5. INTEGRAÇÃO E ARTICULAÇÃO DOS DIFERENTES MÉTODOS DE ANÁLISE DE DADOS

PROLONGAMENTOS DA AFC

Como foi referido no capítulo ABORDAGEM INTUITIVA DA ANÁLISE DE DADOS, a AFC admite prolongamentos interessantes que lhe conferem algum potencial explicativo, mesmo sem fazer intervir outras técnicas no domínio da CLASSIFICAÇÃO, DISCRIMINAÇÃO ou REGRESSÃO.

1. Estimação de Valores Desconhecidos do Quadro de Partida

Os algoritmos da AFC podem ser utilizados para estimar dados desconhecidos no quadro de partida, completando-o de acordo com a estrutura revelada pelas linhas ou colunas completas.

Seja então um quadro de partida como o da Fig. 5.1 em que o elemento $K(i^+, j^+)$ é desconhecido.

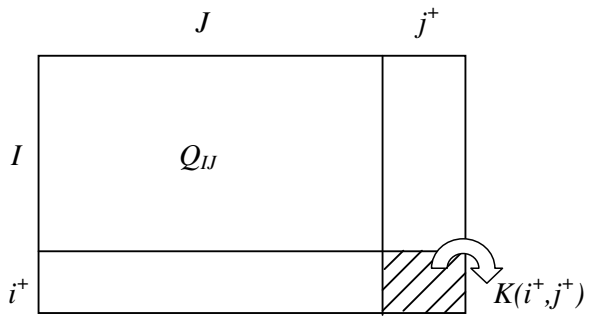

Fig. 5.1 - Quadro de partida em que se desconhece o elemento $K(i^+, j^+)$.

O Quadro Q_{IJ} está completo e pode ser diagonalizado, conduzindo aos valores próprios λ_α e projecções $f'_{j\alpha}$ (com $j \in J$) e $f'_{i\alpha}$ (com $i \in I$).

Consideremos agora j^+ como uma coluna suplementar do quadro de partida Q_{IJ} e i^+ como uma linha suplementar do mesmo quadro.

Então as relações (4.31) e (4.32) dão as projecções $f'_{i^+\alpha}$ e $f'_{j^+\alpha}$ da linha e coluna suplementar no espaço dos factores relativos ao quadro Q_{IJ}.

Aplicando agora a fórmula de reconstituição do quadro de partida (4.37), adaptada ao caso presente, obtém-se

$$f_{i^+j^+} = f_{i^+} \times f_{j^+} \left\{ 1 + \sum_{\alpha=2}^{p} \frac{1}{\sqrt{\lambda_\alpha}} f'_{j^+\alpha} \times f'_{i^+\alpha} \right\} \qquad (5.1)$$

A expressão (5.1) permite obter $f_{i^+j^+}$ e

$$K(i^+ j^+) = f_{i^+j^+} \times K \qquad \text{onde } K = \sum_{i=1}^{n} \sum_{j=1}^{p} K(i, j)$$

Evidentemente que, na estimação do elemento $K(i^+,j^+)$, se pressupõe uma certa estabilidade da estrutura do quadro. Se essa estabilidade puder ser estendida a vários elementos do quadro, a aplicação reiterada de (5.1) permite efectuar uma certa "previsão", dando à AFC uma capacidade que ultrapassa a "descrição".

2. Construção de Índices Sintéticos

Se o primeiro factor resultante de uma AFC explicar uma percentagem considerável da inércia total da nuvem e se a projecção dos indivíduos sobre esse factor for claramente interpretável em termos de uma determinada característica desses indivíduos, é possível, em certos casos, utilizar as projecções das variáveis sobre o factor como um índice sintético que mede a intensidade da característica em estudo.

Assim, por exemplo, é possível encontrar um índice de "classe social" construído com base nas projecções de determinadas variáveis que dão o estatuto socio-económico de uma certa amostra de indivíduos.

O exemplo que se apresenta seguidamente refere-se ao tratamento de um inquérito efectuado a 400 donas de casa de Lisboa em 1982. Das perguntas do inquérito trataram-se neste exemplo apenas aquelas que caracterizam a amostra do ponto de vista sócio-económico (grau de instrução, profissão, zona de residência, tipo de casa, rendimento, posse de determinados bens, etc.). Construiu-se a matriz de Burt correspondente ao cruzamento das modalidades destas perguntas entre si e projectaram-se essas modalidades nos dois primeiros eixos, obtendo-se a Fig. 5.2.

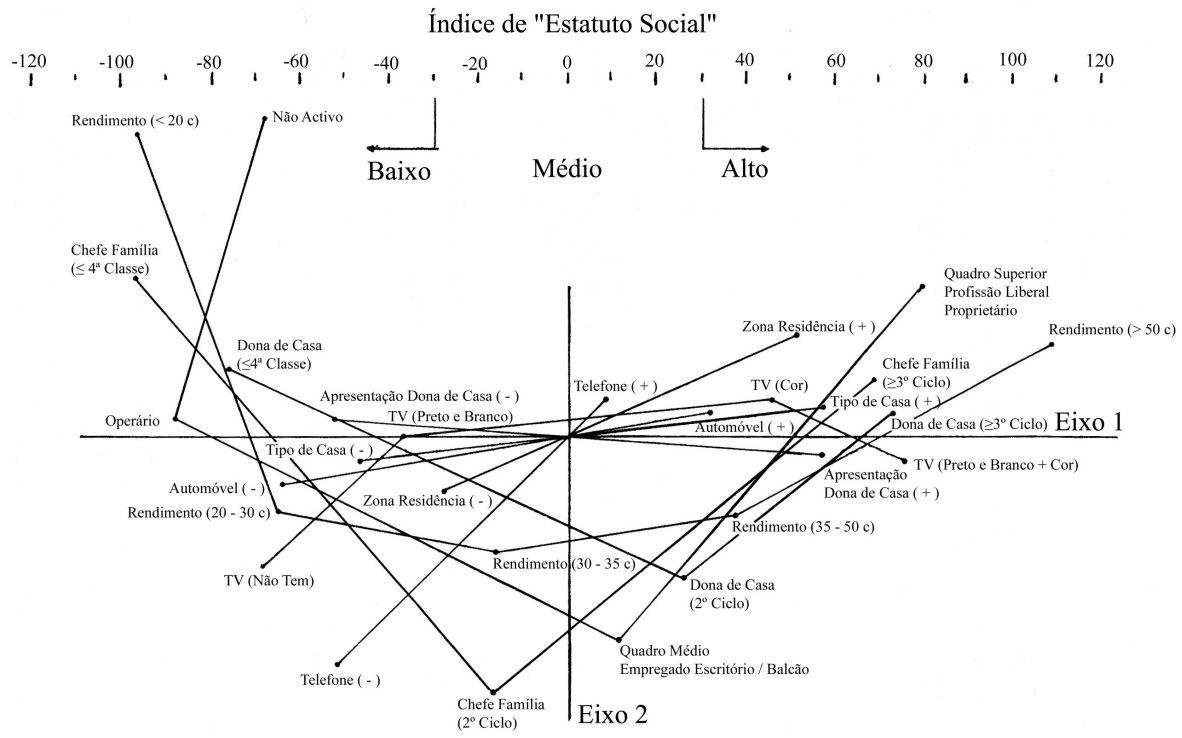

Fig. 5.2 – Projecção nos eixos 1 e 2 das modalidades das variáveis sócio-económicas

O eixo 1 é claramente um eixo de classe (opõe as modalidades associadas às classes "baixas" às modalidades associadas às "classes altas"). O eixo 2 pode interpretar-se como um eixo de idade e não é aqui considerado (elimina-se a variável idade).

Para obter a projecção de um indivíduo em suplementar sobre o eixo 1, pode aplicar-se a expressão (4.31), depois de rearranjada para a codificação disjuntiva completa

de $$f_{i^+1} = \frac{1}{\sqrt{\lambda_1}} \sum_{j=1}^{p} \left(\frac{f_{i^+j}}{f_{i^+}} \right) f'_{j1}$$

obtém-se $$f_{i^+1} = \frac{1}{q\sqrt{\lambda_1}} \sum_{j=1}^{p} \delta_{i^+j} f'_{j_1} \quad ^* \qquad (5.2)$$

onde

q é o número de perguntas

λ_1 é o valor próprio correspondente ao eixo 1

δ_{i^+j} toma o valor 1 se a modalidade *j* existe no indivíduo i^+ e 0 no caso contrário

f'_{j1} é a projecção da modalidade *j* no eixo 1.

Dispondo da tabela onde são dadas as projecções das modalidades no eixo 1 (depois de multiplicadas pela constante $\frac{1}{q\sqrt{\lambda_1}}$), o índice de "classe" é calculado como a simples soma, para todas as perguntas, dos valores da tabela para a modalidade em que o indivíduo se encontra em cada pergunta. A tabela construída para este caso concreto encontra-se no Quadro 5.1.

---

* Notar que $\frac{f_{i^+j}}{f_{i^+}} = \frac{K^+(i,j)}{K^+(i)} = \frac{\delta_{i^+j}}{q}$, porque a soma dos indicadores *booleanos* em linha é constante e igual ao número de perguntas, para a codificação disjuntiva completa.

Quadro 5.1 - Tabela para o cálculo do índice de classe.

Grau de instrução do chefe de família	
4ª classe ou menos	-19
2º ciclo ou menos	-3
3º ciclo ou superior	13
Profissão do chefe de família	
Não activo	-13
Operário	-17
Quadro médio – Emp. Escritório – Emp. balcão	2
Quadro superior – Prof. Liberal – Proprietário	16
Zona de residência	
Baixa	-6
Alta	10
Tipo de casa	
Má	-9
Boa	11
Apresentação da dona de casa	
Má	-10
Boa	11
Rendimento	
-20 contos	-21
20/30 contos	-13
30/35 contos	-3
35/50 contos	7
+50 contos	21
Grau de instrução da dona de casa	
4ª classe ou menos	-15
2º ciclo ou menos	5
3º ciclo ou superior	14
Posse de TV	
Não tem	-13
Tem TV preto a branco	-7
Tem TV cor	9
Tem TV preto e branco e cor	15
Posse de automóvel	
Não tem	-13
Tem	6
Posse de telefone	
Não tem	-10
Tem	2

O índice assim construído varia entre -133 (indivíduo para o qual ocorrem as modalidades mais "negativas") e +119 (indivíduo com as modalidades de maior peso "positivo"). Os limites para a "classe média" foram escolhidos em colaboração com o especialista do estudo (-30, +30).

DISCRIMINAÇÃO

A AFC pode ser utilizada, por si só, como uma técnica de discriminação entre dois grupos estabelecidos *a priori*, evidenciando ainda quais as propriedades que mais contribuem para a separação entre os grupos.

O exemplo de discriminação que se apresenta seguidamente refere-se à imagem dos Hospitais Civis/Clínicas Particulares, para uma amostra de 251 respondentes a um inquérito. Cada entrevistado era convidado a posicionar, numa escala de 1 a 5, a sua opinião sobre dez características de 8 instalações hospitalares específicas. Era-lhe ainda pedido que posicionasse, para a mesma escala e propriedades, a sua imagem sobre o paradigma de Hospital Civil, Clínica Particular e Instalação Hospitalar Ideal.

O modelo de quadro de partida utilizado neste estudo foi o de "notas desdobradas". A nota "positiva" correspondente a uma dada instalação hospitalar para uma certa característica obtém-se pela acumulação dos *scores* atribuídos por todos os indivíduos a essa característica da instalação hospitalar em causa. A nota "negativa" é o complemento para o máximo (neste caso 5×251). Obtém-se assim o quadro de partida que se esquematiza na Fig. 5.3.

	Caract. 1		Caract. 2			Caract. 10		
	(+)	(-)	(+)	(-)		(+)	(-)	
Instalações Hospitalares Específicas								A
Hospital Civil								B
Clínica Particular								
Hospital Ideal								C

Fig. 5.3 - Quadro de partida para a discriminação hospitais civis/clínicas particulares.

Submetendo o quadro A da Fig. 5.3 à AFC e projectando em suplementar os blocos B e C, obtém-se as projecções nos eixos 1 e 2 que se apresentam na Fig. 5.4.

O eixo 1, que explica 95% da inércia da nuvem, é claramente um "eixo discriminante" para a diferenciação entre Hospitais Civis (S. José , Sta. Maria) e os Hospitais ou Clínicas Particulares (Associação Empregados do Comércio, S. João de Deus, CUF, Cruz Vermelha , Hospital Particular, Reboleira).

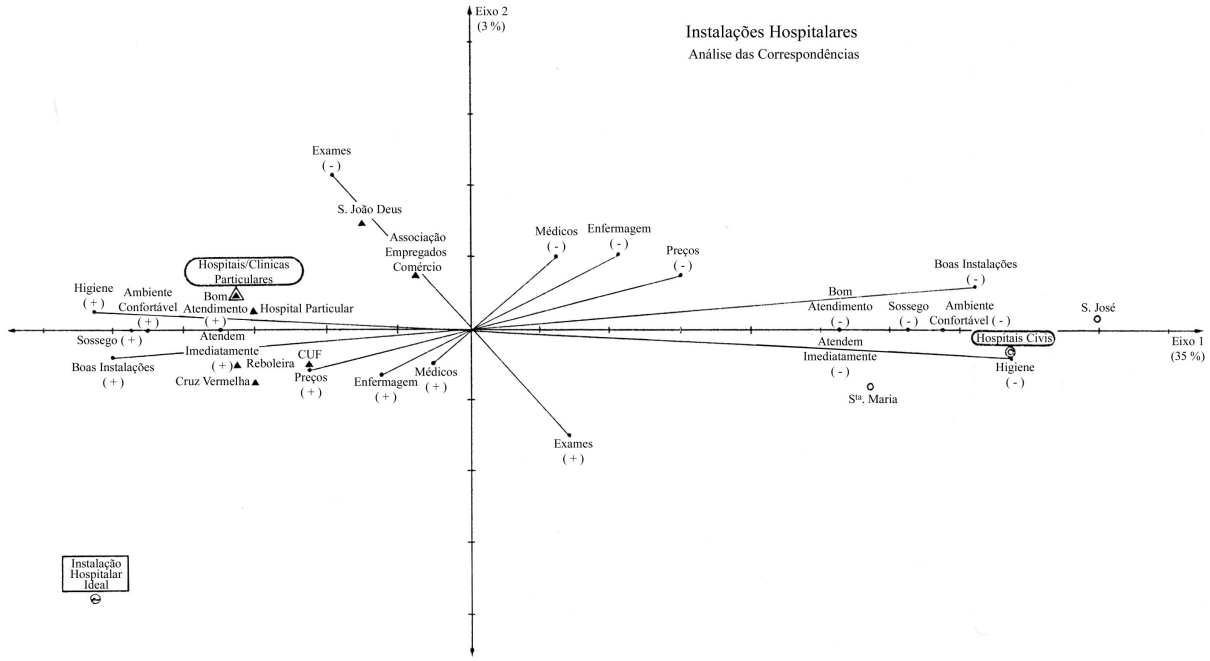

Fig. 5.4 – Projecção nos eixos 1 e 2 das instalações hospitalares e respectivas características (paradigma dos Hospitais Civis, Hospitais/Clínicas Particulares e Instalação Hospitalar Ideal projectados "em suplementar").

Então é possível efectuar agora nova AFC sobre o quadro B da Fig. 5.3 projectando em suplementar A e C. Obtém-se um único eixo onde se projectam todas as instalações hospitalares específicas, o paradigma do hospital ideal e as características (+ e -) que definem os indivíduos submetidos a análise (Fig. 5.5).

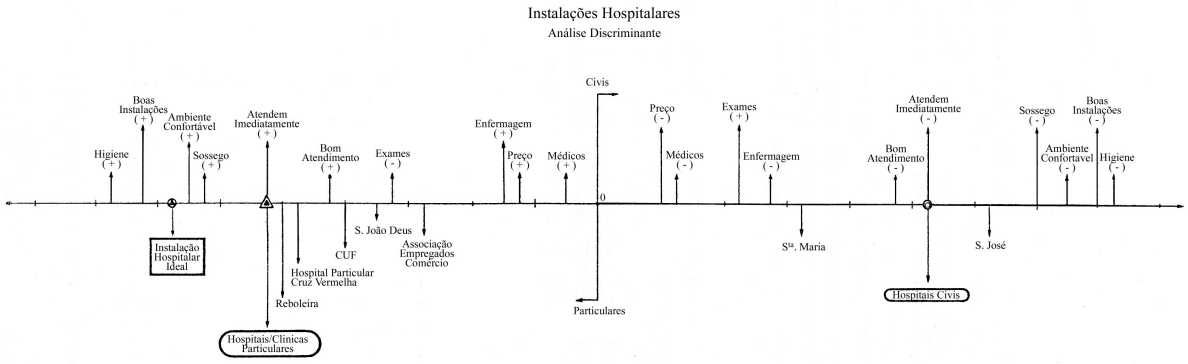

Fig. 5.5 – Discriminação entre Hospitais Civis e Hospitais/Clínicas Particulares baseada na AFC.

Sobre o eixo discriminante é possível agora avaliar quantitativamente, pela diferença de coordenadas, a "distância" de cada instalação hospitalar específica ao paradigma do Hospital Civil, Hospital/Clínica Particular ou Instalação Hospitalar Ideal. Também as características positivas ou negativas se ordenam no eixo discriminante segundo a sua coordenada, a qual mede a importância dessa característica para a criação da ideia de Hospital Civil ou Clínica Particular.

CLASSIFICAÇÃO SOBRE OS FACTORES

Como foi referido no capítulo ABORDAGEM INTUITIVA DA ANÁLISE DE DADOS, a classificação de um conjunto de indivíduos com base em certas propriedades medidas nesses indivíduos, efectuada por um algoritmo de taxonomia numérica, exige o cálculo da distância euclideana entre todos os pares de indivíduos. Se as propriedades que definem esses indivíduos são correlacionadas entre si, a distância euclideana não pode ser calculada pela fórmula habitual $\left(d^2(i,i) = \sum_p \left(x_{ip} - x_{i'p} \right)^2 \right)$ porque o espaço não é ortogonal. Aplicando então previamente um método factorial, obtém-se as projecções dos indivíduos nos eixos factoriais resultantes da ACP. Na Fig.

5.6 apresenta-se o dendrograma simplificado que se obtém pelo algoritmo de classificação. Só estão representados os níveis de agregação dos 5 grupos retidos (corte a 0.95).

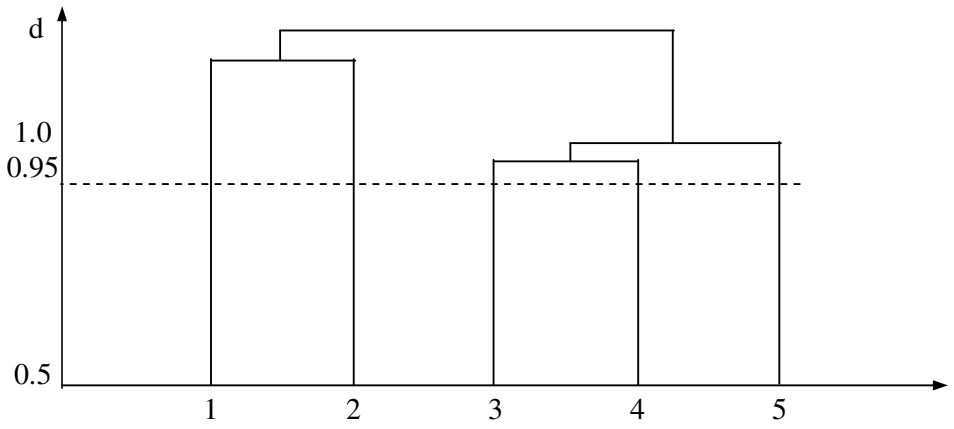

Fig. 5.6 - Dendrograma simplificado da classificação dos indivíduos.

Os valores médios dos teores químicos em cada um dos grupos e no conjunto total podem ser vistos no Quadro 5.2.

Quadro 5.2 - Teores médios por grupo.

Grupos	Enxofre (%)	Ferro (%)	Cobre (%)	Zinco (%)	Chumbo (%)	Arsénio (%)
1	43.73	35.13	0.27	5.76	3.15	0.50
2	44.65	37.64	0.33	4.98	1.78	1.07
3	44.40	40.52	0.53	3.41	1.59	0.53
4	46.47	40.81	0.22	3.54	1.20	0.41
5	46.01	38.17	0.27	7.89	2.55	0.44
Total	45.64	38.54	0.33	4.95	2.01	0.61

Comparando os resultados da Classificação com a projecção dos indivíduos no primeiro plano factorial resultante da ACP (vd. Fig. 3.7), ressalta a impossibilidade, nesse plano, de distinguir os grupos 1 e 5 (A) e 2 e 3 (B).

REGRESSÃO SOBRE OS FACTORES

A técnica de regressão, quando aplicada a um conjunto significativo de dados, fornece um modelo que relaciona uma variável a explicar com uma série de variáveis explicativas. No entanto a validade do modelo depende, como é evidente, da dimensão da amostra onde é calculado. Assim, no limite, a regressão de n variáveis efectuada sobre uma amostra de dimensão n conduz a um "modelo" sem o mínimo significado. Quanto maior for o número de casos, em face do número das variáveis, maior, evidentemente, o significado do modelo.

Para avaliar o significado da regressão pode usar-se o procedimento de decompor o conjunto das amostras disponíveis em dois subconjuntos N_1 e N_2, de dimensão n_1 e n_2. Se for possível estabelecer uma regressão com base em N_1 (se o número de variáveis for muito inferior a n_1), pode testar-se o poder explicativo do modelo encontrado pela reconstituição de N_2 à custa do modelo, análise dos resíduos, etc.. Mas este procedimento tem o inconveniente de não se basear no conjunto total de amostras.

Outro procedimento parte da ideia de que é possível reduzir o número de variáveis significativas e trabalhar sobre a amostra total. Para reduzir o número de variáveis significativas, tem forçosamente de se perder alguma informação. No entanto, para garantir a perda mínima de informação, deve utilizar-se o algoritmo que está na base dos métodos factoriais.

No caso de algumas variáveis serem qualitativas, deve optar-se pela AFC. Se a variável a explicar for quantitativa, pode também aplicar-se a AFC, criando uma série de tabelas de contingência que cruzam as modalidades da variável ordinal (em que a variável a explicar foi codificada) com as modalidades das variáveis explicativas.

O modelo de dados de partida e respectiva transformação para este caso esta representado na Fig. 5.7.

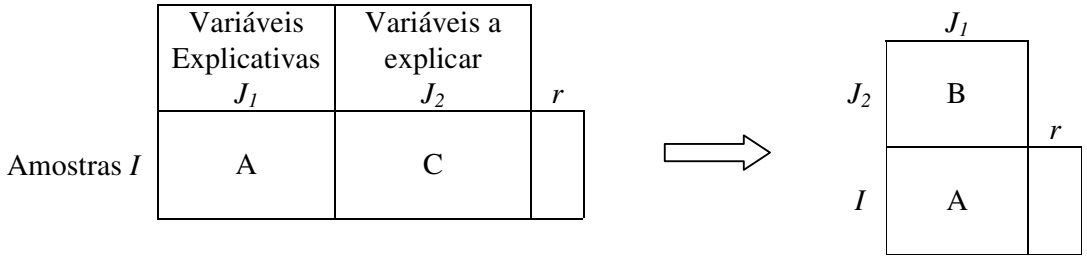

Fig. 5.7 - Transformação da matriz inicial (A ∪ C) na tabela de contingência B e na matriz suplementar A.

Notar que é possível projectar a matriz A (Fig. 5.7) em suplementar sobre os factores resultantes da AFC de B, obtendo-se assim a projecção dos *I* indivíduos. Cada indivíduo contém ainda o valor da variável quantitativa a explicar antes de esta ser categorizada (coluna *r* da Fig. 5.7). Efectuando a regressão dos *I* valores "reais" de *r* sobre os correspondentes valores da projecção em suplementar dos indivíduos nos factores, pode obter-se um modelo com significado sem reduzir a dimensão da amostra, desde que o número pequeno de factores explique uma parte aceitável da inércia de B. Finalmente, pode sempre voltar-se às variáveis explicativas iniciais, através da expressão (4.31), que relaciona as projecções em suplementar dos indivíduos com as variáveis de partida.

O exemplo que se apresenta seguidamente refere-se aos mesmos dados que já foram apresentados em EXEMPLOS DE APLICAÇÃO DA AFC (caso 1).

Para além da AFC já descrita, o passo seguinte foi a projecção dos indivíduos em suplementar sobre os factores resultantes da AFC do quadro de partida da Fig. 4.5 .

Retendo apenas o Eixo 1, que explica 92% da inércia total, pode reescrever-se a expressão (5.2):

$$R' = \frac{1}{q\sqrt{\lambda}} \sum_{j=1}^{11} \delta_j L_j \qquad (5.3)$$

onde R' é a projecção de uma amostra (ensaio) no factor 1

q é o número de variáveis explicativas (5)

δ_j é um código booleano que toma o valor 1 se a modalidade j

ocorre na amostra em causa e 0 no caso contrário

L_j é a projecção da modalidade j no factor 1

Obtendo, através de (5.3), 50 valores de R' e dispondo-se de um conjunto correspondente de valores reais da recuperação R, pode agora ensaiar-se a regressão de R em R' .

Neste caso, verifica-se que a relação $R = \phi(R')$ não é linear (Fig. 5.8) .

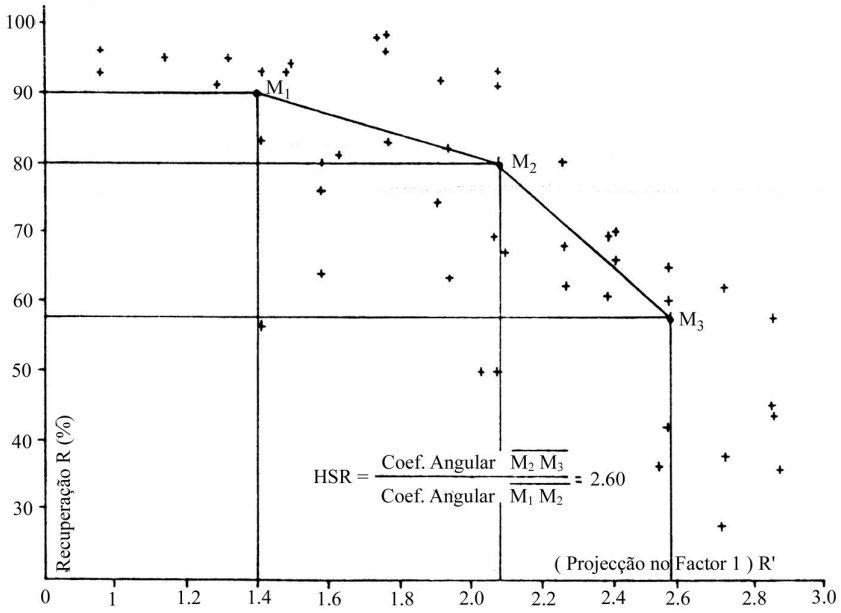

Fig. 5.8 - Representação gráfica de $R = \phi(R')$.

Assumindo uma função potência, obtém-se uma relação linear para o expoente 4.47 (vd. Fig. 5.9).

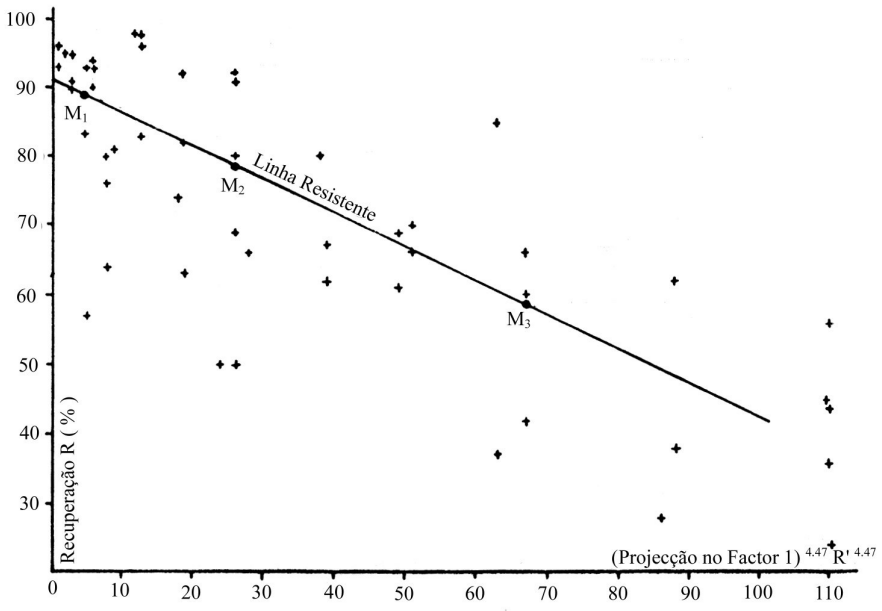

Fig. 5.9 - Linha resistente relacionando R com $R^{4.47}$.

Substituindo (5.3) no modelo encontrado por regressão, obtém-se :

$$R = 91.67 - 0.03\left(\sum_{j=1}^{11}\delta_j L_j\right)^{4.47} \qquad (5.4)$$

Os valores das projecções L_j para cada modalidade das variáveis encontram-se no Quadro 5.3.

Quadro 5.3 - Projecção L_j das modalidades das variáveis no eixo 1.

Variável	Modalidade	Projecção L_j
Presença de mineralização expressa	Sim	0.28
	Não	1.10
Presença de metamorfismo de contacto	Sim	0.94
	Não	0.65
Litologia	Xisto	0.57
	Grés	0.90
Estado de oxidação	Reduzido	0.10
	Oxidado	1.24
Teor de alimentação F ($\%U_3O_8$)	F<0.09	1.11
	0.09≤F≤0.20	0.85
	F>0.20	0.19

Para uma nova amostra, caracterizada por 5 valores correspondentes às modalidades em que as variáveis explicativas se encontram, é possível, usando a equação (5.4), com os valores adequados de L_j retirados do Quadro 5.3, prever a recuperação esperada sem efectuar o ensaio.