ISSN 0873-4275

INSTITUTO NACIONAL DE ESTATÍSTICA

PORTUGAL

REVISTA DE ESTATÍSTICA

VOLUME II

2º QUADRIMESTRE 1999

VOLUME 2
2° QUADRIMESTRE DE 1999

REVISTA DE ESTATÍSTICA
3ª PÁGINA
2º QUADRIMESTRE DE 1999

ÍNDICE

INDEX

- *ARTIGOS*
  ARTICLES:

INFORMAÇÕES
INFORMATIONS:

Revista de Estatística
5ª Página
2º Quadrimestre de 1999

Análise Discriminante com Selecção de Variáveis: 1ª Parte: Descrição

Autor:
António Duarte Silva

VOLUME 2

2° QUADRIMESTRE DE 1999

REVISTA DE ESTATÍSTICA
7ª PÁGINA
2º QUADRIMESTRE DE 1999

ANÁLISE DISCRIMINANTE COM SELECÇÃO DE VARIÁVEIS 1ª PARTE: DESCRIÇÃO

DESCRIPTIVE DISCRIMINANT ANALYSIS WITH VARIABLE SELECTION

Autor: António Pedro Duarte Silva*
Professor Auxiliar – Faculdade de Ciências Económicas e Empresariais
Universidade Católica Portuguesa – Centro Regional do Porto

Resumo:

- Neste trabalho discute-se o problema de selecção de variáveis em Análise Discriminante entendida numa perspectiva descritiva. É feita uma revisão de várias técnicas onde se incluem: métodos informais de selecção implícita, métodos de selecção passo a passo, métodos de comparação entre todos os subconjuntos possíveis e testes estatísticos de adequação. São propostos métodos de comparação entre todos os subconjuntos possíveis baseados em vários índices alternativos, a escolher consoante o ênfase que se pretende dar a diferentes dimensões de separação. As técnicas apresentadas serão ilustradas por um exemplo relativo à descrição das diferenças entre três grupos de bancos a operar em Portugal em 1993.

PALAVRAS-CHAVE:

- *Análise Discriminante, Selecção de Variáveis, Índices Multivariados.*

ABSTRACT:

This paper discusses several issues concerning the problem of variable selection in Descriptive Discriminant Analysis. The topics covered include informal methods for discarding variables, stepwise and all-subsets methods for variable selection, statistical tests of subset adequacy and choice of criteria for variable selection. Methods for all-subsets comparisons are proposed. It is shown how the choice of criteria for comparing subsets is related to the importance given to different dimensions of group separation.

KEY-WORDS:

- *Discriminant Analysis, Variable Selection, Multivariate indexes.*

* O autor agradece a Mário Coutinho dos Santos toda a ajuda prestada com a recolha de dados, interpretação de resultados em termos financeiros, e várias discussões que contribuiram para melhorar a sua qualidade deste trabalho. Os erros e omissões que subsistem são da exclusiva responsabilidade do autor.

VOLUME 2
2° QUADRIMESTRE DE 1999

REVISTA DE ESTATÍSTICA
9ª PAGINA
2º QUADRIMESTRE DE 1999

1. INTRODUÇÃO

A expressão *Análise Discriminante* (AD) é utilizada para designar técnicas estatísticas que têm como objectivo o estudo das diferenças entre grupos bem definidos à partida com base num conjunto relevante de características dos seus elementos. Dentro desta designação genérica encontram-se duas grandes subdivisões: a das técnicas que procuram identificar e interpretar as diferenças existentes entre os grupos e a das técnicas que estudam regras que permitem classificar indivíduos de origem desconhecida num dos grupos existentes.

Na prática, é comum que no mesmo estudo se tenham que interpretar diferenças entre grupos e simultaneamente estabelecer e estudar propriedades de regras de classificação. No entanto, estes dois problemas, embora relacionados, são conceptualmente diferentes e requerem métodos de abordagem distintos. Neste artigo, o estudo das diferenças entre grupos tendo em vista a sua interpretação será designado por *Análise Discriminante Descritiva*[1] (ADD) enquanto que o estudo de regras de classificação será designado por *Análise Discriminante Classificatória* (ADC)[2].

As técnicas clássicas de AD assumem que as características consideradas em cada indivíduo são representadas por um conjunto de variáveis escolhido à partida. Na prática, é comum recolher inicialmente um elevado numero de variáveis, efectuando uma selecção no decorrer da análise, ou simplesmente ignorando para fins de interpretação aquelas variáveis que se revelarem menos importantes ou interessantes. Este tipo de abordagem é muitas vezes baseado em procedimentos ad-hoc de propriedades mal conhecidas. Além disso, não é de todo incomum que depois de se proceder a uma selecção de variáveis, se prossiga a análise ignorando os enviesamentos decorrentes do processo de selecção.

Neste artigo far-se-á uma revisão dos principais métodos de selecção de variáveis em ADD. Inicialmente, discutir-se-ão algumas formas habituais de lidar com este problema, nomeadamente métodos informais de análise e métodos de selecção passo a passo. Em seguida, discutir-se-á o problema de identificar os subconjuntos de variáveis que incluem toda a informação relevante para explicar as diferenças entre os

---

[1] O termo "descritiva" não é aqui utilizado em oposição a "inferencial" como frequentemente acontece em outras técnicas estatísticas. Com efeito, embora técnicas puramente descritivas (nomeadamente técnicas factoriais) sejam empregues em ADD, há também métodos inferenciais que podem ser usados para ajudar a compreender diferenças entre grupos. Nomeadamente, questões do tipo, "Quantas dimensões são necessárias para explicar as diferenças entre os grupos ?" ou "Qual o subconjunto mínimo de variáveis que explica todas as diferenças observadas ?", são tipicamente abordadas com a ajuda de testes de hipóteses baseados em modelos probabilísticos.

[2] Desconhecemos a existência de alguma terminologia portuguesa já estabelecida para distinguir estas duas vertentes de *Análise Discriminante*. Daí a necessidade de definir uma terminologia própria. Aliás, tanto quanto sabemos, são raros os textos em português que fazem uma distinção clara das técnicas de *Análise Discriminante* quanto aos objectivos visados. Tal não é o caso por exemplo da literatura anglo-saxónica, onde técnicas de ADD são geralmente designadas por *Descriptive Discriminant Analysis* enquanto técnicas de ADC tem sido designadas por *Allocation, Classification in Discriminant Analysis* ou *Predictive Discriminant Analysis* (esta última designação é um pouco infeliz, na medida em que pode criar confusão com abordagens Bayesianas). Na literatura de expressão francesa as técnicas de ADD tem sido designadas por *Analyse Discriminante au but Descriptive* ou *Analyse Factoriel Discrimante* enquanto a ADC é designada por *Classement.*

grupos. Por último, propor-se-ão métodos de comparação entre todos os subconjuntos possíveis, sugerindo-se vários índices que poderão ser utilizados para esse efeito. A maioria das técnicas discutidas neste artigo são conhecidas, se bem que algumas estejam ainda pouco divulgadas. A discussão dos índices para a comparação entre todos os subconjuntos possíveis é original. O problema de selecção de variáveis em ADC será discutido num próximo artigo.

As técnicas e problemas discutidos neste artigo serão ilustradas com um exemplo relativo ao estudo das diferenças entre bancos portugueses criados depois de 1984, existentes antes de 1984, e bancos estrangeiros a operar em Portugal. O ano de 1984 foi escolhido devido a corresponder à data de aprovação da denominada "Lei de Delimitação dos Sectores" a qual terminou com as restrições ao acesso dos investidores privados à propriedade empresarial no sector bancário. As variáveis utilizadas serão indicadores de estrutura patrimonial, funcionamento e rendibilidade, construídos a partir de informação contabilística extraída dos balanços e demonstrações de resultados de 1993 publicados no Boletim Informativo da Associação Portuguesa de Bancos. Os dados foram recolhidos e gentilmente cedidos pelo Dr. Mário Coutinho dos Santos, da FCEE da Universidade Católica Portuguesa, Centro Regional do Porto.

2. ABORDAGENS TRADICIONAIS

2.1 NOTAÇÃO E CONCEITOS FUNDAMENTAIS

Considere-se um conjunto de N indivíduos divididos em k grupos e descritos por vectores $\mathbf{x_{gi}} = [x_{gi1}, x_{gi2}, ..., x_{gip}]^T$ ($i = 1,2,...,n_g$; $g = 1,2,..k$), em que x_{gij} representa o valor que a variável X_j assume para o i-ésimo indivíduo do grupo g. Designe-se o número total de indivíduos (observações) por $N = \sum_{g=1}^{k} n_g$, os centroides de cada grupo por $\overline{\mathbf{x}}_g$ e global por $\overline{\overline{\mathbf{x}}}$

$$\overline{\mathbf{x}}_g = \frac{\sum_{i=1}^{n_g} \mathbf{x}_{gi}}{n_g} \qquad\qquad \overline{\overline{\mathbf{x}}} = \frac{\sum_{g=1}^{k} n_g \overline{\mathbf{x}}_g}{N}$$

e as matrizes das somas dos desvios quadráticos e cruzados intra-grupos por $\mathbf{W}$ e entre-grupos por $\mathbf{B}$

$$W = \sum_{g=1}^{k} \sum_{i=1}^{n_g} (x_{gi} - \overline{x}_g)(x_{gi} - \overline{x}_g)^T \qquad B = \sum_{g=1}^{k} n_g (\overline{x}_g - \overline{\overline{x}})(\overline{x}_g - \overline{\overline{x}})^T$$

Designe-se ainda por

$$T=\sum_{g=1}^{k}\sum_{i=1}^{n_g}\left(x_{gi}-\bar{\bar{x}}\right)\left(x_{gi}-\bar{\bar{x}}\right)^T=\mathbf{W}+\mathbf{B}$$

a matriz das soma dos desvios quadráticos e cruzados totais. Admita-se que **W** e **T** são matrizes não singulares e que **B** tem característica $r = \min(p, k\text{-}1)$[3].

As técnicas clássicas de ADD são baseadas na análise dos vectores próprios de $\mathbf{B}\,\mathbf{W}^{-1}$ [4]. Com efeito, é bem sabido que o vectores próprios associados ao primeiro valor próprio de $\mathbf{B}\,\mathbf{W}^{-1}$ (λ_1) definem combinações lineares que maximizam o rácio entre a inércia entre-grupos e a inércia intra-grupos. Mais concretamente, pretendendo-se maximizar o rácio

$$\frac{\sum_{g=1}^{k} n_g\left(\bar{z}_g-\bar{\bar{z}}\right)^2}{\sum_{g=1}^{k}\sum_{i=1}^{n_g}\left(z_{gi}-\bar{z}_g\right)^2}\quad\left(\text{com}\quad \bar{z}_g=\frac{\sum_{i=1}^{n_g} z_{gi}}{n_g}\ \text{e}\quad \bar{\bar{z}}=\frac{\sum_{g=1}^{k} n_g\,\bar{z}_g}{N}\right)$$

entre todas as combinações lineares, $Z = b_1 X_1 + b_2 X_2 + ... + b_p X_p$, então $b_1, b_2, ..., b_p$ deverão ser escolhidos como as coordenadas de um dos vectores próprios de $\mathbf{B}\,\mathbf{W}^{-1}$ associados a λ_1. À função definida por este vector chama-se primeira Função Discriminante Linear (FDL_1) e a combinação linear resultante designa-se habitualmente por Z1. Dado que Z1 só está definida a menos de uma constante de proporcionalidade, é comum normalizá-la de forma a que tenha uma variância amostral (intra-grupos) unitária, ou seja, o vector próprio escolhido deverá garantir que $\sum_{g=1}^{k}\sum_{i=1}^{n_g}(z1_{gi} - \bar{z}\bar{1}_g)^2 / (N - k) = 1$. Aos valores assumidos por Z1 chamam-se "scores" na primeira FDL. A b_j chama-se coeficiente não padronizado de X_j na primeira FDL. Como b_j depende das unidades de medida de X_j, para fins de interpretação é conveniente definir também os coeficiente padronizados (b_j^*) que se obtém multiplicando b_j por $s_w(X_j)$, o desvio padrão intra-grupos de X_j.

$$b_j^* = b_j * s_w(X_j)\ ; \qquad s_w(X_j) = \sqrt{\mathbf{W}_{jj}/(N-k)}$$

---

[3] Note-se que se $k\text{-}1 < p$, como é habitualmente o caso, **B** será uma matriz singular. O mesmo acontece para os produtos matriciais $\mathbf{B}\,\mathbf{W}^{-1}$ e $\mathbf{B}\,\mathbf{T}^{-1}$ que tem a mesma característica que **B**.

[4] É bem sabido que $\mathbf{B}\,\mathbf{W}^{-1}$ e $\mathbf{B}\,\mathbf{T}^{-1}$ têm os mesmos vectores próprios e que os valores próprios de $\mathbf{B}\,\mathbf{W}^{-1}$ (λ_i) e $\mathbf{B}\,\mathbf{T}^{-1}$ (l_i) estão relacionados pela expressão $l_i = \lambda_i / (1+\lambda_i)$. Por conseguinte as técnicas de ADD podem ser apresentadas em termos dos vectores e valores próprios de $\mathbf{B}\,\mathbf{W}^{-1}$ ou de $\mathbf{B}\,\mathbf{T}^{-1}$. Nós utilizaremos indistintamente qualquer um destes produtos, consoante o que a cada momento seja mais conveniente para a exposição.

Diz-se então que a primeira FDL define a primeira dimensão de separação entre os grupos, procedendo-se em seguida à tentativa da sua interpretação em termos de algum conceito teórico subjacente que lhe esteja associado. De igual modo, o vector próprio normalizado associado ao segundo valor próprio de $\mathbf{B}\,\mathbf{W}^{-1}$, define uma combinação linear, não correlacionada intra-grupos[5] com Z1, que maximiza o rácio entre a inércia entre-grupos não explicada pela primeira FDL, e a inércia intra-grupos. Diz-se então que este vector define a segunda dimensão de separação. Prosseguindo deste modo é possível determinar um máximo de r FDLs não correlacionadas[6] (intra-grupos)[7] entre si, que maximizam sucessivamente a "separação" ainda não explicada previamente, e a que se poderão eventualmente associar conceitos teóricos identificados com dimensões de separação. Os valores próprios de $B\,W^{-1}$ são habitualmente interpretados como sendo proporcionais à inércia entre-grupos explicada pela respectiva dimensão de separação (Huberty 1994, 214).

As FDLs são tipicamente interpretadas com base ou nos seus coeficientes padronizados, ou nas suas correlações intra-grupos com as variáveis originais, correlações essas que também se designam por correlações estruturais. Quando o número de variáveis é elevado, é comum ignorar para efeitos de interpretação, aquelas cujos coeficientes padronizados ou correlações estruturais são menores em valor absoluto. A escolha entre coeficientes padronizados e correlações estruturais não é indiferente nem pacífica. Com efeito, é frequente que coeficientes e correlações sugiram interpretações divergentes, não estando completamente resolvido o problema de saber como as conciliar (ver, por exemplo, McKay e Campbell, 1982, 9; Huberty, 1994, 262-264). A utilização dos coeficientes padronizados surge em parte por analogia com procedimentos semelhantes em análise de regressão, com base na ideia intuitiva de que as variáveis que mais fortemente contribuem para a definição de uma FDL são aquelas que mais facilmente a ajudarão a interpretar. A utilização das correlações estruturais surge em consonância com práticas habituais em outras técnicas factoriais, com base na ideia de que variáveis fortemente correlacionadas com uma FDL tenderão a partilhar aquilo que ela representa (ou pelo menos a ser influenciadas por causas comuns). No entanto, tendo em atenção a sua natureza e características próprias, coeficientes e correlações poderão ser utilizados quer de forma alternativa quer de forma complementar. Nomeadamente, FDLs que fundamentalmente representam contrastes, poderão estar fracamente correlacionadas com todas as variáveis originais. Nesse caso os coeficientes padronizados serão mais

---

[5] A correlação intra-grupos entre as variáveis Z1 e Z2 define-se pela fórmula:

$$r_{w(Z1,Z2)} = \frac{\sum_{g=1}^{k}\sum_{i=1}^{n_g}(Z1_{gi} - \overline{Z}\,\overline{1}_g)(Z2_{gi} - \overline{Z2}_g)}{\sqrt{\sum_{g=1}^{k}\sum_{i=1}^{n_g}(Z1_{gi} - \overline{Z}\,\overline{1}_g)^2 \sum_{g=1}^{k}\sum_{i=1}^{n_g}(Z2_{gi} - \overline{Z2}_g)^2}}$$

De forma semelhante é possível definir correlações entre-grupos (r_B) e correlações totais (r_T) substituindo desvios intra-grupos por desvios entre-grupos ou desvios totais.

[6] Mais propriamente, as variáveis definidas pelas FDLs não estão correlacionadas entre si. A fim de simplificar a exposição, recorreremos frequentemente ao abuso de linguagem que consiste em chamar "correlações com FDLs" às correlações com as variáveis definidas por essas FDLs.

[7] Pode-se demonstrar que as FDLs tem correlações entre-grupos e totais igualmente nulas (Kobilinski 1990). No entanto, iremos focar a exposição nas correlações intra-grupos, uma vez que elas são as correlações mais importante em problemas de ADD.

úteis para a sua interpretação. Por outro lado, uma variável relativamente mal representada numa FDL pode estar fortemente correlacionada com ela, facilitando a identificação de um conceito que lhe esteja associado. Finalmente, a própria posição relativa dos "scores" de cada indivíduo nas FDLs ou das suas médias por grupo pode auxiliar a sua interpretação.

2.2 EXEMPLO: CARACTERIZAÇÃO DE DIFERENÇAS ENTRE INSTITUIÇÕES BANCÁRIAS A OPERAR EM PORTUGAL

A fim de tentar caracterizar as principais diferenças, em Dezembro de 1993, entre as instituições bancárias nacionais criadas antes e depois da aprovação da denominada "lei de Delimitação de Sectores" (1984) e as instituições bancárias estrangeiras a operar em Portugal, construíram-se vários indicadores de desempenho económico-financeiro a partir dos respectivos balanços e demonstrações de resultados. Incluíram-se na análise todas as instituições bancárias a operar em Portugal em 1993, com excepção do Banco Comercial de Macau e de instituições que, ainda que legalmente instituidas como Bancos, não exerciam à data em Portugal, uma actividade bancária relevante. O Banco Comercial de Macau foi excluído porque estando em fase de desagregação após a sua aquisição pelo BCP nesse mesmo ano, apresentava valores atípicos para vários indicadores. Ao todo foram analisadas 33 instituições: 14 bancos nacionais anteriores a 1984, 7 bancos nacionais posteriores a 1984 e 12 bancos estrangeiros. A lista das instituições analisadas pode ser consultada no quadro 1.

QUADRO 1

INSTITUIÇÕES BANCÁRIAS E SCORES CENTRADOS DAS FDLs

VARIÁVEIS: LR, CCG, ln TRCC, GE, ln SB, TMA, TMR, MF, MN, RCPE, RCPD, ln EB, ALE, RBA, RBCP, RA, RCP

INSTITUIÇÃO	TIPO	Z1	Z2
ABN	ESTRANGEIRO	5.075	0.933
BANIF	NOVO	-3.163	-2.064
BARCLAYS	ESTRANGEIRO	3.607	-0.926
BANCO DO BRASIL	ESTRANGEIRO	4.078	-0.190
BBI	ANTIGO	-3.244	4.397
BBV	ESTRANGEIRO	2.678	-0.365
BCA	ANTIGO	-2.495	2.060
BCI	NOVO	-1.478	-0.941
BCP	NOVO	-2.160	-3.359
BESCL	ANTIGO	-2.093	1.767
BEX	ESTRANGEIRO	2.837	-0.300
BFB	ANTIGO	-2.414	0.195
BFE	ANTIGO	-1.307	-0.007
BIC	NOVO	-2.416	-3.057
BNC	NOVO	-0.139	-3.100
BNP	ESTRANGEIRO	4.947	0.776
BNU	ANTIGO	-0.882	1.780
BPA	ANTIGO	-2.520	2.347
BPI	NOVO	-3.175	-3.344
BPSM	ANTIGO	-3.380	1.105
BTA	ANTIGO	-2.255	2.184
BTO	ESTRANGEIRO	5.623	-0.047
CGD	ANTIGO	-3.989	-0.008
CHEMICAL	NOVO	-1.645	-4.185
CITI	ESTRANGEIRO	2.786	0.546
CL	ESTRANGEIRO	3.754	1.406
CPP	ANTIGO	-1.458	0.963
DBI	ESTRANGEIRO	3.761	-0.411
GENERALE	ESTRANGEIRO	2.249	-0.204
HISPANO	ESTRANGEIRO	4.527	-0.422
UBP	ANTIGO	-2.807	0.635
MELLO	ANTIGO	-1.101	0.601
MG	ANTIGO	-1.800	1.234

Cada banco foi inicialmente descrito por um conjunto de 17 indicadores que caracterizam diversos aspectos da sua estrutura patrimonial, funcionamento, e rendibilidade. Esses indicadores constam do quadro 2. O quadro 3 apresenta várias estatísticas descritivas das distribuições desses indicadores em cada um dos grupos considerados, bem como o valor das estatísticas F* resultantes de Análises de Variância (ANOVA) efectuadas para cada um deles.

Relativamente às estatísticas F*, convém notar que, para vários indicadores, a distribuição nula habitual (F de Snedecor) não deve ser válida, uma vez que os

pressupostos da normalidade e igualdade de variâncias não parecem razoáveis. Como se pode verificar pelos resultados de testes de normalidade de Kolmogorov-Smirnov (com a correcção de Lilliefors) e de homogeneidade de variâncias de Levene, sumarizados no quadro 2, estas hipóteses foram frequentemente rejeitadas ao nível de significância de 5%. Por essa razão, incluiu-se como equivalente não-paramétrico de F*, o valor das estatísticas χ^{2*} resultantes da realização de testes de Kruskall-Wills. Manteve-se, apesar disso, a apresentação das estatísticas F* uma vez que elas podem ser simplesmente interpretadas como medidas descritivas univariadas da capacidade de cada variável em separar os grupos.

Importa aqui referir que grande parte das técnicas de ADD que vão ser discutidas neste artigo podem ser entendidas como técnicas essencialmente exploratórias, justificadas com base em argumentos não-paramétricos. No entanto estas técnicas, por um lado podem ser negativamente afectadas pela existência de "outliers", e por outro lado assumem implicitamente variâncias e covariâncias idênticas para todos os grupos. Por estas razões, é discutível a utilização das técnicas clássicas de ADD na presença de distribuições com caudas demasiado pesadas, e/ou matrizes de covariância substancialmente diferentes de grupo para grupo. Por vezes em ADD, é conveniente recorrer a testes de hipóteses que admitem explicitamente os pressupostos de normalidade multivariada e igualdade das matrizes de covariância. Para reduzir desvios em relação a estes pressupostos ou simplesmente para evitar condições adversas, é comum recorrer a transformações de variáveis. A utilização de transformações tem no entanto a desvantagem de tornar a interpretação dos resultados menos directa, sendo recomendável apenas na presença de condições particularmente adversas ou de desvios substanciais em relação aos pressupostos clássicos. Na presença de desvios moderados, é geralmente preferível trabalhar com as variáveis originais nomeadamente porque muitos dos testes usados em ADD[8] são relativamente robustos (ver Seber 1984, 440-442). Nesta aplicação, embora a hipótese de nomalidade seja frequentemente violada, uma análise dos coeficientes de assimetria (que variam entre –1.495 e 2.639) e achatamento (variando entre -2.607 e 7.158) não revela distribuições com caudas suficientemente pesadas para que, no nosso entender, se justifique a utilização de medidas correctoras. No entanto, para alguns indicadores existem diferenças marcadas quanto à dispersão por grupo, tendo-se procedido a transformações logarítmicas em três variáveis (TRCC, RCPD e EB) em que este problema era mais grave. Pelas razões apontadas, sempre que recorrermos a testes de hipóteses, deveremos interpretar os resultados obtidos apenas como indicações, não sendo os valores de prova ("p-values") a referir nem a distribuições nulas das estatísticas utilizadas, estritamente válidos.

A fim de tentar encontrar dimensões que explicassem as principais diferenças entre estes três grupos de bancos, calcularam-se os coeficientes das duas FDLs. Os valores destes coeficientes encontram-se no quadro 4, juntamente com as correlações estruturais.

[8] A maioria dos testes de hipóteses utilizados em ADD são testes de Análise Multivariada de Variância (MANOVA) e Análise Multivariada de Covariância (MANCOVA). Como se tornará evidente ao longo do texto, existe uma ligação estreita entre ADD e MANOVA, abordando estas duas metodologias, ainda que sob prespectivas diferentes, problemas fortemente relacionados.

QUADRO 2

INDICADORES DE ESTRUTURA PATRIMONIAL, FUNCIONAMENTO E RENDIBILIDADE

INDICADOR	ABREV.	DEFINIÇÃO*
Liquidez Reduzida	**LR**	L PF
Capacidade Creditícia Geral	**CCG**	A PF
Transformação dos Recursos de Clientes em Crédito	**TRCC**	A RC
Grau de Endividamento	**GE**	D FP
Solvabilidade Bruta	**SB**	FP AL
Taxa Média das Aplicações	**TMA**	JA AF
Taxa Média dos Recursos	**TMR**	JP PF
Margem Financeira	**MF**	RF AF
Margem de Negócio	**MN**	PB AF
Relevância dos Custos Pessoal	**RCPE**	CP CA
Relevância Custos no Produto	**RCPD**	CO PB
Número de Empregados por Balcão	**EB**	NP NB
Activo Líquido por Empregado	**ALE**	AL NP
Rendibilidade Bruta do Activo	**RBA**	RBT AL
Rendibilidade Bruta Capitais Próprios	**RBCP**	RBT KP
Rendibilidade do Activo	**RA**	RL AL
Rendibilidade dos Capitais Próprios	**RCP**	RL KP

*VARIÁVEIS DE GESTÃO BANCÁRIA

1. DO BALANÇO (Valores finais)		**2. DA CONTA DE EXPLORAÇÃO**	
	SIMB		**SIMB**
1.0- Cx. Dep. Bancos Centrais	L	2.1- Juros e Proveitos Equiparados	JA
1.1- Crédito s/ Inst. Crédito		2.2- Juros e Custos Equiparados	JP
1.2- Crédito s/ Clientes (Bruto)	A	2.3- Resultado Financeiro	RF
1.3- Títulos Rendimento Fixo (Bruto)		2.4- Outros Resultados Correntes	ORC
1.4- Activo Financeiro (Bruto)	AF	2.5- Produto Bancário	PB
1.5- Activo Bruto	AB	2.6- Custos com Pessoal	CP
1.6- Activo Líquido	AL	2.7- Outros Gastos Administrativos	OGA
1.7- Débitos à Vista	DV	2.8- Custos Administrativos	CA
1.8- Débitos a Prazo	DP	2.9- Resultado Bruto Exploração	RBE
1.9- Débitos Repres. Títulos	DT	2.10- Resultados Extraordinários	RX
1.10- Passivos Subordinados	PS	2.11- Resultado Bruto Total	RBT
1.11- Passivo Financeiro	PF	2.12- Amortizações e Provisões	DAP
1.12- Fundos Próprios	FP	2.13- Resultados antes Impostos	RAI
1.13- Capital, Reservas e Res.Transitados	KP	2.13- Impostos sobre Lucros	I
1.14- Recursos de Clientes e Títulos	RC	2.15- Resultado Líquido	RL
3. OUTROS DADOS			
	SIMB		
3.1- Número de Balcões Domésticos	NB		
3.2- Número de Empregados Domésticos	NP		

QUADRO 3

ANÁLISE UNIVARIADA

		NOVOS	ANTIGOS	ESTRANG.	F* (p-value)	K-W χ^{2}* (p-value)
LR	MÉDIA	12.666	15.261	9.934	**2.717** (0.082) (IVR)	**3.945** (0.139)
	D. P.	6.372	3.891	7.214		
	C. ASSIM.	-0.462	-0.233	0.263		
	C. ACHAT.	1.428	0.739	-1.380		
CCG	MÉDIA	75.691	68.191	94.341	**6.774** (0.004) (NR)	**12.955** (0.002)
	D. P.	12.106	10.055	26.706		
	C. ASSIM.	-1.418	0.989	2.283		
	C. ACHAT.	1.780	1.489	6.491		
TRCC	MÉDIA	146.266	86.575	263.507	**9.163** (0.001) (NR ; IVR)	**21.145** (< 0.001)
	D. P.	104.544	21.024	154.966		
	C. ASSIM.	2.505	2.168	0.795		
	C. ACHAT.	6.395	6.611	-0.769		
ln TRCC	MÉDIA	4.846	4.438	5.420	**16.294** (< 0.001) (NR ; IVR)	**21.145** (< 0.001)
	D. P.	0.505	0.212	0.574		
	C. ASSIM.	2.200	1.277	0.379		
	C. ACHAT.	5.071	3.528	-1.645		
GE	MÉDIA	1.065	0.569	0.951	**1.024** (0.371) (NR ; IVR)	**0.830** (0.660)
	D. P.	0.943	0.456	1.130		
	C. ASSIM.	0.443	0.799	1.428		
	C. ACHAT.	-2.414	0.159	1.694		
SB	MÉDIA	11.455	5.642	11.692	**3.809** (0.034) (NR)	**9.274** (0.010)
	D. P.	5.207	1.990	9.122		
	C. ASSIM.	0.589	1.273	2.271		
	C. ACHAT.	0.187	1.657	6.771		
TMA	MÉDIA	12.765	12.493	11.638	**1.031** (0.369) (IVR)	**1.791** (0.408)
	D. P.	0.908	0.930	2.832		
	C. ASSIM.	-1.495	0.266	0.577		
	C. ACHAT.	2.466	-1.197	0.035		
TMR	MÉDIA	9.767	8.752	9.884	**1.497** (0.240) (IVR)	**2.292** (0.318)
	D. P.	1.451	0.867	2.594		
	C. ASSIM.	0.048	0.501	0.952		
	C. ACHAT.	0.612	-0.388	0.834		
MF	MÉDIA	3.697	4.019	2.694	**2.945** (0.068) (NR)	**13.337** (0.001)
	D. P.	0.780	0.753	2.113		
	C. ASSIM.	-0.359	2.381	2.328		
	C. ACHAT.	-2.607	7.158	7.119		
MN	MÉDIA	5.091	5.042	3.592	**3.329** (0.049) (NR)	**12.449** (0.002)
	D. P.	0.854	1.178	2.166		
	C. ASSIM.	0.277	2.176	2.290		
	C. ACHAT.	-2.086	5.213	6.956		
RCPE	MÉDIA	61.316	68.999	54.551	**13.656** (< 0.001)	**15.663** (< 0.001)
	D. P.	4.727	5.843	9.094		
	C. ASSIM.	-0.026	-0.160	-0.531		
	C. ACHAT.	-2.205	0.177	-0.854		

RCPD	MÉDIA	60.840	65.336	74.760	**0.502** (0.611) (IVR)	**0.261** (0.877)
	D. P.	23.593	9.478	48.196		
	C. ASSIM.	-0.944	-0.462	1.996		
	C. ACHAT.	2.655	-0.309	5.081		
ln RCPD	MÉDIA	4.002	4.169	4.164	**0.398** (0.675) (NR ; IVR)	**0.261** (0.877)
	D. P.	0.574	0.152	0.552		
	C. ASSIM.	-2.117	-0.778	0.614		
	C. ACHAT.	5.171	0.297	0.001		
EB	MÉDIA	28.839	22.290	21.697	**0.418** (0.662) (NR ; IVR)	**3.588** (0.166)
	D. P.	36.251	5.781	9.432		
	C. ASSIM.	2.639	1.508	0.053		
	C. ACHAT.	6.973	2.587	-0.804		
ln EB	MÉDIA	2.999	3.077	2.974	**0.160** (0.853) (NR)	**3.588** (0.166)
	D. P.	0.759	0.235	0.501		
	C. ASSIM.	2.583	0.961	-0.608		
	C. ACHAT.	6.742	0.731	-0.896		
ALE	MÉDIA	1021.630	446.299	1280.675	**2.593** (0.092) (NR ; IVR)	**7.107** (0.029)
	D. P.	1109.128	350.341	1286.471		
	C. ASSIM.	1.758	2.322	1.585		
	C. ACHAT.	2.494	6.273	1.302		
RBA	MÉDIA	0.0300	0.0281	0.0209	**1.397** (0.263)	**7.599** (0.022)
	D. P.	0.0086	0.0104	0.0174		
	C. ASSIM.	0.510	2.207	1.972		
	C. ACHAT.	-1.395	6.201	5.779		
RBCP	MÉDIA	0.298	0.549	0.214	**9.841** (0.001)	**16.194** (< 0.001)
	D. P.	0.119	0.265	0.129		
	C. ASSIM.	0.574	1.589	0.314		
	C. ACHAT.	-0.976	3.244	0.295		
RA	MÉDIA	0.0083	0.0048	0.0015	**1.861** (0.173) (NR)	**2.956** (0.228)
	D. P.	0.0087	0.0030	0.0100		
	C. ASSIM.	1.653	0.234	-1.688		
	C. ACHAT.	3.104	-1.330	4.941		
RCP	MÉDIA	0.0796	0.0877	0.0202	**3.152** (0.057) (NR)	**5.047** (0.080)
	D. P.	0.0644	0.0560	0.0897		
	C. ASSIM.	0.102	0.620	-2.374		
	C. ACHAT.	-1.835	-0.942	6.969		

Legenda:

ln -- Logaritmo Neperiano.

NR -- Hipótese de normalidade rejeitada (para $\alpha = 0.05$) por um teste de Kolmogorov-Sminorv com a correção de Lilliefors.

IVR -- Hipótese de igualdade de variâncias rejeitada (para $\alpha = 0.05$) por um teste de Levene.

QUADRO 4

FUNÇÕES DISCRIMINANTES LINEARES

VARIÁVEIS: LR, CCG, ln TRCC, GE, SB, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA, RBCP, RA, RCP

		FDL1			FDL2	
VAR	Coef. não Padron.	Coef. Padron.	Corr.	Coef. não Padron.	Coef. Padron.	Corr.
LR	0.29077	1.690	-0.129	0.01388	0.081	0.098
CCG	0.02791	0.511	0.216	0.02859	0.523	-0.084
ln TRCC	6.37803	2.791	0.325	1.30648	0.572	-0.201
GE	0.90519	0.777	0.044	-0.92802	-0.796	-0.134
SB	-0.51589	-3.165	0.115	-0.58152	-3.568	-0.217
TMA	-4.43787	-8.278	-0.084	-3.72430	-6.947	-0.039
TMR	4.15851	7.455	0.076	3.52955	6.328	-0.129
MF	6.06368	8.582	-0.144	5.57241	7.887	0.045
MN	-1.74939	-2.748	-0.155	-1.12865	-1.773	-0.016
RCPE	-0.07429	-0.523	-0.285	-0.21417	-1.508	0.241
ln RCPD	5.22460	2.263	0.018	-0.31996	-0.139	0.092
ln EB	-1.45883	-0.701	-0.027	1.34778	0.648	0.037
ALE	0.00070	0.669	0.115	-0.00169	-1.605	-0.136
RBA	55.83055	0.733	-0.098	-73.93866	-0.971	-0.039
RBCP	3.89523	0.774	-0.215	9.64063	1.916	0.286
RA	98.57053	0.735	-0.098	16.32608	0.122	-0.114
RCP	-9.94804	-0.713	-0.151	-5.54440	-0.397	0.018

QUADRO 5

SCORES MÉDIOS DAS FDLs EM CADA GRUPO (SCORES CENTRADOS NA ORIGEM)

VARIÁVEIS: LR, CCG, ln TRCC, GE, ln SB, TMA, TMR, MF, MN, RCPE, RCPD, ln EB, ALE, RBA, RBCP, RA, RCP

	NOVOS	ANTIGOS	ESTRANGEIROS
Z1	-2.025	-2.268	3.827
Z2	-2.864	1.375	0.067

Os scores das FDLs e as sua médias em cada grupo encontram-se representados nos quadros 1 e 5. A primeira FDL explica 76.7% da inéricia entre-grupos (λ_1 = 9.214) e a segunda os restantes 23.3% (λ_2 = 2.799).

Uma primeira análise dos coeficientes padronizados, sugere que a primeira FDL representa essencialmente um contraste entre a *Taxa Média das Aplicações* versus a *Taxa Média de Recursos* e a *Margem Financeira*. No entanto, esta interpretação é discutível, uma vez que sendo a *Margem Financeira* aproximadamente igual à diferença entre a *Taxa Média das Aplicações* e a *Taxa Média de Recursos*, estas três variáveis tendem a anular-se. Seguem-se em ordem de importância, os coeficientes relativos à *Solvabilidade Bruta* e à *Margem de Negócio*, ambos com sinal negativo, e aos logaritmos da *Transformação dos Recursos de Clientes em Crédito* e da *Relevância dos Custos do Produto* com sinal positivo. Nenhuma dos variáveis originais está fortemente correlacionada com a primeira FDL. As correlações estruturais mais importantes dizem respeito ao logaritmo da *Transformação dos Recursos de Clientes em Crédito* (r = 0.325) e à *Relevância dos Custos com o Pessoal* (r = -0.285). Conjugando estes resultados, Z1 poder-se-á interpretar como um indicador ligado à estrutura de exploração. "Scores" elevados nesta variável indicam estruturas mais leves e gestões mais activas dos créditos a clientes. Esta dimensão permite sobretudo distinguir os bancos estrangeiros dos bancos nacionais.

Os três coeficientes padronizados mais importantes na segunda FDL sugerem igualmente um contraste entre a *Taxa Média das Aplicações* versus a *Taxa Média de Recursos* e a *Margem Financeira*. No entanto, pelas razões já acima apontadas a importância destas variáveis é mais aparente que real. O quarto coeficiente mais importante é o coeficiente associado à *Solvabilidade Bruta* com sinal negativo, seguindo-se a *Rendibilidade Bruta dos Capitais Próprios* com sinal positivo. A variável mais fortemente correlacionada com Z2 é a *Rendibilidade Bruta dos Capitais Próprios* (r = 0.286). Estes resultados sugerem que Z2 é essencialmente um indicador de estrutura patrimonial. Scores elevados nesta variável indicam uma estrutura caracterizada por um peso reduzido peso dos capitais próprios. Esta dimensão permite distinguir sobretudo os bancos nacionais posteriores a 1984 dos bancos nacionais anteriores a esta data, verificando-se que os bancos mais recentes mostram uma maior tendência a recorrer a capitais próprios.

2.3 TESTES DE INFORMAÇÃO ADICIONAL

A existência de FDLs expressas a partir de um número elevado de variáveis, grande parte das quais acabam na prática por ser ignoradas, levanta a questão de saber se não se poderá efectuar a análise com base unicamente nas variáveis que de facto parecem contribuir para as diferenças observadas. Esta questão é importante na medida em que a inclusão de variáveis irrelevantes ou redundantes pode introduzir um grau considerável de variabilidade amostral dificultando o reconhecimento das verdadeiras causas de separação. Para determinar se um determinado subconjunto de variáveis é irrelevante podem utilizar-se os chamados testes de informação adicional (Rao 1973, Seber 1984, 471-472) que são na realidade casos particulares de testes multivariados de análise de covariância (MANCOVA). Nomeadamente, suponha-se que se pretende testar se um determinado subconjunto de variáveis, Q, contem toda a informação relevante para a separação dos grupos, ou de forma equivalente se o seu

complementar, $\overline{Q}$, não contribui para as diferenças. Esta hipótese pode formalizar-se da seguinte forma

$$H_0: \mathbf{E}(\mathbf{X}_{g\overline{Q}} \mid \mathbf{X}_{gQ}) = \mathbf{E}(\mathbf{X}_{g'\overline{Q}} \mid \mathbf{X}_{g'Q}) \qquad \forall g,g'=1,2,...,k$$

ou seja, as médias condicionais das variáveis não incluídas em Q são idênticas para todos os grupos. No caso de só existirem dois grupos (k = 2) esta hipótese é equivalente à hipótese de que as distâncias de Mahalanobis (populacionais) entre as médias de grupos baseadas em Q, sejam idênticas às distâncias de Mahalanobis que consideram o conjunto completo de variáveis (Krishnaia 1982). Quando se verifica H_0, diz-se que $\overline{Q}$ não contem informação adicional para a separação entre os grupos e que o conjunto Q é um "conjunto adequado".

Para testar H_0 é conveniente estabelecer as partições de **W** e **T**

$$W=\begin{bmatrix} W_{QQ} & W_{Q\overline{Q}} \\ W_{\overline{Q}Q} & W_{\overline{Q}\overline{Q}} \end{bmatrix} \qquad T=\begin{bmatrix} T_{QQ} & T_{Q\overline{Q}} \\ T_{\overline{Q}Q} & T_{\overline{Q}\overline{Q}} \end{bmatrix}$$

e definir a matrizes de somas de desvios quadráticos e cruzados condicionais

$$\mathbf{W}_{\overline{Q}|Q} = \mathbf{W}_{\overline{Q}\overline{Q}} - \mathbf{W}_{\overline{Q}Q}\mathbf{W}_{QQ}^{-1}\mathbf{W}_{Q\overline{Q}}\,;\mathbf{T}_{\overline{Q}|Q} = \mathbf{T}_{\overline{Q}\overline{Q}} - \mathbf{T}_{\overline{Q}Q}\mathbf{T}_{QQ}^{-1}\mathbf{T}_{Q\overline{Q}}\,;\,\mathbf{B}_{\overline{Q}|Q} = \mathbf{T}_{\overline{Q}|Q} - \mathbf{W}_{\overline{Q}|Q}$$

Admitindo que os vectores aleatórios $X_g = [X_{g1},...,X_{gp}]^T$ seguem distribuições normais multivariadas com matrizes de covariância idênticas, H_0 pode ser testada com base nas estatísticas MANCOVA habituais, nomeadamente o maior valor próprio de $\boldsymbol{B}_{\overline{Q}|Q}\boldsymbol{W}_{\overline{Q}|Q}^{-1}$ (primeiro valor próprio de Roy), a soma dos valores próprios de $\boldsymbol{B}_{\overline{Q}|Q}\boldsymbol{W}_{\overline{Q}|Q}^{-1}$ (traço de Lawley-Hotelling), a soma dos valores próprios de $\boldsymbol{B}_{\overline{Q}|Q}\boldsymbol{T}_{\overline{Q}|Q}^{-1}$ (traço de Bartlett-Pillai), ou o produto dos complementares para a unidade dos valores próprios de $\boldsymbol{B}_{\overline{Q}|Q}\boldsymbol{T}_{\overline{Q}|Q}^{-1}$ (lambda de Wilks).

Para que estes testes sejam estatisticamente válidos, para alem da verificação dos pressupostos clássicos atrás referidos, é ainda necessário que o subconjunto Q tenha sido escolhido à partida. Na prática, testes de informação adicional são frequentemente utilizados com o conjunto Q sugerido por uma análise de coeficientes padronizados ou correlações estruturais. O principal problema desta estratégia reside no facto que quando Q é sugerido pelos dados, a probabilidade (sob H_0) de se rejeitar a sua adequação tende a ser menor do que a probabilidade expressa no nível de significância nominal.

Voltando ao exemplo da secção anterior, a interpretação final de FDL_1 e FDL_2 baseou-se fundamentalmente no conjunto $Q1$ formado pelas seguintes variáveis: ln TRCC, SB, MN, RCPE, ln RCPD, RBCP. Levanta-se a questão de saber se as restantes variáveis acrescentam alguma informação à capacidade explicativa de $Q1$. O

quadro 6 apresenta o valor das estatísticas MANCOVA referentes aos testes sobre a informação adicional de $\overline{Q}$. Qualquer que seja a estatística considerada, rejeita-se sempre, ao nível de significância de 5%[9], a hipótese de *Q1* ser um conjunto adequado. Por conseguinte, há evidência de que o conjunto de variáveis que foi, de facto, utilizado para interpretar as duas dimensões de separação não contem toda a informação relevante para explicar as diferenças entre estes três grupos.

QUADRO 6

TESTES DE INFORMAÇÃO ADICIONAL

VARIÁVEIS: LR, CCG, GE, TMA, TMR, MF, ln EB, ALE, RBA, RA, RBCP

CRITÉRIO	EST. MULTIV.	F*	GR. DE LIB.	P.VALUE
ROY	2.808	4.212	10 ; 15	0.005
WILKS	0.115	2.736	20 ; 28	0.007
BARTLETT-PILLAI	1.301	2.792	20 ; 30	0.010
LAWLEY- HOTELLING	4.100	2.665	20 ; 26	0.006

2.4 MÉTODOS DE SELECÇÃO PASSO A PASSO

Muitas vezes, à partida não é claro quais serão os melhores candidatos para *menores conjuntos adequados. Pretendendo-se minimizar a realização de escolhas* subjectivas no decorrer da análise, é comum recorrer a métodos de selecção automática, que num certo sentido pretendem deixar "os dados falar por si". Os mais populares destes métodos são métodos de selecção passo a passo baseados em testes de informação adicional. Nomeadamente, considere-se a seguinte factorização da estatística global de Wilks, $\Lambda = |\mathbf{W}| / |\mathbf{T}|$ (ou seja, Λ é a estatística de Wilks relativa à hipótese nula de não existência de diferenças entre os grupos)

$$\Lambda = \Lambda_{j_1} * \Lambda_{j_2|j_1} * \Lambda_{j_3|j_1,j_2} * \ldots * \Lambda_{j_p|j_1,j_2,\ldots,j_{p-1}}$$

Em que

$$\Lambda_{j_1} = \frac{\mathbf{W}_{j_1 j_1}}{\mathbf{T}_{j_1 j_1}}, \quad \Lambda_{j_i|Q} = \frac{\mathbf{W}_{j_i|Q}}{\mathbf{T}_{j_i|Q}} = \frac{\mathbf{W}_{j_i j_i} - \mathbf{W}_{j_i Q}\mathbf{W}_{QQ}^{-1}\mathbf{W}_{Qj_i}}{\mathbf{T}_{j_i j_i} - \mathbf{T}_{j_i Q}\mathbf{T}_{QQ}^{-1}\mathbf{T}_{Qj_i}}$$

Aqui, $\Lambda_{j_i|Q}$ é a estatística de Wilks relativa a um teste sobre a informação adicional de X_j dado Q. Se um teste, dado um conjunto, Q com q variáveis, tivesse

[9] Note-se no entanto, que nenhum dos valores de prova (p-values) indicados no quadro 6 são exactos porque, para além das aproximações assintóticas habituais (que neste caso só não necessárias para a estatística de Wilks), os pressupostos de normalidade multivariada e igualdade de matrizes de covariâncias não se verificam. Com efeito, vimos na secção 2.2 que as condições necessárias (mas não suficientes) de normalidade univariada e igualdade de variâncias não se verificavam para várias variáveis. Se essas condições se verificassem, para garantir a validade destes testes teríamos ainda que verificar a normalidade multivariada e a igualdade das matrizes populacionais de covariância.

sido decidido à partida, $F^* = [(N-k-q) / (k-1)] * [(1-\Lambda_{j_i|Q}) / \Lambda_{j_i|Q}]$, seguiria sob a hipótese nula uma distribuição F de Snedecor com *k-1* e *N-k-q* graus de liberdade. Métodos ascendentes de selecção partem do conjunto vazio e escolhem a cada passo a variável $\mathbf{X}_{j_i} \in \overline{Q}$ que maximiza o valor de F*, para possível inclusão em *Q*. Métodos descendentes, partem do conjunto formado por todas as variáveis e escolhem a cada passo a variável $\mathbf{X}_{j_i} \in Q$ que minimiza o valor de F*, para possível eliminação. Existem ainda métodos mistos que combinam estas duas estratégias. Em qualquer dos casos, pontos críticos das respectivas distribuições F, são muitas vezes utilizados como valores de referência para decidir quando prosseguir com a inclusão/eliminação de variáveis ou terminar o processo. No entanto, devido à existência de selecção e à realização de várias comparações, inferências baseadas na distribuição F não são de facto válidas e estes procedimentos só podem ser justificados de uma forma heurística. Note-se que apesar de este método ter sido apresentado em termos da estatística de Wilks, ele poderia ter sido formulado em função de qualquer das outras estatísticas de informação adicional habituais. Como $\mathbf{W}_{j_i|Q}, \mathbf{T}_{j_i|Q}$ e $\mathbf{B}_{j_i|Q} = \mathbf{T}_{j_i|Q} - \mathbf{W}_{j_i|Q}$ tem dimensão (1*1), $\mathbf{B}_{j_i|Q}\,\mathbf{W}_{j_i|Q}^{-1}$ só tem um valor próprio e todas as estatísticas clássicas conduzem a resultados equivalentes. Existem no entanto métodos de selecção passo a passo que utilizam critérios de selecção diferentes, nomeadamente a maximização da estatística global de Lawley-Hotelling (este critério também é referido como critério de Rao) ou critérios baseados em distâncias de Mahalanobis (ver McKay e Campbell, 1982, 13-14).

No exemplo que temos vindo a utilizar, um método ascendente (puro) de selecção passo a passo usando $\Lambda_{j_i|Q}$ como critério de selecção, e a comparação entre F* e o 90º percentil da respectiva distribuição F como critério de paragem sugere o conjunto *Q2* = {CCG, ln TRCC, MN}. Os coeficientes padronizados da primeira FDL são respectivamente 0.748 (CCG), 0.585 (ln TRCC) e -0.828 (MN). As correlações estruturais são (pela mesma ordem), 0.453, 0.695 e -0.308. Estes valores são compatíveis com a interpretação anterior da primeira dimensão de separação, que continua a distinguir sobretudo os bancos estrangeiros dos bancos nacionais anteriores a 1984 (scores médios de 1.862 e -1.319 respectivamente) ocupando os bancos nacionais posteriores a 1984 uma posição intermédia (score médio -0.493). No entanto, a segunda dimensão de separação tem agora uma expressão muito reduzida (explica apenas 1.8% da inércia entre-grupos) parecendo ser apenas o resultado de variação amostral.

Um método descendente (puro) de selecção usando os mesmos critérios, sugere o conjunto *Q3* = {LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP}. Os coeficientes das respectivas FDLs, correlações estruturais e posições relativas dos seus scores médios encontram-se nos quadros 7 e 8. A primeira FDL explica 77.6 % (λ_1 = 8.346) da inércia entre-grupos e a segunda os restantes 22.4 % (λ_2 = 2.408). Um teste de informação adicional revela que este conjunto pode ser considerado adequado (p-value > 0.50). As interpretações das FDLs e as posições relativas dos scores médios de cada grupo são, neste caso, idênticas ás interpretações baseadas na análise do conjunto completo de 17 variáveis.

QUADRO 7

FUNÇÕES DISCRIMINANTES LINEARES

VARIÁVEIS: LR, ln TRCC, GE, SB, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP

		FDL1			FDL2	
VAR	Coef. não Padron.	Coef. Padron.	Corr.	Coef. não Padron.	Coef. Padron.	Corr.
LR	0.32409	1.883	-0.137	0.05549	0.322	0.099
ln TRCC	6.99993	3.063	0.344	2.67158	1.169	-0.201
GE	1.29104	1.108	0.049	-0.70319	-0.604	-0.142
SB	-0.44495	-2.730	0.124	-0.58599	-3.596	-0.228
TMA	-4.55795	-8.502	-0.087	-3.71700	-6.933	-0.046
TMR	4.21365	7.554	0.082	3.43491	6.158	-0.136
MF	6.00060	8.493	-0.152	5.71432	8.088	0.042
MN	-1.26609	-1.989	-0.163	-1.52258	-2.392	-0.025
RCPE	-0.05767	-0.406	-0.303	-0.18357	-1.293	0.246
ln RCPD	4.60097	1.993	0.017	1.02920	0.446	0.100
Ln EB	-1.64765	-0.792	-0.029	1.29428	0.622	0.038
ALE	0.00068	0.643	0.122	-0.00159	-1.514	-0.141
RBCP	3.55562	0.707	-0.230	7.58427	1.507	0.298

QUADRO 8

SCORES MÉDIOS DAS FDLs EM CADA GRUPO
(SCORES CENTRADOS NA ORIGEM)

VARIÁVEIS: LR, ln TRCC, GE, SB, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP

	NOVOS	ANTIGOS	ESTRANGEIROS
Z1	-1.804	-2.216	3.638
Z2	-2.682	1.246	0.110

2.5 COMPARAÇÕES ENTRE TODOS OS SUBCONJUNTOS POSSIVEIS

McCabe (1975) defendeu que uma comparação entre todos os subconjuntos possíveis de variáveis, seria preferível à utilização de algoritmos de selecção passo a passo, uma vez que estes algoritmos podem não identificar contribuições para a separação decorrentes de combinações de variáveis que sejam substancialmente diferentes da soma das respectivas contribuições individuais. Este problema é mais marcante em algoritmos ascendentes do que em algoritmos descendentes (McKay e Campel 1982, 15). McCabe (1975) sugeriu usar o valor da estatística global de Wilks, entendido neste contexto como um mero índice de proximidade entre os grupos, como

critério de comparação e mostrou como o algoritmo de Furnival (1971) para comparação entre todos os subconjuntos possíveis em Análise de Regressão, pode ser adaptado para esse efeito. De acordo com McCabe (1975), é possível identificar os melhores (segundo o valor de Λ) subconjuntos de variáveis num "tempo razoável", desde que o número total de variáveis (p) não ultrapasse 20. Utilizando modernos computadores pessoais e substituindo o algoritmo de Furnival pelo algoritmo de Furnival e Wilson (1974), este limite poderá ser estendido, pelo menos até às 30 variáveis[10] (ver Duarte Silva 1998, para detalhes).

Para o exemplo que temos vindo a descrever, o quadro 9a) apresenta os dois melhores subconjuntos de cada dimensão de acordo com o valor de Λ. Os gráficos 1 e 2, mostram a evolução do índice $\tau^2 = 1-\Lambda^{1/r}$ para o melhor subconjunto de cada dimensão. Da análise do quadro 9a) ressaltam os seguintes resultados: O valor de Λ tende a decrescer de forma clara desde $q = 1$ até $q = 13$. A partir de $q = 13$, o valor de Λ parece estabilizar. As diferenças entre os valores de Λ para os dois melhores subconjuntos de cada dimensão também são claramente mais marcadas para $q = 13$ (assumindo o valor 0.0065) do que para $q > 13$ (onde nunca ultrapassa 0.0006). Estes resultados sugerem o melhor (de acordo com Λ) subconjunto com 13 variáveis, *Q3,* como um bom candidato para selecção. Curiosamente, *Q3* foi igualmente o conjunto seleccionado pelo método descendente de selecção passo a passo referido na secção 2.4. Em geral, não é possível no entanto garantir que um método de selecção passo a passo vá escolher o melhor subconjunto de uma dada dimensão de acordo com algum índice de separação (ou proximidade) nem um número de variáveis a partir do qual esse índice tenda a estabilizar. Ainda neste exemplo, o método ascendente de selecção passo a passo descrito em 2.4 tinha seleccionado o conjunto *Q2* = {CCG, ln TRCC, MN} que além de parecer ignorar a segunda dimensão de separação e utilizar um número de variáveis aparentemente bastante reduzido (Λ está longe de estabilizar para $q = 3$), está, de acordo com Λ, algo distante do melhor subconjunto da sua dimensão, *Q4* = {CCG, SB, MF}, uma vez que Λ assume os valores 0.274 e 0.300 para os conjuntos *Q4* e *Q2,* respectivamente. Finalmente, convém referir que testes de informação adicional permitem rejeitar a adequação de *Q2,* qualquer que seja a estatística utilizada (p-value < 0.02).

[10] O esforço computacional de algoritmos de pesquisa entre todos os subconjuntos possíveis cresce exponencialmente com p. Para o algoritmo de Furnival, que é um algoritmo de pesquisa exaustiva, o esforço dobra com a introdução de cada nova variável. Para $p = 20$, dependendo do computador utilizado, uma pesquisa completa poderá levar de algumas dezenas de segundos a alguns minutos, enquanto para $p = 30$ o tempo requerido será da ordem das horas ou das dezenas de horas. O algoritmo de Furnival e Wilson é um algoritmo de enumeração implícita onde o esforço computacional depende das diferenças entre os vários subconjuntos de variáveis quanto ao critério de comparação. Nas pior das hipóteses, se todos os subconjuntos tiverem contribuições semelhantes para a separação dos grupos, o esforço destes dois algoritmos será da mesma ordem de grandeza. No entanto, se alguns subconjuntos se destacarem dos restantes, pesquisas para $p = 30$ poderão demorar poucos minutos, e para p entre 40 e 50 poderão ser feitas em menos de 48 horas. Para detalhes sobre algoritmos de pesquisa entre todos os subconjuntos possíveis ver Duarte Silva (1998).

3. IDENTIFICAÇÃO DE SUBCONJUNTOS ADEQUADOS

3.1 TESTES DE HIPÓTESES SIMULTÂNEOS: A METODOLOGIA STP

Como se viu pela discussão apresentada na secção anterior, os métodos tradicionais de selecção de variáveis são fundamentalmente métodos heurísticos, que não são capazes de responder de uma forma estatisticamente válida ao problema da identificação dos subconjuntos adequados. Designe-se por $A = \{Q : Q \text{ é adequado}\}$ o conjunto dos subconjuntos de variáveis que contem toda informação relevante para explicar as diferenças entre os grupos. Nesta secção ir-se-á discutir o problema de fazer inferências sobre A. Em particular, far-se-á uma revisão de métodos que permitem encontrar conjuntos, A_α, que incluem todos os elementos de A com uma probabilidade não inferior a $1-\alpha$. Ao longo da discussão, admitir-se-ão os pressupostos de normalidade multivariada e igualdade das matrizes de covariância (populacionais) intra-grupos.

Como vimos na secção anterior, a adequação de um subconjunto particular pode ser testada com base numa estatística MANCOVA. Ao tentar fazer inferências sobre A é necessário testar simultaneamente a adequação de vários subconjuntos, o que *requer o ajustamento dos níveis de significância individuais. Uma primeira abordagem* para este problema consistiria em fazer testes de informação adicional para todos os subconjuntos possíveis com níveis de significância individuais, $\alpha_{IND} = \alpha / (2^p-1)$, ajustados pelo método de Bonferroni. No entanto, dado o elevado numero (2^p-1) de testes simultâneos a realizar, este procedimento seria demasiado conservador, o que neste caso teria como consequência a inclusão em A_α de muitos conjuntos inadequados, para ser útil na prática. Existem no entanto métodos menos conservadores. Nomeadamente, Mckay e Campbell (1982) descrevem uma metodologia designada por STP (do inglês, Simultaneous Test Procedure) desenvolvida inicialmente por Gabriel (1968, 1969) e adaptado por Mckay (1977) para problemas de selecção de variáveis Análise Discriminante. O primeiro passo desta metodologia consiste em testar, com base numa das estatísticas MANOVA habituais, se a totalidade das variáveis originais revelam de facto diferenças entre os grupos. Este teste é conduzido ao nível global de significância, α. Se a hipótese nula, de médias idênticas para todos os grupos, for aceite não faz sentido continuar a análise. No caso de se concluir pela existência de diferenças significativas, em seguida prossegue-se com o cálculo, para cada subconjunto Q, de uma estatística MANCOVA respeitante à informação adicional contida em $\overline{Q}$. Essa estatística é comparada com o mesmo ponto crítico usado para o teste MANOVA inicial. Note-se que este ponto crítico é diferente daquele que seria usado para um teste MANCOVA individual decidido à partida. Por exemplo, se os testes forem conduzidos com base na estatística de Wilks, o nível de significância individual (α_{IND}) pode ser calculado através da igualdade

$$\Lambda(p-q,k-1,N-k-q;\ \alpha_{IND}) = \Lambda(p,k-1,N-k;\ \alpha)$$

onde $\Lambda(d,m_H,m_E;\ \alpha)$ é o ponto crítico da distribuição de uma estatística Λ com parâmetros d, m_H e m_E (ver Seber 1984, 40-42), a um nível de significância α. Como o ponto crítico, $\Lambda(p,k-1,n-k;\ \alpha)$, para o teste MANOVA global, é inferior ao ponto crítico $\Lambda(p-q,k-1,n-k-q;\ \alpha)$, para o teste MANCOVA sobre a adequação de Q, α_{IND} é inferior a α, sendo esta diferença entre α_{IND} e α que garante a protecção global da

bateria de testes. Testes baseados em qualquer das outras três estatísticas habituais são igualmente possíveis, sendo nesse caso o ponto critico efectivamente utilizado superior àquele que corresponderia a um único teste de informação adicional. Em qualquer dos casos, α_{IND} é sempre inferior a α, ou seja utilizam-se testes individuais conservadores a fim de se garantir uma protecção global ao nível α. No entanto, para se evitarem testes individuais demasiado conservadores é habitual escolherem-se valores relativamente altos para o nível de significância global.

Há três aspectos que se devem ter em consideração para a escolha da estatística em que se baseiam os STP. Em primeiro lugar, as diferenças entre α_{IND} e α variam consoante a estatística escolhida. Em particular, quando os STP são baseados no primeiro valor próprio de Roy, os testes individuais são menos conservadores do que em STP baseados em qualquer outra estatística (McCabe 1982, 19). Com base nesta característica, Gabriel (1968) descreve os STP baseados no primeiro valor próprio como sendo os STP mais *resolventes*. Os STP baseados na estatística de Wilks tendem a ser os menos *resolventes* enquanto os STP baseados nos traços de Lawley-Hotteling ou Bartllet-Pillai tendem a ter um comportamento intermédio. O segundo aspecto a considerar é a potência dos testes individuais. É sabido que o facto da estatística de Roy ignorar todos os valores próprios de $\mathbf{B}\,\mathbf{W}^{-1}$ à excepção do primeiro, pode prejudicar seriamente a sua potência (Seber 1984, 414-416). Em particular, se as diferenças entre grupos não poderem ser explicadas ao longo de uma única dimensão, a potência da estatística de Wilks pode ser substancialmente superior à potência da estatística de Roy, sendo esta diferença de potências geralmente mais importante do que as diferenças devidas à *resolução* (McCabe 1982,). Finalmente, um terceiro aspecto de ordem mais prática tem a ver com razões de ordem computacional. Nomeadamente, dado que os STP consideram implicitamente todos os subconjuntos possíveis a sua aplicação pode tornar-se impraticável para um número moderado de variáveis. Existem no entanto duas formas de tentar ultrapassar este problema. Em primeiro lugar, reconhecendo uma propriedade de coerência dos STP, nenhum subconjunto de um conjunto excluído de $\mathbf{A}_\alpha$ pode fazer parte de $\mathbf{A}_\alpha$, nem todos os subconjuntos tem que ser testados quanto à sua adequação. Esta propriedade traduz o principio intuitivo de que a informação adicional de um subconjunto $\overline{Q}$ nunca pode diminuir quando se adicionam variáveis a $\overline{Q}$, e é uma consequência do facto de as estatísticas de informação adicional de Bartlett-Pillai, Lawley-Hotelling e Roy, nunca diminuírem quando se eliminam variáveis de Q, enquanto a estatística de Wilks nunca aumenta nas mesmas circunstâncias. Em segundo lugar, utilizando o algoritmo de Furnival-McCabe é possível calcular o valor da estatística de Wilks para todos os subconjuntos possíveis (Λ_Q)[11] de forma eficiente. Por sua vez, as estatísticas de Wilks sobre a informação adicional de $\overline{Q}$ dado Q $(\Lambda_{\overline{Q}|Q})$ podem obter-se a partir da factorização $\Lambda = \Lambda_Q * \Lambda_{\overline{Q}|Q}$. Para as restantes estatísticas multivariadas não existem factorizações equivalentes que permitam realizar todos os testes de informação adicional de uma forma eficiente.

[11] Formalmente Λ_Q define-se como $|W_{QQ}| / |T_{QQ}|$, ou seja Λ_Q é a estatística de Wilks para testar a existência de diferenças entre-grupos nas médias de Q.

3.2 EXEMPLO

Para o exemplo descrito na secção 2.2, a estatística de Wilks relativa à hipótese nula de igualdade dos valores esperados de todos os indicadores para os três grupos de bancos assume o valor $\Lambda^* = 0.02577$. Os parâmetros da distribuição de Λ são neste caso, $d = p = 17$, $m_H = k\text{-}1 = 2$ e $m_E = N\text{-}k = 30$. Recordando que, para uma estatística $\Lambda(d,m_H,m_E)$, quando $\min(d,m_H) = 2$, a transformação $F^* = [(1 - \Lambda^{1/2}) / \Lambda^{1/2}] * [(m_E - d+1)/(|m_H - d|+2)]$ segue sob a hipótese nula uma distribuição F de Snedecor com $2(|m_H - d|+2)$ e $2(m_E - d+1)$ graus de liberdade (Seber 1984, 43), o ponto crítico da distribuição de Λ para um teste ao nível de significância, $\alpha = 0.10$, é igual a 0.1145. Neste caso, há evidência clara (p-value < 0.001) para se concluir pela existência de diferenças reais entre os grupos. A metodologia STP a um nível global de significância de 10%, leva a incluir no conjunto $A_{0.10}$ todos os subconjuntos, Q , para os quais $\Lambda_{\overline{Q}|Q}$ seja *superior a 0.1145, ou de forma equivalente* Λ_Q *seja inferior a* 0.2251. Esta estratégia corresponde a incluir subconjuntos em $A_{0.10}$, em função de testes de informação adicional conduzidos aos níveis individuais de significância de 0.074 ($q = 1$), 0.058 ($q = 2$), 0.040 ($q = 3$), 0.028 ($q = 4$) e a níveis inferiores a 0.02 para $q \geq 5$. De acordo com este critério $A_{0.10}$ não deverá incluir nenhum conjunto formado por menos de 4 variáveis. Incluem-se em $A_{0.10}$ dois subconjuntos formados por 4 variáveis, a saber *Q5* = {CCG, SB, MF, ln RCPD} e *Q6* = {CCG, SB, MF, RA}. De entre os conjuntos com um número de variáveis superior a 4 dever-se-ão incluir (entre outros) todos aqueles que contenham *Q5* ou *Q6* como subconjuntos. Uma descrição exaustiva de $A_{0.10}$ seria por um lado extremamente complexa devido ao elevado número de conjuntos envolvidos, e por outro lado de reduzida utilidade dado o objectivo habitual de procurar subconjuntos simples que contenham toda a informação relevante.

As variáveis incluídas em *Q5* e *Q6*, *Capacidade Creditícia Geral*, *Solvabilidade Bruta*, *Margem Financeira* e o logaritmo da *Relevância dos Custos no Produto* (*Q5*) ou a *Rendibilidade do Activo* (*Q6*), são variáveis que capturam os aspectos fundamentais das duas dimensões de separação. Aparentemente, estas variáveis seriam suficientes para explicar todas as diferenças entre os grupos considerados. Impõem-se no entanto alguns comentários. Em primeiro lugar, convém notar que nos testes de informação adicional, o erro de 1ª espécie, consiste em considerar como inadequado um conjunto que é adequado. Ou seja, a probabilidade que é controlada, é a probabilidade de não se identificarem alguns conjuntos inadequados. No nosso exemplo, podemos nomeadamente afirmar (com a probabilidade de erro controlada a 10%) que nenhum conjunto formado por menos de 4 indicadores é por si só suficiente para explicar todas as diferenças entre os grupos considerados. Já a afirmação de que *Q5* ou *Q6* são subconjuntos adequados, corresponde a hipóteses cuja probabilidade de erro não foi controlada. Esta é, bem entendido, a estratégia habitual, considera-se um conjunto como adequado a não ser que haja evidência em contrário. Porém, no caso da metodologia STP esta estratégia levanta alguns problemas. Por um lado, dado que esta abordagem utiliza testes individuais conservadores, a probabilidade de se cometerem erros de 2ª espécie pode, por essa razão, ser demasiado elevada. Neste exemplo, tal não parece ser um problema grave, dado que os níveis de significância individuais para subconjuntos com menos de 5 variáveis estão próximos dos 5% habituais. Por outro lado, em testes ligados a métodos de selecção de variáveis, as consequências de se cometerem erros de 2ª espécie podem ser mais graves do que noutras situações, uma vez que eles levam a que não se incluam na análise variáveis importantes. Em particular, quando se utilizam amostras de dimensão reduzida, não existindo evidência

suficiente para se concluir que um determinando subconjunto é inadequado, pode não ser claro se tal é devido de facto à adequação do respectivo conjunto. Em situações deste género, é conveniente tentar determinar de uma forma exploratória se a inclusão de variáveis adicionais é capaz de enriquecer a análise.

3.3 OUTRAS METODOLOGIAS DE TESTES SIMULTÂNEOS

Os STP de Gabriel e Mckay não constituem a única estratégia para fazer inferências sobre A controlando o nível de significância global de todos os testes realizados. Uma abordagem alternativa, sugerida por Spjøtvoll (1978) para análise de regressão mas directamente aplicável neste contexto, consiste em numa primeira fase realizar para todos os subconjuntos testes de informação adicional ao nível α ignorando que se estão a realizar vários testes em simultâneo. Em seguida, incluem-se em A_α todos os conjuntos que foram considerados adequados por estes testes e ainda todos os conjuntos que incluam membros de A_α como subconjuntos, ainda que a hipótese de adequação tenha sido inicialmente rejeitada. Desta forma, a propriedade da coerência, que era verificada automaticamente nos STP é aqui assegurada "à força". Pode-se provar (Spjøtvoll 1978) que os A_α assim construídos incluem todos os conjuntos adequados com uma probabilidade não inferior a 1-α. Este procedimento tem ainda a vantagem de ser menos conservador que os STP de Gabriel e Mckay. As suas principais desvantagens são as seguintes : (i) A inclusão em A_α para cada subconjunto, Q, não depedende apenas de Q mas também de todos os subconjuntos de Q. (ii) Ao contrário do que acontece nos STP, os níveis de significância individuais de cada teste não são conhecidos.

Relativamente ao exemplo que temos vindo a utilizar, a estratégia de Spjøtvoll baseada na estatística de Wilks levaria a que se incluíssem em A_α todos os conjuntos para os quais $\Lambda_{\bar{Q}|Q}$ seja superior aos pontos críticos de uma distribuição de Λ com parâmetros $d = p-q = 17-q$, $m_H = k-1 = 2$ e $m_E = N-k-q = 30-q$, Para $\alpha = 0.10$ e $q \leq 4$ os pontos críticos são 0.123 ($q = 1$), 0.133 ($q = 2$), 0.145 ($q = 3$), 0.157 ($q = 4$), 0.170 ($q = 5$), 0.187 ($q = 6$) e 0.205 ($q = 7$). Os menores conjuntos a incluir em $A_{0.10}$ são os seguintes conjuntos de 7 variáveis: $Q7$ = {ln TRCC, SB, TMA, TMR, MF, ln RCPD, ALE}, {ln TRCC, SB, TMA, TMR, MF, MN, ln RCPD}, $Q8$ = {ln TRCC, SB, TMA, TMR, MF, MN, ln RCPD} e $Q9$ = {CCG, ln TRCC, SB, TMA, TMR, MF, ln RCPD}. Verificamos então, que a utilização de testes menos conservadores permite neste caso reduzir consideravelmente o número de conjuntos que se aceitam como adequados.

4. DIMENSÕES DE SEPARAÇÃO E ÍNDICES DE COMPARAÇÃO ENTRE SUBCONJUNTOS

Um segundo problema dos métodos tradicionais de selecção de variáveis é o facto de estes métodos não partirem de um conceito claro e universal de "separação entre grupos" conducente a uma única medida objectiva deste conceito. Com efeito, iremos aqui argumentar que em ADD o conceito da "separação" providenciada por um determinado conjunto de variáveis deverá ser entendido tendo em atenção a capacidade desse conjunto para satisfazer os objectivos da análise, ou seja a sua capacidade para explicar cabalmente todas as diferenças entre os grupos que se considerem relevantes. Tendo em atenção esses objectivos, para problemas diferentes os índices de separação mais adequados poderão também ser diferentes. Nesta secção iremos discutir o problema de escolher um índice adequado para comparar subconjuntos de variáveis em ADD. Ao longo da discussão adoptaremos uma perspectiva geométrica evitando recorrer a pressupostos baseados em modelos probabilísticos.

4.1 ADD NUMA PERSPECTIVA GEOMÉTRICA

Suponha-se que se pretende avaliar a qualidade de um determinado subconjunto, Q, formado por q variáveis tendo em vista a explicação das diferenças entre os grupos. Sejam $\mathbf{X}_Q$ a matriz $(N*q)$ das observações em Q, $\mathbf{G}$ uma matriz $(N*k)$ de indicadores do grupo de origem e $\mathbf{1}_N = [1 \ldots 1]^T$ um vector coluna (com N linhas) constituído unicamente por 1's. Sejam ainda, Ω o subspaço de $\Re^N$ gerado pelas colunas de $\mathbf{G}$, ω o subspaço de Ω gerado por $\mathbf{1}_N$, $\Omega^\perp$ e $\omega^\perp$os complementos ortogonais de Ω e ω em $\Re^N$, $\omega^p = \omega^\perp \cap \Omega$ o complemento ortogonal de ω em Ω, e γ o subspaço de $\omega^\perp$ gerado pelas colunas centradas (em relação às médias globais) de $\mathbf{X}_Q$. Então, $\mathbf{X}_Q$ tem uma representação única em termos das suas projecções ortogonais em Ω ($\mathbf{P}_\Omega \mathbf{X}_Q$) e $\Omega^\perp$ ($\mathbf{P}_{\Omega^\perp} \mathbf{X}_Q$) enquanto $\mathbf{P}_\Omega \mathbf{X}_Q$ tem uma representação única nas suas projecções ortogonais em ω ($\mathbf{P}_\omega \mathbf{P}_\Omega \mathbf{X}_Q = \mathbf{P}_\omega \mathbf{X}_Q$) e ω^p ($\mathbf{P}_{\omega^p} \mathbf{P}_\Omega \mathbf{X}_Q = \mathbf{P}_{\omega^p} \mathbf{X}_Q$). Essas representações permitem escrever $\mathbf{X}_Q = \mathbf{P}_\Omega \mathbf{X}_Q + \mathbf{P}_{\Omega^\perp} \mathbf{X}_Q = \mathbf{P}_\omega \mathbf{X}_Q + \mathbf{P}_{\omega^p} \mathbf{X}_Q + \mathbf{P}_{\Omega^\perp} \mathbf{X}_Q$, em que $\mathbf{P}_\Omega \mathbf{X}_Q$ é a matriz das médias por grupo, $\mathbf{P}_{\Omega^\perp} \mathbf{X}_Q$ é a matriz dos desvios em relação às médias por grupo, $\mathbf{P}_\omega \mathbf{X}_Q$ é a matriz das médias globais e $\mathbf{P}_{\omega^p} \mathbf{X}_Q$ é a matriz dos desvios das médias por grupo em relação às médias globais.

Quanto maior for a separação entre grupos maior será a importância da parcela $\mathbf{P}_{\omega^p} \mathbf{X}_Q$ nesta representação. Com efeito, as técnicas clássicas de ADD podem ser apresentadas no contexto de uma Análise de Correlação Canónica entre γ e ω^p. Nomeadamente, nessa análise o vector de $\Re^N$ definido pelos scores (centrados) dos indivíduos na primeira FDL, define a primeira direcção canónica em γ, e o primeiro valor próprio de $\mathbf{B}\,\mathbf{T}^{-1}$ (l_1) iguala o cosseno quadrado do ângulo entre a primeira FDL e ω^p (Masson 1990). Nesse sentido l_1 pode ser entendido como uma medida da separação ao longo da dimensão definida pela primeira FDL. De igual modo, a i-ésima FDL (FDL_i) define a i-ésima direcção canónica em γ, enquanto o i-ésimo valor próprio de $\mathbf{B}\,\mathbf{T}^{-1}$ (l_i) iguala o cosseno quadrado entre

FDL_i e o vector que define a i-ésima direcção canónica em ω^{p}[12]. Desta forma, é possível definir índices de separação entre os grupos a partir dos valores próprios de $\mathbf{B}\,\mathbf{T}^{-1}$. Estes índices medem a proximidade em $\Re^N$ entre as posições relativas de ω^p e do espaço gerado pelos elementos (centrados) de $\mathbf{X}_Q$, havendo uma correspondência entre a importância dada a l_i e o ênfase atribuído à dimensão de separação definida pela i-ésima FDL.

4.2 INDICES DE COMPARAÇÃO ENTRE SUBCONJUNTOS

Para situar neste contexto a estratégia habitual de avaliar a separação entre grupos através do valor da estatística de Wilks (Λ), convém notar a equivalência entre a minimização de Λ e a maximização do índice de separação que lhe está associado

$$\tau^2 = 1 - \Lambda^{1/r} = 1 - \left(\prod_{i=1}^{r} (1 - l_i) \right)^{1/r}$$

e que é simplesmente o complementar para a unidade da média geométrica dos senos quadrados dos ângulos canónicos entre γ e ω^p. Ou seja, τ^2 é um índice que por um lado considera todas as dimensões de separação possíveis, mas por outro lado tende a dar um maior ênfase às primeiras dimensões, uma vez que uma média geométrica de valores compreendidos entre 0 e 1, tende a dar maior importância (pelo menos em comparação com uma média aritmética) aos valores mais próximos de zero, e as FDLs estão ordenadas pela da sua "proximidade" (medida aqui pelo ângulo canónico respectivo) com ω^p. Põe-se então a questão de saber se τ^2 é o índice mais adequado para todos os problemas de ADD. Vai-se aqui argumentar que sendo em geral, τ^2 um índice "razoável", para alguns problemas poderão haver outros índices mais apropriados. Nomeadamente, se todas as diferenças relevantes poderem ser explicadas ao longo de uma única dimensão de separação, um índice mais apropriado será l_1, o cosseno quadrado do ângulo entre FDL_1 e ω^p, uma vez que nesse caso os restantes valores próprios de $\mathbf{B}\,\mathbf{T}^{-1}$ resultam ou de ruído resultante da variação amostral, ou de dimensões de separação consideradas como pouco interessantes pelo analista. Por outro lado, se todas as dimensões de separação possíveis representarem diferenças reais e forem consideradas como igualmente importantes, um índice mais adequado será

$$\xi^2 = \frac{\mathrm{tr}(\mathbf{B}\mathbf{T}^{-1})}{r} = \sum_{i=1}^{r} \frac{l_i}{r},$$

[12] Esta identidade resulta da igualdade entre os quadrados dos coeficientes de correlação canónica e os valores próprios de $\mathbf{B}\,\mathbf{T}^{-1}$ e da interpretação geométrica de um coeficiente de correlação.

que iguala a média aritmética dos cossenos quadrados de todos os ângulos canónicos. O leitor mais atento, certamente reconhecerá que enquanto τ^2 é o índice multivariado associado com a estatística de Wilks, l_1 e ξ^2 são os índices associados respectivamente com as estatísticas de Roy e de Bartlett-Pillai. Torna-se então *natural indagar como se situa neste contexto o índice*

$$\varsigma^2 = \frac{tr(\mathbf{B}\mathbf{W}^{-1})}{r + tr(\mathbf{B}\mathbf{W}^{-1})} = 1 - \frac{r}{\sum_{i=1}^{r} \frac{1}{1-l_i}}$$

associado com a estatística de Lawley-Hotelling. Notando que a relação entre ς^2 e os valores próprios de $\mathbf{B}\,\mathbf{T}^{-1}$ permite exprimir ς^2 como o complementar para a unidade da média harmónica dos senos quadrados de todos os ângulos canónicos, verificamos que ς^2 é um índice semelhante a τ^2, mas que ainda põe maior ênfase nas dimensões definidas pelas primeiras FDLs.

4.3 NÚMERO E IMPORTÂNCIA DAS DIMENSÕES DE SEPARAÇÃO

Vemos então que a questão da escolha de um índice apropriado da contribuição de $\mathbf{X}_Q$ para a separação entre os grupos está intimamente ligada ao ênfase que se pretende dar a cada dimensão de separação. No caso de só existirem dois grupos, só é possível definir uma dimensão de separação e todos os índices dão resultados equivalentes[13]. Porém no caso de existirem mais do que dois grupos o problema pode não ter uma resposta evidente. Por um lado, é importante distinguir as FDLs que representam dimensões reais de separação, das FDLs que estão apenas associadas a variabilidade amostral. Por outro lado, é necessário determinar a importância que cada dimensão real tem para o analista.

Quando se assumem os pressupostos clássicos de normalidade multivariada e igualdade das variâncias e covariâncias intra-grupos a primeira destas questões pode ser respondida com a ajuda de testes de hipóteses conhecidos, os chamados testes de dimensionalidade. Estes testes formalizam a hipótese de que só *t* FDLs correspondem a "dimensões reais" de separação da seguinte forma. Sejam

$$\mu_g = E(\mathbf{X}_g) \qquad \bar{\mu} = \sum_{g=1}^{k} \frac{n_g \mu_g}{N}$$

[13] Que são ainda equivalentes aos resultados que se obtêm se tomar como índice de separação a distância de Mahalanobis entre os centroides de cada grupo, $D = \left[(\bar{x}_1 - \bar{x}_2)^T S^{-1} (\bar{x}_1 - \bar{x}_2)\right]^{1/2}$ em que S = W.(N-2) é a matriz amostral de covariâncias intra-grupos.

Esta equivalência resulta das igualdades $D^2 = \frac{N(N-2)}{n_1 n_2}\lambda_1 = \frac{N(N-2)}{n_1 n_2}\frac{l_1}{1-l_1}$ que definem relações monótonas entre D e os valores próprios positivos de $\mathbf{B}\,\mathbf{W}^{-1}$ e $\mathbf{B}\,\mathbf{T}^{-1}$.

as médias populacionais, por grupo e globais dos vectores aleatórios $\mathbf{X}_g$. Sejam ainda

$$\Sigma = E\left[(X_g - \mu_g)(X_g - \mu_g)^T\right] \quad \Theta = \sum_{g=1}^{k} n_g (\mu_g - \bar{\mu})(\mu_g - \bar{\mu})^T$$

as matrizes de desvios populacionais "intra" e "entre" grupos. Então o problema de determinar o número de "dimensões reais" de separação é equivalente à identificação da característica de matriz de não-centralidade $\Psi = \Theta\ \Sigma^{-1}$ (ver, por exemplo, Krishnaiah, 1982)[14]. Com efeito, se Ψ tiver característica t ($t \leq r$), então as t primeiras correlações canónicas populacionais são positivas enquanto as r-t últimas são iguais a 0. Esse resultado é geralmente interpretado como significando que existem exactamente t "dimensões reais" de separação. A hipótese nula de que a característica de Ψ é igual ou inferior a t pode ser testada com base em várias funções dos últimos r-t valores próprios de $\mathbf{B}\ \mathbf{W}^{-1}$ (λ_{t+1}, ..., λ_r). Por exemplo, uma estatística comum, devida a Bartlett (1947), é

$$T = \left(N - 1 - \frac{q+k}{2}\right) \sum_{i=t+1}^{r} ln(1+\lambda_i)$$

Sob a hipótese nula, T segue assintóticamente uma distribuição do Qui-quadrado com $(q - t)*(k - 1 - t)$ graus de liberdade (ver Krishnaiah, 1982, para uma descrição de outras funções de λ_{t+1}, ..., λ_r que podem ser usadas como estatísticas em testes de dimensionalidade).

A realização de testes de dimensionalidade pode, no entanto, não ser suficiente para determinar quais (nem a importância) as dimensões a considerar. Em primeiro lugar, os pressupostos clássicos dos testes de dimensionalidade raramente estão satisfeitos na prática, pelo que as suas conclusões são frequentemente questionáveis. Apesar destes testes serem relativamente robustos e por conseguinte defensáveis desde que não existam desvios substanciais aos seus pressupostos, há problemas a que eles não conseguem responder de uma forma completamente satisfatória. Nomeadamente, tal como acontece com os testes de informação adicional, os testes de dimensionalidade só são válidos quando a análise é baseada num conjunto de variáveis escolhido à partida havendo enviesamentos de efeitos mal conhecidos quando se efectua uma selecção prévia de variáveis. Finalmente, pode acontecer que algumas dimensões de separação, ainda que estatisticamente significativas, sejam consideradas como "pouco interessantes" e sejam por essa razão ignoradas pelo analista, ou que não haja evidência suficiente para concluir pela existência real de algumas dimensões que o analista considera suficientemente importantes para considerar. Situações do primeiro tipo poderão acontecer quando as ultimas FDLs, embora representando

[14] Estamos a supor que Σ é uma matriz não singular. Note-se ainda que a característica de Ψ é idêntica à característica de Θ. No entanto, dado que por um lado a separação entre os grupos pode ser descrita a partir dos valores e vectores próprios de Ψ e por outro lado os testes de dimensionalidade são baseados em valores próprios de $\mathbf{B}\ \mathbf{W}^{-1}$ que pode ser considerado como uma estimativa de Ψ (a menos de uma constante de proprocionalidade), é habitual caracterizar o número de dimensões de separação a partir de Ψ.

dimensões reais, tenham uma contribuição negligenciável para a separação dos grupos, ou quando elas estejam associadas a conceitos que não sejam relevantes para o estudo em questão. Situações do segundo tipo acontecem tipicamente, quando o reduzido numero de observações disponíveis não permita estabelecer a existência de dimensões associadas a FDLs cuja interpretação é teoricamente consistente e que caracterizem conceitos importantes para o estudo.

Por conseguinte, a escolha de dimensões a considerar é, no nosso entender, uma escolha eminentemente subjectiva em que se deve ter em atenção a natureza do problema e os objectivos do análise. Em geral, das r dimensões de separação possíveis, as primeiras s ($s \leq t \leq r$) corresponderão a dimensões reais com interesse e as seguintes t-s corresponderão a dimensões reais mas sem interesse, e as ultimas serão apenas o resultado de variação amostral. Um índice de separação ideal, deverá considerar as s dimensões relevantes de forma equilibrada e ignorar as restantes. Havendo dúvidas quanto ao número de dimensões a reter, o que poderá acontecer devido a testes de dimensionalidade inconclusivos, ou à existência de interpretações alternativas para as FDLs, índices que tal como τ^2 consideram todas as dimensões possíveis mas atribuem maior ênfase às primeiras, podem ser considerados como um compromisso razoável. No entanto, por vezes poderá haver interesse em utilizar índices que atribuiam maior ou menor ênfase nas primeiras dimensões de separação. Dois índices conhecidos com essas características são respectivamente ζ^2 e ξ^2. Frequentemente, mais do que a escolha de um único índice pode ser particularmente útil a comparação das ordenações dos "melhores" subconjuntos, resultantes da escolha de vários índices que dão ênfases diferentes a cada dimensão de separação. Essa comparação pode dar pistas importantes para uma melhor compreensão das diferenças entre os grupos, compreensão essa que constitui o objectivo último da análise.

Finalmente, um aspecto prático a ter em consideração reside na possibilidade de existirem limitações em termos computacionais quanto à capacidade de se identificarem e ordenarem os "melhores" subconjuntos de acordo com cada índice, num tempo razoável. Como vimos atrás, para ordenações baseadas em τ^2 o algoritmo de *Furnival--Furnival e Wilson--McCabe* pode ser utilizado, o que permite ultrapassar este problema para um número "moderado" de variáveis. Como é demonstrado em Duarte Silva (1998), estes algoritmos continuam a ser adaptáveis para ordenações baseadas em ζ^2 ou ξ^2, e no caso de r ser menor ou igual a 3, para ordenações baseadas em qualquer função monótona dos valores próprios de $\mathbf{B}\,\mathbf{T}^{-1}$.

4.4 EXEMPLO

Para o exemplo relativo às diferenças entre as instituições bancárias a operar em Portugal em 1993 os quadros 9 a) -- 9 d) mostram os dois melhores subconjuntos de cada dimensão de acordo com os índices τ^2, ζ^2, ξ^2 e l_1. A evolução destes índices para os melhores subconjuntos está representada nas figuras 1 e 2. É particularmente interessante comparar a evolução e os melhores subconjuntos seleccionados pelos índices l_1 e ξ^2. Nomeadamente, l_1 tem um crescimento claro desde $q = 1$ até $q = 4$, continuando a crescer de forma mais moderada de $q = 5$ até $q = 11$ e estabilizando para $q > 11$. Recordando que l_1 é um índice que só considera a primeira dimensão de separação, a sua evolução sugere que é possível capturar os aspectos principais dessa dimensão com 4 variáveis, sendo necessárias 11 variáveis para a descrever de uma forma completa.

Figura 1
EVOLUÇÃO DOS INDICES DE SEPARAÇÃO
(melhores subconjuntos)

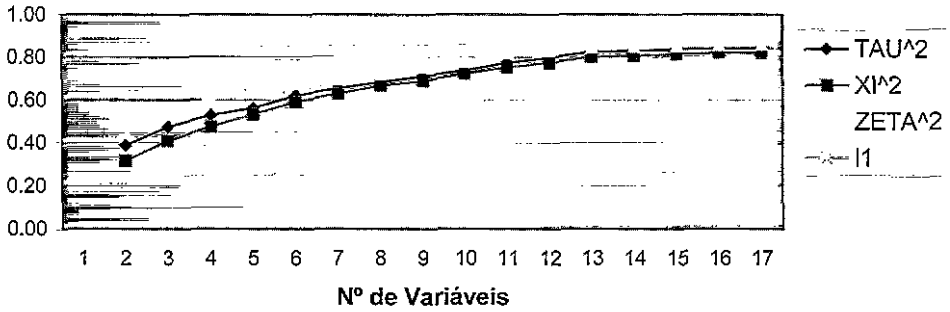
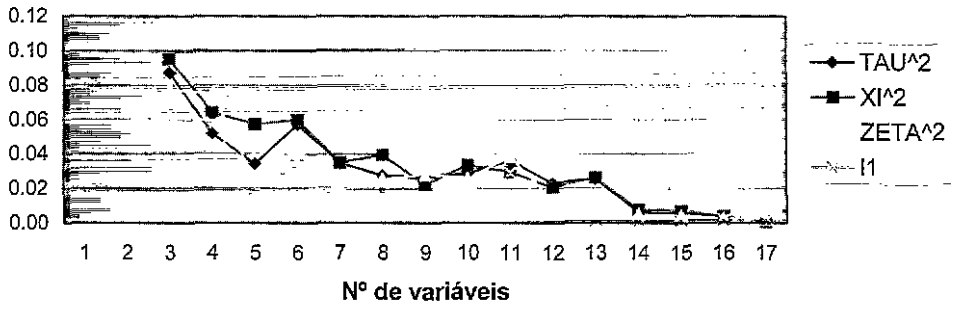


Figura 2
VARIAÇÃO DOS INDICES DE SEPARAÇÃO
(melhores subconjuntos)



A escolha das variáveis incluídas nos melhores subconjuntos (de acordo com I_1) para $q = 4$, *Q10* = {LR, ln TRCC, GE, MN}, e $q = 11$, *Q11* = {LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, ln RCPD, ln EB, ALE}, é consistente com essa hipótese. A evolução do índice ξ^2 é no entanto, diferente (ver figuras 1 e 2). Este índice mantém um claro crescimento até $q = 13$ só começando a estabilizar a partir deste ponto. Recordando que ξ^2 é o índice que pondera as duas dimensões de separação de forma mais equilíbrada, estes resultados sugerem que são necessárias 13 variáveis para as representar a ambas de forma completa. É de realçar que o melhor subconjunto com 13 variáveis de acordo com ξ^2 é o conjunto *Q3* que já havia sido proposto por um método descendente de selecção passo a passo, e pela comparação entre todos os subconjuntos baseada em Λ (ou τ^2). Este conjunto incluiu todas as variáveis importantes para a interpretação das duas dimensões de separação, nomeadamente todos os elementos de *Q11* e a *Rendibilidade Bruta dos Capitais Próprios*, que não fazendo parte de *Q11* é particularmente útil para a interpretação de FDL_2. Curiosamente o melhor conjunto de 13 variáveis de acordo com ζ^2 também é *Q3*, o que já não acontece quando se utiliza I_1 para índice de comparação. No entanto, a evolução de ζ^2 já não sugere a escolha de *Q3* de forma tão clara, podendo igualmente optar-se por um conjunto com 11 variáveis (ver figuras 1 e 2), aparecendo neste caso *Q10* em primeiro lugar. Como tínhamos visto na secção 4.2 de entre os índices τ^2, ξ^2 e

ζ^2 este último é aquele que está mais próximo de l_1, não sendo portanto surpreendente que l_1 e ζ^2 sugiram a escolha de subconjuntos semelhantes.

QUADRO 9
DOIS MELHORES SUBCONJUNTOS DE CADA DIMENSÃO

a) SEGUNDO O CRITÉRIO DE WILKS

q	Q	A	$T.U^2$
1	ln TRCC	0.4793	0.521
	RCPE	0.5234	0.477
2	CCG, MF	0.3726	0.390
	SB, MF	0.3819	0.382
3	CCG, SB, MF	0.2740	0.477
	CCG, MF, RBCP	0.2957	0.456
4	CCG, SB, MF, ln RCPD	0.2223	0.529
	CCG, SB, MF, RA	0.2238	0.527
5	LR, ln TRCC, GE, MN, RBCP	0.1909	0.563
	CCG, SB, MF, ln RCPD, ALE	0.1919	0.562
6	ln TRCC, SB, TMA, TMR, MF, ln RCPD	0.1442	0.620
	LR, ln TRCC, GE, MF, MN, RBCP	0.1522	0.610
7	ln TRCC, SB, TMA, TMR, MF, ln RCPD, ALE	0.1191	0.655
	ln TRCC, SB, TMA, TMR, MF, MN, ln RCPD	0.1199	0.654
8	ln TRCC, SB, TMA, TMR, MF, MN, ln RCPD, ALE	0.1008	0.683
	ln TRCC, SB, TMA, TMR, MF, MN, ln RCPD, RBA	0.1016	0.681
9	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, ln RCPD	0.0847	0.709
	ln TRCC, SB, TMA, TMR, MF, MN, ln RCPD, ALE, RBCP	0.0863	0.706
10	LR, ln TRCC, SB, TMA, TMR, MF, MN, RCPE, ln RCPD, RBCP	0.0689	0.738
	ln TRCC, SB, TMA, TMR, MF, MN, RCPE, ln RCPD, ALE, RBCP	0.0695	0.736
11	LR, ln TRCC, SB, TMA, TMR, MF, MN, RCPE, ln RCPD, ALE, RBCP	0.0512	0.774
	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, RBCP	0.0538	0.768
12	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ALE, RBCP	0.0412	0.797
	LR, ln TRCC, SB, TMA, TMR, MF, MN, RCPE, ln RCPD, ALE, RBA, RBCP	0.0446	0.789
13	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP	0.0314	0.823
	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ALE, RBA, RBCP	0.0379	0.805
14	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP	0.0292	0.829
	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA, RBCP	0.0294	0.828
15	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA, RBCP	0.0273	0.835
	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP, RCP	0.0278	0.833
16	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA, RBCP, RCP	0.0262	0.838
	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP, RA, RCP	0.0268	0.836

b) SEGUNDO O CRITÉRIO DE BARTLETT-PILLAI

q	Q	tr (BT^{-1})	XI^2
	ln TRCC	0.5207	0.521
1	RCPE	0.4766	0.477
	SB, MN	0.6328	0.316
2	SB, MF	0.6327	0.316
	CCG, SB, MF	0.8231	0.412
3	CCG, MF, RBCP	0.7894	0.395
	CCG, GE, MN, RBCP	0.9513	0.476
4	CCG, MN, ALE, RBCP	0.9344	0.467
	CCG, GE, MF, MN, RBCP	1.0660	0.533
5	SB, TMA, TMR, MF, RA	1.0450	0.523
	ln TRCC, SB, TMA, TMR, MF, ln RCPD	1.1858	0.593
6	CCG, ln TRCC, MF, MN, ALE, RBCP	1.1399	0.570
	ln TRCC, SB, TMA, TMR, MF, ln RCPD, ALE	1.2559	0.628
7	SB, TMA, TMR, MF, RCPE, ALE, RA	1.2390	0.620
	ln TRCC, SB, TMA, TMR, MF, RCPE, ALE, RA	1.3348	0.667
8	ln TRCC, SB, TMA, TMR, MF, MN, ln RCPD, ALE	1.3089	0.654
	ln TRCC, SB, TMA, TMR, MF, RCPE, ln RCPD, ALE, RA	1.3791	0.690
9	ln TRCC, SB, TMA, TMR, MF, MN, ln RCPD, ALE, RBCP	1.3735	0.687
	ln TRCC, SB, TMA, TMR, MF, MN, RCPE, ln RCPD, ALE, RBCP	1.4463	0.723
10	ln TRCC, SB, TMA, TMR, MF, MN, RCPE, ALE, RBCP, RA	1.4163	0.708
	LR, ln TRCC, SB, TMA, TMR, MF, MN, RCPE, ln RCPD, ALE, RBCP	1.5056	0.753
11	ln TRCC, SB, TMA, TMR, MF, MN, RCPE, ALE, RBA, RBCP, RA	1.4817	0.741
	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ALE, RBCP	1.5469	0.773
12	LR, ln TRCC, SB, TMA, TMR, MF, MN, RCPE, ln RCPD, ALE, RBA, RBCP	1.5410	0.770
	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP	1.5996	0.800
13	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln EB, ALE, RBCP, RA	1.5725	0.786
	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP, RA	1.6149	0.807
14	CCG, LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP	1.6145	0.807
	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA, RBCP	1.6290	0.815
15	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP, RCP	1.6270	0.814
	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA, RBCP, RCP	1.6370	0.819
16	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA, RBCP, RA	1.6334	0.817

c) SEGUNDO O CRITÉRIO DE LAWLEY-HOTELLING

q	Q	$\pi(BW^{-1})$	$ZETA^2$
	ln TRCC	1.0863	0.521
1	RCPE	0.9104	0.477
	CCG, MF	1.6838	0.457
2	SB, MF	1.5799	0.441
	CCG, SB, MF	2.2969	0.535
3	CCG, MF, RCP	2.2736	0.532
	LR, ln TRCC, GE, MN	3.1176	0.609
4	CCG, SB, MF, ln RCPD	3.0111	0.601
	CCG, SB, MF, ln RCPD, ALE	3.4697	0.634
5	CCG, ln TRCC, MN, ln RCPD, RBA	3.4283	0.632
	LR, ln TRCC, GE, ln RCPD, ln EB, ALE	4.1914	0.677
6	LR, CCG, ln TRCC, MN, ln RCPD, RBA	4.1833	0.677
	LR, CCG, ln TRCC, GE, MN, ln RCPD, RBA	4.6158	0.698
7	LR, ln TRCC, GE, ln RCPD, ln EB, ALE, RBCP	4.6026	0.697
	LR, ln TRCC, SB, GE, MN, ln RCPD, ln EB, ALE	5.2277	0.723
8	ln TRCC, SB, TMA, TMR, MF, MN, ln RCPD, RBA	5.1843	0.722
	LR, ln TRCC, GE, TMA, TMR, MF, MN, ln RCPD, RBCP	6.0374	0.751
9	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, ln RCPD	6.0081	0.750
	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, ln RCPD, RBA	6.9734	0.777
10	LR, ln TRCC, SB, GE, TMA, TMR, MF, ln RCPD, ln EB, ALE	6.9483	0.776
	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, ln RCPD, ln EB, ALE	8.8418	0.816
11	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, RBCP	8.2584	0.805
	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, ln RCPD, ln EB, ALE, RBA	9.6887	0.829
12	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, ln RCPD, ln EB, ALE, RBCP	9.5196	0.826
	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP	10.7540	0.843
13	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA	10.1110	0.835
	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP	11.1857	0.848
14	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA, RBCP	11.1685	0.848
	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA, RBCP	11.5756	0.853
15	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP, RCP	11.4097	0.851
	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA, RBCP, RCP	11.8339	0.855
16	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP, RA, RCP	11.7670	0.855

d) SEGUNDO O CRITÉRIO DE ROY

q	Q	λ_l	l_l
	ln TRCC	1.0863	0.521
1	RCPE	0.9104	0.477
	CCG, MF	1.6840	0.627
2	SB, MF	1.5552	0.609
	CCG, MF, RCP	2.2705	0.694
3	CCG, ln TRCC, MN	2.2027	0.688
	LR, ln TRCC, GE, MN	3.0324	0.752
4	CCG, SB, MF, ln RCPD	2.8394	0.740
	CCG, ln TRCC, MN, ln RCPD, RBA	3.3722	0.771
5	LR, ln TRCC, GE, MN, ln EB	3.2470	0.765
	LR, CCG, ln TRCC, MN, ln RCPD, RBA	4.1267	0.805
6	LR, ln TRCC, GE, ln RCPD, ln EB, ALE	4.0928	0.804
	LR, ln TRCC, GE, ln RCPD ln EB, ALE, RBCP	4.4245	0.816
7	LR, CCG, ln TRCC, MN, ln RCPD, ALE, RBA	4.3727	0.814
	LR, ln TRCC, SB, GE, MN, ln RCPD, ln EB, ALE	4.9925	0.833
8	LR, CCG, ln TRCC, GE, MN, ln RCPD, ln EB, ALE	4.8121	0.828
	LR, ln TRCC, GE, TMA, TMR, MF, MN, ln RCPD, RBCP	5.1958	0.839
9	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, ln RCPD	5.0599	0.835
	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, ln RCPD, RBA	5.9790	0.857
10	LR, ln TRCC, SB, GE, TMA, TMR, MF, ln RCPD, ln EB, ALE	5.9701	0.857
	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, ln RCPD, ln EB, ALE	7.7497	0.886
11	LR, ln TRCC, GE, TMA, TMR, MF, MN, ln RCPD, ln EB, ALE, RBCP	7.0717	0.876
	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, ln RCPD, ln EB, ALE, RBA	8.5044	0.895
12	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, ln RCPD, ln EB, ALE, RBCP	8.0755	0.890
	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, ln RCPD, ln EB, ALE, RBA	8.6428	0.896
13	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA	8.5632	0.895
	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA	8.7152	0.897
14	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA, RCP	8.6723	0.897
	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA, RCP	8.8891	0.899
15	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA, RBCP	8.8670	0.899
	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP, RA, RCP	9.0560	0.901
16	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA, RBCP, RCP	9.0368	0.900

Estes resultados ilustram a forma como a escolha do índice utilizado para comparar subconjuntos, influencia a importância dada às diferentes dimensões de separação. Antes de se escolher um índice particular, é conveniente tentar determinar quantas dimensões são de facto necessárias para separar os grupos. Tal poderá ser feito com a ajuda dos testes de dimensionalidade discutidos na secção 4.3. Neste caso, um teste de dimensionalidade (utilizando todas as variáveis) relativo à hipótese nula de que a matriz de não-centralidade, Ψ, tem característica 1, resulta num valor para a estatística T de 29.362 (p-value = 0.022). Por conseguinte, a um nível de significância de 5%, pode concluir-se pela existência de duas dimensões reais de separação. Havendo interesse em distinguir os bancos relativamente a estas duas dimensões que foram consideradas como igualmente importantes, privilegiaram-se neste caso as comparações baseadas em ξ^2. A comparação entre as ordenações de subconjuntos motivadas por ξ^2 e l_1 ajuda por sua vez a isolar as variáveis mais fortemente associadas a cada dimensão de separação. No caso de análises centradas na interpretações das diferenças associadas unicamente à estrutura de exploração, o índice privilegiado seria l_1.

5. CONCLUSÕES

A descrição de diferenças entre grupos a partir de um grande número de variáveis é um problema complexo que requer uma combinação inteligente de técnicas de ADD, com metodologias de selecção e comparação de subconjuntos de variáveis e conhecimentos substantivos sobre o problema a analisar. Este problema é tradicionalmente abordado ou através de métodos informais que partem da análise do conjunto completo de variáveis, ignorando posteriormente aquelas variáveis que se revelarem menos interessantes, ou através de selecções prévias baseadas em métodos de selecção passo a passo. A análise informal do conjunto completo de variáveis é geralmente um bom ponto de partida para compreender a natureza das diferenças mais importantes. No entanto, esta análise é eminentemente subjectiva e pode ser influenciada por um grande número de variáveis irrelevantes ou redundantes que apenas acrescentam ruído. Métodos automáticos de selecção passo a passo podem contribuir para ultrapassar estes dois problemas. Porém, estes métodos usam algoritmos heurísticos que não garantem a identificação dos subconjuntos mais apropriados. Por outro lado, embora alguns do métodos de selecção passo a passo mais importantes se baseiem em testes de hipóteses formais, esses testes são aí utilizados de uma forma que não permite a realização de inferências estatisticamente válidas. Metodologias de testes simultâneos, permitem realizar inferências sobre o conjunto A formado por todos os subconjuntos "adequados", os seja, os subconjuntos que incluem toda a informação relevante para explicar as diferenças observadas. Estas metodologias são particularmente úteis para identificar subconjuntos inadequados. Para o problema de identificar os subconjuntos adequados, estes métodos não conseguem dar uma resposta completamente satisfatória, uma vez que apenas controlam a probabilidade de se considerar como inadequados conjuntos adequados (erro de 1ª espécie). A escolha entre vários candidatos a conjuntos adequados pode fazer-se de uma forma exploratória comparando o valor de vários índices de separação. Por sua vez, a comparação de ordenações de subconjuntos sugeridas por índices que dão ênfases diferenciados às diferentes dimensões de separação, pode igualmente ajudar a compreender melhor a natureza das diferenças observadas.

Todas as técnicas descritas neste artigo podem ser facilmente generalizáveis a um contexto mais geral. Suponha-se que se dispõe de um conjunto de N observações em p variáveis, X_1, X_2, ..., X_p, cujos valores esperados podem ser representados num espaço vectorial $\Omega \subseteq \Re^N$ de dimensão u. Seja ω um subspaço de Ω com dimensão s, ω^p com dimensão $t = u - s$, o complemento ortogonal de ω em Ω e γ o espaço de dimensão $r = \min(t, p)$ gerado pelas projecções de X_1, X_2, ..., X_p no complemento ortogonal de ω em $\Re^N$. Então, é bem sabido que qualquer vector $x_i = [x_{i1}, x_{i2}, ..., x_{ip}]$ tem representações únicas, $x_i = a_i + b_i = a_i + c_i + d_i$ com $a_i \in \omega$, $b_i \in \gamma$, $c_i \in \omega^p$ e $d_i \in \Omega^\perp$, em que a_i, b_i, c_i e d_i são as projecções ortogonais de x_i em ω, γ, ω^p e $\Omega^\perp$. Nesta formulação, Ω pode ser definido a partir de um modelo linear apropriado,

$E(X) \in \omega$ define uma hipótese de referência supostamente falsa pretendendo-se descrever e interpretar desvios em relação a essa hipótese. Tal poderá ser feito a partir da análise dos vectores próprios de HE^{-1} ou HT^{-1} (que são idênticos), em que H, E e T = H + E são as matrizes das somas de quadrados e produtos cruzados para os vectores c_i (H), d_i (E), e b_i (T). Estes vectores próprios definem as direcções em γ, resultantes de uma análise de correlação canónica entre γ e ω^p e os valores próprios de HT^{-1} igualam os cosenos quadrados dos respectivos ângulos canónicos.

Esta formulação estende as técnicas de ADD à análise de qualquer "efeito multivariado" definido a partir de modelos e hipóteses lineares. No caso dos problemas clássicos de Análise Discriminante, Ω é o espaço gerado por k variáveis binárias que indicam os grupos de origem e a condição $E(X) \in \Omega$, especifica que $E(X_i)$ apenas depende do grupo a que o indivíduo i pertence. A hipótese de referência, é a igualdade das médias por grupo, o que equivale a impor que todos os vectores de valores esperados pertençam a um espaço unidimensional, ω. "Desvios" em relação à hipótese de referência são equivalentes a diferenças entre grupos. Notando a ligação entre a ADD clássica e os modelos MANOVA a um factor, alguns autores (por exemplo, Masson 1990) mostram como a formulação geral apresentado no parágrafo anterior, permite estender a aplicação das técnicas de ADD à interpretação de qualquer "efeito" considerado como significativo por uma MANCOVA ou MANOVA a mais de um factor. Para aplicar as técnicas discutidas neste artigo a problemas de selecção de variáveis visando a interpretação de qualquer "efeito" neste contexto geral, basta substituir as matrizes B e W por matrizes H e E apropriadas, mantendo-se a validade de todos os resultados e técnicas apresentadas.

REFERÊNCIAS BIBLIOGRÁFICAS

BARTLETT, M.S. (1947), "Multivariate Analysis", *Journal of the Royal Statistical Society Suppl.*, Vol 9, 176-190.

DUARTE SILVA, A.P. (1998), "Efficient Screening of Variable Subsets in Multivariate Statistical Models", *FCEE – Universidade Católica Portuguesa, C.R. Porto*, WP-98-004.

FURNIVAL, G.M. (1971), "All Possible Regressions with Less Computation" *Technometrics*, Vol. 13, 403-408.

FURNIVAL, G.M. E WILSON, R.W. (1974), "Regressions by Leaps and Bounds", *Technometrics*, Vol 16, 499-511.

GABRIEL, K.R. (1968), "Simultaneous Test Procedures in Multivariate Analysis of Variance", *Biometrika*, Vol 55, 489-504.

GABRIEL, K.R. (1969), "Simultaneous Test Procedures – Some Theory of Multiple Comparisons", *Annals of Mathematical Statistics*, Vol 40, 224-250.

HUBERTY, C.J. (1994), *Applied Discriminant Analysis*, Nova Iorque, John Wiley.

KOBILINSKY, A. (1990), "Analyse Factoriel Discriminante", in *Analyse Discriminante sur Variables Continues, G. Celeux (ed)*, 65-80, INRIA.

KRISHNAIAH, P.R. (1982), "Selection of Variables in Discriminant Analysis", in *Handbook of Statistics, Vol 2, P.R. Krishnaiah e L.N. Kanal (ed)*, 883-892, North-Holland Publishing Company.

MASSON, J.P. (1990), "Discrimination e Analyse de Variance", in *Analyse Discriminante sur Variables Continues, G. Celeux (ed)*, 81-99, INRIA.

MCCABE, G.P. (1975), "Computations for Variable Selection in Discriminant Analysis", *Technometrics*, Vol 17, 103-109.

MCKAY, R.J. (1977), "Simultaneous Procedures for Variable Selection in Multiple Discriminant Analysis", *Biometrika*, Vol 64, 283-290.

MCKAY, R.J. E CAMPBELL, N.A. (1982), "Variable Selection Techniques in Discriminant Analysis I. Description", *British Journal of Mathematical and Statistical Psychology*, Vol 35, 1-29.

RAO, C.R. (1973), *Linear Statistical Inference and its Applications, 2nd Ed.*, Nova Iorque, John Wiley.

SEBER, J.A.F. (1984), *Multivariate Observations*, Nova Iorque, John Wiley.

SPJØTVOLL, E. (1977), "Alternatives to Plotting C_p in Multiple Regression", *Biometrika*, Vol 64, 1-8.

Revista de Estatística
43ª Página
2º Quadrimestre de 1999

A Dinâmica Populacional das Cidades do Continente Português

Autor:
Henrique Albergaria

VOLUME 2

2º QUADRIMESTRE DE 1999

REVISTA DE ESTATÍSTICA
45ª PÁGINA
2º QUADRIMESTRE DE 1998

A DINÂMICA POPULACIONAL DAS CIDADES DO CONTINENTE PORTUGUÊS

THE DEMOGRAPHIC EVOLUTION OF PORTUGUESE CITIES

Autor: Henrique Albergaria
- Director Regional do INE

RESUMO:

- Neste artigo analisam-se alguns aspectos da dinâmica das cidades portuguesas do continente com base em informação demográfica que cobre um período de mais de um século. Essa reflexão incide essencialmente sobre a evolução da posição hierárquica desses centros urbanos procurando-se, por outro lado, esboçar um quadro comparativo dos sistemas de cidades das cinco regiões do Continente.

PALAVRAS-CHAVE:

- *Urbanização, cidades, demografia*

ABSTRACT:

- In this paper the author studies the dynamics of Portuguese urban spaces, based on demographic data series starting in 1864. This study focus on the changes in the ranking of urban centres comparing the different patterns of the five Portuguese mainland regions.

KEY-WORDS:

- *Urbanization, cities, demography*

VOLUME 2
2° QUADRIMESTRE DE 1999

REVISTA DE ESTATÍSTICA
47ª PÁGINA
2º QUADRIMESTRE DE 1999

A DINÂMICA POPULACIONAL DAS CIDADES PORTUGUESAS DO CONTINENTE

O principal objectivo deste artigo é o de analisar alguns aspectos da dinâmica das cidades portuguesas do Continente com base em informação demográfica que cobre um período de mais de um século. Essa reflexão incidirá essencialmente sobre a evolução da posição hierárquica desses centros urbanos procurando-se, por outro lado, esboçar um quadro comparativo dos sistemas de cidades das cinco regiões do Continente.

O apuramento dos dados necessários à realização do estudo exigiu que se estabelecesse um critério para identificar a área correspondente a cada uma das cidades e a sua correspondência com as unidades territoriais consideradas na colecta das estatísticas oficiais. Ora, não tendo esta questão sido ainda suficientemente estudada no nosso país, houve que ensaiar alguns critérios de modo a seleccionar aquele que poderia ser escolhido para definir a cidade do ponto de vista estatístico. Por isso começaremos por abordar os aspectos metodológicos relacionados com o conceito estatístico de cidade e as suas implicações práticas, servindo a análise que se lhe segue sobretudo para ilustrar o potencial de uma área de investigação em que os trabalhos escasseiam.

1 - POR UMA DEFINIÇÃO ESTATÍSTICA DO CONCEITO DE CIDADE

1.1 - PORQUÊ ESTATÍSTICAS URBANAS?

O século XX poderá vir a ficar na história da humanidade, como o século da mudança de uma sociedade rural para uma sociedade urbana: em 1900 menos de 10% da população mundial vivia em cidades enquanto que actualmente metade da população do planeta é urbana. O aumento da população urbana durante os últimos cem anos foi, pois, verdadeiramente exponencial, passando de 148 milhões de habitantes em 1900 a 2200 milhões em 1990 (POLÈSE, 1998).

É natural que neste contexto e em particular nos países mais desenvolvidos onde a taxa de urbanização ultrapassa os 70%, surjam repetidas provas da importância que a escala urbana foi adquirindo na sociedade actual, tanto ao nível dos processos de decisão como ao nível da análise dos problemas. Não surpreende também que a própria Comissão Europeia (COMISSÃO EUROPEIA, 1997), veja nas cidades os motores do crescimento económico, da competitividade e do emprego, embora alerte para a concentração crescente nas áreas urbanas dos problemas do desemprego, ambiente, criminalidade, pobreza e exclusão social.

É óbvio que esta acelerada concentração urbana gera uma procura de informação específica incluindo naturalmente aquela que se relaciona com os disfuncionamentos das áreas urbanas (criminalidade, poluição, congestionamentos, etc.) e que se destina a sustentar a análise, o planeamento e a tomada de decisão que tem por âmbito essas áreas. Esta é a razão de fundo pela qual se assiste a um reforço da importância das estatísticas urbanas na generalidade dos países incluindo, naturalmente, nos menos desenvolvidos, povoados também de grandes metrópoles que não param de crescer.

1.2 - A DIFICULDADE DOS CONCEITOS

Se é certo que a concentração urbana gera novas necessidades de informação estatística, não é menos verdade que a natureza eminentemente pluridisciplinar do objecto de estudo, dá origem a múltiplas perspectivas que remetem para espaços e a áreas territoriais distintos. As polémicas à volta da definição do conceito de urbano (e de rural), é um bom exemplo dessa diversidade de olhares embora, com o passar do tempo, se constate que o modo de vida urbano se está espalhar a toda a sociedade, esbatendo as dissemelhanças e tornando cada vez mais difícil distinguir o que é urbano do que é rural.

Contudo, quaisquer que sejam as particularidades de cada disciplina ou de cada analista, subsiste a questão concreta da delimitação territorial das áreas urbanas. Na prática, o critério mais habitualmente utilizado, embora nem sempre isoladamente, é o de considerar urbanas as áreas com uma densidade populacional relativamente elevada. Para um economista, este é um critério satisfatório pois, na sua óptica, a especificidade dos espaços urbanos nasce fundamentalmente da proximidade dos agentes económicos e dos contactos frequentes que essa concentração proporciona às *famílias e aos sectores de actividade (SULLIVAN, 1995).*

Também relativamente ao conceito de cidade surge o mesmo tipo de dificuldades. A cidade tanto pode ser considerada simplesmente como um lugar de forte concentração de pessoas e actividades, um espaço caracterizado pela continuidade da construção, um lugar de concentração de poder político, um lugar caracterizado pelo modo de vida, etc. (CAMAGNI, 1992). A discussão sobre o conceito de cidade é, na realidade, mais complexa pois apesar das cidades terem nascido há mais de 5000 anos no Médio Oriente e se terem espalhado progressivamente por todo o mundo (MUMFORD, 1961; BAIROCH, 1988), pode questionar-se a pertinência de um conceito que pretende abarcar formas tão distintas e que vão desde as cidades-estado da Grécia até às grandes metrópoles actuais, e nas quais é difícil vislumbrar uma essência comum (BAUMONT, 1996)).

Assim, apesar do que se referiu relativamente à densidade populacional, não será de estranhar que os critérios utilizados para definir as áreas urbanas variem de país para país, em função também da organização política, administrativa e territorial de cada um. Por exemplo, nos países que compõem a União Europeia, encontramos essa diversidade tanto ao nível da unidade territorial utilizada (aglomeração geográfica ou unidade administrativa) como nos limiares da população. Por razões óbvias é necessário definir conceitos harmonizados à escala europeia e por isso assiste-se actualmente a um gradual processo de convergência (EUROSTAT, 1992)

1.3 - A CIDADE EM PORTUGAL

Em Portugal é a Assembleia da República que decide da elevação dos lugares às categorias de vila e cidade. Segundo o art. 13º da lei 11/82 de 2 de Junho, "uma vila só pode ser elevada à categoria de cidade quando conte com um número de eleitores, em aglomerado populacional contínuo, superior a 8000 e possua, pelo menos, metade dos seguintes equipamentos colectivos: a) instalações hospitalares com serviço de permanência b) Farmácias c) Corporação de bombeiros d) Casa de espectáculos e centro cultural e) Museu e biblioteca f) Instalações de hotelaria g) estabelecimento de ensino preparatório e secundário h) Estabelecimento de ensino pré-primário e infantários".

No entanto (art. 14º), "importantes razões de natureza histórica, cultural e arquitectónica poderão justificar uma ponderação diferente dos requisitos enumerados nos art. 12º e 13º".

Constata-se, em primeiro lugar, que a margem de manobra consentida pelo art. 14º tem dado azo a um cada vez maior relaxamento das condições exigidas pelo art. 13º. Se considerarmos as vilas que foram elevadas à categoria de cidade depois da publicação da lei encontramos mais de metade que estão longe de cumprir o quesito relativo ao número de eleitores definido no art. 13º.

Para além dessa questão cujo efeito prático é o de tornar mais heterogéneo o grupo de unidades territoriais que designamos por cidades, uma outra se coloca relacionada com os limites das cidades. Com efeito, por razões várias, nem sempre os limites definidos pelo INE para os lugares cidade coincidem com os limites que a autarquia considera para a cidade. Em geral, os órgãos autárquicos identificam esses limites com o perímetro urbano definido nos Planos Directores Municipais. Por altura da revisão periódica dos PDM esses limites podem ser alterados.

Em suma, a insuficiente articulação entre o sistema estatístico oficial e as autarquias nesta matéria introduz uma complexidade dispensável na obtenção da informação necessária à gestão e análise das áreas urbanas.

1.4 - UMA PROPOSTA DE CONCEITO ESTATÍSTICO DE CIDADE

Enquanto não for possível estabelecer as regras que permitam facilmente converter a área da cidade em unidades territoriais estatísticas, torna-se necessário definir um conceito estatístico da cidade que possibilite a realização de análises e estudos.

O critério que aqui se defende parte do conceito de lugar cidade definido pelas estatísticas oficiais. Como é óbvio, esses lugares, definidos pela continuidade da construção, podem repartir-se por várias freguesias. Nessas condições parece aceitável considerar como aproximação estatística da cidade o território correspondente ao conjunto de freguesias pelas quais ela se espalha, eliminando

aquelas em que a percentagem da população da cidade aí residente não atinge os 50% da população da freguesia.

Um argumento em favor desta metodologia é a de estar de acordo com a prática de vários países da Europa nos quais, sempre que os limites das unidades administrativas não coincidem com os limites estatísticos, se aplica a citada regra dos 50%, *isto é, toma-se como aproximação o nível territorial correspondente ao mais* baixo nível administrativo.

A escolha da freguesia como unidade territorial de referência oferece várias vantagens entre as quais a de constituir a base de referência para diversas fontes de informação. Assinale-se, por outro lado, que em nenhuma freguesia do país existe mais de uma cidade.

Em termos de análise, os enviusamentos associados a esta opção metodológica não serão significativos, principalmente se o objectivo for analisar a evolução de variáveis, estudar repartições ou hierarquias. Para outros fins, por exemplo análise da actividade económica ou do emprego, e para algumas cidades em particular, a situação poderá ser diferente.

1.5 - RESULTADOS DA APLICAÇÃO DO MÉTODO

Um primeiro exercício realizado com a base de dados construída para este estudo, consistiu na comparação da população das cidades com a população do conjunto das freguesias pelas quais se repartem. Nesse ensaio, agruparam-se as cidades em classes, desde o grupo que inclui as cidades cuja população corresponde exactamente à soma da população das freguesias pelas quais se espalham (100%), até ao grupo que inclui as cidades que, para o mesmo indicador, apresenta um valor inferior a 20%. Os resultados, inscritos no QUADRO 1, suscitam alguns comentários.

Desde logo, para um número muito significativo de cidades (66 em 111), que no seu conjunto representa mais de 85% da população citadina do Continente, a diferença entre a respectiva população e a população total das freguesias pelas quais se repartem é, em média, muito pequena, atingindo um valor máximo compreendido entre 10 e 20 pontos percentuais para apenas 17 das 66 cidades referidas.

No outro extremo estão 16 cidades cuja população representa menos de 50% da população das freguesias pelas quais essas cidades se repartem. Note-se, no entanto, que a população deste conjunto de cidades representa no total apenas cerca de 1,8% da população que vive nos lugares cidade do Continente.

QUADRO 1. RELAÇÃO ENTRE A POPULAÇÃO DAS CIDADES DO CONTINENTE E A POPULAÇÃO DAS FREGUESIAS PELAS QUAIS ELAS SE REPARTEM

Classe	Taxa de coincidência	Número de Cidades	Pop. da Freg. Resid. na Cidade	População Resid. na Freguesia	Diferença
1	2= (4/5) x 100	3	4	5	6= 5-4
1	Igual a 100	19	1 277 004	1 276 671	-333
2	Entre 90 e 100	30	829 798	878 573	48 775
3	Entre 80 e 90	17	228 707	270 091	41 384
4	Entre 70 e 80	12	235 937	315 548	79 611
5	Entre 60 e 70	7	48 190	73 103	24 913
6	Entre 50 e 60	10	75 385	135 014	59 629
7	Entre 40 e 50	1	3 734	8 818	5 084
8	Entre 30 e 40	4	11 273	30 569	19 296
9	Entre 20 e 30	5	15 741	56 292	40 551
10	Menos de 20	6	17 624	164 572	146 948
Total		111	2 743 393	3 209 251	465 858

Verifica-se, por outro lado, que mesmo admitindo que as cidades correspondem em termos populacionais às freguesias pelas quais se repartem, essa opção metodológica só é insatisfatória para um máximo de 30% das cidades do Continente (taxa de coincidência inferior a 70%) que no seu conjunto representam apenas 6,3% da população citadina do Continente.

O exercício seguinte realizado com a base de dados consistiu na aplicação da metodologia aqui defendida baseada na exclusão das freguesias que não cumprem a regra dos 50%. Deparámos com três tipos de situações distintas (QUADRO 2):

Para um primeiro grupo de 80 cidades (grupo A), verificou-se que aplicação da regra conduzia a definir o território de cada cidade exactamente coincidente com o território do conjunto das freguesias pelas quais se repartia, sem excluir nenhuma.

Para um segundo grupo de cidades (grupo B), a aplicação da regra conduziu à definição de um território que correspondia ao território de apenas algumas das freguesias pelas quais elas se espalhavam (aquelas em que mais de metade da população pertencia ao lugar cidade).

Finalmente, para um terceiro grupo de cidades (grupo C) não foi possível, aplicar a regra dos 50%, quer por as cidades se repartirem por uma só freguesia da qual, em termos populacionais representam uma parte inferior a 50%; quer devido ao facto de alguns lugares cidades das estatísticas oficiais estarem actualmente especialmente desajustados à realidade que são supostos descrever. Assim, para este conjunto de 15 cidades (QUADRO 3) optou-se por considerar que a sua população correspondia à população das freguesias pelas quais essas cidades se repartiam mesmo quando os números indicavam que se tratava de uma pequena parte dessa população.

Em suma poder-se-á dizer que as distorções causadas pela aplicação deste critério não parecem excessivas e sobretudo podem ainda vir a ser a minoradas brevemente tendo em conta o projecto BGRI (Base Geográfica de Referenciação de Informação) em curso no INE.

QUADRO 2. GRAU DE APLICABILIDADE DO CRITÉRIO ESTATÍSTICO

Tipo	Nº de Cidades	Pop. da Freg. Resid. Cidade	População Resid. na Freguesia	População Estimada Cidades	Diferença
1	2	3	4	5	6=4-3
A	80	2 279 827	2 440 274	2 440 274	160 447
B	16	417 954	518 077	425 519	100 123
C	15	45 612	250 900	250 900	205 288
Total	111	2 743 393	3 209 251	3 116 693	465 858

QUADRO 3. CIDADES DO GRUPO C

Cidades	Pop. Freguesia Resid. na Cidade	%	População Estimada Cidades	População Residente Freguesias	Diferença
1	2	3=2/4x100	4	5	6=5-2
Águeda	3 841	39,2	9 792	9792	5 951
Albufeira	4 324	28,1	15 373	15373	11 049
Alcácer do Sal	3 734	42,3	8 818	8818	5 084
Amora	7 122	15,7	45 278	45278	38 156
Ermesinde	5 690	16,5	34 415	34415	28 725
Felgueiras	1 816	26,6	6 835	6835	5 019
Lixa	1 205	23,6	5 097	5097	3 892
Loures	5 636	28,7	19 636	19636	14 000
Marco Canaveses	282	9,9	2 843	2843	2 561
Paços de Ferreira	1 525	35,3	4 320	4320	2 795
Pombal	4 760	37,2	12 805	12805	8 045
Rio Tinto	1 260	3,1	40 907	40907	39 647
Seixal	2 652	9,5	28 026	28026	25 374
Vale de Cambra	1 147	31,4	3 652	3652	2 505
Valongo	618	4,7	13 103	13103	12 485
Total	45 612		250 900	250 900	205 288

2. A DINÂMICA POPULACIONAL DAS CIDADES DO CONTINENTE

A escolha das cidades como unidades de análise levanta por vezes algumas reservas. Uma opinião, que encontra eco em vários investigadores é que "entre nós cidades são aglomerados definidos como tal por via legislativa e a elevação de muitas

das povoações, se deve a causas aleatórias, as cidades apresentam uma grande diversidade de características que tornam duvidoso o seu tratamento conjunto e não aconselham a sua consideração como elemento privilegiado de estudo". (SALGUEIRO, 1992). Não partilhamos essa opinião, fundamentalmente porque é a natureza dos problemas e o nível de análise ou intervenção que determina a escala territorial mais adequada. Por exemplo, o conceito de região urbana em sentido lato, que entra em linha de conta com os movimentos pendulares e que corresponde grosso modo às bacias de emprego, é certamente um conceito incontornável para o desenvolvimento de análises territoriais que incidam sobre a organização do espaço, a localização das actividades, a determinação de zonas de influência, etc. Contudo, o conceito de cidade, a área que em Portugal é em geral delimitada pelo perímetro urbano, também constitui uma escala importante de análise territorial e um espaço de planeamento específico o que justifica, portanto, a sua individualização.

2.1 ALGUNS ASPECTOS DA DINÂMICA GLOBAL DAS CIDADES DO CONTINENTE

No fim de 1998, havia oficialmente em Portugal 121 cidades, das quais 111 no Continente. No QUADRO 4 estão agrupados as cidades segundo a sua dimensão (população) em momentos diferentes do tempo.

Verifica-se em primeiro lugar que a dimensão média das cidades do Continente é muito pequena (menos de 30 000 habitantes) o que é coerente com o facto de 70% das cidades terem menos de 20 000 habitantes.

No que respeita à evolução entre 1864 e 1991 constata-se que a população que vive nas áreas agora consideradas cidades triplicou durante o período, valor muito acima do crescimento populacional global do Continente durante o mesmo período (144,8%).

Curiosamente, as cidades de Lisboa e Porto propriamente ditas cresceram a uma taxa mais moderada mas, é claro, este resultado não pode ser interpretado isoladamente, fora do contexto da dinâmica das respectivas áreas metropolitanas que, obviamente, apresentam taxas de crescimento bem superiores.

Constata-se também que entre 1981 e 1991 a população global das cidades do Continente diminuiu em termos absolutos devido sobretudo ao decréscimo significativo da população nas cidades de Lisboa e Porto. A explicação estará na crescente terciarização das zonas centrais dessas cidades e na sua desertificação em termos residenciais em favor das periferias mais próximas, à semelhança do que tem acontecido em muitas cidades europeias onde, aliás, essa tendência está em retrocesso actualmente.

QUADRO 4. EVOLUÇÃO DA POPULAÇÃO DAS CIDADES DO CONTINENTE SEGUNDO A DIMENSÃO

Dimensão das cidades	Nº	1991	1981	1900	1864	Variação 1991-81	Tx. cresc. 1991/1864
1	2	3	4	5	6	7=3-4	8=(3-6)/6
Menos de 10000 hab.	39	241722	215203	397951	360719	26519	-0,33
De 10 a 20000 hab.	36	518851	533607	217561	92836	-14756	4,59
De 20 a 30000 hab.	17	413731	409964	51665	46348	3767	7,93
De 30 a 100000 hab.	16	844204	843769	32 98	89349	435	8,45
Mais de 100000 hab.	3	1098185	1259319	517939	190311	-161134	4,77
Total	111	3116693	3261862	1217604	779563	-145169	3,00

2.2 EVOLUÇÃO DA REPARTIÇÃO REGIONAL DA POPULAÇÃO DAS CIDADES DO CONTINENTE

Conforme se pode observar no *QUADRO* 5, existem diferenças significativas entre as cinco regiões NUT's II do Continente relativamente aos valores médios da população. Os valores mais elevados pertencem às regiões Lisboa e Norte em virtude do peso das duas áreas metropolitanas.

Por outro lado, constata-se que todas as muito pequenas cidades estão situadas a Norte do Tejo (menos de 5000 habitantes) bem como a maioria das pequenas (entre 5000 e 10000 habitantes). Os valores das densidades populacionais são muito diferenciados, fruto nomeadamente do tipo de povoamento e da superfície média das freguesias em cada região. A acção conjugada destes dois factores faz com que este indicador tenho um valor mais de vinte vezes superior no Norte ao que se regista no Alentejo.

QUADRO 5. AS CIDADES DO CONTINENTE POR REGIÃO

	Norte	Centro	Lisboa	Alentejo	Algarve	Continente
Menos de 5000	*8*	*3*	*0*	*0*	*0*	*11*
De 5000 a 10000	9	7	4	6	1	28
Mais de 10000	23	12	24	6	8	73
Nº de cidades	40	22	28	12	9	111
Pop. cidades	953333	403306	1426625	165949	167480	3116693
Pop. média	23 833	18 332	50 951	13 829	18 609	28 078
Área freguesias	535	1 083	1 056	2 262	690	5 626
Dens. freguesias	1 782	372	1 351	73	243	554

Fonte: INE, Censos 91

Os dados inscritos no *QUADRO* 6 permitem analisar a distribuição regional da população das cidades do Continente ao longo de mais de um século.

Os dados sugerem, em primeiro lugar, a ocorrência de mudanças significativas no que se refere ao peso relativo das diferentes regiões. A região de Lisboa e Vale do Tejo, que desde o início do período detém uma posição dominante, viu gradualmente

reforçada a sua importância no conjunto até atingir, em 1981, o seu valor máximo: nesse ano, metade da população citadina do Continente vivia na região de Lisboa. Em 1991 o peso relativo de Lisboa e Vale do Tejo diminuiu, embora no cômputo geral seja a única região que em termos relativos viu reforçada a posição que detinha em 1864.

A região Norte, independentemente de algumas flutuações ao longo do tempo, conseguiu manter praticamente inalterada em 1991 a parte que detinha em 1864, isto é, um pouco menos dos 31%. Sublinhe-se, também, que o conjunto de cidades do Norte, registou uma tendência forte de crescimento relativamente ao recenseamento imediatamente antes do último ao contrário do que sucedeu na região de Lisboa e Vale do Tejo.

No que se refere à região Centro, constata-se, durante o período, uma diminuição significativa, em termos relativos, da população das suas cidades que passa de 15,9% para 12,9% do total. Merece também a pena assinalar que pela primeira vez ao longo de mais de cem anos, se assistiu em 1991 a um crescimento da parte da população das cidades do Centro no total do Continente.

Alentejo e Algarve são as regiões nas quais se registou em termos relativos uma diminuição acentuada da população citadina entre 1864 e 1991. No entanto, as situações parecem bem diferenciadas na medida em que o Algarve apresenta, em duas décadas consecutivas, uma dinâmica de crescimento que contrasta significativamente com a regular diminuição, em termos relativos, da população das cidades do Alentejo. Assim, enquanto a parte da população das cidades do Algarve no total do Continente entre 1970 e 1991 aumenta de 4,1 para 5,4%, a do Alentejo diminui de 5,5 para 5,3% no mesmo período.

QUADRO 6. EVOLUÇÃO DA POPULAÇÃO CITADINA DO CONTINENTE POR REGIÃO 1864-1991

	1991	1981	1970	1960	1950	1940	1920	1900	1890	1864
Norte	953 333	933 944	780 348	742 284	667 239	599 664	448 501	386 457	346 597	248 255
Centro	403 306	391 167	322 054	304 275	282 848	249 483	188 885	172 741	163 692	123 596
Lisboa	1 426 625	1 606 348	1 348 317	1 207 706	1 100 729	952 555	665 756	481 347	416 138	277 419
Alentejo	165 949	175 030	149 386	165 181	161 956	143 275	102 241	89 919	83 390	66 022
Algarve	167 480	155 373	111 205	121 218	121 736	112 692	89 008	87 140	85 725	64 271
Continente	3 116 693	3 261 862	2 711 310	2 540 664	2 334 508	2 057 669	1 494 391	1 217 604	1 095 542	779 563

QUADRO 7. EVOLUÇÃO DA REPARTIÇÃO DA POPULAÇÃO CITADINA DO CONTINENTE 1864 - 1991

	1991	1981	1970	1960	1950	1940	1930	1920	1911	1900	1890	1864
Norte	30,59	28,63	28,78	29,22	28,58	29,14	28,96	30,01	31,15	31,74	31,64	31,85
Centro	12,94	11,99	11,88	11,98	12,12	12,12	12,45	12,64	13,43	14,19	14,94	15,85
Lisboa	45,77	49,25	49,73	47,54	47,15	46,29	45,63	44,55	42,10	39,53	37,98	35,59
Alentejo	5,32	5,37	5,51	6,50	6,94	6,96	6,97	6,84	7,08	7,38	7,61	8,47
Algarve	5,37	4,76	4,10	4,77	5,21	5,48	5,98	5,96	6,24	7,16	7,82	8,24
Continente	100	100	100	100	100	100	100	100	100	100	100	100

Fonte: Dados calculados a partir da informação censitária do INE

3. ALGUMAS REFLEXÕES SOBRE A DINÂMICA DAS SUBSISTEMAS REGIONAIS

3.1. A REDE DE CIDADES DA REGIÃO NORTE

Comecemos por apreciar, através dos dados relativos à população, algumas características da rede de cidades da região Norte (QUADROS 8 e 9).

Em primeira lugar, constata-se que a população a viver nas cidades da Região Norte praticamente quadruplicou, passando de 248.255 habitantes em 1864 a 953.333 em 1991.

Em seguida, constata-se que as posições relativas das cidades na hierarquia da região Norte sofreram oscilações significativas durante o período, a traduzir o efeito de factores de natureza diversa que foram afectando o crescimento de cada um dos centros urbanos. Nessas oscilações é possível vislumbrar as épocas de esplendor e de crescimento de cada cidade bem como os períodos mais sombrios de estagnação. Naturalmente que essas flutuações reflectem-se na hierarquia urbana aqui avaliada pela população.

Uma leitura simplificada da evolução ocorrida nos últimos 120 anos é-nos dada pelo QUADRO 9. Aí se pode verificar que das 40 cidades consideradas, 3 mantiveram a sua posição, 18 melhoraram-na e 19 desceram de ranking.

Dentre as cidades que mantiveram a sua posição na tabela, salienta-se naturalmente o Porto, mas também Braga desde sempre a segunda cidade do Norte.

Do conjunto de cidades que melhorou a sua posição registaram-se vários casos de subidas fulgurantes, a maioria das quais, como seria de esperar, pertencentes à área metropolitana: Ermesinde, Maia, S. João da Madeira, Trofa, Espinho e Santo Tirso subiram mais de dez posições entre 1864 e 1991. Matosinhos (9 posições), Fafe (7), Rio Tinto (6), Paços de Ferreira (5), Gondomar (5 posições), Felgueiras (3) e Paredes (3) são ainda casos de subidas significativas que merecem ser destacadas.

Das cidades que descerem de posição, destacam-se os casos de Vila Nova de Foz Côa (18), Lamego (15), Penafiel e Peso da Régua (14), Barcelos (13) e Chaves e Lixa (10). Mas porventura mais significativa é a descida de três capitais de distrito: Viana de Castelo (8 posições), Bragança (5) e Vila Real (3) a demonstrar que as dinâmicas económicas foram mais fortes que do que o impulso económico que lhes advém do seu estatuto de capital administrativa.

QUADRO 8. EVOLUÇÃO DA POPULAÇÃO DAS CIDADES DA REGIÃO NORTE ENTRE 1864 E 1991

Cidade	1991	1981	1970	1960	1950	1940	1920	1900	1890	1864
Área Metrop. Porto	591497	602018	512503	478484	433816	394094	294319	238868	207103	132756
Porto	302472	327368	306176	303424	281406	258548	202310	166729	146454	89349
Braga	85878	72122	52048	50831	46407	41549	30603	32498	31399	25669
V. Nova de Gaia	68566	62469	50219	44997	38003	34208	22881	19169	16587	10676
Guimarães	52982	53705	45165	43100	35615	30348	19935	19567	17714	15494
Rio Tinto	40907	47616	36895	27100	22269	18738	11105	7890	7569	4785
Ermesinde	34415	29555	15111	12197	9229	7375	4403	2733	2486	1396
Matosinhos	29798	30471	24317	24804	22294	21101	12276	7591	4910	3115
Maia	25885	18114	11654	9390	8162	6941	5144	3712	3404	2512
Póvoa de Varzim	23851	23729	17555	17696	16957	14664	12569	13291	12403	10012
Gondomar	20622	18881	14520	11182	9474	8882	6565	4889	4459	3553
Vila do Conde	19990	20613	16390	12771	11295	9710	7217	5530	5244	4356
S. João da Madeira	18452	16444	14195	11921	9266	7424	4407	3115	2876	2221
Bragança	16079	14379	10001	8662	8818	6595	5370	5310	5839	5093
Viana do Castelo	15562	15447	13451	14371	14023	13869	10717	10090	9765	9727
Vila Real	13809	12860	11202	10672	9285	7917	6232	6661	6014	4760
Valongo	13103	10351	7871	6124	6738	5914	3605	3643	3587	3002
Santo Tirso	12996	11610	10343	10428	8039	6715	4672	3546	2899	1905
Espinho	11888	12851	11795	8799	7989	8013	6244	3691	0	0
Fafe	11584	9871	8128	7126	6855	5966	4698	3615	3071	2080
Chaves	11453	11938	10594	12490	12239	9501	6851	6406	7730	4871
Trofa	11304	10372	7886	6023	4835	3586	2230	1540	1413	1007
Lamego	10630	11267	9671	10206	10288	10384	9086	9544	8840	7702
Peso da Régua	10277	10632	9002	8803	9258	9380	8045	6247	5501	4895
Oliveira de Azeméis	9679	8692	7683	5953	5268	4323	3270	2822	2699	2280
Amarante	8289	7214	6109	5522	4943	4453	3763	3457	3511	2965
Sta. Maria da Feira	8231	5966	5193	4220	3780	3436	2699	2650	2397	2098
Mirandela	8189	8156	5320	5979	5108	4159	2054	2974	2554	1890
Penafiel	7446	7014	5861	6022	6005	5429	5759	4997	4645	4016
Felgueiras	6835	5514	4545	4204	3318	2816	2116	2158	1766	1243
V. Nova de Famalicão	5243	4036	3190	3530	3356	3100	2284	2170	1917	1502
Paredes	5123	4340	3079	2672	2423	2290	1667	1394	1254	965
Lixa	5097	5215	4562	4368	4060	3656	3186	2907	2697	2416
Barcelos	4371	3807	4084	5420	4718	4780	3734	3483	3312	2639
Paços de Ferreira	4320	4123	3070	2549	2009	1640	1221	1002	861	657
Vizela	3799	3280	2489	2455	2055	1719	1207	1227	1059	673
Vale de Cambra	3652	3652	3325	2861	2468	2075	1374	935	868	778
Vila Nova de Foz Côa	2974	3710	2457	4129	4120	3825	3136	3571	3274	2867
Marco de Canavezes	2843	2578	1913	1665	1773	1716	1333	1197	1032	673
Esposende	2789	2189	1533	1751	1760	1629	1603	1524	1599	1499
Miranda do Douro	1950	1793	1746	5867	1331	1290	930	982	988	914
Total	953333	933944	780348	742284	667239	599664	448501	386457	346597	248255

QUADRO 9. EVOLUÇÃO DA CLASSIFICAÇÃO HIERÁRQUICA DAS CIDADES DA REGIÃO NORTE

Cidade	1991	1981	1970	1960	1950	1940	1930	1920	1900	1890	1864
Porto	1	1	1	1	1	1	1	1	1	1	1
Braga	2	2	2	2	2	2	2	2	2	2	2
V. Nova de Gaia	3	3	3	3	3	3	3	3	4	4	4
Guimarães	4	4	4	4	4	4	4	4	3	3	3
Rio Tinto	5	5	5	5	6	6	6	7	8	9	11
Ermesinde	6	7	9	11	16	17	21	22	29	28	31
Matosinhos	7	6	6	6	5	5	5	6	9	14	16
Maia	8	11	14	17	18	18	17	18	17	19	21
Póvoa de Varzim	9	8	7	7	7	7	7	5	5	5	5
Gondomar	10	10	10	13	12	13	13	13	16	16	15
Vila do Conde	11	9	8	9	10	10	10	11	13	13	13
S. João Madeira	12	12	11	12	14	16	20	21	25	24	24
Bragança	13	14	18	20	17	20	16	17	14	11	8
Viana do Castelo	14	13	12	8	8	8	8	8	6	6	6
Vila Real	15	15	15	14	13	15	15	15	10	10	12
Valongo	16	22	23	22	22	22	25	25	19	17	17
Santo Tirso	17	18	17	15	19	19	18	20	22	23	27
Espinho	18	16	13	19	20	14	14	14	18	40	40
Fafe	19	23	21	21	21	21	22	19	20	22	26
Chaves	20	17	16	10	9	11	11	12	11	8	10
Trofa	21	21	22	23	27	30	31	31	33	33	33
Lamego	22	19	19	16	11	9	9	9	7	7	7
Peso da Régua	23	20	20	18	15	12	12	10	12	12	9
Oliveira Azeméis	24	24	23	26	24	26	26	26	28	25	23
Amarante	25	26	25	28	26	25	24	23	24	18	18
Sta. Maria Feira	26	28	28	31	31	31	30	29	30	29	25
Mirandela	27	25	27	25	25	27	27	33	26	27	28
Penafiel	28	27	26	24	23	23	19	16	15	15	14
Felgueiras	29	29	30	32	33	33	33	32	32	31	32
V. N. Famalicão	30	33	33	34	32	32	32	30	31	30	29
Paredes	31	31	34	36	35	34	34	34	35	34	34
Lixa	32	30	29	30	30	29	29	27	27	26	22
Barcelos	33	34	31	29	28	24	23	24	23	20	20
Paços de Ferreira	34	32	35	37	37	38	39	38	38	39	39
Vizela	35	37	36	38	36	36	36	39	36	35	38
Vale de Cambra	36	36	32	35	34	35	38	36	40	38	36
Vila N. Foz Côa	37	35	37	33	29	28	28	28	21	21	19
Marco Canavezes	38	38	38	40	38	37	37	37	37	36	37
Esposende	39	39	40	39	39	39	35	35	34	32	30
Miranda do Douro	40	40	39	27	40	40	40	40	39	37	35

3.2. A REDE DE CIDADES DA REGIÃO CENTRO

Observemos agora algumas das características da rede de cidades da região Centro.

Em primeira lugar, constata-se que a população a viver nas cidades da Região Centro cresceu 226% durante o período considerado, (de 123.596 habitantes em 1864

a 403.306 em 1991), taxa que se situa bastante aquém da valor homólogo calculado para o Continente (300%).

Á semelhança de outras regiões, também no Centro houve alterações significativas nas posições relativas das cidades entre 1864 e 1991: das 22 cidades consideradas, 3 mantiveram a sua posição, 10 melhoraram-na e outras 9 desceram de ranking.

Das cidades que mantiveram a sua posição na tabela, salienta-se naturalmente Coimbra, cujo lugar no topo da hierarquia nunca esteve ameaçado embora no último recenseamento se registe uma aproximação das cidades que se lhe seguem na classificação.

Das cidades que melhoraram a sua posição, registam-se alguns casos de subidas fulgurantes, por exemplo a Marinha Grande que subiu 11 posições (mas que actualmente está em relativa perda de velocidade) ou Esmoriz (8 posições). Contudo, importa sobretudo assinalar as subidas significativas de Castelo Branco (4 posições), Aveiro (3), Leiria (3) e Guarda (3), todas capitais de distrito.

Das cidades que desceram de posição, o caso da Covilhã (4 posições) chamará porventura mais a atenção pelo facto de se tratar de uma cidade com uma dimensão importante no contexto regional e por se constatar uma tendência persistente de perda relativa de importância nas últimas décadas.

QUADRO 10. EVOLUÇÃO DA POPULAÇÃO DAS CIDADES DO CENTRO ENTRE 1864 E 1991

Cidade	1991	1981	1970	1960	1950	1940	1920	1900	1890	1864
Coimbra	86751	88804	68187	65031	60062	50771	36755	29933	28505	20679
Aveiro	37391	32851	26976	24067	22173	19035	13459	12442	11279	8292
Figueira da Foz	31662	25728	22290	23621	22913	20274	16561	14681	12589	8954
Castelo Branco	27004	23570	20792	17616	14865	12763	8798	7400	6712	6046
Marinha Grande	26628	25783	18860	15699	13092	10430	7059	5574	4825	3125
Leiria	25878	22173	15582	13928	13780	11268	8451	7227	6976	4933
Viseu	20659	20070	17636	16961	12613	12785	8268	8121	8101	6639
Covilhã	20571	23052	25606	23595	21385	19044	14030	15542	17559	8862
Guarda	18847	17948	13573	12787	11586	9391	7090	6197	6020	4182
Ílhavo	15204	14201	11181	12646	13114	12134	12691	13163	11276	8210
Ovar	14124	18783	16126	14128	13333	12799	10552	10976	11190	10359
Pombal	12805	12409	12441	9973	11353	10480	7374	5798	4318	4262
Esmoriz	9890	8538	7945	5955	5341	4240	3528	3079	2621	1952
Águeda	9792	12230	9371	8345	7522	6452	4381	3807	3938	3561
Mangualde	8570	8146	4616	6972	7223	6543	4996	5160	5038	4162
Fundão	7070	5792	5328	5651	5400	4783	3614	3182	2801	2375
Tondela	6797	7778	7190	7633	7542	7314	5639	5490	5219	4563
Seia	6465	5675	4173	3457	3340	3728	3269	2759	2637	2199
Cantanhede	6322	7534	6990	6630	6374	6027	4817	4296	4434	3953
Gouveia	3937	3944	2652	4215	4359	4135	3288	3400	3150	2600
Oliveira do Hospital	3510	2965	2141	2092	2166	1919	1601	1608	1537	1454
Pinhel	3429	3193	2398	3273	3312	3168	2664	2906	2967	2234
Total	403306	391167	322054	304275	282848	249483	188885	172741	163692	123596

QUADRO 11. EVOLUÇÃO DA CLASSIFICAÇÃO HIERÁRQUICA DAS CIDADES DA REGIÃO CENTRO

Cidade	1991	1981	1970	1960	1950	1940	1930	1920	1900	1890	1864
Coimbra	1	1	1	1	1	1	1	1	1	1	1
Aveiro	2	2	2	2	3	4	3	4	5	4	5
Figueira da Foz	3	4	4	3	2	2	2	2	3	3	3
Castelo Branco	4	5	5	5	5	7	7	7	8	9	8
Marinha Grande	5	3	6	7	9	11	11	12	12	13	16
Leiria	6	7	9	9	6	9	9	8	9	8	9
Viseu	7	8	7	6	10	6	8	9	7	7	7
Covilhã	8	6	3	4	4	3	4	3	2	2	4
Guarda	9	10	10	10	11	12	10	11	10	10	12
Ílhavo	10	11	12	11	8	8	6	5	4	5	6
Ovar	11	9	8	8	7	5	5	6	6	6	2
Pombal	12	12	11	12	12	10	12	10	11	15	11
Esmoriz	13	14	14	17	18	18	18	18	19	21	21
Águeda	14	13	13	13	14	15	15	16	16	16	15
Mangualde	15	15	18	15	15	14	16	14	14	12	13
Fundão	16	18	17	18	17	17	17	17	18	19	18
Tondela	17	16	15	14	13	13	13	13	13	11	10
Seia	18	19	19	20	20	20	20	20	21	20	20
Cantanhede	19	17	16	16	16	16	14	15	15	14	14
Gouveia	20	20	20	19	19	19	19	19	17	17	17
Oliveira do Hospital	21	22	22	22	22	22	22	22	22	22	22
Pinhel	22	21	21	21	21	21	21	21	20	18	19

3.3. A REDE DE CIDADES DA REGIÃO DE LISBOA E VALE DO TEJO

A população das cidades da Região de Lisboa e Vale do Tejo foi multiplicada por cinco durante o período em análise (de 277.419 habitantes em 1864 a 1.426.624 em 1991), o que corresponde ao maior crescimento regional do Continente (QUADRO 12).

Na Região de Lisboa registaram-se alterações especialmente significativas na hierarquia das cidades durante o período 1864 a 1991 (QUADRO 13). Aí se pode verificar que das 25 cidades para as quais foi possível reconstituir a série completa (ficaram de fora Amadora, Entroncamento e Queluz) apenas Lisboa manteve a sua posição.

QUADRO 12. EVOLUÇÃO DA POPULAÇÃO DAS CIDADES DE LISBOA E VALE DO TEJO

Cidade	1991	1981	1970	1960	1950	1940	1920	1900	1890	1864
Área Metr. Lisboa	1241325	1431437	1206326	1076679	968804	835471	587265	415783	355851	232417
Lisboa	663394	807937	769044	802230	783226	694389	484664	351210	300964	190311
Amadora	132319	124014	94069	47355	18789	9762	4062	0	0	0
Setúbal	85289	89867	60280	50966	50455	45345	37002	21722	17891	12728
Queluz	60370	48112	27679	15746	7968	4967	0	0	0	0
Odivelas	53531	84624	51037	27423	6772	3696	2635	1746	1592	1562
Barreiro	47901	46251	35756	23433	22190	19983	10859	5118	3682	2917
Amora	45278	34589	18028	7361	5044	3745	2701	2055	1263	1119
Seixal	28026	21873	12650	9426	8683	7248	5585	3650	3405	3567
Santarém	25019	23169	19099	18561	17113	13993	9918	8443	8143	5964
Alverca do Ribatejo	24168	24092	15192	7618	4665	3323	2736	1973	1786	1705
Montijo	24145	27257	32552	21947	19403	12287	9171	8113	7156	4666
Almada	22550	42684	42757	31523	17804	10755	11478	7749	6745	4011
Caldas da Rainha	21133	18394	13886	11185	11821	9605	6837	4605	4687	2268
Torres Vedras	19923	19096	14668	13091	12307	11908	8392	6853	6078	4135
Loures	19636	32874	13736	7623	6089	5013	4428	4829	4794	4515
Tomar	18636	18835	14837	12974	12250	11445	8053	6710	6063	4112
V. Franca de Xira	18487	19318	14459	13404	11228	10305	7498	5517	4696	4065
Sacavém	16231	27945	19087	10624	6488	4653	4446	2101	1877	1251
Peniche	15304	15455	12557	11388	10611	8780	5429	2781	2924	2963
Entroncamento	14226	11976	9421	7355	6804	6577	0	0	0	0
Torres Novas	12512	11427	10274	8578	9547	8304	6307	5568	5042	3867
Almeirim	10907	10632	8887	8902	11849	10535	7127	5941	4768	3181
Abrantes	10841	9628	8317	8172	11339	10309	7299	5815	5469	4863
Rio Maior	10424	10774	9681	9032	8402	6760	5372	4685	4199	3400
Cartaxo	9014	8526	6783	6665	6280	6947	5491	7171	6629	5177
Fátima	7213	7169	5898	5852	4719	3890	2536	2044	1760	1601
Alcobaça	5121	5305	3870	5166	4526	4227	2661	2323	2172	1458
Ourém	5027	4525	3813	4106	4357	3804	3069	2625	2353	2013
Total	1426625	1606348	1348317	1207706	1100729	952555	665756	481347	416138	277419

À semelhança do que se verificou na área metropolitana do Porto, também em Lisboa as cidades que mais significativamente subiram na hierarquia regional situam-se na sua área metropolitana: Odivelas, Amora, Queluz, Barreiro e Sacavém são os exemplos mais significativos. Fora da zona de imediata influência da capital apenas

haverá a destacar o caso de Caldas da Rainha que subiu sete posições durante o período.

As descidas em Lisboa e Vale do Tejo (e portanto também as subidas) tiveram uma amplitude muito maior do que nas outras regiões: 12 das 17 descidas que constam no QUADRO 14 foram iguais ou superiores a 5 posições. Cartaxo, Abrantes, Rio Maior, Loures e Torres Novas encabeçam a lista das cidades que mais desceram na hierarquia de cidades da região entre 1864 e 1991.

QUADRO 13. EVOLUÇÃO DA CLASSIFICAÇÃO HIERÁRQUICA DAS CIDADES DE LISBOA E V. DO TEJO

Cidade	1991	1981	1970	1960	1950	1940	1930	1920	1900	1890	1864
Lisboa	1	1	1	1	1	1	1	1	1	1	1
Amadora	2	2	2	3	5	12	15	20	20	20	20
Setubal	3	3	3	2	2	2	2	2	2	2	2
Queluz	4	5	8	9	18	21	23	24	24	24	24
Odivelas	5	4	4	5	20	27	26	27	28	27	23
Barreiro	6	6	6	6	3	3	3	4	13	16	17
Amora	7	8	11	23	24	26	28	25	24	28	28
Seixal	8	14	18	16	16	16	17	14	17	17	13
Santarém	9	13	9	8	7	4	4	5	3	3	3
Alverca Ribatejo	10	12	12	22	26	28	25	23	27	25	22
Montijo	11	11	7	7	4	5	5	6	4	4	6
Almada	12	7	5	4	6	8	10	3	5	5	11
Caldas da Rainha	13	18	16	14	11	13	13	12	16	14	18
Torres Vedras	14	16	14	11	8	6	6	7	7	7	8
Loures	15	9	17	21	23	20	19	19	14	11	7
Tomar	16	17	13	12	9	7	7	8	8	8	9
V. Franca de Xira	17	15	15	10	13	11	12	9	12	13	10
Sacavém	18	10	10	15	21	22	22	18	23	23	27
Peniche	19	19	19	13	14	14	11	16	18	18	16
Entroncamento	20	20	22	24	19	19	21	21	21	21	21
Torres Novas	21	21	20	19	15	15	14	13	11	10	12
Almeirim	22	23	23	18	10	9	8	11	9	12	15
Abrantes	23	24	24	20	12	10	9	10	10	9	5
Rio Maior	24	22	21	17	17	18	16	17	15	15	14
Cartaxo	25	25	25	25	22	17	18	15	6	6	4
Fátima	26	26	26	26	25	24	27	28	25	26	25
Alcobaça	27	27	27	27	27	23	20	26	22	22	26
Ourém	28	28	28	28	28	25	24	22	19	19	19

3.4. A REDE DE CIDADES DO ALENTEJO

A população das cidades do Alentejo aumentou de 66.022 habitantes em 1864 para 165.949 em 1991 (QUADRO 14). Apesar deste aumento significativo, a taxa de crescimento da população das cidades do Alentejo foi a mais baixa de todas as regiões.

A análise da evolução da posição hierárquica das cidades do Alentejo (QUADRO 15) mostra em primeiro lugar que Évora foi desde sempre e de forma indiscutível a primeira cidade da região. À semelhança das outras regiões, também no Alentejo se registaram alterações significativas no posicionamento hierárquico das diferentes cidades. Das 12 cidades consideradas 5 subiram de posição, 2 mantiveram e 5 desceram.

Das que subiram as maiores subidas foram as de Vendas Novas (5 posições), Sines (4 posições) e de Ponte de Sôr (3 posições). De registar também o bom comportamento de duas capitais de distrito, Beja e Portalegre que subiram duas posições.

A maior descida foi a de Estremoz (8 posições), seguindo-se um lote de quatro cidades que desceram duas posições (Elvas, Alcácer do Sal, Santiago do Cacém e Moura).

QUADRO 14. EVOLUÇÃO DA POPULAÇÃO DAS CIDADES DO ALENTEJO ENTRE 1864 E 1991

Cidade	1991	1981	1970	1960	1950	1940	1920	1900	1890	1864
Évora	42399	41102	34954	34145	31243	26416	16133	16004	15352	11078
Beja	22061	22193	18364	18040	16893	14145	10515	8839	8396	6640
Portalegre	16096	15824	12477	13374	13153	11422	9858	11899	10600	6609
Elvas	13393	12505	9729	11036	11107	10771	9082	11462	11077	7974
Sines	11253	12075	7150	8866	9534	8859	5586	3988	3580	3148
Montemor-o-Novo	10194	11246	9436	13115	12678	12318	9485	7176	6643	6058
Vandas Novas	9846	10933	8587	9675	10943	9051	5529	3107	2546	1863
Ponte de Sôr	9170	11611	10445	13010	12782	10802	6698	3847	3096	2196
Alcácer do Sal	8818	12131	13203	14733	14700	12515	8670	5953	6000	5693
Moura	8643	9259	9540	12126	11510	9610	6991	5946	5173	5451
Estremoz	8037	9375	9413	10122	10768	10015	8591	7510	6958	6646
Santiago do Cacém	6039	6776	6088	6939	6645	7351	5103	4188	3969	2666
Total	165949	175030	149386	165181	161956	143275	102241	89919	83390	66022

QUADRO 15. EVOLUÇÃO DA CLASSIFICAÇÃO HIERÁRQUICA DAS CIDADES DO ALENTEJO

Cidade	1991	1981	1970	1960	1950	1940	1930	1920	1900	1890	1864
Évora	1	1	1	1	1	1	1	1	1	1	1
Beja	2	2	2	2	2	2	2	2	4	4	4
Portalegre	3	3	4	4	4	5	4	3	2	3	5
Elvas	4	4	6	8	8	7	6	5	3	2	2
Sines	5	6	11	11	11	11	9	10	10	10	9
Montemor-o-Novo	6	8	8	5	6	4	5	4	6	6	6
Vandas Novas	7	9	10	10	9	10	11	11	12	12	12
Ponte de Sôr	8	7	5	6	5	6	8	9	11	11	11
Alcácer do Sal	9	5	3	3	3	3	3	6	7	7	7
Moura	10	11	7	7	7	9	10	8	8	8	8
Estremoz	11	10	9	9	10	8	7	7	5	5	3
Santiago do Cacém	12	12	12	12	12	12	12	12	9	9	10

3.5. A REDE DE CIDADES DO ALGARVE

Também a população das cidades do Algarve aumentou significativamente passando de 64.271 habitantes em 1864 para 167.480 em 1991 (QUADRO 16) mas a taxa de crescimento do conjunto das cidades foi bem superior (300% contra 161%).

A análise da evolução da posição hierárquica das cidades do Algarve (QUADRO 17) revela em primeiro lugar um aspecto inédito em relação às outras regiões que é o facto de a sua capital nem sempre ter sido a cidade mais importante do Algarve. Com efeito só a partir de 1930 é que Faro subiu ao topo da hierarquia quedando-se antes, a maior parte das vezes, pela quarta posição.
As alterações no posicionamento hierárquico durante o período 1864 a 1991, envolveu 6 cidades numa região que só conta com 9. As que subiram foram Portimão, Albufeira e Faro, todas cidades do litoral, e as que desceram foram Loulé (6 posições), Tavira (4) e Silves (1), todas cidades do interior.

QUADRO 16. EVOLUÇÃO DA POPULAÇÃO DAS CIDADES DO ALGARVE *ENTRE 1864 E 1991*

Cidade	1991	1981	1970	1960	1950	1940	1920	1900	1890	1864
Faro	39661	35628	22331	24877	22085	20100	12825	11336	9373	8097
Portimão	31223	26268	18452	17145	16684	14679	9154	7897	6961	5499
Olhão	25733	22894	16081	19291	20092	17935	14588	12965	11469	8811
Albufeira	15373	11979	7840	8416	8517	7760	7390	5816	4871	4078
Lagos	14378	12860	9967	10008	9526	9277	9572	8236	8381	7257
Tavira	11278	12046	9985	12046	13837	12267	11033	12242	11746	10343
Loulé	10978	10755	7926	9325	10796	11288	9268	12732	18984	12146
Silves	10674	9925	8309	9014	10237	10398	9570	9692	8396	5047
V. Real St. António	8182	13018	10314	11096	9962	8988	5608	6224	5544	2993
Total	167480	155373	111205	121218	121736	112692	89008	87140	85725	64271

QUADRO 17. EVOLUÇÃO DA CLASSIFICAÇÃO HIERÁRQUICA DAS CIDADES DO ALGARVE

Cidade	1991	1981	1970	1960	1950	1940	1930	1920	1900	1890	1864
Faro	1	1	1	1	1	1	1	2	4	4	4
Portimão	2	2	2	3	3	3	3	7	7	7	6
Olhão	3	3	3	2	2	2	2	1	1	3	3
Albufeira	4	7	9	9	9	9	9	8	9	9	8
Lagos	5	5	3	6	8	7	7	4	6	6	5
Tavira	6	6	5	4	4	4	4	3	3	2	2
Loulé	7	8	8	7	5	5	6	6	2	1	1
Silves	8	9	7	8	6	6	5	5	5	5	7
V. Real St. António	9	4	4	5	7	8	8	9	8	8	9

CONCLUSÃO

Neste artigo foram apresentados os primeiros resultados de um projecto de investigação em curso sobre a dinâmica das cidades do Continente baseada na exploração de dados demográficos apurados em todos recenseamentos efectuados em Portugal desde 1864. Nesta fase do trabalho, duas conclusões principais emergem da análise dos resultados dos exercícios realizados com a informação estatística recolhida para o efeito.

A primeira, de natureza metodológica, é que a proposta de associar a cidade estatística à soma das freguesias pelas quais ela se reparte, excluindo aquelas em que a população da cidade não atinge a fasquia dos 50%, parece ser uma base aceitável para o desenvolvimento da investigação na medida em que não introduz, salvo em casos pontuais, enviusamentos significativos na análise, conforme ficou demonstrado nos teste empíricos efectuados.

A segunda conclusão, tem a ver com a riqueza da informação recolhida e o potencial de ensinamentos que ela encerra e que, para já, só foi possível analisar em parte. Mesmo assim, os exercícios efectuados revelam um território em rápida mutação, onde as cidades conhecem dinâmicas muito diferenciadas, quer no contexto regional em que foram apresentadas, quer numa óptica das suas presumidas interdependências. Das profundas alterações no povoamento a que se assistiu durante o último século, sobressai claramente o processo de reforço progressivo das áreas metropolitanas de Lisboa e Porto, cada uma com a sua constelação de cidades, numa teia de interacções e complementaridades que definem espaços estruturados com lógicas territoriais fortes e dinâmicas de concentração que parece difícil contrariar.

REFERÊNCIAS BIBLIOGRÁFICAS

BAIROCH, P. (1988), Cities and Economic Development: From the Dawn of History to the Present, University of Chicago Press

BAUMONT, C., BEGUIN, H. e HURIOT, J.-M. (1996) "Définir la ville", Comunicação apresentada no XXXII° Colóquio da ASRDLF, Berlim 2 a 4 de Setembro de 1996

CAMAGNI, R. (1992), Economia Urbana: Principi e Modelli Teorici, La Nuova Italia Scientifica, Roma,

COMISSÃO EUROPEIA (1997), A questão urbana: orientações para um debate europeu documento disponível na internet no seguinte endereço: www.inforegio.cec.eu

INSTITUTO NACIONAL DE ESTATÍSTICA (1998), Tipologia de áreas urbanas, 10p.

MUMFORD, L. (1961), The City in History: Its Origins, Its Transformations, and Its Prospect, Harcourt, Brace, & World, Nova Iorque

O'SULLIVAN, A (1995), Urban Economics, Irwin, Boston, (3ª ed.) (data1ª edição: 1992)

POLESE, M., (1998), Economia Regional e Urbana, APDR, Coimbra, (versão original publicada pela Economica, Paris, 1994)

PUMAIN, D. e outros (1992), Le concept statistique de la ville en Europe, OPOCE

SALGUEIRO, T. B. (1992), A Cidade em Portugal, Edições Afrontamento, Porto.

Revista de Estatística
67ª Página
2º Quadrimestre de 1999

Factorial Correspondence Analysis: an application to a three-dimensional contingency table

Autor:
Regina Soares

VOLUME 2
2º QUADRIMESTRE DE 1999

REVISTA DE ESTATÍSTICA
69ª PÁGINA
2º QUADRIMESTRE DE 1999

FACTORIAL CORRESPONDENCE ANALYSIS: AN APPLICATION TO A THREE-DIMENSIONAL CONTINGENCY TABLE

ANÁLISE FACTORIAL DE CORRESPONDÊNCIAS: APLICAÇÃO A UM QUADRO DE CONTINGÊNCIAS TRIDIMENSIONAL

Autor: Regina Maria Agostinho Soares
- Gabinete de Estudos e Conjuntura do Instituto Nacional Estatística

ABSTRACT:

- The purpose of this paper is to illustrate the application of Correspondence Analysis to a three-way table. This methodology is considered very effective in the analysis of a two-way table - the contingency table - and so the question of using these techniques to a three-way table was a natural sequence. Namely, we are interested in the study of a collection of two-way tables defined by the same couple of variables observed in different points of time.

Key-Words:

- *Correspondence Analysis; Contingency Tables; Three-dimensional contingency tables; Three way contingency tables.*

RESUMO:

- O objectivo deste documento é ilustrar a aplicação da Análise Factorial de Correspondências a quadros tridimensionais. Uma vez que esta metodologia é considerada muito eficaz para a análise de quadros bidimensionais - tabelas de contingência - a sequência natural é ver a possibilidade da aplicação desta técnica a quadros tridimensionais. Nomeadamente, interessa estudar um conjunto de tabelas de contingência observadas em diferentes pontos do tempo.

PALAVRAS-CHAVE:

- *Análise de correspondências; tabelas de contingência; tabelas de contingência tridimensionais*

VOLUME 2

2º QUADRIMESTRE DE 1999

1. INTRODUCTION

The cross-tabulation of categorical data is perhaps the most commonly encountered and simple form of analysis in statistical research, but interpreting a contingency table with a great number of rows and/or columns can become very complex.

Correspondence Analysis (CA) is a technique with which is possible to find a multidimensional representation of the dependencies between rows and columns in a low dimensional space. It allows the construction of an orthogonal system of axes (called factors) where observations (rows of the table) and variables (columns of the table) can be simultaneously displayed, making easy to discover the salient information included in a given contingency table. In this system proximity between observations or between variables is interpreted as strong similarity. Proximity between observations and variables is interpreted as strong relationship. The usual output from a correspondence analysis includes the "best" two-dimensional representation of the data, the co-ordinates of the plotted points and a measure of the amount of information retained in each dimension (called the *inertia*).

2. CORRESPONDENCE ANALYSIS

Let $\mathbf{X}$ be an $I \times J$ contingency table. The rows and columns of this table correspond to different categories of two different characteristics and the entries x_{ij} give the frequency with which row category i occurs together with column category j.

It is convenient to base the graphical representation of association in contingency tables on a suitable centred and scaled matrix, in such a way that the image of the points in the plane of projection defined by the graphical axes are similar to the real distances between the points. To find the set of axes which preserve the maximum of *inertia* of the points, we start by construct a table of proportions $\mathbf{P} = \{p_{ij}\}$ by dividing each element of $\mathbf{X}$ by n (number of total frequencies in $\mathbf{X}$).

The row total is

$$r_i = \sum_{j=1}^{J} p_{ij} = \sum_{j=1}^{J} \frac{x_{ij}}{n}, i = 1,2,...,I$$

The column total is

$$c_j = \sum_{i=1}^{I} p_{ij} = \sum_{i=1}^{I} \frac{x_{ij}}{n}, j = 1,2,...,J$$

Next, subtracting the product of the row total and the column total for each entry centres **P**

$$\tilde{p}_{ij} = p_{ij} - r_i c_j, i = 1,2,...,I; j = 1,2,...,J$$

Let Let $\mathbf{D}_r = diag(r_1, r_2, ..., r_I)$ and $\mathbf{D}_c = diag(c_1, c_2, ..., c_J)$ and construct the scaled matrix

$$\mathbf{P}^* = \mathbf{D}_r^{-\frac{1}{2}} \tilde{\mathbf{P}} \mathbf{D}_c^{-\frac{1}{2}}$$

The *(i,j)th* entry of $\mathbf{P}^*$ is

$$p^*_{ij} = \frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}}, i = 1,2,...,I; j = 1,2,...,J$$

Define the singular decomposition of $\mathbf{P}^*$:

$$\mathbf{P}^* = \mathbf{U}\Lambda\mathbf{V}'$$

where $\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I}$, $rank(\mathbf{P}^*) = rank(\tilde{\mathbf{P}}) \leq J - 1$ and the matrix $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_{J-1})$ contains, in its diagonal, the singular values ordered from largest to smallest.

Set $\tilde{\mathbf{U}} = \mathbf{D}_r^{\frac{1}{2}} \mathbf{U}$ and $\tilde{\mathbf{V}} = \mathbf{D}_c^{\frac{1}{2}} \mathbf{V}$ .The singular value decomposition of $\tilde{\mathbf{P}}$ is

$$\tilde{\mathbf{P}} = \mathbf{P} - \mathbf{r}\mathbf{c}' = \tilde{\mathbf{U}}\Lambda\mathbf{V}' = \sum_{j=1}^{J-1} \lambda_j \tilde{\mathbf{u}}_j \tilde{\mathbf{v}}_j$$

where $\tilde{\mathbf{u}}_j$ and $\tilde{\mathbf{v}}_j$ are, respectively, the *jth* column vector of $\tilde{\mathbf{U}}$ and $\tilde{\mathbf{V}}$. In this representation the singular vectors are normalised to have unit lengths in the metrics $\mathbf{D}_r^{-1}$ and $\mathbf{D}_c^{-1}$:

$$\tilde{\mathbf{U}}'\mathbf{D}_r^{-1}\tilde{\mathbf{U}} = \tilde{\mathbf{V}}'\mathbf{D}_c^{-1}\tilde{\mathbf{V}} = \mathbf{I}$$

The columns of $\tilde{\mathbf{U}}$ define the co-ordinate axes for the points representing the column profiles of **P**. The columns of $\tilde{\mathbf{V}}$ define the co-ordinate axes for the points representing the row profiles of **P**.

Now it is possible to calculate the co-ordinates of the row profiles

$$\mathbf{Y} = \mathbf{D}_r^{-1}\tilde{\mathbf{U}}\mathbf{\Lambda}$$

and the co-ordinates of the column profiles

$$\mathbf{Z} = \mathbf{D}_c^{-1}\tilde{\mathbf{U}}\mathbf{\Lambda}$$

The first two columns of **Y** contain the pairs of co-ordinates of row points in the best two dimensional representation of the data and the first two columns of **Z** contains the pairs of co-ordinates of column points of the best two dimensional representation of the data. The points of these two sets of co-ordinates can be superimposed in the same graphic, since the lines and columns represent objects of the same kind.

3. CHI-SQUARE ANALYSIS OF ASSOCIATION

The χ^2 statistic for measuring the degree of association between the row and column variables in a contingency table with I rows and J columns is

$$\chi^2 = \sum_{i,j} \frac{(x_{ij} - E_{ij})^2}{E_{ij}}$$

where x_{ij} is the observed frequency for the *(i,j)th* cell and $E_{ij} = nr_ic_j$ is the expected frequency for that cell, if the row variable is unrelated to the column variable. The χ^2 tests the goodness of fit for the independence model and is asymptotically a χ^2 with *(r-1) (c-1)* degrees of freedom if the independence hypothesis holds.

We can write this statistic

$$\chi^2 = n\sum_{i,j} \frac{(p_{ij} - r_ic_j)^2}{r_ic_j} = \sum_{i,j} p_{ij}^{*2}$$

or

$$\frac{\chi^2}{n} = trace\left(\mathbf{D}_r^{-1}\left(\mathbf{P}-\mathbf{r}\mathbf{c}'\right)\mathbf{D}_c^{-1}\left(\mathbf{P}-\mathbf{r}\mathbf{c}'\right)'\right) = trace\left(\mathbf{P}^*\mathbf{P}^{*'}\right) = \sum_{i,j} p_{ij}^{*2}$$

The χ^2-distance concept can be used in interpreting the configuration of the points. It tells us that when two rows are close together, their profiles must be similar and they should be related in a similar manner to the columns. On the other hand, if two rows are far apart, they are related in a different way to the columns.

The profile of the column marginal c_j is the weighted average of the row points - the mean row profile - and it is located in the origin of the space. In a similar way the profile of the row marginal r_i is the weighted average of the column points - the mean column profile - and is also located in the origin of the space. When a point is near the centre of the space, its profile is similar to the column marginal c_j. When two row points are in opposite directions from the centre they deviate in opposite ways from the profile of the columns marginal.

We can approximate the row (column) profiles:

$$\tilde{\mathbf{P}} = \mathbf{P} - \mathbf{r}\mathbf{c}' = \sum_{k=1}^{J-1} \lambda_j \tilde{\mathbf{u}}_j \tilde{\mathbf{v}}_j'$$

and, for some $K \leq J-1$

$$\mathbf{P} \cong \mathbf{r}\mathbf{c}' + \sum_{k=1}^{K} \lambda_k \tilde{\mathbf{u}}_k \tilde{\mathbf{v}}_k' = \hat{\mathbf{P}}$$

The rows of $\hat{\mathbf{P}}$ define a subspace that is the closest to the rows of $\mathbf{P}$, in terms of a weighted sum of squares distance. The same argument can be used for the columns of $\hat{\mathbf{P}}$ and $\mathbf{P}$.

It is also possible to write $\mathbf{P}$ in terms of the matrices of co-ordinates $\mathbf{Y}$ and $\mathbf{Z}$

$$\mathbf{P} = \mathbf{r}\mathbf{c}' + \mathbf{D}_r \mathbf{Y} \Lambda^{-1} \mathbf{Z}' \mathbf{D}_c$$

or

$$p_{ij} = r_i c_j + r_i c_j \sum_{k=1}^{K} \frac{y_{ik} z_{jk}}{\lambda_k}$$

In a generalized least squares sense $\hat{\mathbf{P}} = \mathbf{r}\mathbf{c}'$ is the best approximation to $\mathbf{P}$ under the hypothesis of the independence model and $\frac{\chi^2}{n}$ is the discrepancy between $\mathbf{P}$ and $\hat{\mathbf{P}}$. Correspondence analysis attempts to picture this discrepancy.

4. INERTIA

The total *inertia* is the weighted sum of the squared distance of the row (column) profiles to the centroid - the row (column) mean profile. It is a measure of the overall variation, or differences in the points representing the row (column) profiles.

It can be shown that the *inertia* associated with the row points is the same as the *inertia* associated with the column points, that is:

$$inertia = \sum_{i=1}^{I} r_i \left(\tilde{r}_i - \mathbf{c}\right)' \mathbf{D}_c^{-1} \left(\tilde{r}_i - \mathbf{c}\right) = \sum_{j=1}^{J} c_j \left(\tilde{c}_j - \mathbf{r}\right)' \mathbf{D}_c^{-1} \left(\tilde{c}_j - \mathbf{r}\right)$$

where $\tilde{r}_i$ is the row profile

$$\tilde{r}_{ij} = \frac{x_{ij}/n}{\left(\sum_j x_{ij}\right)/n} = \frac{p_{ij}}{r_i}$$

and $\tilde{c}_j$ is the column profile

$$\tilde{c}_{ij} = \frac{p_{ij}}{c_j}$$

and

$$inertia = \sum_i r_i \sum_j \frac{\left(\frac{p_{ij}}{r_i} - c_j\right)^2}{c_j} = \sum_j c_j \sum_i \frac{\left(\frac{p_{ij}}{c_j} - r_i\right)^2}{r_i}$$

$$= \sum_i \sum_j \frac{\left(p_{ij} - r_i c_j\right)^2}{r_i c_j} = \frac{\chi^2}{n}$$

or

$$inertia = \frac{\chi^2}{n} = trace\left(\mathbf{D}_r^{-1}(\mathbf{P} - \mathbf{r}\mathbf{c}')\mathbf{D}_c^{-1}(\mathbf{P} - \mathbf{r}\mathbf{c}')' \right) = \sum_{k=1}^{J-1} \lambda_k^2$$

One can show that $\mathbf{Y}'\mathbf{D}_r\mathbf{Y} = \mathbf{Z}'\mathbf{D}_c\mathbf{Z} = \mathbf{\Lambda}^2$, so the weighted sum of squares of the co-ordinates of the row points along the *Kth* axis and the weighted sum of squares of the co-ordinates of the column points along the *Kth* axis are each equal to λ_k^2. The proportion of *inertia* explained by the first *K* factors *(K≤J-1)* is given by

$$\frac{\sum_{k=1}^{K} \lambda_k^2}{\sum_{k=1}^{J-1} \lambda_k^2}$$

In order to decide how many factors to retain in the analysis we use the proportion of explained *inertia*, after checking the general meaning of all factors. Usually the researchers keep only the first two or three factors if they can summarise a great percentage of total *inertia*. However, if an axis has a relatively small eigenvalue but it has a meaningful interpretation, we should not discard it, because it can often help to make a fine point about the data.

Then we would like to know which points have contributed the most to the building of each axis. These values are often known as **CTR** or **absolute contributions**. The *inertia* of each factor can be decomposed into the components due to each row profile and one can see which are the rows with higher contribution to the *inertia* of the factor. The rows with greater contribution determined the orientation of the corresponding factor. These values are important to interpret the factors.

Graphically, the further a point is from the origin the bigger is its contribution to the *inertia*, the smaller its marginal weight.

We can also know the contribution of each cell to the total *inertia* - the cells with values far from the theoretical ones will give the biggest contributions. The contribution of the cell*(i,j)* to total *inertia* can be computed with the following formula:

$$\frac{(p_{ij} - r_i c_j)^2}{r_i c_j \sum \lambda_k}$$

where $\sum \lambda_k$ is the sum of the eigenvalues of the factor.

The inertia of each point *i* given by the factors is known as **relative contributions**, **COR** or **QLY** (local quality). We can see how closely each profile vector is to a factor - the squared cosine of the angle formed by a row radius vector and the factor is similar to a squared coefficient of correlation and thus it measures how well the display approximates the profile true positions. To interpret the angle between a point and a factor we can say the closest a point is to the factor, the best this

factor explains its distance from the mean profile. For example, a **COR** 90% means that the factor explains 90% of the *inertia* of *th* point. Conversely, if the angle is close to 90° the factor doesn't explain that point.

The graphical representation of the plane created by the first two or three factors, as the synthetically representation of a big figures table, is of the great interest to data analysis. There are several methods to graphically represent the analysis (ex: two graphics - one for column points and other for the row points; both row and column points in the same graph, but with different scales; row and column points in the same graph with the same scale) and we need to understand their limitations before interpreting them.

Usually the two-dimensional representation figures both rows and columns points in the space created by two axis, at the same scale. The scale intersection of the two axes is known as *mean centroid* and corresponds to the mean profile.

We must be careful with the interpretation of these graphs. For instance, we cannot deduce from the proximity of a row and a column point that they are strongly correlated in the data. This kind of analysis is only possible with points of the same space - within row points or within column points. For the points belonging to the same cloud, if two points are close we can say that they have similar profile. One way to avoid this problem is to base the analysis on the values of **CTR**, **COR** and **QLY**. Another way is interpreting the angle between a row point and a column point, taking the origin as summit:

- If the angle is *acute* the two characteristics for which the points stand are correlated.
- If the angle is *obtuse* the points are negatively correlated.
- If there is a *right* angle the points do not interact.

Correspondence analysis allows also the representation of subsets of variables or observations as "illustrative elements", so that they can be situated with regard to all other active variables or observations. One could wish to add some points in a graph which we think may help the interpretation, but we do not want them to enter in the construction of the factors. This points will be projected in the new axes set. They do not contribute to the *inertia*, but some software gives us some indications as if they were active points. Examples of these "supplementary points" are:

- a point different from all the others but that can help interpret the others;
- an outlier whose co-ordinates can change the graphic's shape and hide the information of the other points;
- a category that is useful to be subdivided to compare the structures (men/women; different classes of revenue, different points of time...)

5. THE CORRESPONDENCE ANALYSIS AND ITS APPLICATION TO A THREE-WAY TABLE

Sometimes numerical statistical data sets involve variables that must be studied according to one specific variable. If the variable is time these data sets are called chronological data sets.

The primary interest of Factorial Correspondence Analysis is in the presentation of the structure of the observed data, but we may be interested in observing the evolution of that structure, that is, in the analysis of chronological contingency tables.

The three-dimensional contingency table arises from the cross classification of the categories associated with three different characteristics.

Let I,J,T be the three different characteristics, where T refers time.

The entries x_{ijt} give the frequency with which, row category i occurs together with column category j in time t. The table of proportions is obtained dividing x_{ijt} by the number of total frequencies n. The techniques proposed to analyse the three way tables consist, first, in obtaining the two dimensional tables for which the cell frequencies are denoted $p_{ij.}, p_{.j.}$, and $p_{..t}$. Since the problems never arise symmetrically in relation to the three variables, our concern is to compare the two way tables and see if there are common trends and structures.

There are three possible methodologies to study these three way tables:

1. Factorial Correspondence Analysis of the tables' sum, using the two-way different tables as "supplementary elements".
2. Factorial Correspondence Analysis of juxtaposed tables, completed with multiple codes.
3. The Within Analysis which allows the study of conditional relations.

The supplementary tables in the CA of their sum

We start with a table S whose columns are the sum of the (j,t) columns of the T $(t=1,2,...,T)$ tables. CA of this table gives us a new set of axes where we are going to display the profiles of the T tables. The factors of S are the centre of gravity of the column profiles of the T tables, and will display the common trends of the T tables, if they exist.

As we project the column profiles in the factors of their centre of gravity, we can study the deviation of the (j,t) columns of each table to the mean profile of these columns in the factor.

The differences between the T tables profiles may not be visible in these projections if the deviations between their profiles are small, when comparing with the deviations between different profiles, or if the profiles are orthogonal to the mean structures.

We can also join to this analysis t supplementary rows and t supplementary columns - the sum of the rows (columns) of each of the \$t\$ tables - to see the projections of the mean profiles of the T tables.

Comments:

- With this technique compare the row (column) profiles gives us an incomplete answer, since we are comparing them in the factors of a point mean cloud.
- If does not exist a strong enough common structure the table of the sum is a mix of different trends. A predominant table can also influence it.
- Differences between tables are not clear, once the significance of the common structures and their differences are not measured.
- The factors of each of the T tables are not in the analysis and cannot be compared.

Another way of doing this analysis is to take one of the t tables as reference, for example the first year, and perform correspondence analysis for this data set. We arrange the other data sets in rows, and then all these data sets are classified as supplementary elements of the reference data analysis. Thus, there are as many points for each observation as there are years, so that each observation point is labelled by its reference year, and we can describe the year-by-year evolution of each observation. If we arrange the data sets in columns we will have as many points for each variable as there are years and with each variable point labelled by its reference year, we can describe the year-by-year evolution of the variables. The results of the two methods can be presented on the same graph.

Factorial Correspondence Analysis of Juxtaposed Tables

In this analysis we can have T horizontally juxtaposed tables or we can have T vertically juxtaposed tables which will give us the reversal variable analysis and the sum of the tables is now the supplementary table.

The factors are computed from the total active columns. The Huygens principle shows that the *inertia* of a cloud of points, composed of several sub-clouds can be decomposed by

$$\text{Total } \textit{inertia} = \text{between } \textit{inertia} + \text{within } \textit{inertia}$$

where the between *inertia* is the *inertia* of the centre of gravity of the sub-clouds and the within *inertia* is the *inertia* of each cloud with respect to its centre of gravity.

In the CA of the juxtaposed tables we can compare the set of factors because we have a common reference. However, we must have in mind that with the row juxtaposed tables we can only compare the factors defined to this set, and that it is not possible compare the factors defined by the set of columns.

Comments:

- On the juxtaposed tables we do not have symmetry of both analyses (columns and rows). If we reverse the juxtaposed tables we find a different problem to analyse.
- In this analysis we can find mixed factors that express simultaneously the two kinds of dispersions: the within and the between inertia. This will rend difficult the interpretation of the factors.
- If there exists a common strong structure the CA results of the juxtaposed tables are very similar to the results of the CA of the sum of the tables, and we can analyse that structure. On the other hand, if the differences are more important, the CA represents well these differences and badly the common structure.
- In this analysis the differences between the *(j,t)* columns of the *T* tables are relevant in the computation of the factors and so they are more visible that in the CA of the sum of the tables. However, if the differences between the columns of the same table are greater than the differences between the *(j,t)* columns of the *T* tables these will not be visible. We cannot compare the rows if we have juxtaposed the tables horizontally, or the columns if we have juxtaposed the tables vertically.
- The CTR's gives a good measure of the differences between the *(j,t)* columns of the *T* tables. It is not possible to have a similar measure to the rows unless we analyse the reverse table.
- The factors of the columns of the juxtaposed table are a good reference to the set of the column factors of each table and so it is possible compare these factors.

The within analysis

Neither the CA of the sum of the tables nor the CA of the juxtaposed tables allow the systematic analysis of the differences between the *(j,t)* columns of the *T* tables. To the first analysis only the between inertia is taken into account to the computation of the factors where the columns profiles will be displayed and to the second analysis both between and within inertia are used to compute the factors.

To analyse the differences between the profiles one needs that the factors are computed taking into account only the between inertia.

This can be achieved if we study a cloud of points derived from the CA of the juxtaposed table, by centring to the origin all the sub-clouds of points of the *(j,t)* columns of the *T* tables, so that all the between inertia of the initial cloud is suppressed and only rests the within inertia.

In this analysis the CA is generalised to a model which is different from the independence model, but the metrics and the weights are the same. We are going to analyse two clouds of rows and columns with a dual relation, where the co-ordinates of the points of this clouds are the differences between the row (column) profiles and the model. We can use a classical software of CA if we introduce the new model instead of the independence model. The table to be analysed is the following :

$$p_{ij} - m_{ij} + p_{i.} + p_{.j}$$

where m_{ij} is the entry of the new model, $p_{i.} = m_{i.}$ and $p_{j.} = m_{j.}$.

The model table has the same dimension of the juxtaposed table and its construction is subject to the following constraints:

The profiles of the *(j,t)* columns of the *T* tables are mixed together with the mean profile $\frac{p_{ij.}}{p_{.j.}}$. The margins $m_{i..}$ and $m_{.jt}$ must equal the margins of $\mathbf{P} = \{p_{ijt}\}$. We can obtain the *(j,t)* column by multiplying its profile $\frac{p_{ij.}}{p_{.j.}}$ by $p_{.jt}$..

$$m_{ijt} = \frac{p_{ij.}}{p_{.j.}} * p_{.jt}$$

It is easy to show that the margins are equal, $m_{ij.} = p_{ij.}$ and $m_{.jt} = p_{.jt}$, as the homologous columns are:

$$\frac{m_{ijt}}{m_{.jt}} = \frac{m_{ij.}}{m_{.j.}} = \frac{p_{ij.}}{p_{.j.}}$$

In this table all the *(j,t)* columns are proportional once they are all proportional to the mean profile $\frac{p_{ij.}}{p_{.j.}}$. We can group this columns because the distance between rows will be the same and we will have the table sum with entry $\frac{m_{ij.}}{p_{.j.}}$, where the *j* column profile is in the centre of gravity of the *(j,t)* columns. So, the row distances induced with the table m_{ijt} coincide with the part inter-*J* induced by the juxtaposed table.

This model expresses the independence between *I* and *T* conditional to *J*:

$$\frac{m_{ijt}}{m_{.j.}} = \frac{m_{ij.}}{m_{.j.}} = \frac{m_{.jt}}{m_{.j.}}$$

The entry of the table to analyze by the within analysis is:

$$r_{ijt} = p_{ijt} - \frac{p_{ij.}p_{.jt}}{p_{.j.}} + p_{i..}p_{.jt}$$

The column *(j,t)* profile is:

$$\frac{r_{ijt}}{r_{.jt}} = \frac{p_{ijt}}{p_{.jt}} - \frac{p_{ij.}}{p_{.j.}} + p_{i..}$$

or

$$\frac{r_{ijt}}{r_{.jt}} - p_{i..} = \frac{p_{ijt}}{p_{.jt}} - \frac{p_{ij.}}{p_{.j.}}$$

if we take the centre of gravity of the cloud as origin.

The row profiles, taking the centre of gravity of the cloud as origin are:

$$\frac{r_{ijt}}{r_{i..}} - p_{.jt} = \frac{p_{ijt}}{p_{i..}} - \frac{p_{ij.}p_{.jt}}{p_{i..}p_{.j.}}$$

Comments:

The within analysis allows us to study:

- The relation between two variables conditional to a third one.
- The differences between the *(j,t)* columns of the *T* tables.
- The differences of the evolution of the row profiles.

Example of application

The data used in this example are from the employment survey of INE.

Example: *For this example we used the* ***CA of the sum of the tables****. The three way contingency table to analyse respects the men employed. The categories observed are degrees (variables J), ranged from 1 (no degree) to 7 (pos-graduate) and occupations (observations I), for the years 1983 and 1994.*

We begin our analyse by the *inertia* explained. The first factor explains 61.3% of the total *inertia*. In other words, this factor summarises 61.3\% of the distance to independence (*inertia*) of the table. The second factor explains 28.96% of the total *inertia*. The plan created by factor 1 and 2 summarises 90.26% of the total *inertia* and we can consider this plan gives us a close information of the data.

For the plan the column points which have greater contribution (CTR) to its *inertia* are degrees 6,1,3 and 4. The column points which are better explained by the factors (COR) are degrees 2,6,7, 1,3,4. The only column point that is not very well displayed in the plane is Degree 5 only 33% of its *inertia* QLY is explained.

The row points with greater CTR are occupations Scientific (CI), Administrative (AD), Trade (TR), Agriculture (AG) and the row points with greater COR are occupations Scientific (CI), Production (PD), Administrative (AD), Trade (TR), Agriculture (AG). All the row points have QLY greater than 0.5 and so, we can consider that all of them are well displayed in the plane.

EMPLOYMENT BY DEGREE AND OCCUPATION - MEN

Degrees	Co-ordinates		CTR		COR		QLY
	Factor 1	Factor 2	Factor 1	Factor 2	Factor 1	Factor 2	
1	0.04	-0.54	0.10	60.40	0.00	0.93	0.93
2	0.12	0.01	4.70	0.10	0.60	0.00	0.60
3	0.14	0.29	2.10	19.10	0.18	0.74	0.92
4	0.01	0.56	0.00	18.30	0.00	0.66	0.66
5	-0.42	0.29	1.50	1.40	0.23	0.10	0.33
6	-1.62	-0.01	78.10	0.00	1.00	0.00	1.00
7	-2.70	0.43	13.40	0.70	0.93	0.02	0.95
			100	100			

1 -NO DEGREE 2- 4 YEARS 3- 9 YEARS 4- 11 YEARS
5- 14 YEARS 6- GRADUATE 7- POS-GRAD

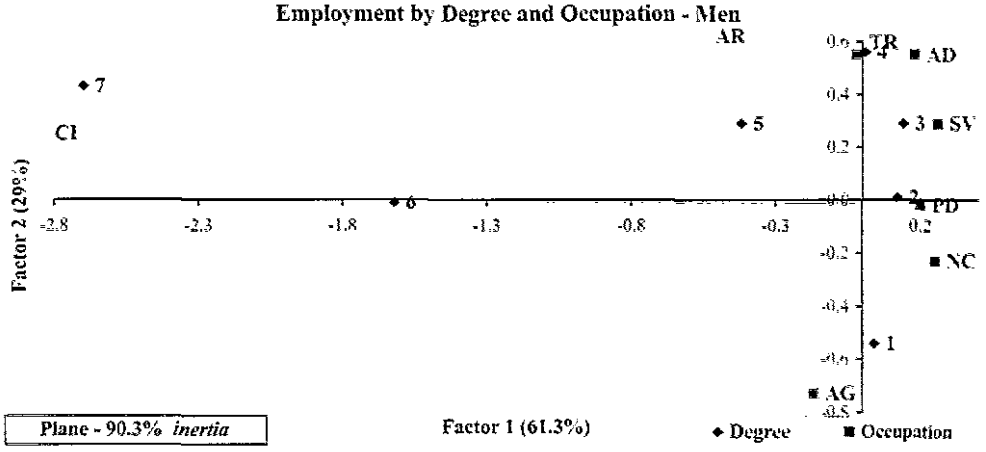
EMPLOYMENT BY DEGREE AND OCCUPATION - MEN

Profession	Co-ordinates		CTR		COR		QLY
	Factor 1	Factor 2	Factor 1	Factor 2	Factor 1	Factor 2	
CI	-2.84	0.25	88.80	1.40	0.99	0.01	1.00
AD	0.18	0.55	0.90	17.60	0.09	0.83	0.92
TR	-0.02	0.55	0.00	17.90	0.00	0.57	0.57
SV	0.26	0.29	1.60	4.20	0.38	0.48	0.86
AG	-0.17	-0.73	1.40	54.50	0.05	0.87	0.92
PD	0.20	-0.02	5.40	0.10	0.57	0.01	0.58
ÑC	0.25	-0.23	1.10	1.90	0.45	0.37	0.82
AR	-0.55	0.62	0.90	2.4	0.36	0.45	0.81
			100	100			

CI- CIENTIFIC AD- ADMINISTRATIVE TR- TRADE
SV- SERVICES AG- AGRCULTURE PD- PRODUCTION
NC- NO CLASSIFIED AR- ARMY

As we can see on the graph, we have the degrees ordered from the higher to the lower (left to right) along the first factor and the occupations ordered from the higher qualification to the lower classification. The second factor (top to bottom) opposes people with medium degrees of education to those who have the lowest and the

highest degree. We can see that people with a medium qualification (Administrative and Trade) opposes those with lower qualification (Agriculture).

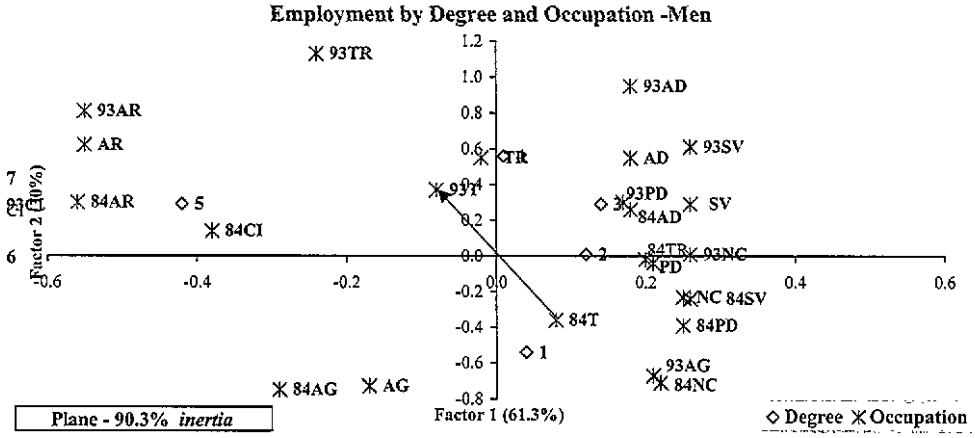


Our purpose now is to see the evolution of the relation between the occupations and the degrees. We displayed as supplementary or illustrative elements two tables: one for the year of 1984 and the other for the year of 1993.

EMPLOYMENT BY DEGREE AND OCCUPATION - MEN

Profession	Co-ordinates		CTR		COR		QLY
	Factor 1	Factor 2	Factor 1	Factor 2	Factor 1	Factor 2	
CI	-2.84	0.25	88.80	1.40	0.99	0.01	1.00
AD	0.18	0.55	0.90	17.60	0.09	0.83	0.92
TR	-0.02	0.55	0.00	17.90	0.00	0.57	0.57
SV	0.26	0.29	1.60	4.20	0.38	0.48	0.86
AG	-0.17	-0.73	1.40	54.50	0.05	0.87	0.92
PD	0.20	-0.02	5.40	0.10	0.57	0.01	0.58
ÑC	0.25	-0.23	1.10	1.90	0.45	0.37	0.82
AR	-0.55	0.62	0.90	2.4	0.36	0.45	0.81
			100	100			
SUPPLEMENTARY ELEMENTS							
84T	0.08	-0.36	0	0	0.03	0.70	0.73
84CI	-0.38	0.14	0	0	0.42	0.05	0.47
84AD	0.18	0.26	0	0	0.18	0.39	0.57
84TR	0.21	-0.04	0	0	0.19	0.01	0.20
84SV	0.26	-0.24	0	0	0.18	0.15	0.33
84AG	-0.29	-0.75	0	0	0.12	0.78	0.90
84PD	0.25	-0.39	0	0	0.16	0.41	0.57
84ÑC	0.22	-0.71	0	0	0.08	0.85	0.93
84AR	-0.56	0.30	0	0	0.53	0.15	0.68
93T	-0.08	0.37	0	0	0.03	0.70	0.73
93CI	-3.68	0.28	0	0	0.99	0.01	1.00
93AD	0.18	0.95	0	0	0.02	0.59	0.61
93TR	-0.24	1.13	0	0	0.02	0.45	0.47
93SV	0.26	0.61	0	0	0.11	0.60	0.71
93AG	0.21	-0.67	0	0	0.09	0.87	0.96
93PD	0.17	0.30	0	0	0.17	0.56	0.73
93ÑC	0.26	0.01	0	0	0.43	0.00	0.43
93AR	-0.55	0.81	0	0	0.21	0.46	0.67

If we consider the total employment of 1984 and 1993 as the mean of each year, we can say that a positive evolution has occurred. In fact, the total of 1984 is near the lowest degrees and not very classified occupations and 1993 is nearer the third/fourth degree (as it was expected once the school age increased to sixteen years) and of the occupations which need more qualifications. Notice that this occupations are related with services activities, which had increased in this lapse of time.



As general rule we can say that, in 1993, for the same occupation men need a higher degree than in 1984. The exception is for men whose occupation is Agriculture. For those who have Scientific occupations they are all far from the others and close Pos-graduate (7) degree.

6. REMARKS

- To apply factorial correspondence analysis to a three way table we can have different methods and we must choose according to the purposes we want to attain.
- The permutation of the role of the three variables increases the number of possible ways to analyse the data. We can consider to analyse the three binary margins, with the homologous columns as supplementary elements; three analysis of juxtaposed tables (*I* with *JT*), (*J* with *IT*), (T with *IJ*), completed with appropriate codes; six within analysis, depending on which of the variables we are conditioning on; the BURT table analysis attained by the juxtaposition of the tables of the paired off variables.
- The usual output from a correspondence analysis includes the "best" two-dimensional representation of the data. This gives us a big help in interpreting the results of correspondence analysis, but we must support the interpretation on the relative and absolute contributions values in order to confirm the conclusions.

VOLUME 2

2º QUADRIMESTRE DE 1999

Revista de Estatística
87ª Página
2º Quadrimestre de 1999

OVERVIEW OF PORTUGUESE 2001 CENSUSES: Concerning Development Strategy, Enumeration Structure, Basic Geography Units and Use of Administrative Registers As a Support Control

Autor:
Fernando Casimiro

VOLUME 2

2º QUADRIMESTRE DE 1999

REVISTA DE ESTATÍSTICA
89ª PÁGINA
2º QUADRIMESTRE DE 1999

OVERVIEW OF PORTUGUESE 2001 CENSUSES: CONCERNING DEVELOPMENT STRATEGY, ENUMERATION STRUCTURE, BASIC GEOGRAPHY UNITS AND USE OF ADMINISTRATIVE REGISTERS AS A SUPPORT CONTROL[15]

OS CENSOS 2001 EM PORTUGAL: PERSPECTIVA GLOBAL SOBRE A ESTRATÉGIA DE DESENVOLVIMENTO, A ESTRUTURA EXECUTIVA, AS UNIDADES GEOGRÁFICAS DE BASE E A UTILIZAÇÃO DE DADOS ADMINISTRATIVOS COMO ELEMENTOS DE CONTROLO

Autor: Fernando Casimiro
- Director do Gabinete dos Censos 2001[16] - Instituto Nacional de Estatística

ABSTRACT:

- Population and Housing Censuses for 2001 (2001 Censuses) are being prepared following a development strategy that includes the analysis and approval of Global Programme and Dissemination Programme by High Statistical Council (HSC). Then, besides the historical and methodological framing, have been defined main objectives to get as statistical units and respective variables and modalities to be observed.

  Enumeration structure should rely on local authorities (municipalities and parishes), with support and frame defined by National Statistical Institute (NSI), and on the identification of enumerator's area made with more precise cartography than that used in 1991. On other hand, electronic processing for questionnaires must be much more automatic than those occurred in 1991; must be used scanning, automatic and assisted coding for alphabetic fields; deterministic and probabilistic editing are being increased.

  Basic geographic units (statistical sections and subsections) are defined over maps produced by Geographic Basis for Information Reference (BGRI); main objective for this project consists of digitising and updating boundaries of statistical sections and subsections used on '91 Censuses. Given the coincidence of methodological structure used by '91 and 2001 cartography, it should be possible to see the decade evolution on territory occupancy by human settlements and delimitation of localities.

  Finally, administrative data on population and housing belonging to main administrative registers must be fully used to: create a warning system, to be used on the data collection phase which allows to estimate an acceptable size for data collected on each statistical unit; produce comparative analysis with 2001 Census data, when files are conceptually and methodologically closer.

[15] Paper prepared for European Workshop on the Preparation Census Fieldwork – Rome, 12-14 April 1999

[16] National Statistical Institute
Avenida António José De Almeida
P1000 Lisboa
Portugal
E-mail: fernando.casimiro@ine.pt

KEY-WORDS:

- *Population and Housing Censuses; '91 Censuses; 2001 Censuses; Global Programme; Dissemination Programme; scanning; administrative register.*

RESUMO:

- Os Recenseamentos da População e Habitação de 2001 (Censos 2001) estão a ser preparados seguindo uma estratégia de desenvolvimento que passou pela análise e aprovação do Programa Global e do Plano de Difusão no Conselho Superior de Estatística. Assim, para além do enquadramento histórico e metodológico destes recenseamentos, foram definidos os objectivos a atingir, bem como as unidades estatísticas, variáveis e modalidades a observar.

  A estrutura de recolha de dados vai passar, sobretudo, por uma forte participação das autarquias locais nas várias tarefas e por uma identificação ainda mais precisa das áreas de trabalho de cada recenseador, através da utilização de cartografia de maior qualidade do que a utilizada nos Censos 91. Por outro lado, todo o tratamento electrónico dos questionários está previsto ser significativamente mais automatizado do que em 1991, nomeadamente através da utilização da leitura óptica e da codificação automática e assistida dos campos alfabéticos e incremento das validações determinísticas e probabilísticas.

  As unidades geográficas de base vão ser definidas e delimitadas nas cartas a produzir no âmbito da Base Geográfica de Referenciação da Informação, cujo principal objectivo consiste na digitalização e actualização dos limites estatísticos e administrativos das secções e subsecções estatísticas utilizadas nos Censos 1991. Dada a coincidência entre a estrutura metodológica utilizada na preparação da cartografia de 1991 e 2001, vai ser possível analisar a evolução da ocupação do território pela habitação e delimitação dos lugares entre aquelas duas datas.

  Finalmente, os dados sobre população e habitação existentes nos principais ficheiros administrativos vão ser amplamente utilizados com dois objectivos: constituição de um sistema de alerta, a utilizar durante as operações do terreno, que permita determinar um intervalo aceitável para a previsão dos dados a recolher sobre cada unidade estatística; produção de uma ampla análise comparativa dos dados dos Censos 2001 com os dados provenientes das fontes administrativas e que sejam passíveis de aproximação conceptual e metodológica.

PALAVRAS-CHAVE:

- *Recenseamentos da população e habitação; Censos 91; Censos 2001; Programa Global; Programa de Difusão: BGRI; BGRE; leitura óptica; ficheiro administrativo.*

1. INTRODUCTION

Portugal has been conducting two simultaneous censuses, population and housing, since 1970. So there are two census sets what means that the next round of censuses (2001 Censuses) corresponds to 14th Population Census and 4th Housing Census.

In fact Portugal only has started with an independent housing census and fully accomplished with the international recommendations, in 1970 with the first Housing Census.

Since 1981, under the perspective to join to the European Community and having realised that spring is the best period of the year to collect census data, because of the weather conditions and population's movement at Christmas Holidays, we changed the census day from the end of the year finishing in zero to the Spring on the year finishing in one.

Population and Housing Censuses starts with preparing a Global Programme which includes the description of the main purposes to reach with these censuses, the identification and definition of each statistical unit to be observed and the respective variables (topics) and modalities and some of the main support instruments like cartography, data control and evaluation, legislation, publicity campaign and time schedule for each activity in the project.

Parallel to writing Global Programme is designed the Dissemination Programme which includes the strategy to disseminate census data and identifies each support and form to be used for. One of the main features of this Programme is the tabulation programme and the lowest level of disaggregation for each table in the version available and the version for publication.

The responsibility to prepare the first version of the Global and Dissemination Programmes belongs to the organisational unit in charge of Population and Housing Censuses, which is, for next censuses, 2001 Censuses Bureau.

We should say that these are the standardised steps to start the preparation of census and are similar to any other statistical project designed for data collection.

2. DEVELOPMENT STRATEGY - WHAT HAVE BEEN HAPPENING FOR 2001 POPULATION AND HOUSING CENSUSES?

2001 Census Bureau has been created as a specific unit of National Statistical Institute (NSI) in the beginning of 1998, and one of the first tasks was to prepare the referred Programmes and the proposal for specific legislation. These three

"instruments" constitutes the basis to start the analysis with main representatives of users, belonging to High Statistical Council (HSC).

In the HSC are represented every central government ministries, employers unions, trade unions, local authorities, universities, autonomous regions and consumers representatives.

2.1. LEGISLATION

According to statistical legislation every statistical projects and respective legislation wherever it exists, must be analysed by that Council which created a special section to follow 2001 Censuses until the end; this special section should be closed with the approval of final report of the statistical operation.

The first act for that special statistical section of the High Statistical Council was the analysis of specific legislation proposed by NSI in order to be after approved by Central Government and Parliament. Even Central Government needs a specific authorisation from Parliament in order to approve this legislation, because it includes regulations to regional and local authorities for which only Parliament has constitutional power to make legislation.

Specific legislation is needed for these censuses due to the assigned responsibility to regional and local authorities on data collection and delimitation of administrative and statistical boundaries. Legal responsibility to data collection belongs to local authorities with the support of NSI, and this is a way to get a strong involvement of these authorities on this task. On the other hand, some administrative boundaries are not so easy to find in the territory when deciding if a building belongs to a specific parish or not.

Now that proposal for legislation is under the responsibility of Central Government in consultation with Regional Governments, Local Authorities Associations and Personal Data Protection Committee. NSI expects to have the final approval of this legislation by the end of the first half on this year.

The main reason to have ready the legislation with this anticipation to 2001 comes from the need to update the administrative and statistical boundaries in accordance with local authorities; this boundaries are being digitised by NSI, as we can see in topics 2 and 3 of this workshop.

2.2. GLOBAL PROGRAMME

Global Programme for 2001 Censuses has been prepared according to five main principles:

- Maintain the census data sets;
- Comply with the international recommendations specially for core variables;
- Satisfy, as far as possible, new user's needs;
- Anticipate, as soon as possible, the date to dispose final data;

- Produce and disseminate consistent quality indicators, which should allow users to know and fully accept the coverage and content rates for these censuses.

2.2.1. CENSUS DATA SETS

Analysis of Global Programme has been recently finished in the section of HSC. Concerning the census data sets, mainly those coming from 1981 and 1991, was decided to keep almost all of them, in spite of changes made in some variables related mainly with building and family statistical units. In the building they were added some variables related with earthquake risk and the availability of urban solid waste collection.

For dwelling was decided to add two new variables: period of the renting contract, if dwelling is rented, and types of heating. Data on period of renting contract allows a better understanding of rented housing because rules on housing turnover are different depending on the date of this contract. In spite of type of heating is a core recommended variable, Portugal has not collected this kind of data due to the existing mild climate and a low frequency of housing heating. However nowadays there is a growing number of dwellings equipped with systems to keep an inside steady temperature which represents a new standard of housing conditions.

In the family, changes have been made mainly due to the new "statute" of children which allows a child to belong to parent's family independently of having or not been married before.

For resident persons, only two main changes have been decided: to draw back variables related to date of last and first marriages and number of children born alive, asked to women older than 12 years, and to change the minimum age limit from 12 to 15 years old to be economic active population. Main reasons for first change are related to the fact that Family and Fertility Survey was recently conducted and this kind of data can be produced in a more deeply way through that survey than it could be in a census. The change of minimum age limit to be economic active is due to legal and in fact improvement of a minimum nine years school leaving and the legal minimum age of 16 years old to start working. On the other way population census is not now the best adequate statistical solution to find out people working below that age limit.

2.2.2. COMPLY WITH THE INTERNATIONAL RECOMMENDATIONS

Every core topics of international recommendations should be observed in each statistical unit, with only one exception: legal marital status. In fact, since 1981 we use the prevalence of "de facto marital status" if there is difference to the legal one. Main reason for that is the objective to collect data which allows to analyse the " the facto" situation of couples without any kind of unacceptable questions on behalf of enumerator. On the other hand, during the questionnaires checking process enumerator is not faced with inconsistencies between marital status and family relationship among

members of same household or family, especially if there is more than one family in the household.

Over and above the core topics, are observed in these censuses 26 of 52 non-core topics internationally recommended, plus 11 topics which correspond to specific national needs. From these national 11 topics, 6 belong to the building and 5 to the dwelling.

2.2.3. SATISFY NEW USER'S NEEDS

A lot of new user's needs have been identified during the analysis of 2001 Censuses Programme in the HSC.

Users are usually expecting from population and housing censuses the opportunity to get ready a substantial amount of data that would answer all their statistical needs. This situation concerns an increasing number of local and regional users, which are not satisfied with data coming from surveys, because of their sampling error when dealing with data for regional levels below NUTS II.

Even sometimes their needs are not covered by statistical information system, what helps to press census to respond those data needs.

Some of the topics raised during this discussion are: accessibility and safety in the building; state of repair of water, electricity, gas and swage disposal networks in the building and in the dwelling; closed condominiums; urban quality of life concerning green areas, social settlements, parking areas, etc.; statistical data on professional training and same sex couples. Otherwise, it was suggested to the HSC that some of the existing surveys should reduce the time lag between each edition, and other surveys conducted on the basis of a not regular periodicity should carry on a regular one.

As referred before, new topics accepted are mainly related to earthquake risk on buildings and two variables for dwellings (type of heating and period of the renting contract).

2.2.4. ANTICIPATE, AS SOON AS POSSIBLE, THE DATE TO DISPOSAL FINAL DATA

One of the main constraints for census data is time lag between census day and the moment when final data are ready to be used by customers.

Last censuses only had final data available two years and a half past census day what has meant a lack of refreshes for some important data. Because the census data processing is a heavy task with a lot of temporary workers and a sophisticated system of editing, two ways are under development to shorten this elapsed time:

- Complete scanning for questionnaires;
- Deeply development of C91 system, which has been used for the first time in '91Census and allows the automatic and assisted coding of alphabetic answers.

Now, four basic questionnaires are designed to be used by scanner and results from the first test are really encouraging. We have tried to combine the easy filling for respondents with technical conditions needed for scanning and for the moment there were no negative reactions. If there is a successful combination, this means that we can spend about 3 months doing what in the recent past has been done by 15 months.

C91 system was developed to perform two main objectives:

- Assume that an alphabetic description written accordingly to respective classification should be coded automatically;
- When description needs to be coded by an operator, the system assumes the decision taken by the operator and should automatically repeat it when the same situation occurs.

From the experience of 91'population census we have ready dictionaries with coded descriptions, written by population and enumerators, for occupations, industry, countries, municipalities and field of study. We expect these dictionaries can significantly shorten time needed to code these alphabetic fields.

2.2.5. PRODUCE AND DISSEMINATE CONSISTENT QUALITY INDICATORS, WHICH SHOULD ALLOW USERS TO KNOW AND FULLY ACCEPT THE COVERAGE AND CONTENT RATES FOR THESE CENSUSES

Data from '91 Censuses have been very surprising even for opposite reasons. While resident population was about 4,5% below the estimated, increasing only 0.3% for the decade, dwellings had an increasing rate of 22%, including dwellings occupied as usual residence with 10% increasing for the decade too. Coverage rate measured by post enumeration survey was 99% for resident population and 99,4% for dwellings. However we feel that users have assumed the suspicion that undercoverage was higher that 1% for population.

So, given that important surprise, we are convinced that the 2001 Censuses will be subject to very careful observation on behalf of their main users.

In a way, despite the fact that ten years will have gone by in the meantime, the 2001 censuses data will end up being an important factor of evaluation of the '91 Census given that no significant and unexpected demographic "accident" has taken place nor is foreseen to take place in the country's demographic evolution process.

Besides the usual controls applied on fieldwork, we have scheduled a control and evaluation system with two main elements:

- "Warning system", based on a set of indicators coming from demographic surplus, electoral roll, geographic information system supported by digitised maps with black cells (buildings), scholar population, post addresses and electricity customers, for each possible and most disaggregated level; this system should allow us to have an expected figure for each covered statistical unit at the respective lowest level; if that figure is significantly not attained or is surpassed, it must be investigated a comprehensive reason for that;

- Post Enumeration Survey (PES), with coverage and content purposes, for a sample of each NUTS II region - Nomenclature of Territorial Units for Statistics; PES have to be the quality benchmark of 2001'Population an Housing Censuses; to assure the guaranty of independence on the meaning of *final results of PES, we foresee a partnership with an independent, external* and prestigious scientific organisation.

We hope these measures can play an effective role on assuring users that 2001 quality indicators should be out of question.

2.3. DISSEMINATION PROGRAMME

Dissemination Programme has been designed with the description of every census products associated to a strategy to disseminate their data.

Because of the long time period between each census, users expect census data faster and faster, some of them even in the census day. Something like political elections!

In fact that is not exactly possible but we must use every alternative to give users some data as soon as possible. For answering these expectations we had designed tabulation programme using several steps to dispose census data according to data processing programme and on an increasing reliability of respective data. It means that provisional data are much more reliable than preliminary ones, because checking process is more exhaustive than that used for preliminary data.

So, 2001 censuses data tabulation have been organised in three steps:

- **Preliminary data**, produced on the basis of administrative controls; when a enumerator ends the data collection in his statistical section he must count the questionnaires in order to be paid for that; these counts will be the input for the preliminary figures about population, households, dwellings and buildings;
- **Provisional data**, produced on the basis of files getting out of first level editing rules; provisional data are made of eight different tables covering every primary statistical unit (building, dwelling, household and person) and variables used for these tables are those ones which are not subject to strong editing rules:
  - Type of building, main use, number of dwellings, availability of solid wastes collection and period of construction, **for buildings**;
  - Type of living quarter, occupancy status, electricity, water supply system, toilet facilities, bathing facilities, type of sewage disposal system **for dwellings**;
  - Private households by size, **for households**;
  - Sex, age group, de facto marital status, literacy, school attendance, educational attainment, **for persons**.
- **Final data**, after ending every electronic processing steps.

Final data available for users will correspond to four different kinds of products:

- **Tables** belonging to the tabulation plan (105 basic tables), which should be ready up to the lowest level of disaggregation (statistical subsection which corresponds to 108.000 for all over the country); on the other hand, each basic table must have several versions to be published or only available to paper or electronic consultation;
- **A file ("ficheiro-síntese")** with the most important aggregated counts for each statistical subsection (108.000); this file is made of 91 different counts for each statistical subsection; those counts could be aggregated for any upper level (statistical section, locality, parish, municipality and every NUTS regions – Nomenclature of Territorial Units for Statistics);
- **A CD-Rom** including a long set of census data (from 1864 up to 2001) and a specific Geographical Information System for data from 1991 and 2001 censuses; these two last censuses have a comparable and digitised cartography, which allows the evaluative analysis on the territory occupation;
- **A central file**, which would allow users to make their own tabulation using Internet network.

Statistical secret is the limit to the users access to census data.

High Statistical Council approved Dissemination Programme in the end of March this year.

2.4. PREPARATION PHASES

Preparation phases include cartography, tests of questionnaires and pilot survey.

2.4.1. CARTOGRAPHY

From '91 Censuses we have ready the cartography support, which consists on the division of whole territory belonging to each parish (the lowest administrative unit) into statistical sections (enumeration areas); and each statistical section is divided into statistical subsections. Every supports of this cartography were on paper and we are changing them to a digitised basis.

More details about cartography are in chapter on topic "Basic geographic units".

2.4.2. QUESTIONNAIRES TESTS

Two questionnaire tests have been scheduled with the following main objectives:

- Evaluate the public reaction to the questionnaires content;
- Reach the best way to combine, in the questionnaires, technical conditions for scanning and easy questions to self-enumeration;
- Evaluate scanning performance;
- Test the remuneration system.

First test was conducted on last October and main conclusions are:

- Questionnaire design with two columns is more acceptable; this means that models used on previous censuses still remain more acceptable than those using only one column;
- They were not detected strong negative reactions to content of questionnaires;
- Self-enumeration rate depends more on person's availability to fill the questionnaires than to question's difficulty;
- Questionnaires designed under technical conditions for scanning are accepted by population;
- Scanning performance was very encouraging, albeit the number of questionnaires used have been low (a sample of 500 for each type of questionnaire).

In April this year, a new questionnaire test is foreseen with a larger sample (about 35000 people), corresponding to a group of complete parishes located in every 7 NUTS II regions.

2.4.3. PILOT SURVEY

Pilot survey should be done one year before census day what means March 2000. Its main objective is to make a complete rehearsal of the final census operations.

3. ENUMERATION STRUCTURE, CONTENTS, PRODUCTION PROCESS (ENUMERATION UNITS, NEW DATA CAPTURE TECHNIQUES, EDITING AND IMPUTATION)

As we have seen before, 2001 Censuses cover population and housing for which we use 5 statistical units: building, dwelling, household, family and person. While one of them (family) is derived, all other units should be observed on the basis of a specific questionnaire. Family unit is derived from the existing family relationships between members of respective household (child/parent/spouse), asked for everyone in the household questionnaire.

Fieldwork is mainly composed by two phases:

- Identification of each building and dwelling units to be observed and delivery of dwelling and individual questionnaires; in this phase we ask population to carefully read the questionnaires and respective instructions and to fill in them on census day; usually the delivery of questionnaires starts about two weeks before census day;
- Data collection, which starts on census day.

3.1. CONTENT AND FORM OF THE QUESTIONNAIRE FOR EACH STATISTICAL UNIT

3.1.1. BUILDING

Building questionnaire is an A4 format with the respective questions occupying only the front page. This questionnaire consists of 14 questions, over and above the geographical identification, concerning the following topics: type of building, main use, number of floors, ground floor building characteristics, relative position to other neighbour buildings and to the respective block, number of dwellings, availability of solid wastes collection, period of construction, structural materials used, materials used to cover external walls, type of roof and respective materials used for and repair needs on the structure, on the roof and on the external walls.

Only 2 of the 14 questions ask for numerically fulfilment; all others ask for mark fulfilment.

Only enumerator must fulfil this questionnaire.

3.1.2. DWELLING

Dwelling questionnaire is a double A4 format with the respective questions occupying only first two pages (front and back pages). Pages 3 and 4 consist of instructions to be used for self-fulfilment by a household's person and remain as protection cover for all questionnaires belonging to the respective dwelling. By the questionnaire preparation for scanning, pages 3 and 4 must be detached using an existing prick of.

The content of this questionnaire consists of 16 questions, over and above the geographical identification, concerning the following topics: type of living quarter, occupancy status, electricity, water supply system, toilet facilities, bathing facilities, type of sewage disposal system, kitchen, type of heating, number of rooms, financial loan costs for owner-occupied dwellings, rent form, period of renting contract, amount size of rent and type of ownership.

Only one of the above 16 questions is filled numerically; using marks fills every other one.

3.1.3. HOUSEHOLD

Household questionnaire is an A4 format with the respective questions occupying front and back pages. Must be filled in only by enumerator during the questionnaire collection phase.

This questionnaire consists in a list of persons belonging to a household, irrespective of being residents or only temporarily presents, and for each person we ask for: name, relationship to the representative (head) of household, identification of spouse (if living in same household), identification of father and/or mother for persons without spouse and/or without their own children living in same household.

The capacity of this questionnaire goes up to 36 persons a household.

Name is an alphabetical field and other fields are numerical.

3.1.4. PERSON

Individual questionnaire is an A4 format with the respective questions occupying front and back pages. Must be filled in by respondents or by enumerators during the questionnaire collection phase.

The content of this questionnaire consists of 30 questions covering the following topics, over and above the geographical identification: name, sex, place of usual residence, place where found at time of census, date of birth, de facto marital status, place of birth, country of citizenship, place of residence one year prior to census day, place of residence five years prior to census day, literacy, school attendance, educational attainment, educational qualifications, field of study for university qualifications, place of work, length of journey to work or school, mode of transport journey to work or school, main source of livelihood, current activity status, time usually worked, occupation, main tasks on main occupation, status in employment, industry, number of persons working in the enterprise, religion (as free answer).

For the extreme chance, there are 8 alphabetical fields, because:

- Some of these fields should not be answered due to the age limit;
- Other ones accept marks for the expected majority of respondents;
- Only personal name, field of study for those persons having a completed university level, occupation and industry must be filled in always on alphabetical.

Date of birth and geographical identification are numerical and all the rest of individual questionnaire fields must be filled in by marks.

3.2. QUESTIONNAIRES ORGANISATION AFTER FIELD WORK

At the end of fieldwork, questionnaires must be organised according to hierarchical order of each statistical unit:

- Every one of the individual questionnaires belonging to the same household must be ordered, on a complete sequence, from 1 to N into the household, with the household questionnaire over all of them;
- The complete collection of questionnaires belonging to each household should be put inside the respective dwelling questionnaire, which forms a cover to the respective household and individual questionnaires; if there is more than one household in a dwelling, the households must be ordered from 1 to N into the dwelling and kept inside the dwelling questionnaire according to this order;
- All the dwelling questionnaires (with the respective household and individual questionnaires inside) belonging to one building are ordered from 1 to N into the building with respective building questionnaire over them;
- For each building there is an auxiliary building cover where are kept inside all the respective questionnaires according to order we have seen before; this building cover has the function of enumerator's report book too;
- Because buildings are numbered from 1 to N into the statistical subsection, all the auxiliary building covers, with the respective questionnaires inside, are numbered according to the respective building number and kept inside an auxiliary statistical subsection cover, ordered according to building number;
- All the auxiliary statistical subsection covers, with the respective content, are kept inside a box with the identification of statistical section.

3.3. DATA CAPTURE

Data capture should be made by using scanning. Test results of April 1999 must constitute the benchmark to take decisions on the number and location of scanners and respective infrastructure, namely the number of workstations including ICR processing and editing stations.

Scanning should be made for 4 questionnaires, as seen before, plus 2 auxiliary sheets: one for statistical subsection identification; another one for transcription of data belonging to people present but not resident in collective dwellings.

Scanning will cover all the content of questionnaires, which means the use of recognition for marks, numerical and alphabetical fields.

Results from the last October's test were positively surprising on performance attainment; only one recognition engine has been used and software can be easily improved for better results:

- Marks do not constitute a major problem for this technology;

- For numerals recognition rates were sized by 86,7 and 97,8%, because numeric fields of questionnaires have filling in different quality levels;
- Alphabetic fields have recognition rates sized between 64,1 and 90,6%.

So we feel that using an improved technology (several engines, carpet system, updated scanners), which is on the way for the next test (April this year), recognition results could be much better than those obtained in the past test.

3.4. EDITING, CODING AND IMPUTATION

Editing, coding and imputation are tasks, which we are trying to automate to the maximum under the objective of reducing elapsed time between data collection and final results. To do that and according to the experience of last censuses (1981 and 1991), we are developing the imputation system based on cold and hot deck rules (deterministic and probabilistic), and the coding system called C91, which allows automatic and assisted coding for alphabetic fields.

For the first editing phase (data just coming from scanning) main rules concern recognition quality, questionnaires hierarchical order and blank questions that can be filled in.

The existing file was built with alphabetical descriptions made in the individual questionnaire of '91 census and coded in that time, which covers about 450.000 different respondent self-made and enumerators written descriptions for each one of occupation and industry variables. So, with this file we expect to save a lot of time doing this task. Recent developments of this system are concentrated on descriptions with 5 or more frequencies (about 30.000 for each of those files), which represent 80% of respective statistical universe.

So, the processing data sequence is: scanning, recognition, first editing phase (to solve recognition problems and other questions that need immediate access to questionnaires), coding (occupation, industry, field of study for university qualification, nationality, place of prior residence), second editing phase (mainly incoherences between questions), third editing phase, data specialising and final data tabulation.

Third editing phase must be done with cold deck and hot deck rules (deterministic and probabilistic) on a totally automated way. By this editing phase we foresee also to check counts of each primary variable and eventually correct any missing inconsistency, after automatic rules processing.

Data specialising concerns the production of a file, for each statistical unit, including primary answers and derived classifications; for example, socio-economic group classification or age for each person.

No imputation for missing statistical units is foreseen.

4. BASIC GEOGRAPHIC UNITS

Geographic units are those coming from the Geographic Basis for Information Reference (BGRI). This geographic basis was built up for '91 Censuses and consists on the division of the whole territory of each lowest administrative unit (parish) into statistical sections and subsections with a delimitation supported on the best available cartography. The whole statistical subsection belongs to a unique statistical section and a whole statistical section belongs to a unique parish.

So, the statistical organisation of the Portuguese territory is:

- Portugal
  - NUTS I (3 regions)
    - NUTS II (7 regions)
      - NUTS III (30 regions)
        - NUTS IV (308 municipalities)
          - NUTS V (≅4240 parishes)
            - Statistical sections (≅14000)
              - Statistical subsections (≅108000)

Statistical section is defined as a continuous area belonging to a unique parish and having about 300 dwellings; however each parish must have at least one statistical section irrespective of having less than 300 dwellings. Statistical subsection correspond more to a "homogenised" portion of territory than a number of dwellings, what means that one statistical subsection may have from 0 to 300 dwellings.

In fact a block is always a statistical subsection irrespective of having or not dwellings, because constitutes a "homogenised" portion of territory. The same happens to a small locality, which constitutes at least one statistical subsection; if a locality could be divided into several statistical subsections, so the respective locality results can be the sum of every statistical subsection belonging to that locality.

The complete identification sequence for each statistical subsection is:

- District – 2 digits
  - Municipality – 2 digits into the District
    - Parish – 2 digits into the Municipality
      - Statistical section – 3 digits into the Parish
        - Statistical subsection – 2 digits into the Statistical Section + 4 digits for Locality to which subsection belongs

Digital map identification for each statistical subsection is made of 15-digit sequence, which include administrative and statistical boundaries, locality name and delimitation. Enumerators only use first 11-digit sequence, because updating process of BGRI makes correspondence between locality and statistical subsection.

District was the prior administrative division of territory, used by statistics, and it corresponds to a group of municipalities. NUTS classification, at levels I, II and III, uses municipality as base unit to constitute those regions; so it is always possible to change from District/Municipality to NUTS/Municipality classification. The reason why District/Municipality classification is kept belongs to less number of digits used (4 against 8 on NUTS).

While in 1991 the geographic basis was supported by maps hand designed on the basis of several root maps and scales, for 2001 it was decided to update and digitise the BGRI over a national and standardised cartography according to a sequence of steps. This work is done using the ArcInfo software with the future purpose of building up a Geographical Information System and a basis for street tracks.

Fieldwork organisation, as to statistical census data from 2001 censuses, should be prepared according to areas and regions supported by this geographic infrastructure.

4.1. UPDATING BGRI

BGRI updating is a combined process among National Statistical Institute (NSI) and the most important national cartography producers, with co-operation of municipalities, which have the responsibility for local planning. However, final responsibility and property of BGRI belongs to NSI.

One of the followed principles by this updating process concerns the preoccupation that the final result of the 2001 BGRI delimitation be compatible with the '91 version, as much as possible, in a way to allow evolution analysis between the two moments of the decade. So it will be possible to see how territory occupation has changed by ten years.

Comparison process is made starting with the statistical subsection. If there is no change on the respective delimitation between 1991 and 2001, comparison can be made fully at this level; if 91 delimitation has been changed by the updating process, a minimal group of statistical subsections, with delimitation changes, must be added up to find out the minimal area, which can be compared into correspondent boundaries.

4.1.1 STEP ONE

First step corresponds to digitisation of delimitation of every statistical section and subsection belonging to '91 BGRI, over the national and standardised cartography (1/25000 and 1/10000). However final support file must be ready over cartography on 1/25000 scale.

Because administrative boundaries always belong to a section and subsection delimitation, it means that this digitisation also allows having a full scale digitised map with all kinds of boundaries.

4.1.2. STEP TWO

By this step we mean the editing of new territory cover with a proposal for an updated delimitation of statistical sections and subsections, which should be confirmed and updated again during the next phase. Inserting '91 BGRI delimitation over updated cartography it provokes immediately a value added to BGRI, because it is possible only by that to see new settlements on the territory.

4.1.3. STEP THREE

This step corresponds to a local validation and updating new delimitation made in the step two for statistical sections and subsections, in a paper support. Specialised people, who belong to National Statistical Institute, make this validation with the municipalities and parish's collaboration.

4.1.4. STEP FOUR

Fourth step deals on editing, over the digitised version of step two, updating corrections made by step three.

4.1.5. STEP FIVE

After corrections made on step four there are produced the final maps in a sized scale between 1/2500 and 1/5000, which may constitute the support to data collection by enumerators.

For each statistical section are produced two maps:

- **Parish Panoramic**, which permits to see delimitation of all statistical sections belonging to the parish; with this map, enumerator can easily see where is his section and the respective boundaries and codes for numeric identification;
- **Statistical Section Panoramic**, which permits to see delimitation of all statistical subsections belonging to respective section; with this map, enumerator can easily see where is each subsection, the respective boundaries and codes for numeric identification, which will be used in the 2001 Census questionnaires.

4.2. GEOGRAPHICAL FIELDWORK ORGANISATION

Fieldwork organisation is based on that geographical basis. This means that each statistical section is assigned to one enumerator, who will receive the two maps

referred above to support and identify each smaller area (statistical subsection) in his section.

For the parish level there is a co-ordinator, who will be the responsible by all the respective enumerators, even for controlling data collection quality; parishes with more than 6 statistical sections should have sub-coordinators, each one controlling 5-7 enumerators.

At the municipality level there will be a municipality delegate, chosen by the municipality, whom is responsible by all the municipality work; for larger municipalities it can be more than one municipality delegate.

At regional level (NUTS II), responsibility for 2001 Censuses belongs to NSI's Regional Directorates, which will indicate regional delegates.

Training process is designed according to this administrative and geographical breakdown but must end at the municipality level. So there is a starting training phase given by national co-ordination to regional co-ordinators; regional co-ordinators do training to municipality delegates, and these last ones must train parish co-ordinators and enumerators.

4.3. FINAL DATA BREAKDOWNS

Final data may use every geographic breakdowns listed before, and all the tables belonging to tabulation plan must be ready down to statistical subsection, which is the lowest level for data disaggregation. However there are some limitations for current use: majority of scheduled tables should be ready for immediate use only down to parish level; for those disaggregation levels each demand should be evaluated to see if there are or not questions concerning statistical secrecy.

4.4. CONFIDENTIALITY AND DATA PROTECTION

For census purposes we have assumed that figures below 3 units could not be under statistical secrecy down to parish level. Below that level (statistical section and subsection) was prepared, since 1981, a specific product called "Ficheiro-Síntese", which permits to have ready data for aggregated variables, for which have been assumed not being under statistical secrecy.

In the near future, probably something has to be changed on this matter, because concerns of population on private life are growing up more and more. Thus, we are studding it on reaching the best way to satisfy users preserving individual confidentiality.

5. USE OF ADMINISTRATIVE REGISTERS AND OTHER SOURCES AS SUPPORT TO FIELDWORK AND QUALITY CONTROL

Population and Housing Censuses are statistical operations, which should constitute a benchmark for statistical data on this subject.

In Portugal, existing administrative registers are used mainly for administrative purposes and often don't comply with statistical concepts and they have important updating problems.

One of the most used administrative registers for statistical purposes is civil registration of births and decease, which allow natural balance, and there is no problem concerning statistical concepts and data quality.

For migration movements, there is only an administrative register for immigration of non-nationals; so it means that immigration of nationals and illegal people is not under administrative control. On the other hand, emigration is almost all out of administrative control, even because the most important part of them goes to countries of European Union.

However, for population estimates, net migration is the other one component that is not so feasible as natural balance; so, concerning net migration, final estimates for each decade are built up with census data.

We may ask why migration trends are not enough measured by population surveys, as for Labour Force Survey or any other one. In fact, for a country like Portugal with a strong migration movement, to evaluate the immigration, including return migration is not so difficult; however evaluation of emigration is not so easy because there is no anyone in respective dwellings to answer survey questions. Then, future european migration statistics wouldn't be excused from using an interchange administrative model that permits to know, to the origin country, that a citizen is asking for a residence authorisation in another country.

Concerning main population registers (Electoral Register and Civil Identification Register), there is a strong feeling that a lot of people living outside Portugal keep their legal residence in Portugal due to several reasons.

On the other hand, if census data are not according to yearly population estimates users often suspect from census data, what represents a real challenge to population census to explain reasons why these data are feasible.

Housing data are not so difficult to control because buildings and dwellings are much more "stable" and easy to find than people are.

So, to avoid surprises and suspicions like those occurred in '91 Census, we are designing a control and evaluating programme that uses administrative data to check every step on the collection phase and to give to the users a global analysis about data odds due to different sources.

5.1. WARNING SYSTEM

As we have seen before (chapter 2.2.5), this system combines data from different sources to estimate an acceptable size for each statistical unit, at parish level. Those estimated sizes should be used to define if data collected by 2001 census fieldwork are acceptable or not; if not a special supervising process should be unchained to check census data.

Electoral Register for population estimates, postal domestic addresses and private electricity customers, both for dwelling estimates, should play an important role on this system.

This warning system must concentrate on a reduced number of administrative sources to avoid a large conflict data range. On the other hand, accuracy levels among administrative registers are very much different and they cover different populations, which are not complementary to each other. So we concentrate on those registers, which are most used by local and regional authorities to check census data.

5.1.1. FOR POPULATION

Electoral Register has been updated recently, but includes every one living outside, for whom there is no any other electoral registration. So every one, who has changed his "de facto" residence to elsewhere and did not update the new electoral address, *still remains as officially resident where the electoral registration belongs.* This occurs mainly with emigrants that are living and working outside but they do not change the electoral registration address, because they have temporary working contracts or even because they do not want to change their "official" residence from Portugal; sometimes they keep more than one legal residence because there is no any way to check it.

So, even after the last updating process of electoral register, there is a difference for about more 10% population with 18 or more years of age on electoral register than NSI population estimate has. Then using electoral register's figures at the parish level to estimate higher size limit for population older than 18 or more years, we are quite sure that only few parishes could be outside this checking process, even because not enrolled population, as elector is not expressive.

For people below 18 years old, school population and birth registration must be used to estimate this age group at parish level. However school population enrolled at the compulsory and the secondary education is not disaggregated to parish, but only by municipality, what implies the statistical partition by parish into the respective municipality.

Another one main concern is younger people; children belonging to first two years of age often are forgotten to be enumerated. A special attention should be asked to enumerators and on the publicity campaign to enumerate younger children

5.1.2. FOR HOUSING UNITS

Housing units are more "stable" than people and data from '91 Census have not been so surprising as to those for population. So we can start with data of last '91 Housing Census adjusted to building licenses issued by local authorities during last decade.

Otherwise housing units stock could be checked with electricity private consumers register, which assure updated data for every active consumer. Nevertheless major problem on this register concerns the disaggregation to parish: albeit data at municipality level are accurate, there is an expressive number of consumers that have no information on which parish they belong. So we have to use a statistical method to adjust this data, at least, to the smaller administrative areas.

Another source, not exactly an administrative one, refers to data on postal addresses: under the basis of an agreement with PTT company we can access to counts of private postal addresses made by postmen, some of them disaggregated to statistical subsection in the urban areas.

5.2. COMPARATIVE ANALYSIS

Many users make comparative analysis of census data with data belonging to other sources, most of them with different concepts and time reference periods.

To help users and anticipate criticism to census data we decided to make a broad comparative analysis of main comparable data sources reminding where are differences and their causes.

Data belonging to electoral register, pupils and students, employees from yearly administrative declaration and legally resident foreigners are the most used to comparative analysis with census data. All these data could be used only after data processing once is really difficult to check every individual characteristic by data collection phase.

Electoral and legally resident foreigner's registers have similar problems: people with "de facto" residence different from the registered one. Because main source of foreign immigration to Portugal are African ex-colonies there are two sources of errors when considering these registers: people that keeps their legally residence in Portugal but in fact is living outside due to business or any other reason; people "de facto" living in Portugal, but not enrolled in foreigners register (illegal ones). Because census questions don't ask about legally situations, only Post Enumeration Survey content errors allow adjusting this data for comparative analysis.

Employees from yearly administrative declaration (Quadros de Pessoal) is an administrative source used by Labour Ministry to control people belonging to enterprises and the respective establishments. They have been used to produce some labour statistics on wages, time worked, size of enterprises, etc, and was based on the

identification of each employee belonging to the enterprise in October each year. Besides there was some lack of coverage, in the comparative analysis to '91census data we realise that there was a strong consistency for employees belonging to enterprises with 10 or more people and the industry classified from C to K, according to NACE (Rev.1); this is the enterprises group with more consistent coverage by "Quadros de Pessoal". For this group of enterprises corresponding to about 1,6 million people, for a total of 2,1 million employees covered by this administrative source, the difference between '91Census data and "Quadros de Pessoal" was 1,4% more people in "Quadros de Pessoal". Regarding the estimated coverage error of -1% for resident population on census data, we can assume that a strong level of comparison consistency should be expected for these two completely independent sources.

Revista de Estatística
111ª Página
2º Quadrimestre de 1999

Estratégias de Difusão: Que meios para que público?

Autor:
Carlos Dias

VOLUME 2
2º QUADRIMESTRE DE 1999

REVISTA DE ESTATÍSTICA
113ª PÁGINA
2º QUADRIMESTRE DE 1999

ESTRATÉGIAS DE DIFUSÃO: QUE MEIOS PARA QUE PÚBLICO?

DISSEMINATION STRATEGIES: WHICH MEANS FOR WHICH PUBLIC?

Autor: Carlos Sebastião Afonso Dias
- Mestre em Estatística e Gestão de Informação pelo Instituto Superior de Estatística e Gestão de Informação da Universidade Nova de Lisboa
e
- Chefe do Serviço de Difusão e Gestão de Informação na Direcção Regional de Lisboa e Vale do Tejo do Instituto Nacional de Estatística

RESUMO:

- Tendo o INE, à semelhança dos outros INE's, adoptado, desde a sua reestruturação ocorrida em 1989, uma política de difusão orientada para o utilizador, tem vindo a desenvolver, um conjunto de produtos e serviços, como resposta às suas solicitações.

  O principal objectivo deste artigo, é o de mostrar a experiência do INE, apresentando alguns dos seus produtos e serviços estratégicos, desenvolvidos ou ainda em desenvolvimento, direccionados a públicos específicos e com recurso às novas ferramentas de difusão, que as Tecnologias da Informação nos proporcionam.

PALAVRAS-CHAVE:

- *Difusão, Estratégia, Informação, Estatística*

ABSTRACT:

- Since 1989, when reorganisation has occurred, INE has adopted, in line with other national statistical institutes, a dissemination policy based on the user's needs and has developed several products and services, as answer to their demands.

  The main goal of this document is to show the experience of INE, presenting some of its products and services, developed or currently under development, directed at specific publics and using the dissemination tools with which

KEY-WORDS:

- *Dissemination, Strategies, Information, Statistics*

VOLUME 2
2º QUADRIMESTRE DE 1999

I. INTRODUÇÃO

Durante muitos anos os utilizadores de informação estatística não foram consultados no planeamento das actividades dos órgãos oficiais de estatísticas. Os produtores de estatísticas elaboravam os seus programas em função do que eles próprios pensavam interessar aos utilizadores.

Nas últimas décadas assiste-se nos INE's dos países mais desenvolvidos a uma mudança progressiva de estratégia. Do fornecimento de dados estatísticos orientados pela oferta disponível passa-se a uma estratégia oposta, isto é, produzindo estatísticas em função da procura efectiva.

Constitui actualmente um dos principais desafios dos INE's a adopção de estratégias de difusão orientadas para as reais necessidades dos utilizadores e baseadas, cada vez mais, na personalização.

No sentido de definir as suas políticas de difusão dirigidas para o utilizador, os INE's têm como tarefas prioritárias conhecer os seus utilizadores, quais as suas necessidades e como disponibilizar a informação nos vários suportes.

Como as estatísticas sempre acompanharam de perto as evoluções das Tecnologias da Informação (é interessante lembrar que o primeiro computador para aplicações comerciais apareceu no *United States Census Bureau* (USCB), para tratar informação dos Censos), as mudanças verificadas nos últimos anos têm provocado alterações significativas na difusão da informação estatística.

Tendo o INE, à semelhança dos outros INE's, adoptado, desde a sua reestruturação ocorrida em 1989, uma política de difusão orientada para o utilizador, tem vindo a desenvolver, um conjunto de produtos e serviços, como resposta às suas solicitações.

O principal objectivo deste documento, é o de mostrar a experiência do INE, apresentando alguns dos seus produtos e serviços estratégicos (**Quadro 1**), desenvolvidos ou ainda em desenvolvimento, direccionados a públicos específicos e com recurso às novas ferramentas de difusão, que as Tecnologias da Informação nos proporcionam.

Quadro 1 – Produtos e Serviços estratégicos e principais destinatários

		DESTINATÁRIOS				
		Administração	Ensino	Agentes económicos	Agentes de informação	Internacionais
PRODUTOS E SEVIÇOS	Destaque					
	INFOLINE					
	CiberINE					
	ALEA					
	FAXSTAT					
	CONSTAT					
	GEFSTAT					
	TROIA					

II. SERVIÇOS

1. WEB SITE DO INE

Tendo acompanhado de perto a evolução das Tecnologias da Informação, não constituiu surpresa que o INE fosse um dos INE's pioneiro na utilização da Internet para difusão de informação estatística. Em Maio de 1995, disponibiliza as suas primeiras páginas na Internet. Um ano depois, tem já o seu próprio domínio, www.ine.pt, com cerca de 50 páginas e uma média de 30 acessos diários.

O Web site evoluiu e em Janeiro de 1997, a informação a difundir é organizada em 2 áreas: uma parte promocional (livre acesso) e o serviço INFOLINE (acesso restrito).

A componente promocional do Web site permite ao utilizador conhecer o INE , os principais indicadores (ao nível de NUTS III), toda a Meta-informação associada à informação estatística, produtos e serviços disponíveis e possibilita a troca de mensagens.

De referir a atribuição, pelo segundo ano consecutivo, do prémio do TOP 100 nacional para o Web site do INE.

1.1. DESTAQUE

O Destaque é um dos principais serviços disponível no Web site e constitui o meio privilegiado de difusão de informação à Comunicação Social. Antecedendo a saída das publicações, são apresentadas no Destaque, com uma periodicidade diária,

sínteses da informação produzida pelo INE. A disponibilização da informação a divulgar para toda uma semana é agendada na 6ª feira da semana anterior.

Para além dos agentes de informação, todos os utilizadores podem aceder gratuitamente aos Destaques, quer através do Web site, ou nas instalações do INE, em suporte papel.

1.2. INFOLINE

O INFOLINE - INFormação estatística On-LINE, é um serviço de acesso restrito, possibilitando a consulta e importação de informação estatística do INE Este serviço está direccionado para um público generalista, isto é, destina-se a todo o tipo de utilizadores de informação estatística.

A implementação do projecto foi faseada no tempo, tanto em termos dos meios tecnológicos envolvidos bem como dos conteúdos da informação.

Numa primeira fase a informação a disponibilizar incluiu toda a informação estatística publicada (ou a publicar) durante o ano, isto é, uma fotografia de todas as Publicações editadas em papel pelo INE.

A forma de apresentação representava uma imagem dos quadros estatísticos publicados (cerca de 5.000), somente diferindo destas pelas capacidades de organização e pesquisa (por temas e palavras-chave) que a Internet possibilitam. Associada aos dados, está também acessível a Meta-informação (Notas de síntese, Conceitos e Metodologias), permitindo aos utilizadores menos familiarizados com a estatística uma compreensão da informação e da terminologia utilizada.

Numa segunda fase, implementam-se novas facilidades, como o acesso a Base de Dados e geração de páginas dinâmicas, isto é, as páginas que são apresentadas ao utilizador são geradas no momento, resultantes da pesquisa efectuada.

Conjuntamente com as melhorias dos meios tecnológicos são adicionados novos conteúdos bem como outras formas de aceder à informação:

Pesquisa por Unidade Territorial – Permite ao utilizador aceder a um vasto conjunto de indicadores (cerca de 400) que caracterizam uma determinada região, respeitantes aos diversos níveis geográficos (NUTSI,II, III e IV). Recentemente foram incluídos cerca de 50 indicadores até ao nível de freguesia - NUTS V.

Séries Cronológicas – Estão disponíveis dados e os principais indicadores em séries retrospectivas, que variam entre os 5 e os 20 anos, com desagregação, quando possível, até ao nível de concelho (NUTS IV). Está também acessível a Meta informação associada à informação. Estão actualmente disponíveis cerca de 11.000 séries.

Estudos - Permite o acesso a estudos e análises estatísticas, editadas e divulgados pelo INE em diversas publicações; estão organizados

por área temática, o que facilita a sua consulta. Existem actualmente cerca de 200 artigos disponíveis.

É sempre possível visualizar o resumo do artigo antes de efectuar a sua consulta ou importação.

Sistema de tarifação

A tarifação utilizada no INFOLINE tem como base o volume de informação, em Kbytes, transferido (quer na consulta quer no download), sendo o valor a pagar dependente do formato dos ficheiros.

O sistema de pagamento adoptado, funciona em moldes semelhantes ao utilizado com os cartões telefónicos. O utilizador adquire um **Crediline**, com um valor em múltiplos de 25 Euros que corresponderá a um crédito, do qual serão descontados os valores correspondentes à informação consultada.

Ao adquirir o Crediline é fornecida ao utilizador um user-id e password que permite identificá-lo perante o sistema (autorizando o acesso) e navegar no INFOLINE.

Um sistema de gestão dos Credilines e das contas dos utilizadores faz o controlo automático e possibilita conhecer os seus consumos por tipo de informação, datas de acesso bem como o seu saldo (*tarifação detalhada*).

Estatísticas de Acesso

O sistema de controlo de acessos fornece também um conjunto de estatísticas que podem proporcionar análises interessantes:

- total de acessos
- número médio de acessos diários, semanais e mensais
- volume de dados transferidos
- *volume médio de transferências diárias, semanais e mensais*
- número médio de páginas consultadas
- *temas ou sub-temas com maior procura*

Estas análises permitem caracterizar melhor os nossos utilizadores, bem como as suas necessidades e poderão fornecer elementos para o planeamento de acções e novos produtos a implementar.

Actualmente o *Web site* do INE tem uma média de 400 acessos diários e são consultadas cerca de 600 páginas no INFOLINE.

A receptividade ao serviço INFOLINE tem sido muito grande, existem mais de 2.000 utilizadores registados, e estão diariamente, a surgir novos aderentes. Já foram elaborados protocolos com as Universidades e com outros organismos, para contas especiais que permitem o acesso de todos os alunos e professores ao serviço.

PERSPECTIVAS DE EVOLUÇÃO

Uma evolução, a curto prazo, em termos do conteúdo da informação será uma desagregação da informação a um nível mais detalhado. Em termos das unidades geográficas por exemplo, vai permitir obter informação ao nível de secção e subsecção estatística. É evidente que qualquer que seja o nível de detalhe a atingir devem existir mecanismos que garantam a aplicação do princípio do segredo estatístico.

O passo seguinte, em termos evolutivos, a médio prazo, poderá ser a realização de quadros a pedido, possibilitando ao utilizador a definição dos seus próprios cruzamentos, não o condicionando a escolhas pré-definidas.

Outra evolução prevista é a apresentação da informação sob a forma de cartogramas, permitindo a integração da informação geográfica com dados alfanuméricos.

Um aspecto importante em termos evolutivos enquadra-se no âmbito da comercialização. Com o evoluir da tecnologia e da possibilidade de transacções comerciais seguras através da Internet, poderá vir a utilizar-se uma tarifação on-line, isto é, o utilizador paga em tempo real, o volume da informação transferida. O pagamento processa-se de forma automática, através do cartão de crédito ou de um método seguro de transacção, no momento da recepção da informação.

Outra possibilidade adicional no caso dos quadros a pedido poderá ser o fornecimento de um orçamento prévio, possibilitando ao utilizador o conhecimento antecipado do valor a pagar pela satisfação do seu pedido.

2. CIBERINE

Dando continuidade à experiência de sucesso que constituíram os quiosques da EXPO 98, que permitiam aos utilizadores navegar pelas estatísticas do INE e de todo o Mundo, instalaram-se doze quiosques electrónicos no Hall de entrada do Edifício-sede em Lisboa.

Este serviço é dirigido a um público específico – estudantes, universitários e investigadores, que constituem cerca de 80 % dos utilizadores que se deslocam às nossas instalações.

Como objectivos globais, do ponto de vista da difusão pretende-se :

- possibilitar uma maior divulgação da informação estatística produzida pelo INE e das suas actividades;
- disponibilizar aos utilizadores, de uma forma orientada, a consulta on-line de informação estatística;
- obter indicadores de consulta.

Do ponto de vista dos utilizadores espera-se:

- facilitar o acesso à informação tornando-o mais rápido;
- poder consultar de forma gratuita a informação disponível no Web site do INE e aceder aos sites de estatística de todo o Mundo;
- possibilitar a aquisição da informação do *INFOLINE* em papel e disquete.

Como o maior volume de informação que se encontra disponível corresponde ao *INFOLINE*, a sua consulta gratuita nas instalações do INE é um importante meio promocional, que incentiva e motiva os utilizadores a adquirir um Crediline, para futura consulta a partir de suas casas, locais de trabalho, estudo, etc. Não só o objectivo promocional do INFOLINE está presente, simultaneamente fomenta-se a utilização da informação estatística e reforça-se a imagem da instituição, mostrando-se que o INE aposta na inovação e nas novas tecnologias, necessidade que tem sido reclamada muitas vezes pelos utilizadores.
Ao utilizador é permitido, para além da consulta da informação, imprimir e gravar a informação que está a visualizar. Enquanto a consulta de toda a informação é gratuita, a impressão e o download de ficheiros é tarifado.

Este serviço foi inaugurado em Setembro de 1998 e recebe actualmente uma média diária de 20 utilizadores, na sua maioria estudantes.

Perspectivas de Evolução

Para além das evoluções do próprio INFOLINE, poderá também estender-se os conteúdos do CiberINE a outras áreas, como o acesso à informação de todo o arquivo bibliográfico do espólio documental do INE e consulta de um vasto conjunto de Publicações (em avançado estado de deterioração, exemplares únicos e séries mais importantes) que estão actualmente em fase de digitalização.

3. ALEA – Projecto Acção Local de Estatística Aplicada

O projecto ALEA – Acção Local Estatística Aplicada – constitui-se como um contributo para a elaboração de novos suportes de disponibilização de instrumentos de apoio ao ensino da Estatística para aos alunos e professores do Ensino Básico e Secundário com o objectivo de melhorar a literacia estatística. Consiste na construção de uma página no Web site do INE , criação de CD-ROM e outros produtos multimédia.

Este projecto nasceu de uma ideia conjunta da Escola Secundária Tomaz Pelayo e do INE, assente nas necessidades e estruturas que os intervenientes possuem. O projecto foi apoiado e patrocinado pelo Fundo Social Europeu (FSE) e Programa de Desenvolvimento Educativo para Portugal (PRODEP).
O papel do INE neste projecto prende-se com as suas funções de difusão de dados, bem como a do fomento da utilização da informação estatística destinada, neste caso, a um público específico – estudantes e professores, facultando condições para a sua compreensão em moldes tão objectivos quanto possível.

4. FAXSTAT

O serviço FAXSTAT do INE funciona através de FAX e é direccionado para os agentes económicos. O utilizador liga para um número de FAX e recebe automaticamente na sua empresa os principais indicadores económicos produzidos pelo INE.

III. PRODUTOS

1. CONSTAT

O CONSTAT – CONcelho em ESTATística é um produto, em suporte CD ROM, resultante da colaboração entre o IPLB (Instituto Português do Livro e das Bibliotecas) e o INE, tendo em vista proporcionar ao público frequentador das Bibliotecas Municipais o acesso a informação estatística apresentada de uma forma inovadora, simples e atractiva.
A aplicação informática desenvolvida possibilita o acesso a um vasto conjunto de dados, permitindo caracterizar a realidade económica, social e cultural dos concelhos de Portugal Continental.

A informação foi seleccionada segundo duas perspectivas:

- dados do Concelho (NUTS IV) – apresentando a informação do município de cada Biblioteca, desagregada sempre que possível ao nível da Freguesia (NUTS V);
- comparações Nacionais – que permitem proceder a uma análise comparativa de indicadores para os 275 concelhos do Continente.

A informação está organizada em áreas temáticas, retratando os vários aspectos das diferentes áreas geográficas. Cada um dos temas inclui um quadro-síntese.

A informação é apresentada de uma forma atractiva e diversificada, com recurso a quadros, cartogramas e gráficos (pirâmides etárias, gráficos de barras, gráficos de linhas e outros).

Para além da informação estatística, são apresentadas graficamente outras informações de interesse sobre os concelhos, nomeadamente as principais Vias de Comunicação, Aspectos Naturais, Localidades e Pontos de Interesse Turístico.

Esta forma de apresentar a informação estatística associada a outras aspectos de interesse do concelho tem tido uma muito boa aceitação por parte dos utilizadores.

2. GEFSTAT

A Meta-informação assume cada vez maior importância na relação entre os utilizadores e o estaticista, no sentido de um diálogo cada vez mais fluido. Nesse sentido, e vindo de encontro às necessidades de alguns utilizadores que pretendiam conhecer o tipo de inquéritos que o INE e outras entidades delegadas utilizam, foi desenvolvido o GEFSTAT – Gestão de Fontes Estatísticas.

Contém cerca de 3000 páginas, com todos os inquéritos estatísticos digitalizados e inclui ferramentas de pesquisa para todas as variáveis inquiridas. É possível visualizar todos os inquéritos e variáveis, bem como imprimir e exportar os dados .

Este produto, em suporte de CD ROM, é direccionado a um público especializado e que necessite conhecer a informação que o INE recolhe, as metodologias utilizadas e os conceitos associados.

3. TROIA

O TROIA – TRade Operators Information and Analisys é um produto em CD ROM direccionado para um público especialista, principalmente os agentes económicos.

Estão representadas neste CD-ROM as operações do Comércio Intracomunitário e do Comércio com Países Terceiros, em valor e volume, realizadas no território nacional entre Janeiro de 1994 e junho de 1998.

Este produto já vai na 3ª versão, sendo o CD ROM da 1ª versão o primeiro CD ROM editado em português e em Portugal. A 2ª versão foi desenvolvida integralmente pelo INE. O TROIA tem algumas analogias com o COMEXT, produzido pelo EUROSTAT, mas apresenta uma melhor qualidade gráfica.

IV. SIDIE

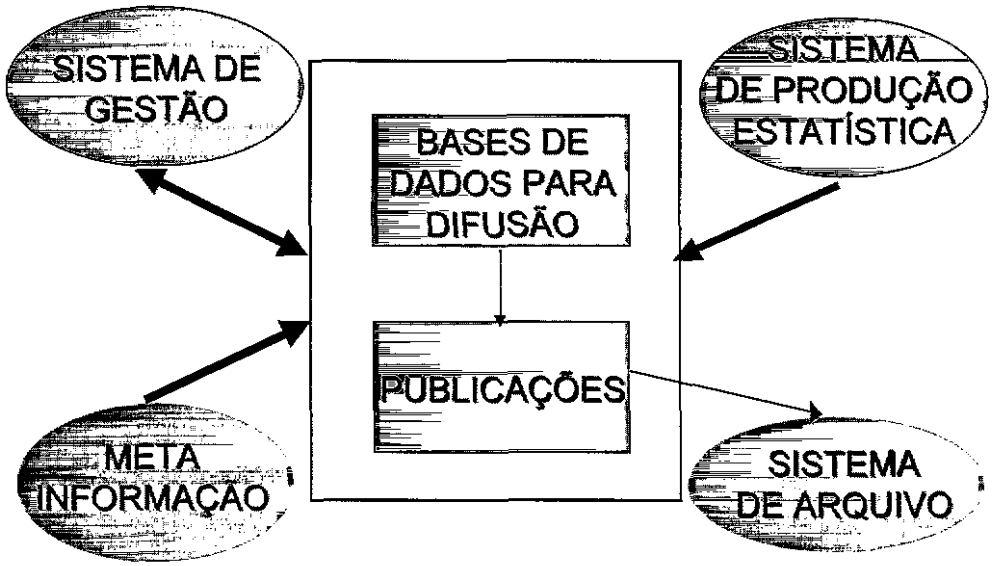
De forma a dar suporte a todos os serviços e produtos e conhecer em detalhe os utilizadores e avaliar as suas necessidades, o INE está a desenvolver um Sistema Integrado de Difusão de Informação Estatística (SIDIE).

Com o sistema a implementar pretende-se:

- possibilitar o acompanhamento do percurso de todos os pedidos de informação estatística, publicada e não publicada, desde a sua solicitação pelo utilizador até à sua entrega (orçamento, execução e facturação);
- manter o arquivo contendo o histórico de pedidos;

- constituir uma base nacional de clientes que possibilite, em qualquer momento, fornecer indicadores sobre tipo de pedidos, valores facturados, etc.;
- possibilitar a obtenção de indicadores de consulta de informação por tipos de informação, por variáveis, por tipo de utilizador, etc. ;
- organizar de forma coerente e integrada as Bases de Dados de Difusão a utilizar por todas as Direcções Regionais. O sistema deverá ser flexível de forma a possibilitar ligação aos datamarts de difusão a implementar na continuidade do projecto INFOLINE;
- A implementação do sistema será faseada.

ARQUITECTURA GLOBAL DO SISTEMA



Sistema de Produção Estatística – Inclui toda a informação proveniente da produção estatística e que alimenta as Bases de Dados para Difusão;

Sistema de Gestão – Inclui as Bases de Dados de Clientes, Pedidos, Tarifas e Facturação bem como os procedimentos de actualização e elaboração de relatórios contendo indicadores, orçamentos, facturas, etc.;

Meta Informação – Inclui os repositórios de Meta informação (conceitos, metodologias e nomenclaturas);

Sistema de Arquivo – Inclui o armazenamento das Publicações.

V. CONCLUSÕES

Ao longo deste artigo foram abordadas algumas experiências do INE relacionadas com a difusão de informação estatística e principalmente a aposta que a Internet representa como alternativa relativamente aos meios tradicionais de difusão.

Os desenvolvimentos do *Web site* e do INFOLINE, em particular, foram a aposta dos serviços a disponibilizar, utilizando este meio de difusão. O projecto ALEA, direccionado para as escolas e ao ensino da estatística também constitui uma prioridade do INE.

A aposta estratégica em termos de difusão de informação estatística, nos próximos anos, passará pela utilização da Internet como meio privilegiado de difusão. O INE deverá acompanhar estas tendências e apostar na divulgação de informação através da Internet, dando continuidade aos projectos já existentes.

No desenvolvimento de produtos, o INE elegeu como suporte privilegiado o CD ROM, tendo produzido o CONSTAT, GEFSTAT e o TROIA.

De forma a sustentar o desenvolvimento de novos produtos e serviços, o INE está a implementar um Sistema Integrado de Difusão, que possibilitará, a médio prazo, a gestão de todos os pedidos bem como a organização da informação estatística em *datamarts* - Bases de Dados específicas para Difusão.

Os produtos e serviços, apresentados neste documento, orientados para as reais necessidades dos utilizadores, reflectem a estratégica da difusão que o INE tem vindo a desenvolver nos últimos anos e constituirão o principal desafio para a próxima década.

REVISTA DE ESTATÍSTICA
125ª PÁGINA
2º QUADRIMESTRE DE 1999

INFORMAÇÕES

VOLUME 2

2º QUADRIMESTRE DE 1999

ACTIVIDADES E PROJECTOS IMPORTANTES NO ÂMBITO DO SISTEMA ESTATÍSTICO NACIONAL

IMPORTANTS ACTIVITIES AND PROJECTS IN THE SCOPE OF THE NATIONAL STATISTICAL SYSTEM

SOBRE SISTEMAS DE CODIFICAÇÃO E PROCESSOS "INTELIGENTES" DE COMPARAÇÃO DE EXPRESSÕES*

INTRODUÇÃO

Em 1991, foi desenvolvido para o recenseamento da População e da Habitação um sistema de codificação misto que integrava um subsistema de codificação automática e um subsistema de codificação assistida, designado então por C91.

Apesar das taxas de codificação automática conseguidas com este sistema terem atingido os 80% em algumas categorias, o C91 foi desenvolvido em tais moldes que nunca poderá voltar a ser utilizado em qualquer outra operação estatística.

No entanto, o aumento em produtividade e na consistência dos dados que este sistema trouxe ao processo fez com que os técnicos do INE/DSII continuassem a envidar esforços no sentido de criar um processo completamente independente do tipo de informação a processar.

Em 1991 toda a informação recolhida durante o recenseamento foi digitada manualmente, de acordo com o modelo mais clássico de recolha de dados existente.

Hoje em dia o INE está envolvido em processos de recolha de dados que recorrem a informação capturada por toda a espécie de meios, designadamente pela Internet e/ou recorrendo a soluções ICR, pelo que se tornou imperativo encontrar um sistema que lhes desse resposta.

Neste artigo começaremos por partilhar a nossa visão do que deve ser um sistema de codificação, que subsistemas deve integrar, e que tipo de funcionalidades tem de providenciar.

Depois descreveremos a nossa primeira aproximação – o "PortDix", um algoritmo de pesquisa baseado na assinatura fonética das expressões alfabéticas.

---

* Autores: Nuno Eurico Ferreira da Silva e Mário João Figueiredo, Junho de 1999

Em seguida falaremos um pouco sobre um método de codificação que é praticamente independente do tipo de dado que é suposto processar.

Por fim traçamos, ainda que genericamente, o caminho que pretendemos percorrer no futuro com este sistema.

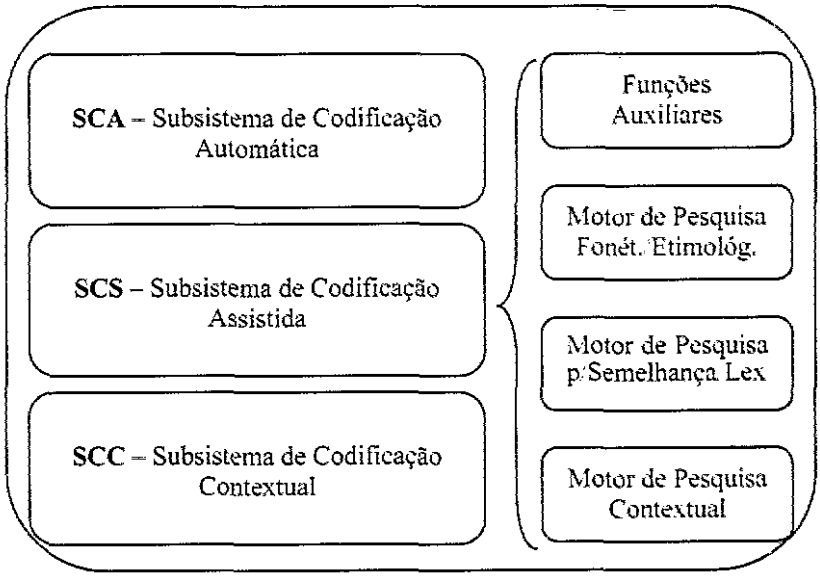
UM NOVO CONCEITO DE SISTEMA DE CODIFICAÇÃO

O conceito de sistema de codificação evoluiu bastante desde o último recenseamento, sobretudo devido à mudança entretanto ocorrida nos métodos de recolha da informação.

Presentemente o nosso objectivo é, como temos vindo a explicitar, conseguir separar o sistema de codificação do objecto a codificar.

De facto, aquilo que desenvolvemos, tal como se pode ver na figura seguinte, é uma colecção de motores de pesquisa e um conjunto de funções auxiliares, que combinamos de acordo com as circunstâncias para fazer face ao problema de codificação que se coloca em cada momento.

SICOD – Sistema Integrado de Codificação



FUNÇÕES AUXILIARES

Designamos por funções auxiliares todos os processos que permitam fazer com que os nossos motores funcionem da forma mais eficiente possível. Na prática são esta funções que preparam e assistem o trabalho de um dado motor, que determinará o grau de semelhança entre duas expressões e/ou que decidirá se lhe deve ou não atribuir o respectivo código.

Desenvolvemos basicamente dois tipos de funções:

Funções de Normalização: são utilizadas para realizar toda a espécie de alterações aos textos fornecidos pelos nossos inquiridos.

Temos condições para:

- eliminar pontuação e canonizar a acentuação e outros modificadores de caracteres;
- eliminar qualquer caracter irrelevante para o nosso contexto;
- eliminar pedaços de texto pouco significativos.

Funções de Gestão de Índices: utilizadas para criar e manter as estruturas complexas que eventualmente usamos para levar a cabo as nossas pesquisas. Dado que temos motores de pesquisa de vários tipos, as nossas funções de gestão têm de ser especialmente potentes e, sobretudo, muito versáteis e rápidas.

MOTOR DE PESQUISA FONÉTICO/ETIMOLÓGICA

Na prática estamos a falar de um motor de pesquisa que determina a semelhança entre expressões a partir da sua assinatura fonética ou, por outras palavras, tirando partido da semelhança entre a forma como se pronunciam. A semelhança torna-se etimológica se forem privilegiados os princípios das palavras, ignorados os géneros e plurais, etc.

Este tipo de motor é particularmente eficiente quando estamos a lidar com informação recolhida manualmente e digitada à posteriori – ou qualquer ambiente em que o erro prevalecente é ortográfico.

MOTOR DE PESQUISA POR SEMELHANÇA LEXICOGRÁFICA

Este motor foi a primeira aplicação do nosso conceito de "ToolBox". Evoluiu para qualquer coisa de muito poderoso e quase independente do tipo de dados a codificar.

Originalmente foi desenvolvido para trabalhar em complemento ao motor de ICR que estávamos a usar no INE (ou qualquer outro semelhante), embora com a afinação apropriada se possa explorar noutros tipos de ambiente – é particularmente bom a identificar valores com erros de edição em posição randómica.

MOTOR DE PESQUISA CONTEXTUAL

É neste motor que pretendemos investir os nossos esforços nos próximos tempos. O seu propósito será lidar com as respostas ambíguas a questões como a Profissão, o Ramo de Actividade Económica, ou mesmo a Localidade.

De acordo com as nossas perspectivas estará em funcionamento já durante a próxima operação censitária e permitirá classificar alguém que se descreve a si próprio como um pintor, podendo ser realmente um Artista Plástico ou um Trabalhador da Construção Civil.

Para atingir este objectivo o motor tem de estar dotado de um conjunto de funcionalidades que permitirão ao estaticista:

- estabelecer as relações entre campos de um dado questionário;
- descrever os termos dinâmicos em que cada relação tem lugar.

Na prática, e voltando ao nosso exemplo, poderíamos ligar a Profissão ao Ramo de Actividade Económica e às Habilitações Literárias, estabelecendo como regras, por exemplo, que alguém que trabalha para uma construtora civil ou um sub-empreiteiro e só tem escolaridade obrigatória é, de facto, um trabalhador da construção civil.

Adicionalmente – e porque o conjunto de motores de reconhecimento de expressões (fonético/etimológico, semelhança lexicográfica, proximidade de digitação, etc.) não resolve 100% dos casos (por falta de grau de certeza) = o contexto tem o potencial para os suplementar.

Por exemplo, e no caso anterior, se a profissão, tal como (parcialmente) reconhecida pelo ICR fosse "P??TOR", os restantes dados impediriam "pastor" de ser alternativa válida.

O suplemento de segurança (grau de certeza) introduzido por uma consistente análise contextual, tende a reduzir quer a taxa de insucesso quer quaisquer falsos positivos.

INFORMAÇÃO AUXILIAR

Para manter todos estes motores a funcionar somos forçados a assegurar, sob as mais diversas formas, toda a espécie de informação auxiliar. Assim sendo, mantemos:

- tabelas de caracteres a ignorar e/ou normalizar;
- tabelas de partículas a ignorar;
- tabelas de semelhança (fonéticas, gráficas, etc.);
- ficheiros de parametrização do funcionamento dos motores;
- tabelas de decisão;

etc.

SICOD – SISTEMA INTEGRADO DE CODIFICAÇÃO

Do nosso ponto de vista um sistema de codificação é uma entidade dinâmica que, para se tornar verdadeiramente produtiva, tem de sobreviver à operação estatística na qual é aplicado, ou melhor, um e o outro têm de ser completamente independentes.

Tal sistema tem de estar dotado de capacidade de:

- Codificar informação de diversas fontes (Teclado, Internet, ICR, etc.);
- Processar informação originada por diferentes tipos de questões;
- Deixar o próprio utilizador poder estabelecer e afinar o seu funcionamento.

O que estamos a construir já não é um sistema de codificação, *mas um conjunto de ferramentas que permite montar sistemas de codificação.*

A combinação, completamente definida pelo utilizador, das funções e motores de pesquisa que temos vindo a desenvolver, permite ao estaticista ter à sua disposição um sistema de codificação desenhado para cada operação.

Este paradigma constitui a base do desenvolvimento e da implementação de todo o nosso sistema. Um produto orientado para o utilizador, versátil e potente, com capacidade para codificar qualquer tipo de questão – independentemente da sua proveniência.

A APROXIMAÇÃO "TOOLBOX"

Ao iniciarmos o desenvolvimento deste projecto o nosso objectivo era evitar, tanto possível, repetir o mesmo tipo de erros cometido em 1991.

Assim estabelecemos por objectivo:

- a separação completa entre o sistema de codificação e o objecto a codificar;
- a liberdade de lidar com dicionários de respostas naturais ou com nomenclaturas oficiais;
- colocar o controlo do sistema, tanto quanto possível, nas mãos do seu utilizador.

Desta forma a aproximação a um dado problema de codificação passa a ser efectuada em duas fases.

1. A análise do problema e a identificação das ferramentas e do método adequado;

2. A combinação daqueles elementos num sistema desenvolvido por medida para resolver a situação.

Gostaríamos de deixar claro que já não estamos a desenvolver sistemas de codificação. Criámos um conceito, uma colecção, ou porque não dizê-lo, uma caixa de ferramentas, que pode ser usada para montar o sistema que mais se adequar a cada situação.

O MÉTODO PORTDIX

Aquilo que decidimos designar por método PortDix teve a sua génese na análise dos dados recolhidos durante o censo de 1991.

A observação crítica das respostas alfabéticas recolhidas durante esta operação estatística mostra que, numa operação de recolha clássica, cometem-se basicamente dois tipos de erros:

- Erros de sintaxe cometidos pelos próprios inquiridor/inquirido, quer por lapso, quer por falta de conhecimento (ex.: de Português);
- Erros de digitação cometidos pelo operador que recolhe os dados (ex.: troca uma tecla por outra).

Já em 1908 foi determinado que um algoritmo que transforme uma dada expressão num código que traduza o seu som, permite mais tarde estimar a semelhança entre expressões, ainda que elas não tenham sido escritas exactamente da mesma forma. Assim sendo, e tomando por base este algoritmo desenvolvido no início do século (o Soundex), criámos o PortDix que funciona de acordo com o mesmo paradigma: preservação da assinatura fonética das palavras, adaptando-o – naturalmente – à Língua Portuguesa.

O motor que construímos é baseado neste método e permite realizar pesquisas de toda a espécie de maneiras, designadamente:

- Procurar expressões que contenham:
  - todo um conjunto de palavras exactas;
  - um dado número de palavras exactas;
  - todo um conjunto de palavras (foneticamente) semelhantes;
  - um dado número de palavras (foneticamente) semelhantes;

mantendo ou não o ordenamento destas palavras.

TESTES

Com o propósito de medir a eficiência deste motor, seleccionámos aleatoriamente cerca de 20400 respostas à questão Profissão Principal do último recenseamento e sujeitámo-las ao nosso processo, comparando-as com a CNP 1994.

Para cada expressão comparada com a CNP 94 foi automaticamente produzida uma lista com 5 candidatos, devidamente prioritizados – análogos a uma lista/combo-box para uma escolha pelo utilizador. Sempre que um destes candidatos verificava o código atribuído em 1991, a codificação era dada como bem sucedida.

Globalmente conseguimos codificar correctamente **78%** das questões sendo que, em **65%** dos casos, a primeira sugestão da lista era a correcta.

CODIFICAÇÃO POR APROXIMAÇÃO

Durante o último ano temos estado envolvidos em vários processos de recolha de dados recorrendo a tecnologia ICR, nomeadamente:

- o "Teste de Outubro de 1998" do Censos 2001;
- o "Teste de Abril de 1999" do Censos 2001;
- o Inquérito aos Nados Vivos do DEDS.

Os problemas colocados pela codificação das respostas recolhidas no decorrer destes inquéritos mostraram-nos que os processos tradicionais de codificação não produzem o tipo de resultados que estamos habituados a esperar dos nossos sistemas.

Os motores de ICR cometem muitos erros. Confundem um caracter com outro, apagam um caracter aqui e ali, consideram alguns caracteres completamente irreconhecíveis e, por vezes, colam mesmo algumas palavras.

Embora estes erros nada tenham a ver com os erros cometidos pelos inquiridos, ou pelos operadores de registo, eles obedecem a regras, o que os torna identificáveis e nos permite tirar partido da sua ocorrência.

Assim sendo, conseguimos traçar uma espécie de tabela de semelhança gráfico/óptica de raiz topológica entre caracteres (por exemplo: U ⇔ V ou H ⇔ N, etc.) que nos permite tirar partido do padrão de erros dos motores para determinar a semelhança entre expressões.

Este motor de pesquisa associado a um conjunto de funções que ordenam lexicamente os candidatos produzidos e que, através da ponderação das suas diferenças para a expressão pesquisada, produzem um candidato apropriado, constitui agora o nosso mais usado processo de codificação.

TESTES

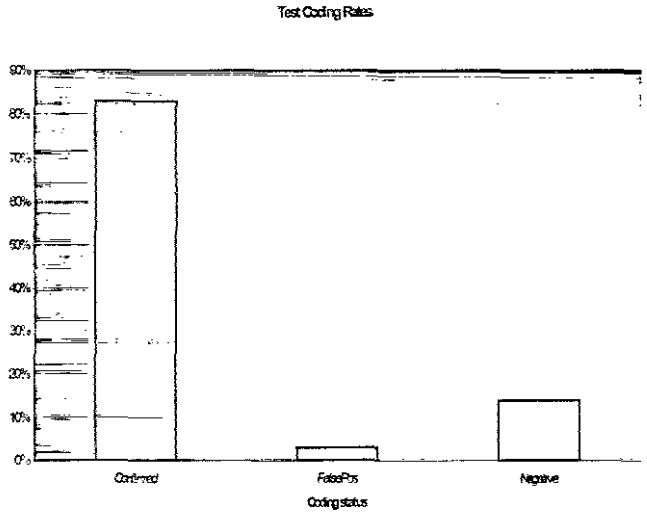
Atendendo a que à data da redacção deste artigo não existiam ainda resultados reais, volumosos e devidamente verificados, partilhamos convosco os resultados de um teste.

Podemos no entanto afirmar que no "Teste de Abril" do Censos 2001, para a questão Profissão Principal os resultados rondaram os 68% de codificação automática, embora ainda não tenha sido possível determinar com rigor a taxa de falsos positivos.

Já no teste que levámos a cabo, um processo especial pré-adulterou (de forma tão semelhante quanto possível ao motor de ICR) cerca de 20000 respostas à questão profissão, recolhidas aleatoriamente do último recenseamento.

Em seguida sujeitámos estas expressões ao processo de reconhecimento e verificámos, tal como se pode ver na figura seguinte, que:

```
------------------------
Total Read..: 0000020000
Total Found.: 0000019506
Texts%Found.: 98%
------------------------
Confirmed...: 0000016730
 Conf%Rate..: 83%
FalsePositiv: 0000000524
 FPos%Rate..: 03%
TrueNegative: 0000002725
 Negt%Rate..: 14%
------------------------
======================
Search  Time: 2:21:24
Elapsed Time: 2:22:05
======================
```



- 83% das expressões foi codificada correctamente;
- 3% das expressões foi codificada incorrectamente;
- 14% das expressões não foi codificada.

CONSIDERAÇÕES FINAIS

Nos sub-sistemas de codificação acima descritos nenhuma incursão foi feita no *reino da análise semântica. O – presentemente em desenvolvimento – sub-sistema de* contexto irá por aí. Necessariamente muito próximo das peculiaridades de cada operação estatística, pode no entanto ser concebido como uma estrutura de métodos muito independente e adaptável.

A abordagem de "ToolBox" pode bem vir a provar ser o instrumento conceptual que permita breves análises – e rápidos desenvolvimentos e implementações – de sub-sistemas de pesquisa aproximada, altamente adaptáveis.

No futuro, a integração de meios de recolha (Teclado, Internet ou ICR/OCR) com Sistemas de Codificação Automática e/ou Assistida, cuidadosamente complementados por Análise Mecânica de Contexto, prometem incrementos sólidos e mensuráveis na eficiência da recolha de dados – em escalas industriais.

Um uso muito mais confiante de quesitos abertos em operações estatísticas irá melhorar a sua convivialidade – e em última análise a qualidade das respostas. Simultaneamente os nossos métodos integrados de codificação reduzirão dramaticamente os custos das respectivas recolhas.

VOLUME 2

2º QUADRIMESTRE DE 1999

CONGRESSOS, SEMINÁRIOS, COLÓQUIOS E CONFERÊNCIAS

CONGRESS, SEMINARS AND CONFERENCES

No Estrangeiro:
Abroad:

1999

- 06 - 24 de Setembro
  School on Modern Statistical Methods in Medical Research, Trieste, Itália.
  Informações: International Centre for Theoretical Physics. P. O. Box 586, 34100 Trieste, Italy;
  E - mail: smr1122@ictp.trieste.it
  or
  E - mail: sci info@ictp.trieste.it
  WWW: http://www.ictp.trieste.it

- 13 - 16 de Setembro
  44th Annual Conference of the German Society of Medical Informatics, Biometry and Epidemiology (GMDS), Heidelberg, Germany.
  Informações: *Norbert Victor*, Department of Medical Biometry, Institute for Medical Biometry and Informatics, University of Heidelberg, Im Neuenheimer Feld 305, D-69120 Heidelberg, Germany
  or
  Lutz Edler, Biostatistics Unit, German Cancer Research Center, IM Neuenheimer Feld 280, D-69120 Heidelberg, Germany; Fax: 49 6221 564195
  E - mail: GMDS-ISCB99@dkfz-heidelberg.de
  or
  WWW: http://www.dkfz-heidelberg.de/biostatistics/GMDS-ISCB99

- 13 - 17 de Setembro
  Heidelberg Congress Week: Joint Conference of GMDS – ISCB 99, Heidelberg, Germany.
  Informações: *Norbert Victor*, Department of Medical Biometry, Institute for Medical Biometry and Informatics, University of Heidelberg, Im Neuenheimer Feld 305, D-69120 Heidelberg, Germany
  or
  Lutz Edler, Biostatistics Unit, German Cancer Research Center, IM Neuenheimer Feld 280, D-69120 Heidelberg, Germany; Fax: 49 6221 564195
  E - mail: GMDS-ISCB99@dkfz-heidelberg.de
  or
  WWW: http://www.dkfz-heidelberg.de/biostatistics/GMDS-ISCB99

- 14 - 17 de Setembro
  20th Annual Conference of the International Society of Clinical Biostatistics (ISCB), Heidelberg, Germany.
  Informações: *Norbert Victor*, Department of Medical Biometry, Institute for Medical Biometry and Informatics, University of Heidelberg, Im Neuenheimer Feld 305, D-69120 Heidelberg, Germany
  or
  Lutz Edler, Biostatistics Unit, German Cancer Research Center, IM Neuenheimer Feld 280, D-69120 Heidelberg, Germany; Fax: 49 6221 564195
  E - mail: GMDS-ISCB99@dkfz-heidelberg.de
  or
  WWW: http://www.dkfz-heidelberg.de/biostatistics/GMDS-ISCB99

- 14 - 18 de Setembro
  Second European Conference on Highly Structured Stochastic Systems, Pavia, Itália.
  Informações: WWW: http://www.unipv.it/hsss99/hsss.html

- 22 – 24 de Setembro
  Third International Conference on Survey and Statistical Computing, Edimburgo, Escócia.
  Informações: *Diana Elder*, Adminitrator, ASC, P.O. Box 60, Besham, Bucks, HP5 3QH, UK; Telf./Fax: 44-1494-973033
  E - mail: asc@essex.ac.uk

- 30 de Setembro – 01 de Outubro
  Biométrie et Epidémiology 99. Topic: Diffusion of rarely-used statistical methods in Epidememiology., Vannes, França.
  Informações: *Jean-François Petiot*, Institut Universitaire de Techonologie, 8 rue Montaigne, 56017 Vannes, France; Telf.: 33-6-10126396; Fax: 33-2-97463190;
  E - mail: epibiostat@iu-vannes.fr
  URL: http://www.iut.iu-vannes.fr

- 05 – 06 de Outubro
  PLS'99 International Symposium on PLS Methods., HEC campus, Paris France.
  Informações: *Concesa Oliver*, CISIA-CERESTA Symposium PLS'99, 1 avenue Herbillon – 94160 Saint-Mande (France), Telf.: ++33 1 43 74 95 20; Fax: ++33 1 43 74 17 29;
  E - mail: cisia@calva.net
  Web: www.cisia.com

- 22 – 23 de Outubro
  Workshop on "Correlated data modeling: the estimating equations approach"., Trieste, Itália.
  Informações: *Dario Gregori*, Department of Economics and Statistics, University of Trieste, P. le Europa 1, 34127 Trieste, Italy; Telf.: 39-40-6767927; Fax: 39-40-567543;
  E - mail: gredar@univ.trieste.it
  URL: http://dises17.univ.trieste.it/Wuts/wuts.htm

- 28 – 31 de Outubro
  International Conference on Nonresponse, Portland, Oregon, USA.
  Informações: *Joint Program in Survey Methodology*,
  URL: http://www.bsos.umd.edu/jpsm
  Specific questions can posed to: icsn@survey.umd.edu;
  URL: http://www.jpsm.umd.edu/icsn99;
  Telf./Fax:44-1494-973033;
  E - mail: asc@essex.ac.uk

* - Novas entradas – *Denotes new Entries.*

VOLUME 2
2º QUADRIMESTRE DE 1999

ACÇÕES DESENVOLVIDAS PELO INE NO ÂMBITO DA COOPERAÇÃO BILATERAL E MULTILATERAL

ACTIONS ACHIEVED BY NSI IN THE SCOPE OF BILATERAL AND MULTILATERAL COOPERATION

(DE 1 DE MAIO A 31 DE AGOSTO DE 1999):

a) Cooperação desenvolvida com os PALOP e Macau:

No âmbito do Programa Estatístico da Comunidade dos Países de Língua Portuguesa (CPLP), foi elaborado o Ante-projecto de Desenvolvimento dos Sistemas de Informação sobre Estatísticas da Educação dos Países Africanos de Língua Portuguesa, encontrando-se em período de apreciação pelas diversas entidades envolvidas.

Realizou-se na cidade da Praia, Cabo Verde, a 8ª Reunião Internacional sobre Estatísticas Sociais dos Países de Língua Oficial Portuguesa – RIESLOP, na qual o Dr. J. Graça Costa foi eleito Presidente.

No quadro do projecto comum sobre Classificações, Conceitos e Nomenclaturas, realizou-se, no período em apreço uma missão de assistência técnica ao INE de Moçambique, pelo Dr. Saraiva Aguiar, para formação dos utilizadores da Classificação de Actividades Económicas (CAE), e estágios de desenvolvimento dos trabalhos da Classificação Nacional de Bens e Serviços (CNBS), em que participaram o Dr. Francisco Rodrigues, do INE de Cabo Verde, Dr. Jorge Utui, do INE de Moçambique e Dra. Isabel Mendes do INEC da Guiné-Bissau.

No âmbito do projecto-piloto para a implementação do Sistema de Contas Nacionais das Nações Unidas (SCN-93), foi realizada uma missão de assistência técnica ao INE de Moçambique pelo Dr. Idílio Freire.

Realizou-se em Abidjan um Seminário sobre o Índice de Preços no Consumidor na UEMOA, no qual o Dr. Daniel Santos participou, tendo apresentado o seu relatório de avaliação do software CHAPO.

No período em apreço, e no âmbito da cooperação bilateral com os PALOP, foram realizadas as seguintes acções:

CABO VERDE

Na execução do programa de cooperação foram realizadas; uma missão de identificação no âmbito do projecto Índices do Comércio Externo, uma visita de trabalho do Senhor Presidente do Conselho Nacional de Estatística de Cabo Verde, Dr. Edgard Chrysostome Pinto e um estágio no âmbito do projecto Tecnologias de Informação e Informática e uma missão de assistência técnica no âmbito do projecto Estatísticas das Empresas.

A missão de identificação no âmbito do projecto Índices do Comércio Externo, realizada pela Dra. Ana Antunes, teve como principal componente a preparação do plano de trabalho do INE de Cabo Verde neste domínio.

A visita de trabalho, efectuada pelo Dr. Edgard Chrysostome Pinto incidiu na organização e funcionamento do Conselho Superior de Estatística (CSE) e incluiu ainda diversos contactos ao nível do Departamento de Coordenação e Contas Nacionais, Serviço do Secretariado do CSE, ISEGI e CESD-Lisboa.

O estágio no âmbito do projecto Tecnologias de Informação e Informática, realizado pelo Eng. João Baptista Lopes de Pina, incluiu a frequência de duas acções de formação teórica no domínio da Administração e Configuração de Redes em Windows NT e a execução de diversos trabalhos práticos no Departamento de Sistemas de Informação e Informática.

A *missão de assistência técnica no âmbito do projecto Estatísticas das* Empresas, realizada pela Engª. Júlia Cravo, teve como principal componente a preparação da análise de resultados e resolução de algumas questões metodológicas do Recenseamento Empresarial de Cabo Verde.

GUINÉ-BISSAU

Em virtude da incapacidade para assegurar o normal regresso de alguns técnicos do INEC à Guiné-Bissau, foram prolongados os respectivos estágios no INE, nas áreas de elaboração de Estudos Prévios para o Programa de Cooperação Estatística no quadro da CPLP, Contas Nacionais e Estatísticas da Agricultura e Pescas.

SÃO TOMÉ E PRÍNCIPE

Na execução do programa de cooperação foram realizados estágios nos domínios das Estatísticas Demográficas e Sociais e Registo Nacional de Pessoas Colectivas.

O estágio no âmbito do projecto Estatísticas Demográficas e Sociais, realizado pela Dra. Armilinda Pereira no Departamento de Estatísticas da População, incidiu na análise dos diversos produtos estatísticos nacionais nesta área, incluindo contactos com a Direcção Geral de Saúde, Gabinete de Estudos do Ministério da Justiça e Departamento de Estatística do Ministério da Educação.

Em continuidade do projecto do Registo Nacional de Pessoas Colectivas, o estágio realizado pelo Dr. Adelino de Freitas teve como objectivo dar continuidade à preparação da organização dos serviços no INE de São Tomé e Príncipe, na sequência dos trabalhos no âmbito do Ficheiro de Unidades Estatísticas, tendo incluído a realização de diversas reuniões de trabalho na Direcção Regional de Lisboa e Vale do Tejo e no Departamento de Coordenação e Contas Nacionais.

MACAU

Durante o período considerado foi realizada, pelo Dr. Albano Miranda, uma missão de assistência técnica à DSEC – Direcção de Serviços de Estatística e Censos, no domínio do Inquérito aos Orçamentos Familiares.

b) Cooperação desenvolvida com os PECO, no quadro do Programa PHARE:

No âmbito do Programa PHARE de Assistência Técnica aos Países da Europa Central e Oriental (PECO), realizaram-se, durante o período mencionado em epígrafe, nove acções de cooperação.

Três dessas acções realizaram-se no âmbito de **Projectos Piloto destinados aos países PHARE.**

No quadro do projecto piloto Exaustividade das Contas Nacionais, a Drª Ana Leal, realizou uma missão de assistência técnica. Esta acção teve lugar em Londres (14-15 de Junho de 1999) e consistiu numa reunião entre todos os peritos da U.E: envolvidos no projecto e o Eurostat. O objectivo desta reunião foi a apresentação, por parte dos peritos, de relatórios intermédios.

No âmbito do projecto piloto Estatísticas das Finanças Públicas, a Drª Ana Leal realizou uma missão ao GUS-Polónia (18 a 21 de Maio) e uma missão ao NSO-Eslováquia (24 a 26 de Maio), cujo o objectivo foi o de definir os trabalhos a realizar pelos técnicos destes dois organismos estatísticos no âmbito do projecto.

No âmbito da **Cooperação Bilateral**, efectuaram-se cinco acções de cooperação com o *National Commission for Statistics* da Roménia.

As primeiras duas acções, que tiveram lugar entre 3 a 7 de Maio, consistiram em dois estágios ao INE no âmbito dos Índices de Produtos Agrícolas e Contas Económicas da Agricultura. A primeira acção insere-se num novo projecto e teve como objectivo identificar os trabalhos que o INE realiza neste âmbito. A segunda acção consistiu na apresentação da nova metodologia para as Contas Económicas da Agricultura e sua relação com as Contas Nacionais e Trimestrais.

A terceira acção, que revestiu a forma de missão, foi realizada pela Drª Emília Saleiro, no âmbito do Programa PHARE Nacional - Contas Nacionais/Regionais, e teve como objectivos a identificação e análise das fontes e métodos utilizados pela Roménia para o cálculo das Contas Regionais. Esta acção teve lugar no período de 14 a 18 de Junho.

O Engº Pinto Martins realizou uma missão à Roménia, na semana de 5 a 9 de Julho, no âmbito do Programa PHARE Nacional - Difusão e teve como objectivo a identificação do equipamento tipográfico utilizado pela NCS para a impressão de publicações estatísticas e levantamento das necessidades nesta área, com vista à preparação de um concurso internacional de aquisição de equipamento tipográfico.

Por último, e ainda inserido no projecto identificado no parágrafo anterior, o Dr. Paulo Mateus participou, nos dias 5 e 6 de Agosto, numa reunião entre o INSEE (França) e o NCS com vista ao planeamento das actividades e definição dos aspectos contratuais relativos à participação conjunta INE/INSEE neste projecto.

Outros

O Dr. João Morais participou num Seminário destinado aos países PHARE sobre Estatísticas Estruturais das Empresas, realizado na Polónia entre 7 a 9 de Junho, tendo apresentado a experiência portuguesa ao nível da aplicação do Regulamento Comunitário sobre Estatísticas Estruturais das Empresas.

FUNDAMENTO, OBJECTO E ÂMBITO

O INE, consciente de como uma cultura estatística é essencial para a compreensão da maioria dos fenómenos do mundo actual, e da sua responsabilidade na divulgação do conhecimento estatístico, fazendo-o chegar ao maior número possível de leitores, tendo reconhecido a necessidade de dar um passo nesse sentido, passa a editar quadrimestralmente a presente Revista de Estatística destinada a divulgar:

a) Numa perspectiva científica, artigos originais sobre temas especializados da estatística, tanto pura como aplicada, bem como sobre estudos e análises nos domínios económico, social e demográfico;

b) Informações sobre actividades e projectos importantes no âmbito do Sistema Estatístico Nacional;

c) Informações sobre congressos, seminários, colóquios e conferências de interesse estatístico ou afim;

d) Informações sobre acções desenvolvidas pelo INE no âmbito da cooperação bilateral e multilateral.

Para tal, são adoptadas as seguintes formas de contribuição para publicação na Revista:

- Quanto aos artigos referidos em a), contribuições da iniciativa dos próprios autores e por convite do Conselho Editorial, pertencentes ou não ao INE;

- Quanto às informações referidas em b), c) e d), contribuições dos departamentos do INE.

As contribuições por iniciativa dos próprios autores serão objecto de avaliação de mérito científico pelo Conselho Editorial, que decidirá ou não pela respectiva publicação.

Para a elaboração e envio das contribuições para publicação na Revista são adoptadas as Normas de Apresentação de Manuscritos que figuram na última página.

Os autores dos artigos publicados, a que se refere a alínea a), receberão uma contribuição financeira paga pelo INE, de montante a fixar por despacho da Direcção mediante proposta do Director da Revista.

FOUNDATION, SUBJECT MATTER AND SCOPE

INE is conscious of how statistical awareness is essential to the understanding of the majority of phenomena in the present world and is aware of its responsibility to disseminate statistical knowledge, making it available to the widest possible range of readers. INE has recognised the need to take a step in that direction and will begin publication of this *Statistical Review* three times yearly, designed to provide the following:

a) Within a scientific perspective, original articles on specialised areas of statistics, both *pure and applied, as well as studies and analyses within the sphere of economics,* social issues and demographics;

b) Information on activities and projects within the scope of the National Statistical System;

c) Information on congresses, seminars and conferences of a statistical or related nature;

d) Information on activities developed by INE within the scope of bilateral or multilateral co-operation;

The following approaches for contributing material for publication in the review have been adopted:

- In relation to the articles referred to in section a), contributions are made by the authors themselves and by invitation of the Editorial Committee, whether they are employees of INE or not;

- In relation to the information referred to in section b), c) and d); contributions are from departments of INE.

The Editorial Committee who has sole discretion in deciding whether or not the material will be published will assess the scientific merit of contributions made on the initiative of the authors themselves.

The preparation and delivery of material for publication in the Review are subject to the Rules for Submitting Manuscripts presented on the last page.

The authors of the published articles referred to in section a) will receive pecuniary compensation from INE in an amount to be determined by resolution of the Board on the recommendation of the Director of the Review.

NORMAS DE APRESENTAÇÃO DE MANUSCRITOS

Nos termos da alínea b) do nº. 3 do Artigo 5º do Regulamento da *Revista de Estatística* do Instituto Nacional de Estatística, o Conselho Editorial aprovou as seguintes **Normas de Apresentação de Manuscritos**:

1. Os originais dos artigos serão enviados ao Director da Revista pelos respectivos autores, devendo ser escritos em português e não terem sido ainda totalmente publicados, ou estar em processo de edição em qualquer outra publicação.
2. Poderão também ser apresentados artigos escritos em inglês, cabendo ao Director da Revista a decisão sobre a sua aceitação.
3. Quanto à *avaliação do mérito científico* dos artigos:
   a) Os artigos apresentados por iniciativa dos respectivos autores serão submetidos à avaliação do mérito científico pelo Conselho Editorial, com garantia do anonimato tanto do autor como dos avaliadores;
   b) Os autores receberão a informação sobre o resultado da avaliação num prazo máximo de trinta e cinco dias, com indicação, nos casos de avaliação positiva, do número da *Revista* em que serão publicados, e nos casos de avaliação negativa com a devolução do artigo apresentado e respectiva *disquette*, com indicação do(s) software(s) adicional(ais) eventualmente utilizado(s) na produção do documento original.
4. Os artigos aceites para publicação na *Revista de Estatística* serão igualmente divulgados no *site* do INE na *Internet*.
5. Os originais, com uma extensão não superior a trinta páginas, serão processados em *Word for Windows*, integralmente a preto e branco, e entregues em suporte papel acompanhado da respectiva *disquette*.
6. Na apresentação dos originais, os autores respeitarão ainda as seguintes normas:
   6.1. Quanto à *estrutura*:
      a) O texto deve ser dactilografado em formato A_4, com utilização do tipo de letra *Times New Roman* - 11, e com as seguintes margens: *top*: 2,5 cm, *bottom*: 2 cm, *left*: 2,5 cm, *right*: 5 cm;
      b) A primeira página conterá exclusivamente o título do artigo, bem como o nome, morada e telefone do autor, com indicação das funções exercidas e da instituição a que pertence, devendo, no caso de vários autores, ser indicado a quem deverá ser dirigida a correspondência da Revista;
      c) A segunda página conterá, em português e inglês, unicamente o título e um resumo do artigo, com um máximo de cem palavras, seguido de um parágrafo com indicação de palavras-chave até ao limite de quinze;
      d) Na terceira página começará o texto do artigo, sendo as suas eventuais secções ou capítulos numeradas sequencialmente;
   6.2. Quanto a *referências bibliográficas*:
      a) Os autores eventualmente citados no texto do artigo serão indicados entre parênteses curvos pelo seu nome seguido da data da respectiva publicação e, se for caso disso, do número de página (p. ex.: Malinvaud, 1989, 23);
      b) As referências bibliográficas serão listadas, por ordem alfabética dos apelidos dos respectivos autores, imediatamente a seguir ao final do texto, de acordo com a fórmula seguinte:

ANDERSON, C.W., and TURKMAN, K.F, (1995) "Sums and maxima of stationary sequences with heavy tailed distributions", *Sankhya*, Vol. 57, Series A, pp.1-10.

6.3. Quanto à *revisão de provas e publicação*:

a) Uma vez aceite o artigo e antes da sua publicação, receberá o autor dois exemplares de provas para revisão, um dos quais será devolvido ao Director da Revista no prazo máximo de uma semana contado da data da sua recepção;

b) Serão da responsabilidade dos respectivos autores as consequências de eventuais modificações da versão inicial aceite, bem como de atrasos na revisão das provas, que impossibilitem a publicação no número da Revista previsto, reservando-se o Conselho Editorial o direito de decidir a data da sua publicação futura;

c) Uma vez publicado o artigo, o autor receberá vinte exemplares da sua versão impressa e um exemplar do respectivo número da *Revista*.

7 Para informações adicionais contactar o Secretariado de Redacção:

Eduarda Liliana Martins
Instituto Nacional de Estatística
Avª. António José de Almeida, nº. 5 – 9º.
1 000 Lisbon - Portugal
Tel.: +351 1 842 61 00 (3905) Fax.: +351 1 842 63 66 e-mail: liliana.martins@ine.pt

RULES FOR SUBMITTING MANUSCRIPTS

Within the terms of sub-section a of no. 3 of Article 5 of the regulations of the *Statistical Review* of the National Statistical Institute (INE), the Editorial Committee has approved the following **Rules for Submitting Manuscripts**:

1. The original articles will be sent to the Review Director by the respective authors. They should be written in Portuguese, they should not have already been published in their entirety nor should they be in the process of being published in any other publication.
2. Articles may also be submitted in English to the Review Director who will decide whether to accept them.
3. In relation to the *evaluation of the scientific merit* of the articles:
   a) The Editorial Committee will assess all articles submitted on the initiative of the respective authors on the basis of their scientific merit. The identity of both the author and the Committee members will be strictly confidential;
   b) The authors will receive information regarding the results of the evaluation within a maximum period of thirty-five days. If the article is accepted, the Committee will indicate the issue number of the *Review* in which the article will be published. If the article is not accepted, it will be returned along with the respective diskette, with the information on the addicional(s) software(s) eventually used in the production of the original document.
4. The articles accepted for publication in the *Statistical Review* will also be made public on the INE Internet site.
5. The original articles having no more than thirty pages must be processed in *Word for Windows*, completely at black and white, and they will be delivered in hard copy as well as on diskette.
6. With the presentation of the original articles, the authors must also respect the following rules:
   6.1 In relation to the *structure*:
      a) The text shall be printed on A4 format paper utilising the font *Times New Roman* size 11 and with the following margins: top: 2.5 cm, bottom: 2 cm, left: 2,5 cm, right: 5 cm;
      b) The first page shall contain only the title of the article as well as the name, address and telephone number of the author, indicating the position held and the institution that he/she belongs to. In the case of various authors, it is necessary to indicate the person to whom all correspondence received by the *Review* should be forwarded;
      c) The second page shall contain only the title and a abstract of the article in Portuguese and English with the maximum of one hundred words followed by a paragraph indicating key words up to the limit of fifteen;
      d) The third page will begin the text of the article with its respective sections or chapters sequentially numbered;
   6.2 Regarding *bibliographical references*:
      a) Authors who are cited in the text of the article shall be indicated in parentheses with their name followed by the date of the respective publication and, if necessary, the page number (ex.: Malinvaud, 1989, 23);

b) All bibliographical references will be listed in alphabetical order by the surnames of the respective authors, immediately following the end of the text, as in the following example:

ANDERSON, C.W., and TURKMAN, K.F., (1995) "Sums and maxim of stationary sequences with heavy tailed distributions", *Sankhya*, Vol. 57, Series A, pp. 1-10.

6.3 Regarding *proof-reading and publication*:

a) Once the article is accepted and prior to its publication, the author will receive two copies for review. One of these copies will be returned to the Director of the Review within a maximum period of one week from the date of its reception;

b) The consequences of subsequent changes to the accepted first version are the responsibility of the respective authors as well as any delays in proof-reading that make its publication in the planned issue of the Review impossible. The Editorial Committee reserves the right to decide upon the date for future publication;

c) Once the article is published, the author will receive twenty copies of his/her printed version and a copy of the respective issue of the *Review*.

7. For further information kindly contact the Editorial Secretary:

Eduarda Liliana Martins
Instituto Nacional de Estatística
Avª. António José de Almeida, nº. 5 – 9º.
1 000 Lisbon - Portugal
Tel.: +351 1 842 61 00 (3905) Fax.: +351 1 842 63 66 e-mail: liliana.martins@ine.pt

BOLETIM DE ENCOMENDA

Nome _____ Data de nascimento: __/__/__

Profissão/Função _____ Instituição/Empresa _____

Telef.:_____ Fax: _____

DESEJO RECEBER OS EXEMPLARES DA REVISTA DE ESTATÍSTICA:

Em casa . ☐ Na Instituição/empresa ☐

Morada para envio: _____

Localidade: _____ Código Postal: _____

Autorizo débito no cartão Visa ☐ ou Mastercard ☐

nº.: ☐☐☐☐ ☐☐☐☐ ☐☐☐☐ ☐☐☐☐

Valor da transacção: 6. 900$00 Validade do cartão __/__/__

☐ Junto cheque nº_____ à ordem do *INSTITUTO NACIONAL DE ESTATÍSTICA* sobre o Banco _____

Data: __/__/__ Assinatura: _____

OS DADOS RECEBIDOS SERÃO PROCESSADOS AUTOMATICAMENTE E DESTINAM-SE AOS ENVIOS RELACIONADOS COM A SUA ASSINATURA, RESPECTIVAS OPERAÇÕES ADMINISTRATIVAS E ESTATÍSTICAS, E À EVENTUAL APRESENTAÇÃO DE OUTROS PRODUTOS E SERVIÇOS DO INSTITUTO NACIONAL DE ESTATÍSTICA.

BOLETIM DE ENCOMENDA

Nome _____ Data de nascimento: __/__/__

Profissão/Função _____ Instituição/Empresa _____

Telef.:_____ Fax: _____

DESEJO RECEBER OS EXEMPLARES DA REVISTA DE ESTATÍSTICA:

Em casa . ☐ Na Instituição/empresa ☐

Morada para envio: _____

Localidade: _____ Código Postal: _____

Autorizo débito no cartão Visa ☐ ou Mastercard ☐

nº.: ☐☐☐☐ ☐☐☐☐ ☐☐☐☐ ☐☐☐☐

Valor da transacção: 6. 300$00 Validade do cartão __/__/__

☐ Junto cheque nº_____ à ordem do *INSTITUTO NACIONAL DE ESTATÍSTICA* sobre o Banco _____

Data: __/__/__ Assinatura: _____

OS DADOS RECEBIDOS SERÃO PROCESSADOS AUTOMATICAMENTE E DESTINAM-SE AOS ENVIOS RELACIONADOS COM A SUA ASSINATURA, RESPECTIVAS OPERAÇÕES ADMINISTRATIVAS E ESTATÍSTICAS, E À EVENTUAL APRESENTAÇÃO DE OUTROS PRODUTOS E SERVIÇOS DO INSTITUTO NACIONAL DE ESTATÍSTICA.

AUTORIZADO PELOS CTT
NO SERVIÇO NACIONAL

RSF

NÃO PRECISA DE SELO

INSTITUTO NACIONAL DE ESTATÍSTICA
SECÇÃO VENDA DE INFORMAÇÃO

Av. Antonio Jose de Almeida
1000-043 LISBOA

AUTORIZADO PELOS CTT
NO SERVIÇO NACIONAL

RSF

NÃO PRECISA DE SELO

INSTITUTO NACIONAL DE ESTATÍSTICA
SECÇÃO VENDA DE INFORMAÇÃO

Av. Antonio Jose de Almeida
1000-043 LISBOA


VA049902