Henrique Garcia Pereira

*Análise Social*, vol. XXIII (98), 1987-4.°, 733-746

# Tratamento informático de questionários: o ponto de vista da análise factorial das correspondências*

## 1. INTRODUÇÃO

A dimensão das matrizes obtidas como resultado dos questionários habitualmente utilizados pelo investigador em ciências sociais exige a recorrência a modernos métodos computacionais, para garantir que se atinja, em tempo útil, um certo significado estatístico. No entanto, a escolha do programa de cálculo mais adequado a cada caso concreto não pode ser deixada ao «especialista» em informática, sob pena de o investigador do problema em causa perder o domínio sobre as próprias conclusões do estudo.

Dada a multiplicidade das grandezas que podem dar conta da variabilidade dos fenómenos a explicar através da análise dos resultados do questionário, e atendendo ao pequeno custo marginal de uma pergunta suplementar (fixada a dimensão da amostra), o investigador tem tendência a multiplicar, por vezes de uma forma incontrolada, o número de questões e a diversificar a sua forma e conteúdo, na esperança de «surpreender» certos aspectos insuspeitados da realidade complexa e multifacetada que pretende estudar. Esta atitude, se permite, por um lado, alargar o domínio dos conhecimentos empíricos, revelando eventualmente aspectos novos do fenómeno em estudo e não se limitando à simples tentativa de verificação de certas hipóteses preestabelecidas, apresenta, por outro, o perigo de diluir a capacidade de penetração no próprio objecto do inquérito, pelo efeito perturbador provocado pelo «ruído» associado a variáveis «laterais» incontroláveis.

Ultrapassado o clássico obstáculo do cálculo, através do recurso à informática, pode considerar-se que os aspectos positivos da recolha «maciça» de informação com o mínimo de hipóteses *a priori* se sobrepõem aos negativos, desde que o investigador disponha de um quadro metodológico coerente que lhe permita avançar passo a passo na formulação de hipóteses e domine um conjunto de técnicas estatísticas robustas de filtragem, capazes de eliminar as eventuais redundâncias e atenuar o ruído,

* A versão original deste texto serviu de base para uma conferência realizada no ISCTE, integrada nas cadeiras de Sociologia de Trabalho e Métodos e Técnicas de Investigação Sociológica. Agradecem-se ao Dr. Santos Lima, à Dr.ª Maria João Rodrigues e ao Eng.° António Jorge de Sousa as críticas sugeridas pela leitura do manuscrito.

fazendo emergir as estruturas significativas. De facto, o carácter multidimensional dos modelos utilizáveis em ciências sociais exige que a realidade seja apreendida segundo diferentes ângulos; por outro lado, a multiplicidade de abordagens de um mesmo tema só pode enriquecer a sua investigação. Mas, para tirar partido da importante massa de dados colhida no questionário, é necessário efectuar uma retroacção permanente entre os dados, a sua codificação, as técnicas de tratamento e a conceptualização dos modelos de interpretação, o que exige que o esquema linear clássico *dados→processamento→interpretação* seja abandonado em favor de um sistema interactivo com retroacção que permita, sem recomeçar o processo de início, retomar a cadeia de processamento em qualquer ponto, ensaiando diferentes hipóteses e alterando certos passos da análise. Tais hipóteses e procedimentos alternativos são, em geral, sugeridos no decorrer do trabalho de interpretação de *outputs* intermédios.

A atitude indutiva com retroacção que aqui é defendida inscreve-se no ponto de vista da área da estatística multidimensional designada genericamente por *análise de dados,* na acepção de J. P. Benzécri, 1973. Trata-se de uma metodologia rigorosa e coerente, dispondo de um conjunto de regras claras de codificação e interpretação, que exigem a construção de módulos de programas de computador flexíveis e versáteis, articulados e encadeados segundo diferentes modelos, capazes de responder em tempo útil aos diferentes ensaios de tratamento da informação sugeridos pelo próprio método.

Neste trabalho pretende-se ilustrar, através de exemplos de aplicação, a metodologia da análise de dados aplicada a questionários, evidenciando as suas vantagens relativamente ao tratamento clássico recorrendo apenas a percentagens e tabulações.

## 2. APURAMENTO DOS RESULTADOS DE UM QUESTIONÁRIO — UTILIZAÇÃO DAS TABULAÇÕES

Numa primeira abordagem, pode segmentar-se o conjunto de perguntas contidas na maioria dos questionários em dois grandes grupos: aquele que contém as «variáveis de classificação», referentes ao estatuto socieconómico-demográfico da população inquirida (idade, sexo, rendimento, ocupação profissional, local de residência) e as questões ligadas ao próprio objecto do inquérito (variáveis factuais e/ou de opinião associadas ao tema a investigar). Em geral, o apuramento do inquérito consiste fundamentalmente em utilizar as variáveis do primeiro grupo para estabelecer grelhas de tabulação que permitam «explicar» o comportamento das variáveis do segundo grupo. As tabulações cruzam as variáveis duas a duas[1], dando origem a um quadro de dupla entrada, contendo o número de casos em que ocorre intersecção das partições de cada variável.

Por exemplo, num questionário destinado aos estudantes do ISCTE (cf. St. Maurice, 1986) estabeleceu-se como um dos primeiros objectivos encontrar a relação entre a *idade* (segmentada em três categorias) e o facto de os inquiridos serem ou não *estudantes-trabalhadores.* Sendo 59 o número de questionários, construiu-se uma matriz de $59 \times 2$, cuja primeira

[1] Com a eventual utilização de filtros pode chegar-se a relações ternárias, como exemplifica Pires de Lima, 1981, pp. 92-100.

coluna contém um símbolo para a modalidade «idade» de cada indivíduo (—25, 25/30, +30) e a segunda um símbolo para a modalidade *sim* ou *não* da pergunta relativa ao facto de se tratar de um estudante-trabalhador. Um programa de tabulação muito simples permite acumular a frequência de co-ocorrências das 6 combinatórias das modalidades das duas perguntas e produz uma tabela como a que se apresenta no quadro n.° 1.

[QUADRO N.° 1]

| Idade | Ser estudante-trabalhador | | |
|---|---|---|---|
| | Sim | Não | Total |
| —25 anos | 2 | 44 | 46 |
| 25-30 anos | 2 | 6 | 8 |
| +30 anos | 4 | 1 | 5 |
| Total | 8 | 51 | 59 |

No quadro n.° 1 está contida a informação completa relativa às duas perguntas em análise — frequências absolutas (que podem ser transformadas em percentagens) de ocorrências cruzadas e número total de casos de cada modalidade. Trata-se da *tabela de contingência,* que resume o apuramento relativo às perguntas «idade» e «estudante-trabalhador». Se as duas questões fossem independentes (isto é, se não houvesse relação entre o facto de o indivíduo ser estudante-trabalhador e a idade), a probabilidade de co-ocorrência de cada par de modalidades ($P_{ij}$) seria dada pelo produto das probabilidades de ocorrência de cada uma delas ($P_i$ e $P_j$):

$$P_{ij} = P_i . P_j \begin{cases} i = 1, 2, 3 \text{ (modalidades da variável idade);} \\ j = 1, 2 \text{ (sim ou não ao facto de ser estudante-trabalhador).} \end{cases}$$

Aplicando a relação anterior aos totais em linha e coluna do quadro n.° 1, obtém-se, na hipótese de independência, o quadro n.° 2.

[QUADRO II]

| Idade | Ser estudante-trabalhador | | |
|---|---|---|---|
| | Sim | Não | Total |
| —25 anos | 6 | 40 | 46 |
| 25-30 anos | 1 | 7 | 8 |
| +30 anos | 1 | 4 | 5 |
| Total | 8 | 51 | 59 |

Para visualizar o afastamento relativamente à hipótese da independência pode utilizar-se uma forma gráfica como a que se exemplifica na fig. 1.

Os programas de apuramento usados tradicionalmente produzem apenas tabelas cruzando as variáveis aos pares (como a que se apresentou no quadro n.° 1). No caso geral obtém-se pois uma sequência infindável de tabelas cruzando todas as combinatórias das variáveis de partida [sendo $q$ o número de perguntas, é necessário produzir e interpretar $q(q-1)/2$ quadros para esgotar todos os cruzamentos possíveis — um pequeno questionário com 10 perguntas dá lugar a 45 tabulações]. Para tirar partido da informação disponível é pois necessário um trabalho de interpretação fasti-

dioso e demorado, o qual, aliás, não leva em conta as relações que possam existir entre os próprios elementos da grelha de tabulação (por exemplo, ao cruzar sequencialmente a variável «ser ou não estudante-trabalhador» com a «idade», com «o curso que frequenta» e com as «habilitações literárias do pai», não se detectam as interdependências nas três variáveis de base).

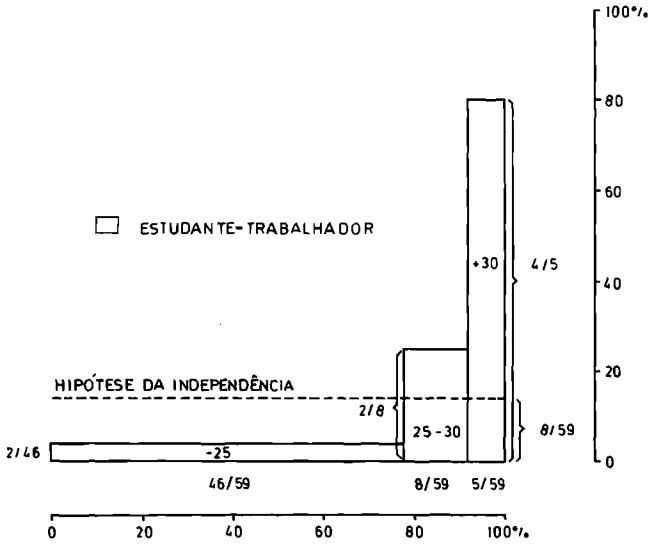



Fig. 1 — Tabela de contingência apresentada sob a forma gráfica, contendo os desvios para a hipótese de independência

Assim, na impossibilidade de analisar globalmente a informação recolhida, e não dispondo de um método claro de selecção das tabulações mais pertinentes para o objectivo em causa, o investigador acaba, muitas vezes, por produzir resultados triviais que se traduzem em algumas percentagens e certos cruzamentos simples, que não acompanham, de modo nenhum, a complexidade do fenómeno a analisar, nem sequer exploram minimamente os dados recolhidos.

Noutros casos pode até acontecer que as tabulações seleccionadas sem critério objectivo possam induzir em erros graves, que comprometem o próprio objecto do estudo.

Parece pois necessário dispor de uma metodologia genérica que trate globalmente os dados de partida, eliminando as redundâncias e combinando as variáveis «observadas» num pequeno número de factores interpretáveis, capazes de reproduzir os traços fundamentais do fenómeno em estudo. É nesta linha que surgem, nos anos 60, os métodos de *análise de dados,* de que o paradigma é a *análise das correspondências* (Benzécri, 1973, Benzécri, 1980). Trata-se de um método factorial capaz de hierarquizar a informação disponível por ordem decrescente do seu grau de explicação do fenómeno em estudo e produzir *variáveis compósitas* que resumem as relações existentes entre os atributos «observados».

A aplicação da análise das correspondências ao apuramento de questionários implica a articulação de uma série de etapas, algumas das quais são puramente automáticas (realizadas pelo computador), mas que exigem a intervenção permanente do especialista do tema a investigar.

De facto, há uma interdependência e retroacção contínua entre as diferentes fases de uma análise de dados. Da concepção do questionário até à interpretação, passando pela codificação das variáveis, há uma série de decisões a serem tomadas a cada passo, no sentido de, por um lado, assegurar a coerência estatística do método e, por outro, fazer emergir as estruturas inteligíveis que respondem ao objectivo do inquérito.

Em problemas de uma certa dimensão, essas decisões têm de ser tomadas por uma equipa pluridisciplinar que domine, simultaneamente, o objecto da pesquisa, o método matemático e os algoritmos de programação, de modo a tornar *flexível* o poderoso e rápido instrumento de cálculo e organização da informação de que se dispõe modernamente — o computador.

Nesta abordagem, a máquina não é tomada como uma caixa preta que produz um conjunto rígido de resultados através de *packages* preestabelecidos, mas sim como um auxiliar precioso capaz de responder instantaneamente aos ensaios de tratamento que a linha de investigação prosseguida vai sugerindo, havendo uma interacção permanente entre a conceptualização, o processamento e os dados.

## 3. CODIFICAÇÃO DAS VARIÁVEIS

Em geral, para surpreender diferentes facetas da realidade, o investigador concebe perguntas de natureza diversa — só a combinação de um certo número de questões (os factores que surgem da análise das correspondências) permite, muitas vezes, penetrar na complexidade do objecto da investigação. Formalmente, cada item do questionário constitui uma variável (ou observável) para a qual se estabelece um conjunto de modalidades de resposta: $S = \{S_1, S_2 \ldots S_k \ldots S_r\}$.

Conforme a natureza de $S$, assim as variáveis se dizem *nominais* —quando $S$ não tem estrutura *a priori* (por exemplo, a profissão, a religião, o estatuto jurídico de uma empresa)—, *ordinais* —quando $S$ admite uma relação de ordem (por exemplo, o nível de escolaridade, a idade, a dimensão de uma empresa)—, ou *numéricas*—quando $S$ pode ser expresso por um número real (escala de proporção ou de intervalo), munido da sua estrutura própria, que permite todas as operações aritméticas habituais (por exemplo, o salário de um trabalhador, o número de horas semanais de trabalho, o volume de vendas de uma empresa).

Desde já se pode verificar, pelos exemplos dados, que não há fronteiras rígidas entre variáveis ordinais e numéricas (o salário, expresso por um número real, pode ser transformado na presença de uma certa modalidade de salário—entre 20 e 30 contos, por exemplo; a idade, tomada como variável ordinal —entre 20 e 30 anos—, pode ser transformada na idade exacta do indivíduo). Muitas vezes, em perguntas relativas à opinião sobre determinado tema, $S$ é uma escala arbitrária destinada a matizar um interesse, uma aptidão, um grau de participação, uma frequência de utilização. Outro caso muito comum é o das perguntas abertas, em que o investigador não pode ou não quer prever *a priori* as modalidades de resposta e deixa ao inquirido a liberdade de produzir um texto sugerido pela questão posta (por exemplo, «digas o que lhe ocorre» ao ler determinada frase). No caso de perguntas abertas há que analisar previamente as respostas antes de qualquer codificação, tipificando a resposta em modalidades sugeridas por

uma análise de conteúdo, ou pela simples classificação empírica do material produzido pelos inquiridos.

Em perguntas de natureza diferente, e para assegurar o tratamento conjunto de todas as variáveis, há que garantir a coerência estatística do tratamento, através de uma codificação unificadora, usando um critério bem definido (mas flexível, de modo a poder, em qualquer passo do tratamento, alterar as fronteiras entre modalidades). Esse critério designa-se por *codificação disjuntiva completa* e consiste em estabelecer *todas* as modalidades possíveis de cada pergunta (incluindo as «sem resposta»)[2], passando as variáveis numéricas a ordinais através da definição de intervalos, adoptando escalas ou *scores* para as perguntas de opinião não dicotómicas e tipificando coerentemente as modalidades relativas às perguntas abertas. A codificação diz-se *disjuntiva* porque as modalidades são mutuamente exclusivas e *completa* porque a cada indivíduo é atribuída necessariamente uma modalidade de resposta. O procedimento da codificação disjuntiva completa consiste em transformar a informação bruta retirada do questionário num quadro rectangular em que cada inquirido ocupa uma linha e a cada modalidade de resposta corresponde uma coluna. Para cada pergunta (bloco de $r$ colunas) codifica-se como *1* a intersecção da linha $i$ com a coluna $k$ se o indivíduo de ordem $i$ escolher a modalidade de ordem $k$ e como *O* todas as outras modalidades da mesma pergunta. É através deste sistema de codificação que é possível tratar *conjuntamente todos* os tipos de variáveis, não fazendo depender o tratamento da «forma» sob a qual as perguntas são formuladas pelo especialista.

Este sistema de codificação assegura que, seja qual for a natureza das observáveis, a soma em linha dos valores que surgem na tabela é constante e igual ao número de perguntas $q$, o que se traduz numa homogeneidade estatística necessária para o processamento subsequente.

Sendo $n$ o número de inquiridos, $q$ o número de perguntas e $r(j)$ o número de modalidades da pergunta de ordem $j$, o número total de colunas da matriz de dados é:

$$p = \sum_{j=1}^{q} r(j)$$

Por exemplo, um questionário que contenha as seguintes perguntas: SEXO (2 modalidades), IDADE (3 modalidades), LOCAL DE RESIDÊNCIA (3 modalidades) e OPINIÃO SOBRE O NUCLEAR (2 modalidades), será codificado segundo o modelo da fig. 2 (Pereira, 1984):

| Perguntas | Sexo | | Idade | | | Residência | | | Aprova o nuclear | |
|---|---|---|---|---|---|---|---|---|---|---|
| Modalidades | M | F | 20/30 | 30/40 | >40 | Lisboa | Porto | Província | Sim | Não |
| Indivíduo 1 ..... | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| Indivíduo 2 ..... | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| Indivíduo 3 ..... | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |

Fig. 2 — Codificação disjuntiva completa

[2] O problema de como tratar as «não respostas» tem de ser analisado caso por caso. Se o facto de haver um número significativo de «não respostas» a uma certa pergunta estiver ligado, de algum modo, ao objecto do estudo (o que só se pode verificar *a posteriori*), estas devem ser integradas numa modalidade de resposta com o mesmo estatuto de qualquer outra. Se, pelo contrário, se verificar que as «não respostas» resultam de factores aleatórios, estas devem ser distribuídas aleatoriamente pelas outras modalidades e suprimidas da análise.

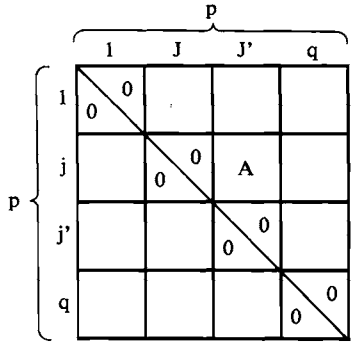
Trata-se de uma matriz de 10 colunas cuja soma em linha é sempre 4 (número de perguntas) e cuja soma em coluna dá a frequência absoluta de cada modalidade de resposta. Para cada pergunta, a soma das frequências absolutas das suas modalidades é sempre igual ao número de indivíduos submetidos ao questionário ($n$), e portanto o total em linha e em coluna reproduz $nq$. Esta propriedade é importante, visto que, deste modo, o questionário pode ser tomado como justaposição de *tabelas de contingência.*

Foi de facto com base em tabelas de contingência (quadro que dá a frequência absoluta de co-ocorrências das modalidades de duas variáveis — cf. exemplo no quadro n.° 1) que a teoria da *análise das correspondências* foi desenvolvida por J. P. Benzécri, no âmbito das aplicações em linguística, nos anos 60. Posteriormente, com os trabalhos de Lebart, 1975, o método generalizou-se ao apuramento de questionários sob a designação de «análise das correspondências múltiplas», visto que há que considerar, *simultaneamente,* um conjunto multidimensional de variáveis e ter em consideração o sistema de interdependências entre as diferentes modalidades de todas as perguntas.

## 4. A ANÁLISE DAS CORRESPONDÊNCIAS MÚLTIPLAS

Quando a matriz de partida é constituída pela justaposição de tabelas de contingência (como é o caso de qualquer questionário), a análise das correspondências múltiplas permite encontrar os *eixos factoriais* (hierarquizados por ordem decrescente da sua contribuição para a explicação da variabilidade dos dados), construídos através das combinatórias das variáveis de partida que melhor se ajustam à estrutura dos dados.

Após a fase de análise exploratória das variáveis e sua codificação (a qual pode ser guiada por *outputs* intermédios entretanto obtidos em etapas posteriores), o passo seguinte no tratamento do inquérito é construir *todas* as tabelas de contingência que cruzam as perguntas duas a duas. Esse conjunto de $q(q-1)/2$ tabelas, em vez de ser analisado independentemente, como é prática habitual no apuramento por tabulações, é organizado segundo um formato especial, designado por *matriz de Burt.* Trata-se de uma matriz quadrada e simétrica, de dimensões $p \times p$ (onde $p$ é o número total de modalidades de todas as perguntas), dividida em $q \times q$ blocos (um bloco por cada par de perguntas). A matriz de Burt é calculada como o




A matriz de Burt é simétrica. Os blocos *jj* são matrizes diagonais que contêm o número total de indivíduos distribuídos pelas modalidades de *j;* os blocos do tipo A são tabelas de contingência cruzando as modalidades de *j* com *j'*

Fig. 3 — Matriz de Burt

produto $T'T$ (onde $T$ é o quadro $n \times p$ obtido por codificação disjuntiva completa e $T'$ designa a transposta de $T$, de dimensão $p \times n$) e tem o formato esquematizado na fig. 3.

Se a matriz de Burt for submetida a um programa de análise das correspondências, obtém-se um conjunto de $p$—$q$ eixos factoriais que organizam de forma hierarquizada toda a informação contida no questionário (o primeiro eixo factorial tem mais «importância» do que o segundo para a explicação dos dados, e assim sucessivamente).

Cada eixo factorial designa-se por «vector próprio» e a ele está associado um «valor próprio», que mede a sua contribuição para a explicação da variabilidade dos dados[3]. Uma vez encontrados os eixos factoriais, arranjados por ordem decrescente da sua importância, é possível projectar as modalidades das $q$ perguntas nos primeiros $m$ eixos factoriais (o número de eixos a reter é escolhido tendo em conta a percentagem de explicação fornecida por esses eixos e a inteligibilidade dos resultados obtidos), chegando-se assim a uma imagem aproximada do quadro de partida[4].

Partindo das projecções das modalidades nos $m$ eixos, apresenta-se o *output* do método da análise das correspondências sob a forma gráfica, cruzando cada par de eixos em diagramas cartesianos que são interpretáveis com base na contribuição de cada modalidade para o eixo e nas proximidades e oposições entre projecções. Os gráficos planos cruzam os factores por ordem decrescente da sua importância e permitem pois interpretar e escolher apenas os mais significativos, eliminando aqueles cuja intervenção na compreensão do fenómeno é considerada desprezável pelo investigador.

A análise das correspondências permite ainda efectuar factorizações separadas de certos blocos da matriz de dados, projectando «em suplementar» os restantes, sobre os eixos resultantes dos primeiros. Os blocos a projectar em suplementar não intervêm na construção dos factores, mas dão a posição das modalidades relativas às perguntas desses blocos, no espaço dos eixos resultantes dos blocos «principais», efectuando-se assim uma espécie de regressão qualitativa das variáveis suplementares sobre as principais.

Uma aplicação habitual da «projecção em suplementar» liga-se com a própria estrutura do inquérito —em vez de cruzar cada pergunta do bloco de «opinião» com cada variável de «estatuto socieconómico», projectam-se em suplementar *todas* as perguntas do primeiro bloco sobre os eixos resultantes da factorização do segundo, estudando-se assim globalmente o sistema de relações entre os dois blocos, sem perder a estrutura interna de cada um deles.

Relativamente ao apuramento por tabulações, a análise das correspondências apresenta a vantagem óbvia de permitir seleccionar as combinatórias de variáveis significativas e suas relações, produzindo uma base gráfica

---

[3] O algoritmo que produz os valores e vectores próprios de uma matriz quadrada e simétrica pode ser escrito numa rotina de cálculo que existe em todas as bibliotecas de cálculo científico (cf. Kaiser, 1972, que publica uma rotina em Fortran).

[4] A perda de informação resultante do facto de se basear a análise apenas em $m$ eixos é quantificada por

$$\sum_{l=1}^{p-q} \lambda_l - \sum_{l=1}^{m} \lambda_l, \text{ onde } \lambda_l \text{ é o valor próprio de ordem } l$$

sobre a qual as estruturas presentes nos dados podem ser interpretadas de uma forma global, hierarquizando a sua importância na explicação da variabilidade do fenómeno em estudo.

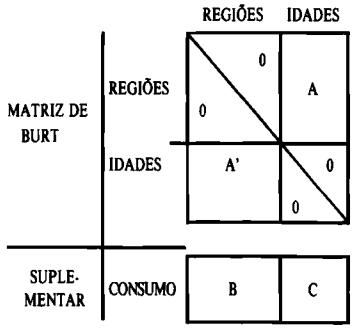
Quando o investigador selecciona certos eixos, pode basear a sua decisão em critérios quantitativos, os quais permitem, como *output* final, efectuar certas tabulações mais significativas, sem ser necessário analisar sequencialmente todos os cruzamentos possíveis, ou desprezar alguns com base em atitudes impressionistas ou em hipóteses *a priori,* geralmente de verificação duvidosa ou difícil.

## 5. EXEMPLOS DE APLICAÇÃO

A aplicação dos métodos de análise de dados ao apuramento de questionários para as ciências sociais tem sido efectuada por diferentes autores. No domínio particular da sociologia do trabalho é clássico o exemplo apresentado por Philipe Cibois[5].

Alguns exemplos de casos reais tratados pelo autor vão ser apresentados seguidamente, com o objectivo de ilustrar as potencialidades do método da análise das correspondências no tratamento de questionários.

Considere-se em primeiro lugar um exemplo didáctico de uma matriz de Burt de pequenas dimensões (Pereira, 1984, p. 86). Para relacionar o consumo de um determinado produto com a idade e a região de proveniência de um conjunto de 1939 indivíduos construiu-se em primeiro lugar a matriz de Burt que cruza as regiões com a idade. A essa matriz foram justapostas «em suplementar» as tabelas de contingência (consumo × × regiões).




A — Tabela de contingência cruzando regiões e idades.
A' — Matriz A transporta
B — Tabela de contingência cruzando consumos e regiões
C — Tabela de contingência cruzando consumos e idades

Fig. 4 — Modelo dos dados de partida para a análise de correspondências

Sujeitando a matriz de Burt ao programa de análise de correspondências, obtêm-se dois eixos principais (vectores próprios), que contêm 74% da informação de partida.

[5] Cf. Cibois, 1984, pp. 247-342. Consultar também Nicolau, 1977, p. C-1, que trata os resultados de um inquérito psicossociológico relativo à satisfação no trabalho numa empresa industrial do Togo.

Projectando em suplementar nesses eixos a variável a explicar (consumo do produto), obtém-se o gráfico da fig. 5.

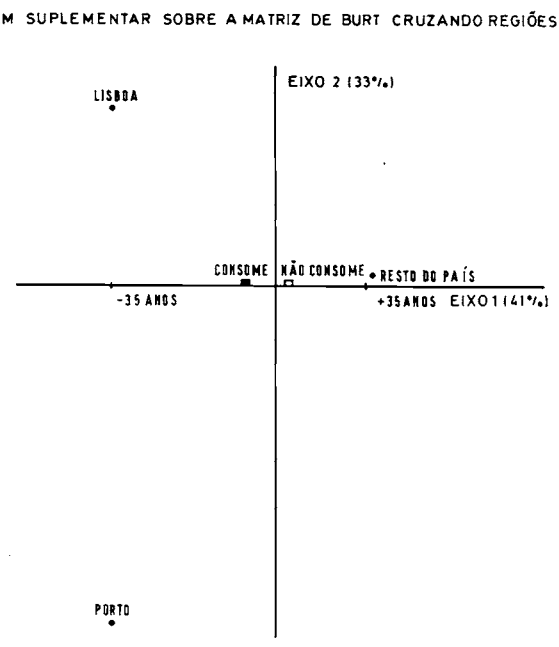



Fig. 5 — Projecção em suplementar do «consumo» sobre os eixos factoriais resultantes da análise das correspondências da matriz de Burt (regiões × idade)

A fig. 5 pode ser interpretada de um modo simples: o eixo 1 opõe as regiões «urbanas» às «rurais» e os «jovens» aos «adultos». O eixo 2 separa Lisboa do Porto. Como o consumo está ligado ao eixo 1 (a projecção das modalidades dessa variável tem projecção nula no eixo 2), poderá dizer-se que o consumo está associado com a modalidade —35 anos da variável idade e com as regiões «urbanas», não havendo distinção entre Lisboa e Porto do ponto de vista dos hábitos de consumo do produto em causa.

Consideremos seguidamente um exemplo mais complexo, que permite ilustrar as potencialidades da análise das correspondências como método poderoso de apuramento dos resultados de questionários. Trata-se de um inquérito efectuado em 1982 a 400 agregados familiares de Lisboa, cujo primeiro objectivo era a construção de um índice quantitativo de estatuto social, capaz de sintetizar um conjunto de indicadores de diferente natureza, apurados através da observação do entrevistador ou das respostas dos inquiridos (Pereira, 1984, p. 90). Os indicadores socieconómicos considerados foram 10 (nível de escolaridade do chefe de família e da dona de casa, profissão do chefe de família, zona de residência, tipo de casa, apresentação da dona de casa, rendimento dos agregados, posse de TV, automóvel e telefone). Codificadas as variáveis em 29 modalidades, segundo o modelo exemplificado na fig. 2, foi construída a matriz de Burt, que sintetiza as relações entre todas as modalidades de todas as variáveis, e aplicou-

-se a análise das correspondências a essa matriz, o que permitiu detectar uma estrutura particular nas projecções nos dois primeiros eixos — trata-se do efeito Guttman[6] (ver fig. 6). Verifica-se que as sucessivas modalidades das variáveis ordinais se dispõem segundo uma forma aproximadamente parabólica, ao longo do eixo 1. A menos da inflexão que se nota para as categorias «rendimento <20 contos» e «não activo»[7], pode dizer-se que, através de fronteiras estabelecidas num único eixo (o eixo 1), é possível distinguir claramente três «níveis de estatuto social» (baixo, médio e alto). A análise das correspondências fornece uma tabela de pesos (positivos e negativos) a atribuir às modalidades de cada variável para a construção de um índice de «classe» (*I*), dado por:

$$I = \sum_{j=1}^{q} \sum_{k=1}^{r(j)} \delta_{jk} L_{jk}$$

onde $\delta_{jk}$ é um código booleano que toma o valor 1 se a modalidade $k$ ocorre na variável $j$ e zero no caso contrário; $L_{jk}$ é o peso da modalidade $k$ da variável $j$; $r(j)$ é o número de modalidades da variável $j$.

Os pesos $L_{jk}$ são proporcionais às projecções das modalidades no eixo 1. Por simulação de agregados familiares «típicos» de cada «classe social» e dos casos intermédios foi possível estabelecer os limites —30 e +30 do estatuto social «médio» (cf. fig. 6), o que permite afectar um elemento desconhecido, caracterizado pelo vector booleano das suas 10 variáveis características, a uma das «classes» consideradas. Este exemplo permite avaliar a capacidade «explicativa» da análise das correspondências — partindo de uma bateria de indicadores qualitativos e quantitativos, foi possível resumir, no valor tomado por um único índice[8], as características que definem cada um dos tipos de «classe social» que ocorrem na população considerada.

A análise das correspondências permite ainda efectuar a discriminação de uma população em grupos, calculando o peso de cada modalidade das variáveis qualitativas na função discriminante. Apresenta-se seguidamente um exemplo deste tipo de discriminação, designada por baricêntrica, visto que cada grupo é representado pelo centro de gravidade dos indivíduos que a ele pertencem.

Com o inquérito já referido (St. Maurice, 1986), efectuado aos dois cursos do ISCTE (Sociologia e Organização e Gestão de Empresas), pretendia-se avaliar, num certo passo da análise, o modo como os dois cursos se diferenciam, do ponto de vista do consumo cultural dos seus alunos. Tomando o bloco de perguntas relativas ao tipo de teatro de que os inquiridos afirmam gostar e respectiva graduação (1 — gosta muito, 2 — gosta pouco, 3 — não gosta), foi possível construir uma matriz de 2 linhas (1 por curso) por 12 colunas (3 colunas por tipo de teatro — ligeiro, revista, clássico, de intervenção), contendo a frequência de inquiridos que escolhe uma dada modalidade para cada tipo de teatro. Submetendo esta matriz à aná-

[6] Cf. Greenacre, 1984, pp. 226-233.

[7] Para distinguir os «não activos» dos «operários» haverá que introduzir as projecções no eixo 2.

[8] Pode considerar-se que a análise das correspondências fornece um método corrente de cálculo de índices ou variáveis compósitas, no sentido apresentado por Ferreira de Almeida e Madureira Pinto, 1982, pp. 31-139.

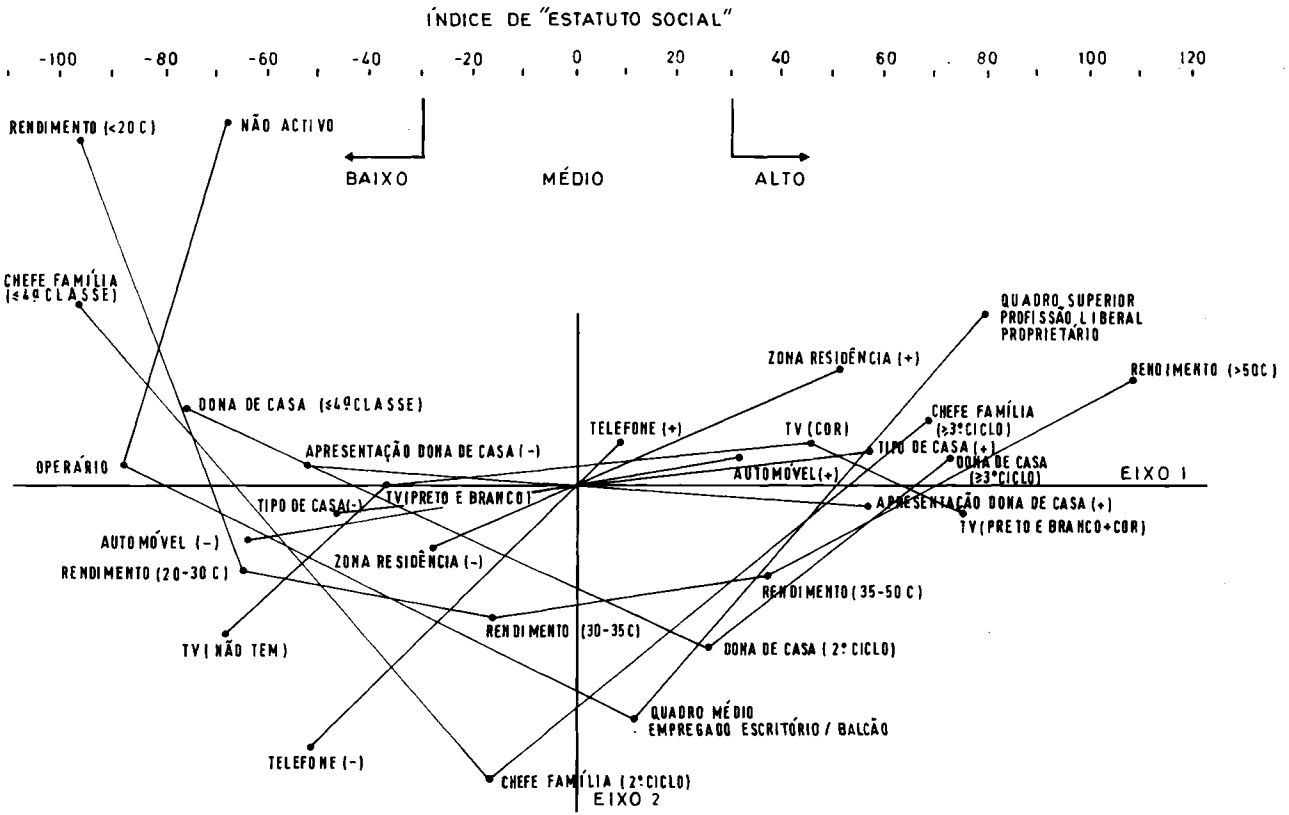



Fig. 6 — Construção de um índice de «estatuto social» através da análise das correspondências múltiplas

lise das correspondências, obtém-se um único eixo discriminante, onde se projectam os centros de gravidade dos cursos (OGE e SOC, cf. fig. 7) e as modalidades de cada tipo de teatro.

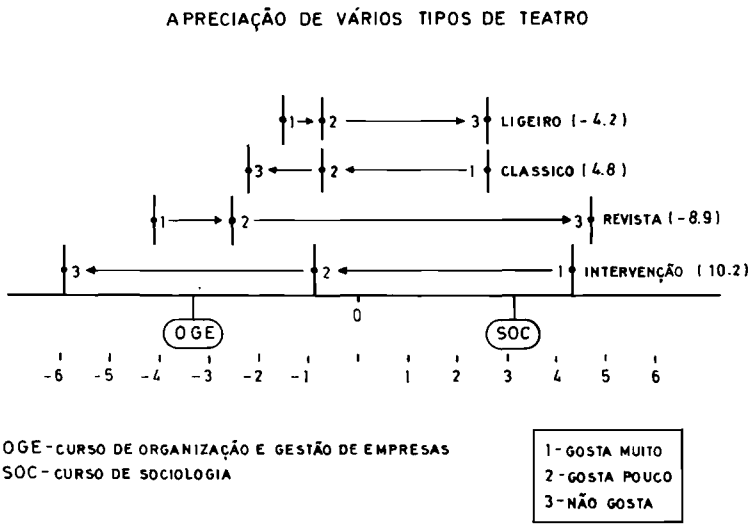



Fig. 7 — Poder discriminante do tipo de teatro nos cursos de OGE e Sociologia

Na escala do eixo discriminante é possível medir o poder diferenciador de cada tipo de teatro para os dois cursos em causa: verifica-se que o teatro de intervenção é o mais discriminante, conduzindo a uma distância OGE↔SOC de 10.2; em segundo lugar, e de sinal contrário, surge o teatro de revista (a modalidade «gosta muito» projecta-se na vizinhança do ponto OGE, ao contrário do anterior, em que essa modalidade se projecta junto do centro de gravidade do grupo de sociologia); em terceiro lugar, com poder discriminante aproximadamente igual a metade do teatro de intervenção, mas com o mesmo sinal, encontra-se o teatro clássico; finalmente, com o menor poder discriminante, surge o teatro ligeiro, com o mesmo sinal do teatro de revista, mas com distância OGE↔SOC aproximadamente igual a metade da que separa os dois cursos para o teatro de revista.

## REFERÊNCIAS BIBLIOGRÁFICAS


ALMEIDA, J. Ferreira de, e PINTO, J. Madureira (1982), *A Investigação em Ciências Sociais,* Lisboa, Presença.

BENZÉCRI, J. P. (1973), *L'Analyse des Données,* Paris, Dunod, 2 vols.

BENZÉCRI, J. P. (1980), *Pratique de l'Analyse des Données,* Paris, Dunod, 1980, 3 vols.

CIBOIS, P. (1980), *La Représentation Factorielle des Tableaux Croisées et des Données d'Enquête: Étude de Méthodologie Sociologique,* thèse 3ème cycle, Paris, CNRS, 1980.

GREENACRE (1984), *Theory and Applications of Correspondence Analysis,* Londres, Academic Press.

KAISER, H. (1972), «The JK method: a Procedure for Finding the Eigenvectors and Eigenvalues of a Real Symetric Matrix», in *Computer Journal,* vol. 15, n.° 3, pp. 271-273.

LEBART, L. (1975), «Orientation du Dépouillement de Certaines Enquêtes par l'Analyse des Correspondences Multiples», in *Consommation,* n.° 2, 1975, pp. 73-96.

LIMA, M. Pires de (1981), *Inquérito Sociológico,* Lisboa, Presença, 1981.
NICOLAU, F. da Costa (1977), *Contributions au Traitement Automatique des Données Multidimensionelles par l'Analyse des Correspondences et la Classification Automatique,* Thèse 3ème cycle, Université de Paris VI, 1977.
PEREIRA, H. Garcia (1984), *Análise de Dados para o Tratamento de Quadros Multidimensionais,* Centro de Valorização de Recursos Minerais, IST, 1984 *(roneo).*
ST. MAURICE, A. (1986), *Análise de Dados em Sociologia, Uma Pesquisa Empírica,* Instituto Superior de Ciências do Trabalho e da Empresa, provas de aptidão pedagógica e capacidade científica.