE = 'ARSO' AND EMP = 'FRX' THEN FAC = 'CCAF    ';
        ELSE IF TE = 'ARSO' AND EMP = 'NOR' THEN FAC = 'CCAN    ';
        ELSE IF TE = 'ARSO' AND EMP = 'DEN' THEN FAC = 'CCAD    ';
        ELSE IF TE = 'MUIA' AND EMP = 'FRX' THEN FAC = 'CCMAF   ';
        ELSE IF TE = 'MUIA' AND EMP = 'NOR' THEN FAC = 'CCMAN   ';
        ELSE IF TE = 'MUIA' AND EMP = 'DEN' THEN FAC = 'CCMAD   ';
    END;

    ELSE IF ADJ = 'mf' THEN DO;
        IF      TE = 'PURO' AND EMP = 'FRX' THEN FAC = 'CCPFmf  ';
        ELSE IF TE = 'PURO' AND EMP = 'NOR' THEN FAC = 'CCPNmf  ';
        ELSE IF TE = 'PURO' AND EMP = 'DEN' THEN FAC = 'CCPDmf  ';
        ELSE IF TE = 'ARSO' AND EMP = 'FRX' THEN FAC = 'CCAFmf  ';
        ELSE IF TE = 'ARSO' AND EMP = 'NOR' THEN FAC = 'CCANmf  ';
        ELSE IF TE = 'ARSO' AND EMP = 'DEN' THEN FAC = 'CCADmf  ';
        ELSE IF TE = 'MUIA' AND EMP = 'FRX' THEN FAC = 'CCMAFmf ';
        ELSE IF TE = 'MUIA' AND EMP = 'NOR' THEN FAC = 'CCMANmf ';
        ELSE IF TE = 'MUIA' AND EMP = 'DEN' THEN FAC = 'CCMADmf ';
    END;

    ELSE IF ADJ = 'f ' THEN DO;
        IF      TE = 'PURO' AND EMP = 'FRX' THEN FAC = 'CCPFf   ';
        ELSE IF TE = 'PURO' AND EMP = 'NOR' THEN FAC = 'CCPNf   ';
        ELSE IF TE = 'PURO' AND EMP = 'DEN' THEN FAC = 'CCPDf   ';
        ELSE IF TE = 'ARSO' AND EMP = 'FRX' THEN FAC = 'CCAFf   ';
        ELSE IF TE = 'ARSO' AND EMP = 'NOR' THEN FAC = 'CCANf   ';
        ELSE IF TE = 'ARSO' AND EMP = 'DEN' THEN FAC = 'CCADf   ';
        ELSE IF TE = 'MUIA' AND EMP = 'FRX' THEN FAC = 'CCMAFf  ';
        ELSE IF TE = 'MUIA' AND EMP = 'NOR' THEN FAC = 'CCMANf  ';
        ELSE IF TE = 'MUIA' AND EMP = 'DEN' THEN FAC = 'CCMADf  ';
    END;

    IF ADJ = 'm ' THEN DO;
        IF      TE = 'PURO' AND EMP = 'FRX' THEN FAC = 'CCPFm   ';
        ELSE IF TE = 'PURO' AND EMP = 'NOR' THEN FAC = 'CCPNm   ';
        ELSE IF TE = 'PURO' AND EMP = 'DEN' THEN FAC = 'CCPDm   ';
        ELSE IF TE = 'ARSO' AND EMP = 'FRX' THEN FAC = 'CCAFm   ';
        ELSE IF TE = 'ARSO' AND EMP = 'NOR' THEN FAC = 'CCANm   ';
```

```
        ELSE IF TE = 'ARSO' AND EMP = 'DEN' THEN FAC = 'CCADm  ';
        ELSE IF TE = 'MUIA' AND EMP = 'FRX' THEN FAC = 'CCMAFm ';
        ELSE IF TE = 'MUIA' AND EMP = 'NOR' THEN FAC = 'CCMANm ';
        ELSE IF TE = 'MUIA' AND EMP = 'DEN' THEN FAC = 'CCMADm ';
    END;

    ELSE IF ADJ = 'g ' THEN DO;
        IF      TE = 'PURO' AND EMP = 'FRX' THEN FAC = 'CCPFg  ';
        ELSE IF TE = 'PURO' AND EMP = 'NOR' THEN FAC = 'CCPNg  ';
        ELSE IF TE = 'PURO' AND EMP = 'DEN' THEN FAC = 'CCPDg  ';
        ELSE IF TE = 'ARSO' AND EMP = 'FRX' THEN FAC = 'CCAFg  ';
        ELSE IF TE = 'ARSO' AND EMP = 'NOR' THEN FAC = 'CCANg  ';
        ELSE IF TE = 'ARSO' AND EMP = 'DEN' THEN FAC = 'CCADg  ';
        ELSE IF TE = 'MUIA' AND EMP = 'FRX' THEN FAC = 'CCMAFg ';
        ELSE IF TE = 'MUIA' AND EMP = 'NOR' THEN FAC = 'CCMANg ';
        ELSE IF TE = 'MUIA' AND EMP = 'DEN' THEN FAC = 'CCMADg ';
    END;

    ELSE IF ADJ = 'mg' THEN DO;
        IF      TE = 'PURO' AND EMP = 'FRX' THEN FAC = 'CCPFmg ';
        ELSE IF TE = 'PURO' AND EMP = 'NOR' THEN FAC = 'CCPNmg ';
        ELSE IF TE = 'PURO' AND EMP = 'DEN' THEN FAC = 'CCPDmg ';
        ELSE IF TE = 'ARSO' AND EMP = 'FRX' THEN FAC = 'CCAFmg ';
        ELSE IF TE = 'ARSO' AND EMP = 'NOR' THEN FAC = 'CCANmg ';
        ELSE IF TE = 'ARSO' AND EMP = 'DEN' THEN FAC = 'CCADmg ';
        ELSE IF TE = 'MUIA' AND EMP = 'FRX' THEN FAC = 'CCMAFmg';
        ELSE IF TE = 'MUIA' AND EMP = 'NOR' THEN FAC = 'CCMANmg';
        ELSE IF TE = 'MUIA' AND EMP = 'DEN' THEN FAC = 'CCMADmg';
    END;
END;

IF NTAM = 'CMC' THEN DO;

    IF ADJ = '  ' THEN DO;
        IF      TE = 'PURO' AND EMP = 'FRX' THEN FAC = 'CMCPF  ';
        ELSE IF TE = 'PURO' AND EMP = 'NOR' THEN FAC = 'CMCPN  ';
        ELSE IF TE = 'PURO' AND EMP = 'DEN' THEN FAC = 'CMCPD  ';
        ELSE IF TE = 'ARSO' AND EMP = 'FRX' THEN FAC = 'CMCAF  ';
        ELSE IF TE = 'ARSO' AND EMP = 'NOR' THEN FAC = 'CMCAN  ';
        ELSE IF TE = 'ARSO' AND EMP = 'DEN' THEN FAC = 'CMCAD  ';
        ELSE IF TE = 'MUIA' AND EMP = 'FRX' THEN FAC = 'CMCMAF ';
        ELSE IF TE = 'MUIA' AND EMP = 'NOR' THEN FAC = 'CMCMAN ';
        ELSE IF TE = 'MUIA' AND EMP = 'DEN' THEN FAC = 'CMCMAD ';
    END;

    ELSE IF ADJ = 'mf' THEN DO;
        IF      TE = 'PURO' AND EMP = 'FRX' THEN FAC = 'CMCPFmf';
        ELSE IF TE = 'PURO' AND EMP = 'NOR' THEN FAC = 'CMCPNmf';
        ELSE IF TE = 'PURO' AND EMP = 'DEN' THEN FAC = 'CMCPDmf';
        ELSE IF TE = 'ARSO' AND EMP = 'FRX' THEN FAC = 'CMCAFmf';
        ELSE IF TE = 'ARSO' AND EMP = 'NOR' THEN FAC = 'CMCANmf';
        ELSE IF TE = 'ARSO' AND EMP = 'DEN' THEN FAC = 'CMCADmf';
        ELSE IF TE = 'MUIA' AND EMP = 'FRX' THEN FAC = 'CMCMAFmf';
        ELSE IF TE = 'MUIA' AND EMP = 'NOR' THEN FAC = 'CMCMANmf';
        ELSE IF TE = 'MUIA' AND EMP = 'DEN' THEN FAC = 'CMCMADmf';
    END;

    ELSE IF ADJ = 'f ' THEN DO;
        IF      TE = 'PURO' AND EMP = 'FRX' THEN FAC = 'CMCPFf ';
        ELSE IF TE = 'PURO' AND EMP = 'NOR' THEN FAC = 'CMCPNf ';
        ELSE IF TE = 'PURO' AND EMP = 'DEN' THEN FAC = 'CMCPDf ';
        ELSE IF TE = 'ARSO' AND EMP = 'FRX' THEN FAC = 'CMCAFf ';
        ELSE IF TE = 'ARSO' AND EMP = 'NOR' THEN FAC = 'CMCANf ';
        ELSE IF TE = 'ARSO' AND EMP = 'DEN' THEN FAC = 'CMCADf ';
        ELSE IF TE = 'MUIA' AND EMP = 'FRX' THEN FAC = 'CMCMAFf ';
        ELSE IF TE = 'MUIA' AND EMP = 'NOR' THEN FAC = 'CMCMANf ';
```

```
        ELSE IF TE = 'MUIA' AND EMP = 'DEN' THEN FAC = 'CMCMADf ';
    END;

    ELSE IF ADJ = 'm ' THEN DO;
        IF      TE = 'PURO' AND EMP = 'FRX' THEN FAC = 'CMCPFm  ';
        ELSE IF TE = 'PURO' AND EMP = 'NOR' THEN FAC = 'CMCPNm  ';
        ELSE IF TE = 'PURO' AND EMP = 'DEN' THEN FAC = 'CMCPDm  ';
        ELSE IF TE = 'ARSO' AND EMP = 'FRX' THEN FAC = 'CMCAFm  ';
        ELSE IF TE = 'ARSO' AND EMP = 'NOR' THEN FAC = 'CMCANm  ';
        ELSE IF TE = 'ARSO' AND EMP = 'DEN' THEN FAC = 'CMCADm  ';
        ELSE IF TE = 'MUIA' AND EMP = 'FRX' THEN FAC = 'CMCMAFm ';
        ELSE IF TE = 'MUIA' AND EMP = 'NOR' THEN FAC = 'CMCMANm ';
        ELSE IF TE = 'MUIA' AND EMP = 'DEN' THEN FAC = 'CMCMADm ';
    END;

    ELSE IF ADJ = 'g ' THEN DO;
        IF      TE = 'PURO' AND EMP = 'FRX' THEN FAC = 'CMCPFg  ';
        ELSE IF TE = 'PURO' AND EMP = 'NOR' THEN FAC = 'CMCPNg  ';
        ELSE IF TE = 'PURO' AND EMP = 'DEN' THEN FAC = 'CMCPDg  ';
        ELSE IF TE = 'ARSO' AND EMP = 'FRX' THEN FAC = 'CMCAFg  ';
        ELSE IF TE = 'ARSO' AND EMP = 'NOR' THEN FAC = 'CMCANg  ';
        ELSE IF TE = 'ARSO' AND EMP = 'DEN' THEN FAC = 'CMCADg  ';
        ELSE IF TE = 'MUIA' AND EMP = 'FRX' THEN FAC = 'CMCMAFg ';
        ELSE IF TE = 'MUIA' AND EMP = 'NOR' THEN FAC = 'CMCMANg ';
        ELSE IF TE = 'MUIA' AND EMP = 'DEN' THEN FAC = 'CMCMADg ';
    END;

    ELSE IF ADJ = 'mg' THEN DO;
        IF      TE = 'PURO' AND EMP = 'FRX' THEN FAC = 'CMCPFmg ';
        ELSE IF TE = 'PURO' AND EMP = 'NOR' THEN FAC = 'CMCPNmg ';
        ELSE IF TE = 'PURO' AND EMP = 'DEN' THEN FAC = 'CMCPDmg ';
        ELSE IF TE = 'ARSO' AND EMP = 'FRX' THEN FAC = 'CMCAFmg ';
        ELSE IF TE = 'ARSO' AND EMP = 'NOR' THEN FAC = 'CMCANmg ';
        ELSE IF TE = 'ARSO' AND EMP = 'DEN' THEN FAC = 'CMCADmg ';
        ELSE IF TE = 'MUIA' AND EMP = 'FRX' THEN FAC = 'CMCMAFmg';
        ELSE IF TE = 'MUIA' AND EMP = 'NOR' THEN FAC = 'CMCMANmg';
        ELSE IF TE = 'MUIA' AND EMP = 'DEN' THEN FAC = 'CMCMADmg';
    END;
END;

IF NTAM = 'CRE' THEN DO;
    IF  TE = 'PURO' AND EMP = 'FRX' THEN FAC = 'CREPF ';
    ELSE IF TE = 'PURO' AND EMP = 'NOR' THEN FAC = 'CREPN ';
    ELSE IF TE = 'PURO' AND EMP = 'DEN' THEN FAC = 'CREPD ';
    ELSE IF TE = 'ARSO' AND EMP = 'FRX' THEN FAC = 'CREAF ';
    ELSE IF TE = 'ARSO' AND EMP = 'NOR' THEN FAC = 'CREAN ';
    ELSE IF TE = 'ARSO' AND EMP = 'DEN' THEN FAC = 'CREAD ';
    ELSE IF TE = 'MUIA' AND EMP = 'FRX' THEN FAC = 'CREMAF';
    ELSE IF TE = 'MUIA' AND EMP = 'NOR' THEN FAC = 'CREMAN';
    ELSE IF TE = 'MUIA' AND EMP = 'DEN' THEN FAC = 'CREMAD';
END;

IF (NCON EQ 'SCO' OR NCON EQ 'MST' OR NCON EQ 'WST') AND TE NE 'AREN' THEN FAC = 'CLU';
ELSE IF NCON EQ 'SCO' AND TE EQ 'AREN' THEN FAC = 'ARN   ';
ELSE IF NCON NE 'SCO' AND TE EQ 'AREN' THEN FAC = 'ARNCON ';
IF     PF = 'GST' THEN FAC = 'GST    ';
ELSE IF PF = 'BIO' THEN FAC = 'BIO    ';
ELSE IF PF = 'REC' THEN FAC = 'REC    ';
ELSE IF PF = 'SLX' THEN FAC = 'SLX    ';

KEEP POCO COD PROF NGRAN ESP NCON TE AB QU NTAM ADJ NESPC EMP FAC;
PROC PRINT;
RUN;
```

PROGRAMA P.3

Fácies Sedimentares Simplificadas

```
DATA FAC;
    SET IN.TRAB1;

IF   NTAM = 'C  ' THEN  FAC = 'C   ';
ELSE IF NTAM = 'CC ' THEN  FAC = 'CC  ';
ELSE IF NTAM = 'CMC' THEN  FAC = 'CMC ';
ELSE IF NTAM = 'CRE' THEN  FAC = 'CRE ';
ELSE IF (NCON EQ 'SCO' OR NCON EQ 'MST' OR NCON EQ 'WST') AND TE NE 'AREN' THEN FAC = 'CLU';
ELSE IF NCON EQ 'SCO' AND TE EQ 'AREN' THEN FAC = 'ARN    ';
ELSE IF PF = 'GST' THEN FAC = 'GST     ';
ELSE IF PF = 'BIO' THEN FAC = 'BIO     ';
ELSE IF PF = 'REC' THEN FAC = 'REC     ';
ELSE IF PF = 'SLX' THEN FAC = 'SLX     ';

KEEP POCO COD PROF NGRAN ESP NCON TE AB QU NTAM ADJ NESPC EMP FAC;
RUN;
```

PROGRAMA P.4

Análise de Correspondência no Modulo STAT do SAS

```
*** PROC CORRESP                : realiza análise de correspondência simples ***
*** ALL                         : imprime os valores das estatísticas esperadas e observadas ***
*** DATA                        : especifica o arquivo de entrada de dados ***
*** OUTC                        : especifica o arquivo de saída de dados ***
*** TABLES                      : especifica as variáveis a serem analisadas ***

PROC CORRESP ALL DATA=TRAB1 OUTC=COOR;
TABLES NTAM, NMTE;
RUN;

*** PROC CORRESP MCA            : realiza análise de correspondência múltipla ***
*** OBSERVED                    : imprime a tabela de contingência das freqüências observadas ***

PROC CORRESP MCA OBSERVED DATA=IN.TRAB1 OUTC=COOR;
TABLES NGRAN NMTE NTAM NESPC NEMP;
RUN;

*** PREPARAÇÃO DOS RESULTADOS DA ANÁLISE DE CORRESPONDÊNCIA PARA ***
*** CONSTRUÇÃO DO GRÁFICO - USO DO PROC GPLOT E ANNOTATE FACILITY ***

DATA COOR;
    SET COOR;
    LABEL          X = 'DIMENSÃO 1'
        Y = 'DIMENSÃO 2';
    X = DIM1;      /* variável x */
    Y = DIM2;      /* variável y */
    XSYS = '2';    /* especifica área entre x máx e x mín */
    YSYS = '2';    /* especifica área entre y máx e y mín */
    TEXT =_NAME_;          /* posição do texto no gráfico */
    SIZE = 0.7;    /* tamanho do texto no gráfico */
    KEEP X Y XSYS YSYS TEXT SIZE;

GOPTIONS DEVICE=VGA;
AXIS1 V=(F=ITALIC) LABEL=(F=ITALIC H=0.7 'DIMENSAO 1')
LENGTH=14 CM ORDER=-2 TO 2 BY 1;
AXIS2 V=(F=ITALIC) LABEL=(F=ITALIC H=0.7 'DIMENSAO 2')
LENGTH=14 CM ORDER=-2 TO 2 BY 1;

PROC GPLOT DATA=COOR UNIFORM;
SYMBOL1 V=NONE;
 PLOT Y*X=1 / ANNOTATE=COOR FRAME CFRAME=GRAY
          HAXIS=AXIS1 VAXIS=AXIS2
          HREF=0 VREF=0;

GOPTIONS DEVICE=CGM;
FILENAME GRAFOUT 'GRAFOUT.GSF';
GOPTIONS GSFNAME=GRAFOUT GSFMODE=REPLACE NODISPLAY;
AXIS1 V=(F=ITALIC) LABEL=(F=ITALIC H=0.7 'DIMENSAO 1')
LENGTH=8 CM ORDER=-2 TO 2 BY 1;
AXIS2 V=(F=ITALIC) LABEL=(F=ITALIC H=0.7 'DIMENSAO 2')
LENGTH=8 CM ORDER=-2 TO 2 BY 1;
PROC GPLOT DATA=COOR UNIFORM;
SYMBOL1 V=NONE;
 PLOT Y*X=1 / ANNOTATE=COOR FRAME
          HAXIS=AXIS1 VAXIS=AXIS2
          HREF=0 VREF=0;
RUN;
```

PROGRAMA P.5

Análise de Correspondência no SAS/IML

```
title 'Análise de correspondência da tabela de contingência N';
data lagfeia;
input C1 C2 C3 C4;                    /* entrada de dados */
cards;
2   1   5   0
0   4   1   1
0   9   0   1
14  20  27  15
16  40  30  22
37 101  43  84
10  92  42  85
426  6  42 104
;
run;

proc iml;
use lagfeia;
read all var {C1 C2 C3 C4} into N;
print N;                               /* tabela de contingência */
UM = J(4,1,1);                         /* matriz de uns */
MU = J(8,1,1);                         /* matriz de uns */
F = T(MU)*N*UM;                        /* numero total de elementos de N */
print F;
P = (1/F)*N;                           /* matriz de correspondência */
print P;
R = P*UM;                              /* somatório linhas de P=massa linhas */
print R;
C = T(P)*MU;                           /* somatório colunas de P=massa colunas*/
print C;
Dr = diag(R);                          /* matriz diagonal dos elementos de r */
Dc = diag(C);                          /* matriz diagonal dos elementos de c */
print Dr;
print Dc;
RR = inv(Dr)*P;                        /* matriz perfil-linha  */
print RR;
CC = inv(Dc)* t(P);                    /* matriz perfil-coluna */
print CC;
CENR = t(RR)*R;                        /* centróide das linhas */
print CENR;
CENC = t(CC)*C;                        /* centróide das colunas */
print CENC;

INI = trace (Dr*(RR-MU*t(CENR)) * inv(Dc) * t(RR-MU*t(CENR)));
print INI;                             /* inércia das linhas */
INJ = trace (Dc*(CC-UM*t(CENC)) * inv(Dr) * t(CC-UM*t(CENC)));
print INJ;                             /* inércia das colunas */
```

```
INT = trace (inv(Dr) * (P-CENC*t(CENR)) * inv(Dc) * t(P-CENC*t(CENR)));
print INT;                          /* inércia total */


A = P - CENC*t(CENR);               /* matriz centrada */
print A;

/* Transformação da matriz A */
B = inv(sqrt(Dr)) * A * inv(sqrt(Dc));
print B;

/* Decomposição Ordinária da matriz transformada B */
call svd (u, s, v, B);
print u, s, v;

/* calculando a matriz de posto 2 */
rank = 2;
U1 = U[,1];
U2 = U[,2];
UU = U[,1:2];
print UU;

V1 = V[,1];
V2 = V[,2];
VV = V[,1:2];
print VV;

SU = S[1:rank,];
print SU;
Da = diag(SU);
print Da;

/* coordenadas dos eixos principais */
N2 = sqrt(Dr) * UU;                 /* nuvem linha */
print N2;

M2 = sqrt(Dc) * VV;                 /* nuvem coluna */
print M2;

/* matriz identidade */
I = t(N2) * inv(Dr) * N2;
print I;
II = t(M2) * inv(Dc) * M2;
print II;

/* coordenadas dos pontos */
F = inv(Dr) * N2 * Da;              /* dos pontos linha */
print F;

G = inv (Dc) * M2 * Da;             /* dos perfis coluna */

print G;
quit;
```

ANEXO 1

Arquivo de saída do Programa P.4 PROC CORRESP do SAS

Tabela de Contingência

	AREN	ARSO	MUIA	PURO	Sum
AA	2	1	5	0	8
BB	0	4	1	1	6
CC	0	9	0	1	10
DD	14	20	27	15	76
EE	16	40	30	22	108
FF	37	101	43	84	265
GG	10	92	42	85	229
SCO	426	6	42	104	578
Sum	505	273	190	312	1280

Valores esperados da estatística quiquadrado

	AREN	ARSO	MUIA	PURO
AA	3.156	1.706	1.187	1.950
BB	2.367	1.280	0.891	1.463
CC	3.945	2.133	1.484	2.438
DD	29.984	16.209	11.281	18.525
EE	42.609	23.034	16.031	26.325
FF	104.551	56.520	39.336	64.594
GG	90.348	48.841	33.992	55.819
SCO	228.039	123.277	85.797	140.888

Valores observados menos valores esperados

	AREN	ARSO	MUIA	PURO
AA	-1.156	-0.706	3.813	-1.950
BB	-2.367	2.720	0.109	-0.463
CC	-3.945	6.867	-1.484	-1.437
DD	-15.984	3.791	15.719	-3.525
EE	-26.609	16.966	13.969	-4.325
FF	-67.551	44.480	3.664	19.406
GG	-80.348	43.159	8.008	29.181
SCO	197.961	-117.277	-43.797	-86.888

Contribuição total à estatística quiquadrado

	AREN	ARSO	MUIA	PURO	Sum
AA	0.424	0.292	12.240	1.950	14.906
BB	2.367	5.783	0.013	0.146	8.310
CC	3.945	22.111	1.484	0.848	28.388
DD	8.521	0.886	21.902	0.671	31.980
EE	16.617	12.496	12.172	0.711	41.995
FF	43.645	35.006	0.341	5.830	84.822
GG	71.454	38.137	1.886	15.256	126.733
SCO	171.850	111.569	22.357	9.658	315.434
Sum	318.824	226.280	72.396	35.069	652.569

Perfis das linhas

	AREN	ARSO	MUIA	PURO
AA	0.250000	0.125000	0.625000	0.000000
BB	0.000000	0.666667	0.166667	0.166667
CC	0.000000	0.900000	0.000000	0.100000
DD	0.184211	0.263158	0.355263	0.197368
EE	0.148148	0.370370	0.277778	0.203704
FF	0.139623	0.381132	0.162264	0.316981
GG	0.043668	0.401747	0.183406	0.371179
SCO	0.737024	0.010381	0.072664	0.179931

Perfis das colunas

	AREN	ARSO	MUIA	PURO
AA	0.003960	0.003663	0.026316	0.000000
BB	0.000000	0.014652	0.005263	0.003205
CC	0.000000	0.032967	0.000000	0.003205
DD	0.027723	0.073260	0.142105	0.048077
EE	0.031683	0.146520	0.157895	0.070513
FF	0.073267	0.369963	0.226316	0.269231
GG	0.019802	0.336996	0.221053	0.272436
SCO	0.843564	0.021978	0.221053	0.333333

Inércia e Decomposição da estatística quiquadrado

Valores Singulares	Inércia Principal	Qui-Quadrado	Porcentagem
0.67737	0.45883	587.300	90.00%
0.19686	0.03875	49.605	7.60%
0.11062	0.01224	15.663	2.40%
0.50982	652.569	(Graus de liberdade = 21)	

```
18  36  54  72  90
----+---+---+---+---
**********************
**
*
```

Coordenadas das linhas

	Dim1	Dim2
AA	0.25857	1.33826
BB	1.04152	-0.16107
CC	1.22088	-0.64244
DD	0.42291	0.49179
EE	0.56056	0.24398
FF	0.55481	-0.11032
GG	0.72729	-0.10783
SCO	-0.73837	-0.02269

Resumo estatístico para os pontos-linha

	Qualidade	Massa	Inércia
AA	0.997064	0.006250	0.022842
BB	0.801996	0.004687	0.012734
CC	0.670445	0.007813	0.043502
DD	0.999797	0.059375	0.049006
EE	0.961197	0.084375	0.064354
FF	0.999674	0.207031	0.129982
GG	0.976781	0.178906	0.194207
SCO	0.999960	0.451563	0.483372

Contribuição Parcial à inércia dos pontos-linha

	Dim1	Dim2
AA	0.000911	0.288828
BB	0.011082	0.003138
CC	0.025380	0.083202
DD	0.023144	0.370545
EE	0.057785	0.129598
FF	0.138889	0.065014
GG	0.206246	0.053677
SCO	0.536563	0.005998

Índice das coordenadas que mais contribuem com a inércia dos pontos-linha

	Dim1	Dim2	Melhor
AA	0	2	2
BB	0	0	1
CC	0	2	2
DD	0	2	2
EE	0	2	2
FF	1	0	1
GG	1	0	1
SCO	1	0	1

Coseno ao quadrado para os pontos-linha

	Dim1	Dim2
AA	0.035882	0.961183
BB	0.783262	0.018734
CC	0.525059	0.145386
DD	0.425032	0.574765
EE	0.808115	0.153082
FF	0.961653	0.038021
GG	0.955772	0.021010
SCO	0.999017	0.000943

Coordenadas das Colunas

	Dim1	Dim2
AREN	-0.793227	-0.013679
ARSO	0.891236	-0.124511
MUIA	0.419275	0.451894
PURO	0.248750	-0.144104

Resumo estatístico para os pontos-coluna

	Qualidade	Massa	Inércia
AREN	0.996928	0.394531	0.488568
ARSO	0.977006	0.213281	0.346752
MUIA	0.997290	0.148437	0.110940
PURO	0.735243	0.243750	0.053740

Contribuição parcial à inércia para os pontos-coluna

	Dim1	Dim2
AREN	0.541035	0.001905
ARSO	0.369222	0.085320
MUIA	0.056871	0.782165
PURO	0.032871	0.130610

Índices das coordenadas que mais contribuem com a inércia dos pontos-coluna

	Dim1	Dim2	Melhor
AREN	1	0	1
ARSO	1	0	1
MUIA	0	2	2
PURO	0	2	2

Coseno ao quadrado para os pontos-coluna

	Dim1	Dim2
AREN	0.996632	0.000296
ARSO	0.958302	0.018704
MUIA	0.461356	0.535934
PURO	0.550495	0.184747