I. S. E. G.
Biblioteca
HD30.2.V54
2003 5302

Universidade Técnica de Lisboa

INSTITUTO SUPERIOR DE ECONOMIA E GESTÃO

Departamento de Economia

O processamento analítico, elemento crucial da análise de maturidade dos sistemas de informação para a Economia

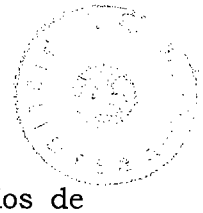
Luís Manuel Ribeiro Vieira

Júri:

Presidente	Reitor da Universidade Técnica de Lisboa
Vogais	Doutor José Dias Coelho, professor catedrático da Faculdade de Economia da Universidade Nova de Lisboa
	Doutor Vitor Manuel da Silva Santos, professor catedrático do Instituto Superior de Economia e Gestão da Universidade Técnica de Lisboa
	Doutor José António Correia Pereirinha, professor catedrático do Instituto Superior de Economia e Gestão da Universidade Técnica de Lisboa
	Doutor Luís Alfredo Martins do Amaral, professor associado da Escola de Engenharia da Universidade do Minho
	Doutor Manuel de Jesus Farto, professor associado do Instituto Superior de Economia e Gestão da Universidade Técnica de Lisboa
	Doutor António Maria Palma dos Reis, professor associado do Instituto Superior de Economia e Gestão da Universidade Técnica de Lisboa
	Doutor Ilídio Rodrigues Antunes, professor auxiliar do Instituto Superior de Economia e Gestão da Universidade Técnica de Lisboa

Dissertação orientada pelo Professor Doutor Manuel de Jesus Farto e co-orientada pelo Professor Doutor Ilídio Rodrigues Antunes, apresentada no Instituto Superior de Economia e Gestão para obtenção do grau de Doutor em Economia

Lisboa, Outubro de 2003



Resumo

Esta dissertação revê e reavalia a utilização de uma teoria de estádios de desenvolvimento dos sistemas de informação num ambiente onde as redes de computadores têm um papel determinante e onde as organizações que disponibilizam informação para a economia constituem o objecto de análise.

Considerando a formulação dos estádios de desenvolvimento numa constante mutação através da alteração das variáveis de análise, desenvolvemos uma métrica aplicável à chamada era das redes onde se consideraram relevantes as características do sítio e as características da informação que o sítio disponibiliza.

Ao longo de três anos, recolhemos dados sobre dezasseis organizações que disponibilizam informação para a economia. Os dados relativos ao primeiro ano (1999) permitiram-nos pôr a hipótese da existência de quatro estádios: início da disponibilização de dados e publicações, procura de actualidade, procura da integridade temporal e abertura ao analista, embora só os três primeiros tivessem uma manifestação efectiva. Os dados relativos aos anos seguintes (2000 e 2001) confirmaram o sentido da evolução, mas não conseguiram evidenciar uma efectiva concretização da abertura ao analista. Assim, concluímos que se torna necessária uma transformação cultural com múltiplas dimensões que leve à remoção dos obstáculos à transição para aquele estádio.

Palavras Chave: Internet, Sistemas de Informação para a Economia, Métrica, Maturidade, Estádio de desenvolvimento, Processamento Analítico

Abstract

This dissertation reviews and re-evaluates the use of a stages of growth theory for information systems in an environment where the computer networks have a determinative paper and whose object are organizations that make available information about economy.

Considering the formulation of stages of growth in a constant mutation through modification of variables of analysis, we develop a metric applicable to networks era, where the characteristics of the site and the characteristics of the information that the site makes available are the relevant elements.

Throughout three years, we have been collecting data on sixteen organizations. The data relative to the first year (1999) allowed to put the hypothesis of the existence of four stages: beginning of offering of data and publications, search of actuality, looks for integrity in time and opening to the analyst, even so the three first ones only had an effective manifestation. The data relative to following years (2000 and 2001) confirmed the direction of the evolution, but they were not able to evidentiate an effective concreteness of the opening to the analyst. Thus, we conclude that a cultural transformation with multiple dimensions becomes necessary to led to the removal of the obstacles to the transition for that stage.

Key Words: Internet, Information Systems for the Economy, Metric, Maturity, Stage of Growth, Analytical Processing

Índice

Resumo ..... **i**
Abstract ..... **ii**
Índice ..... **iii**
Lista de Quadros ..... **viii**
Lista de Figuras ..... **xi**
Lista de Expressões ..... **xiii**
Lista de Expressões ..... **xiii**
Agradecimentos ..... **xv**
Glossário ..... **xvii**
Introdução **1**
1. Problemática da maturidade dos sistemas de informação e sua referência à área de economia 5
1.1. Dados do problema 6
1.2. Problemas por resolver 13
1.3. Objecto de análise 14
1.4. Métodos de Análise 15
1.5. Experimentação 16
1.6. Produtos da dissertação 17
2. Planificação global da dissertação 17
Capítulo I - Questões da Maturidade dos Sistemas de Informação **23**
1. A Maturidade como estádio final do processo de desenvolvimento dos sistemas de informação 23
1.1. O modelo de Nolan 27
1.2. O modelo de Cash, McFarlan e McKenney 28
1.3. O modelo de Earl 29
1.4. O modelo de Bhabuta 30
1.5. O modelo de Hirscheim 31
1.6. O modelo de Sutherland e Galliers 33
2. A maturidade e a questão da evidência empírica 35
2.1. A curva orçamental em forma S 36
2.2. Sistemas de Custeio 36
2.3. Qualificações do gestor da área 37
2.4. Administração de Dados (AD) 37
2.5. Localização da Função SI na organização 38
2.6. Aplicação de Benchmarks 38
2.7. Uso de planeamento e controle 39
2.8. Steering Committees 39
2.9. A efectiva evidência empírica 39

3. Questões da articulação dos SI com as organizações decorrentes dos autores da teoria de estádios 41
3.1. Processamento de dados externos 42
3.2. Integração de aplicações 42
3.3. Inovação e uso estratégico dos sistemas de informação 43
3.4. Gestão dos dados e da informação como um recurso da organização 45
3.5. Relação com o utilizador 47
3.6. Novas funções 50
3.7. Novas qualificações 51
4. Era das redes, sociedade da informação, economia digital, economia do conhecimento 52
5. Novos elementos para análise da maturidade dos sistemas de informação 56
5.1. Guias e roteiros para a informação 57
5.2. Integridade temporal 57
5.3. Multidimensionalidade 58
5.4. Descoberta de conhecimento em bases de dados 58
6. Contribuições para a análise da maturidade dos SI na era das redes 59
Resumo do Capítulo I 61
Capítulo II - Processamento Analítico 63
1. Bases de Dados Analíticas 64
1.1. Bases de dados 65
1.2. Características essenciais das BDAs 67
1.3. Outras características das BDAs 68
1.4. Estrutura de uma Base de Dados Analítica 69
1.5. Especificação 70
1.6. Do Modelo Entidade Associação ao Modelo Multidimensional 76
2. Extracção de Informação de Bases de Dados 77
2.1. Interrogação 78
2.2. Análise Multidimensional 82
3. Descoberta de Conhecimento em Bases de Dados 87
3.1. Classificação 87
3.2. Processos Difusos 100
3.3. Pesquisa de Associações 106
3.4. Processos de pesquisa de Clusters 116
4. Outros processos utilizados na dissertação 125
4.1. A análise factorial 125
4.2. Análise discriminante 131
Resumo do capítulo II 135
Capítulo III - Organizações responsáveis pela Informação para a Economia em Portugal 137

1. A disponibilização e utilização de informação para a Economia como um processo 139
1.1. Organização do processo de disponibilização 140
1.2. Organização do processo de investigação 155
1.3. Consulta aos SIE 157
1.4. Avaliação dos dados e informação disponibilizada pelo SIE 157
1.5. Pedidos, sugestões e definição de requisitos 159
1.6. Revisão do processo de investigação 160
1.7. Extracção dos dados 160
1.8. Outras actividades 160
2. O Sistema Estatístico Nacional 161
2.1. Composição do Sistema Estatístico Nacional 161
2.2. Princípios orientadores do actual Sistema Estatístico Nacional 163
2.3. Difusão da Informação Estatística 164
3. Os Sistemas de Informação para a Economia e os factos económicos de que se constituem missão 165
3.1. Os SIE de âmbito alargado 165
3.2. Os SIE de âmbito restrito 172
3.3. A Matriz SIE / Factos Económicos 203
4. Da heterogeneidade dos utilizadores à heterogeneidade da informação dos SIE .. 205
4.1. Importância dos SIE para a análise do ambiente 207
4.2. Os SIE e as decisões estratégicas sobre o produto 208
4.3. As dificuldades da análise do grupo competitivo 211
4.4. A viabilização da análise do consumidor 212
5. A questão da disponibilização de informação incompleta em Economia 213
6. Avaliação dos sítios da Administração Directa e Indirecta e sua relação com a maturidade dos SIE 214
Resumo do Capítulo III 220
Capítulo IV – Aplicações de Processamento Analítico suportadas por Sistemas de Informação para a Economia em Portugal 223
1. Processos de interrogação em informação demográfica e social 224
1.1. Modelação 225
1.2. Processos analíticos suportados 226
1.3. Aspectos criticáveis na informação demográfica e social providenciada pelo INE em termos de unidades territoriais 230
2. Processos de análise multidimensional 230
2.1. Um modelo da análise multidimensional para o IDE 230
2.2. Um modelo da análise multidimensional para o comércio externo português 243
3. Processos híbridos 262
3.1. Retrato de unidades territoriais 263

3.2. Cobrança de imposto sobre o rendimento 266
4. Processos de pesquisa de associações 269
4.1. A pesquisa de associações no âmbito da análise do IDE 270
4.2. Resultados da pesquisa 271
5. Breve reflexão sobre a viabilidade do processamento analítico suportado pelos SIE 276
Resumo do Capítulo IV 278
Capítulo V – Métricas para análise da maturidade dos sistemas de informação para a Economia 281
1. A Maturidade dos Sistemas de Informação no contexto da Análise da Economia 282
1.1. Primeira abordagem à maturidade dos SI 283
1.2 - A maturidade dos SI como resultado de uma analogia com o neurónio de McCulloch e Pitts 283
1.3. A maturidade no tempo 286
1.4. A maturidade no contexto da análise da Economia 291
2. Uma base de dados relativa à maturidade dos Sistemas de Informação para a Economia 297
2.1 Modelo conceptual 298
2.2. Modelo lógico 299
2.3. Implementação 300
2.5. Do valor dos indicadores ao índice de maturidade 301
2.4. Indicadores relativos à disponibilização interactiva de informação pelos SIE 303
3. A maturidade dos SIE numa perspectiva difusa 307
3.1. Obtenção de árvore de decisão relativa à maturidade dos SIE 308
3.2. A classificação de maturidade dos SIE numa perspectiva difusa 310
3.3. Valor do melhor compromisso para avaliação da maturidade 315
3.4. A maturidade dos SIE no ano 2000 com base no modelo difuso desenvolvido para 1999 316
3.5. Avaliação da correcção do modelo 319
3.6. Proposta de um processo de análise de maturidade dos SIE em perspectiva difusa 321
Resumo do Capítulo V 324
Capítulo VI - Estádios de desenvolvimento dos Sistemas de Informação para a Economia na era das redes 327
1. Identificação da dimensões de análise da maturidade dos SIE 328
1.1. Adequação da Análise Factorial 328
1.2. Dimensões de análise 335
2. Identificação dos segmentos de informação para a Economia 340
3. Diferenciação dos segmentos de informação para a Economia 345

4. Hipótese sobre os estádios de desenvolvimento dos Sistemas de Informação para a Economia 350
5. Será a abertura ao analista uma utopia? 353
5.1. A suavização das diferenças entre os SIE nos anos de 2000 e 2001 354
5.2. A associação dos SIE aos clusters através de uma análise discriminante.. 356
5.3. A necessidade de um «corte epistemológico» 363
5.4. Conclusão sobre os estádios de desenvolvimento dos SIE na era das redes 368
6. Contribuição para o desenvolvimento do estádio de *abertura ao analista* 370
Resumo do Capítulo VI 373
Conclusões 377
Da teoria 377
Da consciência 379
Da contradição 380
Da prática 381
Da descoberta de conhecimento 383
Do avanço e do retrocesso 384
Da predição 385
Dos estádios 387
Da destruição criadora 388
Referências 389

Lista de Quadros

Quadro 1. Tópicos relevantes por autor 26
Quadro 2. Equilíbrio Óptimo Controlo-Autonomia em Nolan 44
Quadro 3. Investimento Directo Estrangeiro em Portugal 92
Quadro 4.Tabelas Obtidas por Partição da Tabela Original 93
Quadro 5. Tabelas Obtidas por Partição da Tabela Intermédia 93
Quadro 6. Simulação da determinação da entropia 98
Quadro 7. Diferenciação de métodos de partição 99
Quadro 8. Transacções do Período 110
Quadro 9. Produtos Envolvidos em Transacções 110
Quadro 10. Frequência dos Produtos - 1 111
Quadro 11. Frequência dos Produtos - 2 111
Quadro 12. Quadro de Regras Potenciais - 1 111
Quadro 13. Frequência dos Produtos - 3 112
Quadro 14.Quadro de Regras Potenciais - 2 112
Quadro 15. Quadro de Regras Potenciais - 3 113
Quadro 16. O algoritmo Apriori 115
Quadro 17. Comparação Apriori / SIMILIS 116
Quadro 18. Medidas de distância mais utilizadas 119
Quadro 19. Coeficientes de Associação 120
Quadro 20. Outros Coeficientes de Semelhança 121
Quadro 21. Resumo comparativo dos critérios de aglomeração 124
Quadro 22. Metadados de apoio à pesquisa e navegação 146
Quadro 23. Metadados de apoio à interpretação 147
Quadro 24. Descrição dos vários tipos de formatos 149
Quadro 25. Tratamento de documentos por formatos baseados em ficheiros de linguagem natural 152
Quadro 26. Séries Longas para a Economia Portuguesa pós II Guerra Mundial 171
Quadro 27. Algumas publicações disponibilizadas pelo DPP 193
Quadro 28. Estatísticas disponibilizadas pelo GEPE 195
Quadro 29. O documento Economia Portuguesa do DGEP 198
Quadro 30. Matriz SIE / Factos Económicos 204
Quadro 31. Os SIE considerados 203
Quadro 32. A implementação da sociedade da informação em Portugal 216
Quadro 33. Os termos da declaração para a disponibilização interactiva 218
Quadro 34. Resumo da avaliação dos sítios da Administração Pública 220
Quadro 35. Roteiro de Análise para o IDE 235
Quadro 36. Actividades económicas segundo a hierarquia GEPE 251
Quadro 37. Utilização da Área Territorial 264
Quadro 38. Elementos sobre os Dados Utilizados 271
Quadro 39. Frequências de Acções de Investimento por Sector 272

Quadro 40. Frequências de Acções de Investimento por Combinação de Sectores ... 272
Quadro 41. Frequências de Acções de Investimento e Desinvestimento por Sector .. 273
Quadro 42. Combinação de Acções de Investimento e Desinvestimento 273
Quadro 43. Associações entre Investidores 274
Quadro 44. Associações entre Empresas..... 275
Quadro 45. Valores para os Ponderadores no Tempo 288
Quadro 46. Valores para as variáveis de análise e índice de maturidade 289
Quadro 47. Valores para as variáveis de análise e índice de maturidade 290
Quadro 48. Caracterização de Maturidade dos SI 296
Quadro 49. Maturidade dos SIE (1999-2001)..... 304
Quadro 50.Ponderadores utilizados 305
Quadro 51. A maturidade dos SIE por critério de análise (1999-2001) 306
Quadro 52. Regras Relativas à Maturidade dos SIE 310
Quadro 53. Valores Típicos dos Termos 311
Quadro 54. Definição dos Termos 312
Quadro 55. Dados Relativos aos SIE..... 312
Quadro 56. Explicitação *Difusa* de Resultados..... 313
Quadro 57. Explicitação dos Resultados *Melhor Compromisso*..... 314
Quadro 58. Explicitação dos Resultados *Solução Mais Plausível*..... 315
Quadro 59. Explicitação dos Resultados Melhor Compromisso 316
Quadro 60. Predição Difusa de Maturidade dos SIE em 2000 317
Quadro 61. A Maturidade dos SIE em 2000 de acordo com Métrica e Modelo..... 318
Quadro 62. Resultados de teste de diferença de médias..... 319
Quadro 63. Frequências relativas a Maturidade e Predição..... 319
Quadro 64. Matriz de Acertos e Erros para a Classe 1 320
Quadro 65. Matriz de Acertos e Erros para a Classe 2 320
Quadro 66. Matriz de Acertos e Erros para a Classe 3 320
Quadro 67. Matriz de Acertos e Erros para a Classe 4 321
Quadro 68. Matriz de Acertos e Erros para a Classe 5 321
Quadro 69. Testes de Normalidade 328
Quadro 70. Estatísticas Descritivas 329
Quadro 71. Matriz de Correlação 331
Quadro 72. KMO e Teste de Bartlett 332
Quadro 73. Matriz Anti-imagem..... 333
Quadro 74. Comunalidades 334
Quadro 75. Variância Total Explicada..... 335
Quadro 76. Matriz de pesos dos factores..... 336
Quadro 77. Matriz de pesos dos factores após rotação 337
Quadro 78. Processo de Aglomeração 342
Quadro 79. Estatísticas Descritivas para os Clusters..... 343
Quadro 80. Expressão dos SIE nos Factores Retidos 345

Quadro 81. Testes de Normalidade para os Factores 346
Quadro 82. Teste de Homogeneidade de Variâncias 346
Quadro 83. Resultados do ANOVA 346
Quadro 84. Multiplas Comparações de Bonferroni 347
Quadro 85. Descrição dos Clusters nos Factores Retidos 348
Quadro 86. Resumo de Análise 349
Quadro 87. Teste de homogeneidade de variâncias 357
Quadro 88. Variáveis excluídas da análise 358
Quadro 89. Sumário da análise discriminante 359
Quadro 90. Poder discriminante das funções discriminantes 359
Quadro 91. Centróides dos grupos 360
Quadro 92. Coeficientes das funções de classificação 360
Quadro 93. Transição entre grupos com base nas equações classificatórias 361
Quadro 94. Resumo da Transição 361
Quadro 95. Evolução da distribuição por estádios 362
Quadro 96. Novos coeficientes das funções de classificação 363
Quadro 97. Caracterização dos estádios em termos de variáveis de análise 369
Quadro 98. Caracterização do valor informacional de um SIE 373

Lista de Figuras

Figura 1. Circuito económico em termos conceptuais 11
Figura 2. Lógica de desenvolvimento da dissertação 18
Figura 3. Evolução das Despesas em TI ao longo dos Estádios de Nolan 27
Figura 4. As Múltiplas Curvas de Aprendizagem de Earl 29
Figura 5. As trocas de dados na Economia 53
Figura 6. O processo da descoberta de conhecimento 58
Figura 7. Uma visão conjunta das contribuições para a maturidade dos SI 60
Figura 8. Diagrama Estrela 73
Figura 9. Diagrama Floco de Neve 74
Figura 10. Operadores da Álgebra Relacional 81
Figura 11. A Rotação em Bases de Dados Multidimensionais 83
Figura 12. O *Slice and Dice* em Bases de Dados Multidimensionais 84
Figura 13. O *Drill-Down* em Bases de Dados Multidimensionais 86
Figura 14. Árvore de decisão 91
Figura 15. Árvore de Decisão Gerada por *Query* 94
Figura 16. Função de Pertença do conjunto difuso entre 5 e 8 102
Figura 17. Função de Pertença do número difuso 4 103
Figura 18. Função de Pertença da União 103
Figura 19. Função de Pertença da Intersecção. 104
Figura 20. Algoritmo geral para a pesquisa de associações 114
Figura 21. Concretização de uma análise factorial 126
Figura 22. Concretização de uma análise discriminante 132
Figura 23.O processo de disponibilização e utilização da informação sobre a Economia 140
Figura 24. A espiral da agonia da economia agrária 172
Figura 25. Quatro décadas da evolução do preço da gasolina super 174
Figura 26. Estrutura do SISED 176
Figura 27. Diagrama de cruzamentos de informação de Emprego 177
Figura 28. Diagrama de cruzamentos de informação de IRCT 178
Figura 29. Formas diferentes do ciclo de vida 209
Figura 30. Modelação da informação demográfica e social 225
Figura 31. A associação poder de compra vs taxa de divórcio 228
Figura 32. Dendrograma relativo a indicadores demográficos e sociais por NUTs ... 229
Figura 33. Representação Genérica do MAMIDE 238
Figura 34. Representação mais Detalhada do MAMIDE 238
Figura 35. Um modelo conceptual multidimensional para o Comércio Internacional254
Figura 36. Evolução global do *grau de abertura* no período 1980-1999 258
Figura 37. A evolução do *grau de abertura* por grande espaço 259
Figura 38. Evolução do *grau de abertura* por grande espaço e país 260
Figura 39. Modelo Conceptual do Retrato por Unidade Territorial 263

Figura 40. Hierarquia Presente em Retrato por Unidade Territorial 263
Figura 41. Utilização da àrea por Unidade Territorial 265
Figura 42. Utilização da Área por tipo 265
Figura 43. Modelo Conceptual da Cobrança de Imposto sobre Rendimento 267
Figura 44. Diagrama Estrela relativo à Cobrança de Imposto sobre Rendimento 267
Figura 45. Concentração do Rendimento e do Imposto (1994) 268
Figura 46. Modelo para Análise da Maturidade dos SI 284
Figura 47. Função limiar com delimitador fixo 285
Figura 48. Variantes da função limiar linear 285
Figura 49. Variantes da função logística 286
Figura 50. Função de Maturidade no Tempo sem Modificações das Observações 289
Figura 51. Função de Maturidade do SI mais realista 291
Figura 52. Modelo Entidade Associação relativo à Maturidade dos SIE 298
Figura 53. Esquema de determinação dos índices de maturidade 301
Figura 54. Árvore de Decisão relativa à Maturidade dos SIE 309
Figura 55. Modelo Conceptual do Sistema Difuso de maturidade dos SIE 311
Figura 56. Caixas de Bigodes Relativas a Variáveis de Análise 330
Figura 57. Scree Plot 336
Figura 58. Gráfico das Componentes após Rotação 338
Figura 59. Os SIE nos Factores Retidos (1) 339
Figura 60. Os SIE nos Factores Retidos (2) 339
Figura 61. Os SIE nos Factores Retidos (3) 340
Figura 62. Dendrograma 343
Figura 63. Descrição Gráfica dos Clusters 344
Figura 64. Os *clusters* em termos de factores 349
Figura 65. Representação de Estádios de Desenvolvimento dos SIE 352
Figura 66. Maturidade dos SIE no período 1999-2001 354
Figura 67. Maturidade dos SIE no período 1999-2001 no ambiente de disponibilização 355
Figura 68. Caixas de bigodes com anterior estrutura de clusters 356
Figura 69. Evolução de um SIE em estádios 368

Lista de Expressões

Expressão 1. Especificação de uma relação 71
Expressão 2. Função de Entropia 96
Expressão 3. Função de Gini 96
Expressão 4. Desenvolvimento da Função de Gini 96
Expressão 5. Função de Ganho 97
Expressão 6. Função de Pertença 101
Expressão 7. Definição Trivial de um Conjunto 101
Expressão 8. Alternativa à definição trivial de um conjunto 101
Expressão 9. Definição trivial de um conjunto representado em intervalo contínuo. 101
Expressão 10. Definição de Conjunto Difuso 102
Expressão 11. Alternativa à Definição de Conjunto Difuso 102
Expressão 12. Alternativa à Definição de Conjunto Difuso para Universo Contínuo 102
Expressão 13. União de Conjuntos Difusos 103
Expressão 14. Alternativa à Definição de União de Conjuntos Difusos 103
Expressão 15. Intersecção de Conjuntos Difusos 103
Expressão 16. Alternativa à Definição da Intersecção de Conjuntos Difusos 103
Expressão 17. Negação de um Conjunto Difuso 104
Expressão 18. Forma Paramétrica do Operador MinMax 105
Expressão 19. Forma Paramétrica do Operador MinAvg 105
Expressão 20. Forma Paramétrica do Operador Gamma 105
Expressão 21. Propriedades das Métricas - Simetria 118
Expressão 22. Propriedades das Métricas – Desigualdade Triangular 118
Expressão 23. Propriedades das Métricas – Diferenciabilidade de não idênticos 118
Expressão 24. Propriedades das Métricas - Indiferenciabilidade de idênticos 118
Expressão 25. Distância Euclidiana 119
Expressão 26. Quadrado da Distência Euclidiana 119
Expressão 27. Distância absoluta 119
Expressão 28. Distância de Minkowski 119
Expressão 29. Distância de Mahalanobis 119
Expressão 30. Distância de Chebishev 119
Expressão 31. Coeficientes de Emparelhamento - Semelhança 120
Expressão 32. Coeficientes de Emparelhamento - Distância 120
Expressão 33. Coeficientes de Jaccard - Semelhança 120
Expressão 34. Coeficientes de Jaccard - Distância 120
Expressão 35. Coeficiente de Gower 121
Expressão 36. Outros Coeficientes de Semelhança: Peso duplo às presenças e ausências simultâneas 121
Expressão 37. Outros Coeficientes de Semelhança: Peso duplo às situações discordantes, inclusão das ausências simultâneas 121

Expressão 38. Outros Coeficientes de Semelhança: Peso duplo às presenças simultâneas, exclusão das ausências simultâneas 121
Expressão 39. Outros Coeficientes de Semelhança: Peso duplo às situações discordantes, exclusão das ausências simultâneas 121
Expressão 40. Outros Coeficientes de Semelhança: Quociente entre presenças simultâneas e situações discordantes, exclusão das ausências simultâneas.... 121
Expressão 41. Distância entre dois Grupos – vizinho mais próximo 122
Expressão 42. Distância entre dois Grupos – vizinho mais afastado..... 123
Expressão 43. Modelo factorial 125
Expressão 44. Lambda de Wilks 133
Expressão 45. Função discriminante para dois grupos 133
Expressão 46. Função de classificação 134
Expressão 47. BCG - Quota de Mercado Relativa..... 210
Expressão 48. Medida da disponibilização de informação incompleta 214
Expressão 49. Ajustamento por Regressão..... 228
Expressão 50. Grau de Cobertura das Importações pelas Exportações..... 248
Expressão 51. Grau de Abertura ao Exterior..... 248
Expressão 52. Peso do Défice Externo no PIB 248
Expressão 53. Taxa de Cobertura a nível de Produto 249
Expressão 54. Taxa de Penetração das Importações..... 249
Expressão 55. Taxa de Penetração das Importações a nível de Produto..... 249
Expressão 56. Condições de Entrega - Valor FOB 250
Expressão 57. Condições de Entrega - Valor CIF..... 250
Expressão 58. Valor do Processo de Maturidade 283
Expressão 59. Maturidade - Vector de Distâncias 283
Expressão 60. Maturidade - Distância Global 283
Expressão 61. Índice de Maturidade 284
Expressão 62. Valor da Maturidade 284
Expressão 63. Função Geradora de Ponderadores 287
Expressão 64. Índice de Maturidade no Tempo 287
Expressão 65. Determinação dos Índices de Maturidade - 1..... 302
Expressão 66. Determinação dos Índices de Maturidade - 2..... 302
Expressão 67. Índice de Desenvolvimento da Maturidade 371
Expressão 68. Valor Informacional em Determinado Momento..... 371
Expressão 69. Valor Informacional de um SIE num Momento..... 372
Expressão 70. Restrição Relativa aos Ponderadores 372
Expressão 71. O Valor Informacional em Especificação Simplificada..... 373
Expressão 72. Redefinição da Restrição Associada aos Ponderadores 373

Agradecimentos

Ao atingir o final deste longo período dedicado aos estádios de desenvolvimento dos Sistemas de informação para a Economia, constato que tive a colaboração de muitas pessoas sem a qual seria impossível este resultado. Pedindo desculpa por algum involuntário esquecimento, aqui deixo registado o meu reconhecimento.

À Dra. Rita Araújo e à administração do ICEP pela disponibilização dos dados desta organização relativos ao Investimento Directo Estrangeiro em Portugal no ano de 1996 que foram utilizados na ilustração de um processo analítico de pesquisa de associações.

Ao Dr. Carlos Costa pelos esclarecimentos relativos à utilização da base de dados de comércio internacional do GEPE e pela disponibilização de um manual de notas metodológicas relativas ao mesmo. A sua contribuição foi muito útil para a apresentação de um processo de análise multidimensional suportado num Sistema de Informação para a Economia em Portugal.

Ao Dr. João Manuel Santos da Direcção de Informação do ICEP pela disponibilização de um *login* e *password* para um acesso temporário ao produto EconomiaNet tendo por objectivo a sua apreciação crítica.

À Dra Elsa Farto e ao SAS Institute em Portugal por terem proporcionado a participação na conferência de utilizadores do SAS, CAUS2001, o que tornou possível o enriquecimento da explicitação relativa à pesquisa de associações através da inclusão de alguma pesquisa efectuada em Portugal.

A Ana Maldonado do Núcleo de Difusão Estatística do Banco de Portugal por ter disponibilizado dados relativos ao IDE no período 1996-99.

Ao Dr. Fernando Naves por ter proporcionado a utilização do produto Media que possibilitou a apresentação de um *Modelo de Análise Multidimensional para o IDE.* Os resultados fazem parte dos processos analíticos integrados na dissertação e o documento completo é integrado nos anexos à dissertação, tendo sido sujeito a apresentação ao Seminário do Departamento de Economia do ISEG em Janeiro de 2000.

Ao Dr. Pedro Miguel dos Reis Filipe que validou aspectos de coerência e correcção ortográfica em associação a disponibilizar-se para testar a clareza de exposição e, aí, propor um sem número de melhorias. O Dr. Pedro Filipe é um jovem investigador na área de História que, neste momento, frequenta o Mestrado de História e Cultura Pré-Clássica na Faculdade de Letras da Universidade de Lisboa.

Ao Professor Doutor António Luís Silvestre pelo seu incomensurável apoio ao nível de conceitos, aplicações e testes da área de Estatística Multivariada sem esquecer muitos aspectos de validação ortográfica que decorreram daquela actividade essencial.

Uma palavra final de muito apreço para os orientadores, Professor Doutor Manuel de Jesus Farto e Professor Doutor Ilídio Rodrigues Antunes. Foi o seu apoio, o seu acompanhamento, o seu empenho que tornou possível elaborar, validar e concluir esta dissertação.

Luís M. Ribeiro Vieira

Glossário

Estádio de Desenvolvimento. Conjunto coerente de sinais relativos à concretização em termos de maturidade de um objecto de estudo, por exemplo, um Sistema de Informação para a Economia. Neste documento, descobriremos quatro estádios fundamentais: *Disponibilização Interactiva de Dados e Publicações, Procura de Actualidade, Procura de Integridade Temporal* e *Abertura ao Analista*

Internet. Rede de redes ligando computadores a computadores que partilham protocolos *TCP/IP (Transmission Control Protocol/Internet Protocol)* cada um dos quais corre software para fornecer informação e/ou para fazer acesso ou ver informação, utilizada como mecanismo para a sua disseminação e para a colaboração e interacção de pessoas e dos seus computadores sem preocupação com a localização geográfica.

Maturidade. Fase suprema de desenvolvimento na concretização de um conceito. Em termos de estádio de desenvolvimento dos sistemas de informação para a Economia corresponderá ao estádio de *Abertura ao Analista.*

Métrica. Forma de valorização da concretização de um conceito que se assemelha a medida se a esta atribuirmos um sentido metodológico e de processo.

Processamento Analítico. Conjunto de técnicas de análise de dados, suportadas em computadores, que visam a sua exploração numa perspectiva multidimensional e de descoberta de conhecimento que pode ser expresso sob as mais variadas formas de que são exemplo uma árvore de decisão, uma regra de produção, uma rede bayesiana, uma equação de regressão, uma associação, um cluster, etc.

Sistema de Informação para a Economia. Organização cuja função essencial é a disponibilização de informação para os agentes económicos, designadamente, para os analistas da Economia nas empresas, no Estado, nas famílias e no exterior, actividade que pode desempenhar com autonomia técnica e de gestão, utilizando meios organizados de acordo com conceitos, metodologias e técnicas herdados das ciências da computação e da informação.

Introdução

Nas sociedades actuais, onde o desenvolvimento económico é acompanhado e viabilizado pela expansão da utilização de sistemas computacionais, a informação, transformada num dos seus recursos essenciais, esconde-se sob a forma de uma imensa acumulação de dígitos binários. Esse recurso, para ser usado, carece de processos de pesquisa, selecção, avaliação, processamento e distribuição.

À medida que cresce a participação da informação no processo de produção dos diversos bens, as organizações integram-na no conjunto das suas preocupações estratégicas. Elas sabem que o sucesso está associado à qualidade e quantidade da informação que envolve a sua actividade.

Esta tomada de consciência, pelas organizações, da sua dependência relativamente ao recurso informação, levou ao surgimento de novas profissões, novas actividades e ao desenvolvimento de todo um conjunto de meios que, eles próprios, se constituíram num subsector de actividade económica.

Aos analistas da informação é exigido, cada vez mais, um conhecimento profundo de matérias relacionadas com a sua área de actividade e, por outro lado, com a pesquisa, selecção, avaliação e processamento de dados.

Reflexo disso, o ensino de Economia e Gestão, áreas de actividade impregnadas de analistas da informação, integra novas preocupações associadas à estruturação dos dados que vão ser utilizados como matéria prima para uma análise exploratória ou para processos mais complexos.

Para além disso, autores recentes afirmam que, escondidos no ambiente de dados, se encontram filões informacionais que, para serem descobertos em ordem a se tirar partido deles, carecem de ser pesquisados com ferramentas próprias para a sua detecção.

Curiosamente, dados essenciais ao sucesso das organizações são os que as relacionam com o seu ambiente e, como tal, não são dependentes directamente da sua actividade. Esses dados, a existirem, têm de ser fornecidos por organizações externas onde os Sistemas de Informação para a Economia (SIE) desempenham um papel de relevo, para além de outras

organizações que, submetida às relações de produção dominantes, vivem da produção da informação.

Tal como aconteceu com os sistemas de informação internos às organizações, também os SIE têm tido uma capacidade para dominar a sua oferta, isto é, eles podem não produzir a informação que é necessária ao funcionamento para a economia no seu todo e justificar a sua existência pela satisfação de um ou outro requisito informacional que algum utilizador melhor colocado relativamente ao poder de decisão lhes encomende.

Na realidade, os utilizadores dos SIE manifestam elevada *heterogeneidade*, podendo ser listados quatro tipos de utilizadores: governo, organizações, analistas da economia e estudantes. Reconhece-se, com facilidade, que um destes grupos pode instrumentalizar os SIE e dar-lhes um carácter que não seja o que mais interessa numa perspectiva global da sociedade.

O reconhecimento da criticidade do recurso informação permite, ainda, que os que detêm o poder no interior dos SIE possam favorecer uma ou outra organização na sua luta pela sobrevivência. A problemática desta dissertação prende-se com esta questão: *como garantir que a actividade dos SIE é neutra e que contribui globalmente para o sucesso das organizações em vez de favorecer algumas em detrimento de outras?*

A informação é um recurso como outros que a nossa sociedade já satisfaz com relativa eficiência: a água, captada em múltiplos reservatórios dignos de confiança e distribuída através de uma rede que envolve todo o país, e a energia eléctrica que carece de um processo de produção com tecnicidade elevada e precisa igualmente de uma rede pública. Tal como estes e outros produtos da actividade económica, a informação necessita de um processo de produção e de um processo de distribuição.

Onde a informação se distingue de outros recursos é numa característica notável do seu processo de produção: a sua matéria prima, os dados, não desaparece no produto final, podendo ser utilizada noutros momentos em nova produção. Por outro lado, curiosamente, a informação pode ser matéria prima para o processo de produção da informação. Assim, dados e informação podem gerar mais informação, constituindo uma espiral que se opõe aos rendimentos decrescentes.

A informação sobre a Economia tem muito de específico e muito a ver com a melhor e mais conhecida informação de gestão. Esta relação pode ser estabelecida ao nível das suas características e dos problemas que enfrenta.

Entre as suas características específicas reconhecemos, quase de imediato, uma *maior orientação para os factos*. A informação para a economia tem por objecto elementos de análise da realidade económica onde, no mundo actual, o desemprego, a distribuição do rendimento, o comércio internacional, a contabilidade nacional, o investimento estrangeiro, os recursos naturais, constituem instâncias importantes.

Em segundo lugar, as variáveis que nos interessam em Economia são respeitantes a *grandes agregados*, aproximando-se assim da chamada informação de apoio à gestão estratégica das organizações, sendo por isso detentoras de um maior grau de agregação.

Em terceiro lugar, há que fazer ressaltar a sua *externalidade*: a informação sobre a economia não interessa à entidade que a produz, mas a departamentos governamentais, a unidades económicas e outras organizações e a analistas externos.

A última característica, não a de menor valor, é a que se refere à *elevada importância da componente cronológica*: para além das relações entre variáveis, é importante que os diferentes factos sejam vistos em termos de evolução no tempo e sujeitos a análise de tendências no sentido de, em termos de política, lhes poder ser definida uma resposta mais racional.

De entre os problemas que a informação sobre a Economia hoje enfrenta não podemos deixar de citar, em primeiro lugar, uma elevada *redundância*: embora não sendo uma característica natural, a redundância, traduzida na proliferação de cálculo da mesma variável em múltiplos departamentos de informação para a Economia é, hoje, uma realidade que deve ser tida em conta. Ela tem como consequência, pelo menos, alguma confusão entre aqueles que procuram analisar a realidade económica.

Em segundo lugar, são inúmeras as *dificuldades para chegar aos dados* que o analista económico enfrenta: são relativamente conhecidas as críticas que se

levantam à burocracia, aos estreitos limites de divulgação e até às redes interpessoais de difusão.

A *disponibilização de informação incompleta* é uma prática generalizada que se manifesta na agregação de dados com perda da base que a possibilitou, sendo possível, por exemplo, no que respeita a três variáveis conhecer o valor da sua intersecção duas a duas e não ser possível chegar à base que teoricamente deveria ter providenciado as agregações correspondentes.

É vulgar a *incoerência entre fontes oficiais*, traduzindo-se em séries cujos elementos constituintes tiveram na base critérios de obtenção distintos quer entre departamentos diferentes quer entre várias publicações do mesmo departamento.

A *instabilidade das estatísticas*, que faz com que as mais básicas séries mudem várias vezes em curto espaço de tempo, é outro problema que o analista da Economia deve enfrentar.

Por fim, o enquadramento geral do sistema estatístico nacional não parece satisfatório para alguns autores: *a autonomia do INE não está suficientemente cuidada, sendo criticável a delegação de competências suas em departamentos ministeriais* (Barreto, 1996).

Apesar dos significativos investimentos residentes ao nível do sistema estatístico nacional materializados em computadores, *software*, recursos humanos e actividades de consultores, este conjunto de problemas leva a que, no momento crucial em que se procura analisar a evolução de uma variável relativamente simples, surjam múltiplos problemas, obstáculos e inexactidões.

Que tem a informação para a Economia de especial para não deter a resposta hoje já conseguida para a informação gerada internamente com que as organizações funcionam?

Nesta dissertação vamos demonstrar que estas questões não estão devidamente tratadas porque os sistemas de informação (SI) não estam maduros para que esse processo tenha uma concretização eficaz.

Acreditamos que os grandes problemas dos SI têm a ver com a maturidade das organizações para os enfrentar. Tal como as organizações tiveram o seu

Hammer, o seu Davenport e o seu Inmon[1], a informação para a Economia precisa de alguém que a vire para os factos, para os analistas e para o tempo.

O tema desta dissertação prende-se com a maturidade dos SIE e, como se referiu, está indissoluvelmente ligado à potencialidade de sucesso das organizações. Assume um carácter eminentemente transversal ao relacionar questões da área de Sistemas de Informação com questões da área de Gestão e procurar uma aplicação no campo da Economia.

1. Problemática da maturidade dos sistemas de informação e sua referência à área de economia

Nesta secção, explicitamos a problemática que esta dissertação visa tratar. Num primeiro momento, vamos apresentar aspectos do seu enquadramento teórico que se concretizam nos dados do problema.

Na nossa perspectiva, uma multiplicidade de elementos legitima as preocupações relativas à maturidade dos sistemas de informação e, em particular, dos que se destinam a produzir informação sobre a economia.

O primeiro desses elementos é a existência de uma teoria de estádios de desenvolvimento dos sistemas de informação que, desde a década de setenta, vem resistindo e mantendo alguma capacidade explicativa sobre o nosso objecto de estudo. Um elemento importante associado a essa teoria é a existência de reformulações provenientes de um dos que mais contribuíram para o seu desenvolvimento e que se materializaram na consideração de uma evolução em termos de eras, sendo a actual denominada como *era das redes*.

O desenvolvimento de *bases de dados analíticas* suportadas nas recomendações de Inmon (1994) e materializando-se em novos repositórios capazes de suportarem interrogações extremamente complexas é outro traço que deve ser considerado. Apesar de estas bases de dados possibilitarem um acesso mais fácil à informação e permitirem outros níveis de análise, não deixa de ser importante determinar como se materializam os *guias para a informação* primária e, neste contexto, recomendações como as de Davenport (1994) trazem novos elementos, possibilitando o que Codd et al (1993)

[1] Estas referências a nomes associados ao processo de transformação estrutural das organizações na década de 90 serão posteriormente decifradas, mantendo-se esta formulação para o texto exibir um estilo que noutra pespectiva lhe seria retirado.

designou por *processamento analítico interactivo*. Finalmente, é indispensável uma referência ao papel dos *dados externos* contribuição essencial inserta em Bhahuta (1988), um autor associado à teoria de estádios.

Após a apresentação deste conjunto de elementos que se traduzem em dados para o nosso estudo, explicitamos, num segundo momento, os problemas que, pensamos, estão por resolver e vão ser objecto de uma resposta pela dissertação. Dispomos de uma teoria de estádios, mas será ela satisfatória? E, já que nos situamos na área de Economia e na era das redes, poderemos avançar com alguma hipótese sobre a evolução dos sistemas de informação para a Economia na era das redes?

Em terceiro lugar, definimos o objecto da nossa análise. Falamos de Economia, falamos de organizações, falamos de processos analíticos, mas os verdadeiros elementos que se constituem em objecto de análise são as organizações que disponibilizam informação para a Economia. Neste contexto, é necessário explicitá-las.

Sendo a problemática desta dissertação a maturidade dos sistemas de informação que disponibilizam informação para a economia, a sua observação carece de passar por alguma explicitação. Por essa razão vamos nomear as áreas de análise que nos vão servir para apoiar as nossas considerações e julgar sobre a maturidade dos sistemas de informação seleccionados para análise.

Finalmente definimos o produto da dissertação indissoluvelmente ligado ao segundo momento, isto é aos problemas de maturidade dos SIE.

1.1. Dados do problema

Consideramos que a maturidade dos SI está relacionada com o seguinte conjunto de elementos:

- a teoria de estádios de Nolan (1978) e a sua transfiguração na era das redes descrita em Nolan (1995);

- a concepção de sistemas de informação centrados nas pessoas derivada de Davenport (1994)

- o processamento analítico interactivo descrito Codd et al (1993);
- as bases de dados analíticas derivadas de múltiplos trabalhos de Inmon, designadamente, Inmon e Hackathorn (1994).

1.1.1. Nolan

Existirá um modelo para o desenvolvimento dos sistemas de informação? Nas últimas décadas, esta questão tem vindo a ser analisada e para alguns autores a resposta é afirmativa com a particularidade de o culminar deste processo ser, em princípio, uma fase de maturidade.

Como demonstrámos em Vieira (1997), o processo de maturação das organizações em termos de Sistema de Informação (SI) pode ser analisado utilizando o conceito de estádio de desenvolvimento. Este conceito é útil por fornecer um meio para avaliar o momento actual do sistema de informação e, nesse contexto, para apoiar acções que conduzam à sua transformação.

Nolan (1995) apresenta uma visão do desenvolvimento da utilização de computadores nas organizações assente em três eras: a era do processamento de dados, a era micro e a era das redes.

A era do processamento de dados, que considerou dominante entre 1960 e 1980, respeita à utilização de sistemas de processamento transaccional que automatizaram os sistemas manuais anteriormente existentes. Poder-se-á dizer que a teoria de estádios, tal como surgiu, teria uma aplicação directa nesta era.

A era micro, associada pelo autor ao período entre 1980 e 1995, potenciou a capacidade de profissionais como engenheiros, analistas financeiros, gestores que utilizam o computador para computações analíticas, acesso a dados e impressão de desenhos ou gráficos. Em ambas estas eras, a utilização de computadores foi primariamente no interior da organização.

Em sobreposição à era micro existe um mercado externo emergente em que o computador é integrado em produtos (automóveis) e serviços (cartões inteligentes).

A era das redes, iniciada em 1990 e sem se antever o seu termo, que correntemente se sobrepõe à era micro, emergiu a partir da fusão de computadores e telecomunicações. Nesta era, as firmas estão a integrar computadores internos e externos de modo a poderem modificar a sua estrutura para terem formas eficiente de competição flexível no mercado global. Esta integração coloca pressões extraordinárias às empresas que necessitam de manter os serviços antigos em processamento ao mesmo tempo que estão a desenvolver interconexões e serviços necessários para o novo ambiente.

É na era das redes que devemos incluir a integração das organizações com os sistemas de informação para a economia. Esta era só surge em toda a sua plenitude quando as trocas de informação, no fim de contas, o processo de integração de dados externos, utilizarem todas as suas potencialidades. Neste contexto, podemos assumir que a posição dos SIE perante este processo poderá ter diferenças significativas e, se aceitarmos os estádios de Nolan para a era de processamento de dados, poderemos procurar outros estádios numa realidade diferente. Assim, devemos concluir que o caminho iniciado por Nolan se mantém e, hoje em dia, constitui uma força importante para a explicação da realidade.

Mas os dados do problema não se esgotam nas concepções de Richard Nolan.

1.1.2. Davenport

Um artigo sobre uma concepção de sistemas de informação centrados nas pessoas, Davenport (1994), critica aspectos da visão actual dominante a nível das tecnologias de informação e do planeamento de sistemas de informação chamando a atenção para a necessidade de esta manter múltiplos significados para os seus utilizadores, de ser conveniente a distinção entre o global e o particular e a existência na organização de pessoas com a função de guias para a informação.

Para o autor, (1) a informação desenvolve-se gradual e naturalmente em múltiplas direcções, tomando múltiplos significados, (2) as pessoas não partilham facilmente a informação e (3) a mudança a nível de tecnologias de informação só por si não transformará a cultura de informação na empresa.

O conceito de arquitectura da informação - proveniente de metodologias como o BSP (Business Systems Planning) destinadas a planeamento estratégico dos Sistemas da Informação - terá eclipsado uma visão da informação centrada nas pessoas portadora de múltiplos factores de distinção.

Os defensores de uma visão centrada na arquitectura da informação apresentam como fundamentos que os ambientes informacionais podem ser desenhados para toda a organização sem atender a particularidades e que a organização terá um núcleo de peças de informação - cliente, produtos e transações - em torno dos quais podem ser desenvolvidos os sistemas chave.

Naturalmente existiriam inúmeras vantagens desta visão, pois estes projectos de SI procuravam estruturar a partilha de informação entre múltiplas aplicações de computador e minimizar a redundância. Assim, os executivos poderiam tomar decisões baseados na informação comum.

A realidade é que poucos projectos de sistemas identificados nos planos foram implementados e a maioria dos planos estratégicos foram fechados sem implementação. Uma mudança rápida do negócio conduzia a um sistema desactualizado no momento da conclusão (1 ou 2 anos após o início). Como razões para o fracasso apontam-se dois factos fundamentais: (1) poucas empresas empreenderam tal planificação com alguma preocupação pelo modo como as pessoas actualmente usam a informação e (2) os gestores obtêm 2/3 da informação face a face ou por conversa telefónica e a restante de documentos que vêm da organização e não estão no sistema computacional.

Assim, Davenport (1994) defende o equilíbrio entre Globalismo (significado para toda a organização) e Particularismo (indivíduos e grupos definem a informação de modo que faça sentido para eles). Uma função importante da gestão no futuro será dar ao Particularismo o seu lugar e definir quais os aspectos de informação da organização que são globais.

A existência de múltiplos significados para a informação acarreta alguns problemas. Ela implica mudanças básicas no comportamento dos que fornecem e dos que usam a informação, traduzindo-se em maior necessidade de participação e alocação de tempo quando é necessário definir significados comuns. Se a partilha torna mais fácil para os empregados da companhia obter informação crítica também abre caminho a partes interessadas externas

(concorrentes, poder, piratas de computadores) e, por outro lado os empregados podem fugir com ficheiros, planos, etc.

Na verdade estes problemas são apenas uma pequena parte dos que uma organização pode esperar. Os verdadeiros problemas têm a ver com três questões fundamentais:

1. Desejo de poder: as tecnologias da informação, mais do que atribuir, retiram poder aos empregados que antes controlavam determinados processos.

2. Como escolher entre todas as alternativas que possibilitam a partilha?

3. Como evitar o crescimento da não informação?

As soluções específicas para a transformação da cultura de informação na organização, decorrentes deste autor, poderão ser estendidas aos SIE, considerando-se outras propostas, como o acesso a documentos do negócio e a disponibilização de ferramentas *groupware*, eventualmente não adaptáveis neste contexto.

1.1.3. Codd

É generalizada a convicção de que a Economia não é uma área rica em matéria prima para os SI e, no interior destes, para a especificação de modelos de dados ou de processos de tratamento.

Esta formulação radica numa associação muito comum dos SI ao processamento transaccional vulgarmente designado por *OLTP* (*On-Line Transactional Processing*) o qual é de certo modo contraditório com o processamento de grandes agregados de dados como são as grandezas económicas (embora possa ser a sua base).

Apesar de ser concebível um sistema em que qualquer transacção entre agentes económicos (Estado, empresas e indivíduos) fosse registada centralmente dando origem a um serviço nacional de registo de transacções, este sistema é, na prática, inviável sendo confiado descentralizadamente aos agentes que temporariamente comunicam agregados para departamentos do Sistema Estatístico Nacional ou para o Sistema Fiscal.

Em contrapartida, a Gestão, com as transacções cliente-produto apareceu como o objecto privilegiado de aplicação de SIs.

No entanto, um conceito abstracto como o de circuito económico é bem mais rico quando considerado do ponto de vista dos sistemas de informação em termos de modelo conceptual dos dados do que quando visto apenas em termos macroeconómicos.

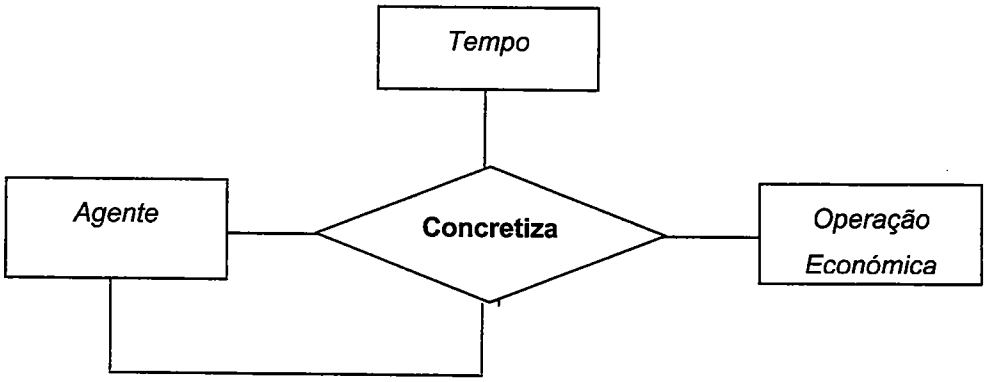
O diagrama apresentado na figura seguinte explicita, conceptualmente, o circuito económico bem melhor do que o utilizado vulgarmente pelos economistas por três razões:

1. Utiliza categorias mais abstractas: agente é mais abstracto do que família ou empresa ou Estado que são mera concretização dele.

2. Permite a introdução do tempo no modelo e, se necessário, do espaço económico, sendo, portanto, representável a um nível global.

3. Vê o circuito como relação entre agentes, permitindo a compreensão do registo quádruplo a nível de contabilidade nacional.

Figura 1. Circuito económico em termos conceptuais



Esta relutância de muitos economistas relativamente aos SI faz com que o próprio ensino de Sistemas de Informação nas áreas de Economia reproduza a ideia anterior e esteja permanentemente atravessado por modelos provenientes das áreas de Gestão quando é necessário exemplificar.

Podemos demonstrar, com um lote elevado de problemas, que, apesar de existirem alguns fundamentos, esta ideia não é a mais correcta e os princípios de modelização de dados e processos têm aqui uma aplicação tão legítima

como em outras áreas. Em Economia existem múltiplas áreas susceptíveis de serem analisadas como os recursos energéticos, o comércio externo, o desemprego, o ambiente, a conjuntura e a contabilidade nacional, isto sem sermos exaustivos.

Por outro lado, esta problemática apresenta a vantagem de obrigar à implementação de regras e princípios provenientes do corpo teórico da ciência económica.

Mas se a ideia ainda pode colher o acordo de alguns ao nível do transaccional, quando pensamos em realidades actuais oferecidas com o processamento analítico da informação, nomeadamente, a multidimensionalidade de análise e a descoberta de conhecimento em bases de dados, ela deve ser abandonada.

Aqueles que, nas áreas de Economia, procuram informação utilizando sistemas computacionais integram-se em algumas das seguintes categorias:

(1) os que fazem um acesso básico aos dados, muitas vezes através de relatórios predefinidos;

(2) os que utilizam um interface muito fácil com os dados com exibição gráfica de resultados e utilização de várias fontes, incluindo textos e fontes externas;

(3) os que procuram informação orientada por assuntos da área de análise, onde a multidimensionalidade é importante e se desdobra por vários níveis de consolidação, podendo ainda fazer recurso a estatísticas e modelização relativamente robusta para suportar a previsão, a relação entre variáveis, a análise multivariada e

(4) os que, a partir de observações de variáveis económicas, procuram elaborar e testar modelos explicativos do seu comportamento.

Ora, o processamento analítico, descrito em Codd (1994), cobre áreas de utilizadores referidas em (2) e (3) e tem uma intersecção não vazia com (4).

1.1.4. Inmon

O que é a informação económica senão, potencialmente, uma monumental base de dados analítica (BDA)?

Inmon tem uma definição muito exaustiva relativamente a este conceito que apresentaremos no capítulo II. Os SIE podem viabilizar a construção, no interior das unidades económicas e outras organizações, de bases de dados analíticas orientadas para os assuntos da economia naqueles domínios que mais se revelem determinantes para a análise do meio ambiente a nível de tendências.

Tais bases de dados, de acordo com as concepções de Codd, poderão permitir processos analíticos e um tratamento dos dados susceptível de produção de conhecimento nas mais variadas formas de representação: árvores de decisão, regras de produção, *clusters*, associações materializam alguns dos resultados possíveis.

Na sequência da descrição feita, parece pacífico que os sistemas de informação para a Economia são uma origem fundamental dos dados que viabilizam o processo de tomada de decisão nas organizações. As características que os viram para os analistas da Economia e para as organizações podem ser consideradas como as que marcam, hoje, mais profundamente a apreciação da sua maturidade. Assim, uma contribuição fundamental deste documento é a consideração desta ligação entre a viabilização do que designamos por processamento analítico e a disponibilização de dados pelos SIE.

Pensamos que, partindo destes elementos e do seu desdobramento em critérios de análise, é possível gerar métricas que nos apoiem na caracterização da maturidade de um Sistema de Informação em geral e, em particular, de um SIE.

1.2. Problemas por resolver

Apesar das contribuições para uma nova visão sobre a disponibilização de informação para os analistas e de novas recomendações sobre a forma de organização do processo, não encontrámos nas referências mencionadas elementos para concluir acerca de alguns problemas associados aos estádios de desenvolvimento dos sistemas de informação para a Economia.

Questão 1. (Questão importante no processo de evolução em estádios) Atingido um estádio, será possível um retorno a um estádio anterior? Se se der uma resposta pela negativa, a realidade desmente-nos: há

sistemas de informação que, no passado, foram modelo e, no momento actual, não cumprem devidamente a sua missão. Mas, se a resposta for positiva, há que procurar meios para responder: se há um retorno no processo de disponibilização de informação, como podemos proceder à sua análise, como poderemos medi-lo e justificá-lo?

Nesta dissertação vamos dar uma resposta a esta primeira questão. Essa resposta parte da observação de que, para se julgar do estádio actual de um sistema de informação, os autores terão utilizado um conjunto de critérios, determinantes na sua perspectiva. É natural que os critérios utilizados há duas décadas sejam completamente distintos dos critérios na era das redes. Isto permite pôr a questão: como evolui a ponderação de um critério? A resposta a esta questão deve responder também a outra: como considerar novos critérios para a análise de maturidade dos sistemas de informação?

Questão 2. (Questão que tem a ver com a formulação de uma hipótese sobre a evolução dos sistemas de informação para a economia na era das redes).

Será possível identificar um conjunto de estádios nesta era? Existirão ferramentas que possamos utilizar numa identificação de estádios? Esta questão associa-se ao conjunto de métodos que esta dissertação utiliza. Em vários momentos nos referiremos à *análise de clusters* como uma ferramenta útil numa diversidade de ambientes. A sua utilização levará à descoberta de possíveis estádios no desenvolvimento de sistemas de informação para a Economia.

Os problemas formulados nas duas questões (transformações nos critérios e estádios na era das redes no âmbito dos sistemas de informação para a economia) são os eleitos para obter uma resposta nesta dissertação.

1.3. Objecto de análise

Neste documento analisamos os sistemas de informação para a economia na sua relação com os analistas de organizações externas que carecem de dados para se integrarem melhor no processo produtivo global.

Uma demorada busca sobre a situação em Portugal levou-nos a uma selecção inicial de dezasseis organizações, com uma vocação vincada para o serviço público, que disponibilizam informação para a economia e, por via disso, se podem tornar como os elementos a observar, interrogar e a partir do qual serão elaboradas as conclusões do documento.

Essas organizações eram as que no início desta dissertação de certa maneira esgotavam o universo da disponibilização interactiva de informação para a Economia, podendo a situação actual ser relativamente diferente, mas a captação inicial do fenómeno permitiu-nos, como iremos descrever, conclusões que, no momento actual, seria extremamente mais difícil descobrir.

Uma justificação para a escolha de organizações vocacionadas para o serviço público tem a ver com a nossa convicção de que a informação, tal como outros recursos, tem de circular através de meios que possibilitem o processo e deve estar ao alcance daqueles que carecem dela para o seu processo de trabalho. Isto não significa que não deva ser paga. Significa, apenas, que não devem ser admitidas práticas que possibilitem a sua apropriação por interesses económicos que possam transformar o relacionamento económico em algo menos são.

1.4. Métodos de Análise

Os métodos utilizados na dissertação encontram o seu fundamento numa duplicidade de carácter em termos práticos.

Um primeiro grupo de métodos é destinado a determinar a viabilização de certo tipo de análises caso a informação disponibilizada pelos SIE detenha determinadas características. Esta consideração radica em todo o conjunto de potencialidades provenientes das concepções de Codd e Inmon e neles incluímos a análise multidimensional e a pesquisa de associações métodos que serão aplicados a dados provenientes dos SIE.

Um segundo grupo de métodos está associado à produção de resultados que são um produto específico da dissertação e, como tal, nos quais se baseiam directamente as suas conclusões. Integramos aqui, as facilidades integradas

em folhas de cálculo, a extracção de informação de base de dados com SQL[2], a análise de clusters, a análise em componentes principais e elementos de estatística descritiva.

Os dois grupos não têm uma intersecção vazia como se pode detectar na descrição feita sendo a utilização da primeira mais cabalmente esclarecida no ponto seguinte.

1.5. Experimentação

Definidos os problemas, os objectos de análise e os métodos a utilizar, cabe referir que, neste documento, apoiaremos as conclusões em casos de aplicação e em observações periódicas dos objectos de análise.

Em todos os casos se procurará responder positivamente a uma de duas questões:

1. Será possível com a informação disponibilizada pelos Sistemas de Informação para a Economia utilizar determinado método de análise?

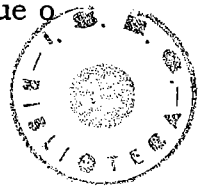
2. Ilustrará o caso uma utilização de processos analíticos?

Os casos seleccionados dizem respeito a múltiplas áreas da Economia: Economia Internacional, Demografia, Fiscalidade. Na área de Economia Internacional existem alguns tópicos que tornam possível a utilização dos métodos referidos, concretamente a nível de Comércio Internacional e de Análise do Investimento Directo Estrangeiro. Nas outras áreas referidas encontraremos processos que poderemos classificar como híbridos, isto é detendo características de várias categorias de processos analíticos *puros*.

Nestes casos demonstraremos a legitimidade do método através da formulação teórica da solução. Assim, em três casos apresentados, há uma referência ao problema que está a ser examinado e à solução preconizada. Em todos se esboçam modelos para a sua compreensão conceptual. Procura-se de seguida obter resultados através da utilização de dados dos Sistemas de Informação para a Economia, independentemente desses dados terem sido ou não

[2] Structured Query Language: linguagem de interrogação de bases de dados da família relacional, desenvolvida pela IBM e cuja exploração se iniciou na década de 80.

disponibilizados em linha. Finalmente, julga-se o resultado e os meios que o possibilitaram.



1.6. Produtos da dissertação

A definição do problema e a experimentação com os métodos referidos, que constituem os dados originais da dissertação, servirão para formular conclusões acerca das interrogações levantadas sobre o problema dos critérios de análise de maturidade e apoiarão uma hipótese sobre a evolução dos sistemas de informação para a economia na era das redes.

Uma visão dos estádios de desenvolvimento de sistemas de informação apoiada na existência de um ciclo de vida das variáveis de análise do tipo nascimento, crescimento, maturidade e declínio permitirá resolver o problema posto inicialmente. Demonstrar-se-á que, pelo facto de se modificarem os critérios de análise, forçosamente uma métrica relativa à maturidade deverá apresentar valores diferentes mesmo não havendo modificação do sistema de informação. Não há modificação no sistema de informação, mas há modificação nos olhos que o vêm pelo que ele é diferente em termos de maturidade.

Esta análise, apoiada em considerações sobre o ciclo de vida, recorrerá a uma analogia com desenvolvimentos da área de redes neuronais o que permitirá introduzir a problemática não linear nas funções resultantes.

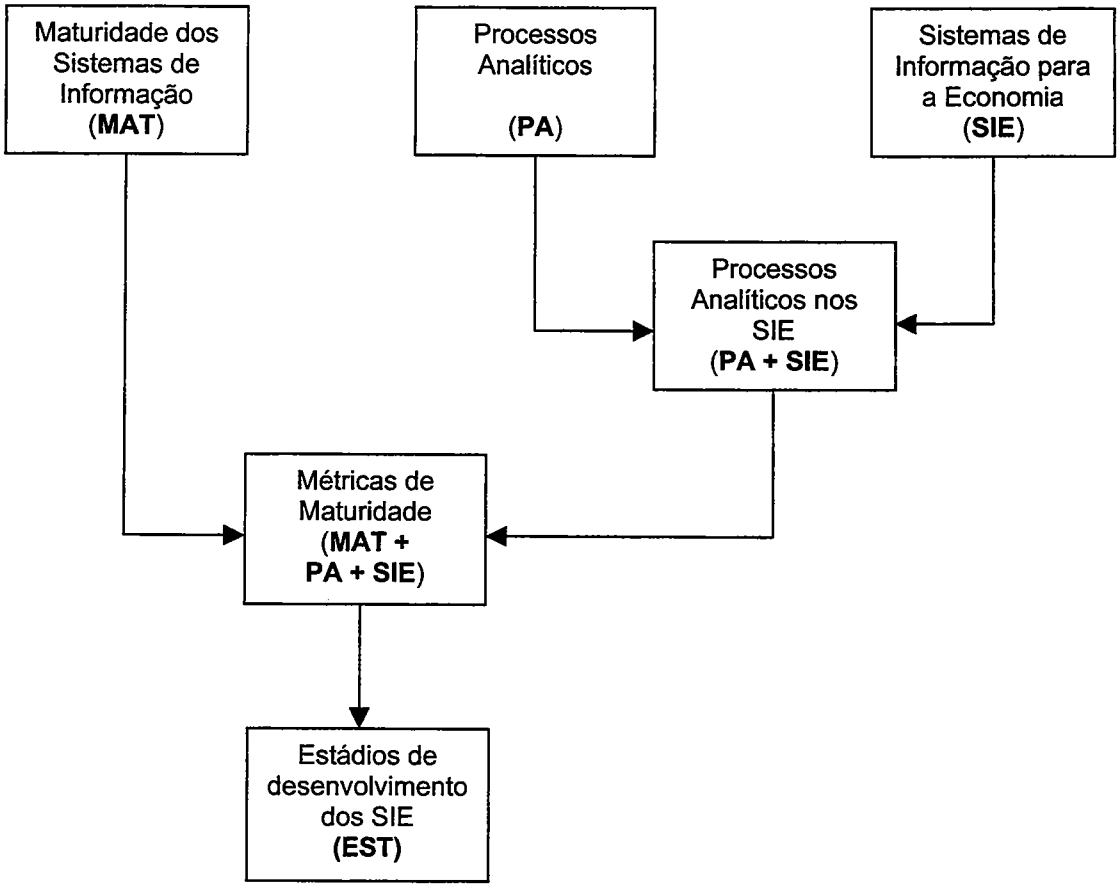
Uma hipótese sobre a evolução dos SIE na era das redes em Portugal, alicerçada na análise de clusters, defenderá a existência de um conjunto de estádios onde serão relevantes inicialmente, a disponibilização em linha e gradativamente, a procura de actualidade, a capacidade de processamento e a abertura ao analista. As organizações objecto de análise serão, então, associadas a cada um dos estádios o que tornará possível a formulação de recomendações para um desenvolvimento futuro das mesmas.

2. Planificação global da dissertação

A lógica de construção desta dissertação baseia-se no tratamento isolado, numa primeira fase de aspectos conceptuais, métodos e objecto da análise e sua integração progressiva até à possível geração de conclusões gerais.

O diagrama seguinte ilustra o caminho que vamos percorrer.

Figura 2. Lógica de desenvolvimento da dissertação



Nestas condições, decidimos organizar a dissertação em seis capítulos, precedidos de uma introdução e seguidos de conclusões gerais. Em cada capítulo, procuraremos concluir com um resumo que poderá ou não ser retomado nas conclusões gerais.

Explicitamos, de seguida, mais detalhadamente, o conteúdo de cada um dos componentes referidos.

Introdução

Nesta grande divisão da dissertação, a presente, procedemos à apresentação do seu conteúdo em torno dos objectivos da investigação, métodos utilizados, objecto da dissertação e resultados.

Capítulo I - Questões da Maturidade dos Sistemas de Informação

Como o seu nome indica, este capítulo vai tratar o problema da maturidade dos sistemas de informação em geral. A perspectiva é válida

para qualquer tipo de sistema de informação, pois foi nesse campo que surgiram as principais contribuições para a análise da maturidade. Por esse motivo, o capítulo é despojado de qualquer referência ao objecto da dissertação: os sistemas de informação para a economia. Uma designação alternativa seria *a maturidade dos sistemas de informação em si mesmo.*

O objectivo do capítulo é relacionar as questões de maturidade tal como elas se manifestaram nos primeiros autores da teoria de estádios com um conjunto de problemas suscitado pela natureza particular do desenvolvimento nas sociedades actuais.

No final, deverá ter ficado clara a contribuição de um conjunto de autores relativamente vasto para o problema e esboçado um lote inicial de critérios para análise de maturidade dos sistemas de informação nos tempos actuais.

Capítulo II – Processamento Analítico

O objectivo deste capítulo consiste na apresentação, relativamente formal, de todos os métodos utilizados na dissertação.

Integrámos no capítulo os métodos utilizados como caracterizadores do processamento analítico e os que se integram na designação e terão sido por nós utilizados para obtenção de resultados quer sejam ou não integrados num processo de pesquisa de informação associado a um SIE.

Tal apresentação, procurará sempre fornecer elementos fundamentais, designadamente:

- uma caracterização do método;
- o tipo de resultado que, em alguns casos, poderá ser designado por a forma específica de representação do conhecimento utilizada;
- a transformação prévia aos dados necessária para viabilizar a aplicação;

- uma difusão sucinta, descrevendo, por vezes, os algoritmos disponíveis.

Capítulo III – Organizações responsáveis pela informação para a Economia em Portugal

Examinamos, neste capítulo, um conjunto de organizações que disponibilizam informação sobre a Economia.

A nossa perspectiva é a de que esta informação será utilizada por analistas económicos integrados em processos de investigação da realidade económica. Tal perspectiva permite-nos esboçar uma visão do que podemos considerar como ambiente da relação do SIE com o analista que, no fundo, é o quadro de fundo da relação possibilitadora ou não do acesso à informação.

O exame feito às organizações permitirá detectar algumas áreas de actuação das mesmas, eventuais redundâncias e práticas de disponibilização incompleta que serão modeladas e especificadas formalmente.

Capítulo IV - Aplicações de Processamento Analítico suportadas por Sistemas de Informação para a Economia em Portugal

Como afirmámos anteriormente, este é o primeiro capítulo que relaciona dois tópicos abordados anteriormente na perspectiva *em si mesmo.*

O seu objectivo consiste na procura de processos analíticos nos SIE apresentados no capítulo anterior e na concepção de processos analíticos aplicáveis à informação que os SIE disponibilizam.

Dar-se-á, ainda, relevo ao surgimento de processos analíticos com características de vários processos puros. Pela sua natureza, em alguns momentos, o capítulo assume um carácter experimental tendente a percepcionar a viabilização de um determinado processo.

No final do capítulo, conheceremos o processamento analítico disponibilizado (sob a forma de resultados) e possibilitado pelos SIE.

Capítulo V – Métricas para análise da maturidade dos Sistemas de Informação para a Economia

Neste capítulo iniciamos a apresentação dos resultados do processo de pesquisa. Tais resultados assumem, num primeiro momento, a forma de uma contribuição para análise da maturidade dos SIE. Com a continuação da análise possibilitada por tal resultado é possível analisar empírica e profundamente a maturidade dos SIE. Tal análise, materializada num processo de avaliação, é conseguida através da utilização de métricas.

Um primeiro momento de reflexão é suscitado por uma tentativa de determinar a maturidade dos SIE a partir de um número inferior de critérios do que os utilizados na análise, através da intervenção de um processo analítico de natureza difusa.

Capítulo VI – Estádios de Desenvolvimento dos Sistemas de Informação para a Economia

Os resultados da perspectiva anterior poderão ser utilizados na procura de organizações que mantenham relações de proximidade em termos de critérios de análise.

Justifica-se, assim, a existência de um novo capítulo destinado à divulgação de resultados, agora centrado naqueles que provêem dos instrumentos criados no capítulo anterior e para os quais se porá a hipótese de existência em termos de estádios de desenvolvimento. A confirmação de tal hipótese surgirá após uma *análise de clusters* que permitirá estabelecer agrupamentos a nível dos SIE.

O exame gráfico dos agrupamentos poderá viabilizar a formulação de uma hipótese sobre os estádios de desenvolvimento dos sistemas de informação para a economia e compreender os efeitos do tempo sobre tal formulação.

Admitindo que a transição entre estádios não seja pacífica, repetir-se-á a sua visualização e interpretação em anos posteriores e serão apontadas medidas para aceleração da transição.

Conclusões Gerais

Finalmente, estaremos em condições de formular conclusões acerca do processo de evolução da maturidade dos sistemas que disponibilizam informação para a Economia e do papel que o processamento analítico aí desempenha. Tais conclusões serão suportadas por todos os capítulos embora, naturalmente, as provenientes dos dois últimos sejam mais relevantes.

O trabalho apresentado neste documento integra um conjunto de anexos onde serão apresentados listagens de programas utilizados, listagens de comandos em linguagens de natureza variada como, por exemplo, de quarta geração, resultados de processos analíticos quando não se justifique a sua inclusão no documento principal e eventuais provas de contactos efectuados para viabilizar a investigação ou até para testar determinados níveis de relacionamento dos SIE com utilizadores. Serão também referidos resultados de trabalhos semelhantes, realizados no país, onde os objectos de análise sejam diferentes.

Capítulo I - Questões da Maturidade dos Sistemas de Informação

Este primeiro capítulo tem por objectivo a revisão da literatura em torno da problemática da dissertação.

As referências que nos interessam respeitam a contribuições para a teoria de estádios de desenvolvimento dos sistemas de informação e materializam-se na primeira secção tratada.

É, no entanto, extremamente importante, proceder a um confronto entre esta teoria e todo um conjunto de visões críticas alicerçadas na questão da evidência empírica. Assim, a segunda secção deste capítulo, procederá a uma revisão de questões que terão servido para que um grupo de académicos atacasse o modelo de Nolan, sendo complementada pelo testemunho da continuação da sua utilização em anos recentes o que foi reconhecido e interpretado pelos críticos.

Na terceira secção, examinamos questões associadas à articulação dos elementos anteriores, procurando respostas para as referências que defenderam o processamento de dados externos e para todo um conjunto de problemas relevantes como a capacidade de processamento, a abertura ao analista, as novas qualificações presentes e a inovação.

Os novos problemas postos pelos sistemas de informação na era das redes, consubstanciados na gestão da informação e do conhecimento, na integridade temporal, na análise multidimensional e na descoberta de conhecimento em bases de dados, terminam o capítulo, abrindo perspectivas para a reformulação de uma teoria de estádios na era actual. Esta reformulação, despida de alguns aspectos sujeitos a crítica demolidora, como o formato da curva de despesas em tecnologias de informação e o seu carácter determinante, poderá apresentar uma nova consistência e apoiar uma interpretação centrada nos SIE.

1. A Maturidade como estádio final do processo de desenvolvimento dos sistemas de informação

Nesta secção, forneceremos uma visão geral sobre a contribuição de vários autores que, posteriormente, será escalpelizada no sentido de responder a um

conjunto de questões, problemas reais nos sistemas de informação de hoje. Por fim, apresentaremos os elementos críticos que poderão atribuir a estes modelos um lugar bem definido nas concepções sobre sistemas de informação.

Entre os modelos mais importantes de análise dos estádios de desenvolvimento dos SI encontram-se:

- o modelo de seis estádios de Nolan (1979);
- o ciclo inovação / implementação de Cash, McFarlan e McKenney (1983);
- a perspectiva multi-tecnológica de Earl (1993);
- a relevância dos mercados e dos dados externos de Bhabuta (1988);
- a contribuição de Hirschheim (1988) e
- a visão holística de Galliers e Sutherland (1994).

Apesar da distância no tempo, ou talvez por causa dela, o modelo de Nolan é o mais largamente conhecido e utilizado, embora os outros autores tragam novos e importantes elementos para a análise. Por outro lado, o próprio modo como o que é relevante em termos de SI se vai transformando induz a necessidade de transformação dos elementos de análise.

Com efeito, é legítimo pormos a questão: existindo um padrão de desenvolvimento para os SI, existindo múltiplos autores com contribuições para esse padrão de desenvolvimento, em que momento é considerada uma fase de maturidade dos Sistemas de Informação como algo presente? Poderá acontecer que esse momento não seja conscientemente referido pelos autores o que permite pensar na existência de elementos que, progressivamente, se vão materializando e que não sejam mais do que, finalmente, o despertar para o tratamento da questão. Designaremos esse processo por *maturação* e vamos procurar alguns dos atributos que nos permitirão caracterizá-lo.

Aceitando um processo de maturação dos SI em geral e, naturalmente, dos que têm por objecto os dados relativos à Economia, põe-se o problema de como modelizar este processo de modo a conseguirmos perspectivar um retorno para fases mais recuadas. A questão está em que, de acordo com a teoria de estádios, alcançada uma fase esta parece deste logo consumada embora sem garantia de sustentabilidade.

Exemplo 1. Na década de setenta, a existência de uma equipa de programação com elevada qualidade técnica, utilizando linguagens da terceira geração e monitores de teleprocessamento, com certeza, contribuiria para uma elevada classificação do SI em termos de maturidade. Admita-se a manutenção desta equipa à entrada do ano 2000 sem modificações radicais noutros elementos de análise. Claramente, há aqui um processo de obsolescência do SI que não é captado em qualquer das contribuições referenciadas na teoria de estádios.

Enfim, sendo pacífica a existência de um processo de maturação para os SI, devemos reflectir em quais são os critérios que permitem analisar esse processo no contexto da informação para a Economia, parecendo-nos que este campo do conhecimento deve deter alguma especificidade.

Todos os autores referenciados consideram que os sistemas de informação passam por um conjunto de estádios de desenvolvimento e estes são, desde logo, reveladores do tipo de preocupações com que se defrontaram.

Assim, Nolan (1979) usa um conjunto de seis estádios que apelidou de *início*, *contágio*, *controlo*, *integração*, *administração de dados* e *maturidade* os quais denunciam a necessidade de controlar o processo de desenvolvimento de sistemas de informação, a preocupação com a integração de aplicações e com a administração dos dados como um recurso organizacional. Pode ainda deduzir-se da sua contribuição um avanço no processo de gestão do sistema de informação na transição da gestão tecnológica para a gestão da informação.

A perspectiva de McFarlan, constante em Cash et al (1983), é mais virada para as tecnologias de informação, relativamente às quais é apresentado um modelo que passa por *pesquisa e identificação*, *aprendizagem*, *racionalização/controlo* e *maturidade*.

A enumeração de estádios relativa a Earl[3], *início*, *controlo*, *apoio ao negócio*, *direcção colegial*, *procura da vantagem estratégica* e *integração TI/negócio* permite detectar novos e importantes elementos. Com efeito, Earl (1993) inicia uma importantíssima discussão relativa à vantagem competitiva presente ao

[3] A designação dos estádios em Earl e Bhabuta é da nossa responsabilidade. Earl limitou-se a enumerar os estádios de uma forma sequencial: I, II, III, IV, V , VI

longo de toda a década de oitenta e relaciona o sistema de informação com o apoio ao negócio.

Bhabuta[4] (1988) foi portador de uma das contribuições mais interessantes na matéria deste documento ao referir a importância dos dados externos no processo de tomada de decisão estratégica nas organizações e utilizou um modelo de quatro estádios que denominámos *procura da eficiência, procura da eficácia, análise do ambiente* e *interorganizacional.*

Os estádios de Hirscheim (1988) são pouco significativos em termos de conteúdo o que será esclarecido noutra secção (*entrega, reorientação* e *reorganização*), parecendo a sua contribuição menos conseguida que a de outros autores.

Por fim, Sutherland e Galliers (1994) utiliza um conjunto de designações com uma conotação fortemente descritiva para os seus estádios: *adocracia,* significando um desenvolvimento e utilização mais baseada nas convicções técnicas dos seus agentes do que em necessidades organizacionais, *início das fundações, ditadura centralizada,* referindo o início do processo de planeamento, *dialéctica democrática e cooperação,* acentuando a necessidade de colaboração com o utilizador final, *oportunidade interempresarial,* já encontrada em Bhabuta (1988) e *relações harmoniosas e integradas.*

Considerando o simples enumerar dos estádios de cada um dos autores referenciados e escassos elementos adicionais insertos nos pontos seguintes é desde já possível apresenta uma matriz da contribuição em função de alguns conceitos chave do problema tratado neste documento.

Quadro 1. Tópicos relevantes por autor

	Nolan	McFarlan	Earl	Hirscheim	Bhabuta	Galliers e Sutherland
Maturidade	√	√				
Vantagem Estratégica			√	√		
Integração TI/Negócio			√	√		
Análise do Ambiente					√	
Interorganizacional					√	√

[4] A designação original utilizada por Bhabuta foi Fase 1, Fase 2, Fase 3 e Fase 4

O quadro anterior permite concluir da existência de um movimento, no tempo, que consiste na deslocação dos critérios de análise da maturidade dos sistemas de informação de critérios centrados na organização para critérios que a relacionam com os seus concorrentes (vantagem estratégica) ou com outras organizações (interorganizacional).

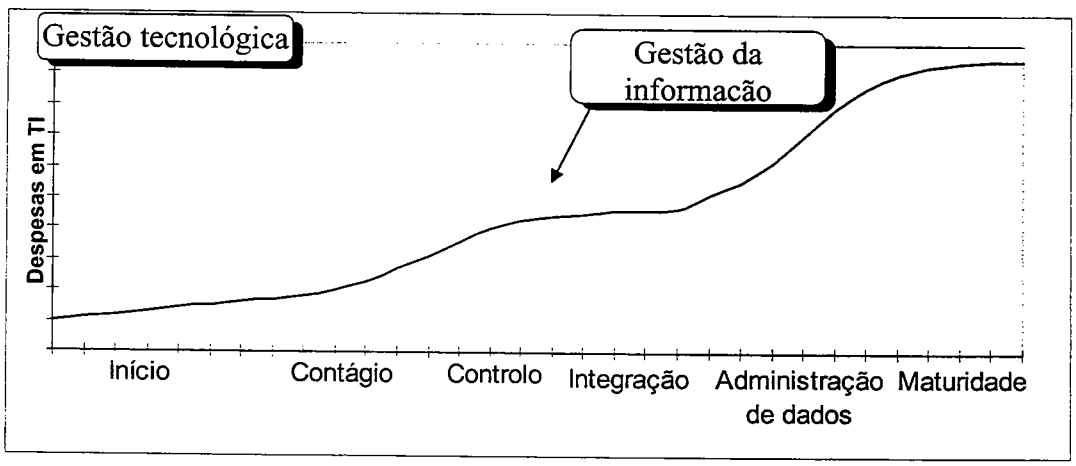
1.1. O modelo de Nolan

O modelo original de Nolan com quatro estádios apresentado em Gibson e Nolan (1974) evoluiu para um modelo com seis estádios em Nolan (1979), o aplicado mais vulgarmente. Tal como os modelos que se lhe seguiram, baseia-se na premissa de que *na utilização e gestão de tecnologias de informação, as organizações passam por um número de fases de desenvolvimento identificáveis*. Estas fases de desenvolvimento são usadas para determinar o nível de maturidade das organizações com o objectivo de identificar problemas chave associados à evolução futura do SI.

Gibson e Nolan (1974) defendeu que o estádio de desenvolvimento podia ser identificado primariamente pela análise das despesas em TI (investimentos e custos de exploração), postulando que estas despesas seguiriam no tempo uma *curva S*. Sustentou ainda que esta curva parecia representar o *caminho de aprendizagem* relativamente à utilização geral de tecnologias da informação (TI) na organização.

A expansão do modelo original para um modelo com seis estádios foi operada com a adição de dois novos estádios entre controlo e maturidade, que Nolan (1979) designou por integração e administração de dados.

Figura 3. Evolução das Despesas em TI ao longo dos Estádios de Nolan



Ao aproximar-se o fim do estádio de controlo, a organização transforma o seu estilo de gestão do SI passando da pura *gestão tecnológica* para a *gestão da informação*. Esse ponto é marcado no gráfico acima por uma seta.

Além das despesas em processamento de dados existem mais quatro processos de desenvolvimento que podem ser analisados para identificar o estádio de maturidade relativamente ao uso de TI:

- a amplitude do *portfolio* aplicacional na organização (movendo-se de sistemas financeiros e contabilísticos de utilização geral a nível operacional para sistemas de informação de gestão);
- a organização do processamento de dados (que passa de centralizado, à porta fechada para a gestão da informação como um recurso, na maturidade);
- a actividade de planeamento e controlo do processamento de dados (foco mais interno nos primeiros estádios para um foco externo nos últimos);
- o nível de participação do utilizador (que passa de uma atitude reactiva nos dois primeiros estádios para uma força condutora nos estádios intermédios e para uma associação na maturidade).

1.2. O modelo de Cash, McFarlan e McKenney

Não pondo directamente em causa o modelo de Nolan, a contribuição de Cash et al (1983) procura generalizá-la e proceder à sua aplicação quer se trate de bases de dados (a que o modelo de Nolan surge indubitavelmente ligado), quer a redes locais ou à utilização de novas estações de trabalho que permitam a concepção assistida por computador, reconsiderando os quatro estádios originais de Gibson e Nolan (1974). A designação utilizada é a de ***aprendizagem organizacional*** que é considerada como um processo de quatro fases:

- identificação da tecnologia e investimento;
- aprendizagem e adaptação tecnológica;
- racionalização/controlo de gestão e
- maturidade/transferência tecnológica num âmbito alargado.

A teoria de estádios transfigura-se assim num modelo muito geral, eternamente aplicável, perdendo por esse facto a sua capacidade de explicação contingencial.

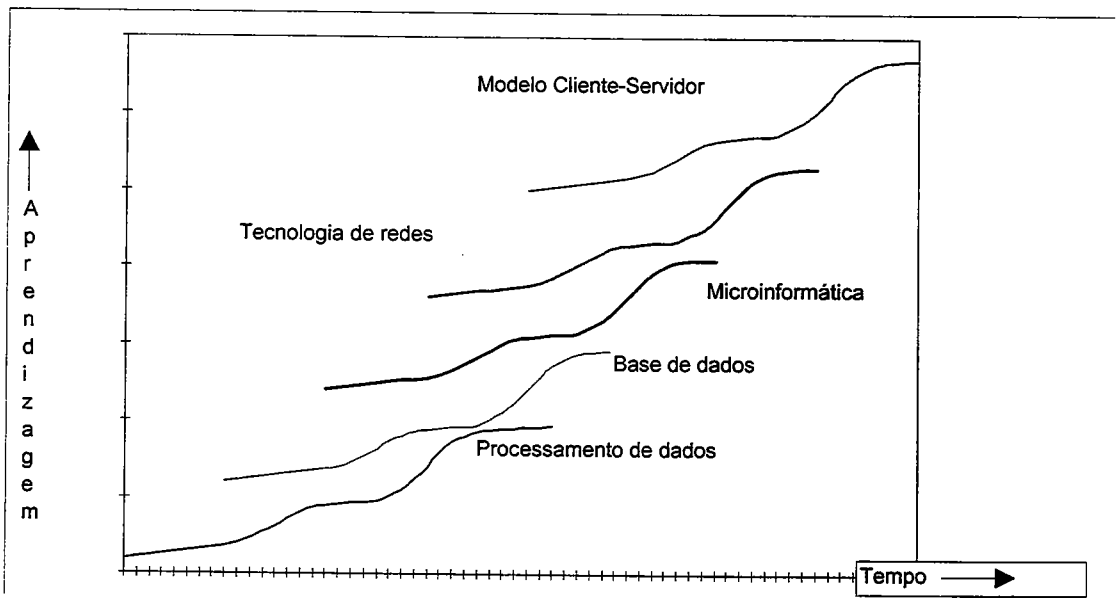
1.3. O modelo de Earl

Diferentemente do modelo de Nolan, o de Earl centra-se nos *estádios que as organizações atravessam a planear os seus sistemas de informação.*

Earl (1993) ilustra a modificação no planeamento de sistemas de informação concentrando a atenção no que é considerado a tarefa primária do processo, nomeadamente o objectivo principal, a força condutora do processo de planeamento, a ênfase metodológica e o contexto no qual o planeamento tem lugar.

Neste processo, as organizações enfrentarão múltiplas curvas de aprendizagem, encontrando-se em estádios diferentes num dado momento relativamente a cada uma delas.

Figura 4. As Múltiplas Curvas de Aprendizagem de Earl



Na sua essência o argumento de Earl é o seguinte:

- As organizações iniciam os seus esforços de planeamento pela tentativa inicial de atingir o estado de arte corrente na utilização e gestão de tecnologias de informação.

- Progressivamente, a atenção desloca-se da gestão das tecnologias de informação para passar a ter uma ligação mais sólida com os objectivos do negócio.

- Finalmente, a orientação toma um foco estratégico, mantendo-se um equilíbrio relativamente à *composição das equipas de planeamento* (utilizando recursos humanos da área de sistemas de informação, da gestão e dos utilizadores), ao tipo de informação ambiental/organizacional (desenvolvimento de sistemas interorganizacionais) e à diversidade de aproximações adoptadas (sendo aceites múltiplos métodos).

1.4. O modelo de Bhabuta

Baseado num trabalho anterior de Gluck et al (1980), que propõe um processo com quatro estádios de evolução para o planeamento estratégico e um modelo de assimilação e difusão de TI algo similar ao postulado por McFarlan, Cash e McKenney[5], Bhabuta desenvolveu um modelo que *procura seguir o progresso no sentido do planeamento estratégico de sistemas de informação* (Bhabuta, 1988).

A argumentação de Bhabuta encontra suporte na controvérsia de que as *estratégias baseadas na melhoria da produtividade (e os sistemas de informação necessários para as suportar) tornar-se-ão o paradigma dominante nos mercados competitivos e turbulentos da década de 90* (Bhabuta,1988).

O seu modelo é mais abrangente que o de Nolan ou Earl, na medida em que procura juntar elementos de formulação de estratégia, sistemas de informação e mecanismos pelos quais a função de sistemas de informação é gerida.

Assim, o resultado do seu trabalho traduz-se numa interligação do planeamento estratégico com os sistemas de informação e a organização da função de sistemas de informação.

Para tal, começa por identificar quatro fases evolucionárias do planeamento estratégico - (1) planeamento financeiro básico, (2) planeamento baseado em previsões, (3) planeamento orientado externamente e (4) gestão estratégica - e, configurando a inovação como um dos elementos fundamentais de suporte à

[5] Ver 1.2. neste capítulo

estratégia, traça algumas das suas características quando relacionadas com essas fases: (i) produtividade do nível operacional e difusão da inovação, (ii) foco na inovação (nicho) e no nível de produtividade táctico/operacional, (iii) foco na inovação e produtividade estratégica (foco qualidade) e (iv) inovação sistemática e produtividade.

As consequências sobre o sistema de informação organizacional são imediatas forçando-o a considerar a integração de dados externos no processamento, a procura do interorganizacional e o suporte às necessidades do negócio.

Este percurso traduz-se em marcos ao nível da problemática dos valores, sendo considerados como fundamentais na associação a cada estádio[6] (1) o respeitar o orçamento, (2) o prever o futuro (3) o pensar estrategicamente e (4) o gerir o futuro.

1.5. O modelo de Hirscheim

O modelo de Hirscheim, derivando também dos trabalhos de Nolan, resulta de investigação levada a cabo na primeira metade de 1986 sobre a evolução da gestão de sistemas de informação em algumas organizações inglesas. Como resultado desta pesquisa, concluiu que as organizações onde a gestão de topo compreendeu que os sistemas de informação são vitais para o negócio passam por três estados evolucionários que designou por entrega, reorientação e reorganização.

O estádio de entrega caracteriza-se pela percepção sobre a capacidade da função responsável por sistemas e tecnologias de informação (SI/TI) para *entregar os bens e serviços*. Apesar da seriedade presente neste processo há muitas vezes insatisfação relativamente à qualidade dos sistemas (tecnologias da informação, aplicações) disponíveis e à função responsável pelos mesmos, juntamente com juízos sobre os custos em tecnologias de informação e de desenvolvimento e a consistência das políticas de *hardware* e *software*.

Muitas vezes esta fase é iniciada pela mudança do anterior *gestor de processamento de dados* por um elemento recrutado externamente com uma

[6] Recorde-se que anteriormente considerámos as designações seguintes para os estádios de Bhabuta: procura da eficiência, procura da eficácia, análise do ambiente e interorganizacional

boa imagem e substancial experiência que manterá no entanto uma perspectiva de gestão tecnológica.

A ênfase nesta fase está na entrega de sistemas e, de acordo com isso, o novo executivo gasta a maior parte do seu tempo com matérias internas ao departamento. O papel principal é restaurar a credibilidade da função e criar confiança nos utilizadores e na gestão de topo de que a função está realmente a suportar as necessidades actuais e corre eficientemente. Nesta fase, a formação é escassa, mas, quando é providenciada, é interiorizada, sendo dirigida aos recursos humanos da área de tecnologia de informação.

No estádio de reorientação, a gestão de topo (ou o director ultimamente responsável por SI) muda o foco da atenção da entrega de serviços básicos de sistemas de informação para a *exploração de TI para obter vantagem competitiva.* É feita uma tentativa para alinhar o investimento SI/TI com a estratégia do negócio. É nesta fase de reorientação que o negócio é tratado pelo computador. Com esta mudança de direcção/ênfase é possível nomear um gestor de sistemas de informação acima do gestor de processamento de dados. Tipicamente, o novo lugar é preenchido por um elemento interno da organização, um executivo sénior que tenha dirigido uma unidade do negócio ou que tenha estado activo num papel da empresa tal como marketing ou formulação de estratégia. Ele poderá ter experiência limitada em processamento de dados, mas será respeitado pela gestão de topo dada a sua capacidade para gerar mudança. Nesta segunda fase a atenção centra-se no mercado, no ambiente externo à empresa, na utilização de TI para vantagem competitiva e na propagação da cadeia de valor através de sistemas interorganizacionais.

Por fim, no estádio de reorganização, o gestor de sistemas de informação é posto perante a *gestão de interfaces e relações entre a função de sistemas de informação e o resto da organização.* Algumas áreas estarão estrategicamente dependentes da área de sistemas de informação, outras verão nela mais um papel de suporte. Algumas disporão de significativa capacidade em TI, particularmente com o desenvolvimento da utilização de computadores pelo utilizador final e alguns executivos do negócio estarão a liderar o desenvolvimento de tecnologias e de sistemas aplicacionais. Cada vez mais, a gestão do sistema de informação terá necessidade de articular a capacidade no

centro e a desagregação por unidades/funções do negócio. Estas relações transformadas e em mutação requerem uma gestão cuidada e muitas vezes reorganização e mais uma vez a atenção centra-se em assuntos internos (organizacionais), em oposição a externos (mercado).

1.6. O modelo de Sutherland e Galliers

Para estes autores, os maiores defeitos do modelo de Nolan referem-se à ausência de foco organizacional e de gestão e às *assunções simplistas e subjectivas* em que se baseou (Galliers e Sutherland, 1994).

Esse modelo pouca ajuda terá propiciado ao gestor de TI que procurasse criar uma função de sistemas de informação com sucesso numa organização o que terá sido corrigido por trabalhos subsequentes de Earl, Bhabuta e Hirscheim. Em todos com excepção do último caso, todavia, os modelos descreviam mais como uma organização se situava em determinado estádio de maturidade no planeamento de TI do que o necessário para progredir de forma a atingir o estádio de maior maturidade.

Para analisar o estádio da utilização e gestão do sistema de informação nas organizações, Galliers e Sutherland carecem de um conjunto de factores chave recorrendo para isso aos utilizados por McKinsey & Company, vulgarmente designados por *os sete 'Ss'*[7].

O primeiro factor é o designado por *estratégia*, entendida como plano ou conjunto de acções que conduzam a alocação de recursos escassos de uma empresa no tempo para alcançar objectivos identificados.

Os autores utilizam ainda o factor *estrutura* que lhes serve como caracterização do modelo organizacional (funcional, descentralizado, etc).

Os *sistemas* são entendidos como os relatórios procedimentais e processos rotineiros.

Outro factor é o designado por *recursos humanos*, conceito através do qual procedem à descrição *demográfica* das categorias de pessoal importantes na empresa (engenheiros, empreendedores, *MBAs*, etc.).

[7] Esta designação deriva da proveniência anglo-americana dos elementos considerados: strategy, structure, systems, staff, style, skills, superordinate goals.

O factor *estilo* corresponde à caracterização de como os gestores chave se comportam para conseguir os objectivos da empresa e também o estilo cultural da organização.

O sexto factor, *qualificações*, respeita às capacidades distintivas de pessoal chave ou da empresa como um todo e finalmente os *princípios* serão os conceitos guia significativos que a empresa transmite aos seus membros, os valores comuns ou cultura da empresa.

O desenvolvimento da maturidade em tecnologias de informação numa organização pode ser representado por seis estádios cada um com o seu conjunto particular de condições associado aos sete 'Ss'.

Assim, a prioridade no estádio *adocracia* é a aquisição de hardware e software em prejuízo de recursos humanos e formação. As aplicações são simples e, vulgarmente, relacionam-se com aspectos financeiros do negócio (por exemplo, sistemas de contabilização). A estratégia normalmente utilizada neste estádio relaciona-se com a aquisição de *packages standard* e, em muitas ocasiões, mais do que proceder ao desenvolvimento interno (o que seria a norma antes dos anos 80), podem ser contactados fornecedores externos para desenvolver aplicações específicas.

No estádio seguinte, *início das fundações*, os recursos humanos da informática já integrados num quadro permanente, procuram descobrir e satisfazer as necessidades do utilizador. Uma das características desta fase é o olhar-se o futuro sempre como se fosse uma simples extensão do passado. Se, no primeiro estádio, se instalaram sistemas predominantemente financeiros, agora as organizações tendem a *concentrar o desenvolvimento de aplicações noutras áreas do negócio.* Os sistemas não servem exclusivamente a área financeira, sendo, no entanto, sistemas destinados à utilização por áreas operacionais. Não é feito qualquer planeamento eficaz, embora o pessoal informático possa argumentar que planeia pelo menos o seu próprio trabalho. O único planeamento considerado relaciona-se com o processo anual de orçamentação. A natureza *bottom-up* da percepção das necessidades informáticas conduz a um grande atraso nos sistemas a desenvolver (*backlog* aplicacional) e a pedidos para maiores aumentos nos gastos em processamento de dados.

O estádio 3 (*ditadura centralizada*) procura corrigir os desequilíbrios causados pela natureza *ad hoc* do desenvolvimento inicial e da investida cega dos sistemas do estádio anterior. A necessidade de planeamento é reconhecida e defendida por membros da gestão (incluindo algum pessoal informático). Mas este estádio é marcado pelo *conflito entre conseguir o controlo central e mover-se na direcção da computação pelo utilizador final,* problema que deixa a área de informática num estado de confusão com uma coordenação fraca entre o departamento de processamento de dados e os que o utilizam a tecnologia.

No estádio de *dialéctica democrática e cooperação,* acentua-se a procura de integração e coordenação. A postura deixa de ser uma *torre de marfim* defensiva para se integrar na perturbação do mundo real da organização.

O estádio de *oportunidade interempresarial* leva à procura activa de oportunidades para o uso estratégico de tecnologias da informação a fim de proporcionar uma vantagem competitiva à organização. Esta estratégia envolve uma pesquisa substancial no ambiente envolvente. As forças que conduzem o processo de desenvolvimento das tecnologias de informação olham, predominantemente, para o exterior sendo as operações internas delegadas com sucesso aos outros gestores.

No estádio de *relações harmoniosas e integradas*, a gestão é posta perante a questão de manter a vantagem estratégica que tinha sido árdua de alcançar no estádio anterior. Isto envolve uma avaliação constante de todos os usos de tecnologia de informação quer na organização quer no exterior. Planeamento interactivo que envolve a análise de condições futuras e presentes é o foco da formulação de estratégia.

2. A maturidade e a questão da evidência empírica

A apresentação de um conjunto de estádios que procurava explicar a evolução das organizações em termos de utilização e gestão das tecnologias de informação teve um impacto significativo no mundo académico. Muitos investigadores procuraram elementos que lhes permitissem refutar ou confirmar as posições Richard Nolan. Um dos grupos mais importantes surgiu no Canadá, na Simon Frazier University, e encontrou em Izak Benbasat a sua expressão mais elaborada, tendo sujeitado alguns aspectos da concepção em análise a uma crítica demolidora, designadamente:

- a curva orçamental em forma S;
- a utilização de sistemas de custeio;
- as qualificações do gestor da área de processamento de dados / sistemas de informação;
- a administração de dados;
- a localização da função SI na organização;
- a aplicação de *benchmarks*;
- o uso de planeamento e controlo e
- a constituição de *steering committees*.

Nos pontos seguintes, vamos examinar, a partir de um documento que resumiu o processo, Benbasat et al (1984), cada um destes elementos em particular e fazer um balanço sobre o seu impacto em termos da utilização posterior da teoria de estádios

2.1. A curva orçamental em forma S

Nolan (1973;1) tinha sugerido que as despesas com a computação exibiriam a forma *S* característica de uma curva de aprendizagem no tempo:

> *The points of inflection of a generalized S-shaped, computer budget curve provided the basis for the stage hypothesis... It is proposed that the computer budget curve will serve as a useful surrogate for representing the growth phenomenon for use of computers in organizations.*

De acordo com Benbasat et al (1984), os investigadores Lucas e Sutton, em 1977, numa actividade suportada por entrevista e questionário a 29 organizações, terão concluído pela não confirmação desta hipótese.

2.2. Sistemas de Custeio

De acordo com Nolan (1979), à medida que a maturidade das firmas se desenvolvia do estádio III para o estádio VI, estas deveriam ter adoptado sistemas de custeio dos utilizadores.

Em 1979, Drury and Bates, através de uma investigação suportada por questionário a 64 organizações, citada por Benbasat et al (1984), terão igualmente concluído pela não confirmação de tal hipótese.

2.3. Qualificações do gestor da área

Nolan (1973;2) tinha indicado que quanto mais se desenvolvia a maturidade dos sistemas de informação nas organizações, mais o responsável da área de processamento de dados requereria qualificações de ordem geral e de gestão em vez de qualificações como especialista e de ordem técnica.

Benbasat, Dexter e Mantha realizaram em 1980 um questionário a 35 organizações cujos resultados levaram à não confirmação deste ponto, do qual é dado testemunho também em Benbasat et al (1984).

2.4. Administração de Dados (AD)

No modelo original, desenvolvido antes de os conceitos de AD serem largamente compreendidos ou aceites, a emergência de uma função de AD formal foi uma das características de uma organização *madura* do estádio IV. Nolan (1975) notou, todavia, que existiriam indubitavelmente mais curvas em forma *S* à medida que novos conceitos e técnicas emergissem. Noutras versões do modelo, Nolan (1978) tinha descrito uma nova curva *S* correspondente às despesas com a tecnologia de base de dados.

Ele previu que a assimilação da tecnologia de base de dados requeria quatro estádios evolucionários semelhantes aos do modelo original. De acordo com Nolan, a sobreposição dos dois padrões de quatro estádios produzia um modelo de seis estádios relativo à evolução da função *data resource*. O estádio IV original mudou de nome para *Integração* e *Administração de Dados* tornou-se no quinto estádio imediatamente antes do estádio final de *Maturidade*. Assim, em qualquer formulação, a hipótese de estádios de Nolan associava a existência de uma função formal AD com uma actividade SI perto do fim da escala da maturidade.

De acordo com Benbasat et al (1984), este ponto foi investigado por Goldstein e McCrinck em 1981, que, a partir de um questionário a 273 organizações concluíram pela sua não confirmação.

2.5. Localização da Função SI na organização

Gibson and Nolan (1974;1) tinha posto a hipótese de que no estádio Iniciação, a unidade SI seria colocada na área cujas aplicações fossem computadorizadas em primeiro lugar, e, depois, à medida que a organização atingia a maturidade o SI tornar-se-ia uma unidade autónoma separada.

> *... at the time when the investment decision and location are made, there should be clear signals of the prospect of relocation at some later date. The initial approval for the investment could well carry a specific plan for the creation of an autonomous computer and management information system after a few years... The purpose, then, is the clear understanding at the outset that the resource will become something different al a point in the future and that its initial location is within an existing department for purposes of economic justification and accountability.*

Esta posição de Nolan foi parcialmente confirmada em 1982 pelos investigadores Ein-Dar e Segev através de uma investigação suportada por 45 entrevistas, conforme é testemunhado por Benbasat et al (1984).

2.6. Aplicação de **Benchmarks**

O desenvolvimento de um plano e de uma estratégia eficaz para o processamento de dados é um processo com três passos de acordo com Nolan [1979]:

- a gestão deve determinar onde se situa a companhia em termos de evolução da função de processamento de dados e analisar as forças e fraquezas que assentam em estratégias de processamento de dados;

- deve escolher uma estratégia de processamento de dados que se enquadre bem com a estratégia de negócio;

- deve elaborar as linhas de um plano de desenvolvimento do processamento de dados para os próximos três a cinco anos.

Em 1983, uma investigação de Drury suportada por questionário a 144 organizações, referida em Benbasat et al (1984), concluiu pela não confirmação de tal hipótese.

2.7. Uso de planeamento e controle

Nolan (1979) distingue entre dois grandes grupos de técnicas de planeamento e controle relativas ao processamento de dados: (1) para gestão tecnológica nos estádio I a III e (2) para gestão do recurso dados nos estádios IV a VI.

O mesmo trabalho de Drury referido no ponto anterior terá concluído por uma confirmação parcial desta posição.

2.8. Steering Committees

Steering committees são outra técnica de planeamento e controlo advogada por Nolan (1979): eles deveriam ser constituídos em estádios avançados de maturidade.

Mais uma vez, uma investigação empírica de Drury, em 1983, suportada por um questionário a 67 organizações, concluiu pela não confirmação da hipótese de Nolan, conforme refere Benbasat et al (1984).

2.9. A efectiva evidência empírica

Poder-se-ia pensar, na sequência deste facto, que a concepção dos estádios de desenvolvimento dos sistemas de informação seria abandonada. Tal não aconteceu quer no mundo académico quer no mundo comercial. Múltiplos factos testemunham esta opinião:

- Nolan (1978) foi reeditado pela Harvard Press em 1992 numa colectânea de textos dedicada à gestão de sistemas de informação.
- A análise de estádios de desenvolvimento dos sistemas de informação passou a surgir em múltiplas obras como um método de planeamento estratégico dos SI.
- A IBM incorporou a análise de estádios em práticas de análise dos sistemas de informação.
- Múltiplos autores reutilizaram o conceito de maturidade na sua análise do SI das organizações ou da sua relação com outras tecnologias.

Grover e Goslar (1993) é um dos exemplos desta prática na sua análise da iniciação, adopção e implementação de tecnologias de telecomunicações nas organização dos Estados Unidas da América. Se o próprio título do texto

lembra imediatamente o modelo de Cash et al (1992), a sua relação com as concepções de Nolan é clara na passagem seguinte:

> *The increasing convergence of telecommunications and computing suggests that mature IS groups would tend to be proactive in evaluating and implementing telecommunications technologies. The hypotheses proposed involve a positive relationship between IS maturity and initiation, adoption, and implementation of telecommunication technologies.*

Os estádios de Nolan estão de novo presentes em Karimi et al (1996) ao examinar a resposta estratégica das empresas à globalização. Tal pode-se detectar na passagem seguinte integrada nas conclusões:

> *...the IT maturity variables (IT integration and IT organization) have impact on the change in IT investment. Thus, depending on competitive strategy, to encourage investment in IT, the IT managers must be prime movers in developing planning processes for linking information strategies to business needs and in encouraging experimentation with new technology. In order to enhance "IT organization," the IT managers should not only personally attempt to understand the business, but should also insist that other IT personnel gain a good understanding of the business and the firm. They must ensure that the structure of the IT function is aligned with the organizational structure, create an environment where user ideas are given attention in IT planning, and promote good relationships with users.*

Passados cerca de quinze anos sobre as críticas demolidoras à teoria de estádios o próprio Benbasat, em Benbasat e Zmud (1999), vem reconhecer a sua permanência e procurar uma razão para a mesma ao nível da comunicação no interior da organização:

> *Why is it that the "Stages Hypothesis" continues to be received well by practitioners, in spite of having been criticized by some academic researchers on theoretical and empirical grounds? It is because it provides a conceptual language enabling an IS manager to identify where his/her firm is positioned in a "stages" sense. This allows the manager to better grasp the current challenges facing the firm and the appropriate tactics for overcoming them, to predict what is likely to happen as the firm*

transcends to successive stages, and, most importantly, to communicate these notions to other executives.

O certo é que, apesar desta opinião limitadora sobre a hipótese dos estádios de desenvolvimento, múltiplas contribuições dos autores associados à teoria de estádios permanecem válidas para interpretar o processo de desenvolvimento dos SI e da sua gestão como vamos apresentar na secção seguinte.

3. Questões da articulação dos SI com as organizações decorrentes dos autores da teoria de estádios

O exame detalhado da contribuição dos autores referidos para a questão dos *estádios de desenvolvimento dos sistemas de informação* permite apresentar alguns contributos importantes organizados em torno de alguns temas.

A integração de dádos e sistemas aplicacionais é um problema presente desde muito cedo nas preocupações relativas ao processamento da informação.

Como vimos, o processamento de dados externos veio a revelar-se como algo necessário relativamente mais tarde embora com uma acuidade extremamente acentuada devido à sua relação com a sobrevivência da organização.

Curiosamente, as preocupações com a inovação surgiram relativamente cedo no processo de desenvolvimento dos sistemas de informação. Nolan (1979) dá-nos uma curiosa visão do problema que reproduziremos num ponto desta secção.

A consideração de dados e informação como um recurso organizacional e a necessidade do seu planeamento e gestão como outros recursos já é um problema mais recente e justificou o surgimento de disciplinas na área.

Ainda mais recentes são as discussões sobre o uso estratégico das tecnologias da informação fazendo-as deter um papel importante no processo de geração de vantagem competitivas.

Estes elementos justificam que a relação com o utilizador se tenha modificado, sendo acompanhada pelo surgimento de novas profissões e novas qualificações.

Nas secções seguintes vamos descrever mais detalhadamente estes elementos.

3.1. Processamento de dados externos

Este tema introduz uma das contribuições fundamentais de Bhabuta ao chamar a atenção para a importância de processamento de dados externos.

Inicialmente, o *processamento de dados exclusivamente internos* servia satisfatoriamente o planeamento financeiro básico.

O desenvolvimento da necessidade de previsão do futuro exigiu um *processamento ad hoc de dados externos* o qual passou a ser integrado no conjunto de serviços providenciados pela área de sistemas de informação.

O refinamento do processo de planeamento organizacional conduziu à *análise sistemática de dados externos* culminando, em termos de gestão estratégica, com a *ligação de actividades tácticas e operacionais a análise de dados externa*

É um facto que, hoje em dia, a análise de dados externos é determinante para a sobrevivência das organizações, introduzindo uma dificuldade adicional no processo de integração de aplicações.

Por outro lado, cabe referir que o processamento de dados externos está iminentemente relacionado com a capacidade da organização para o fazer o que é condicionado pelo formato que esses dados assumem e pelas formas de comunicação dos mesmos. Isto significa que a oportunidade interorganizacional também está relacionada com este tópico.

Retomaremos esta questão em outros locais deste documento, nomeadamente no Capítulo III ao abordar a questão do apoio dos SIE ao processo de tomada de decisão nas organizações.

3.2. Integração de aplicações

Para Nolan, os sistemas aplicacionais vão evoluindo de acordo com um processo de integração que culmina num sistema aplicacional que é o *reflexo dos fluxos de informação organizacionais.*

Assim ele considera os seguintes grandes momentos em termos de evolução: (1) aplicações para redução dos custos funcionais, (2) proliferação de sistemas aplicacionais, (3) actualização da documentação e reestruturação das aplicações existentes, (4) adaptação das aplicações existentes à tecnologia de

base de dados, (5) organização da integração de aplicações e (6) a integração de aplicações como reflexo dos fluxos de informação.

Por isso fornece os alicerces para a distinção operada por Galliers (1987) em termos de objectivos dos sistemas informacionais aqui tomados no sentido de aplicações: sistemas operacionais (transaccionais), sistemas de controlo de gestão e sistemas de informação estratégica.

Para Nolan é importante a integração de sistemas para garantir a coerência entre a informação que serve a gestão de topo e os níveis operacionais, isto é a informação que integra, por exemplo, um sistema de informação de apoio à decisão estratégica deve colher parte do seu conteúdo nos sistemas aplicacionais.

Para o autor existe uma relativa facilidade de integração ao nível operacional a qual vai diminuindo à medida que se consideram os sistemas de controlo de gestão e os sistemas de informação estratégica.

Esta conclusão é reforçada pelo facto de muita informação que o nível estratégico carece não provir da componente informática do sistema de informação (Davenport, 1994) e, por outro lado, integrar uma larga fatia de dados externos (Bhabuta, 1988).

Notamos quanto a este ponto a convergência Nolan, Davenport e Bhabuta que estará presente em todo este documento.

3.3. Inovação e uso estratégico dos sistemas de informação

Para Nolan, a aprendizagem organizacional é influenciada pelo ambiente onde tem lugar. Um possível ambiente pode designar-se por *controlo*, um segundo pode designar-se por *autonomia* (Nolan, 1979).

No ambiente de controlo, todos os sistemas financeiros e de gestão do desempenho - incluindo planeamento, orçamento, gestão de projectos, análise do desempenho de pessoal, e contabilidade de custos - são utilizados para assegurar que as actividades de processamento de dados são eficazes e eficientes.

No ambiente de autonomia, todavia, estão ausentes controlos sofisticados. Em vez disso o incentivo à utilização de processamento de dados de um modo experimental está presente (por exemplo, os analistas de sistemas podem ser associados aos utilizadores sem nenhum encargo para os seus orçamentos).

Quando a gestão permite a autonomia organizacional nas actividades de processamento de dados, atribui mais recursos ao processamento de dados do que os necessários para obter o trabalho concluído. O gasto adicional procura outro objectivo - *a criação de inovação.*

O equilíbrio entre controlo e autonomia é importante para o desenvolvimento de aproximações de gestão apropriadas para cada estádio da aprendizagem organizacional. Por exemplo, um desequilíbrio de alto controlo e autonomia pequena nos estádios mais iniciais pode impedir o uso de tecnologia de informação na organização; por outro lado, um desequilíbrio de pouco controlo e grande autonomia nos estádios finais pode conduzir a um crescimento explosivo no orçamento de processamento de dados e a sistemas ineficientes.

Para Nolan, a gestão tecnológica e a transição para a gestão da informação legitimam o surgimento de elementos de autonomia em estádios específicos.

Quadro 2. Equilíbrio Óptimo Controlo-Autonomia em Nolan

Estádios	Autonomia Organizacional		Controlo		Objectivo dos sistemas de controlo
	Computador	Dados	Computador	Dados	
Estádio 1	Baixo		baixo		
Estádio 2	Alto		baixo		Facilitar o crescimento
Estádio 3	baixo	Baixo	alto	Baixo	Conter a procura
Estádio 4		Alto		Baixo	Equilibrar oferta e procura
Estádio 5		Baixo		Alto	Conter a procura
Estádio 6		Alto		Alto	Equilibrar oferta e procura

Fonte: Nolan (1979)

Assim, a evolução do controlo ao longo dos estádios do SI traduz-se num processo de afirmação do planeamento que culmina em Nolan com o planeamento estratégico de dados como recurso: (1) controlo fraco, (2) controlo

ainda mais fraco, (3) planeamento formalizado e controlo (4) planeamento sistematizado e controlo, (5) dados partilhados e sistemas comuns e (6) planeamento estratégico dos dados como recurso.

Pensar a missão do SI surge muitas vezes confundido com a utilização das TI. Apesar de tudo Earl consegue traçar um quadro que faz evoluir essa categoria entre o *satisfazer pedidos* e o *garantir uma ligação estratégica ao negócio.*

Esta perspectiva é coerente com a de outros autores, como McFarlan, que, na década de 80, sustentaram a necessidade de suportar a procura de vantagem competitiva pelo uso de TI. Esta afirmação é demonstrável pela enumeração da evolução da missão em estádios de acordo com Earl: (1) prover serviço determinado através de pedidos ao responsável do SI, (2) controlar a utilização de TI através da limitação desses pedidos, (3) garantir o suporte ao negócio acordando prioridades, (4) providenciar planeamento detalhado (5) descobrir oportunidade para utilização de vantagem estratégica e (6) *garantir a ligação estratégica negócio/TI.*

Num paralelismo interessantíssimo com este autor, em Galliers e Sutherland torna-se claro um processo de amadurecimento da organização ao longo do tempo, passando sucessivamente por: (1) aquisição de *packages,* (2) descobrir e satisfazer as necessidades do utilizador, (3) pôr um plano em prática, (4) integrar e coordenar, (5) *procurar activamente oportunidades para a utilização de TI* e (6) manter a vantagem estratégica.

3.4. Gestão dos dados e da informação como um recurso da organização

Para Nolan, no início todo o processamento é centralizado à porta fechada. Com o tempo a área informática transforma-se em guarda da informação. Estabelecida a utilidade do computador é conseguida a confiança, sendo possível, então, a implementação do conceito de *gestão de dados como um recurso.* Por fim, há responsáveis pelo processamento de dados em níveis apropriados na organização.

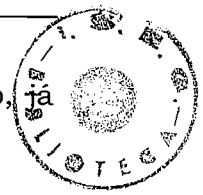
Para este autor o tipo de gestão a nível de sistema de informação passa, sucessivamente, por (1) especialização para aprendizagem tecnológica, (2) programadores orientados ao utilizador, (3) gestão intermédia, (4)

estabelecimento de utilidade do computador e de equipas de contabilização, (5) administração de dados e (6) gestão de dados como um recurso.

No estádio inicial foram automatizados alguns sistemas operacionais numa área funcional tipicamente a contabilidade.

No estádio designado por *contágio*, os níveis superiores de gestão, cada vez mais, vêem frustadas as suas tentativas para obter informação dos sistemas computacionais da empresa para suportar as necessidades de apoio à decisão. Em todos os níveis há sistemas não económicos ou de automatização inviável apesar dos desejos dos gestores por dados melhores e mais rápidos. A organização encoraja a inovação e a utilização extensiva da tecnologia de processamento de dados mantendo um controlo baixo e uma autonomia elevada. Enquanto a penetração alargada da tecnologia é conseguida pela expansão nos sistemas operacionais, os problemas são criados por programadores inexperientes que trabalham sem levar em conta sistemas de controlo de gestão eficazes. Estes problemas tornam-se alarmantes quando os sistemas de base não conseguem suportar os sistemas de mais alto nível particularmente o processamento de encomendas, o controlo de produção e os sistemas de controlo orçamental. A manutenção dos sistemas existentes, relativamente pobres em desenho começa a ocupar de 70% a 80% do tempo produtivo dos programadores e analistas de sistemas.

No estádio de *controlo*, pode observar-se uma mudança básica na orientação da gestão do computador para a gestão dos recursos de dados. Esta mudança na orientação é um resultado directo da análise de como colocar mais ênfase ao expandir as actividade de processamento de dados nas necessidades de controlo de gestão e planeamento em oposição às necessidades de consolidação e coordenação nas actividades de processamento de dados em si mesmas. Esta modificação serve também para manter o processamento de dados flexível para responder a novas questões da gestão em matéria de controlo ou ideias para planeamento. Este estádio caracteriza-se pela reconstrução e profissionalização da actividade de processamento de dados para a tornar mais estável na organização e pelas tentativas iniciais para a contabilização das despesas do utilizador em matéria informática, tentativas que, usualmente, tomando a forma de encargos pelos serviços, conduzindo,



quase sempre, à confusão e alienação sem ganhos reais, apesar de tudo, já não regridem.

Consequentemente, neste processo, os utilizadores vêem poucos progressos no desenvolvimento de novos sistemas de controlo enquanto o departamento de processamento de dados está em reconstrução embora eles sejam contabilizados de uma forma arbitrária e tenham poucos meios para influenciar os custos o que contribui para uma frustração geral. À medida que este estádio se aproxima do fim, a reconstrução do departamento de processamento de dados fica concluída e torna-se possível utilizar tecnologia de base de dados em muitas aplicações chave, tais como encomendas, contabilidade e planeamento de materiais. Além disso, a utilidade do computador e da rede atingiram um ponto que permite a prestação de serviços de alta qualidade aos utilizadores. A concretização disto marca a transição para o estádio 4 que Nolan designa por *integração.*

A chegada de terminais interactivos e vários suportes e a assistência necessária permitiram usar e tirar proveito da tecnologia de base de dados. O utilizador já aceitou o custo dos serviços de processamento de dados e, agora, percebido o seu valor real, quer mais suporte e está disposto a pagar seja lá qual for o custo. As despesas em informática crescem de um modo que se assemelha ao estádio de *contágio* quando se pensava que isto se não repetiria e surge nova onda de problemas. A redundância de dados complica o uso de sistemas de controlo e planeamento. Sobem os pedidos para melhor controlo e mais eficiência.

A administração dos dados surge no estádio 5 e Nolan aplica essa designação ao estádio.

Finalmente, durante o estádio de *maturidade,* o *portfolio* aplicacional fica completo e a sua estrutura reflecte os fluxos de informação na organização.

3.5. Relação com o utilizador

De acordo com Nolan a participação do utilizador no desenvolvimento e utilização do SI nunca ultrapassa uma participação no processamento sendo a sua evolução caracterizada como (1) *mãos de fora,* (2) entusiasmo superficial, (3) responsabilização mantida arbitrariamente, (4) aprendizagem da

responsabilização, (5) efectiva responsabilização e (6) aceitação da junção com a responsabilização pelo processamento dos dados.

Em termos mais gerais esta evolução do utilizador passa por o mesmo poder ser classificado como:

- *Reactivo*: o utilizador final é superficialmente envolvido. O computador provê mais, melhor e informação mais rápida do que as técnicas manuais.
- *Força motora*: o utilizador final está directamente envolvido com a entrada de dados e o seu uso. É responsável pela qualidade dos dados e pelo valor acrescentado final.
- *Participatório*: o utilizador final e o processamento de dados são conjuntamente responsáveis pela qualidade dos dados e pelo desenho eficaz de aplicações de valor acrescentado.

Earl também capta a evolução do utilizador para um posicionamento mais participatório que o leva inclusivamente à constituição de coligações estratégicas.

No entanto, não atribui a relevância suficiente a este critério para o considerar isoladamente; logo as referências ao utilizador surgem no interior de critérios que considera mais significativos, nomeadamente *força condutora* e *contexto*.

Consideremos, em primeiro lugar, a posição de Earl quanto ao utilizador em termos do seu critério *força condutora*. Ao atingir o *estádio 3*, que designámos por *apoio ao negócio*, ele considera progressivamente:

- a condução do processo pode ser garantida por associação utilizador/SI na sua concretização;
- executivo e SI conduzem, conseguindo o envolvimento do utilizador;
- formam-se coligações estratégicas SI / utilizador suportadas pelo nível executivo.

Ainda de acordo com Earl, o contexto em que decorre o planeamento de SI vai-se alterando de forma a poder ser considerado, no final, de *maturidade e colaboração*.

A relação com o utilizador é tratada por Galliers e Sutherland com algum paralelismo e semelhança relativamente a Earl. Inicialmente, o estilo predominante na utilização de TI é o de *ser desconhecedor* e, mais significativo, sem consciência de o ser. O sistema opera num vácuo virtual e num alheamento total em relação ao modo como afecta a organização, os seus processos e os recursos humanos. Na sua perspectiva pessoal os únicos problemas com alguma relevância serão de ordem técnica: nada mais é significativo seja lá aquilo a que diga respeito. Muito deste estilo pode ser atribuído ao uso de contratados externos. Estes tipicamente mostram pouco interesse pela organização por que são contratados (eles não estarão ali muito tempo e o seu desenvolvimento no futuro não depende da organização ou da sua gestão).

No estádio *início das fundações*, o estilo predominante é não incomode (*estou demasiado ocupado na instalação deste novo sistema neste momento*). A pressão é real e os recursos humanos da área de sistemas de informação mostram-no. A sua orientação ainda é técnica. Assumem que o quer que estejam a fazer é o que devem fazer para servir a organização. O seu esforço centra-se em construir o sistema tão rapidamente e tão tecnicamente correcto quanto possível. O envolvimento com outros recursos especialmente na construção de sistemas não é bem-vindo porque os utilizadores *mudam de ideias acerca do que pretendem.* Por outras palavras, não apreciam a natureza mutável das necessidades em informação neste estádio.

O estilo predominante no estádio de *ditadura centralizada* é de abrogação (ou, pelo menos delegação) de responsabilidade, a partir do departamento informático para outras pessoas na organização vulgarmente o utilizador final. O ponto de vista assumido é que o último pode fazer o que quiser quanto a aquisição de equipamento informático ou desenvolvimento de *software* desde que o pague. Qualquer falha ou erro do sistema é um problema do utilizador. Similarmente a gestão de processamento de dados procurará na gestão sénior direcção, concordância e guia para novos desenvolvimentos.

O modo de trabalho do estádio anterior (defensivo) evolui para a cooperação e a colaboração. O responsável de tecnologias de informação é deliberadamente escolhido como uma pessoa que pode assegurar que o departamento trabalha em conjunção e para benefício do resto da organização. Uma das tarefas mais

importantes que lhe são confiadas é instilar o sentido de cooperação através da organização.

No estádio de *oportunidade interempresarial*, a ideia fundamental baseia-se na necessidade de sistemas de informação que conduzam a uma vantagem estratégica para a organização. Tais sistemas são muito difíceis de justificar numa base de análise de custo-benefício tradicional. Eles requerem o suporte dos elementos poderosos da organização para assegurar que são implementados (e mesmo assim correm o risco de ser parados a meio do desenvolvimento).

Finalmente, o estilo é de interdependência, sendo os elementos de TI apenas uma parte da equipa de negócio, trabalhando conjuntamente para criar e manter a organização em sucesso.

3.6. Novas funções

Os recursos humanos resumem-se, inicialmente, a um pequeno número de programadores. Alguns podem ser empregados internos, mas muitas vezes são utilizados contratados externos. A compra de programas-produto significa que poucas pessoas serão necessárias internamente.

O *início das fundações* marca o aparecimento de um gestor informático que vulgarmente reporta ao director financeiro ou equivalente. Para além dos programadores provenientes do estádio 1, surgirão analistas de sistemas: pessoas encarregadas de se assegurar de uma adequada compreensão dos requisitos do utilizador e do desenho de sistemas apropriados.

No estádio de *ditadura centralizada* o departamento informático não só retém os anteriores elementos (programadores e analistas de sistemas) como também lhes junta elementos para o planeamento de sistemas de informação e administração de base de dados. Na zona final do estádio o gestor informático pode mudar de designação para gestor de sistemas de informação ou gestor de tecnologias de informação, modificação de designação que pode acompanhar a do departamento.

Ao tradicional *staff* de analistas e programadores junta-se o Analista do Negócio no estádio de *dialéctica democrática e cooperação*. Estes recursos são utilizados pelos departamentos linha que servem, mas interagem muito

fortemente com os restantes recursos do departamento de processamento de dados. Surge um gestor de alto nível usualmente ao nível do vice-presidente.

A *oportunidade interempresarial* traz consigo uma nova função que *combina o planeador do negócio e dos sistemas de informação.* Estas pessoas são responsáveis por reconhecer e planear sistemas de informação estratégicos para a organização como um todo e para as unidades do negócio individuais. Eles tiveram alguns anos de experiência quer no negócio quer na área de tecnologias de informação. No fundo, podem ter vindo de qualquer área mas definitivamente a sua disciplinaridade é cruzada, multifacetada.

Finalmente, o responsável de TI será um membro do quadro de directores tomando parte activa na definição de direcções estratégicas. As decisões estratégicas terão assim desde o início parecer deste elemento mais do que após terem sido tomadas.

3.7. Novas qualificações

Na *adocracia,* as habilitações associadas tendem a ser de natureza técnica e baixo nível. O acento é verdadeiramente na tecnologia em oposição a questões organizacionais do negócio ou informacionais. As únicas habilitações a que o pessoal do utilizador tem acesso referem-se muito especificamente a aplicações particulares quer seja um *package* ou um desenvolvimento à medida. Os computadores e aplicações tendem a ser tão fechados que o pessoal não informático sentirá que é extremamente difícil conseguir as condições necessárias para usar os poucos sistemas que existem. O treino providenciado pela organização praticamente não existe.

Com o *início das fundações,* para além da programação e instalação de equipamento computacional, os recursos da área informática reúnem competências relacionadas com a concepção, desenvolvimento e instalação de sistemas completos. Assim, os conhecimentos de metodologias de desenvolvimento, técnicas estruturadas e relacionados tornam-se importantes.

A maior capacidade demonstrada no estádio de *ditadura centralizada* é a gestão de projectos. Os projectos que forem suportados centralmente serão normalmente bem controlados, seguindo estritamente as linhas da gestão de

projectos. A ênfase é posta no assegurar que os sistemas que estão a ser construídos o são em tempo e dentro do orçamento.

No estádio seguinte, os recursos a nível de pessoal informático mudam radicalmente. Embora a capacidade técnica seja ainda requerida ela perde importância ênfase relativamente ao conhecimento do negócio e à necessidade de uma boa articulação com o resto da organização. O tema mais importante é a integração organizacional.

O estádio de *oportunidade interempresarial* é aquele em que a área de tecnologias de informação deixa de ser um serviço menor e uma unidade de suporte para ser uma parte integrante da operação com sucesso da organização. O perfil exigido para gerir esta transição é a do executivo sénior. Capacidades interempresariais e de marketing são também os requisitos básicos para assegurar o sucesso neste estádio. Utilizadores de TI muito sabedores tornam-se um lugar comum. As organizações de sucesso utilizam estas pessoas no seu potencial total, porque não existe qualquer reacção defensiva relativamente a departamentos utilizadores conhecedores de tecnologias da informação em processo.

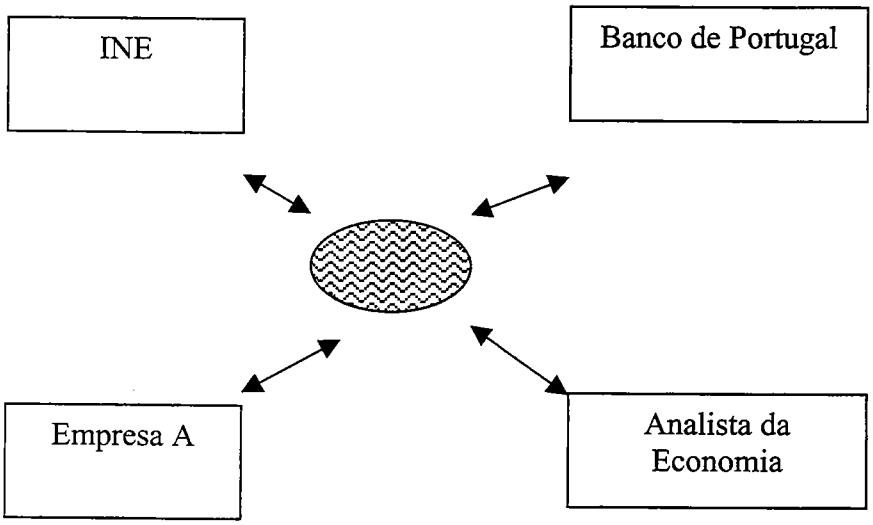
Todas as capacidades requeridas de um membro do quadro juntamente com as de um gestor sénior que compreende as potencialidades das tecnologias de informação tal como o negócio são necessárias no estádio de *relações harmoniosas e integradas*. O pessoal de TI estará bastante mais sintonizado com as necessidades e aspirações das unidades estratégicas de negócio com as quais trabalham. O tipo de necessidades passa por questões educacionais, de credibilidade, de estratégia e de relacionamento.

4. Era das redes, sociedade da informação, economia digital, economia do conhecimento

O título desta secção utiliza diferentes designações para aspectos de uma realidade que têm muito em comum. Manteremos o termo *era das redes* já que deriva de Nolan e Croson (1995), elemento impulsionador deste documento. Em devido momento estabeleceremos o seu relacionamento com a economia do conhecimento.

A era das redes acompanha o dia a dia do processo de trabalho das organizações no início do terceiro milénio. Portugal não é uma excepção. O desenvolvimento do comércio electrónico, da economia digital é apenas um aspecto, talvez o de maior impacto visual associado ao novo paradigma. Mas outros aspectos vão sendo modificados e deles não podemos excluir o modo como os analistas e as organizações se posicionam perante a Economia. Cada vez mais uns e outros fazem parte de um extensa teia onde a troca de dados e o seu processamento local vão detendo uma importância acrescida. A figura seguinte, com uma intenção meramente ilustrativa, permite visualizar esta realidade.

Figura 5. As trocas de dados na Economia



Apesar da sua simplicidade e do seu carácter ilustrativo, algumas conclusões podem ser tiradas perante esta figura:

- Sendo disponibilizados dados sobre um determinado tema económico, se o sistema funcionar, qualquer analista da economia lhe pode ter acesso;

- as prestações electrónicas de dados não são apenas num sentido, sendo de admitir que as comunicações de dados para o INE, o Banco de Portugal e outros serviços generalizem a utilização de redes;

- a extracção de dados dos serviços que os difundem já não necessita do intermediário local, abolindo eventuais focos de poder suscitados pelos esconderijos de informação.

Como nenhuma era tem as suas características num estado puro, é de admitir que os primeiro sinais da era das redes tenham surgido em momentos anteriores e que ela própria apresente alguns resquícios do passado e integre elementos de uma potencial era futura.

No entanto, não devemos ignorar uma substancial mudança que o desenvolvimento económico vai fazendo afirmar-se: na linguagem dos economistas, deparamos com uma nova fonte de criação de riqueza, com um factor de produção mais importante: o conhecimento (Shwartz *et al*, 2001).

Conhecimento significa como devemos fazer as coisas e como poderemos fazê-las melhor (Davenport e Prusak, 1998). Saber quem, saber como, saber o quê e saber porquê tornam-se perspectivas determinantes de novas formas económicas em que as Tecnologias de Informação desempenharão com certeza um papel extremamente importante.

Igualmente relevante é o facto de uma economia baseada no conhecimento não estar sujeita a alguns preconceitos. Com efeito o conhecimento é abundante, não é escasso. Por outro lado, contrariamente a formulações que acompanharam a humanidade durante décadas, o conhecimento gera rendimentos crescentes. A aplicação de conhecimento gera mais conhecimento susceptível de aplicação e esta espiral parece não ter fim.

Qual o papel e quais as preocupações que os sistemas de informação poderão assumir na sociedade do conhecimento que se avizinha? Em primeiro lugar, os sistemas de informação poderão participar em processos de criação de conhecimento: o nosso documento descreve várias formas de descoberta de conhecimento em bases de dados, uma actividade crítica da nova sociedade. Por outro lado, podem assumir um papel activo que colabore na afirmação desse tipo de sociedade, um papel que pode passar por um conjunto de estádios naturalmente diferentes dos postulados para a era de processamento de dados. Esse é o objecto essencial deste documento, mas, para já, vamos procurar de que modo, actualmente, se afirmam o que considerámos os seus elementos relevantes.

A Integração de Aplicações mantém toda a sua importância no mundo actual. Ela, no fim de contas, viabiliza outros atributos que aparentemente são mais importantes. Não há informação robusta para os níveis superiores de gestão se

ela não estiver garantida. Naturalmente, outros aspectos de integração vão sendo relevantes e teremos ocasião de chamar a atenção para o problema que designamos por integridade temporal e para a própria actualidade da informação.

O Processamento de Dados Externos é outra das características marcantes da nova era o que significa que pode ser considerado um embrião do futuro quando surgiu na era de processamento de dados o que, quase quinze anos depois, atribui enorme importância às concepções de Bhahuta.

Mais do que no passado, a Inovação tem enorme importância nas organizações actuais e a era das redes segue as potencialidades de criação de valor utilizando facilidades de informação. O conceito Inovação deve agora estender-se aos serviços que disponibilizam dados para as empresas fazendo com que estes modifiquem o seu tradicional modo de inserção no processo de produção.

Naturalmente, a pressão sobre a gestão dos dados e da informação como um recurso da organização manter-se-á e as organizações que não procederem de acordo terão sérias dificuldades em subsistir. O processo tenderá agora a envolver os dados externos fundamentais para a criação de bases de dados analíticas. É de admitir, no entanto, a existência de acções de pedidos de informação nos serviços que a disponibilizam relativamente a elementos ou resultados que antes não eram considerados e com um grau de variação elevado. Assim, as organizações que disponibilizam informação sobre a Economia deverão tender para conciliar a resposta a estes pedidos com a necessidade de alguma estabilidade em ordem a poderem satisfazer a procura habitual.

A nova era manterá e desenvolverá as oportunidades para o uso estratégico das tecnologias da informação podendo ter efeitos que alterem a pressão de fornecedores e clientes e o modo como se concorre. Com a globalização os efeitos temporais de ganhos de competitividade tenderão a atenuar-se, gerando uma bola de neve de inovação / utilização estratégica.

Marcante será a relação com o utilizador, que ao nível dos sistemas de informação para a Economia, será o analista ou a empresa no seu conjunto. Este já é mais do que participatório e os fornecedores de informação deverão

procurar interiorizar as suas necessidades e apoiar os seus processos de extracção de dados e informação.

Naturalmente, tudo será acompanhado por novas funções e novas qualificações suscitadas pela *economia em rede* e pela disponibilização interactiva de informação.

Profissionais do comércio electrónico, cuja missão é dominar as tecnologias e práticas nessa área, nomeadamente relativamente ao desenvolvimento de transações *on-line* utilizando *sites* seguros, de acordo com normas actualizadas, *standards* de criptografia e emissão de *standards* de autenticidade, verão o seu peso crescer devido às pressões na sua área a que não será estranho a disponibilização de dados como um produto ou um serviço. O mesmo acontecerá com *desenhadores de sites* e *administradores de servidores*. Concretamente, estes terão por missão dominar os procedimentos de configuração, manutenção e optimização de estruturas de suporte à disponibilização de informação. Finalmente, será crucial o papel dos *administradores de redes* cuja vocação essencial será o dominar as arquitecturas de rede e identificar componentes da infra-estrutura, bem como monitorar e analisar o desempenho da rede.

5. Novos elementos para análise da maturidade dos sistemas de informação

A problemática de maturidade descrita em secções anteriores, apesar de permanecer, encontrou os seus fundamentos na era do processamento de dados. Actualmente, há que pensar se novos elementos para análise daquela questão poderão ser encontrados em íntima relação com os sinais de uma nova economia.

A década de noventa foi relativamente fértil nessa área o que nos permitiu encontrar quatro nove elementos de maturidade que passaremos a examinar em pontos específicos, designadamente:

- guias e roteiros para a informação;
- a integridade temporal;
- a multidimensionalidade e
- a descoberta de conhecimento em bases de dados.

5.1. Guias e roteiros para a informação

Davenport (1994) apresenta algumas soluções específicas para a transformação da cultura de informação na organização, designadamente ao nível de:

- introdução de roteiros de informação (descrição da localização e disponibilidade de informação usada mais largamente) que incluirão pessoas que a possuem ou supervisionam e que poderão interpretá-la, descrever o seu uso e limitações e dirigir para outras fontes e
- desenvolvimento de guias para a informação (pessoas funcionando como guias gerais para os recursos em informação disponíveis numa organização o que se traduz em redução do tempo de acesso à mesma pela generalidade dos empregados).

O desenvolvimento posterior do pensamento deste autor leva-o à criação e desenvolvimento do conceito de Gestão do Conhecimento (*Knowledge Management*).

5.2. Integridade temporal

Inmon tem uma definição muito exaustiva relativamente ao conceito de base de dados analítica (BDA) que denominou *Data Warehouse*. Uma BDA é *uma colecção de dados orientada para os assuntos do negócio da empresa, integrada, variável com o tempo e não volátil para suporte dos gestores, nos processos de tomada de decisão* (Inmon, 1994).

A orientação para os assuntos da empresa significa que os dados são organizados em torno dos assuntos da empresa, em vez de continuarem acoplados às aplicações. Assuntos típicos das organizações são os clientes, produtos, postos de venda, etc.

Ora este conceito pode ser transposto para os factos económicos materializados nas tendências de evolução relativas aos principais produtos ou a elementos demográficos e sociais. Por exemplo, o Instituto Nacional de Estatística considera mais de vinte áreas temáticas na organização da informação que disponibiliza sobre a Economia.

Por outro lado, é visível no autor a necessidade de garantir a integridade temporal dos dados uma característica fundamental dos tempos futuros.

5.3. Multidimensionalidade

Para Codd et al (1993), OLAP (On-line Analytical Processing) é o nome dado à análise dinâmica necessária para criar, manipular, animar e sintetizar informação com modelos de análise de dados, onde deverá ser possível:

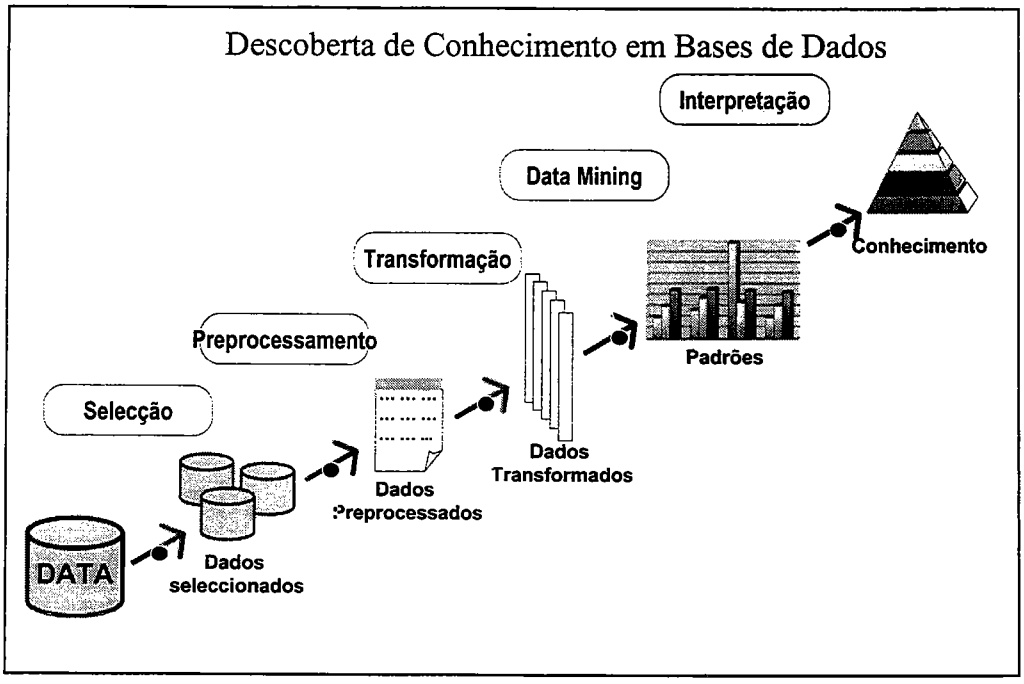
- projectar cenários, através da manipulação de grandes quantidades de dados com parâmetros diferentes;
- utilizar um número ilimitado de dimensões e caminhos de consolidação;
- especificar condições e expressões entre dimensões;
- discernir novas ou inantecipadas relações entre variáveis.

5.4. Descoberta de conhecimento em bases de dados

Fayadd et al (1996) descreve o processo de descoberta de conhecimento em bases de dados, designado em termos de linguagem original por *knowledge discovery in databases* (*KDD*) como um conjunto sequencial de actividades: selecção de dados, preprocessamento dos dados, transformação do dados, *data mining* e interpretação dos resultados.

A figura seguinte explicita este processo e mostra que, no final de cada uma das actividades consideradas, se encontra um resultado que é utilizado na fase seguinte.

Figura 6. O processo da descoberta de conhecimento



Fonte: Fayyad et al (1996)

Mantivemos a expressão *data mining*, específica da fase em que são aplicados algoritmos de descoberta aos dados transformados, pois explicita melhor do que qualquer expressão portuguesa o seu objectivo.

A manutenção da expressão *data mining* no processo do KDD permite ainda considerar a actividade assim designada como um dos seus subprocessos, esclarecendo antecipadamente uma posição que, como veremos ao longo da dissertação, não é comum a todos os que tratam o tema – por exemplo, Saporta (2000), que apresentamos no ponto 5. do capítulo IV.

Apesar de tudo, já encontrámos em múltiplos documentos uma designação significativamente expressiva, *mineração de dados*, que por vezes utilizaremos.

6. Contribuições para a análise da maturidade dos SI na era das redes

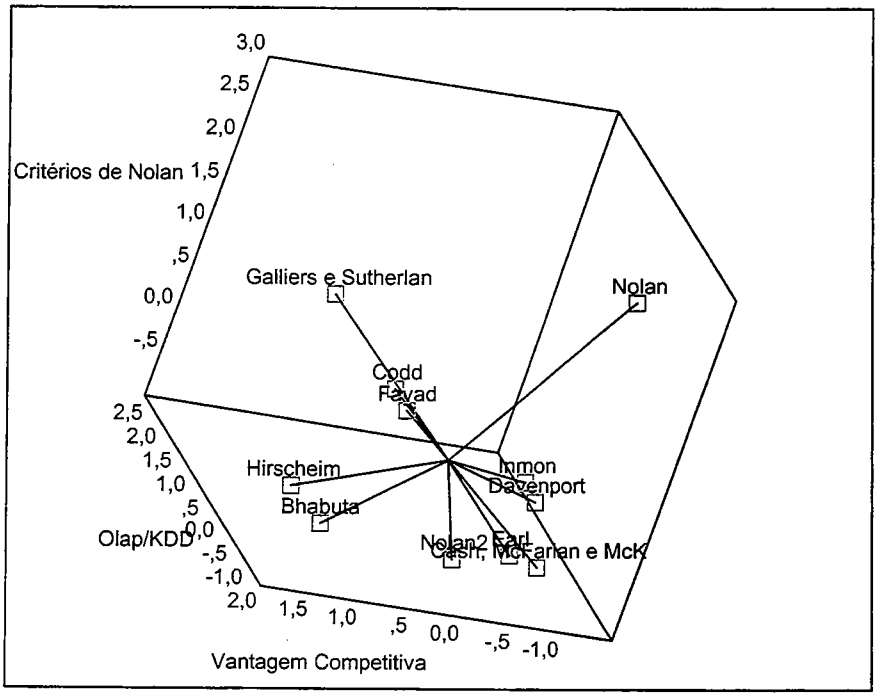
Aqui chegados, parece-nos reunido todo o material que pode apoiar a elaboração de uma teoria de estádios relacionados com o desenvolvimento dos SIE.

Nas páginas anteriores, apresentámos a concepção original dos estádios de desenvolvimento dos SI, analisámos a posição dos seus adversários, vimos a linha de desenvolvimento da argumentação dos seus seguidores e procurámos todo um conjunto adicional de elementos que nos permitisse uma integração no que Nolan designou por era das redes.

Na figura seguinte, apresentamos uma visão de conjunto sobre este percurso que espelha alguma diferenças entre as posições dos autores em termos de três temas fundamentais:

- os critérios originais de Nolan;
- os tópicos avançados de bases de dados materializados na análise multidimensional e na descoberta de conhecimento em bases de dados e
- a questão da vantagem estratégica.

Figura 7. Uma visão conjunta das contribuições para a maturidade dos SI



Nota: as escalas não têm significado neste contexto, existindo por necessidade de elaboração e utilização de uma Análise de Componentes Principais, o meio técnico para chegar à figura

A figura mostra as diferenças entre os múltiplos autores, embora da mesma tenhamos retirado as posições dos críticos de Nolan e acrescentado a nova visão deste autor suportada em textos mais correntes.

Como é natural, em termos de utilização dos critérios de Nolan é ele quem apresenta a maior contribuição, a qual é representada por uma linha iniciada na origem e que vai até *Nolan.*

Em termos de discussão sobre a vantagem competitiva a qual dominou algumas polémicas da década de 80 e encontra uma síntese interessante em Lederer e Mendelow (1988), encontramos em Hirscheim e Bhabuta as posições mais consequentes.

Finalmente, Codd e Fayyad trazem-nos uma contribuição importante em termos de tópicos avançados de bases de dados a qual será determinante em termos de julgamento de maturidade dos SI na época actual.

Na figura, manifesta-se ainda a capacidade de formulação holística de Sutherland e Galliers visível na sua situação entre os melhores contribuintes para os dois tópicos iniciais.

Resumo do Capítulo I

Neste capítulo apresentámos algumas formulações utilizadas na análise do processo de maturação dos sistemas de informação no que designámos por era do processamento de dados.

Tais formulações manifestam-se insuficientes para compreendermos idêntico processo numa sociedade baseada no conhecimento, viabilizada parcialmente pelas tecnologias de informação e da comunicação, onde a informação externa às organizações da qual a disponibilizada pelos sistemas de informação para a economia é parte importante, assume um papel crucial.

É necessário integrar os dados externos nos processos de apoio à decisão suportados pelas tecnologias de informação no interior das organizações que lutam pela sobrevivência. Após a sua integração, eles poderão apoiar processos de pesquisa operacional, análise multidimensional e descoberta de conhecimento em bases de dados sendo portanto necessário garantir a sua compatibilização com outros dados já residentes no sistema informático.

Por outro lado, admite-se que a mesma informação seja disponibilizada através de meios que possibilitem imediatamente os resultados anteriores.

As novas características da informação, os novos processos que visa suportar, o formato e os meios utilizados para a sua disponibilização justificam que se pense em todo um novo conjunto de critérios que permita analisar a maturação das organizações que disponibilizam informação para a economia no contexto de uma sociedade baseada no conhecimento.

Um novo quadro de requisitos, formulado a partir de preocupações antigas (processamento de dados externos, integração de aplicações, inovação e uso estratégico da informação, gestão dos dados como recurso) e de novas preocupações (foco no utilizador, multidimensionalidade, integridade temporal, descoberta de conhecimento em bases de dados), vai-se afirmando e justifica uma nova grelha para a avaliação dos SI, novas métricas e nova concepção de estádios.

Capítulo II - Processamento Analítico

Os ficheiros tradicionais, as folhas de cálculo e as bases de dados constituem os suportes essenciais ao processamento analítico. Qualquer destas estruturas é relativamente conhecida. Se um ficheiro tradicional se traduz numa mera colecção de registos, uma folha de cálculo pela sua capacidade de relacionamento de células é uma forma superior e mais complexa.

No entanto, pela diversidade, pela problemática subjacente e até por novas tendências subjacentes à utilização e desenho, as bases de dados merecem uma referência mais demorada. Com efeito, defendemos que esta estrutura é a mais indicada como repositório de dados para processos analíticos e até para o processamento estatístico em geral. Tal não significa que não devam ser utilizadas folhas de cálculo ou ficheiros de texto, mas sim que os dados que vão ser usado no processamento devem residir em bases de dados o que lhes garante um nível de qualidade e segurança superiores materializados em integridade referencial e facilidades de interrogação.

Neste capítulo, vamos examinar alguns dos traços fundamentais do processamento analítico através de uma revisão da descrição de processos analíticos utilizados na dissertação. Este exame não é formulado em termos de revisão da literatura, pois os processos não são objecto de discussão. Eles são, antes, um meio para aplicação de ferramentas em diferentes pontos, pelo que se torna válida qualquer descrição suportada por algum texto proveniente de uma entidade idónea.

Por outro lado, é de referir que a utilização de processos se faz com dois objectivos diferentes, como ressalta do que já foi enunciado na Introdução. Num primeiro momento, há que examinar a viabilidade de aplicação de processos analíticos à informação proveniente de Sistemas de Informação para a Economia. Na fase final, é possível aplicar esses ou outros processos analíticos às medidas decorrentes para os SIE e que visam caracterizá-los em função do modo como se articulam com o problema. Neste contexto, pode dizer-se que, no presente trabalho, o processamento analítico é utilizado como ferramenta de análise dos SIE e como objecto de análise.

O capítulo é, assim, elaborado numa perspectiva de apresentação de alguns conceitos e métodos utilizados na dissertação, para o que foi estruturado em três secções principais e uma secção complementar. Na primeira, são apresentados os fundamentos da utilização de bases de dados analíticas. Na segunda secção, descreve-se o processo de extracção de informação de bases de dados consubstanciado em interrogação (a perspectiva mais conhecida) e análise multidimensional (uma perspectiva mais recente). A terceira secção dedica-se a aspectos que, nos últimos anos, têm sentido um substancial desenvolvimento, descrevendo algumas das modalidades de que se reveste a descoberta de conhecimento em bases de dados. Na quarta e última secção, propomo-nos proceder a uma apresentação superficial de dois processos adicionais que, não fazendo parte dos que inicialmente integrámos no processamento analítico, são indispensáveis para alcançar um melhor conhecimento dos SIE, designadamente, a análise factorial de componentes principais e a análise discriminante.

1. Bases de Dados Analíticas

Nesta secção, vamos rever alguns aspectos associados à concepção e desenho de bases de dados, vamos introduzir a noção de bases de dados analíticas e terminamos com uma referência à questão da sua especificação aproveitando a enorme contribuição da abordagem relacional. Uma apresentação completa desta abordagem encontra-se em McFadden (1994). Omitimos referências ao processo de modelação conceptual dada a enorme ligação das bases de dados analíticas a modelos lógicos e à sua construção a partir de sistemas operacionais. Este tipo de modelação deriva de Chen (1976) e baseia-se no desenho do chamado Modelo Entidade Associação (MEA). Este autor estaria com certeza muito longe das implicações e influência que iria ter na área de sistemas de informação quando escreveu:

> *A data model, called the entity-relationship model, is proposed. This model incorporates some of the important semantic information about the real world. A special diagrammatic technique is introduced as a tool for database design. An example of database design and description using the model and the diagrammatic technique is given. Some implications for data integrity, information retrieval, and data manipulation are discussed. The entity-relationship model can be used as a basis for*

unification of different views of data: the network model, the relational model, and the entity set model. Semantic ambiguities in these models are analyzed. Possible ways to derive their views of data from the entity-relationship model are presented.

O problema da derivação do modelo relacional a partir do modelo entidade associação é abordado em McFadden (1994). Apesar de relacionado, esse problema não se integra nos nossos objectivos de tratamento, embora seja aceitável a derivação de um modelo lógico de uma bases de dados analítica a partir de um MEA questão com que fechamos a secção.

1.1. Bases de dados

O termo de *Base de Dados* é usado, hoje em dia, com diversos significados distintos. Informalmente, este conceito é identificado com *uma colecção de dados armazenados em computador* o que não implica qualquer asserção sobre a tecnologia usada ou sobre as opções técnicas quanto à respectiva definição.

Outra utilização frequente do termo Base de Dados é no sentido de *serviço de informação*, isto é uma entidade à qual se pode recorrer para obtenção de um certo tipo de informação (por exemplo comercial, bibliográfica, científica) normalmente de forma remota e através do recurso a uma rede de computadores. Este é porventura a utilização mais frequente do termo, por exemplo, através dos meios de comunicação social.

Num sentido mais técnico, o termo Base de Dados foi usado em Codd (1970) de forma intimamente relacionada com a noção de independência de dados e no contexto da sua proposta da abordagem relacional. Tal como Nolan (1979) é um documento notável em matéria de maturidade dos sistemas de informação, esta referência apresenta à entrada do século XXI, passados trinta anos sobre a sua primeira versão, uma actualidade interessante. O documento está dividido em duas partes: "Relational Model and Normal Form" e "Redundancy and Consistency". A primeira parte inclui uma discussão acerca de questões relacionadas com a independência dos dados e com a normalização e define relações e domínios, para Codd os objectos relacionais essenciais. A segunda parte introduz os operadores relacionais e discute questões relacionadas com a redundância e consistência numa base de dados

relacional. A posição de Codd quanto à redundância virá a ser revista quando, no âmbito de sistemas destinados à obtenção de informação para os níveis táctico e estratégico, apresentar mais uma contribuição revolucionária para a área dos sistemas de informação: o conceito de *Online Analytical Processing*, constituindo com Inmon e Davenport as bases fundamentais para a afirmação da análise multidimensional na sua perspectiva mais integral.

A maior parte das aplicações informáticas existentes, mesmo hoje em dia, é dependente dos dados respectivos (Date, 1996). Isto significa que o conhecimento sobre a organização física em memória secundária dos dados está embebido nos programas de aplicação que manipulam os mesmos dados. Sempre que houver uma modificação nos dados, os programas que os utilizam deverão ser modificados.

Esta situação de dependência entre dados e programas afecta obviamente de forma negativa a complexidade e a capacidade de reformulação destes últimos face a alterações na estrutura dos dados subjacentes. Mais, não viabiliza uma gestão desses mesmos dados enquanto entidade independente dos programas de aplicação que deles fazem uso. O resultado desta situação é materializada num impacto muito negativo na eficiência do processo de produção de programas novos e de reutilização de programas previamente desenvolvidos.

Entende-se por *base de dados física* o conjunto de ficheiros, índices e outras estruturas de armazenamento destinadas a possibilitar o acesso eficiente aos dados armazenados.

Define-se *base de dados lógica* uma representação abstracta dos dados armazenados que não envolve os conceitos usados para descrever a base de dados física.

A *independência física* é definida como a separação entre a base de dados física e a base de dados lógica no sentido de possibilitar que a manipulação desta última não requeira um conhecimento sobre a organização da primeira.

Um outro conceito associado à independência de dados é o designado por *independência lógica.* Os programas de aplicação ou os utilizadores finais, usando uma linguagem de manipulação, acedem a uma vista (*view*) sobre a base de dados lógica. Assim, definem-se, em geral, um conjunto de vistas que

caracterizam a perspectiva dos vários programas de aplicação e/ou grupos de utilizadores sobre a base de dados lógica. A independência lógica traduz o grau de separação entre esta última e as diversas vistas que sobre ela forem definidas.

Assim, no sentido mais técnico define-se *base de dados* como um conjunto de dados armazenado em computador de forma a respeitar a independência física e lógica.

1.2. Características essenciais das BDAs

Uma *base de dados analítica* será uma base de dados que apoia o processamento analítico para o que integra física ou logicamente métricas, dimensões e múltiplos níveis de consolidação.

As *Métricas* são as medidas, normalmente numéricas, que constituem o foco das pesquisas do analista. São as variáveis que são relevantes para o estudo da economia (o PIB, o VAB, o IDE) ou, numa empresa, que têm impacto no negócio como vendas, lucros, custos, preços, orçamentos, inventários ou quotas de mercado.

Os eixos do modelo constituem as *Dimensões*, dado que as métricas não fazem sentido sem perspectivas de análise. *Total de investimento* só tem significado quando lhe adicionamos a perspectiva por investidor, por tipo de operação ou ainda ao longo de um período de tempo. As perspectivas, ou dimensões, fornecem contexto às métricas em análise.

A informação desejada por dimensão deverá ainda poder ser visualizada com diferentes *Níveis de Detalhe*. Estes níveis de detalhe, ou qualificadores de métricas, são designados de atributos ou elementos. Por exemplo, os investimentos deverão poder ser visualizados por país, ou por comunidade. Neste processo de síntese, deverão ser definidos caminhos de consolidação para a informação.

Ao conjunto das métricas, dimensões e níveis de consolidação chamamos *CUBO*. Na realidade, como podemos funcionar com mais do que três dimensões, não se trata de um cubo mas de um *Hipercubo*.

1.3. Outras características das BDAs

Para Inmon, uma BDA, na sua terminologia *Data Warehouse, é* uma colecção de dados orientada para os assuntos do negócio da empresa, integrada, variável com o tempo e não volátil para suporte dos gestores, nos processos de tomada de decisão (Inmon, 1994).

A *orientação para os assuntos da empresa* significa que os dados são organizados em torno dos assuntos da empresa, em vez de continuarem acoplados às aplicações. Assuntos típicos das organizações são os clientes, produtos, postos de venda, etc. Naturalmente, é possível fazer uma extensão à visão de Inmon e aplicar o conceito a nível de Economia onde o desemprego, o investimento directo estrangeiro, o comércio internacional, as contas nacionais constituirão alguns dos temas centrais para a referida orientação.

A *integração* dos dados *é* obtida numa BDA em vários níveis - consistência nos formatos, consistência nos nomes, consistência nas descrições, consistência nas unidades de medida, etc. Perspectivam-se aqui alguns problemas à integração de dados externos pois as suas características (por exemplo, o formato) poderão ser distintas das dos dados internos e não existir um conversor aplicável.

A *variabilidade no tempo* é outra característica distintiva. Ao nível operacional, as atenções estão concentradas nos dados correntes de uma organização. Isto significa que os dados operacionais reflectem sempre a situação actual como, por exemplo, os stocks actuais, o balancete actual, etc. Mas uma organização também necessita dos dados que lhe permitam analisar tendências ao longo de um período, não detectáveis na simples análise dos dados actuais. Uma BDA deve armazenar e disponibilizar a informação histórica, pelo que os dados apenas fazem sentido quando contextualizados com a data da sua criação. As modificações provenientes dos sistemas operacionais darão sempre origem a novos registos.

A *não volatilidade dos dados* termina esta lista de características distintivas. Os dados operacionais são constantemente modificados pelas transacções das aplicações operacionais, de forma a reflectirem sempre a realidade actual. Este tipo de processamento *é,* como vimos, geralmente designado de OLTP. Numa BDA apenas existem 2 tipos de operações: o carregamento periódico de dados

e o acesso de leitura para análise, envolvendo uma larga quantidade de registo que viabiliza o processamento analítico (OLAP).

Assim a BDA nunca será modificada directamente pelas aplicações operacionais, sendo o seu carregamento efectuado periodicamente.

1.4. Estrutura de uma Base de Dados Analítica

Uma BDA integra quatro componentes principais que quase sempre existem: os dados correntes detalhados, os dados históricos detalhados, os dados sumariados e o Metadados.

Os *dados correntes detalhados* constituem a entidade central de uma BDA. É aqui que residem os dados operacionais mais recentes e, geralmente, mais utilizados, após o já referido processo de extracção, onde são transformados e integrados (convertidos, filtrados, modificados estruturalmente, etc.). Esta estrutura contém os dados mais detalhados, sendo o horizonte temporal da sua existência dependente dos objectivos de utilização de cada organização.

Algumas questões de economia são importantes relativamente aos *dados históricos detalhados*. O armazenamento dos dados históricos constitui um dos principais objectivos de uma BDA. Contudo, nem todos os dados necessitam de ter o mesmo nível de disponibilidade. Os dados mais antigos, consequentemente menos utilizados, poderão ser armazenados em meios de maior capacidade de armazenamento e menor custo do que os necessários para o arquivo dos dados recentes.

A pesquisa justifica a existência de *dados sumariados*. Embora um dos objectivos de uma *BDA* seja o de permitir o acesso e pesquisa de informação relevante nos dados elementares das organizações, muitas das pesquisas apenas necessitam de dados com níveis superiores de agregação. Assim, se as exigências do negócio estão concentradas nos assuntos de Marketing, e, consequentemente, na adaptação de produtos aos clientes, será provável que as pesquisas incidam nos dados detalhados da empresa. No entanto, se a pesquisa for orientada para o estabelecimento de objectivos estratégicos de negócio, provavelmente o acesso a dados sumariados será suficiente. A agregação ou particionamento dos dados permite obter ganhos de eficiência em muitas das pesquisas pretendidas pelos gestores de nível superior.

Metadados é o termo utilizado para denominar o elemento arquitectural que descreve a estrutura e o conteúdo de uma BDA. As descrições existentes no Metadados podem ser agrupadas em 2 vertentes fundamentais:

Operacional - Descrição do processo de extracção e carregamento dos dados operacionais. Esta vertente descreve principalmente os aspectos físicos da estrutura.

Informacional - Descrição dos dados existentes do seu agrupamento multidimensional. Esta vertente descreve a estrutura em termos do negócio.

Sem Metadados, o utilizador não conseguiria interagir com os dados existentes, dada a impossibilidade em conhecer o modelo de dados, a estrutura física, as definições dos dados, as relações entre os dados, a origem dos dados e as suas datas de criação e o processo de extracção e transformação.

Geralmente equiparado aos dicionários de dados dos ambientes operacionais, o Metadados tem uma importância muito maior no ambiente informacional. Os dicionários de dados operacionais são utilizados quase exclusivamente pelos administradores de bases de dados. Numa BDA, o Metadados é parte integrante do processo de pesquisa realizado pelos utilizadores, permitindo-lhes determinar os dados existentes, a sua localização e o seu significado.

O Metadados permite ainda que os utilizadores determinem a qualidade das pesquisas e análises realizadas. A origem e o processo de transformação dos dados presentes na BDA, podem constituir indicadores de qualidade.

Metadados é, fundamentalmente, *informação acerca da informação existente.* Sem a sua existência, os dados não terão significado. A qualidade final depende muito do Metadados, pois é através desta entidade que os utilizadores adquirem confiança na informação obtida.

1.5. Especificação

A implementação de uma base de dados analítica começou por utilizar uma tecnologia específica com meios abertamente multidimensionais. A reacção do relacional surgiu e a sua capacidade para o estabelecimento de *standards* de

especificação e definição dinâmica de relacionamentos breve se traduziu na demonstração da existência de um concorrente à altura. Actualmente estas duas formas básicas de implementação e a sua combinação em produtos híbridos abriram um mercado de concorrência feroz a nível intelectual, científica e comercial onde o poder do relacional está longe de estar afastado.

Em termos de especificação, há que reconhecer que a experiência acumulada pela perspectiva relacional em breve conduziu a novos conceitos, a novos diagramas enquanto, se nos situarmos do lado da perspectiva multidimensional, poderemos constatar a ausência de passos significativos.

Para examinarmos a contribuição da implementação relacional em termos de especificação começamos com o conceito tradicional de relações. Em seguida apresentaremos três modelos lógicos decorrentes do esforço de conceptualização da modelação: o diagrama estrela, o diagrama floco de neve e a constelação de factos, todos assentes na existência de um repositório relativo às métricas referidas ao nível mais detalhado da granularidade pretendida. Terminamos com a especificação relacional de um cubo.

1.5.1. Relações

Uma *relação* é uma tabela de dados bidimensional com um nome, consistindo num conjunto de colunas com nome e num número arbitrário de linhas sem nome.

Cada coluna numa relação corresponde a um atributo dessa relação.

Cada linha corresponde a um registo que contém valores para um objecto representado na relação. Apesar de não ter nome, uma linha pode ter um elemento identificador. Este corresponde ao valor de uma coluna especial, um atributo *chave*.

A especificação de uma relação é feita de acordo com a notação seguinte:

$$R=\{\underline{A1}, A2,....A3\} \qquad (1)$$

onde *R* é o nome da relação e *An* o nome das colunas ou atributos considerados. O sublinhado de *A1* significa que este é um atributo identificador.

Ao conjunto formado pela especificação de todas as relações de um sistema chamaremos *Esquema Relacional.*

Propriedades das relações

Para que uma tabela seja uma relação deve obedecer às propriedades seguintes:

1. *As ocorrências nas colunas deverão ser atómicas.* Não podem existir atributos com múltiplos valores ou grupos repetitivos numa relação. Sendo teoricamente possível pensar numa tabela com o valor do IDE da Alemanha em Portugal nos últimos dez anos num único registo, por existir aqui uma repetição do atributo, esta tabela nunca seria considerada uma relação.

2. *As ocorrências de colunas são do mesmo domínio.* Sendo o domínio uma definição dos valores e intervalos que um atributo pode assumir, numa relação todas as ocorrências numa coluna respeitam ao mesmo domínio.

3. *Cada linha é única.* Numa relação não existem linhas idênticas o que assegura que cada linha numa tabela tem significado e que um utilizador pode localizar com facilidade os dados desejados. Esta propriedade é garantida pela designação de uma chave primária para cada relação, isto é uma atributo que identifica inequivocamente as linhas dessa relação.

4. *A sequência das colunas não tem qualquer significado,* elas são referenciadas pelo nome e não pela sua posição na tabela.

5. *A sequência das linhas não tem significado* podendo os utilizadores visualizá-las de acordo com diferentes sequências lógicas.

1.5.2. Bases de Dados Analíticas suportadas por relações

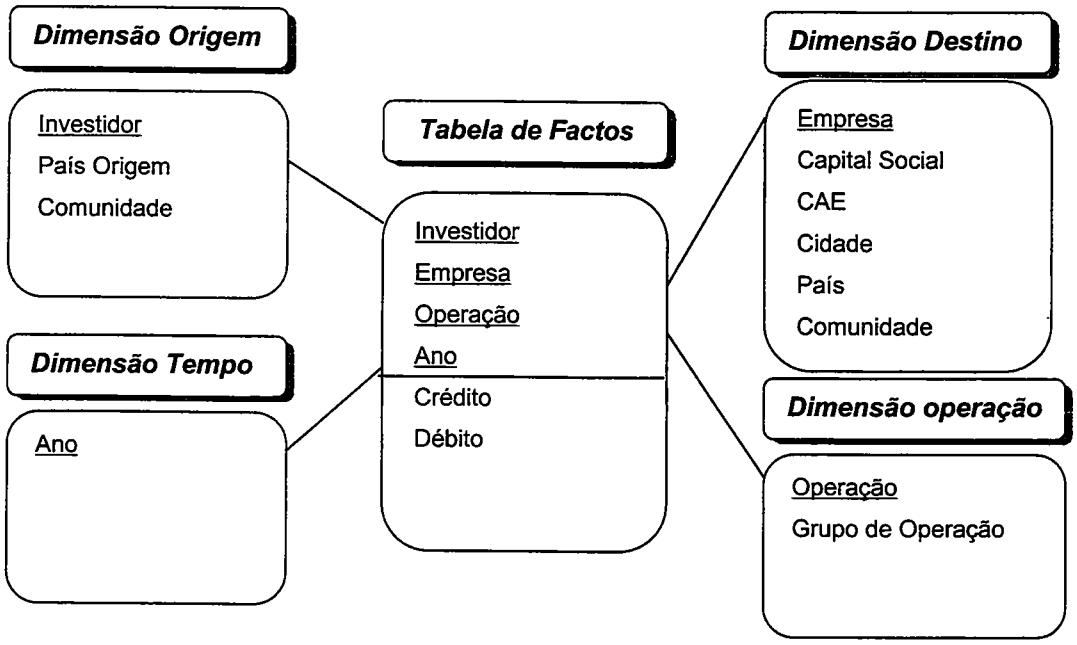
Vulgarmente, o modelo de dados multidimensional é representado por um *Diagrama Estrela* sendo esta a base para a construção de uma *BDA.* A razão para o nome é clara: a base de dados contém uma tabela de factos central e várias tabelas de dimensão organizadas à sua volta.

Enquanto a tabela de factos é grande, especialmente em termos de número de registos, as tabelas de dimensão são por vezes relativamente pequenas. Cada registo da tabela de factos tem um apontador na forma de chave estrangeira para cada uma das tabelas de dimensão.

Por outro lado, cada tabela de dimensão consiste em colunas que representam atributos de dimensão. Estes atributos podem ou não corresponder à hierarquia associada à dimensão.

No diagrama estrela apresentado, enquanto a tabela de factos é altamente normalizada[8], as tabelas associadas às dimensões surgem desnormalizadas.

Figura 8. Diagrama Estrela



A desnormalização existe na *dimensão Origem* dada a dependência funcional *País Origem* → *Comunidade* e na *dimensão Destino* que contem as dependências funcionais *Cidade* → *País* e *País* → *Comunidade*. A primeira daquelas dimensões pode ser normalizada, eliminando a coluna *Comunidade* e criando outra tabela com as colunas *País Origem* e *Comunidade*.

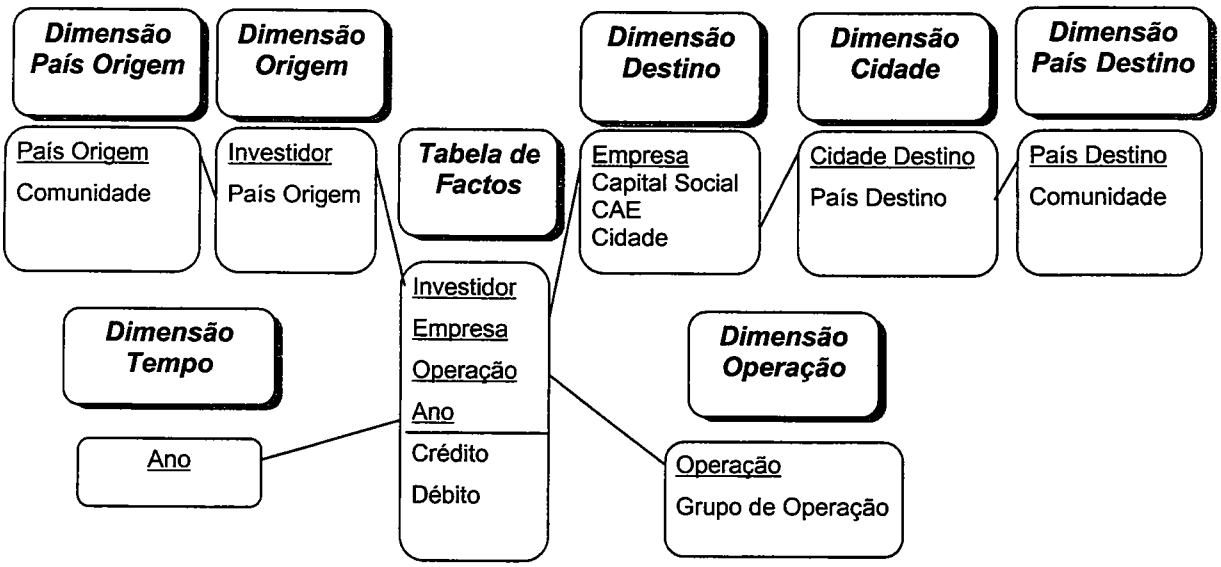
[8] Após a construção do modelo conceptual dos dados (Modelo Entidade/Associação) e a derivação de um modelo lógico relacional obtem-se um Esquema Relacional primário que representa a estrutura da informação de um modo natural e completo. O problema está em que pode não ser portador do mínimo de redundância possível. A Normalização tem por objectivo avaliar a qualidade do esquema relacional primário e transformá-lo (em caso de necessidade) num esquema equivalente, menos redundante e mais estável. Na sequência do trabalho de Codd foram apresentadas muitos passos para a normalização. Geralmente, para os sistemas operacionais, é satisfatória a garantia dos três seguintes:

1. eliminar grupos de atributos repetitivos;
2. eliminar dependências funcionais parciais;
3. eliminar dependências funcionais transitivas.

Para além das dimensões, o *diagrama estrela* apresenta as *métricas de análise*, neste caso o investimento a débito ou a crédito. Todas as medidas (ou métricas de análise) são armazenadas na tabela de factos.

Devido às tabelas de dimensão desnormalizadas, um *diagrama estrela* não fornece um suporte explícito a hierarquias conceptuais, pelo que muitos *BDAs* são desenhados utilizando um *Diagrama em Floco de Neve* onde algumas ou todas as dimensões são totalmente normalizadas. Isto resulta em vantagens em manutenção do esquema e trabalho com hierarquias conceptuais. Todavia, a estrutura desnormalizada oferece um acesso mais fácil aos dados associados às dimensões. Por esta razão, Kimball (1997) desencoraja fortemente a utilização de *diagramas floco de neve.*

Figura 9. Diagrama Floco de Neve



Finalmente, algumas *BDAs* são construídas utilizando *Constelações de Factos*, uma estrutura complexa onde, de acordo com Raden (1995), múltiplas tabelas de factos partilham tabelas de dimensões.

A especificação de bases de dados analíticas deve fornecer informação acerca dos vários elementos que a caracterizam e das possíveis interrelações. Assim, à já referida especificação relacional, há que juntar informação acerca das dimensões e dos níveis de consolidação.

Sendo $D=\{d_1,...d_n\}$ um conjunto de nomes de dimensões, $Di = \{A_{i1}, A_{i2},...A_{in}\}$ o conjunto de nomes de atributos pertencente a cada dimensão, $F=\{A_1,...,A_m\}$ um conjunto de atributos e métricas constantes da tabela de factos e $P(Di)$ =

$\{A_{i1},...A_{in}\}$ uma lista de composição das hierarquias referida a cada uma das dimensões, *(D,F,P)* pode ser considerada, de acordo com Agrawal et al (1996,1), como o *esquema de um cubo.*

1.5.3. Bases de Dados Analíticas suportadas por arrays

Como dissemos, não estão definidos *standards* de especificação para BDAs de implementação multidimensional e a prática dos fornecedores neste domínio tem manifestado uma tal variabilidade que é duvidoso que a sua existência se venha a materializar.

A especificação antes apresentada servirá em casos em que haja uma maior orientação aos dados, pois informa sobre atributos de dimensão, métricas e níveis de consolidação existentes. O problema está em BDAs que, mais orientadas à informação, construam as sua dimensões a partir de estruturas de elementos informacionais, isto é, das próprias instâncias de atributos. Neste caso uma especificação da tabela de factos e a descrição completa dos níveis de consolidação tem de ser feita gerando uma enorme carga de manutenção quando os domínios são muito extensos. Um exemplo é apresentada em seguida.

Exemplo 1.

> **Dimensão Sector**= {Global {Agricultura, Minas e Extractivas {Extracção de Petróleo e Gás}, Indústria {Produtos Alimentares, Têxteis e Vestuário,...},...}}
>
> **Indicador Crédito** = {Sector, Operação, País, Tempo}
>
> Neste caso é incompletamente especificada a dimensão Sector onde se encontram vários níveis de consolidação. O mais elevado denomina-se Global e é subdividido em Agricultura, Minas e Extractivas, Indústria e outros. Por seu lado, Indústria subdivide-se em Produtos Alimentares, Têxteis e Vestuário, etc.
>
> O indicador especificado existe granularizado a um nível de quatro dimensões sendo este o mínimo de detalhe exigido para a informação de carregamento.

O maior defeito da especificação referida encontra-se na excessiva carga de trabalho associada à manutenção do dicionário, pois um novo elemento exige uma nova definição o que implica, por outro lado, uma extrema atenção a dados provenientes do sistema operacional ou a dados externos para detecção de elementos novos.

1.6. Do Modelo Entidade Associação ao Modelo Multidimensional

Poderá ter algum interesse em termos teóricos e práticos a capacidade de derivação de um modelo multidimensional a partir de um modelo entidade associação.

Naturalmente que esta abordagem não resiste a uma crítica associada ao tipo de orientação (uma virada para as aplicações outra virada para os factos), mas o certo é que, em determinado contexto, este conhecimento pode ter um papel catalizador de esforços de mudança.

Para esta derivação podem ser considerados dois passos fundamentais: (1) a representação das associações no modelo multidimensional e (2) a representação das entidades no modelo multidimensional.

1.6.1. Representação das Associações no Modelo Multidimensional

Esta representação dependerá da cardinalidade da associação.

Associações com Multiplicidade Tipo N:N.

Estas associações originam uma *tabela de factos* cuja *chave primária* será composta pelas chaves primárias das entidades que participam na associação. As *métricas* serão qualquer atributo não chave usado para caracterizar a associação desde que seja numérico. Os atributos da associação considerados chave (por exemplo, uma data) contribuirão para a chave da tabela de factos e deverão originar uma tabela dimensão cuja chave primária é formada por esse atributo.

Associações com Multiplicidade Tipo 1:N

Estas associações originam o processo de desnormalização pela adição dos atributos chave e não chave da entidade do lado 1 da associação à

entidade/tabela do lado N. As associações com multiplicidade tipo 1:1 podem ser tratadas como caso especial do 1:N

1.6.2. A Representação das Entidades no Modelo Multidimensional

Neste caso as ***entidades participantes em associações tipo n:n.*** dão origem às *tabelas de dimensão* cuja chave provém do identificador primário da entidade e recebendo todos os atributos da entidade base. A chave desta tabela dimensão deverá contribuir para a chave da tabela de factos.

Já as *entidades participantes em associações tipo 1:N* originam uma transformação diferente. Os atributos da entidade do lado 1 transitam para uma potencial *tabela dimensão* do lado N. Esta tabela é potencial porque pode fazer parte de uma associação em que esteja do lado 1 originando processo semelhante.

2. Extracção de Informação de Bases de Dados

A extracção de informação de bases de dados é uma das primeiras práticas de processamento analítico e pode ser integrada em processos considerados mais elaborados. Por exemplo, uma das suas materializações, a interrogação, pode ser integrada nos processos de pesquisa de árvores de decisão ou de afinidade lógica, sendo possível chegar a árvores de decisão por *query.*

Por outro lado, elementos da análise multidimensional (dimensões, métricas, *drill-down, roll-up*), são incorporados por alguns investigadores em produtos com objectivos mais ambiciosos do que a mera extracção de informação, podendo constituir-se em elementos essenciais para a descoberta de conhecimento como é visível em formulações do *tipo indução orientada por atributos* da escola Simon Fraser University do que Han et al (1997) é uma concretização. Tais desenvolvimentos levam os autores a falar numa linguagem de interrogação para a descoberta de conhecimento em bases de dados[9].

Esta secção traduz-se numa visita a duas materializações da extracção de informação de bases de dados: a interrogação e a análise multidimensional.

[9] Data Mining Query Language (DMQL), conforme Han et al (1997)

2.1. Interrogação

Dado um conjunto de tabelas de atributos *A1, A2,...An*, a interrogação extrai um conjunto de linhas que obedecem a uma condição de pesquisa simples ou complexa (envolvendo uma ou mais tabelas) as quais são constituídas pelo valor dos atributos ou expressões envolvendo atributos referidos no processo.

Vulgarmente está disponível através de *query* directo utilizando uma linguagem para tal efeito, sendo o SQL *(Strutured Query Language)* uma das linguagens mais divulgadas neste campo.

A história do SQL começa com uma implementação realizada pela IBM que tinha por objectivo a implementação do modelo de Codd. Inicialmente o produto teve a designação SEQUEL (a Structured English Query Language) a qual foi modificada por razões legais. A evolução da linguagem e do produto concretizaram-se numa primeira implementação comercial pela Relational Software, Inc., hoje conhecida por *Oracle Corporation.*

O SQL tem vindo a adaptar-se ao desenvolvimento das necessidades dos utilizadores de sistemas de informação e é, ainda, considerado um *standard* dos Sistemas de Gestão de Bases de Dados Relacionais. A primeira apresentação do SQL, ainda SEQUEL, foi proporcionada por Chamberlin e Boyce (1974). É possível encontrar uma explicitação actualizada da linguagem em Date (1996]. O seu autor C. J. Date foi um companheiro de longa data de E. F. Cobb contribuindo para a divulgação da filosofia relacional. Aliás, tal como se pode perguntar o que seria o mundo das bases de dados sem a contribuição de Codd, podemos interrogar-nos do que seria o conhecimento do relacional sem o trabalho enveredado por Date sempre portador de uma carga pedagógica elevada pela simplicidade da linguagem e pelos múltiplos exemplos. Na língua portuguesa, a referência Damas (1999] cobre perfeitamente toda a problemática da linguagem à qual associa a sua utilização em ambientes da terceira geração e em ambientes orientados por objectos.

Entre os seus pontos fortes ressalta a sua articulação profunda com a filosofia relacional e com operações lógicas sobre conjuntos o que, associado à sua expressão em língua inglesa, permite uma compreensão facilitada dos objectivos de qualquer processo que esteja a suportar. Esta facilidade diminui

à medida que o processo se torne mais complexo e necessite de comandos mais elaborados para a sua consecução. Neste capítulo, a interrogação aparecerá exemplificada numa secção específica, mas em vários pontos apresentaremos soluções em SQL para suportarem outro tipo de processos.

Exemplo 2. Considere-se o esquema *Investimento = {País, Sector, Operação, Data, Invest}* relativo ao investimento directo estrangeiro em Portugal. O comando escrito em SQL

Select Pais, SUM (INVEST) from Investimento group by Pais

extrairia o total do investimento directo por país de uma tabela (*Investimento*) onde existisse informação detalhada com esse objectivo.

Exemplo 3. O comando sobre o mesmo esquema

Select Sector, Data, Invest from investimento where Pais = 'Alemanha'

extrairia com detalhe total, da tabela *Investimento*, todas as acções de investimento da Alemanha, independentemente do *Sector, Data e Operação* e só estas.

2.1.1. Operadores Relacionais

De acordo com Codd (1970), a álgebra relacional consiste em oito operadores organizados em dois grupos:

1. o *tradicional conjunto de operações* designadas por união, intersecção, diferença e produto cartesiano (todos eles modificados para ter em conta o facto de que os seus operandos são especificamente relações e não conjuntos arbitrários);

2. os *operadores especiais* designados por restrição, projecção, junção e divisão.

A *Restrição* tem como resultado uma relação que consiste em todos os registos de uma relação que satisfazem uma condição especificada.

A *Projecção* tem como resultado uma relação que consiste em todos os registos que permanecem numa relação depois de terem sido eliminados atributos especificados.

O *Produto Cartesiano* tem como resultado todos os possíveis registos que são uma combinação de dois registos cada um a partir de duas relações especificadas.

A *União* tem com resultado uma relação de todos os registos que aparecem em algumas ou ambas as relações especificadas.

A *Intersecção* tem como resultado os registos que aparecem em ambas as relações especificadas.

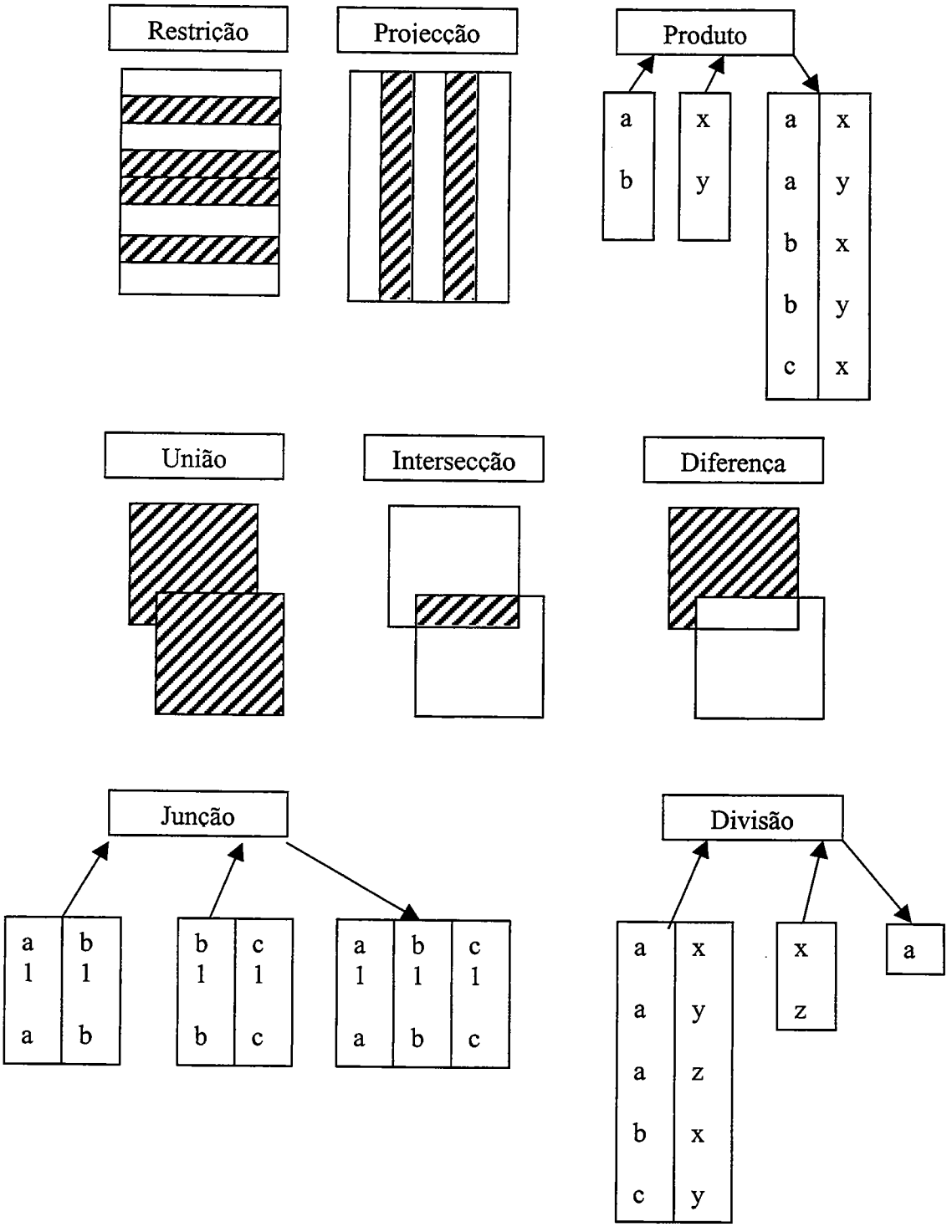
A *Diferença* tem como resultado uma relação que consiste em todos os registos que aparecem na primeira relação e não aparecem na segunda.

A *Junção* tem como resultado uma relação que contém todos os registos possíveis que são uma combinação de dois registos, cada um das relações especificadas de tal modo que dois registos que contribuam para uma dada combinação têm uma valor comum para o atributo comum.

A *Divisão* toma duas relações, uma binária e uma unária, e devolve uma relação com todos os valores de um atributo da relação binária que satisfazem todos os valores na relação unária.

A figura seguinte dá uma imagem simplificada deste conjunto de operadores.

Figura 10. Operadores da Álgebra Relacional



Fonte: Date (1996)

2.2. *Análise Multidimensional*

A visão bidimensional dos dados, apresentada pelas relações, é actualmente insuficiente para produzir uma imagem real ou projectada sobre os factos económicos. Os economistas necessitam de visualizar os dados através de várias perspectivas, ou seja, de várias dimensões.

Exemplo 4. Considere-se uma análise sobre o Investimento Directo Estrangeiro em Portugal. *Investidor, Tipo de operação, Actividade Económica* e *Período de Tempo* são as designadas dimensões que correspondem, genericamente, a perspectivas especificas de análise. Os utilizadores esperam ver a informação através destas dimensões – *Investimento Externo* por *Tipo de Operação* no *Tempo* ou *Investimento Externo* por *País Investidor* - esperando ainda poder alterar dinamicamente estes pedidos. A especificação do cubo associado a este exemplo poderia apresentar o seguinte aspecto:

D={Investidor, Tipo de Operação, Actividade Económica, Período de Tempo}

R={Investidor, Operação, CAE, Ano, Valor do IDE}

P={Investidor (País origem, Investidor), Operação (Grande Operação, Operação), Actividade (Grande Actividade, Actividade económica), Período de Tempo (Período, Ano)}

2.2.1. Operadores Multidimensionais

Os operadores que se traduzem nas principais ajudas ao processo de Análise Multidimensional materializam-se num vasto conjunto de funcionalidades: rotação, *slice and dice, data scoping, drill-down e roll-up.*

A organização dos dados num *array* multidimensional tem uma utilidade inerente. Diferentemente do modo como se vê uma tabela relacional, a visualização dum *array* explicitamente mostra dimensões, posições ao longo das dimensões, intersecções e valores de células. De facto, visualizar um *array* multidimensional é semelhante a visualizar uma folha de cálculo.

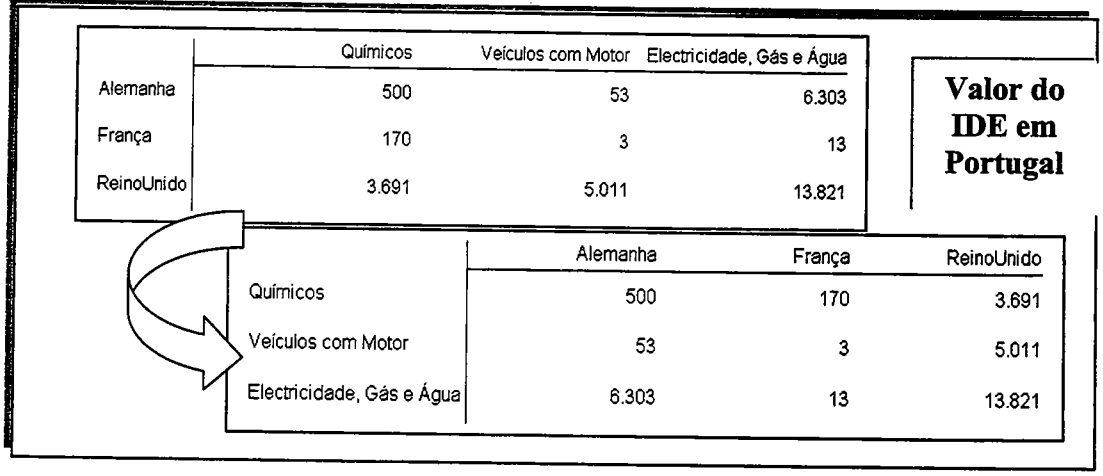
Exemplo 5. Considere-se o exemplo anterior relativo à análise do Investimento Directo Estrangeiro em Portugal. É possível visualizar o Investimento Bruto por Actividade Económica e País de Origem, os dados estão imediatamente disponíveis para visualização. As bases de dados multidimensionais permitem percorrer um *array* tal como se percorre uma folha de cálculo.

Num ambiente relacional, esta forma de visualização estruturada exigiria a elaboração prévia de um *query* complexo à base de dados.

Após ter acesso aos primeiros resultados, o utilizador pode concluir que uma visualização dos dados por País de Origem e Actividade Económica poderá ser mais útil. Num ambiente relacional, esta modificação da forma de ver os dados exigiria um novo *query*, eventualmente complexo, e uma ou nova ordenação. Num ambiente multidimensional a nova visualização corresponde apenas a uma *rotação* da antiga.

A figura abaixo ilustra como a *rotação* funciona num ambiente multidimensional.

Figura 11. A Rotação em Bases de Dados Multidimensionais



O *array* dimensional acima tem duas possibilidades de visualização: *Actividade Económica por País de Origem* ou *País de Origem por Actividade Económica.* Em qualquer dos casos, a intersecção das duas dimensões suporta o valor do IDE em Portugal em 1996.

Naturalmente, podem ser considerados exemplos com três dimensões: Actividade Económica, País de Origem e Tipo de Operação. Este cubo tridimensional tem seis possíveis visualizações de interesse para o utilizador:

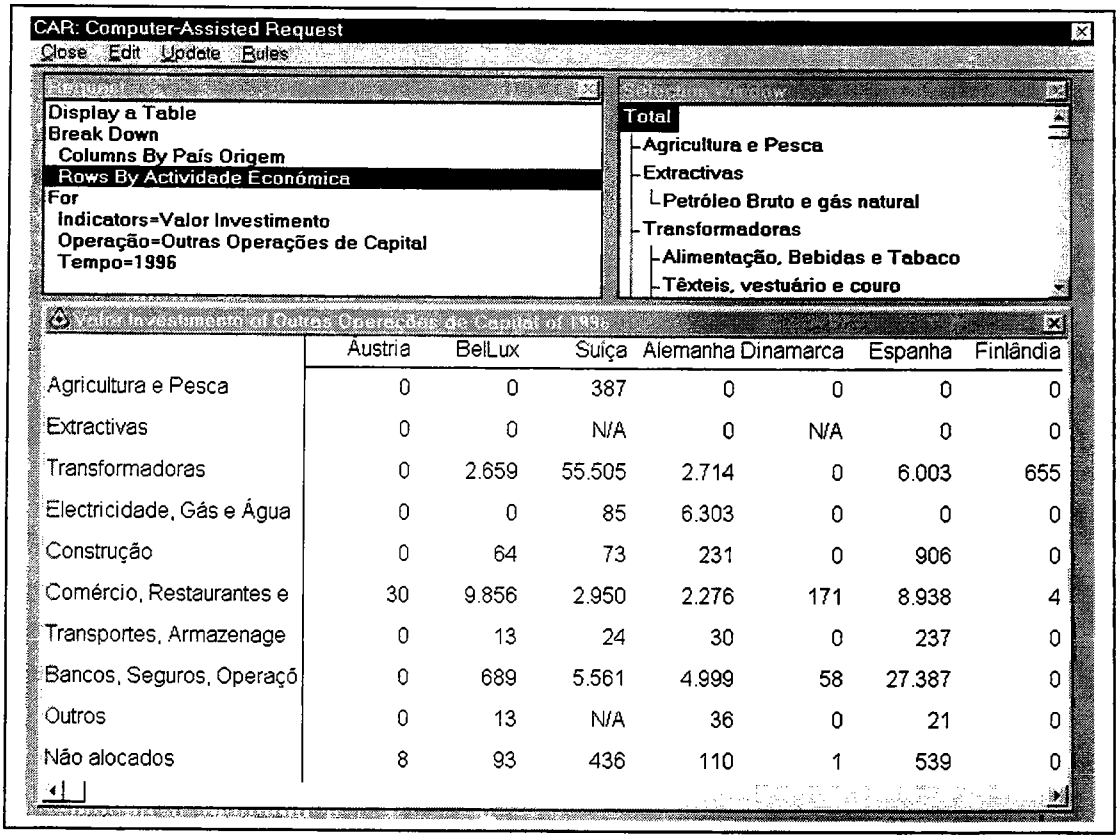
- *País de Origem* por *Actividade Económica* (com *Tipo de Operação* dado);
- *Actividade Económica* por *País de Origem* (com *Tipo de Operação* dado);
- *Actividade Económica* por *Tipo de Operação* (com *País de Origem* dado);
- *Tipo de Operação* por *Actividade Económica* (com *País de Origem* dado);
- *Tipo de Operação* por *País de Origem* (com *Actividade Económica* dada);
- *País de Origem* por *Tipo de Operação* (com *Actividade Económica* dada).

A capacidade para ter acesso a diferentes visualizações dos dados traduz-se em grande flexibilidade para diferentes utilizadores que poderão estar interessados em aspectos específicos da base de dados multidimensional.

Só por si, a *rotação* é insuficiente para definir as dimensões de análise, os valores fixos e os indicadores a visualizar. Assim, as ferramentas de análise multidimensional dispõem de uma facilidade de selecção de ocorrência de dimensões e indicadores e, para além disso, de fixação das dimensões em *background*.

Exemplo 6. Este conjunto de elementos pode ser visualizado na figura seguinte cujos dados podiam ter sido uma base para a escolha das dimensões, do indicador ou métrica de análise e dos valores focados em dimensões de *background*.

Figura 12. O *Slice and Dice* em Bases de Dados Multidimensionais



	Austria	BelLux	Suiça	Alemanha	Dinamarca	Espanha	Finlândia
Agricultura e Pesca	0	0	387	0	0	0	0
Extractivas	0	0	N/A	0	N/A	0	0
Transformadoras	0	2.659	55.505	2.714	0	6.003	655
Electricidade, Gás e Água	0	0	85	6.303	0	0	0
Construção	0	64	73	231	0	906	0
Comércio, Restaurantes e	30	9.856	2.950	2.276	171	8.938	4
Transportes, Armazenage	0	13	24	30	0	237	0
Bancos, Seguros, Operaçõ	0	689	5.561	4.999	58	27.387	0
Outros	0	13	N/A	36	0	21	0
Não alocados	8	93	436	110	1	539	0

Uma base de dados multidimensional implementada com preocupações de correcção permite ao utilizador final situar-se rapidamente no conjunto dos dados pretendido para análise.

Como foi descrito previamente, o utilizador final tem muitas vezes necessidade de visualizar os dados de diferentes pontos de vista (por exemplo, IDE por país de Origem versus IDE por Actividade Económica). Muitas vezes, todavia, as visualizações pretendidas são bastante semelhantes (IDE por país de Origem versus IDE por Comunidade).

As visualizações referidas podem referir-se a diferentes níveis de análise. Um utilizador pode querer examinar, primeiro, o investimento a nível da Comunidade e, depois, examinar a sua repartição pelas países em que esta se desdobra existindo uma relação natural entre um e outro. O mesmo pode dizer-se relativamente dos dados do investimento por actividade económica: esta pode ser considerada a diferentes níveis de agregação.

Uma ferramenta para a análise multidimensional é especialmente desenhada para facilitar o tratamento destas relações naturais. Os operadores *drill-down* e *roll-up* são propositadamente desenhados para implementar estas facilidades. O operador *drill-down* proporciona maior nível de detalhe relativamente à visualização em causa, possibilitando a passagem a um nível inferior na hierarquia associada à dimensão. Na dimensão *Investidor*, um *drill-down* sobre um país possibilitaria a visualização da informação associada ao país. O operador *roll-up* corresponde à operação inversa do *drill-down*, carecendo de informação adicional relativa ao objecto. Um *roll-up* sobre investidor conduziria ao país de origem, informação decorrente do esquema do cubo.

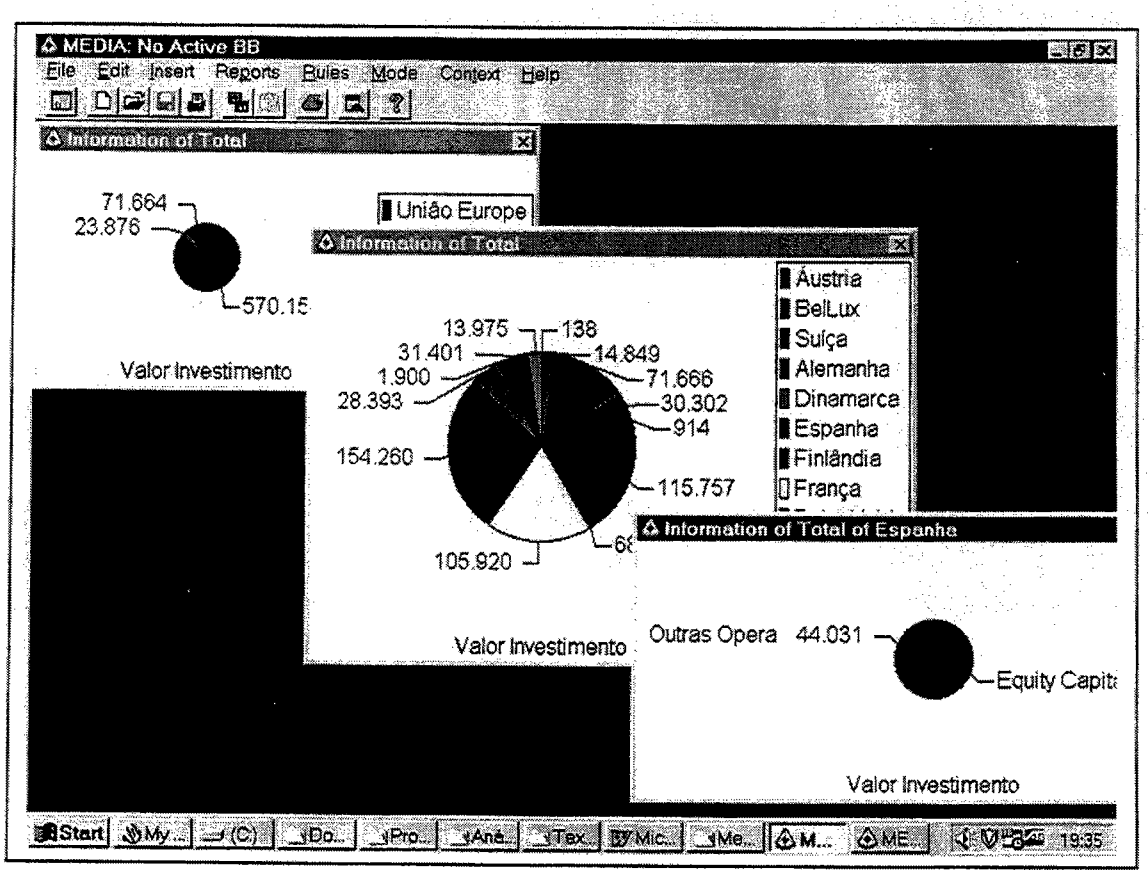
Apesar de ser possível criar dimensões independentes para dois níveis de análise, por exemplo país e comunidade, uma solução mais poderosa e elegante é definir duas relações agregadas na mesma dimensão. Um nível de agregação será *País*, outro será *Comunidade*. Estas duas agregações são, então, vistas como fazendo parte da mesma dimensão.

Exemplo 7. Os países podem organizar-se em comunidades (exemplo, a União Europeia). Assim, *Comunidade* é um nível de agregação mais elevado de uma dimensão a que poderíamos chamar *Origem*. Quando

relacionarmos todos os pontos de comunidade com os dos países que a integram teremos definida uma hierarquia na dimensão *Origem*. O mesmo raciocínio pode ser feito para a *Actividade Económica* e até para o *Tipo de Operação*.

A figura seguinte é a expressão de uma organização desse tipo.

Figura 13. O *Drill-Down* em Bases de Dados Multidimensionais



Neste exemplo, partiu-se da distribuição do IDE em Portugal por três espaços: União Europeia, Fora da União Europeia (Estados Unidos e Japão) e EFTA. Um primeiro *drill-down* foi criado permitindo ver a distribuição daquele montante pelos países da União Europeia. Por fim, considerando um país é possível examinar como o seu investimento se distribuiu por tipo de operação.

Numa base de dados relacional, a alternativa era criar um campo por cada nível na hierarquia ou criar tabelas separadas definindo as relações. Esta solução é menos eficiente do que a perspectiva multidimensional que, naturalmente, compreende os relacionamentos hierárquicos entre diferentes agregações nas dimensões.

Note-se que toda esta análise é feita em moldes interactivos e suportada por selecções através do *mouse*.

3. Descoberta de Conhecimento em Bases de Dados

O objectivo desta secção é a descrição de processos de descoberta de conhecimento em bases de dados utilizados neste documento.

Estes não esgotam o leque de possibilidades que o analista da Economia pode utilizar sendo de referir ainda as perspectivas associadas a frequências alguns ramos das quais utilizam estatística *bayesiana* e as perspectivas equacionais que integram a regressão e o ajustamento de funções com utilização de redes neuronais.

Apesar de tudo, os casos que descrevemos ilustram um leque de possibilidades importante pois, por exemplo, a utilização de redes neuronais pode ser aplicada a processos de classificação. Estes, quer na perspectiva tradicional quer na perspectiva designada por lógica difusa, são os processos com que iniciamos esta secção. Em seguida, descrevemos algoritmos que apoiam a descoberta de associações. Terminamos com referências à análise de *clusters* que será utilizada na procura de estádios de desenvolvimento dos SIE.

Em analogia com o ponto anterior, procuraremos fazer uma abordagem sistemática onde seja possível detectar, num ponto inicial, aspectos de caracterização e exemplificação, num segundo momento, figuras em que se materialize o processo e, num terceiro, algoritmos ou operadores associados ao mesmo.

3.1. Classificação

Na grande maioria das facilidades associadas à descoberta de conhecimento em bases de dados, os dados não podem ser tratados utilizando a estrutura em que residem, carecendo de um processo prévio de transformação. Tal acontece com a *classificação*. Assim, após a caracterização desta técnica de descoberta de conhecimento, procederemos à descrição da transformação que muitos produtos operam para a viabilizar. Em seguida, apresentaremos dois possíveis resultados da classificação, a árvore de decisão e a regra de produção. Terminamos com a explicitação de alguns algoritmos de partição

que poderá ser operada quer em termos de descrição quer em termos de referências.

3.1.1. Caracterização

Dado um conjunto de instâncias de uma estrutura de dados complexa de atributos *A1, A2,... An,* onde um dos atributos tem um domínio de baixa cardinalidade (correspondendo a um número não elevado de classes), a *classificação* associa a cada instância nesse atributo um dos valores do seu domínio.

Vulgarmente, a classificação é feita em dois passos: num primeiro, é dado um conjunto de exemplos ao sistema para este induzir as regras de classificação, em seguida as regras são utilizadas para classificar outros registos

Existem múltiplos meios de se proceder a uma classificação. Entre os mais conhecidos estarão as árvores de decisão, as regras de produção, as redes neuronais e os modelos de lógica difusa. As duas primeiras constituem uma perspectiva lógica tradicional e o conhecimento subjacente ao processo é facilmente apreendido, enquanto a rede neuronal funciona como uma *caixa negra* devido à opacidade do conhecimento proporcionado. Por outras palavras, apesar de o objectivo (a classificação) ser possível as condições em que esta é feita tornam-se conhecidas em árvores de decisão e regras e permanecem opacas no caso das redes neuronais.

3.1.2. Transformação prévia dos dados

Os dados, tal como existem nos repositórios de base, podem carecer de uma transformação de forma a suportarem o processo de classificação. Esta transformação pode preocupar-se com várias problemas, designadamente:

1. com a criação de uma configuração específica para o atributo de baixa cardinalidade que vai servir de base à classificação;
2. com a transformação de um produto cartesiano resultante da junção de duas tabelas com relacionamento 1:N numa única linha através da multiocorrência do atributo associado à tabela do lado N;
3. com a geração de intervalos para a informação de tipo numérico.

Se a última transformação tipo pode ser automaticamente processada pelos produtos de *software* de análise, nos dois casos iniciais, é possível a

necessidade de recorrer às facilidades de interrogação oferecidas pelas bases de dados tradicionais.

No primeiro caso, pode dizer-se que um atributo de baixa cardinalidade pode ser objecto de utilização numa classificação através de uma de duas modalidades:

- o atributo participa com o valor original sem qualquer modificação ao longo do processo de partição e classificação e, neste caso, os valores do domínio de atributo condicionam os resultados, ou

- há uma transformação prévia para um atributo de menor cardinalidade da família 0/1 ou sim/não, tornando-se o processo de classificação em algo que permite clarificar as condições em que uma hipótese previamente lançada se concretiza ou não.

No segundo caso, o que está em causa não é o atributo objecto de classificação, mas os atributos que vão servir como condicionantes da classificação. Para evitar uma multiplicidade de linhas, toda a informação relativa a um determinado elemento deve ser reunida numa só linha através da junção de colunas representando ocorrências de atributos que se situam do lado N numa relação 1:N e que correspondem ao domínio do atributo respectivo. Dessas colunas será valorizada com 1 a correspondente à do caso em análise e com 0 todas as outras. Há assim um crescimento significativo do número de colunas tratado o que pode traduzir-se em alguma degradação e em algum desconforto para quem utiliza método. Quando o número de ocorrências possíveis é de vários registos com atributos diferentes no relacionamento 1:N referido o resultado é um tanto diferente do anterior traduzindo-se em várias colunas com 1 e as restantes com 0.

Exemplo 8. Considere-se um investidor do Reino Unido que realiza investimentos em vários sectores em Portugal. A transformação anterior que passaria por uma junção entre investidor e país, por um lado e entre investidor e movimento, por outro, traduzir-se-ia numa linha com colunas relativas a todos os países origem considerados e só uma delas valorizada em 1 e ainda com colunas relativas a todos os sectores de actividade económica considerados estes com valorizações a 1 nos sectores objecto de investimento.

Designaremos por transformação 1→N (1 valorizado em N) a primeira referida no exemplo anterior e por transformação N→M (N valorizados em M) a segunda. Pyle (1999) apresenta um leque muito variado de transformações aos dados associadas aos diferentes processos de descoberta de conhecimento em bases de dados, incluindo as referidas anteriormente.

3.1.3. Árvores de Decisão

Uma árvore de decisão é uma representação do conhecimento, estruturalmente simples, utilizada com sucesso em sistemas de aprendizagem supervisionada por máquinas.

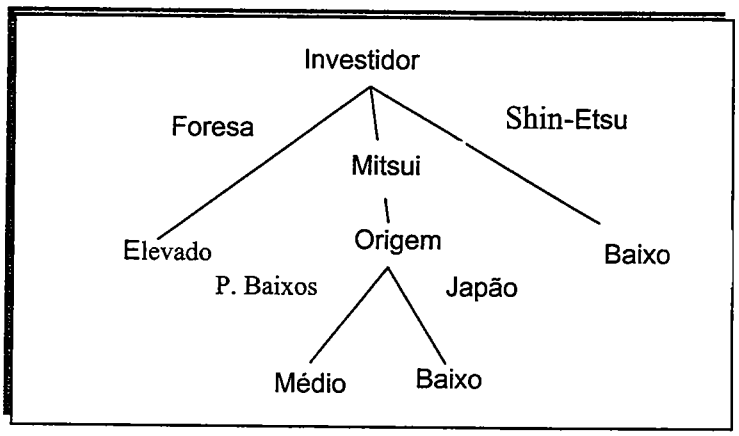
Uma *árvore de decisão* classifica exemplos num número finito das classes. Os nós numa árvore são etiquetados com o nome do atributo que lhes corresponde, as bordas são etiquetadas com valores possíveis para este atributo, e as folhas são etiquetadas com as classes diferentes. Um objecto é classificado seguindo um trajecto pela árvore, fazendo exame das bordas, correspondendo aos valores dos atributos no objecto.

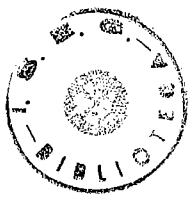
Exemplo 9. Considere-se um pequeno conjunto de treino, armazenando os objectos que descrevem o *Investimento Directo Estrangeiro em Portugal* num momento particular. Os objectos contêm informação relativa a:

Investidor: (Foresa, Mitsui, Shi-Etsu),
País de Origem: (Espanha, P. Baixos, Japão),
Concelho: (Ilhavo, Estarreja),
Operação: (Aquisição de quotas/acções, Aumento de capital) e
Montante.

A informação relativa a *Montante* foi codificada com o objectivo de gerar três classes: elevado (>100000), médio ([50000, 100000]) e baixo (<50000). A tarefa da classificação consiste em construir uma árvore de decisão que classifica todos os objectos no conjunto de treino correctamente. Como se demonstrará é possível construir a árvore seguinte:

Figura 14. Árvore de decisão





3.1.4. Regras de Produção

Os relacionamentos lógicos são usualmente representados como regras de produção. Uma regra de produção é uma afirmação da forma:

***Se** Condição1 **Então** classe1*

Cada caminho numa árvore de decisão corresponde a um termo: uma conjunção de condições envolvendo os atributos. A árvore apresentada na figura anterior é equivalente ao conjunto de regras de produção seguintes:

Default Montante = Baixo.
Se Investidor = Foresa Então Montante = Elevado.
Se Investidor = Mitsui e Origem = P.Baixos Então Montante = Médio.

Neste caso, fazem-se comparações utilizando os valores de campos de uma tabela de dados, isto é, temos uma representação atributo-valor: Investidor é o atributo e Foresa é o valor. Outro exemplo duma expressão atributo-valor é Origem = Espanha, onde Origem é o atributo e Espanha o valor.

As regras de produção usualmente funcionam em tabelas com atributos, tal como a apresentada no próximo quadro, de onde deduzimos a árvore de decisão representada na figura seguinte.

No caso apresentado, a possibilidade de definição de uma classe por defeito simplifica significativamente qualquer especificação.

Outra vantagem das regras é que podem facilmente estar para além das representações atributo-valor, por exemplo, ter instruções que comparem os

valores de dois campos sem expressamente referir quaisquer valores o que não pode ser estabelecido por árvores de decisão.

3.1.5. Algoritmos de Partição

As árvores de decisão exprimem uma forma simples de lógica condicional. Um sistema de árvores de decisão simplesmente particiona uma tabela em tabelas mais pequenas pela selecção de subconjuntos baseada em valores para um dado atributo.

Baseado no modo em como a tabela é particionada, obtemos um algoritmo diferente para árvore de decisão tal como o *Classification and Regression Trees* (CART), ou o *Chisquare Automatic Interaction Detection* (CHAID), ou o C4.5, etc.

Mas, para além destes algoritmos suportados por desenvolvimentos na área de matemática e estatística, é possível encontrarmos métodos expeditos para chegar a resultados expressos nos mesmos moldes, embora não se possa garantir que serão uma solução óptima.

Parsaye (1996) descreve a obtenção de sucessivas tabelas com base num algoritmo de partição de uma tabela original.

Considere-se a tabela seguinte, com dados hipotéticos acerca do *Investimento Directo Estrangeiro em Portugal.*

Quadro 3. Investimento Directo Estrangeiro em Portugal

Investidor	País Origem	Concelho	Operação	Montante
Foresa	Espanha	Ilhavo	Aquisição quotas/acções	> 100000
Foresa	Espanha	Ilhavo	Aquisição quotas/acções	> 100000
Mitsui	P. Baixos	Estarreja	Aumento de Capital	[50000, 100000]
Mitsui	Japão	Estarreja	Aumento de Capital	< 50000
Shin-Etsu	P. Baixos	Estarreja	Aumento de Capital	< 50000
Shin-Etsu	P. Baixos	Estarreja	Aumento de Capital	< 50000

Para obter uma árvore de decisão, seleccione-se o atributo *Investidor* para iniciar a operação de partição e, depois, o atributo *Origem.* Naturalmente, se existirem muitas colunas na tabela, a questão de qual o atributo a seleccionar primeiro torna-se crucial.

A partir do valor do atributo *Investidor* são construídas novas tabelas, neste caso três por ser esse o número de valores diferentes. A partição é feita com base em:

Investidor = Foresa,

Investidor = Mitsui,

Investidor = Shin-Etsu.

As tabelas obtidas serão as seguintes:

Quadro 4.Tabelas Obtidas por Partição da Tabela Original

Investidor	País Origem	Concelho	Operação	Montante
Foresa	Espanha	Ilhavo	Aquisição quotas/acções	> 100000
Foresa	Espanha	Ilhavo	Aquisição quotas/acções	> 100000

Investidor	País Origem	Concelho	Operação	Montante
Mitsui	P. Baixos	Estarreja	Aumento de Capital	[50000, 100000]
Mitsui	Japão	Estarreja	Aumento de Capital	< 50000

Investidor	País Origem	Concelho	Operação	Montante
Shin-Etsu	P. Baixos	Estarreja	Aumento de Capital	< 50000
Shin-Etsu	P. Baixos	Estarreja	Aumento de Capital	< 50000

Das tabelas obtidas a primeira e a terceira estão classificadas porque todos os valores de *Montante* são os mesmos. O mesmo não acontece com a segunda tabela que carece de nova cisão. Esta vai originar duas novas tabelas:

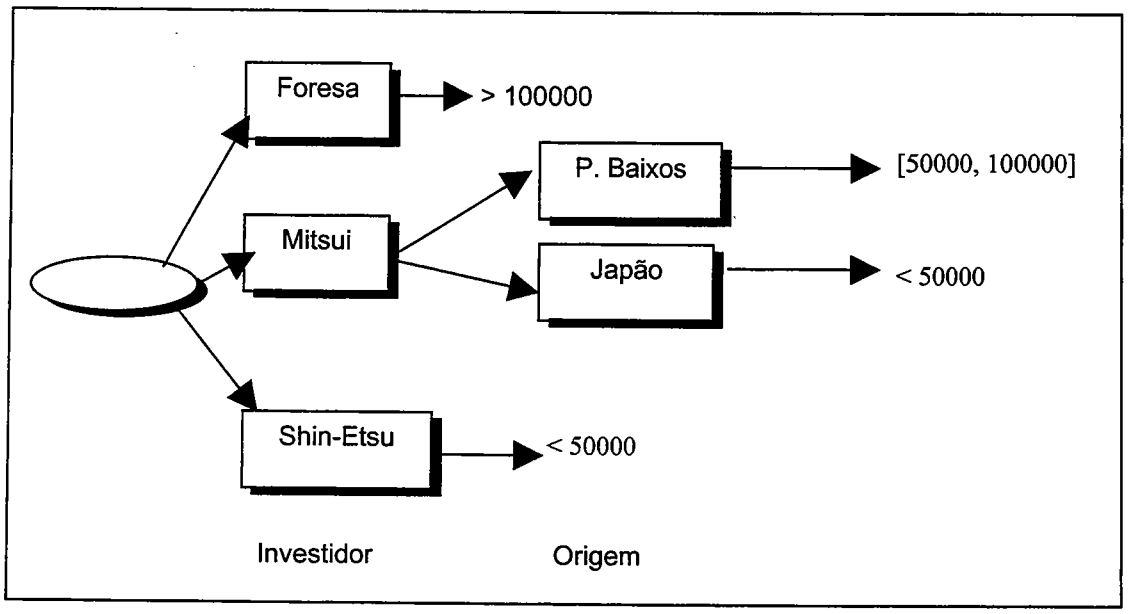
Quadro 5. Tabelas Obtidas por Partição da Tabela Intermédia

Investidor	País Origem	Concelho	Operação	Montante
Mitsui	P. Baixos	Estarreja	Aumento de Capital	[50000, 100000]

Investidor	País Origem	Concelho	Operação	Montante
Mitsui	Japão	Estarreja	Aumento de Capital	< 50000

Neste ponto, todas as tabelas estão classificadas. A árvore de decisão que descreve o processo de classificação, gerada a partir desta tabela, é mostrada pictorialmente na figura seguinte.

Figura 15. Árvore de Decisão Gerada por *Query*



Em muitos casos, incluindo a tabela acima, não existe qualquer atributo melhor para início de selecção, e qualquer que seja o atributo escolhido haverá sempre conhecimento perdido. Por exemplo, os dois factos:

a) *O investidor Foresa aplica montantes elevados,*
b) *Aquisições de quotas/acções são aplicações de montantes elevados*

nunca podem ser obtidos da tabela acima num ciclo de extracção de conhecimento com a árvore de decisão. Podemos obter o facto (a) ou o facto (b) da árvore, não ambos, porque a árvore de decisão selecciona um atributo especifico para executar a partição em cada estádio.

Apesar de ser possível gerar manualmente árvores de decisão e regras, a utilização de software específico torna essa tarefa mais rápida e permite inclusivamente a geração de várias árvores utilizando algoritmos e condicionalismos diferentes. Tal implica um conhecimento mínimo do funcionamento dos mesmos.

A ideia central presente nos algoritmos de partição consiste em dividir recursiva e o mais eficazmente possível os exemplos do conjunto de aprendizagem por testes definidos com a ajuda dos atributos até que se obtenha sub-conjuntos de exemplos contendo apenas (ou quase) exemplos pertencendo todos a uma mesma classe.

Em todos os métodos, encontram-se os três operadores seguintes :

1. **Decidir se um nó é terminal**, quer dizer decidir se um nó deve ser etiquetado como uma folha, o que acontecerá quando, por exemplo: todos os exemplos estiverem na mesma classe, ou existir menos do que um certo número de erros.

2. **Seleccionar um teste a associar a um nó,** o que pode ser feito aleatoriamente, utilizando critérios estatísticos ou indicadores como o de Gini ou o de entropia.

3. **Afectar uma classe a uma folha**. Atribui-se a classe maioritária salvo no caso onde se utilizem funções custo ou risco.

Os métodos vão diferir pelas escolhas efectuadas para estes diferentes operadores, quer dizer sobre o escolha de um teste (por exemplo, utilização do ganho e da função entropia) e o critério de paragem (quando parar o crescimento da árvore, seja quando decidir se um nó é terminal).

O ideal seria encontrar um critério que permitisse parar o crescimento da árvore no bom momento. Infelizmente, no estado actual da investigação, um tal critério não pôde ser encontrado. Além disso, o risco de parar demasiado cedo o crescimento da árvore é mais importante que o de parar demasiado tarde. Por consequência, os métodos utilizados procedem muitas vezes em duas fases: (1) a primeira corresponde aos três passos antes apresentados; (2) numa segunda fase, *desbasta-se* a árvore obtida para ensaiar fazer diminuir o erro real (desbastar uma árvore consiste em suprimir certas sub-árvores).

Consideremos uma estrutura de dados S onde se encontra um conjunto de classes $\{1,...,c\}$ e uma árvore de decisão t. A cada posição p de t corresponde um sub-conjunto da estrutura de dados que é o conjunto dos exemplos que satisfazem os testes da raiz até a esta posição. Por consequência, pode-se definir, para toda a posição p de t, as quantidades seguintes

1. N_p é o cardinal do conjunto dos exemplos associado a p,

2. N_{kp} é o cardinal do conjunto dos exemplos associado a p que são da classe k,

3. $P_{kp} = N_{kp}/N_p$ é a proporção de elementos da classe k na posição p.

A fim de permitir comparar as diferentes escolhas possíveis, definem-se funções que permitam medir o *grau de mistura* dos exemplos entre as diferentes classes. Uma tal função deve verificar a propriedade seguinte: ela deve tomar o seu valor mínimo quando todos os exemplos estão numa mesma classe (o nó é puro) e o seu valor máximo quando os exemplos são equi-repartidos.

Exemplo 10. Se se dispuser de 8 elementos e de duas classes, uma tal função deverá tomar o seu mínimo para os pares (0,8) e (8,0) e o seu máximo para o par (4,4).

Existem diferentes funções que satisfazem estas propriedades, nós citaremos duas: a função de Gini, apresentada por Breiman et al (1984), responsável pelo surgimento do algoritmo CART e a função entropia, comum a outras áreas do conhecimento, mas, neste contexto utilizada por Quinlan (1993) e em múltiplos trabalhos alusivos ao ID3.

$$Entropia\ (p) = -\sum_{k=1}^{c} P_{kp} * \log_2(P_{kp}) \qquad (2)$$

$$Gini\ (p) = 1 - \sum_{k=1}^{c} P_{kp}^2 \qquad (3)$$

$$= 2 * \sum_{k<k'} P_{kp} * P_{k'p} \qquad (4)$$

Consideremos o caso de duas classes e chamemos x à proporção de elementos da classe 1 na posição p.

Tem-se, portanto, $Entropia\ (p) = -\ x \log x - (1-x) \log (1-x)$. Esta função de x toma os seus valores no intervalo [0,1], tem o seu mínimo para $x=0$ e $x=1$ que vale 0 e tem seu máximo para $x = 1/2$ que vale 1.

A função de Gini é definida por $Gini\ (p) = 2x(1-x)$. Esta função de x toma os seus valores no intervalo [0,1/2], tem seu mínimo para $x = 0$ e $x = 1$ que vale 0 e tem o seu máximo para $x=1/2$ que vale 1/2. Estas duas funções são simétricas em relação a $x=1/2$.

Dispõe-se, assim, de funções que permitem medir o grau de mistura das classes para todo o conjunto de dados utilizado e, portanto, para toda a posição da árvore em construção. Chamemos i à função escolhida. Resta definir uma função que permita escolher o teste que deve etiquetar o nó

corrente. Em princípio são possíveis tantos testes quantos os atributos disponíveis. Com este objectivo, introduz-se uma função de ganho por :

$$Ganho(p,t) = i(p) - \sum_{j=1}^{n} P_j \times i(p_j) \qquad (5)$$

onde p designa uma posição, t um teste de aridade n e P_j é a proporção de elementos de S na posição p que estão na posição p_j (que satisfazem o *j-ésimo* ramo do teste).

Se se considerar como função i a função entropia, o termo $i(p)$ representa a entropia actual do nó p, o segundo termo da diferença representa a entropia esperado introduzindo o teste que é igual à soma ponderada das entropias dos novos nós criados. Deseja-se obter as entropias mais fracas possível porque, de acordo com as propriedades da função entropia, se a entropia é fraca, a maior parte dos elementos encontram-se numa mesma classe. Procura-se pois obter o de ganho máximo.

Exemplo 11. Reconsideremos o exemplo em que fizemos a partição da tabela relativa ao Investimento Directo Estrangeiro. Vamos mostrar que o caminho seguido foi o associado ao ganho máximo.

Comecemos por determinar a entropia na raiz da árvore. Nesse momento, não existe qualquer partição pelo que há que considerar todos os registos (exemplos). São seis registos: 2 *elevado*, 1 *médio* e 3 *baixo*. Virá

$$\textit{Entropia na raiz} = -\frac{2}{6}\log_2\frac{2}{6} - \frac{1}{6}\log_2\frac{1}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1{,}459148$$

Estando situados na posição raiz, a partição pode ser operada através de um de quatro atributos: *Investidor*, *País Origem*, *Concelho* e *Operação*. Há assim que determinar o ganho associado a cada um destes atributos. Se escolhermos *Investidor*, obteremos:

- para Foresa, dois registos na mesma classe (montante elevado);
- para Mitsui, dois registos repartidos pelas classes média e baixa;
- para Shin-Etsu dois registos na mesma classe (montante baixo).

Esta situação pode ser representada por (2,0,0) (0,1,1) (0,0,2) que permite determinar as entropias associadas a cada investidor. No

quadro seguinte, resumimos esse cálculo a que juntamos o da subdivisão por outros atributos e o da entropia esperada.

Quadro 6. Simulação da determinação da entropia

Atributo	Cardinal	Entropia	Ponderador	Entropia esperada
Investidor	(2,0,0)	$-\frac{2}{2}\log_2\frac{2}{2}=0$	2/6	0,0000
	(0,1,1)	$-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2}=1$	2/6	0,3333
	(0,0,2)	$-\frac{2}{2}\log_2\frac{2}{2}=0$	2/6	0,0000
País de Origem	(2,0,0)	$-\frac{2}{2}\log_2\frac{2}{2}=0$	2/6	0,0000
	(0,1,2)	$-\frac{1}{3}\log_2\frac{1}{3}-\frac{2}{3}\log_2\frac{2}{3}=0{,}918296$	3/6	0,4591
	(0,0,1)	$-\frac{1}{1}\log_2\frac{1}{1}=0$	1/6	0,0000
Concelho	(2,0,0)	$-\frac{2}{2}\log_2\frac{2}{2}=0$	2/6	0,0000
	(0,1,3)	$-\frac{1}{4}\log_2\frac{1}{4}-\frac{3}{4}\log_2\frac{3}{4}=0{,}811278$	4/6	0,5409
Operação	(2,0,0)	$-\frac{2}{2}\log_2\frac{2}{2}=0$	2/6	0,0000
	(0,1,3)	$-\frac{1}{4}\log_2\frac{1}{4}-\frac{3}{4}\log_2\frac{3}{4}=0{,}811278$	4/6	0,5409

Notas:

1. Na coluna *Cardinal,* apresentamos, para cada ocorrência diferente do atributo utilizado como meio de partição, a cardinalidade de cada classe, leitura que foi feita previamente para o investidor *Foresa.*

2. *Ponderador* corresponde à proporção dos elementos que se repartem pelas três classes na ocorrência particular do atributo escolhido para teste, relativamente ao cardinal de elementos repartidos.

3. *Entropia esperada* corresponde ao produto do ponderador pela entropia calculada.

Vê-se que a *entropia esperada mínima* é a relativa ao atributo *Investidor* que será então o que proporciona um maior ganho em informação. Prosseguindo, ver-se-ia que o maior ganho em informação estaria associado ao *País de Origem*: dois registos que, só nesse caso, se traduziriam em classes diferentes. Nestas condições, deverá ser utilizado o atributo País de Origem na partição das linhas associadas ao investidor *Mitsui*.

Os métodos distinguem-se uns dos outros pela escolha dos operadores, mas também pelos métodos de *desbaste* utilizados. No quadro seguinte apresentamos algumas distinções entre algoritmos disponíveis com relativa frequência em produtos informáticos:

Quadro 7. Diferenciação de métodos de partição

Método	CART	ID3 e C45	CHAID
Autor	Breinman	Quinlan	Kass
Tipo de árvore	A partição origina sempre dois nós descendente (Binária)	Não binária	A partição origina dois ou mais nós descendentes
Tipo de atributo	Extensivo a contínuos	Não categórico na condição; Categórico na predição	Categóricos
Associação de teste a um nó	Regras de Gini e Twoing	Função de ganho ou Função GainRatio baseadas sobre a função entropia	Diferença estatisticamente significativa e Melhor predição
Desbaste	Estimação do erro real Validação cruzada	Estimação do erro real a partir do conjunto de aprendizagem	

3.2. Processos Difusos

No final do século XIX o matemático alemão George Cantor iniciou o estudo e a aplicação da Teoría de Conjuntos, ponto de partida da matemática moderna. Chama-se conjunto a uma colecção de objectos e diz-se que os objectos individuais da colecção são elementos do conjunto ou que pertencem ao conjunto.

Nos anos 30, Lukasiewicz procurou explicar matematicamente o modo de pensar das pessoas e criou a Lógica Multi-Nivel.

Em 1965, Lofti Zadeh, professor de Engenharia Eléctrica e Ciências da Computação na Universidade de Berkeley, baseando-se no trabalho de Lukasiewicz, introduziu a Lógica Difusa baseada numa Teoría de Conjuntos Difusos segundo a qual um elemento não necessariamente pertence ou não pertence a um conjunto, existindo um continuo de graus de pertinência ou pertença.

Esta ambiguidade associada à pertença ou não a um determinado conjunto tem grande afinidade com o modo como os analistas pensam já que frequentemente são forçados a uma conclusão com base em classificadores sem uma rigorosa definição do seu conteúdo (exemplo *elevado, muito elevado, baixo.*

Vamos apresentar algumas particularidades dos processos analíticos difusos, nomeadamente, as referentes às variáveis e operadores difusos, às implementações desses operadores em sistemas computacionais, o que se concretiza na inferência difusa, e à explicitação de resultados de processos difusos.

3.2.1. Características de Processos Analíticos Difusos

O conceito formal de conjunto difuso, fundamento da lógica difusa, foi introduzido em Zadeh (1965). Apesar da sua importância, durante alguns anos, a comunidade científica não reconheceu o real valor da contribuição de Lofti Zadeh. Mas a sua determinação em termos de investigação e divulgação prevaleceu, de tal modo que, passados alguns anos, viria a apresentar um repositório elaborado de conceitos associados a processos difusos em Zadeh (1996), obra em que se encontram trabalhos seleccionados e um resumo

bibliográfico. Nesse momento, já outros autores se debruçavam sobre a problemática difusa e os anos seguintes marcaram efectivas implementações de produtos orientados por esta filosofia, acompanhadas por novas contribuições científicas.

Na teoria clássica dos conjuntos, os conjuntos são ditos *crisp:* um dado elemento do universo do discurso (domínio) pertence ou não pertence ao referido conjunto. Na teoria dos conjuntos difusos existe um grau de pertença de cada elemento a um determinado conjunto.

Um *conjunto difuso A* definido no universo de discurso U é caracterizado por uma função de pertença μA, a qual associa os elementos de U ao intervalo [0,1].

$$\mu A: U => [\,0,1\,] \tag{6}$$

Desta forma, a *função de pertença* associa a cada elemento x, pertencente a U, um número real $\mu A(x)$ no intervalo [0,1] que representa o grau de possibilidade de que o elemento x venha a pertencer ao conjunto A, isto é, o quanto é possível para o elemento x pertencer ao conjunto A.

No caso de conjuntos *crisp*, uma lista contendo os seus elementos é suficiente para a sua definição:

$$A=\{x_1,x_2,x_3,\ldots,x_n\} \text{ ou} \tag{7}$$

$$A = \sum_{i=1}^{n} x_i \tag{8}$$

No caso do conjunto representar um intervalo contínuo, por exemplo, o conjunto dos números reais menores ou iguais a quinze e maiores ou iguais a cinco, teríamos:

$$A = \int_{5}^{15} x \tag{9}$$

Note-se que, aqui, os símbolos Σ e $\int$ são usados para representar a operação de união lógica, e portanto não representam os operadores de somatório e de integração no senso do uso comum.

No caso de conjuntos difusos, a definição só estará completa com o conhecimento do grau de pertença de cada elemento. O grau de pertença $\mu_A(x_i)$ indica o quanto o elemento x_i pertence ao conjunto difuso A. Assim, temos:

$$A = \{ \mu_A(x_1)/x_1 , \mu_A(x_2)/x_2, \mu_A(x_3)/x_3 \ldots \mu_A(x_n)/x_n \} \quad (10)$$

ou

$$A = \sum_{i=1}^{n} \mu_A(x_i) / x_i \quad (11)$$

onde "/" é apenas um separador. De forma similar, quando o universo de discurso U for contínuo, o conjunto nebuloso A é frequentemente escrito como:

$$A = \int_U \mu_A(x) / x \quad (12)$$

Exemplo 12. O conjunto dos números naturais aproximadamente iguais a 10, poderia ser representado pelo conjunto nebuloso A descrito como

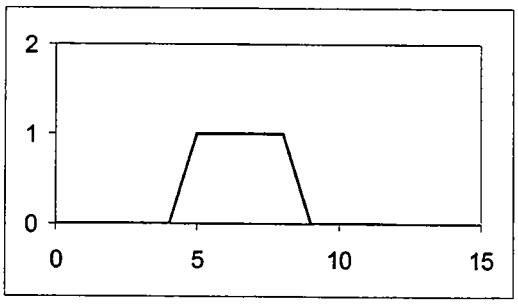
A= {0.2/6, 0.4/7, 0.6/8, 0.8/9, 1/10, 0.8/11, 0.6/12, 0.4/13, 0.2/14 }

Na maioria das aplicações práticas, as funções de pertença são do tipo gaussiano, triangulares, trapezoidais, crescentes ou decrescentes.

3.2.2. Operadores da lógica difusa

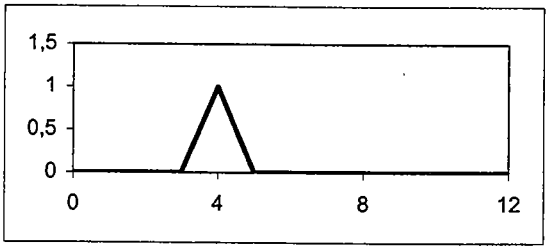
União, Intersecção e Negação são os operadores fundamentais da lógica difusa. Considere-se A um intervalo difuso entre 5 e 8 com uma função de pertença representada de acordo com a figura seguinte.

Figura 16. Função de Pertença do conjunto difuso entre 5 e 8



Considere-se B um número difuso perto de 4 com uma função de pertença representada de acordo com a figura seguinte.

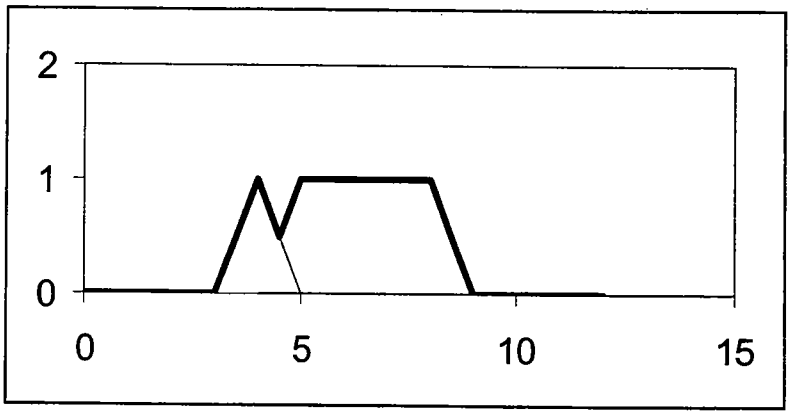
Figura 17. Função de Pertença do número difuso 4



O conjunto difuso resultante da *união* de dois conjuntos difusos é aquele cuja função de pertença é composta pelos máximos das funções de pertença de A e de B para todo x pertencente a U .

$A \cup B = \int [\mu_A(x) \vee \mu_B(x)]/x$ (13) ou $\mu_A(x) \vee \mu_B(x) = Max[\mu_A(x), \mu_B(x)]$ (14)

Figura 18. Função de Pertença da União

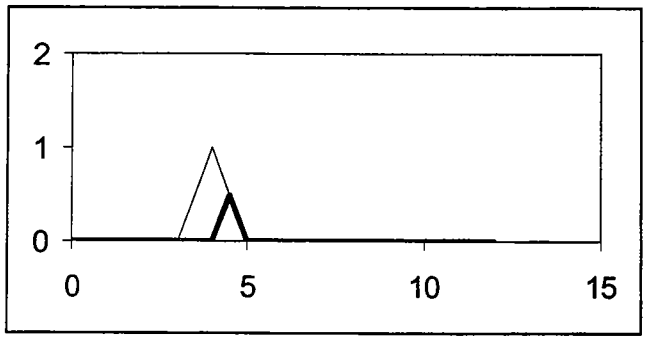


O conjunto difuso resultante da *intersecção* de dois conjuntos difusos é aquele cuja função de pertença é composta pelos mínimos das funções de pertença de A e de B para todo x pertencente a U.

Ela surge representada a traço mais grosso na figura seguinte.

$A \cap B = \int [\mu_A(x) \wedge \mu_B(x)]/x$ (15) *ou* $\mu_A(x) \wedge \mu_B(x) = Min[\mu_A(x), \mu_B(x)]$ (16)

Figura 19. Função de Pertença da Intersecção.



Chama-se *negação* a todos os elementos do universo do discurso que não pertencem ao conjunto dado originalmente. Em termos de lógica difusa, o grau de pertença da negação calcula-se através da diferença para 1.

$$\mu_{\bar{A}}(x) = 1 - \mu_A(x) \qquad (17)$$

3.2.3. Inferência Difusa

A *inferência difusa* é um processo que, avaliando os graus de pertença dos dados de entrada a determinados conjuntos, determina o seu grau de pertença ao conjunto resultado, pela utilização de um conjunto de regras. Estas regras de inferência são regras expressas sob a forma *IF condição THEN expressão* e elas próprias podem ser difusas.

Para se chegar a um resultado, um dos problemas num processo difuso é a obtenção de uma base de regras, podendo ser apresentadas três soluções alternativas:

- a construção pelo analista, o que implica um conhecimento profundo do problema;
- a extracção de bases de dados através de uma classificação e
- a extracção de bases de dados pelo recurso a redes neuronais.

Tal como referimos na secção anterior, a extracção de regras de bases de dados através de uma classificação pode ser feita de utilizando procedimentos manuais ou algoritmos matemáticos como o Classification and Regression Trees (CART), ou o Chisquare Automatic Interaction Detection (CHAID), ou o

C4.5, etc., existindo uma descrição de um procedimento manual em Parsaye (1996).

O recurso a redes neuronais é fundamentado por Kasabov (1996) e Altrock (1995) originando o que se pode designar inferência neuro-difusa.

Dispondo de uma base de regras e conhecendo os valores das variáveis de input, é possível chegar a conclusões utilizando operadores difusos ou derivados dos operadores difusos, nomeadamente, os designados por MinMax, MinAvg e Gamma apresentados por Altrock (1995). Enquanto a regra determina o conjunto a que o resultado pertence, o operador determina o grau de pertença (μ).

A forma paramétrica do operador MinMax é uma combinação linear dos operadores Mínimo e Máximo expressa por:

$$\mu = \lambda.\max_{i=1}^{m}\{\mu_i\} + (1-\lambda).\min_{i=1}^{m}\{\mu_i\} \qquad (18)$$

Na expressão anterior, se λ (considerado um actor de compensação) for 0 obter-se-á o operador AND. Deste modo o resultado de uma avaliação é sempre uma combinação linear do AND e do OR (λ=1).

A expressão do operador MinAvg corresponde a uma combinação linear entre a média dos graus de pertença e o seu mínimo:

$$\mu = \frac{\lambda}{n}.\sum_{i=1}^{m}\mu_i + (1-\lambda).\min_{i=1}^{m}\{\mu_i\} \qquad (19)$$

Entende-se que o operador Gamma é o que melhor representa o processo de avaliação humano (Zimmermann, 1983), sendo a sua expressão matemática a seguinte:

$$\mu = \left(\prod_{i=1}^{m}\mu_i\right)^{(1-\lambda)}.\left(1-\prod_{i=1}^{m}(1-\mu_i)\right)^{\lambda} \qquad (20)$$

3.2.4. Explicitação de Resultados de Processos Difusos

A explicitação dos resultados de um processo analítico difuso pode basicamente ser feita de acordo com três princípios distintos, apresentados em Altrock (1997):

- explicitação difusa;
- melhor compromisso e

- solução mais plausível

No primeiro caso, *Explicitação Difusa*, o resultado consiste na revelação do grau de pertença do resultado a cada um dos conjuntos ou termos possíveis.

O resultado derivado do *Melhor Compromisso* é baseado no valor típico associado a determinada variável linguística. Nesta ponderação os pesos são os graus de pertença.

A *Solução Mais Plausível* resulta da consideração como resultado com probabilidade máxima da solução que detenha a maior probabilidade admitindo-se três hipóteses: ponto médio do máximo, esquerda do máximo e direita do máximo.

3.3. *Pesquisa de Associações*

Os processos de afinidade lógica são distintos dos de lógica condicional quer em termos da linguagem de expressão quer das estruturas de dados que usam.

A Análise de Afinidade (ou Análise de Associação) é a procura de padrões e condições que descrevem como vários itens *se agrupam juntos* ou *acontecem juntos* em séries de eventos ou transacções. Uma regra de afinidade tem a forma:

Quando *Item1* ***Também*** *Item2.*

O problema considerado é encontrar regras de associação a partir de dados binários que, por seu lado, terão sido obtidos a partir de ficheiros de transacções organizados de acordo com os modelos usuais, nomeadamente o modelo relacional.

Esta perspectiva tem tido um considerável desenvolvimento a partir de grupos de investigação do IBM Almaden Research Center (Rakesh Agrawal), do Department of Computer Science da Universidade de Helsinquia (Heikki Mannila e Hannu Toivonnen) e, mais recentemente, do grupo de investigação formado em torno do projecto DBMiner no Canadá (Jiawei Han e outros).

3.3.1. Caracterização e exemplificação

Assuma-se que temos um conjunto $R = \{A_1,...,A_p\}$ de atributos binários, isto é, o domínio de cada A_i é $\{0,1\}$. Uma relação $r = \{t_1,...t_n\}$ no esquema R é uma matriz com colunas R e n linhas, sendo cada linha um vector de comprimento p cujos elementos são 0 e 1.

Uma *regra de associação* em r é uma expressão da forma $X \rightarrow B$, onde $X \subseteq R$ e $B \in R \setminus X$. O significado intuitivo da regra é que se uma linha da matriz r tem um 1 em cada coluna de X, então a linha tende a ter um 1 também na coluna B.

Exemplos de dados em que possam ser aplicadas regras de associação incluem os seguintes:

- Uma base de dados de estudantes numa universidade: as linhas correspondem a estudantes, as colunas a cursos, e um 1 na posição *(e, c)* indica que o estudante *e* frequentou o curso *c*.

- Dados recolhidos a partir de leitores de código de barras em supermercados: as colunas correspondem a produtos, e cada linha corresponde ao conjunto de produtos comprado uma vez.

- Uma base de dados de publicações: as linhas e colunas correspondem ambas a publicações e *(p,p')* = 1 significa que a publicação *p* refere a publicação *p'*.

- Uma base de dados sobre o IDE: as linhas correspondem a investidores, as colunas a sectores e um 1 na posição *(i,s)* significa que houve investimento de *i* no sector *s*.

3.3.2. Transformação prévia dos dados

A pesquisa de associações é possível a partir da transformação de uma tabela original de esquema A={Identificador de transacção, Identificador de elemento} numa tabela vulgarmente designada por Tabela_R com esquema Tabela_R = {Identificador de transacção, E1, E2,...,En} onde o identificador de transacção provém da tabela original e E1, E2, ..., En assumem valores binários (0,1)

significando a ausência ou presença do elemento En na transacção considerada.

A Tabela_R tem assim uma cardinalidade igual ou inferior à tabela original, podendo ser considerada como uma representação condensada do seu conteúdo.

Em determinados sistemas de gestão de bases de dados (exemplo, o MS Access), a determinação da Tabela_R é relativamente fácil através de uma sequência do tipo:

(início)

Insert into Tabela_R

Select identificador de transacção, identificador de elemento, count (*) from A group by identificador de transacção, identificador de elemento; ⇨

(atribuir ao campo resultado de count a designação de valor) ⇨

Update Tabela_R set valor = 1 where valor > 0; ⇨

(apresentar a Tabela_R na forma de crosstab) ⇨

(guardar os resultados anteriores para uso posterior) ⇨

(fim)

3.3.2. Frequência e confiança de Regras de Associação

Dado $W \subseteq R$, representamos por *s(W, r)* a *frequência* de *W* em *r*: a fracção de linhas de *r* que tem um 1 em cada coluna de *W*.

A *frequência da regra* $X \rightarrow B$ em *r* é definida por $s(X \cup \{B\}, r)$, e a *confiança da regra* é $s(X \cup \{B\}, r) / s(X, r)$.

Na descoberta de regras de associação, a tarefa é encontrar todas as regras $X \rightarrow B$ tais que a frequência da regra seja pelo menos um valor dado σ e a confiança seja pelo menos igual a outro valor θ.

Note-se que não há limites predefinidos ao número de atributos do lado esquerdo X duma regra de associação $X \rightarrow B$; isto é importante para que associações não esperadas não sejam desprezadas antes de o processamento se iniciar.

Significa também que o espaço de pesquisa tem um tamanho exponencial no número de elementos da relação de input o que requer algum cuidado com os algoritmos de tratamento.

Diz-se que um subconjunto $X \subseteq R$ é *frequente em r*, se $s(X, r) \geq \sigma$. Uma vez conhecidos todos os conjuntos frequentes de r, encontrar as regras de associação é fácil. Concretamente, para cada conjunto frequente X e cada $B \in X$ verificar se a regra $X \setminus \{B\} \rightarrow B$ tem uma confiança suficientemente alta.

A descoberta de todos os conjuntos frequentes pode ser feita de muitos modos diferentes. Uma abordagem típica é usar o facto de que todos os subconjuntos de um conjunto frequente são também frequentes.

- Primeiro, há que descobrir todos os conjuntos frequentes de tamanho 1 lendo os dados uma vez e registando o número de vezes que cada atributo *A* ocorre.

- Depois, formar conjuntos de tamanho 2 tomando todos os pares de atributos *{B,C}* tais que *{B}* e *{C}* sejam ambos frequentes. A frequência dos conjuntos candidatos é de novo avaliada relativamente à base de dados.

- Uma vez conhecidos os conjuntos frequentes de tamanho 2 são formados os candidatos de tamanho 3: estes são conjuntos *{B,C,D}* tais que *{B,C}*, *{B,D}*, e *{C,D}* sejam todos frequentes. Este processo continua até que não possam ser formados mais conjuntos candidatos.

3.3.3. Algoritmos de descoberta de Associações

Tal como no caso da classificação, é possível utilizar algoritmos com uma elevada componente estatística e matemática para a pesquisa de associações ou algoritmos derivados de uma utilização mais intensiva do SQL.

Na literatura científica existem múltiplas referências ao algoritmo A-Priori de que apresentamos uma descrição na parte final desta secção. Mas as mesmas referências atestam a possibilidade de uma implementação *naïve* da pesquisa de associações que ilustramos com um exemplo.

Exemplo 13. Considere-se que o quadro seguinte representa um ficheiro com as transacções de determinado período identificadas pelo número de transacção.

Quadro 8. Transacções do Período

Número de transação	*Código de Produto*
1	A
1	B
1	C
2	A
2	C
3	A
4	C
4	B
4	D
4	A
5	A

Número de transação	*Código de Produto*
5	D
6	B
7	A
8	A
8	C
8	D
9	B
9	D
10	A
10	B
10	D

A partir deste ficheiro é possível definir o conjunto $R=\{A,B,C,D\}$ e a relação r que terá o seguinte aspecto:

Quadro 9. Produtos Envolvidos em Transacções

Número de transacção	**A**	**B**	**C**	**D**
1	1	1	1	0
2	1	0	1	0
3	1	0	0	0
4	1	1	1	1
5	1	0	0	1
6	0	1	0	0
7	1	0	0	0
8	1	0	1	1
9	0	1	0	1
10	1	1	0	1

Os resultados da primeira iteração serão:

Quadro 10. Frequência dos Produtos - 1

Conjunto	Frequência absoluta
{A}	8
{B}	5
{C}	4
{D}	5

Admitam-se os seguintes valores: $\sigma = 0.1$ e $\theta = 0.8$. Com base nestes valores todos os subconjuntos são frequentes pelo que todos os formados por 2 elementos são candidatos a frequentes. Resulta da segunda iteração a seguinte tabela de frequências:

Quadro 11. Frequência dos Produtos - 2

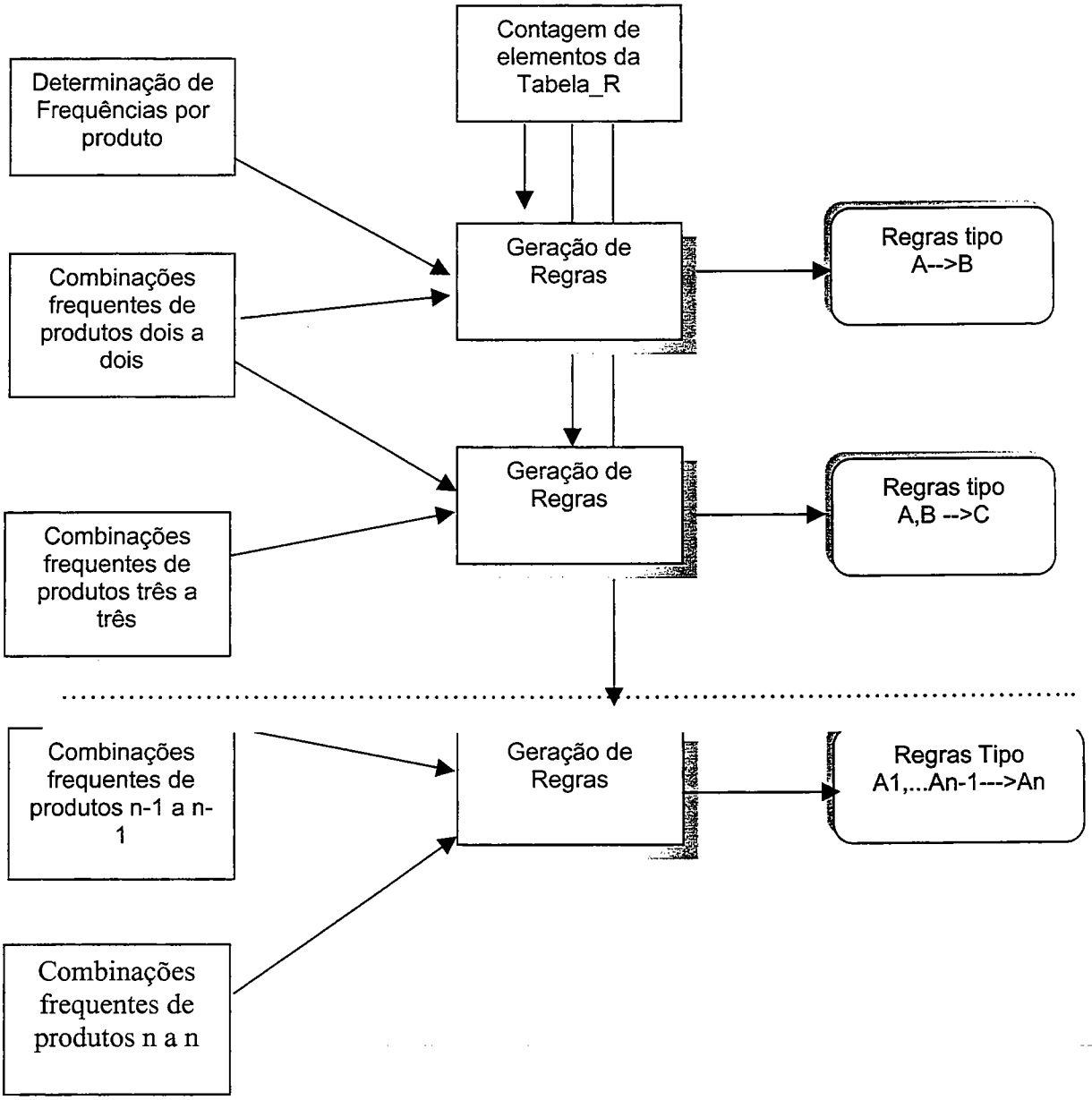
Conjunto	Frequência absoluta
{A,B}	3
{A,C}	4
{A,D}	4
{B,C}	2
{B,D}	3
{C,D}	2

Desta tabela de frequências podem considerar-se as seguintes associações candidatas:

Quadro 12. Quadro de Regras Potenciais - 1

Regra	Confiança
A→B	3/8
A→C	4/8
A→D	4/8
B→C	2/5
B→D	3/5
C→D	2/4
B→A	3/5
C→A	4/4
C→B	2/4
D→A	4/5
D→B	3/5
D→C	2/5

Figura 20. Algoritmo geral para a pesquisa de associações



3.3.3.2. O algoritmo Apriori

O problema da pesquisa de associações é apresentado pela primeira vez em Agrawal et al (1993,2) que especificou um primeiro algoritmo denominado *AIS*, um acrónimo associado ao apelido dos seus autores. Com idêntico objectivo, foi proposto por Houtsma e Swami (1993), no mesmo ano, o algoritmo *SETM*, assente no uso de SQL e exploração de bases de dados relacionais.

O algoritmo *APRIORI* surgiu em 1994, ligado a questões de performance dos anteriores, tendo sido defendido pelo grupo de trabalho de Agrawal. De acordo com Agrawal e Srikant (1994), segue a estrutura iterativa básica apresentada

anteriormente, utilizando um passo de *desbaste* que elimina qualquer subconjunto que não seja considerado frequente. O algoritmo termina num passo t em que não existam conjuntos frequentes de t elementos. Apresentamos no quadro seguinte a estrutura base do algoritmo conforme com a descrição do seu autor.

Quadro 16. O algoritmo Apriori

L_1 = {large 1-itemsets}:
for ($k = 2$; $L_{k-1} \neq 0$; $k++$)
    /*Candidate Item sets generation*/
    Ck = Set of New Candidates;
    /*Support Counting*/
    for all transactions $t \varepsilon D$
        for all k-subsets s of t
            if ($s \varepsilon C_k$) *s.count* + +;
    /*Large itemsets generation*/
    $L_k = \{c \varepsilon C_k \mid c.count \geq$ minimum support$\}$;
Set of all large itemsets$= U_k L_k$;

Fonte: Agrawal e Srikant (1994)

No quadro, L_k representa o conjunto de subconjuntos frequentes com k elementos e C_k é o conjunto de candidatos a subconjuntos frequente com k elementos os quais foram gerados pela junção de L_{k-1} consigo próprio. Antes da inserção em C_k, é feito o teste de frequência pela comparação com o suporte mínimo e o desbaste é realizado se não for satisfeito.

O *APRIORI* é referido em múltiplos trabalhos científicos dedicados ao estudo da pesquisa de associações. Mas, apesar do esforço da equipa de Agrawal pensamos que, pelas características de que esse trabalho se reveste, a divulgação operada nos trabalhos de Toivonen e Jiawei Han tem vindo a contribuir para um melhor conhecimento sobre a natureza da pesquisa de associações.

3.3.3.3. O algoritmo Similis

A pesquisa de associações tem vindo a dar conteúdo a algum trabalho científico em Portugal. É, assim, que se torna possível constatar a existência

de mais um algoritmo nesta área resultado de trabalho de investigadores portugueses.

Este algoritmo, apresentado em Cavique e Themido (2001), tem três passos que resumimos nas linhas seguintes:

1. Com base numa tabela de transações, gerar a matriz binária [M(cliente, produto)].

2. Com base na matriz binária *M*, gerar um grafo ponderado G(V,A) onde o peso de cada aresta representa a semelhança do par de produtos e a frequência de cada produto.

3. Encontrar, no grafo G(V,A), a clique de maior peso de dimensão k, que corresponde ao cabaz de compras mais frequente.

Os autores procedem a uma comparação com o algoritmo Apriori que resumimos no quadro seguinte.

Quadro 17. Comparação Apriori / SIMILIS

	Apriori	SIMILIS
Dados	tabela original	grafo ponderado
Método	exacto (crescimento exponencial)	aproximado
Informação devolvida	perfeita	quase-perfeita
Volume de dados	pequeno	médio e grande
Complexidade temporal depende	do número de produtos, dimensão tabela, dimensão cabaz	do número de produtos

Fonte: Cavique e Themido (2001)

3.4. Processos de pesquisa de Clusters

Os processos de *clustering* são utilizados para procurar grupos de pontos similares num determinado espaço de pesquisa, operação em que usam algoritmos de diversa natureza, concretizando um tipo de aprendizagem designada por *não supervisionada*, isto é, em que não existe um *professor* para o sistema.

Esta procura de pontos similares vai-nos permitir definir grande grupos de Sistemas de Informação para a Economia e ver o seu comportamento no tempo na fase final desta dissertação.

Na secção, começaremos com uma caracterização deste processo analítico, examinando em seguida a transformação prévia dos dados que necessita de concretizar. Terminamos com a descrição de algoritmos para agregação ou desagregação de casos.

3.4.1. Caracterização

Dado um conjunto de *n* indivíduos para os quais existe informação sobre a forma de *p* variáveis, o método de análise de *clusters* procede ao agrupamento dos indivíduos em função da informação existente, de tal modo que os indivíduos pertencentes a um mesmo grupo sejam tão semelhantes quanto possível e sempre mais semelhantes aos elementos do mesmo grupo do que a elementos dos restantes grupos. Este método é também chamado de método de partição, classificação ou taxonomia, embora o termo partição seja mais utilizado para uma das técnicas específicas da análise: aquela em que os indivíduos são divididos por um número preestabelecido de grupos.

Genericamente, a análise de *clusters* compreende cinco etapas: (1) selecção de indivíduos ou de uma amostra de indivíduos a serem agrupados; (2) definição de um conjunto de variáveis a partir das quais será obtida a informação necessária ao agrupamento dos indivíduos; (3) definição de uma medida de semelhança ou distância entre cada dois indivíduos; (4) escolha de um critério de agregação ou desagregação dos indivíduos, isto é, a definição de um algoritmo de partição / classificação; (5) validação dos resultados encontrados.

Discutem-se nos pontos seguintes os aspectos fundamentais de cada uma destas etapas (à excepção da primeira que, naturalmente, não diz respeito apenas ao método, dependendo de cada problema concreto de classificação).

3.4.2. Transformação prévia dos dados

Os métodos de partição exigem uma transformação prévia dos dados em índices de (dis)semelhança.

3.4.2.2. Coeficientes de Associação

Estes coeficientes são particularmente úteis para definir a semelhança entre indivíduos caracterizados por variáveis qualitativas de tipo booleano (variáveis binárias, por exemplo, em que 0 significa ausência da característica para determinado indivíduo e 1 significa presença da característica).

De acordo com Reis (1997), mais de trinta coeficientes de associação foram já propostos por vários autores. Alguns deles merecem tratamento particular: (1) os coeficientes de emparelhamento simples, (2) os coeficientes de Jaccard e (3) o coeficiente de Gower.

Considerem-se os indivíduos i e j caracterizados por p variáveis binárias e admita-se a possibilidade de construção de uma tabela de contingência onde

a corresponde ao número de características que assumem o valor 1 (estão presentes) em ambos os indivíduos,

b o número de características com valor 1 no indivíduo j e 0 no indivíduo i,

c o número de características presentes em i mas ausentes em j, e

d as características simultaneamente ausentes em i e j.

Quadro 19. Coeficientes de Associação

Coeficiente	Expressão	Breve Descrição
Coeficientes de Emparelhamento	$S_{ij} = \frac{(a+d)}{(a+b+c+d)}$ (31)	É a relação entre o número de características presentes e ausentes simultaneamente para os dois indivíduos e o número total de características, varia entre 0 e 1 e mede a semelhança entre cada dois indivíduos.
	$d_{ij} = \frac{(b+c)}{a+b+c+d}$ (32)	É o quociente entre o número de características presentes num dos indivíduos e ausentes no outro e o número total de características, varia entre 0 e 1 e mede a dissemelhança entre os dois indivíduos.
Coeficientes de Jaccard	$s_{ij} = \frac{a}{a+b+c}$ (33)	Estes coeficientes evitam a contribuição da ausência conjunta de uma característica para o cálculo da semelhança ou dissemelhança entre dois indivíduos.
	$d_{ij} = \frac{b+c}{a+b+c}$ (34)	

Quadro 19. Coeficientes de Associação

Coeficiente	Expressão	Breve Descrição
Coeficiente de Gower	$S_{ij} = \frac{\sum_{v=1}^{p} s_{ijv}}{\sum_{v=1}^{p} w_{ijv}}$ (35)	Este coeficiente permite a utilização de variáveis definidas em diferentes escalas de medida (binárias, nominais, ordinais e contínuas) mas torna-se idêntico ao coeficiente de Jaccard quando as variáveis são todas binárias. s_{ijv} – valor da semelhança entre os indivíduos *i* e *j* para a variável *v*[12] w_{ijv} – ponderação a afectar à variável *v*: 1 se a comparação para a variável v for válida 0 se a comparação não for considerada válida; exemplo, uma *não resposta* para a variável em causa

Fonte: Reis (1997)

Ainda de acordo com Reis (1997), muitos outros coeficientes de semelhança foram já propostos (Johnson e Wichern (1988)), cada um com a sua própria justificação:

Quadro 20. Outros Coeficientes de Semelhança

Coeficiente		Justificação
2(a+d) / (2(a+d)+b+c)	(36)	Peso duplo às presenças e ausências simultâneas.
(a+d) / (a+d + 2(b+c))	(37)	Peso duplo às situações discordantes, inclusão das ausências simultâneas.
2a / (2a + b + c)	(38)	Peso duplo às presenças simultâneas, exclusão das ausências simultâneas.
a / (a + 2 (b + c)	(39)	Peso duplo às situações discordantes, exclusão das ausências simultâneas.
a / (b + c)	(40)	Quociente entre presenças simultâneas e situações discordantes, exclusão das ausências simultâneas.

Fonte: Reis (1997)

[12] O cálculo de s_{ijv} para o caso de variáveis binárias, nominais e ordinais é semelhante: se $X_{iv}=X_{jv}$ então $s_{ijv}=1$ senão $s_{ijv}=0$. No caso de variáveis contínuas, Reis (1997) apresenta a expressão $s_{ijv} = 1 - |X_{iv} - X_{jv}| / R_v$, onde R_v é o coeficiente de variação da variável *v*.

3.4.3. Algoritmos de Agregação e Desagregação dos Casos

Escolhida uma medida de distância, surge o quarto problema a resolver em qualquer análise de *clusters*: a escolha do critério de (des)agregação dos indivíduos. Poder-se-á dizer que os vários métodos pretendem responder, de forma diferente, aos seguintes problemas:

- Distância entre indivíduos do mesmo grupo e distância entre indivíduos de grupos diferentes.
- Dispersão dos indivíduos dentro do grupo.
- Densidade dos indivíduos dentro e fora dos grupos.

Os vários métodos de agregação dos indivíduos diferem no modo como definem distâncias entre grupos já formados e outros grupos ou indivíduos por agrupar (que se podem considerar grupos formados por um elemento). O processo de agrupamento de grupos depende da distância entre os grupos. Portanto, diferentes definições destas distâncias poderão resultar em diferentes soluções finais.

Não existe aquilo a que se possa chamar o melhor critério de (des)agregação dos casos em análise de *clusters*. É prática comum utilizar vários critérios e fazer a comparação dos resultados. Se estes forem semelhantes, é possível concluir que se obtiveram resultados com elevado grau de estabilidade e, portanto, fiáveis. Os critérios de agregação mais utilizados são os seguintes:

Single linkage ou critério do vizinho mais próximo

Este critério define como semelhança entre dois grupos a semelhança máxima entre quaisquer dois casos pertencentes a esses grupos. Dados dois grupos (i, j) e *(k),* a distância entre os dois é a menor das distâncias entre os elementos dos dois grupos:

$$d(i,j)k = \min \{ d_{ik}; d_{jk} \} \quad (41)$$

Complete linkage ou critério do vizinho mais afastado

Este critério utiliza o procedimento inverso ao anterior, uma vez que a distância entre dois grupos é agora definida como sendo a distância entre os seus elementos mais afastados ou menos semelhantes.

Dados dois grupos *(i, j)* e *(k)*, a distância entre eles é a maior das distâncias entre os seus elementos:

$$d(i,j)k = \max (dik; djk) \quad (42)$$

Critério da média dos grupos

Esta estratégia de agrupamento define a distância entre dois grupos como sendo a média das distâncias entre todos os pares de indivíduos constituídos por elementos dos dois grupos.

Critério do centróide

Nesta estratégia, a distância entre dois grupos é definida como a distância entre os seus centróides, pontos definidos pelas médias das variáveis caracterizadoras dos indivíduos de cada grupo, isto é, o método do centróide calcula a distância entre dois grupos como a diferença entre as suas médias, para todas as variáveis.

Critério de Ward

Este método pode ser resumido nas seguintes etapas:

1. Calcular as médias das variáveis para cada grupo.
2. Calcular os quadrados das diferenças entre essas médias e os valores das variáveis para cada indivíduo dentro dos grupos.
3. Somar as diferenças calculadas para todos os indivíduos dentro dos grupos.
4. Por último, optimizar a variância mínima dentro dos grupos. A função objectivo que se pretende minimizar é também chamada soma dos quadrados dos erros *(ESS)* ou soma dos quadrados dentro dos grupos *(WSS)*.

No início do processo de agrupamento, cada indivíduo constitui um grupo e *WSS* = 0. Em seguida, são agrupados os dois indivíduos que provocam um aumento mínimo no valor da soma dos quadrados dos erros, passando a existir *n* - 1 grupos; estes *n* - *1* grupos são, então, reexaminados e transformados em *n* - 2 grupos, mas de tal modo que o aumento na função objectivo seja minimizado; o processo continua de forma sistemática até todos os indivíduos formarem um grupo apenas.

Quadro 21. Resumo comparativo dos critérios de aglomeração

Método	Característica	Consequência	Vantagem	Desvantagem
Single Linkage	Qualquer grupo é definido como o conjunto de casos em que qualquer elemento é mais semelhante a pelo menos um outro elemento do mesmo grupo do que a qualquer elemento de outro grupo	Este método torna-se, assim, um sistema contractor do espaço uma vez que cada caso terá mais tendência para se agrupar a um grupo já definido do que para formar o núcleo de um novo grupo	1) é insensível a transformações monótonas da matriz de distâncias 2) não é afectado pela existência de relações nos dados iniciais	Os agrupamentos finais assemelhem-se a cadeias de elementos quando representados num espaço multidimensional
Complete Linkage	Cada grupo passa a ser definido como um conjunto de elementos em que cada um é mais semelhante a todos os restantes elementos do grupo do que a qualquer dos elementos dos restantes grupos	Tem tendência para encontrar *clusters* compactos compostos de indivíduos muito semelhantes entre si	Os resultados da aplicação deste método dão uma visão nítida dos diferentes grupos encontrados	1) Os resultados da aplicação nem sempre apresentam um elevado grau de concordância com a estrutura inicial dos dados 2) Tendência para encontrar grupos esféricos
Média dos Grupos	Um grupo passa a ser definido como um conjunto de indivíduos no qual cada um tem mais semelhanças, em média, com todos os membros do mesmo grupo do que com todos os elementos de qualquer outro grupo		1) evita valores extremos e 2) toma em consideração toda a informação dos grupos	Apresenta tendência para encontrar grupos esféricos
Centróide	A distância entre dois grupos e definida como a distância entre os seus centróides, pontos definidos pelas médias das variáveis caracterizadoras dos indivíduos de cada grupo	Com este método, o centróide do novo grupo é uma combinação ponderada dos centróides dos dois grupos separados, sendo as ponderações proporcionais ao tamanho destes grupos	Combina os *clusters* de tal modo que distância média entre todos os pares possíveis de indivíduos dentro do grupo daí resultante seja mínima	Se os dois grupos forem muito diferentes em termos de dimensão, o centróide do novo agrupamento estará mais próximo daquele que for maior e as características do grupo menor tenderão a perder-se
Ward	Baseia-se na perda de informação resultante do agrupamento dos indivíduos e medida através da soma dos quadrado dos desvios das observações individuais relativamente às médias do grupos em que são classificadas			1) A tendência a criar grupos de tamanho semelhante e a encontrar soluções que podem ser ordenadas a partir dos perfis relativamente às variáveis iniciais 2) A tendência para encontrar grupos esféricos

Fonte: Reis (1997)

4. Outros processos utilizados na dissertação

Nesta secção, vamos apresentar brevemente dois processos utilizados na dissertação, mas que não integramos vulgarmente nos processos analíticos apesar de um deles, o segundo, apresentar grandes afinidades com a classificação: a análise factorial pelo método das componentes principais e a análise discriminante.

O primeiro destes processos é muitas vezes usado para permitir a actuação de outros de uma forma mais rigorosa e é nesse sentido que o vamos aplicar antes de uma pesquisa de clusters. O segundo irá ser utilizado na associação de uma classe aos SIE: primeiro geram-se funções de discriminação, posteriormente associa-se uma classe.

4.1. A análise factorial

Nas linhas que se seguem, apresentamos os objectivos da análise factorial e todo um conjunto de passos para a sua concretização.

4.1.1. Objectivos da análise

A análise factorial procura identificar variáveis ou factores que explicam o padrão de correlações num conjunto de variáveis observadas. A análise factorial é muitas vezes utilizada na redução de dados para identificar um pequeno número de factores que explicam a maior parte da variância observada num número maior de variáveis conhecidas. Pode também ser utilizada para gerar hipóteses relativas a mecanismos causais ou para visualizar variáveis para análises subsequentes.

Como resultado desta análise chegamos a um modelo que se pode escrever nos moldes seguintes:

$$X_1 = b_{11} F1 + b_{12} F_2 + \ldots + b_{1k} F_k + U_1$$

$$X_2 = b_{21} F1 + b_{22} F_2 + \ldots + b_{2k} F_k + U_2 \qquad (43)$$

...

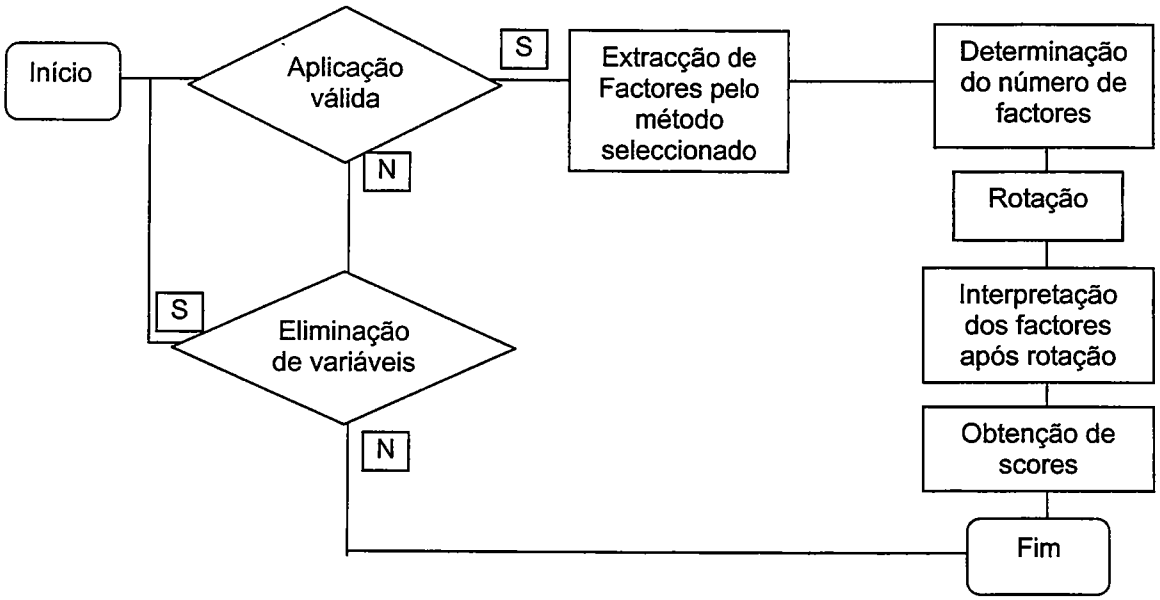
$$X_p = b_{p1} F1 + b_{p2} F_2 + \ldots + b_{pk} F_k + U_k$$

onde os bij (com i=1,...p; j=1,...,k; e k<p), também designados por *factor loadings* são as correlações entre os factores comuns e as variáveis observáveis. Ui, i=1,....,p é o factor único, representando a parte de uma variável que não é explicada pelos factores comuns. Os factores, independentemente da sua espécie, não estão relacionados entre si nem com os de outra espécie.

4.1.2. Passos para a análise

Na realização de uma análise factorial podemos considerar o seguinte conjunto de passos. Os factores obtidos podem ser utilizados em processos posteriores.

Figura 21. Concretização de uma análise factorial



4.1.2.1. Validade de aplicação

A utilização de uma *Análise Factorial* carece de uma validação prévia do processo, através de um conjunto de testes cujo resultado pode levar à eliminação de variáveis ou à sua não concretização.

1. Teste à normalidade das variáveis. Um resultado negativo põe a hipótese de exclusão de análise da respectiva variável.
2. Análise do *diagrama de extremos e quartis* (*caixa de bigodes*). Permite a detecção *outliers* que podem ou não ser mantidos na análise.

3. Exame da *matriz de correlações*. A análise factorial exige uma elevada correlação, se tal não acontecer a realização da análise pode ser posta em causa.

4. Teste de esfericidade de *Bartlett*. De acordo com este teste, não existirá correlação entre as variáveis se a matriz de correlações for a matriz identidade.

5. Exame do indicador designado por *Kaiser-Meier-Olkin* (KMO) que procede à comparação das correlações simples com as parciais observadas entre as variáveis. Um valor não muito afastado de 1 permite considerar que se pode prosseguir a análise factorial.

6. Análise da *matriz anti-imagem* que mede a adequação de cada variável para uso da análise factorial. Se alguma variável tiver na diagonal principal um valor perto de zero pode ser excluída de análise

7. Exame das *comunalidades* ou variância de cada variável explicada pelas factores comuns (as componentes principais se for este o método de extracção) cujo resultado é um valor entre 0 e 1. Leva à exclusão de uma variável da análise se apresentar um valor perto de 0.

4.1.2.2. Extracção de factores

Existem vários métodos para a extracção de factores. O SPSS, um dos produtos que utilizamos nesta dissertação, disponibiliza sete métodos para a concretização da análise factorial, cujas características aqui apresentamos de acordo com o seu *help*.

(1) **Principal components**. A factor extraction method used to form uncorrelated linear combinations of the observed variables. The first component has maximum variance. Successive components explain progressively smaller portions of the variance and are all uncorrelated with each other. Principal components analysis is used to obtain the initial factor solution. It can be used when a correlation matrix is singular.

(2) **Unweighted leastsquares**. A factor extraction method that minimizes the sum of the squared differences between the observed and reproduced correlation matrices ignoring the diagonals.

(3) **Generalized least squares**. A factor extraction method that minimizes the sum of the squared differences between the observed and reproduced correlation matrices. Correlations are weighted by the inverse of their uniqueness, so that variables with high uniqueness are given less weight than those with low uniqueness.

(4) **Maximum likelihood**. A factor extraction method that produces parameter estimates that are most likely to have produced the observed correlation matrix if the sample is from a multivariate normal distribution. The correlations are weighted by the inverse of the uniqueness of the variables, and an iterative algorithm is employed.

(5) **Principal axis factoring**. A method of extracting factors from the original correlation matrix with squared multiple correlation coefficients placed in the diagonal as initial estimates of the communalities. These factor loadings are used to estimate new communalities that replace the old communality estimates in the diagonal. Iterations continue until the changes in the communalities from one iteration to the next satisfy the convergence criterion for extraction.

(6) **Alpha factoring**. A factor extraction method that considers the variables in the analysis to be a sample from the universe of potential variables. It maximizes the alpha reliability of the factors.

(7) **Image factoring**. A factor extraction method developed by Guttman and based on image theory. The common part of the variable, called the partial image, is defined as its linear regression on remaining variables, rather than a function of hypothetical factors.

Nesta dissertação, apenas utilizaremos o método das componentes principais, um método estatístico multivariado que permite transformar um conjunto de variáveis iniciais correlacionadas entre si num outro conjunto de variáveis não correlacionadas (ortogonais), as componentes principais, que resultam de combinações lineares do conjunto inicial (Reis, 1997). As componentes vão ser determinadas por forma que a primeira explique o máximo possível da variância dos dados, a segunda explique o máximo da variância dos dados não explicada pela primeira e assim sucessivamente para as diferentes componentes.

As componentes principais correspondem aos factores comuns da análise factorial quando esta usa o método referido.

A variância dos dados é medida recorrendo ao conceito de variância total que, de acordo com Silvestre (2000), corresponde à soma das variâncias das variáveis utilizadas na análise. Encontramos em Silvestre (2000) e Reis (1997) uma explicação cuidada da extracção de factores a partir de uma matriz de covariâncias e a partir de uma matriz de correlações o que leva a resultados distintos.

Existe software que permite chegar a estes resultados pelo que não vamos apresentar qualquer fundamentação matemática dos mesmos a qual está presente nas referências citadas.

4.1.2.3. Determinação do número de factores necessários

Na literatura, é possível encontrar um conjunto relativamente variado de critérios para responder a esta questão[13].

1. **Proporção da variância explicada**. O número de factores deve ser tal que a proporção da variância total explicada pelos mesmos seja pelo menos igual a um certo valor. De acordo com Silvestre (2000), *0.8 é um valor que os investigadores consideram frequentemente.* Para Reis (1997), que fala num valor superior a 70% da variância total, *este é o critério mais subjectivo, existindo divergências quanto ao limiar mínimo da variância explicada, podendo levar à retenção de factores com peso significativo de apenas uma variável.*
2. **Critério de Kaiser** ou do valor próprio superior à unidade. Segundo este critério, aplicável somente quando se recorre à matriz de correlações para extracção dos factores, devem reter-se tantos factores quantos os valores próprios superiores à unidade (factores cuja variância é superior a 1).
3. **Critério de Joliffe**. Este autor, citado por Silvestre (2000), considera o critério de Kaiser muito exigente pelo que considera a utilização do valor crítico 0.7 em vez de 1.
4. **Critério gráfico do *scree plot***, proposto por Catell, citado por Silvestre (2000). Os valores próprios são representados contra a sua ordem. Unindo os pontos com segmentos de recta obtém-se uma linha poligonal que apresenta um declive acentuado para os primeiros valores próprios passando depois a um declive que se aproxima de zero. Quando se dá uma mudança brusca no declive encontramos a ordem da última componente a considerar.

[13] Para além dos critérios que vamos apresentar, há que dizer que Reis (1997) refere um critério mais formal, a aplicar quando as componentes principais são derivadas a partir de uma matriz de variância e covariância amostral, e que consiste em reter apenas as componentes principais cuja variância é significativamente diferente de zero. O critério, que não será utilizado nesta dissertação, apoia-se num procedimento desenvolvido por Bartlett destinado a testar a hipótese de que os últimos p-k valores próprios da matriz de variância e covariância são iguais.

4.1.2.4. Rotação dos factores com o objectivo de interpretação

A aplicação de um método de rotação tem como objectivo principal a transformação dos coeficientes dos factores principais retidos numa estrutura simplificada consistindo num conjunto de vectores, cada um deles com coeficientes não nulos para um reduzido número de variáveis e sem sobreposições relativamente aos restantes vectores. O conjunto inicial de variáveis fica assim dividido em subconjuntos tão independentes entre si quanto possível (Reis, 1997).

Ainda de acordo com Reis (1997), tal estrutura simplificada é de muito difícil obtenção, não existindo sequer a garantia de que as factores se mantenham não-correlacionadas depois da rotação.

1. **Rotação VARIMAX**. É um método proposto por Kaiser, citado por Reis (1997), que pretende que para cada factor comum existam apenas alguns pesos significativos e todos os outros sejam próximos do zero com o objectivo de maximizar a variação entre os pesos de cada factor comum. O novo conjunto de pesos aumenta a contribuição das variáveis que mais contribuem para a formação do factor e diminui os pesos das que menos contribuem. A obra citada refere que, de acordo com demonstrações feitas por F. Gebhardt em 1968, J. Deleeuw e S. Pruzanski em 1978 e J.M.F. TenBerge em 1984, se prova ser o melhor método de rotação.
2. **Rotação QUARTIMAX**. Este método pretende simplificar as linhas de uma matriz de pesos, tendo por objectivo tornar os pesos de cada variável elevados para um número reduzidos de factores e próximos de zero para todas as restantes factores.
3. **Rotação EQUIMAX**. Esta rotação, também do tipo ortogonal, pretende ser uma solução de compromisso entre os dois métodos anteriores, tendo por objectivo simplificar simultaneamente linhas e colunas em vez de se concentrar nas linhas ou nas colunas da matriz de pesos.
4. **Métodos oblíquos**. Neste caso não se mantém o pressuposto de independência entre as factores sendo-lhes permitido que rodem livremente de maneira a simplificarem o agrupamento das variáveis e a interpretação dos factores.

4.1.2.5. Interpretação dos factores

A identificação e interpretação de cada factor é efectuada a partir dos pesos das variáveis que contribuem para ele. Quanto mais próximo o peso estiver de 1 em valor absoluto, mais forte é a associação entre a variável e o factor, sendo considerados significativos pesos iguais ou superiores a 0,5.

Existirá sempre alguma subjectividade no interpretação associada à combinação linear de variáveis e bem assim à designação atribuída à componente principal.

4.1.2.6. Obtenção da matriz de scores individuais

Consiste na expressão das anteriores variáveis, relativamente a cada indivíduo, em termos de cada factor. A visualização gráfica dos indivíduos utilizando dois eixos permite uma interpretação das relações existentes entre os mesmos.

Os valores assim gerados podem ser utilizados em análises futuras ou substituir os das variáveis iniciais por exemplo na descoberta de clusters.

4.2. Análise discriminante

Passamos a apresentar, nas linhas que se seguem os objectivos da análise discriminante e, em consonância com o ponto anterior, todo um conjunto de passos para a sua concretização.

4.2.1. Objectivos da análise

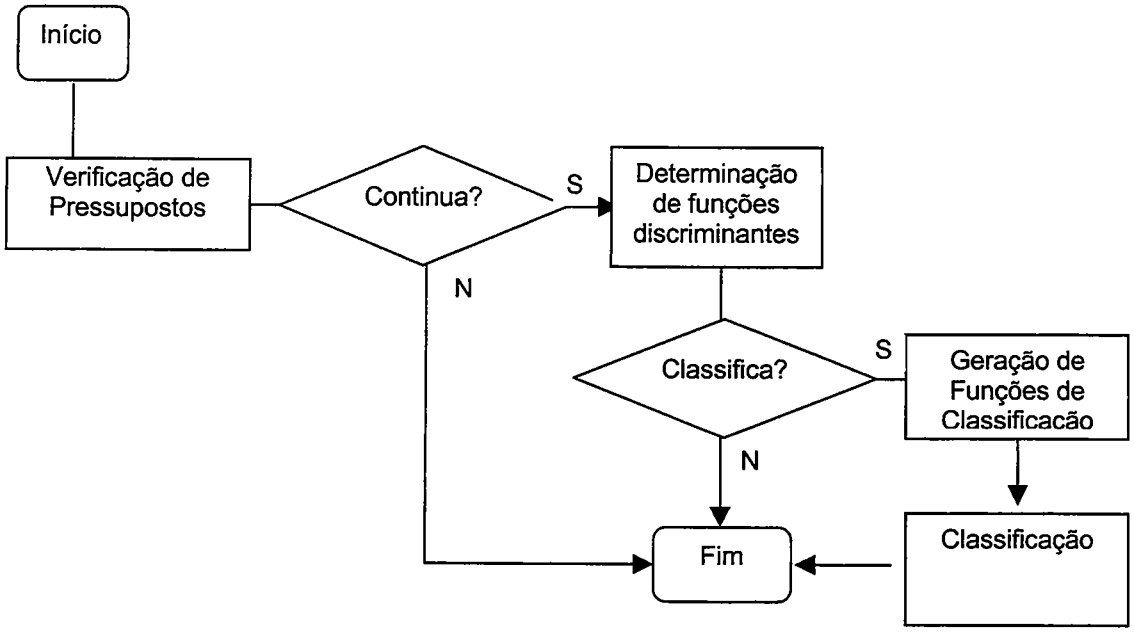
A análise discriminante é um método estatístico multivariado que se emprega *para descobrir as características que distinguem os membros de um grupo dos de outro, de modo que, conhecidas as características de um novo indivíduo, se possa prever a que grupo pertence* (Reis, 1997).

Esta análise foi originalmente desenvolvida na botânica tendo por objectivo fazer a distinção de grupos de plantas com base no tamanho e no tipo de folhas o que tornaria possível, posteriormente classificar as novas espécies encontradas. Uma obra fundamental, citada por Reis (1997), é Fisher (1936) responsável pelo desenvolvimento da análise para dois grupos.

4.2.2. Passos da análise

Na realização de uma análise discriminante, podemos considerar o seguinte conjunto de passos.

Figura 22. Concretização de uma análise discriminante



4.2.2.1. Verificação de pressupostos

Nas referências utilizadas, encontrámos um conjunto de pressupostos e/ou actividades para viabilizar a análise discriminante ou para tomar uma decisão relativamente à sua consecução.

1. Remoção de *outliers*. Uma análise discriminante é altamente sensível à presença de *outliers* ou valores extremos de variáveis que têm um largo impacto nas médias e também aumentam as variâncias podendo erroneamente resultar em significância estatística. Assim, os *outliers* devem ser identificados e removidos antes da análise.
2. Testes de normalidade às variáveis. Assume-se que os dados representam uma distribuição normal multivariada. O exame da sua distribuição em termos gráficos através de histogramas permite visualizar se são ou não normalmente distribuídas. Todavia as violações de normalidade, neste particular, não são usualmente fatais o que significa que os testes de significância resultantes são fidedignos.

3. Homogeneidade das matrizes de variância-covariância. Assume-se que as matrizes de variância/covariância são homogénas entre os grupos. Também neste caso, desvios menores são pouco importantes pelo que é possível uma decisão para prosseguir a análise apesar da violação do pressuposto.

4. Existência de diferenças significativas entre os grupos. A estatística *Λ de Wilks* é uma medida inversa do grau de diferenciação entre os grupos: quanto menor o seu valor, maior esse grau de diferenciação

$$\Lambda = \frac{|W|}{|T|} = \frac{|W|}{|B+W|} \qquad (44)$$

O determinante de W é uma medida da variabilidade dentro dos grupos enquanto o determinante de T nos dá uma medida da variabilidade total. Quanto maior for a semelhança entre os dois determinantes, menores serão as diferenças entre os grupos e mais o valor do *Lambda de Wilks* se aproximará de 1.

4.2.2.2. Determinação de funções discriminantes

O número máximo de funções discriminantes é igual ao número de grupos menos um, ou ao número de variáveis discriminantes, sendo o critério de escolha baseado no menor destes dois valores.

No caso de dois grupos, a função discriminante é uma função linear do tipo:

$$Y = a + b_1 * x_1 + b_2 * x_2 + \ldots + b_m * x_m \qquad (45)$$

onde a é uma constante e $b_1..b_m$ são um conjunto de coeficientes cujo valor representa o seu grau de contribuição para a predição do grupo a que pertence.

No caso de mais de dois grupos, pode ser estimada mais do que uma função discriminante como a anterior e o significado dos coeficientes é idêntico. No entanto, estes coeficientes não nos dizem entre que grupos as respectivas funções discriminam podendo tal interpretação ser obtida a partir de uma representação gráfica ou do exame das médias das funções entre grupos.

Um problema que pode haver necessidade de resolver é a determinação das variáveis a considerar numa análise. Em muitos estudos, um investigador pode querer considerar todas as variáveis no modelo. No entanto, em muitas aplicações, um objectivo chave é identificar um subconjunto útil de variáveis para concretizar a predição. Alguns produtos, como o SPSS, fornecem métodos para construir um modelo de um modo passo a passo (*stepwise discriminant analysis*): fazendo entrar ou removendo em cada passo uma variável do modelo. O método começa por determinar a variável para a qual a média é mais diferente e continua, passo a passo, a juntar a próxima melhor variável.

Reis (1997) apresenta várias estatísticas de apoio à determinação de variáveis a integrar ou remover do modelo, designadamente, o *lambda de Wilks*, a *distância de Mahalanobis*, o *rácio do F mais pequeno*, o *V de Rao* e a *soma da variância não explicada.*

4.2.2.3. Determinação de funções de classificação

As funções de classificação podem ser utilizadas para determinar a que grupo cada caso pertence. Há tantas funções de classificação quantos os grupos. Cada função permite-nos calcular *scores* de classificação para cada caso em cada grupo aplicando a expressão:

$$S_i = c_i + w_{i1} * x_1 + w_{i2} * x_2 + \ldots + w_{im} * x_m \qquad (46)$$

onde i representa o grupo respectivo, $1,2,\ldots\ m$ as variáveis, c_i um valor constante para o gupo i, w_{ij} o peso da variável j no cálculo do *scores* do grupo i, x_j o valor observado do caso respectivo para a variável j.

4.2.2.4. Classificação

Dispondo das funções de classificação associadas a cada grupo, todos os casos são relacionados com cada uma das funções de classificação.

Considera-se que um caso determinado pertence ao grupo para o qual se encontrou o *score* mais elevado.

É possível ter uma avaliação do rigor da predição através da elaboração de uma matriz de classificação onde se mostra para cada grupo o número de casos correctamente classificados e o número de casos mal classificados.

Resumo do capítulo II

Neste capítulo, apresentámos a generalidade dos processos analíticos utilizados e procurados na dissertação em termos de processamento da informação sobre a Economia a partir de bases de dados analíticas.

Uma *base de dados analítica* (BDA) é uma base de dados que apoia o processamento analítico para o que integra física ou logicamente métricas, dimensões e múltiplos níveis de consolidação.

As *métricas* são as medidas que constituem o foco das pesquisas do analista. As perspectivas, ou *dimensões*, fornecem contexto às métricas em análise. A informação desejada por dimensão deverá ainda poder ser visualizada com diferentes *níveis de detalhe*.

Uma BDA é uma colecção de dados orientada para os assuntos do negócio de uma empresa ou para os factos económicos relevantes num espaço económico determinado, integrada, variável com o tempo e não volátil. Integra quatro componentes principais que, quase sempre, existem: dados correntes detalhados, dados históricos detalhados, dados sumariados e Metadados, elemento arquitectural que descreve a sua estrutura e o seu conteúdo.

A especificação de uma BDA depende da arquitectura da sua implementação sendo suportada por modelos lógicos relacionais, multidimensionais e híbridos. A perspectiva relacional, a mais avançada em termos de estandardização, é responsável por vários tipos de diagrama: estrela, floco de neve e constelação de factos, sendo, em determinado contexto, aconselhável a desnormalização de relações.

O processamento analítico materializa-se em processos interrogacionais, multidimensionais e de descoberta de conhecimento suportados pelas BDAs.

Dado um esquema relacional, um processo interrogacional origina uma relação composta por um conjunto de linhas que obedecem a uma condição de pesquisa simples ou complexa.

Um processo de análise multidimensional permite visualizar os dados segundo várias perspectivas de análise (as dimensões), sujeitando-os a operações dinâmicas de agregação ou desagregação proporcionadas pelos níveis de consolidação implementados e descritos no Metadados.

Os processos de descoberta de conhecimento podem ser de diversa natureza. Neste documento, preocupamo-nos essencialmente com a classificação, a pesquisa de associações e a descoberta de clusters.

Dado um conjunto de instâncias de uma estrutura de dados complexa de atributos *A1, A2,... An*, onde um dos atributos tem um domínio de baixa cardinalidade, a *classificação* associa a cada instância nesse atributo um dos valores do seu domínio. Vulgarmente, a classificação é feita em dois passos: num primeiro, é dado um conjunto de exemplos ao sistema para este induzir as regras de classificação, em seguida as regras são utilizadas para classificar outros registos. O conhecimento gerado no primeiro passo pode ser expresso de várias formas (árvores de decisão e regras de produção), existindo vários algoritmos para a sua detecção, inclusivamente manuais.

A pesquisa de associações é a procura de padrões e condições que descrevem como vários itens *se agrupam* ou *acontecem juntos* em séries de eventos ou transacções. Também aqui apresentámos algoritmos manuais e perspectivas decorrentes de trabalho científico inclusivamente na língua portuguesa.

Finalmente, a análise de *clusters* procede ao agrupamento dos indivíduos em função da informação existente, de tal modo que os indivíduos pertencentes a um mesmo grupo sejam tão semelhantes quanto possível e sempre mais semelhantes aos elementos do mesmo grupo do que a elementos dos restantes grupos. Ao contrário da classificação, não usa exemplos.

Referimos ainda a existência de perspectivas difusas associadas à classificação e apresentámos vários métodos de inferência difusa.

Na generalidade dos casos de descoberta de conhecimento em bases de dados, foi mencionada a necessidade de uma transformação prévia do dados: (1) na classificação, a geração de uma classe de baixa cardinalidade, a transformação 1-->N e a transformação N-->M, (2) na pesquisa de associações, a obtenção da designada tabela_R e (3) na análise de clusters, a obtenção de índices de (dis)semelhança.

A importância dos processos apresentados permite-nos pensar que eles detêm um papel de relevo no tratamento da informação sobre a Economia e, como tal, que seja detectável a emergência de novas formas de disponibilização de informação pelas organizações responsáveis nesta área.

Capítulo III - Organizações responsáveis pela Informação para a Economia em Portugal

A existência de organizações que recolhem dados e produzem e divulgam informação para a Economia é uma realidade em todos os países. O facto novo, possibilitado pela proliferação de redes de comunicação, é a disponibilização ser interactiva. Um novo paradigma, que pode ser designado por *economia em rede*, vai ganhando forma e compete aos múltiplos agentes que intervêm a nível económico e a nível do processo de conhecimento da economia, ganhar consciência do facto, racionalizá-lo e fazer com que a sua afirmação seja pacífica e profícua.

Consciente disto, o governo português decidiu, através da Resolução do Conselho de Ministros n.º 95/99, de 25 de Agosto, que todas as Direcções-Gerais e serviços equiparados, bem como os Institutos Públicos nas suas diversas modalidades, ficam obrigados a disponibilizar ao público, em formato digital na Internet, toda a informação que seja objecto de publicação em papel (publicações, formulários, etc.). A informação assim disponibilizada poderá ser livremente utilizada pelo público que a ela acede, desde que se faça menção da respectiva fonte.

Esta é, sem dúvida, uma importantíssima resolução que poderá permitir a todos os que analisam a economia ter um acesso mais fácil aos dados e à informação e permitindo esquecer todos os problemas do passado.

Neste capítulo, vamos definir o objecto de análise desta dissertação em termos de aplicação de análise de maturidade: os sistemas de informação para a economia. No entanto, há que ter a consciência que esta análise tem uma conotação temporal forte: o que é verdadeiro num determinado momento pode não o ser passado algum tempo, embora seja de acreditar que, numa perspectiva de apuramento e difusão da informação, a situação vá melhorando.

Organizámos o capítulo em seis secções. Na primeira, mostramos o modo como se articula o processo de disponibilização de informação sobre a economia com as necessidades do analista.

Na segunda secção, procuramos caracterizar o Sistema Estatístico Nacional (SEN). Fazemo-lo utilizando as próprias palavras que os organismos em que se materializa justificam a sua existência.

A consciência de que, para além das organizações que fazem parte do SEN, outras divulgam informação sobre a Economia levou a que estendêssemos a nossa pesquisa e examinássemos a disponibilização interactiva de dados e publicações organizada em termos de natureza dos seus agentes o que permitiu uma terceira secção cujo objecto são os Sistemas de Informação para a Economia. Esta secção permite-nos examinar os SIE centrados em factos económicos.

Dispondo do conhecimento sobre a informação fornecida pelos SIE em termos temáticos, podemos questionar-nos sobre o seu apoio ao processo de decisão estratégicas nas organizações ao nível da análise SWOT, das decisões sobre o produto, das dimensões competitivas e da análise do consumidor. Esse é o objecto de uma quarta secção que permitirá proceder a um relacionamento de determinados SIE com esses temas ou concluir sobre a ausência de relacionamento.

Na quinta secção, esboçando um indicador relativo à questão da disponibilização incompleta de informação: na realidade, determinado organismo pode parecer estar a divulgar informação sobre a Economia e, após uma análise mais rigorosa, poderemos detectar até que ponto tais dados se aproximam do total de informação relevante. Isto não significa que o analista carece de toda a informação possível, significa, isso sim, que deve ser ele a julgar as suas necessidades, não devendo manifestar-se qualquer imposição na sua satisfação.

Finalmente, retomamos a questão da avaliação do processo disponibilização interactiva de informação para a Economia. Esta avaliação tem a ver dominantemente com as características da informação. Naturalmente as características do sítio na Internet onde tal acontece são importantes, mas esse não é o nosso problema: não vamos avaliar sítios, vamos avaliar informação. Nesse contexto, resumimos a avaliação de sítios recentemente divulgada e marcamos as diferenças para o objectivo desta dissertação.

1. A disponibilização e utilização de informação para a Economia como um processo

Organismos Produtores e Analistas são as entidades fundamentais do processo de disponibilização e utilização de informação para a Economia. Na medida em que existe uma relação de externalidade entre ambos, o que a sua interacção poderá ser resulta de práticas individualizadas as quais assentam fortemente na cultura e no grau de maturidade de cada um dos elementos. Esta fluidez confere ao processo em causa uma diversidade extremamente elevada que vamos tentar compreender e modelar neste capítulo.

Dois elementos fundamentais existem no início de um processo analítico cujo objecto é a Economia: por um lado, uma organização do processo de investigação cuja responsabilidade é do analista e, por outro, uma organização do processo de disponibilização cuja responsabilidade pertence aos Sistemas de Informação para a Economia intervenientes.

Esta organização do processo de disponibilização materializa-se em dados e informação que, no contexto do que nos interessa, assumem uma forma digital e são susceptíveis de comunicação electrónica. Assume-se que quer o analista desconhece os organizadores do processo de informação de que necessita quer estes desconhecem os analistas a quem esta virá a servir.

No contexto do processo de investigação, o analista consulta os SIE susceptíveis de disponibilizarem informação no âmbito do trabalho em elaboração e avalia a informação disponibilizada relativamente às suas necessidades.

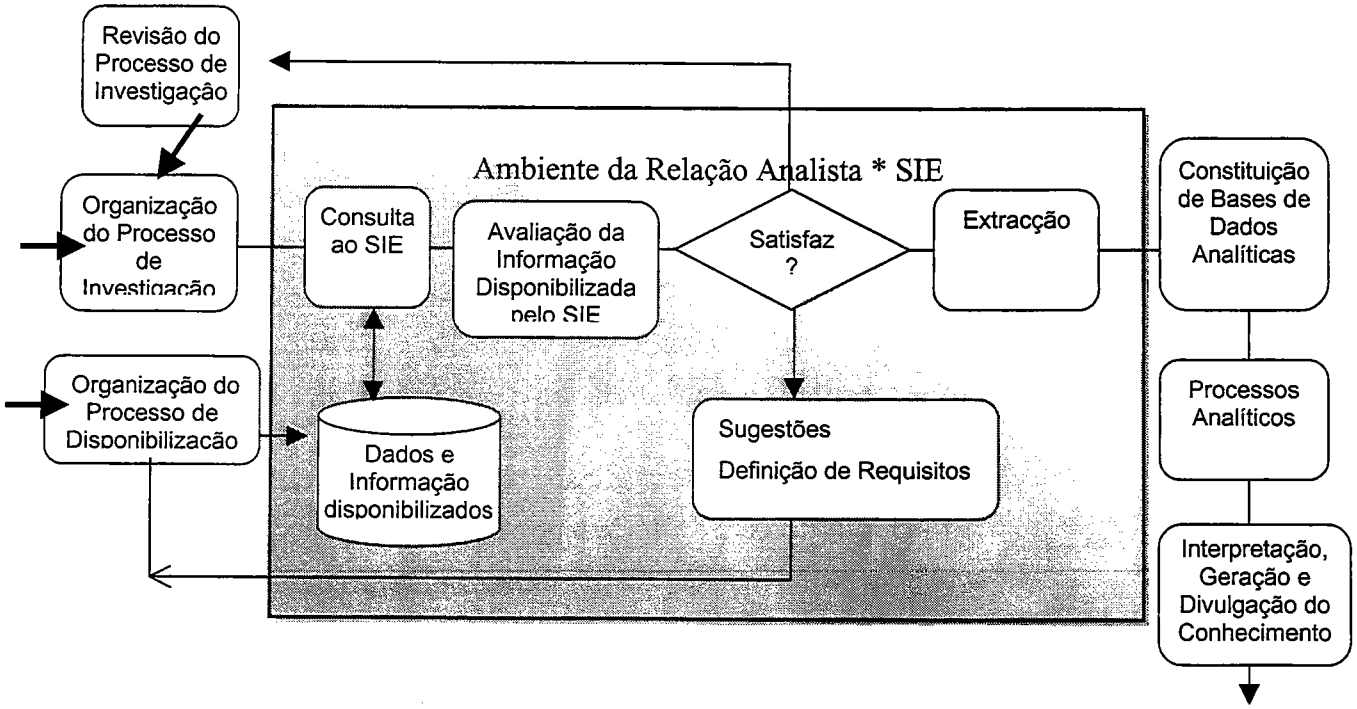
O resultado dessa avaliação conduz a uma de três actividades: 1) pedido, envio de sugestões ou definição de requisitos para o SIE caso o que este disponibilize esteja afastado das necessidades, mas se admita que de algum modo possa haver uma aproximação; 2) revisão do processo de investigação quando o afastamento é de tal ordem que inviabiliza os passos seguintes e 3) extracção dos dados/informação seleccionados e disponibilizados.

Com base nos dados extraídos, o analista constitui bases de dados analíticas, realiza processos analíticos e interpreta os seus resultados em ordem a obter conhecimento sobre a realidade estudada o qual naturalmente será divulgado.

Enquanto as actividades de 1) consulta ao SIE, 2) avaliação da informação, 3) pedido, sugestões, definição de requisitos, e 4) extracção fazem parte do ambiente da relação analista *versus* SIE, todas as outras são externas a este ambiente embora possam por ele ser condicionadas. O interface é proporcionado pela informação disponibilizada pelo SIE naturalmente representada em bases de dados ou em estruturas de dados de complexidade variável.

A figura seguinte explicita este conjunto de actividades tornando claro o ponto de contacto entre SIE e Analista.

Figura 23.O processo de disponibilização e utilização da informação sobre a Economia



Na continuação deste capítulo vamos examinar com maior detalhe algumas destas actividades.

1.1. Organização do processo de disponibilização

A organização do processo de disponibilização é uma actividade interna de um SIE que pode ou não levar em conta as necessidades dos analistas. De acordo com uma maior ou menor orientação para este, é possível responder a um conjunto de questões importantes, designadamente:

- o SIE produz dados ou informação?

- existe disponibilização de informação incompleta?
- como publicita o SIE os seus produtos?
- que formatos usa o SIE na disponibilização de dados e publicações?
- o SIE utiliza algum modelo de anonimização dos dados?

Com o desenvolvimento dos meios computacionais na posse de analistas e particulares, quer em termos de equipamento quer em termos de aplicações, uma maior incidência na disponibilização de dados terá como resultado um alívio dos SIE e uma maior liberdade de tratamento. Ao falarmos em disponibilizar dados, não estamos a considerar que o apuramento de que os SIE eram responsáveis deixa de pertencer á área destes, antes procurando que este apuramento seja divulgado em ficheiros de texto multidimensionais ou tabulares em vez de sujeitos a apresentação resumo em formato *crosstab* ou gráfico. Nesta perspectiva o SIE deixa de se preocupar com a apresentação da informação ficando liberto para a produzir à medida, de acordo com o pedido de analistas ou de outros departamentos.

As vantagens desta prática manifestam-se no facto de deixar de existir disponibilização de informação incompleta num contexto como o preconizado. Admita-se a informação sobre o PIB e um conjunto de cinco atributos ao nível das dimensões de análise: A1, A2, A3, A4, A5. Na nossa perspectiva dados sobre o PIB poderiam ser disponibilizados num ficheiro obedecendo a uma estrutura como {A1, A2, A3, A4, A5, Valor do PIB}. A tentação actual dos SIE não é para apresentar dados de uma forma tão simples mas após a sua passagem por algumas tabulações cruzadas, por exemplo total do PIB no domínio de A1, ou da intersecção de A1 e A2, ou de outra intersecção que algum responsável se lembre.

Entre inúmeros relatórios destinados a apresentar informação, provenientes das mais diversas entidades, não descortinamos algum que apresente completamente os cruzamentos possíveis, optando sempre pelos que, na sua perspectiva, são os mais adequados. Objectivamente tal facto traduz-se em ser impossível ao analista captar o fenómeno em estudo na sua totalidade. Pode argumentar-se que os SIE dispõem de amostras, não da globalidade dos dados, mas a disponibilidade de novos meios computacionais e o sucesso dos algoritmos de mineração implica uma modificação dos processos de recolha.

Existindo uma necessidade de que os dados cheguem às mãos de quem deles carece, é de admitir um processo de publicitação de forma a tornar mais fácil um plano de consultas. No próprio SIE é de admitir a existência de pessoas destinadas a guiar os analistas no acesso aos seus dados, não se tratando de os substituir mas de os fazer chegar mais rápido ao destino.

Que aconteceria se a um grupo representativo de utilizadores de informação estatística (investigadores, estudantes, planificadores, decisores) fosse dada a oportunidade para descrever a sua *máquina de sonho* – o ambiente tecnológico que lhe poderia servir como um suplemento ao seu próprio cérebro, potenciando a sua produtividade bem como a sua qualidade de trabalho?

Ryssevik e Musgrave (1998) acredita que uma lista combinada de requisitos incluiria pelo menos os elementos seguintes:

- todos os dados empíricos disponíveis em linha;
- qualquer sistema integrado que permitisse o acesso e pesquisa a estes dados;
- quantidade significativa de metainformação[14] disponível associada e totalmente integrada com estes dados;
- capacidade para listar e analisar os dados em linha ou para os ter convertidos para a ferramenta de análise favorita;
- garantir a metainformação ao proceder ao *download* dos dados;
- agentes de pesquisa activos procurando na rede e informando da existência de novos dados no seu campo de interesse;
- *hyperlinks* eficientes a partir das fontes de dados para todas as publicações científicas publicadas com base nesses dados;
- um sistema de *feedback* eficiente relativamente à metainformação permitindo ao utilizado juntar a sua própria experiência à memória colectiva da fonte de dados.

[14] Como referimos no capítulo II, trata-se de *Informação acerca da informação*. Na continuação deste documento, dependendo do contexto, muitas vezes utilizaremos com o mesmo significado a expressão Metadados

Embora os meios tecnológicos de construção para desenvolver tais sistemas estejam mais ou menos disponíveis, existe alguma discrepância entre o sonho e a realidade:

- apenas uma fracção de fontes de dados relevantes está disponível em linha;
- não existe nenhum sistema de acesso e pesquisa comum aos dados disponíveis;
- existe alguma ausência de estandardização especialmente do lado do metadados;
- existe pobre integração entre dados e metainformação;
- existem barreiras institucionais (a concorrência);
- existem barreiras jurídicas (salvaguarda da privacidade);
- existem barreiras comerciais (informação com um bem de consumo).

Tomados no seu conjunto estes obstáculos constituem uma barreira muito significativa entre os utilizadores de informação estatística e as fontes de dados. Como o valor latente dos dados apenas pode ser revelado pelo seu uso activo, a situação está longe de ser satisfatória.

Apesar de tudo, desde 1995, num processo que Portugal acompanhou em termos quase exemplares, a Internet surgiu como um veículo de disseminação importante utilizado pelos organismos de estatística, sendo vista actualmente como o principal canal de disseminação no futuro.

De acordo com Podehel (1998), as vantagens da Internet neste contexto são várias:

- uma localização (um sítio associado a uma organização que disponibiliza dados e publicações estatísticas sobre a Economia) permite o acesso à variedade de informação independente do tempo e da distância;
- a libertação periódica da informação mais recente com acesso instantâneo pelo cliente;
- a oportunidade de publicar bastante mais informação do que seria possível em papel;

- a oportunidade para publicar mais informação devidamente contextualizada através do fornecimento de *hyperlinks* para detalhes, notas explicativas, informação previamente publicada, etc.
- a libertação de custos na distribuição física comparada com as publicações em papel onde cada cópia adicional incorre em custos de impressão, processamento da encomenda, envio, contabilização, etc.: na Internet, os custos por ter um cliente adicional a fazer acesso a uma peça existente de informação estão perto de zero para o cliente e para o departamento de estatística[15].

Ainda de acordo com Podehel (1998), há que procurar métodos para que tais páginas sejam criadas e/ou actualizadas de alguma forma dinâmica e automatizada a partir de um conjunto de informação organizada. Esta questão é referida como *data base publishing*. A ideia central deste conceito é a de separar a informação que suporta um site da representação do seu conteúdo nas páginas HTML o que tem duas vantagens:

- à medida que é adicionada nova informação à base de dados, são geradas automaticamente páginas HTML novas ou actualizadas sem qualquer intervenção ou codificação manual;
- separando as duas funções é possível fazer melhorias em qualquer das duas funções sem impacto necessário na outra.

As novas necessidade em termos de métodos de publicação, os requisitos associados aos dados em termos do seu formato e do Metadados justificam que sob o patrocínio das Nações Unidas os organismos de estatística dos diferentes países procurem novas soluções e tornar conciliável o produto da sua actividade. Em 2001, está disponível extenso material metodológico o qual é actualizado e acrescentado numa base anual a partir de uma Conferência de Estatísticos Europeus e da notícia de novos desenvolvimentos em termos de tecnologias de informação e comunicação.

[15] Embora seja verdade que, em contraste com as publicações em papel, o custo marginal por informar um cliente adicional através da Internet é muito baixo senão perto do zero, há custos significativos na operação de um sítio Internet e no desenvolvimento e actualização do seu conteúdo. Em particular, à medida que o conteúdo cresce, o custo de manutenção e actualização de páginas individuais no formato HTML torna-se significativo.

1.1.1. Recomendações para o desenho de sítios na Web relativos a dados estatísticos

UNSC/ECE (2001,1) apresenta um conjunto de recomendações para o desenho de sítios na Internet que respeita aos utilizadores, ao desenvolvimento e manutenção, à pesquisa e navegação, à interpretação e ao processamento posterior dos dados.

Utilizadores. Devem constituir o centro de atenções. É importante investigar quem são, o que realmente precisam, como usam os dados, quais são os seus níveis de competência. Deverá ser organizada alguma espécie de testes de utilização mesmo que seja muito simples. Os estudos de utilização providenciam um retorno que permite melhorar o sítio e, a longo prazo, reduzir os custos com o redesenho. Apesar desta preocupação com os utilizadores e com as suas necessidades, podemos adiantar que este é um dos elementos que na análise futura mais nos surgirá como apresentando maiores carências ao que julgamos por o SIE não investir decisivamente numa transformação da relação, pretendendo sempre ter a última palavra na decisão sobre as características de um produto, isto é, tentando manter o poder.

Desenvolvimento e manutenção. Não deve ser apenas tarefa de especialistas da área das tecnologias de informação e de pessoal especial da disseminação. Um suporte muito forte a nível da gestão de topo é um factor crítico de sucesso. Deve ser assegurada que a competência necessária para desenvolver e manter está assegurada.

Pesquisa e navegação. UNSC/ECE (2001,1) defende que a arquitectura de um sítio que disponibiliza informação para a economia deve providenciar uma navegação compreensível ao longo de todo o sítio. Deve ser fácil encontrar os dados desejados. O tempo de resposta deve ter alta prioridade.

Interpretação. Um suporte em metadados flexível e consistente deve assegurar que os dados estatísticos publicados são transparentes e compreensíveis para os utilizadores, devendo ser evitada qualquer espécie de má interpretação ou compreensão. O metadados deve suportar a comparabilidade dos dados no tempo. Os dados estatísticos publicados devem ser consistentes ao longo da totalidade do sítio.

Processamento posterior. UNSC/ECE (2001,1) evidencia ainda preocupações com o processamento posterior dos dados que o utilizador deve ser capaz de descarregar para o seu ambiente técnico. Estes dados devem ser fornecidos num formato estandardizado bem aceite.

1.1.2. Metadados e informação estatística

As preocupações dos estatísticos com a disponibilização interactiva de informação levaram-nos a definir *standards* em termos de Metadados os quais são apresentados em UNSC/ECE (2000). De acordo com este documento podemos considerar vários tipos de Metadados, designadamente:

- Metadados de apoio à pesquisa e navegação;
- Metadados de apoio à interpretação;
- Metadados de apoio ao processamento posterior.

A primeira categoria pode ser desdobrada em termos de informação geral sobre o sítio e apoio à pesquisa.

Quadro 22. Metadados de apoio à pesquisa e navegação

Informação geral sobre o sítio	Apoio à pesquisa
- listagem de conteúdo do sítio; - questões frequentes, úteis como primeiro guia para usar o sítio; - notícias anunciando novas possibilidades (serviços ou documentos novos, ou uma actualização importante); - descrições das áreas estatísticas tratadas; - descrição da instituição estatística (enquadramento legal, estrutura organizacional etc.); - descrição do sistema estatístico (papel dos diferentes intervenientes, etc.); - publicações de referência, visão genérica dos produtos, publicações gerais; - contactos pessoais ou endereços electrónicos para mais informação; - calendário de entregas; - ligações a outros sítios estatísticos na Web; - facilidades de retorno (via e-mail) permitindo encomendar produtos e serviços disponíveis na Web e que podem também tornar possível: 1. convidar os utilizadores a exprimir os seus pontos de vista acerca de um sítio particular; 2. obter informação dos utilizadores; 3. criar livros de visitantes e sistemas de conferência onde os utilizadores podem enviar as suas próprias contribuições.	- Informação dando uma panorâmica geral sobre um tópico, com ligações (*links electrónicos*) a informação mais detalhada fornecerá um bom ponto de partida para muitos utilizadores sem qualificações. - Uma lista de palavras chave ligando uma linguagem trivial de todos os dias a tabelas ou gráficos estatísticos disponíveis quer directamente na Internet e/ou noutra forma impressa/publicada é uma técnica útil. - Muitos utilizadores esperam uma pesquisa baseada em texto livre. Deve-se, no entanto, contar com alguns problemas de implementação derivados da necessidade de estabelecimento de ligações entre a linguagem de todos os dias e termos técnicos utilizados muitas vezes em tabelas. A sua integração com um Thesaurus é útil. - Muitos utilizadores preferem um sistema hierárquico de classificação e pesquisa para concretizarem o *drill-down* numa área e tabela. No entanto, neste aspecto há que ter em conta que *muitos níveis hierárquicos não são eficientes* (UNSC/ECE, 2000).

Consideremos agora a segunda categoria de Metadados, a que visa apoiar o utilizador em termos de interpretação. Os requisitos relativos à informação de apoio à interpretação dependerão de alguma forma da área considerada, do tipo de informação disponibilizada e dos grupos de utilizadores alvo para algum serviço específico. As necessidade variam substancialmente entre utilizadores gerais, *media*, investigadores, etc.

Permitir ao utilizador em geral fazer uma interpretação rigorosa e evitar uma utilização incorrecta dos dados disponibilizados é a palavra de ordem nesta área.

Quadro 23. Metadados de apoio à interpretação

Conjunto mínimo para interpretação correcta das estatísticas	**Percepção da qualidade e correcção**
– Título / descrição de conteúdo · população estatística; · cobertura geográfica; · unidade estatística observada; · classificações and standards aplicados; - labels para linha/colunas de tabelas e elementos de gráficos; - definições de labels; - unidades de medida; - período de referência; - unidades regionais; - comparabilidade no tempo (quebras em séries, dados omissos); - notas com precauções especiais; - fonte dos dados (agência que compilou os dados); - explanação de símbolos standard em tabelas; - informação sobre *copyright*, restrições de uso; - pontos de contacto para informação adicional;	- comparabilidade com fontes alternativas; - ligações para publicações na imprensa / sumário de descobertas; - descrição de métodos usados na colecção, revisão, cálculo e estimação de estatísticas; - informação sobre fontes de erro e correcção das estatísticas; - descrição do *background* e objectivos das estatísticas; conceitos, variáveis e standards usados.

Finalmente, o Metadados de apoio ao processamento posterior deve ser incluído quando os dados estatísticos são descarregados para ulterior aplicação, apresentando dois requisitos importantes:

1. O Metadados referido na categoria anterior como *conjunto mínimo de metadados para correcta interpretação das estatísticas* deve ou estar junto

aos dados descarregados ou ser de fácil acesso para descarga numa operação separada.

2. Dados e metadados devem permitir processamento posterior utilizando instrumentos apropriados como folhas de cálculo, bases e dados, *packages* de análise estatística ou pata apresentação tabular/gráfica. Tabelas grandes devem ser facilmente descarregadas. Os SIE, ao porem dados na Internet, devem estar a par das possibilidades e limitações dos diferentes formatos relativamente ao *download* de dados estatísticos. Por exemplo

- Tabelas ou texto em formato gráfico (gif, jpg) não são práticos para posterior análise e processamento.
- Devem estar sempre disponíveis formatos gerais e abertos.
- Formatos proprietários (folhas de cálculo, bases de dados, *packages* estatísticos) podem ser úteis para simplificar processamento posterior e incluir Metadados básico. Quando for usado um formato proprietário, deve ser também incluída referência ao software ou uma possibilidade para a sua descarga.

1.1.3. Formatos para disponibilização de dados e informação

As considerações anteriores permitem-nos detectar que os formatos de disponibilização dos dados e do Metadados são extremamente importantes em termos do utilizador final. Tal facto foi percepcionado pela Conferência de Estatísticos Europeus que produziu um documento metodológico nesta área, o UNSC/ECE (2001,3).

De acordo com este documento podemos considerar seis categorias de formatos: o ficheiro simples de dados, o ficheiro gráfico, o ficheiro de linguagem natural, o formato de gestão de dados, formatos mistos e um formato que constitui uma extensão ao HTML denominada XML. O documento nota que nenhum destes formatos terá sido desenvolvido especificamente para dados estatísticos e noticia uma prometedora iniciativa consubstanciada na criação de uma *task force* liderada pelo Fundo Monetário Internacional para desenvolver um *standard* comum para *Statistical Data and Metadata Exchange (SDMX)* baseada no formato XML.

Enquanto este formato destinado a dados estatísticos não surge, vejamos mais detalhadamente as seis categorias enunciadas anteriormente.

Quadro 24. Descrição dos vários tipos de formatos

A. *O ficheiro de dados simples* Trata-se de um ficheiro que contém apenas valores. O Metadados estará algures embebido em código de programa ou descrito separadamente num formato legível. Os valores podem aparecer em posições pré-definidas (*formato fixo*), ou podem ser separados por um código predefinido, usualmente uma vírgula (*separado por vírgulas*). Esta é a forma mais velha de informação digital. Era própria de uma época em que as máquinas tinham uma capacidade de memória muito curta, quando esta era perfurada em cartões e o espaço de trabalho era medido em *kilobytes*. Este formato é muito perigoso pois não dá qualquer garantia de que os dados são utilizados com o metadados apropriado. O risco de uso incorrecto tornou-se particularmente importante com o tamanho e facilidade de acesso da Internet.
B. *O ficheiro gráfico* Esta é uma família de formatos capaz de armazenar figuras (GIF, JPEG, BMP, etc.). É perfeitamente apropriada para informação puramente gráfica, tais como fotografias ou desenhos, significativa para os olhos humanos. Não é apropriada para apresentação de dados, porque não consegue representar os dados elementares internamente. Assim, os formatos gráficos também não são indicados para gerir Metadados.
C. *O ficheiro em linguagem natural* Trata-se de uma categoria de documentos legíveis com Metadados ligeiro. O seu propósito fundamental é apresentar informação textual no papel ou écran. Na sua forma mais simples, o ficheiro de texto ASCII, a máquina vê o conteúdo como um conjunto estruturado de caracteres só em termos de linhas ou parágrafos. Formatos mais sofisticados permitem ao software definir o layout e controlar a estrutura em termos de páginas, parágrafos e capítulos. Este é o caso de formatos associados a processadores de texto como o MS-Word e o WordPerfect. É também o caso de formatos que só admitem leitura como o PDF e Envoy. O HTML também se situa nesta categoria. Embora todos estes formatos tratem informação em linguagem natural, eles não são igualmente indicados para o tratamento de documentos a disponibilizar para descarregar na máquina do utilizador.

D. *Os formatos de gestão de dados*

Esta categoria compreende muitos formatos concebidos para acesso e processamento dos dados por software.

Estes formatos têm metadados integrado[16], e requerem funcionalidades de apresentação para produzir relatórios ou vistas. Um sistema de gestão de bases de dados oferece o exemplo mais típico deste tipo.

O *software* de processamento de dados em geral utiliza formatos *standard* de bases de dados ou define o seu próprio formato proprietário. No contexto da estatística, o SPSS oferece um exemplo de tal formato. Protocolos de transferência como o GESMES também pertencem a esta categoria.

O acesso a tais formatos depende da presença do *software* a que pertencem. Não é razoável contar com a sua presença na máquina do utilizador final, pelo que estes formatos só devem ser utilizados na base de acordos entre o SIE e os utilizadores que o pretenderem.

Algumas organizações que disponibilizam informação para a economia usam formatos proprietários suportados por *software* específico. Tal uso em documentos descarregáveis é aceitável desde que o *software* seja fornecido para *download* no mesmo sítio.

E. *Formatos mistos:*

Esta categoria suporta documentos que podem ser lidos universalmente com metadados ligeiro[17] para torná-los humanamente legíveis e com metadados formal para processamento automático. O expoente principal desta categoria é a folha de cálculo. Estes formatos que suportam apresentação dos resultados e processamento posterior parecem ser largamente aceites pelos utilizadores com o objectivo de apresentação e como uma espécie de formato intermédio para processamento posterior. UNSC/ECE (2001,3) apresenta duas recomendações quanto aos formatos mistos:

- Se a edição do documento pelo utilizador final não fizer parte dos propósitos do documento, os formatos de folha de cálculo e os formatos processador de texto devem ser evitados;
- Ficheiros ASCII quer na sua forma tabular quer de linguagem natural apresentam falta de funcionalidades de apresentação pelo que devem ser sempre evitados.

[16] Neste caso, dados e metadados estão presentes no ficheiro de uma maneira que a sua relação não tem qualquer ambiguidade. Um bom exemplo são as bases de dados.

[17] Neste caso, dados e metadados estão presentes no mesmo ficheiro ou em ficheiros diferentes os quais são associados através de hyperlinks HTML. Todavia a relação entre dados e metadados não é formalmente definida.

F. *O formato XML*

É o formato mais recente e tem características que o tornam muito atractivo para a comunicação de dados. O XML é mais do que apenas um formato, pois fornece ferramentas para definir formatos novos para documentos de uma categoria específica. A estrutura de um tipo de documento pode ser especificada num *Esquema XML*, numa *Definição de tipo de documento* (DTD) ou num *Resource Description Framework* (RDF). O autor, ou o utilizador final, pode especificar o layout num documento separado o que está de acordo com as regras do *XML Style sheet Language* (XSL). Os *browsers* utilizam esta definição, se disponível, com o objectivo de apresentar o conteúdo de um modo que é compatível com as intenções do autor. Todavia, os *browsers* são capazes de proceder a uma apresentação com significado de um documento XML mesmo que não tenham acesso a uma definição de estrutura ou a uma especificação XSL.

A principal vantagem do XML é que o conteúdo de um documento é acessível não só a *browsers* mas a qualquer *software* com capacidade XML que é capaz de usar a definição própria. Isto põe o XML numa posição que lhe permite ultrapassar as funções da maioria dos outros formatos. Ele pode suportar metadados integrado tal como os formatos de base de dados. Pode misturar metadados integrado com metadados ligeiro tal como as folhas de cálculo. Pode gerir linguagem natural com metadados ligeiro tal como os processadores de texto e HTML. Pode também suportar metadados integrado referindo-se a metadados separado como fazem alguns formatos de transferência.

Todavia, o XML suporta a ligação de hipertexto o que possibilita a construção de documentos complexos a partir de múltiplos ficheiros como o HTML. Assim a questão referida para o HTML também se aplica ao XML. Ao publicar dados para *download* devem ser evitados *links* para estruturar o documento. Deve-se ter como objectivo produzir um só ficheiro contendo todos os dados. Isto dito, UNSC/ECE (2001,3) conclui que XML é o formato a escolher sempre que os dados sejam altamente estruturados e o uso de software específico não seja crítico.

A disponibilização de dados e informação em documentos suportados por formatos baseados em linguagem natural é ainda extremamente importante apesar dos novos desenvolvimentos sugeridos pelo formato XML e pelos estudos conducentes a um formato vocacionado para a informação estatística que já referimos. Nestas condições resumimos no quadro seguinte algumas das funcionalidades associadas a cada uma das subcategorias referidas.

Quadro 25. Tratamento de documentos por formatos baseados em ficheiros de linguagem natural

Formatos ASCII (American Standard Code for Information Interchange) É o mais universal, porque é bastante simples. Não requer software específico. Pode ser enviado como linha de dados para o écran ou para a impressora. Todavia este formato é severamente limitado em termos de funcionalidade. Não tem qualquer *layout* de controlo, dispõe de meios muito elementares para a estruturação de texto, tais como a separação de parágrafos com uma linha em branco, apresentando os títulos em maiúsculas, ou iniciando novo capítulo com um separador de página. O único meio para arrumar texto em forma tabular é através do uso de espaços ou caracteres de tabulação. Além disso, não pode envolver objectos gráficos. O formato ASCII usa um número de caracteres muito limitado, 94, podendo com extensões utilizar mais 127 as quais requerem metadados separados para funções de tratamento de página. Tudo isto torna o ASCII desaconselhável para algo que não seja texto não estruturado muito simples, atribuindo este formato aos dados uma imagem ultrapassada e pouco profissional pelo que deve ser evitado na disponibilização profissional de dados estatísticos.
Formatos HTML (HyperText Markup Language da Internet) Pode ser interpretado pelos diferentes tipos de software que gerem a informação disponibilizada na Internet (*browsers*) e espera-se que qualquer pessoa que faça acesso à Internet o possa usar. Além disso, o HTML pode ser considerado como um formato razoavelmente universal. O HTML, todavia, não pode integrar objectos, pode apenas referi-los. As parte diferentes de um documento em HTML são usualmente armazenadas em ficheiros separados, sendo tarefa do *browser* seguir os *links* e apresentar os diferentes elementos juntos como um todo. Isto é o que os *browser* fazem quando actuam em linha. Se um documento HTML precisar de ser descarregado para acesso não em linha, esta potencialidade torna-se problemática. É difícil proceder ao *download* de todas as partes de um documento de tal modo que garanta que a leitura não em linha produza o mesmo resultado que o seu processamento em linha pelo *browser*. Assim, recomenda-se o uso de HTML em documentos exclusivamente tratados em linha. Se, todavia, for utilizado em documentos para *download*, o autor deve ter cuidado na sua estruturação de modo que seja guardado num só ficheiro.

Formatos processadores de texto

Não costumam ser acessíveis sem o *software* a que pertencem.

Os ficheiros MS-Word precisavam do produto MS-Word para acesso e os ficheiros Word Perfect precisavam do produto WordPerfect para o mesmo efeito.

Nas versões actuais, os processadores de texto são capazes de ler ficheiros produzidos por uma quantidade significativa de outros processadores de texto. Todavia, a apresentação do documento que daí resulta não é sempre satisfatória. Pode dizer-se o mesmo do Rich Text Format (RTF), concebido para troca de documentos independentes da plataforma. Embora os processadores de texto de vanguarda suportem este formato, não se está sempre seguro de ver a mesma imagem dum ficheiro RTF num programa diferente daquele que o produziu.

A principal desvantagem dos processadores de texto é a facilidade da sua edição pelos utilizadores que pode levá-los à sua transformação inadvertida por alguém que não esteja sempre consciente das modificações que estão a ter lugar e que pode levar a que tabelas sejam mal formatadas e que as figuram apareçam em páginas diferentes.

Utilizadores subsequentes, desconhecedores disto, consideram que este foi um resultado do trabalho do SIE.

Assim, os formatos de processadores de texto devem ser evitados a não ser que a edição do texto pelo utilizador final faça parte dos objectivos do documento.

Formatos *Read-only*

São flexíveis e acessíveis. Podem integrar toda a funcionalidade de apresentação dos processadores de texto e os programas de acesso aos mesmo são descarregáveis sem encargos.

O formato *read-only* mais popular é o Portable Document Format (PDF), cujo programa de acesso, chamado Acrobat Reader, é fornecido pela Adobe. Para tornar os documentos PDF acessíveis universalmente, basta tornar o Acrobat Reader acessível para *download*.

Este formato, e particularmente o PDF, é a melhor escolha se a integridade do documento for um requisito importante – o que é geralmente o caso dos dados publicados.

Um formato *read-only* protege um documento contra modificações inadvertidas ou *naive*.

O comportamento malicioso apenas pode ser, no entanto, excluído através de certificação baseada em algoritmos de encriptação, o que não é suportado por PDF nem por outro formato largamente utilizado na Internet.

1.1.4. Modelos de anonimização

Contrariamente aos casos anteriores, provém da experiência da construção europeia todo um conjunto de processos relativos à anonimização dos dados testado em casos de disponibilização de microdados para investigação científica. O problema tem a sua origem na necessidade de anonimização, ou seja, na necessidade de "anulação de todos os meios que possam razoavelmente ser utilizados por uma terceira parte para identificar a dita

unidade estatística", conforme consta do regulamento do Conselho Europeu 322/97.

Reconhecendo que, *quando estão em causa universos de grande dimensão e dados recolhidos por amostragem, a eliminação das variáveis de identificação das unidades de registo constitui com frequência, só por si, um procedimento eficaz de anonimização*, Mauritti et al (2002) resume seis modelos de anonimização: (1) a agregação de categorias, (2) a agregação de casos, (3) a construção de subamostras, (4) a disponibilização de microdados em centros autorizados, (5) o acesso indirecto aos dados e (6) o acesso através da Internet.

Agregação de categorias. Este processo consiste na selecção de variáveis contidas nas bases de dados originais que poderiam conduzir à identificação de unidades estatísticas, sujeitando-as a uma agregação de categorias. Permite soluções diversas consoante o tipo de utilização previsto para os dados. Na selecção das variáveis a agregar, há que considerar as perspectivas de análise e os temas específicos de investigação.

Agregação de casos. Consiste em agregar, não as categorias de variáveis, mas os casos em subconjuntos, atribuindo a cada caso o valor de caracterização do subconjunto. Apresenta algumas desvantagens relativamente ao anterior. Entre elas, o facto de o processo de agregação ser mais complexo e introduzir maior indeterminação na análise. Para além disso, o nível de interacção com os dados é limitado, aumentando a barreira à construção e teste de novas variáveis.

Subamostras das bases de dados originais. Este processo usa o princípio de que, com a construção destas subamostras, fica assegurado o anonimato, com a impossibilidade de identificação dos dados.

Disponibilização de microdados em centros autorizados. Conjuga a *agregação de casos* com a disponibilização das bases de dados em centros autorizados onde os investigadores podem permanecer durante determinado período para o desenvolvimento de análises específicas. A *agregação de casos* assume, neste caso, a forma de um cálculo laboratorial seguro, onde é também rigorosamente impedida a alteração dos registos originais. Este método faculta a oportunidade de

tratamento de dados de todos os casos colectados e não apenas de uma subamostra, como acontecia no caso anterior. A implementação deste processo de anonimização é, no entanto, muito dispendiosa, implicando a existência de uma equipa permanente que dê apoio e supervisione os investigadores. Além disso, custos e inconvenientes são também impostos aos investigadores, que têm de se deslocar para trabalharem com todos os dados, podendo implicar que tenham de aí permanecer algum tempo até concluírem a pesquisa

Acesso indirecto aos dados. Este processo materializa-se no acesso indirecto aos dados através da elaboração de pedidos específicos, dirigidos à instituição produtora/gestora das bases de dados, de acordo com as análises que se pretendem conduzir. Este sistema permite solicitar frequências ou cruzamentos entre indicadores. Porém, não permite, por exemplo, a criação de novas variáveis ou a elaboração de testes estatísticos mais sofisticados, designadamente de estatística multivariada, nem possibilita experimentação indirecta de sucessivos tratamentos alternativos, o que o torna menos atractivo para alguns tipos de investigação

Acesso através da Internet. Trata-se de uma situação intermédia entre a referida anteriormente e os vários modelos de acesso directo aos microdados que consiste na disponibilização de bases de dados estandardizadas na Internet. Possibilita a realização de apuramentos simples de frequências, a selecção de categorias de variáveis e/ou segmentos da população, bem como, pedidos de cruzamentos com duas ou mais variáveis. Para além de facultar a experimentação das bases de dados, tem a vantagem de integrar na informação disponibilizada todo um conjunto de variáveis derivadas de uso comum cuja construção implicaria o domínio de algoritmos específicos.

1.2. Organização do processo de investigação

A organização do processo de investigação é uma actividade da responsabilidade do analista que pode influenciar o processo de disponibilização pelo SIE. O analista deve guiar-se pelo máximo de exigência

em termos de dados para chegar aos resultados pretendidos procurando o mais cedo possível:

- definir as necessidades em informação,
- identificar as fontes potenciais,
- definir palavras-chave que possam apoiar uma consulta interactiva,
- elaborar um plano de acesso às fontes

A definição das necessidades em informação é a actividade que procura estabelecer os itens informacionais necessários a um estudo e os atributos caracterizadores que os acompanham e que podem assumir o estatuto de dimensão de análise. Como exemplo, podemos atender ao que neste documento vai ser considerado como fundamental em termos do Modelo de Análise Multidimensional para o Investimento Directo Estrangeiro. Os dados fundamentais reportam ao investimento bruto (também designado por crédito) e ao desinvestimento (débito) em Portugal. Esses dados são necessários ao nível de Sector de Actividade Económica, Operação e Tempo que constituem três dimensões de análise. Adicionalmente poderá ser necessário conhecer alguma lógica do relacionamento ou hierarquização numa dada dimensão de análise, podendo por exemplo ser considerados grupos de operação.

Com a diversidade de fontes hoje existente, a identificação, prévia a uma consulta, das potenciais é um factor que pode acelerar o processo de investigação. Em Portugal, praticamente INE e Banco de Portugal monopolizam os dados disponibilizados aos analistas. Isto não significa que outros organismos não sejam importantes neste processo e que, por vezes, não possam fornecer dados mais adequados a um determinado problema. Como exemplo, podemos apontar o ICEP e o GEPE. O primeiro foi a entidade que nos forneceu dados que nos possibilitaram a construção de um caso relativo ao estudo de associações no processo de análise do IDE. O segundo (GEPE) é um exemplo bem positivo da utilização de facilidades do processamento analítico aplicadas à área de economia, concretamente ao Comércio Externo. Assim, o conhecimento de fontes alternativas àquelas que em princípio estariam mais bem colocadas para um processo de disponibilização, pode revestir-se de utilidade acrescida para o analista.

Sendo em geral desconhecidas fontes e fontes alternativas para o analista inexperiente e sendo pouco divulgada a produção dos SIE, a utilização de palavras-chave numa pesquisa viabilizada pela Internet pode conduzir a fontes mais ou menos fiáveis o que exige um processo relativamente rigoroso de avaliação dos dados disponibilizados.

Enfim, a elaboração de um pequeno plano de acesso às fontes, considerando num momento inicial as mais prováveis em termos de capacidade de fornecimento dos resultados pretendidos, depois as secundárias, depois as internacionais e por fim as resultantes de palavras-chave, sem excluir a possibilidade de uma miscelânea para uma comparação prévia é sem dúvida importante para a pesquisa poder disponibilizar os melhores resultados.

1.3. Consulta aos SIE

No contexto do processamento analítico, a *consulta aos SIE* é operada pelo acesso através de um endereço relativo ao organismo que disponibiliza informação o qual poderá ter sido conhecido através de publicação do SIE, consulta por palavra-chave, contacto com outros analistas, etc.

Em termos de consulta ao SIE, o analista encontra-se em determinado momento perante um conjunto de opções associadas à oferta daquele e de acordo com a qualidade da organização terá resultados ou não com maior facilidade. Assim os aspectos associados à navegação e à informação ao analista, viabilizados pelo próprio SIE, tornam-se elementos determinantes. A existência de mecanismos de busca operando num espaço de pesquisa mais pequeno podem tornar-se extremamente úteis. Como exemplo, consulte-se a página do Banco de Portugal possibilitadora de uma pesquisa orientada por determinadas palavras que podem desde logo possibilitar uma associação com o objecto de pesquisa do analista.

1.4. Avaliação dos dados e informação disponibilizada pelo SIE

A avaliação dos dados e informação disponibilizada pelo SIE é uma actividade da responsabilidade do analista que, tal como a organização do processo de investigação, pode influenciar o processo de disponibilização pelo SIE. O analista deve guiar-se por um conjunto de critérios de avaliação que lhe garantam a melhor e mais rigorosa informação, designadamente:

- atender à potencialidade de processamento posterior dos dados;
- analisar a multidimensionalidade do que é disponibilizado;
- avaliar a integridade temporal;
- medir a diversidade dos dados do SIE;
- ter em linha de conta a actualidade.

Não existe processo analítico sem um processamento de dados posterior à sua extracção da fonte. Esta afirmação revela a importância que assume o formato dos dados extraídos: quanto mais estes se aproximarem do *ASCII* e da estrutura de folha de cálculo, maior o seu valor imediato para o analista. Em contrapartida, do ponto de vista de uma gestão do SIE que procure tirar partido da escassez de dados e de informação, estes formatos são desastrosos. Assim pode existir uma oposição entre a possibilidade de processamento posterior e os objectivos de rendibilização da actividade do SIE. Esta oposição manifesta-se então em múltiplos obstáculos à divulgação dos dados os quais passam a um formato especial (PDF e *postscript* são exemplos) ou incluem caracteres de edição que tornam menos imediato o seu processamento. Esta dificuldade foi encontrada inclusivamente em informação paga ao INE no seu produto INFOLINE. A inserção de caracteres como vírgulas, pontos e brancos inviabiliza o tratamento posterior em folha de cálculo sem um processo de limpeza desses caracteres.

Dados e informação podem ser disponibilizados com maior ou menor nível de detalhe. Por exemplo, o PIB pode ter o seu valor divulgado em termos do país como um todo ou integrar informação acerca da região e do sector de actividade económica. Se, na última perspectiva, se admitir que tal sucederá anualmente, existirão várias dimensões de análise.

Quando a disponibilização temporal assume uma forma coerente e periódica fala-se em integridade temporal. O processamento analítico baseia-se na integridade temporal desde as recomendações de Inmon e Codd.

Pelos seus objectivos, os SIE podem produzir dados e informação com maior ou menor grau de diversidade. Por exemplo, o INE apresenta-nos mais de vinte temas em torno da disponibilização. O Banco de Portugal também apresenta um número elevado. Já, por exemplo, os dados disponibilizados por uma

Direcção Geral da Concorrência e Preços reportam apenas a esta área de actividade.

Mas o aspecto mais importante, depois, naturalmente, do rigor associado aos dados, tem a ver com a sua actualidade. Dados e informação apenas são úteis quando as decisões que se podem basear neles são tomadas em tempo útil. Isto significa que deverão ser muito actuais na maioria dos processos de investigação em Economia. Esta característica deixa de ser importante quando o projecto é de História da Economia onde o próprio passar do tempo permite uma maior objectividade.

1.5. Pedidos, sugestões e definição de requisitos

Quando o analista, após o processo de avaliação da informação disponibilizada pelos SIE, tem consciência de que esta não serve os seus objectivos, mas, simultaneamente sente que pequenas modificações, ou um processo de extracção simples poderiam servi-lo, deverá dirigir-se ao organismo produtor da informação e apresentar-lhe esse problema. Muitos SIE têm consciência dessa necessidade e desenvolvem a interactividade com os seus utilizadores através de caixas de sugestões. Infelizmente, temos vindo a ganhar a ideia de que estas facilidades só têm uma acção apropriada decorrente quando respondem a um processo cujo interesse é do organismo produtor, ou quando são um aplauso. Noutras condições é vulgar o conteúdo de caixas de sugestões ser ignorado ou desencadear acções óbvias.

Um pedido formal de informação com determinadas características, definidas em requisitos, dirigido ao SIE pode ter melhores resultados do que um preenchimento de uma caixa de sugestões em virtude de o interlocutor possivelmente se situar numa função diferente na organização. Esse pedido pode inclusivamente estender-se ao suporte pretendido.

A utilização de conhecimentos pessoais é ainda o meio mais eficaz de se chegar a esta informação escondida do SIE, mas tal processo por razões óbvias deve ser banido de um relacionamento são entre produtores da informação e analistas. Dados e informação deverão jorrar dos seus repositórios desde que exista uma necessidade que possa ser suprida e independentemente de tal processo ser ou não taxado.

1.6. Revisão do processo de investigação

A impossibilidade de satisfação do analista em termos das suas necessidades de informação conduz ao fracasso do projecto de investigação tal como tinha sido previamente definido e a uma eventual redefinição. Esta questão põe-se sempre que, cientes de que os dados que se pretendia originalmente não existem, mas têm uma concretização com uma multidimensionalidade ou uma actualidade ou uma integridade temporal diferentes. A investigação tal como estava definida inicialmente é inviável, mas é possível noutro domínio relativo à mesma variável objecto. Caberá ao analista e a outros elementos com posições de decisão no projecto definir se ele deve prosseguir com os dados que há ou declarar a sua inviabilidade.

Um exemplo com alguma relação com esta descrição encontra-se no caso de estudo que apresentamos relativamente ao tratamento de associações no âmbito do IDE. Este carece de dados categóricos em maior grau do que um modelo de análise multidimensional para o IDE. Convém ter informação sobre investidor e país de origem, sobre a empresa onde se realizou o investimento e o país de destino. Estes dados não faziam parte do quadro de informação que dispúnhamos inicialmente nem eram passíveis de obtenção no Banco de Portugal. O recurso ao ICEP permitiu ultrapassar o problema com uma base menor é certo mas capazes de proporcionar algumas conclusões sobre uma componente importante do processamento analítico.

1.7. Extracção dos dados

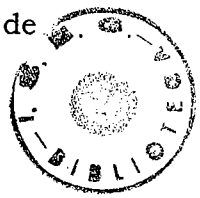
Com esta actividade, o analista importa para o seu ambiente os dados, relatórios e informação disponibilizados pelo SIE em que se materializa a satisfação das necessidades de informação.

Para apoiar este processo muitos SIE constituem repositórios compactados. Uma transferência compactada é mais rápida e menos sujeita a erros e a percalços de transmissão.

1.8. Outras actividades

As últimas actividades associadas ao processo de disponibilização e utilização de informação para a economia envolvem a constituição de bases de dados analíticas, a modelação e utilização de processos analíticos e a interpretação

de resultados e geração e divulgação de conhecimento sendo objecto de descrições mais detalhadas em outros capítulos deste documento.



2. O Sistema Estatístico Nacional

O SEN é o agente mais importante na produção e difusão de informação para a Economia em Portugal. Nesta secção vamos examinar a sua composição, princípios e algumas questões relativas à difusão da informação.

2.1. Composição do Sistema Estatístico Nacional

Nos termos da Lei n.° 6/89, de 15 de Abril, o SEN é constituído pelo Conselho Superior de Estatística (CSE) e pelo Instituto Nacional de Estatística (INE).

O CSE é o órgão do Estado que superiormente orienta e coordena o SEN, competindo-lhe nos termos do artigo 10° da mesma Lei:

a) Definir as linhas gerais da actividade estatística nacional e estabelecer as respectivas prioridades;

b) Garantir a coordenação do SEN, aprovando os conceitos, definições, nomenclaturas e outros instrumentos de coordenação estatística;

c) Apreciar o plano de actividades do INE e o correspondente relatório final;

d) Fomentar o aproveitamento de actos administrativos para fins estatísticos, formulando recomendações com vista, designadamente à utilização nos documentos administrativos das definições, conceitos e nomenclaturas estatísticas;

e) Pronunciar-se, a pedido do Governo, sobre as normas e princípios gerais que devem regular a produção de dados estatísticos referidos na alínea a do número 3 do artigo 14° do presente diploma;

f) Zelar pela observância do segredo estatístico e decidir sobre as propostas de dispensa de segredo estatístico, nos termos do número 5 do artigo 5°;

g) Aprovar a seu Regulamento Interno;

h) Propor delegações de competências do INE em outros serviços públicos ou determinar a cessação das mesmas delegações, nos termos dos números 3 e 4 do artigo 16°.

Nos termos do Artigo 24° da mesma Lei, compete ainda ao CSE apreciar projectos de diplomas que criem serviços de estatística ou contenham quaisquer normas com incidência na estrutura ou funcionamento do Sistema Estatístico Nacional.

O CSE é presidido pelo Ministro que tutela o Instituto Nacional de Estatística (INE), presentemente o Ministro do Equipamento, do Planeamento e da Administração do Território e tem como Vice-Presidente o Presidente do INE.

Em Dezembro de 1999 de acordo com o site do INE (www.ine.pt), o CSE inclui na sua composição serviços públicos, o Instituto Nacional de Estatística, o Instituto Superior de Estatística e Gestão da Informação, o Ministério das Finanças, o Ministério do Equipamento, do Planeamento e da Administração do Território, o Ministério da Justiça, o Ministério da Agricultura, do Desenvolvimento Rural e das Pescas, o Ministério da Economia, o Ministério da Educação, o Ministério da Saúde, o Ministério do Trabalho e da Solidariedade, o Ministério do Ambiente, o Ministério da Ciência e Tecnologia, o Ministério da Defesa Nacional, o Ministério da Cultura, o Ministério dos Negócios Estrangeiros, os Governos Regionais da Madeira e dos Açores, a Associação Nacional dos Municípios Portugueses, o Banco de Portugal, as Confederação Sindicais, as Confederações Patronais, a Associação Portuguesa para a defesa do Consumidor e o Conselho de Reitores das Universidades Portuguesas

Ao INE cabe exclusivamente o exercício de funções de notação, apuramento, coordenação e difusão de dados estatísticos oficiais (artigo 3° da mesma Lei). Contudo, o número 1 do artigo 16° daquele diploma refere que "para a prossecução das suas atribuições, o INE pode delegar funções oficiais de notação, apuramento e coordenação de dados estatísticos noutros serviços públicos". Fica excluída desta delegação de competências a difusão estatística que permanece exclusiva do INE[18].

[18] Neste contexto, os serviços produtores de estatísticas oficiais (entidades com competências delegadas) no âmbito do SEN são o Serviço Regional de Estatística da Região Autónoma dos Açores, a Direcção Regional de Estatística da Região Autónoma da Madeira, o Gabinete de Estudos e Planeamento do Ministério da Justiça, o Departamento de Avaliação Prospectiva e Planeamento do Ministério da Educação, o Observatório das Ciências e das Tecnologias, o Departamento de Estatística do Ministério para a Qualificação e o Emprego, o Instituto de Gestão Financeira da Segurança Social, o Secretariado Nacional para a Reabilitação e Integração das Pessoas com Deficiências, o Gabinete de Planeamento e Política Agro-Alimentar, a Direcção Geral das Florestas e a Direcção Geral das Pescas e Aquicultura.

2.2. Princípios orientadores do actual Sistema Estatístico Nacional

Centralização estatística, coordenação técnica, segredo estatístico, autoridade estatística e autonomia técnica são os cinco princípios orientadores do SEN.

Centralização estatística significa que o exercício de funções de notação, apuramento, coordenação e difusão de dados estatísticos oficiais cabe exclusivamente ao Instituto Nacional de Estatística. Contudo, para a prossecução das suas atribuições, o INE pode delegar funções oficiais de notação, apuramento e coordenação de dados estatísticos noutros serviços públicos.

O princípio da *coordenação estatística* verifica-se em duas vertentes: através do CSE que coordena o Sistema Estatístico e através do INE a quem compete a coordenação técnica criando para isso instrumentos técnicos dos quais se salientam: (1) ficheiros de empresas e estabelecimentos, (2) controlo técnico central dos questionários estatísticos, (3) recomendações internacionais sobre métodos, definições, conceitos e nomenclaturas e (4) sistemas de contabilidade nacional

De acordo com o artigo 5° da Lei do Sistema Estatístico o *segredo estatístico* visa salvaguardar a privacidade dos cidadãos, preservar a concorrência entre os agentes económicos e garantir a confiança dos informadores no sistema estatístico. Assim todas as informações estatísticas de carácter individual colhidas pelo INE são de natureza confidencial, pelo que, não podem ser discriminadamente insertas em quaisquer publicações ou fornecidas a quaisquer pessoas ou entidades, nem delas pode ser passada certidão, constituem segredo profissional para todos os funcionários e agentes que delas tomem conhecimento, nenhum serviço ou autoridade pode ordenar ou autorizar o seu exame. As informações individualizadas sobre pessoas singulares nunca podem ser divulgadas. Salvo disposição legal em contrário, as informações sobre a Administração Pública não estão abrangidas pelo segredo estatístico. As informações sobre cooperativas, empresas públicas e privadas, instituições de crédito e outros agentes económicos não podem ser divulgadas, salvo autorização do Conselho Superior de Estatística, caso a caso, desde que estejam em causa as necessidades do planeamento e coordenação económica ou as relações económicas externas.

No exercício da sua actividade, o INE pode realizar inquéritos e efectuar todas as diligências necessárias à produção de dados estatísticos e pode solicitar informações a todos os funcionários, autoridades, serviços ou organismos e a todas as pessoas singulares ou colectivas que se encontrem em território nacional ou nele exerçam a sua actividade, exceptuando-se informações relacionadas com convicções políticas, religiosas ou outras de idêntica natureza, bem como aquelas que possuam um carácter eminentemente pessoal.

A *autonomia técnica* consiste no poder conferido aos órgãos de definir livremente os meios tecnicamente mais ajustados à prossecução das atribuições do INE, agindo no âmbito da sua competência técnica, com inteira independência.

2.3. Difusão da Informação Estatística

A difusão da informação estatística oficial no âmbito do SEN constitui uma das funções essenciais do Sistema, tendo presente que as necessidades dos diferentes utilizadores - públicos e privados - tendem a alterar-se a um ritmo cada vez mais acelerado.

As novas modalidades e processos de difusão da informação estatística oficial com recurso às novas tecnologias, e a avaliação permanente das necessidades dos diferentes utilizadores, passaram a ter um papel determinante na configuração das operações e acções ao nível dos sistemas estatísticos nacionais. Assim, as preocupações mais recentes dos responsáveis pelo SEN têm a ver com a revisão do plano de publicações estatísticas oficiais visando confiná-las ao seu desígnio essencial de suporte de difusão da informação estatística oficial, de interesse nacional e geral, e por áreas temáticas, para cumprimento das obrigações de serviço público, a intensificação e alargamento da disponibilização da informação estatística oficial via INTERNET, designadamente através do projecto INFOLINE enquanto subsistema de informação em linha; e, finalmente com a intensificação, também à escala do SEN, a utilização de suportes magnéticos e ópticos, bem como a criação de bases de dados temáticas com elevado grau de acessibilidade pelos diferentes utilizadores.

3. Os Sistemas de Informação para a Economia e os factos económicos de que se constituem missão

Uma análise breve da disponibilização de informação para a Economia através da Internet permite detectar a existência de um conjunto de *sites* que operam neste domínio e ultrapassam a descrição do SEN. Apesar deste se assumir como o órgão oficial relativo à estatística, há que passar por esse conjunto de sistemas e averiguar até que ponto uma evolução no conjunto se pode traduzir em melhores produtos para o analista da Economia. Assim, organismos que não são referidos como detendo algum papel no SEN surgem como produtores e difusores de informação, por vezes centrada num único facto económico, em termos, pelo menos, interessantes. Assumimos que o SEN é, sem dúvida, o mais importante produtor e difusor de informação para a Economia, mas existindo outros organismos que também o fazem, a sua união formará o que poderemos designar Sistemas de Informação para a Economia em Portugal.

Nesta secção, vamos relacionar os sistemas de informação para a economia com os factos económicos que se deduzem da sua descrição de missão e que no fundo justificam a sua existência. Uma análise superficial permite, desde logo, detectar uma subdivisão dos SIE em termos que permitem considerá-los como de um âmbito alargado ou de um âmbito restrito. Assim prosseguiremos com uma descrição de missão de um conjunto de organismos que divulgam informação para a economia e com o relacionamento com factos económicos que são objecto dessa divulgação, num primeiro momento, centrados nos SIE de âmbito alargado e, num segundo momento, centrados nos factos económicos.

3.1. Os SIE de âmbito alargado

Consideramos como integradores desta classificação os SIE que têm por objecto um conjunto de factos económicos relativamente distinto. Da análise efectuado, entendemos que apenas o Instituto Nacional de Estatística e o Banco de Portugal se podem integrar nesta categoria.

3.1.1. O Instituto Nacional de Estatística

Tal como se referiu, cabe ao INE exclusivamente o exercício de funções de notação, apuramento, coordenação e difusão de dados estatísticos oficiais. O

INE pode delegar funções oficiais de notação, apuramento e coordenação de dados estatísticos noutros serviços públicos, ficando excluída desta delegação de competências a difusão estatística que permanece exclusiva do INE.

O INE trata a maioria dos temas estatísticos e económicos de interesse sobe a sociedade portuguesa, podendo a informação disponibilizada ser organizada pelas seguintes áreas estatísticas: (1) Ambiente, (2) Administrações Públicas, (3) Agricultura, Produção Animal e Silvicultura, (4) Comércio Internacional, (5) Comércio Interno e Outros Serviços, (6) Condições de Vida das Famílias, (7) Conjuntura Económica, (8) Contas Nacionais e Regionais, (9) Cultura, Desporto e Recreio, (10) Demografia, (11) Emprego e Salários, (12) Estatísticas Gerais, (13) Habitação, Construção e Obras Públicas, (14) Indústria e Energia, (15) Instituições Financeiras e Seguros, (16) Iniciativas de Produção e Estudos Regionais, (17) Pesca, (18) Preços, (19) Protecção Social, (20) Saúde, (21) Transportes e Comunicações, (22) Turismo e Restauração.

O INE disponibiliza um grande número de indicadores estatísticos, mas não o faz em moldes imediatamente utilizável por um analista em termos de processamento. Esta facilidade só se encontra num produto pago, o INFOLINE.

Disponibiliza, ainda, catálogos sobre as publicações mais recentes que faz acompanhar de algumas palavras de descrição e condições de aquisição. De um modo geral, a descrição não é de molde a ter-se uma ideia satisfatória sobre o conteúdo sendo o analista rapidamente dirigido para uma proposta de aquisição.

O serviço de disponibilização interactivo do INE materializa-se num produto denominado INFOLINE. Apesar disso o que é veiculado por este produto pode não ser imediatamente processável devido ao formato utilizado.

Na página inicial do INFOLINE encontra-se a seguinte estrutura: Informação Rápida, Pesquisa por Tema, Pesquisa por Palavra, Pesquisa por Unidade Territorial, Séries Cronológicas, Estudos, Retratos Territoriais.

A *Informação Rápida* é informação estatística comentada e disponibilizada pelo INE no formato FIR - "Folha de Informação Rápida" correspondendo às Notas à

Comunicação Social divulgadas pelo INE. Alguma desta informação é processável enquanto outra está em formato que não o permite.

A opção *Pesquisa por Tema* permite visualizar os temas em que se encontra organizada a informação estatística que dão acesso a (1) uma página de índice da metainformação, isto é comentários analíticos sobre a informação estatística - incluindo gráficos e cartogramas -, conceitos estatísticos, metodologia utilizada, sinais convencionais, questionários utilizados e outra informação qualitativa relevante sobre a informação estatística e (2) a subtemas correspondentes ao tema seleccionado que conduzem a quadros estatísticos que podem ser visualizados, impressos e importados em formato compactado.

O subsistema *Pesquisa por Unidade Territorial* permite a pesquisa de informação estatística relativa a um conjunto de indicadores para os diversos níveis territoriais do país (regiões NUTS I, NUTS II, NUTS III, concelhos e FREGUESIAS - RETRATO DAS FREGUESIAS).

O subsistema *Séries Cronológicas* disponibiliza dados retrospectivos relativos aos principais indicadores estatísticos. O utilizador pode consultar a metainformação associada a cada série cronológica que será complementada futuramente com comentários adicionais. É possível importar cada série para folha de cálculo. A temática associada às séries cronológicas envolve Demografia e Censos, Saúde, Emprego e Desemprego, Protecção Social, Cultura, Desporto e Recreio, Agricultura Produção Animal Silvicultura e Pesca, Industria e Energia, Contas Nacionais, Transportes Armazenagem e Comunicações, Turismo, Comércio Internacional, Comércio Outros Serviços e Preços e Condições de vida das famílias.

O subsistema *Estudos* permite o acesso a artigos divulgados em publicações do INE. Estes artigos encontram-se organizados por áreas temáticas de modo a facilitar a consulta. Antes de visualizar o artigo, o utilizador tem acesso a uma Ficha do Estudo - ainda em fase de desenvolvimento - que faz uma breve caracterização do artigo e apresenta um resumo do mesmo.

O subsistema *Retratos Territoriais* permite a pesquisa de informação estatística relativa a um conjunto de indicadores fundamentais para os diversos níveis territoriais do país (regiões NUTS I, II, III, concelhos e

FREGUESIAS - RETRATO DAS FREGUESIAS). A navegação poderá fazer-se através da selecção da região pretendida sobre um mapa desenhado no ecrã ou através da selecção de uma *list box* que contém todas as unidades territoriais associadas aos respectivos códigos NUTS. Para cada unidade territorial, poderá visualizar informação repartida por quatro tipos de indicadores: Indicadores Genéricos, Indicadores Demográficos, Actividade Económica e Indicadores Sociais.

3.1.2. O Banco de Portugal

De acordo com a respectiva Lei Orgânica, o Banco de Portugal, como banco central da República Portuguesa, faz parte integrante do Sistema Europeu de Bancos Centrais (SEBC) e, nessa qualidade, prossegue os objectivos e participa no desempenho das atribuições cometidas ao SEBC. De acordo com o Tratado da União Europeia o objectivo primordial do SEBC é a manutenção da estabilidade dos preços. Esse objectivo é prosseguido por via da política monetária.

Compete especialmente ao Banco velar pela estabilidade do sistema financeiro nacional, assegurando, com essa finalidade, designadamente, a função de refinanciador de última instância. A prossecução deste objectivo é efectuada através da supervisão das instituições de crédito e das sociedades financeiras.

O Banco emite notas com curso legal e poder liberatório e põe em circulação as moedas metálicas, embora o BCE detenha o direito exclusivo de autorizar a sua emissão. Esta função pode articular-se com outra, mediante a qual compete ao Banco de Portugal regular, fiscalizar e promover o bom funcionamento dos sistemas de pagamentos, designadamente no âmbito da sua participação no SEBC.

Além das funções referidas, compete ainda ao Banco de Portugal gerir as disponibilidades externas do País e agir como intermediário das relações monetárias internacionais do Estado, bem como aconselhar o Governo nos domínios económico e financeiro e a recolha e elaboração das estatísticas monetárias, financeiras, cambiais e da balança de pagamentos designadamente no âmbito da sua colaboração com o BCE.

A actividade do Banco situa-se em torno de um conjunto de temas económicos que se deduzem da estrutura das suas publicações e onde há que considerar: (1) Inflação, (2) Estatísticas Cambiais, (3) Estatísticas Monetárias e Financeiras, (4) Estatísticas da Balança de Pagamentos e (5) Estatísticas das Finanças Públicas.

Nas estatísticas Monetárias e Financeiras é possível encontrarmos o Balanço do Banco de Portugal e de outras instituições financeiras monetárias, Detalhes dos activos e passivos das outras instituições financeiras monetárias face a residentes e não residentes, taxas de juro e rendibilidade dos bancos. Ao nível das estatísticas da Balança de Pagamentos, temas como o Investimento Directo Estrangeiro e do Investimento em Carteira têm também tratamento.

De acordo com a sua Lei Orgânica, o Banco de Portugal assegura a centralização e a elaboração das estatísticas monetárias e financeiras, cambiais e da balança de pagamentos.

A informação detalhada relativa às estatísticas monetárias e financeiras, cambiais e da balança de pagamentos pode ser importada sob a forma de ficheiros de texto, que contêm séries temporais referentes a períodos mais longos do que os publicados no Boletim Estatístico.

A estrutura adoptada segue a hierarquia existente na publicação, desdobrando-se em grupos e subgrupos, contendo cada um deles um conjunto de quadros e/ou documentos para importação de valores.

O Banco de Portugal é responsável por uma série de publicações susceptíveis de consulta interactiva, nomeadamente o Relatório Anual, instrumento indispensável de análise sobre a economia portuguesa, os Indicadores de Conjuntura, o Boletim Estatístico, publicação mensal, onde se agregam não só as estatísticas coligidas pelo Banco de Portugal, mas também outras, oriundas de diversas fontes estatísticas, igualmente importantes para a condução da política monetária e para a execução da política cambial, o Boletim Económico, publicação trimestral, que integra artigos e textos de análise de vários aspectos de política económica e ainda a edição de estudos e investigação económica levada a cabo pelos seus quadros técnicos.

O Relatório Anual está dividido em duas partes distintas. Na primeira parte, é feita a análise integrada da "Economia Portuguesa" ao longo do ano, estando organizada por capítulos sobre as várias áreas de actividade, nomeadamente (1) enquadramento internacional e integração europeia, (2) preços, procura, produção e mercado de trabalho, (3) balança de pagamentos, (4) finanças públicas, (5) situação monetária e cambial, (6) mercados de títulos; (7) sistema bancário e (8) financiamento da economia. Na segunda parte - Relatório e Contas - é divulgada a Actividade do Banco de Portugal, o Balanço e Contas e o Relatório e Parecer do Conselho de Auditoria.

O Boletim Económico inclui análises e estudos realizados no Banco de Portugal. Nele encontramos discursos e intervenções públicas, textos de política e situação económica, artigos sobre economia aplicada.

Os Indicadores de Conjuntura integram um texto síntese, bem como gráficos e quadros relativos aos principais indicadores divulgados ao longo do mês.

O Boletim Estatístico é uma publicação bilingue português/inglês, de periodicidade mensal, onde são divulgadas as estatísticas monetárias e financeiras, cambiais e da balança de pagamentos produzidas no Banco. Inclui ainda um capítulo sobre finanças públicas e outro com as principais estatísticas gerais. Os Suplementos ao Boletim Estatístico apresentam novas séries estatísticas, bem como análises e/ou notas metodológicas sobre a informação estatística.

O Banco Central Europeu publica aproximadamente de quatro em quatro anos um guia descritivo dos sistemas de pagamento e de liquidação de títulos nos estados-membros da União Europeia, designado "Payment Systems in the European Union", ou, mais vulgarmente, "Blue Book". Anualmente, publica-se uma actualização do anexo estatístico (Blue Book Addendum).

Em Abril de 1996 o Fundo Monetário Internacional (FMI) criou o 'Special Data Dissemination Standard' (SDDS) visando reforçar a transparência, integridade, actualidade e qualidade da informação estatística. No âmbito do SDDS é disponibilizada informação sobre: dados macroeconómicos; política de divulgação ao público; política de revisões; metodologias subjacentes à preparação da informação estatística. Portugal aderiu ao SDDS em Outubro

de 1998, podendo ser consultada a informação relativa a Portugal no 'Dissemination Standard Bulletin Board' (DSBB) do FMI, acessível na Internet.

A informação estatística abrangida pelo SDDS, relativa a Portugal, é compilada pelo Ministério das Finanças, pelo Instituto Nacional de Estatística, pela Bolsa de Valores de Lisboa e pelo Banco de Portugal.

A partir de 2001, tornou-se relevante, na disponibilização interactiva proporcionada pelo Banco de Portugal, a publicação das *Séries Longas para a Economia Portuguesa pós II Guerra Mundial* a qual apresenta uma integridade temporal excelente, apesar de a sua actualidade ser relativamente pobre. No quadro seguinte, apresentamos a sua estrutura a qual segue, de certo, modo a estrutura da publicação estatística do Banco.

Quadro 26. Séries Longas para a Economia Portuguesa pós II Guerra Mundial

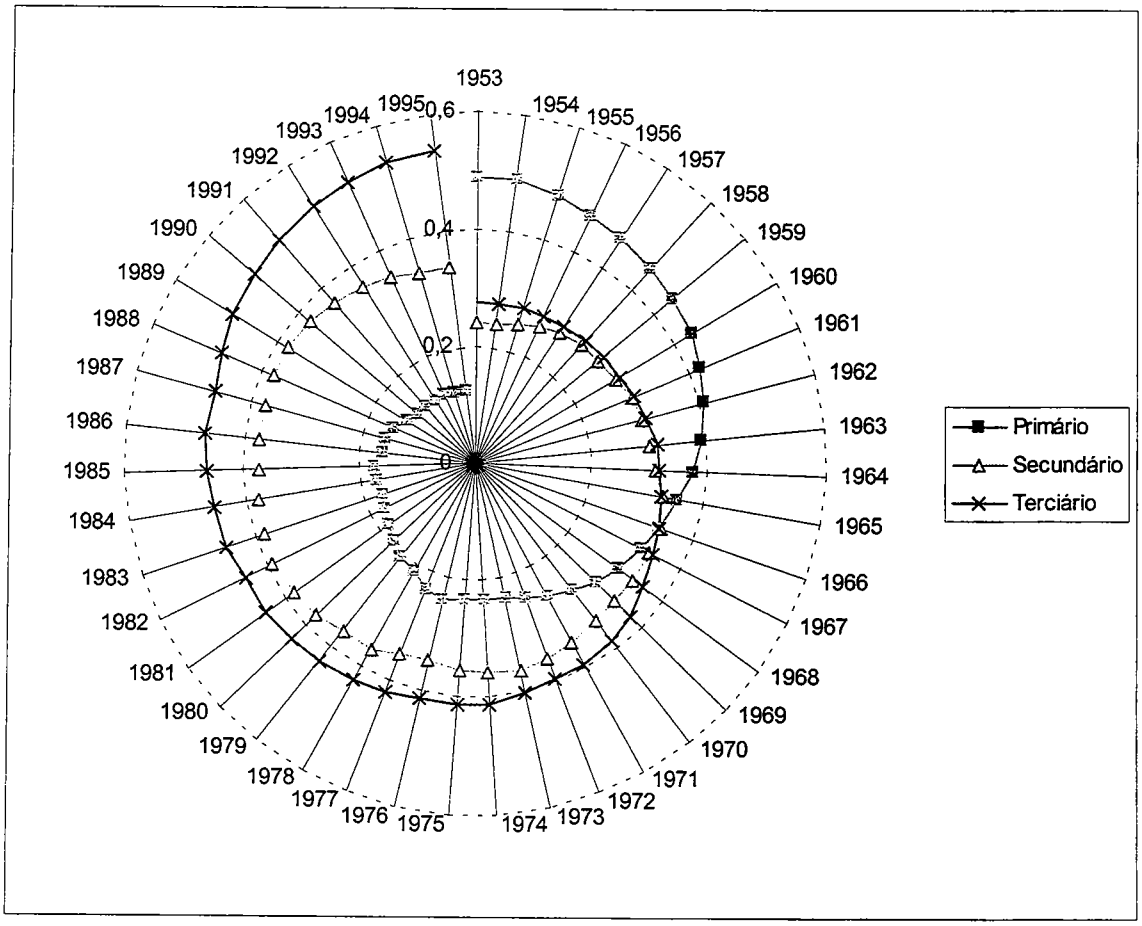
Designação da Série	Componente
Parte I - Estatísticas Monetárias e Financeiras	Balanço do Banco de Portugal
	Síntese Monetária
	Quadros de Taxas de Juro
	Quadros de Emissões de Dívida
Parte II - Estatísticas de Balança de Pagamentos	Balança de Pagamentos Externos de Portugal
	Balança de Transacções Correntes
	Balança de Capitais Não Monetários
	Reservas Oficiais Líquidas
	Balança de Pagamentos da Metrópole com o Ultramar
Parte III - Contas do Sector Público Administrativo	Receitas e Despesas das Adm. Públicas, Central, Estado e Seg. Social
	Receitas e Despesas das Adm. Regional e Local, Fundos Autónomos e Inst. Part. sem fim Lucrativo
	Impostos Correntes e Contr. Sociais Efectivas
	Impostos Ligados à Prod. e Imp. e Remun. dos Ramos não Mercantis
	Transferências Correntes e Contr. Sociais Fictícias
	Transferências de Capital; FBCP e Aqui. Líq. de Terrenos
Parte IV - População, Emprego e Desemprego	População, Emprego e Desemprego
	Emprego Total
	Trabalhadores por Conta de Outrem
Parte V - Produção, Despesa e Rendimento	Produto Interno Bruto
	Despesa Nacional
	Consumo Privado
	Formação Bruta de Capital Fixo
	Exportações de Mercadorias
	Importações de Mercadorias
	Repartição do Rendimento Nacional
	Rendimento e Poupança das Famílias e Adm. Privadas
	Poupança e Investimento

Fonte: Banco de Portugal (*www.bportugal.pt em 18.01.2001*)

Devido à integridade temporal de que são portadoras, estas *Séries Longas* são o testemunho de múltiplos processos de transformação operados na sociedade portuguesa ao longo do século XX.

O gráfico seguinte ilustra, com base nelas, o processo de fuga ao campos que acompanhou a lenta terciarização da Economia.

Figura 24. A espiral da agonia da economia agrária



Fonte: Séries Longas para a Economia Portuguesa pós II Guerra Mundial em ***www.bportugal.pt*** em 18.01.2002

3.2. Os SIE de âmbito restrito

Para além do Instituto Nacional de Estatística e do Banco de Portugal, é possível ter acesso a informação sobre a Economia a partir das páginas na Internet de todo um conjunto de organismos onde se detecta o foco num número restrito de factos económicos, existindo a possibilidade de um mesmo facto ser tratado por vários SIE. Assim uma abordagem centrada no facto, para além de estar de acordo com as considerações de Inmon (1994), permitirá

detectar os SIE que constituem esse facto como objecto da sua vocação informacional.

3.2.1. Energia

Para além do INE, detectou-se a existência de um *site* (www.dge.pt) facultando informação sobre este tema. De acordo com a sua declaração de missão, a Direcção-Geral de Energia (DGE) é o órgão do Estado Português, dependente do Ministério da Economia, responsável pela administração energética, competindo-lhe, designadamente:

a) Realizar os estudos necessários para apoiar as decisões do Governo em matéria de política energética;

b) Propor a legislação de regulamentação do sector e fiscalizar o seu cumprimento;

c) Licenciar instalações e equipamentos que produzam ou utilizem produtos energéticos e estabelecer as condições técnicas que aquelas instalações e equipamentos devem respeitar, através de regulamentos, normas e especificações;

d) Proceder à certificação de produtos, processos e sistemas energéticos e ao reconhecimento das entidades competentes para intervenção no sector energético;

e) Realizar a gestão de programas que se destinem a apoiar iniciativas que se enquadrem na política energética definida pelo Governo;

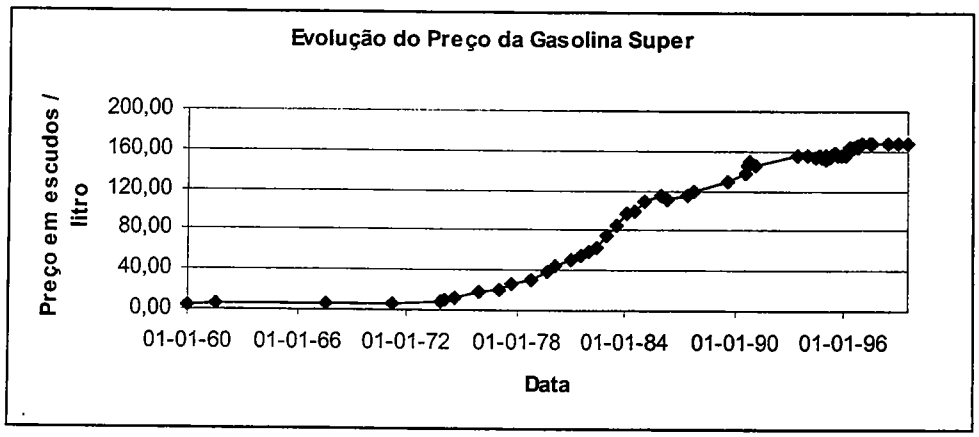
f) Promover a difusão de informação junto dos utilizadores de energia, nomeadamente a informação estatística sobre o sector e a informação que divulgue os temas da política energética nacional.

Em termos económico/estatísticos, as preocupações essenciais da DGE prendem-se com (1) os preços dos combustíveis líquidos e gasosos, (2) o balanço energético e (3) o comércio externo em matéria de energia.

A DGE mantém um conjunto de elementos informacionais, susceptíveis de processamento imediato pelo analista e com uma integridade temporal nalguns casos interessante, nomeadamente, a evolução dos preços dos

combustíveis líquidos (gasolina, gasóleo, petróleo, fuel), a evolução dos preços dos combustíveis gasosos e um conjunto de estatísticas rápidas sobre Balanço energético, Evolução do consumo, Carvão, Energia eléctrica, Petróleo, Gás natural e Conversões. A série relativa à evolução do Preço da Gasolina tem início em 01-01-1960 e a sua representação gráfica é um espelho de uma evolução sugestiva como se mostra na figura seguinte. A evolução do consumo é vista em termos de relação entre dois anos consecutivos desde 1991.

Figura 25. Quatro décadas da evolução do preço da gasolina super



É de realçar que, apesar de os dados utilizados na figura anterior, terem sido obtidos através de extracção do SIE, utilizando um ficheiro criado por este para o efeito, esta facilidade não está generalizada, havendo muitos casos em que a reintrodução de dados teria de ser feita caso os mesmos tivessem de ser utilizados.

Finalmente, a DGE disponibiliza catálogos de publicações que podem ser encomendadas interactivamente e abrangem áreas como (1) economia e conservação de energia, (2) projectos apoiados pelo SIURE, (3) energias renováveis, (4) energia eléctrica, (5) planeamento e estatística, (6) tarifas e preços, (7) combustíveis e (8) publicações didácticas. A evolução da página da DGE veio entretanto a tornar menos visível esta organização. No final de 2001, não se informa sobre os temas existentes apesar de se permitir uma ordenação da lista de publicações pelos mesmos.

3.2.2. Emprego e Formação Profissional

Nesta área o INE e o Departamento de Estatística do Trabalho e Formação Profissional (DETEFP) do Ministério do Trabalho e da Solidariedade (MTS)

serão os interlocutores privilegiados para estatísticas sobre esta área temática. O DETEFP é o serviço de produção, análise e divulgação de estatísticas, para as áreas do emprego, formação profissional e relações laborais, exercendo, nomeadamente as competências delegadas no âmbito do Sistema Estatístico Nacional. A sua página encontra-se em www.detefp.pt.

São competências do DETEFP nas suas áreas de intervenção:

a) Produzir, aperfeiçoar, desenvolver e divulgar informação estatística, em articulação com o Instituto Nacional de Estatística;

b) Efectuar estudos no âmbito de conceitos e metodologias estatísticas;

c) Apoiar a produção de informação estatística específica, pelos serviços responsáveis pelo actos administrativos geradores de informação estatística;

d) Assegurar as relações externas em matéria das suas competências, em articulação com o Departamento para os assuntos Europeus e Relações Internacionais do MTS.

Em 1999, o DETEFP publicava elementos informacionais nas área temáticas de (1) Acidentes de Trabalho, (2) Balanço Social, (3) Execução de Acções de Formação Profissional (4) Greves, (5) Duração do Trabalho (6) Impacto da Acções de Formação Profissional nas Empresas, (7) Ganhos, (8) Emprego Estruturado, (9) Necessidades de Formação Profissional das Empresas, (10) Salários por Profissão e (11) Quadros de Pessoal, que se materializavam em estudos e inquéritos com periodicidades trimestrais, anuais ou outras, sendo os dados que delas resultam integráveis em processamento por estarem em formato de texto. Nessa data, o DETEFP publicitava uma base de dados para o estudo desta área temática, mas a sua utilização não era possível em moldes interactivos. Um acesso ao chamado SISED, limitava-se a informar acerca do número de empresas, estabelecimentos e trabalhadores, dos atributos que os caracterizam no seu sistema e de quatro áreas temáticas susceptíveis de se basearem nela: estrutura empresarial, emprego, salário e ganho e duração do trabalho.

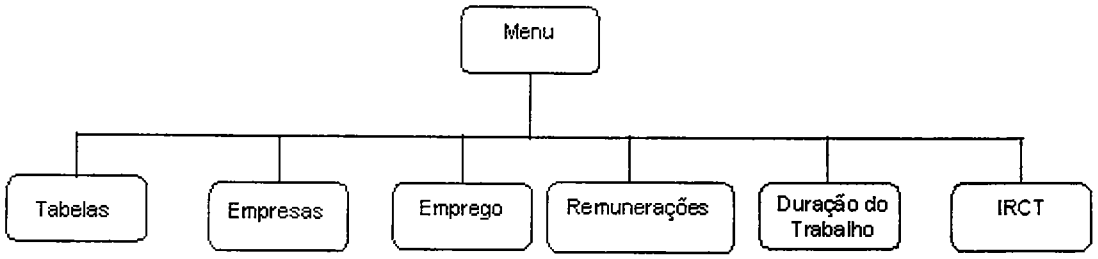
Em Outubro de 2001, a página do DETFP caracterizava-se por uma melhor organização, era portadora de mais informação em termos de relacionamento

com o utilizador embora os produtos disponibilizados não tivessem sofrido grande modificação. Assim as áreas temática envolviam : (1) Acidentes de Trabalho, (2) Balanço Social, (3) Greves, (4) Formação Profissional, (5) Emprego Estruturado, (6) Remunerações, (7) Condições de Trabalho dos Trabalhadores, (8) Quadros de Pessoal, (9) Sistema de Indicadores de Alerta e (10) demografia dos estabelecimentos. Estas áreas eram tratadas por produtos de natureza diversa, designadamente: estatísticas em síntese, publicações estatísticas, o boletim estatístico, estudos técnicos e séries estatísticas. Com excepção do que se referia a publicações estatísticas todos estes produtos se constituíam em ficheiro susceptíveis de extracção pelo utilizador. As séries estatísticas respeitavam ao período 1990-1998. O seu formato (PDF) cria, no entanto, alguns problemas em termos do seu processamento posterior.

O DETFP prometia o atendimento personalizado de clientes e o encaminhamento para outras entidades em questões fora do âmbito de intervenção.

O SISED criava grandes expectativas ao ser prometida a sua disponibilização via Internet. Esta base de dados é estruturada segundo os grandes grupos de Informação de acordo com a figura abaixo:

Figura 26. Estrutura do SISED



O site fornece informação acerca do conteúdo de cada um dos componentes.

Tabelas - Contem toda informação codificada (código e descritivo), por ano, relacionada com a informação constante da base de dados SISED. Esta informação poderá ser consultada separadamente ou, dentro de cada consulta específica como apoio na selecção de variáveis codificadas.

Empresas - Informação estatística sobre Estrutura Empresarial ilustrando a distribuição de Empresas, Estabelecimentos e Pessoas ao Serviço, para uma

determinada localização geográfica (Pais, NUT, Distrito, Concelho), por Actividade Económica e Volume de Vendas, Natureza Jurídica, Capital Social, Dimensão da Empresa e do Estabelecimento.

Emprego - Informação estatística com a repartição dos trabalhadores por conta de outrem (TCO,s) , para uma determinada localização geográfica (Pais, NUT, Distrito, Concelho), cruzando informação de acordo com o diagrama.

Figura 27. Diagrama de cruzamentos de informação de Emprego

AE	NQ, P, H, Dim, GE,
P	NQ, H, GE,
NQ	H

onde

AE Actividade económica
NQ - Nível de Qualificação,
P - Profissão,
H - Habilitação,
Dim - Dimensão da Empresa e do Estabelecimento e
GE - Grupo Etário.

Remunerações - Informação estatística com remunerações médias mensais (base e ganho) relativas a trabalhadores por conta de outrem (TCO,s) com duração do trabalho a tempo completo e remunerações horárias de trabalhadores a tempo completo e incompleto, para uma determinada localização geográfica (Pais, NUT, Distrito, Concelho), cruzando:

AE	NQ, P, H, GE,

Duração do Trabalho - Informação estatística sobre duração do trabalho dos trabalhadores por conta de outrem. A duração do trabalho considerada é para os trabalhadores de acordo com a remuneração auferida no período de referência. Os valores apresentados referem-se assim à duração do trabalho remunerada e não a duração efectiva do trabalho, para uma determinada localização geográfica (Pais, NUT, Distrito, Concelho), cruzando:

AE	NQ, P, Dim

IRCT - Informação estatística sobre a Regulamentação Colectiva de Trabalho e para cada Instrumento de Regulamentação Colectiva (IRCT). Dados relativos ao número de trabalhadores por conta de outrem abrangidos, remunerações médias mensais base e ganho e duração do trabalho efectuado, para uma determinada localização geográfica (Pais, NUT, Distrito, Concelho), com o seguinte diagrama de cruzamento.

Figura 28. Diagrama de cruzamentos de informação de IRCT

AE	D
D	Dim
IRCT	Dim, Cat

Tendo em consideração os objectivos da base de dados como instrumento de consulta on-line, os volumes de informação envolvidos e tempos de resposta, apenas se disponibiliza a informação sobre actividade económica (C.A.E.) desagregada a 1, 2 e 3 dígitos e a informação sobre profissões (C.N.P.) a 1 e 2 dígitos. Caso se pretenda para situações específicas uma maior desagregação, haverá que consultar o Departamento de Estatística.

3.2.4. Comércio Interno

Para além do INE, detectámos vários serviços que publicam informação sobre o comércio, nomeadamente, a Direcção Geral do Comércio e Concorrência (DGCC), a Confederação do Comércio e Serviços de Portugal (CCP) e o Observatório do Comércio.

A Direcção Geral do Comércio e Concorrência

A DGCC (www.dgcc.pt) é um serviço dotado de autonomia administrativa, responsável pela concepção, execução e avaliação da política de concorrência e das políticas sectoriais para o comércio e a distribuição, competindo-lhe:

a) Contribuir para a definição e execução da política de concorrência, enquanto mecanismo de regulação do desenvolvimento económico e empresarial, e acompanhar e avaliar a execução das medidas delas decorrentes;

b) Promover o funcionamento eficiente dos mercados, agindo nos domínios das estratégias e comportamentos dos agentes económicos, das estruturas dos mercados, dos mecanismos de regulação e das políticas sectoriais e do direito da concorrência;

c) Contribuir para a definição e execução das políticas sectoriais para o comércio e a distribuição, promovendo a modernização das estruturas e competitividade sustentada das empresas, atendendo à sua diversidade, especialização e defesa dos interesses dos consumidores;

d) Acompanhar as actividades comerciais e de distribuição, mantendo um conhecimento actualizado das tendências das respectivas condições gerais de funcionamento, por forma a permitir a avaliação dos efeitos das medidas de política sectorial;

e) Dar cumprimento às missões que, no âmbito do direito interno e comunitário de concorrência, são atribuídas à autoridade administrativa nacional de concorrência;

f) Acompanhar a evolução dos preços dos bens e serviços e apreciar os procedimentos relativos aos diferentes regimes de preços e negociar e celebrar, em representação da Administração, as convenções de preços;

g) Contribuir para a elaboração das propostas legislativas e regulamentares necessárias à prossecução dos objectivos das políticas acima referidas e assegurar o cumprimento da legislação aplicável nas suas áreas de intervenção, nos termos definidos por lei;

h) Apoiar o Governo nas negociações e decisões, nas instâncias internacionais, envolvendo as políticas de concorrência e as políticas sectoriais para o comércio e a distribuição, em particular no quadro da União Europeia, com vista à sua adequação aos interesses fundamentais da política económica nacional.

No âmbito das atribuições previstas anteriormente, caberá, nomeadamente, à DGCC:

a) Assegurar um conhecimento actualizado do funcionamento dos mercados nacionais dos vários bens e serviços, a fim de criar ou reforçar as condições necessárias ao funcionamento concorrencial dos mesmos;

b) Participar, sem prejuízo das competências atribuídas a outras entidades, na actividade desenvolvida por organismos e instituições internacionais em matéria de concorrência;

c) Promover a divulgação pelos agentes económicos de informação útil para a definição e formulação das respectivas estratégias empresariais.

A página da DGCC valoriza três temas económico estatísticos fundamentais: (1) Preços, (2) Estabelecimentos Comerciais e (3) Concorrência. Além disso, disponibiliza alguns elementos informacionais relativos a Legislação Nacional Básica relativa ao regime geral da defesa e promoção da concorrência, à legalidade ou ilegalidade de acordos ou práticas concertadas de empresas e à proibição de práticas individuais restritivas do comércio.

Em termos estatísticos dedica especial atenção ao índice de preços no consumidor (IPC) disponibilizando informação sobre o ultimo mês e previsão do mês seguinte e aos preços de produtos perecíveis. Esta informação utiliza um formato de difícil processamento posterior (ficheiro PDF).

A DGCC mantém um ficheiro de recolha de informação onde são identificados os estabelecimentos comerciais, sua localização, número de trabalhadores, área de venda e actividade económica e comercial exercida destinado aos comerciantes por grosso e a retalho, equiparados a retalhistas e a agentes de comércio. Tem por objectivo um melhor conhecimento do mercado e dos comerciantes, reunir informação de base para estudos sectoriais, apoiar a definição de novas políticas comerciais e apoiar ao comerciante na modernização do comércio.

São disponibilizados quadros de informação constante da base de dados do Cadastro Comercial referenciada a Junho de 1998 não processáveis imediatamente. Na versão da página da DGCC existente em 1999 anunciava-se que qualquer deles pode ser obtido em formato Excel 7, por solicitação à DGCC. Actualmente (Outubro de 2001) nada é dito sobre tal problema. Assim podemos acreditar que, em termos de facilidade de processamento posterior, a

informação disponibilizada pela DGCC piora nesse período. Os apuramentos referem-se ao número de estabelecimentos e número de pessoas vistos em múltiplas dimensões: retalhista/grossista, distrito, escalões, estatuto jurídico, actividade económica, etc.

A DGCC edita vários *boletins de preços* relativos a vários produtos: electrodomésticos, mobiliário de madeira, têxteis domésticos, loiça e cutelaria, material escolar, vestuário, bens de consumo corrente (água, electricidade, gás, transportes, comunicações, ensino, saúde e assistência, serviços financeiros, cultura e distracção, serviços relacionados com automóveis, viagens turísticas, alojamento e restauração, serviços pessoais e outros serviços). venda ao público de medicamentos e, bem assim, da legislação publicada com incidência nesses mesmos preços.

A Confederação do Comércio e Serviços de Portugal

De acordo com o *site* www.ccp.pt, a Confederação do Comércio e Serviços de Portugal, tal como decorre dos seus estatutos, representa "as entidades associativas e empresariais do sector terciário, comércio e serviços - pugnando pela sua dignificação e desenvolvimento".

Para além das iniciativas de âmbito global e com o objectivo de prosseguir, o conjunto de atribuições que lhe foram cometidas pelos respectivos estatutos, a CCP, através dos seus Gabinetes, presta ainda aos associados um conjunto de serviços individualizados, de que se destacam, em especial, os seguintes: (1) Elaboração de estudos e projectos, com especial incidência no âmbito dos programas de apoio a modernização do comercio; (2) Serviços de Consultadoria Fiscal; (3) Informação sobre diversos temas de interesse geral ou sectorial através de Circulares CCP; (4) Acesso a publicações da CCP incluindo a revista "Comercio e Serviços"; (5) Informação de carácter económico através de Boletim Económico trimestral a empresas de comércio e serviços (juntamente com o CISEP), Inquérito anual às empresas comerciais; (6) Informações sobre e legislação e normativos comunitários (em colaboração com o Eurogabinete PME).

Embora os dados disponibilizados não sejam directamente processáveis a CCP contribui com informação interessantíssima sobre o sector em Portugal e nos países da União Europeia veiculada através de uma Apresentação Estatística

do Sector onde se encontram os temas Enquadramento Comunitário e Enquadramento Nacional.

A CCP edita várias revistas que são publicitadas no seu site em termos de conteúdo e preços, nomeadamente, os números do Comércio, o Boletim Económico, o Inquérito Anual às empresas comerciais, o Inquérito Anual ao Comercio.

Observatório do Comércio

De acordo com a informação constante no *site* www.obscom.min-economia.pt, o Observatório do Comércio é um organismo independente que funciona no âmbito do Ministério da Economia, criado pela Resolução do Conselho de Ministros do dia 23 de Abril de 1998 que tem como objectivo principal a recolha, tratamento e divulgação de informação relativa à actividade comercial, assim como a realização de estudos técnicos, rigorosos e independentes, que permitam aprofundar o conhecimento sobre diferentes segmentos do sector do comércio e definir as linhas da sua evolução futura.

Assume-se como *um suporte de apoio à decisão dos agentes económicos e à formulação das políticas públicas para o sector.*

Entre os temas económico estatísticos inerentes à prática do Observatório do Comércio encontrámos (1) o comércio (2) as empresas comerciais (3) os quadros de pessoal das empresas comerciais em termos de qualificação, níveis etários, sexo, região, (4) a duração do trabalho e (5) o turismo

Os dados resultado de estudos deste SIE e de fornecimento de outros sistemas são disponibilizados e susceptíveis de processamento imediato ou após descompactação. Estão neste caso os relativos a

- PROCOM - Projectos de Urbanismo Comercial, Investimentos Comerciais da Direcção Geral do Comércio e da Concorrência - Gabinete do Gestor de Interv. Operacional do Comércio e Serviços.

- Unidades Comerciais de Dimensão Relevante, retalhistas da Direcção Geral do Comércio e da Concorrência.

- Estatísticas do Cadastro Comercial da Direcção Geral do Comércio e da Concorrência.

- Quadros de Pessoal do Ministério do Trabalho e da Solidariedade.

É de realçar que os elementos informacionais anteriormente referidos são disponibilizados pelo Observatório do Comércio em formato de folha de cálculo, portanto em condições que facilitam o trabalho do analista e lhe permitem adicionar valor ao produto enquanto nos serviços de origem ou não são anunciados, ou o são sem serem publicados ou são publicados em PDF, formato que dificulta o tratamento.

O sítio disponibiliza alguns estudos realizados e anuncia um número elevado de estudos em curso sobre emprego, formação profissional, inquérito ao comércio retalhista, mercados municipais, *franchising*, etc.

São disponibilizadas ajudas à navegação, ligações de interesse e um fórum de discussão que possibilita a criação de fóruns mesmo em matérias afastadas do comércio. À data da primeira observação do *site* (Dezembro de 1999) esta facilidade parecia ainda dar os primeiros passos. Em Outubro de 2001, já incluía algumas mensagens onde dominavam os pedidos de informação.

3.2.5. Fiscalidade

De acordo com o *site* www.dgci.min-finanças.pt, a Direcção-Geral dos Impostos (DGCI) tem por objectivo fundamental a execução da política fiscal definida pelo Governo. As atribuições da DGCI exercem-se nos domínios da liquidação e cobrança dos impostos, e da fiscalização, justiça, informação e investigação tributárias. Em consequência, compete-lhe, nomeadamente:

a) Assegurar a liquidação e o controlo da cobrança das contribuições e impostos, nos termos das leis tributárias;

b) Exercer a acção de fiscalização tributária;

c) Exercer a acção de justiça tributária e assegurar os interesses da Fazenda Nacional junto dos órgãos judiciais;

d) Estudar e propor medidas fiscais de carácter normativo.

Os temas estatísticos encontrados, prendem-se com a missão da DGCI e referem-se a (1) Execução Orçamental (informação sobre montantes orçamentados e executados do IRS, IRC, IVA , selo e outros), (2) Estatísticas do IRS, (3) Estatísticas do IRC

O *site* disponibiliza alguma informação, nomeadamente a relativa a preços, condições de encomenda e periodicidade sobre as suas publicações embora nenhuma delas seja passível de consulta interactiva quer em termos de índice quer de conteúdo propriamente dito.

Em 1999, era possível a extracção de dados relativos ao IRC e IRS em termos de Estatística Nacional, Distrital, por Concelho e Repartição de Finanças para os anos de 1994, 1995, 1996 e 1997. Em 2001, a mesma informação era proporcionada por uma aplicação que descreveremos melhor no Capítulo IV que integrava alguns operadores multidimensionais, mantendo-se activa a anterior forma de extracção possibilitadora de processamento local.

3.2.6. Contabilidade Pública

De acordo com *www.dgo.pt*, a Direcção-Geral do Orçamento superintende na elaboração e execução do Orçamento do Estado, na contabilidade do Estado e no controlo da legalidade, regularidade e economia da administração financeira do Estado, exercendo a sua acção no sector público administrativo, sobre todos os serviços e organismos da Administração Central, independentemente do seu grau de autonomia.

A Direcção-Geral do Orçamento é o órgão especializado do Ministério das Finanças que está presente em todos os Ministérios, através das suas Delegações de Contabilidade Pública. É a estas que cabe, essencialmente, fiscalizar as despesas dos serviços dos Ministérios no quadro da correcção jurídica e económica das despesas.

A Direcção-Geral, pela importância do controlo da gestão orçamental no quadro da sua organização e funcionamento, desenvolve um papel fundamental no âmbito do sistema do controlo interno da administração financeira do Estado.

Cabe, ainda, à Direcção-Geral do Orçamento esclarecer os serviços públicos quanto à melhor forma de observarem as normas de administração

necessárias a uma racional melhoria da cobrança das receitas e a uma melhor efectivação das despesas.

De acordo com a descrição cronológica inserta no site procede à disponibilização de informação sobre a execução orçamental e o Orçamento do Estado na Internet em simultâneo com a respectiva divulgação em papel

Os temas estatísticos associados à DGO resumem-se a (1) Orçamento do Estado, (2) Execução Orçamental e (3) Conta geral do Estado

A informação relativa a múltiplos elementos (Conta Geral do Estado, o Orçamento do Estado, classificações orçamentais e outra) está disponível para extracção encontrando-se o texto em formato que permite o seu processamento (Word, Excel, eventualmente comprimido). Em termos temporais os documentos são actuais, mas não são disponibilizadas versões para anos anteriores a 1996 o que lhes confere um carácter de integração temporal limitada.

O site permite a escolha de língua, procede à apresentação da DGO e disponibiliza informação de Finanças Públicas, informação sobre a reforma da administração financeira do Estado, Legislação e doutrina, listas de difusão/DGO, arquivo de transferência de documentos, ligações outros sites. Admite pesquisas por palavra e não dispõe de mapa de site.

3.2.7. Turismo

De acordo com o texto incluído no endereço www.dgturismo.pt, a redefinição do papel da Direcção Geral de Turismo (DGT), concluída com a publicação de nova lei orgânica em 1998, visou, sobretudo, adequar os seus serviços ao acompanhamento das novas realidades emergentes no sector, bem como proporcionar-lhe a capacidade para integrar no planeamento, ordenamento e estratégia turísticas outros pressupostos, do âmbito cultural e ambiental, que vêm atraindo crescente atenção.

Subjacente à reforma esteve uma concepção nova do turismo, considerado sector estratégico da economia nacional e alavanca das políticas de desenvolvimento regional.

Enumeram-se seguidamente os principais eixos de actuação da DGT nesta fase:

a) Definição, acompanhamento e execução da política de turismo.

b) Planeamento e ordenamento turístico.

c) Condução de programas de valorização e de ajustamento da oferta turística.

d) Funções regulamentadoras e de controlo das actividades turísticas.

e) Intensificação da investigação sobre o sector.

f) Sistematização da informação sobre o sector e sua divulgação através do recurso às modernas tecnologias de informação.

g) Apoio directo ao Governo nas negociações e decisões em instâncias internacionais.

h) Reforço das parcerias e da assistência aos clientes.

i) Conclusão da modernização da DGT, através da informatização plena dos seus serviços.

Dos assuntos cobertos, dentro da temática TURISMO, destacam-se: (1) economia do turismo, (2) profissionais do turismo, (3) oferta turística, (4) procura turística e (5) hotelaria.

A informação disponibilizada interactivamente reporta-se a uma síntese dos principais resultados estatísticos, metodologia, quadros estatísticos e dados Consolidados

A síntese dos principais resultados estatísticos apresenta valores relativos a um ano relativos a taxas de ocupação média para a hotelaria, dormidas na hotelaria, contagens de passageiros e volume de entradas de visitantes

O segundo elemento, metodologia, permite ao analista conhecer um conjunto de conceitos gerais relativos aos Projectos Estatísticos da Responsabilidade da Direcção Geral de Turismo *(Divisão de Recolha de Informação e Estatística)* o que torna possível conhecer a realização de várias acções como:

1. Inquérito de Fronteira para conhecer os movimentos de entradas em Portugal desagregados por países, por tipo de visitantes e por vias de acesso assim como caracterizar o perfil do Turista.

2. Inquérito ao Turismo no Espaço Rural para Apuramento dos índices de ocupação-cama, além dos dados sobre dormidas e hóspedes entrados.

3. Inquérito aos Circuitos Turísticos traduzido em Inquérito às agências de viagens, para aferir as ocupações dos circuitos turísticos, bem como as nacionalidades dos passageiros.

4. Inquérito aos Preços ao Balcão na Hotelaria para Apuramento dos preços médios por categorias e regiões.

5. Inquérito às Receitas Correntes dos Hotéis para acompanhamento dos preços médios por dormida e dos preços médios reais por quarto duplo.

6. Inquérito de Conjuntura às Perspectivas de Evolução de Curto Prazo com o objectivo de Conhecer a opinião dos empresários do sector sobre as perspectivas de evolução da procura turística a curto prazo.

7. Inquérito aos Parques de Campismo para apuramento de elementos sobre receitas e despesas nos parques de campismo.

8. Sondagem Periódica à Hotelaria para apuramento dos índices médios de ocupação-quarto/cama.

Os quadros estatísticos apresentam apenas os últimos dados estatísticos o que diminui o seu interesse em termos de integridade temporal, Não sendo processáveis estes dados referem-se a Entrada de visitantes estrangeiros, Ocupação de quartos camas, Dormidas na hotelaria e Receitas e Despesas de Turismo

A informação sobre dados consolidados estava ainda em construção à data de visita ao *site*.

O Centro de Documentação da Direcção-Geral do Turismo dispõe de um fundo documental constituído por estudos, relatórios, teses, estatísticas, actas de reuniões, manuais, revistas, recortes de imprensa, legislação nacional e comunitária, num total de mais de 7000 títulos de monografias e de cerca de

900 títulos de publicações periódicas. Todos estes assuntos podem ser pesquisados na base de dados TUR, actualmente constituída por cerca de 7000 referências bibliográficas. As publicações podem ser consultadas na sala de leitura do CD - TUR.

O Inventário de Recursos Turísticos (IRT) é uma base de dados relacional, desenvolvida pela Direcção-Geral do Turismo, que constitui um acervo de informação sobre a oferta turística nacional único no país. O IRT contempla 190 tipos de recursos turísticos, entendendo-se por recurso turístico todo o elemento natural, actividade humana ou seu produto, capaz de motivar a deslocação de pessoas, de ocupar os seus tempos livres ou de satisfazer as necessidades decorrentes da sua permanência. Comporta actualmente cerca de 80 000 recursos individualizados, caracterizados por uma ficha de detalhe, com informação alfanumérica relevante, completada em muitos casos com informação multimédia. A informação do IRT, de âmbito particularmente abrangente, encontra-se sistematizada segundo uma grelha de classificação, que abrange as áreas de Património Natural e Cultural, Equipamentos, Actividades e Eventos.

Relacionado com o IRT, o SIGRT (Sistema de Informação de Gestão de Recursos Turísticos) é um projecto promovido pela DGT (Direcção Geral do Turismo) que se assume como fonte de informação sobre o potencial turístico do país. Na génese do projecto esteve a necessidade de dotar a DGT de instrumentos de apoio à análise e as decisões no domínio do planeamento e do ordenamento turístico.

3.2.8. Apoio ao Exportador e ao Investidor Externo

De acordo com www.icep.pt, o ICEP é o organismo oficial responsável pela promoção da economia portuguesa no mundo, nas vertentes do Investimento, do Comércio e do Turismo de Portugal. A sua actividade está centrada em cinco áreas específicas nas áreas de promoção da imagem global de Portugal, promoção de bens e serviços portugueses, apoio à internacionalização das empresas portuguesas, captação de investimento directo estrangeiro e promoção de Portugal como destino turístico, tanto para o mercado externo como para o interno. No âmbito da sua actividade, o ICEP tem competência para o exercício do seguinte conjunto de funções:

a) recolha, tratamento e difusão de informação especializada sobre Portugal, o turismo português, a actividade das empresas, a oferta dirigida aos importadores e investidores estrangeiros e os mercados externos;

b) acções de sensibilização e de conselho, gere instrumentos de incentivo e desenvolvimento de actividades de promoção, por forma a responder com eficácia às necessidades dos diferentes destinatários dos seus serviços;

c) serviços de apoio especializado nos seguintes domínios da definição e implantação de estratégias de abordagem dos mercados externos, detecção de oportunidades de negócio, elaboração de projectos de cooperação empresarial com parceiros portugueses e informação técnico-comercial sobre procedimentos relativos ao comércio, investimentos e turismo;

d) planeamento e execução anual de acções de promoção em mercados externos, entre as quais:

A designação da organização permite detectar os grandes interesses em matéria económica e estatística, concretamente: (1) Investimento Externo, (2) Turismo e (3) Comércio Externo

O ICEP permite a consulta de uma base de dados relativa a Empresas Exportadoras, Feiras em Portugal e Oportunidade de negócios a qual se encontrava em remodelação no momento da pesquisa (Dezembro de 1999) e portanto inoperacional.

O ICEP publicita na Internet as suas publicações sobre Mercados Externos, a Oferta Portuguesa, Portugal e outras edições possibilitando uma aquisição através do contacto pelo *site*.

Em 1999, o ICEP disponibilizava um conjunto de indicadores relativos à actividade económica na sua área de actuação:

- Indicadores Sócio Económicos de 1993 a 1997 com dados utilizáveis pelo analista relativos a Educação, Saúde, Indicadores de Conforto, Bens de Equipamento e Meios de Transporte;
- Investimento Directo Português no Estrangeiro em formato gráfico;
- Investimento Directo Estrangeiro em Portugal;

- Comércio Externo (1) com dados utilizáveis Entre 1993 e 1997 em termos Globais e Por grupo de Produto em 1998 e (2) Em formato gráfico a Distribuição Geográfica do Comércio Internacional Português;
- Turismo com informação gráfica sobre turistas, dormidas e receitas de turismo para 1997;
- Indicadores Económicos (Valores de 1995 a 1998 e previsão para 1999 e 2000).

Em Outubro de 2001, o sítio do ICEP mostrava-se muito diferente daquele que nos serviu de base para este documento. Alguns dos elementos referidos existiam ainda, mas em formato não processável. Por outro lado, anunciava-se uma base de dados *(EconomiaNet)*, devendo o utilizador satisfazer um pagamento prévio para lhe ter acesso. Esta base de dados cobria já várias áreas da Economia e disponibilizava os dados em formato folha de cálculo. As áreas consideradas eram as de (1) Macroeconomia, (2) Comércio Internacional, (3) Investimento Directo, (4) Turismo, (5) Mercados (6) Produtos e (7) Nova Economia e Serviços. Todos os elementos numéricos disponibilizados tinham a sua fonte primária noutro SIE.

3.2.9. Cenários para a Economia

Vários serviços disponibilizam informação relativa a questões de cenários de desenvolvimento para a Economia Portuguesa como é o caso do Departamento de Prospectiva e Planeamento do Ministério de Administração Interna, do Gabinete de Estudos e Prospectiva do Ministério da Economia e do Departamento de Estudos e Previsão do Ministério das Finanças.

Departamento de Prospectiva e Planeamento

O DPP, Departamento de Prospectiva e Planeamento (ex-DCP), é uma Direcção-Geral do Ministério do Equipamento, do Planeamento e da Administração do Território (Decreto-Lei nº 4/95 de 17 de Janeiro) vocacionada para o estudo, concepção e proposta da estratégia de desenvolvimento. De acordo com o texto do *site www.dpp.pt* em Dezembro de 1999, entre as suas funções principais, encontramos:

a) Realizar e participar em estudos de prospectiva nas áreas política, social e económica, de âmbito internacional, com especial incidência nos de âmbito comunitário.

b) Analisar a evolução económico-social mundial, em especial a das zonas geográficas e sectores com maior relevo para Portugal.

c) Analisar e acompanhar a evolução económico-social do País, em estreito diálogo com outros serviços da Administração e com o sector privado.

d) Preparar cenários e trajectórias possíveis de evolução da economia e sociedade portuguesas e propor as grandes linhas da estratégia de desenvolvimento integrando e articulando as políticas sectoriais e espaciais, em especial para a preparação das Grandes Opções do Plano (GOP).

e) Preparar as grandes linhas de enquadramento dos planos de desenvolvimento económico e o PIDDAC (Programa de Investimentos e Despesas de Desenvolvimento da Administração Central) e avaliar o impacte macroeconómico dos grandes projectos de investimento.

Entre os temas económico/estatísticos susceptíveis de enquadrarem a actividade do DPP encontramos como mais importantes os seguintes: (1) Indicadores Macroeconómicos e (2) Prospectiva

Em 1999, o DPP mantinha uma folha com informação sobre Principais Indicadores Macroeconómicos os quais eram processáveis pelo analista com algumas dificuldades por desformatação após transferência. De um ponto vista temporal encontrava-se informação para os anos entre 1994 e 1998 e os indicadores referem-se a Consumo privado de residentes, Consumo Público, Formação Bruta de Capital Fixo, Procura Interna, Exportações de Bens e Serviços, Procura Global e Importações de Bens e Serviços

Eram ainda disponibilizados valores do Produto Interno Bruto, PNBpm por habitante, Inflação, Índice de preços no consumidor, Deflatores do PIB e do Consumo Privado, salários, produtividade e emprego para além de informação financeira. Num segundo quadro eram disponibilizadas perspectivas de evolução a curto prazo em termos de PIB e Inflação para Portugal e vários países.

Em 2001, a lista de indicadores cresceu, sendo organizados em termos de: (1) principais indicadores macroeconómicos – Portugal (para o período entre 1995 e 2000), (2) estimativas e previsões macroeconómicas - Portugal e (3) perspectivas de evolução de curto prazo - contexto internacional. Estes indicadores eram organizados em documentos PDF.

O DPP publicita duas bases de dados para uso dos seus utentes: CRONOL e Econos. A primeira define-a como um sistema de informação de estatísticas económicas e sociais de âmbito nacional e internacional, constituído por 4 módulos:

- Sistema de Informação de Estatísticas Históricas - inclui tratamento e gestão de dados estatísticos, não actualizáveis, sob a forma de séries cronológicas.

- Sistema de Informação de Estatísticas Correntes - inclui tratamento e gestão de dados estatísticos, sob a forma de séries cronológicas.

- Sistema de Informação de Contas Nacionais base 1977 - inclui todos os quadros das novas contas nacionais portuguesas (Base 77), de 1977 a 1990, sob a forma de séries cronológicas.

- Sistema de Informação de Contas Nacionais base 1986 - inclui todos os quadros das novas contas nacionais portuguesas (Base 86), desde 1986, sob a forma de séries cronológicas.

A ECONOS contém três bases de dados bibliográficas integradas: BIBL, LEGE e CAPE:

- A base de dados BIBL contém fontes de informação no domínio económico-social, metodologias, indicadores e estudos, nacionais (desde 1953) e estrangeiros (desde 1984) recebidas no DPP.

- A base de dados LEGE disponibiliza referências de diplomas publicados no Diário da República referentes a medidas de política económico-social, desde 1984.

- A base de dados CAPE referencia os títulos dos periódicos (portugueses, estrangeiros e de organizações internacionais) existentes no DPP,

apresentando um resumo dos indicadores difundidos, no caso de publicações de carácter estatístico.

- A informação destas Bases de Dados só está acessível no DPP, podendo ser fornecida a solicitação por Email ou Fax.

O DPP apresenta uma descrição das publicações mais recentes (desde 1994) que se encontram organizadas de acordo com a seguinte classificação: (1) e Ensaios de prospectiva e exercícios de cenarização, (2) análise sobre a evolução da situação nacional e internacional, com especial relevo da europeia e (3) estudos sobre a economia e sociedade portuguesas. Algumas destas publicações são disponibilizadas em linha o que é reflectido no quadro seguinte.

Quadro 27. Algumas publicações disponibilizadas pelo DPP

Título da publicação	Descrição
Informação Internacional. Análise Económica e Política	Apresenta uma análise da situação internacional, nas suas vertentes económica e política com vista à elaboração de estudos de prospectiva, especialmente no que se refere à envolvente externa da economia e sociedade portuguesas. Esta publicação encontra-se estruturada em torno de 7 áreas temáticas: "As Grandes Economias", "Economia Internacional"; "Tendências da Economia Mundial"; "Eurotendências"; "Horizontes"; "Questões Geopolíticas Centrais"; e "Em Foco".
Informação Económica	Publicação trimestral constituída por dois volumes, em que o primeiro volume consta de uma análise da conjuntura internacional e nacional, na qual se integra um conjunto de painéis seleccionados de indicadores relevantes para a avaliação da situação em cada domínio de análise, bem como de uma síntese dos principais acontecimentos económico-sociais, identificados quer em termos de medidas de política quer em termos de iniciativas empresariais e o segundo volume constitui um anexo estatístico detalhado com informação anual e infra-anual para um conjunto alargado de variáveis e indicadores socioeconómicos.
Situação Económico-social em Portugal	Apresenta uma síntese da evolução da situação internacional e analisa a evolução macroeconómica em Portugal, relativamente às principais variáveis: procura, actividade produtiva, balança de pagamentos, preços e rendimentos, mercado de trabalho, situação monetária e cambial, mercado de capitais e finanças públicas.
Portugal em Números. Indicadores Económicos e Sociais	Publicação anual (editada em português e inglês), onde se divulgam dados estatísticos sobre o território português, demografia e mercado de trabalho, caracterização global, sectorial e regional da economia.
Distribuição do Rendimento e da Pobreza em Portugal e suas Regiões	O documento analisa as desigualdades na repartição do rendimento, bem como a evolução da pobreza em Portugal e suas regiões (desagregação ao nível de NUTS II) na primeira metade da década de 90.

Fonte: DPP (www.dpp.pt em 18.10.2001)

Cabe ainda referir a existência a acompanhar estas publicações de informação económica numérica consubstanciada em quadros portadores de uma multidimensionalidade e temporalidade razoáveis. O DPP constitui-se, assim, como um sítio de consulta obrigatória para um analista que necessite de informação macroeconómica. O processamento posterior dos dados destes quadros apresenta algumas dificuldades por estarem em formato PDF.

Gabinete de Estudos e Prospectiva da Economia

De acordo com a descrição constante no *site* www.gepe.pt, trata-se de um organismo do Ministério da Economia vocacionado para o estudo das questões económicas numa óptica prospectiva e integrada e para a produção da informação qualificada de natureza envolvente e estratégica que orienta a sua actividade segundo três grandes objectivos:

a) Proporcionar informação de natureza envolvente e estratégica dirigida aos agentes económicos.

b) Fomentar, na base de parcerias e de cooperação, actividades de reflexão e de debate de ideias envolvendo as universidades, centros de investigação e as associações empresariais.

c) Apoiar a fundamentação, configuração e avaliação das políticas económicas relevantes no âmbito do Ministério da Economia.

As sua funções integram:

a) Promover e realizar estudos das questões económicas numa óptica de identificação de trajectórias possíveis de desenvolvimento.

b) Avaliar as políticas relevantes no âmbito do Ministério da Economia.

c) Coordenar, tratar e analisar a informação estatística nas áreas de intervenção do Ministério.

d) Proporcionar informação qualificada aos agentes económicos.

Os temas estatísticos são de grande variedade, englobando questões que vão do Investimento Estrangeiro ao Mercado Automóvel e a questões de inovação.

O GEPE anuncia interactivamente as suas publicações que tratam os mais variados temas económicos. Estas publicações englobam a revista Economia e Prospectiva, textos sobre Temas da Economia, documentos de trabalho e documentos publicados em imprensa da especialidade

È possível obter algumas dessas publicações (documentos de trabalho e imprensa) por transferência imediata. As outras publicações apresentam um *abstract* e preços.

No momento de realização da primeira pesquisa (Dezembro de 1999) à informação disponibilizada pelo GEPE, não foi possível realizar qualquer consulta a dados numéricos por estar em construção todo o processo de disponibilização de estatísticas.

Em Outubro de 2001, a situação era completamente diferente: O GEPE publica informação interessantíssima sobre a Economia Portuguesa num formato que permite aos analistas fazerem os seus próprios estudos a partir dela sem reintrodução, isto é, processável. Para além disso, este organismo situou-se na vanguarda do processo de disponibilização ao proporcionar uma ferramenta multidimensional para consulta do Comércio Externo Português.

Quadro 28. Estatísticas disponibilizadas pelo GEPE

Produto	Descrição
BDCI - Base de Dados Comércio Internacional	Valores anuais dos fluxos de importação e de exportação entre Portugal e o resto do Mundo, por países ou áreas geográficas e por sectores ou tipo de produtos.
Séries	Quadros e Gráficos com séries temporais superiores a 5 anos. - Contas Nacionais - Preços e Salários - Câmbios
Estatísticas Recentes	Quadros e Gráficos com dados anuais recentes, desde 1995. - Sectores no âmbito do Ministério da Economia - Contas Nacionais - População, Emprego e Desemprego - Regiões - Índices de Actividade e de Preços na Indústria - Preços - Investimento Directo Estrangeiro - Tecnologia - Comparações Internacionais
Fichas Informativas	Quadros e Gráficos com comentários sob a forma de legendas e de pequenos textos. - Países - Comércio Internacional - Sectores - Quotas de Mercado

Fonte: GEPE (www.gepe.pt em 18.10.2001)

Cabe, no entanto referir, que, apesar da excelente qualidade da informação disponibilizada pelo GEPE em termos de processamento posterior e de integridade referencial, do ponto de vista da multidimensionalidade de análise ela já é menos satisfatória com excepção do caso da BDCI.

Direcção-Geral de Estudos e Previsão

De acordo com o conteúdo do site www.dgep.pt, no âmbito do artº 14º do DL n.º 158/96, de 3 de Setembro de 1996, que aprova a Lei Orgânica do Ministério das Finanças, a Direcção-Geral de Estudos e Previsão (DGEP) tem por missão fundamental o conselho económico e técnico do Ministro das Finanças em matéria de política macroeconómica, tendo por base a investigação científica teórica aplicada no domínio da economia, bem como o acompanhamento da conjuntura económica numa perspectiva de antecipação dos principais agregados macroeconómicos, permitindo a definição das políticas adequadas à prossecução dos objectivos de estabilização conjuntural e de desenvolvimento económico. Compete à DGEP, designadamente:

a) Realizar trabalhos de investigação científica no domínio da economia, particularmente em matérias relevantes para o apoio da decisão e definição da política económica;

b) Elaborar regularmente análises da conjuntura económica portuguesa e estabelecer estimativas macroeconómicas de curto prazo;

c) Acompanhar a evolução económica e financeira internacional e as diferentes políticas adoptadas;

d) Estudar e propor medidas de política económica, nomeadamente nos domínios das políticas orçamental, fiscal, monetária, cambial e de rendimentos e preços;

e) Elaborar projecções das principais variáveis macroeconómicas, tendo em vista a programação orçamental de médio prazo;

f) Analisar o impacte da evolução dos agregados macroeconómicos relevantes na gestão e controlo da política fiscal e orçamental;

g) Desenvolver e aperfeiçoar a informação estatística relativa à actividade financeira do sector público administrativo, em colaboração com o Instituto Nacional de Estatística e com os serviços do Ministério das Finanças;

h) Participar nos trabalhos preparatórios do Orçamento do Estado e das Grandes Opções do Plano;

i) Acompanhar a actividade do Conselho Económico e Social.

Os seguintes temas económico estatísticos são susceptíveis de encontrarem uma resposta na DGEP (1) Cenários Macroeconómicos, (2) Orçamento do Estado, (3) Programas de Estabilidade e Crescimento (4) Indicadores de conjuntura (5) Indicadores económicos-critérios de convergência (6) Indicadores de Finanças públicas

A DGEP disponibiliza em formato processável um conjunto de dados sobre a Economia Portuguesa e faculta o acesso a um conjunto de *Documentos oficiais nacionais.* Um dos documentos mais interessantes do DGEP é designado por *Economia Portuguesa.* Até 1999 foi utilizado como complemento ao título a expressão *Estabilidade e Crescimento,* posteriormente a essa data passou a utilizar-se *Inovação e Competitividade.* Esta preocupação com a *inovação* está bem patente na edição relativa a Abril de 2001 que dedicou um capítulo a este grande tema conforme podemos constatar no quadro seguinte.

Quadro 29. O documento Economia Portuguesa do DGEP

Grande Tema	Subdivisão do Tema
1 Desenvolvimentos Recentes e Perspectivas	11 Área do Euro - Desenvolvimentos Recentes - Diferenças Económicas na Área Do Euro - Melhoria no Mercado Trabalho - Estabilidade dos Preços - Situação das Finanças Públicas - Situação Monetária e Cambial 12 A Economia Portuguesa - O Actual Ciclo Económico - Procura Agregada - Emprego e Desemprego - Preços e Salários - Balança Corrente e de Capital
2 Mercado de Trabalho	21 População Activa, Emprego e Desemprego 22 Investimento em Capital Humano 23 Salários
3 Finanças Públicas	31 Desenvolvimentos Recentes - Cobrança de Impostos Mais Eficiente - Alteração da Estrutura da Despesa Pública - Redução do Peso da Dívida Pública 32 Financiamento do Estado e Gestão da Dívida 33 Privatizações - Programa de Privatizações 2000–2002
4 Mercado de Capitais	41 Evolução Recente 42 Regulamentação e Supervisão 43 O Sistema Bancário Português -Desenvolvimentos Recentes
5 Reformas no Mercado de Bens e Serviços	51 Acelerar a Integração Económica 52 Promover a Concorrência - Política de Concorrência e Ajudas Estatais - Melhorar o Ambiente Empresarial - Aumentar A Eficiência da Administração Pública 53 Liberalização Das Indústrias de Rede - Telecomunicações - Electricidade - Combustível e Gás - Transportes - Serviços Postais Caixa 9 Principais Medidas de Liberalização: Indústrias de Rede
6 O Desafio da Inovação	61 Um Novo Paradigma 62 Prioridades Políticas – Passado e Presente - Construir A Base do Sistema Científico e Tecnológico - A " Sociedade da Informação" 63 Investir em Capital de Conhecimento - Recursos Financeiros 64 Investimento em Capital de Risco - Recursos Humanos 65 Avaliação 66 Algumas Ideias Para o Futuro Caixa 10 Alguns Dados Sobre Tics em Portugal Caixa 11 Programa Integrado de Apoio à Inovação

Fonte: DGEP (www.dgep.pt em 18.10.2001)

O documento atrás descrito em termos estruturais é acompanhado de um extenso anexo estatístico cuja utilidade se reconhece. A sua integração em processamento posterior é prejudicada pelo formato actual: PDF. É de realçar que o mesmo anexo, em 1999, era disponibilizado em formato folha de cálculo.

Tal *inovação* foi comum à publicação de *Indicadores de conjuntura* e a *Pocket Book*. Será que o DGEP não acredita naquilo que escreve nos seus documentos oficiais?

3.2.10. Justiça

De acordo com o sítio *www.gplp.mj.pt*, o Gabinete de Estudos e Planeamento do Ministério da Justiça (GEPMJ) foi criado pelo Decreto-Lei n.º 238/80, de 18 de Julho, tendo as suas atribuições e orgânica definidas no Despacho Normativo n.º 105/85, de 8 de Novembro.

O GEPMJ é o órgão sectorial de planeamento, estatística e informática da área orgânica da Justiça, de estudo e apoio técnico no tratamento da informação, com atribuições que se situam nas seguintes áreas de planeamento, estatísticas da justiça, coordenação informática, estudos jurídicos de direito da informática e análises jurídico-sociais e selecção e difusão da informação.

O Sistema de Informação Estatística da Justiça (SINEJUS) compreende toda a informação relacionada com a actividade judiciária, registral, notarial, policial, dos organismos de execução das penas e medidas, de intervenção social e de outros, ligados directa ou indirectamente ao sistema de administração da justiça.

Os dados são, por regra, recolhidos mensalmente e por via postal, em instrumentos de notação estruturados, organizando-se posteriormente sob a forma de base de dados.

Enquanto sistema unitário, o SINEJUS subdivide-se em quatro subsistemas principais : estatística judiciária, estatística registral e notarial, estatística policial e dos organismos de apoio à investigação e estatística de execução das penas e medidas e de intervenção social.

Em todos estes domínios existe uma cobertura estatística suficiente face às necessidades de informação reveladas pelos utilizadores, procurando

satisfazer-se a procura através do alargamento do elenco de publicações editadas, da satisfação de pedidos pontuais de informação, da colaboração e participação em projectos de investigação desenvolvidos sobre a temática da administração da justiça e da adopção de novas tecnologias na difusão da informação, nomeadamente por via da Internet e da criação de apresentações.

Constitui preocupação actual o alargamento das Estatísticas da Justiça a novas áreas de intervenção, nomeadamente no domínio da resolução extra judiciária de conflitos (Tribunais arbitrais, Comissão de Protecção de Menores), o aperfeiçoamento dos métodos de recolha e o controlo da qualidade de informação.

Os temas estatísticos fundamentais incluem não apenas elementos quantitativos (1) Segredo Estatístico, (2) Estatísticas de Processos em Tribunais (3) Estatísticas de Crimes, (4) Estatísticas de Registos e Notariado (5) Estatísticas de Serviços Prisionais e (6) Reinserção Social.

Em 1999, o *site* disponibilizava várias publicações no âmbito da justiça sendo de assinalar a que se refere a questões de segredo estatístico num momento em que se pensa em redefinir todo o papel do Sistema Estatístico Nacional.

Alguns elemento informacionais disponibilizados neste contexto eram processáveis embora isso não aconteça com um lote importante de elementos de séries retrospectivas. A integridade temporal não era total e no máximo envolvia dados iniciados em 1994. Encontravam-se assim:

- Séries retrospectivas;
- Movimento de Processos nos Tribunais – 1998;
- Estatísticas dos Registos e do Notariado – 1998;
- Justiça de Menores – 1998;
- Justiça Cível – 1998;
- Estatísticas Criminais – 1998.

Nas Séries Retrospectivas encontrámos informação para vários anos relativa a:

- Tribunais;
- Registos e Notariado (Dados gerais, Registo Civil, Registo Predial, Registo Automóvel) - algumas destas séries têm valores de 1994 a 1998;
- Polícia (Crimes registados pelas autoridades policiais com dados de 1995 a 1998;



- Serviços Prisionais (dados sobre o movimento de reclusos num determinado ano em formato não processável);
- Reinserção Social (para um determinado ano).

Em Outubro de 2001, a situação das estatísticas da Justiça era relativamente diferente, sendo possível encontrar em formato PDF vários conjuntos de estatística, designadamente:

- Estatísticas de 2000 do Gabinete de Política Legislativa e Planeamento:
  - Estatísticas da Justiça - Dados Pré-definitivos 2000;
  - Criminalidade Registada 2000;
  - Estatísticas dos Registos e do Notariado 2000.

- Estatísticas de 1999:
  - Justiça de Menores 1999;
  - Movimento dos Processos nos Tribunais;
  - Estatísticas dos Registos e do Notariado 1999;
  - Justiça em Números;
  - Estatísticas da Justiça 1999;
  - Estatísticas Criminais 1999;
  - Justiça Cível 1999.

A perspectiva retrospectiva parece ter sido perdida, pois não descortinámos as séries associadas. Há que acrescentar a excessiva existência de ficheiros corrompidos que impediu um exame mais completo dos resultados.

3.2.11. Saúde

De acordo com o descrito no seu *site* (www.dgsaude.pt) compete à Direcção Geral da Saúde:

- Promover a saúde dos portugueses, elevando o nível de bem estar físico, mental e social e contribuindo para a melhoria da sua qualidade de vida.
- Desenvolver uma visão estratégica da evolução do sistema de saúde centrado no cidadão, propondo cenários alternativos, susceptíveis de informar a decisão política e as estratégias apropriadas ao desenvolvimento do Serviço Nacional de Saúde e promovendo a aplicação concertada de orientações que concorram para a coesão do sistema de saúde e para o enquadramento dos seus principais actores.

- Influenciar a evolução do sistema de saúde, desenvolvendo mecanismos de objectivação e de influência múltipla e concertada na criação de comportamentos pró-activos, capazes de traduzirem em acções concretas e em "boas práticas" o conhecimento científico adquirido em relação à saúde e aos serviços de saúde.
- Exercer uma posição de liderança no país, no que se refere aos valores, ideias, factos e formas de actuação sobre a saúde Desta forma, a Direcção-Geral da Saúde é uma referências indispensável para todos aqueles que pensam e actuam no campo da saúde.

Neste contexto, cabem-lhe entre outras as funções seguintes:

a) Preparar, lançar e explorar inquéritos de saúde quer de natureza estatística, quer epidemiológica;

b) Assegurar a inserção das estatísticas de saúde no Sistema Estatístico Nacional;

c) Assegurar a realização de inquéritos na área da saúde no quadro do Sistema Estatístico Nacional;

d) Acompanhar e controlar a execução dos programas de investimento do Ministério da Saúde;

No campo da informação estatística, detectam-se como preocupações da DGS as questões relacionadas com (1) Estatísticas de Mortalidade (2) Recursos do Serviço Nacional de Saúde e (3) Formação dos Profissionais de Saúde.

Em Dezembro de 1999, as Estatísticas de Mortalidade constituíam a única informação disponibilizada pelo *site* da DGS. Outros elementos anunciados (doenças transmissíveis, recursos do SNS e formação dos profissionais de Saúde) constituíam endereços não localizados Os dados relativos às Estatísticas de Mortalidade apresentavam relativa integridade temporal e actualidade (referindo-se a 1994 e 1997) A perspectiva abrangia vários aspectos nomeadamente relacionados com causas da morte (escolhido previamente pelo analista), região, idade e sexo.

Em Outubro de 2001, a produção estatística da Direcção Geral de Saúde envolvia os seguintes elementos:

- Estatísticas vitais de Natalidade, mortalidade infantil e perinatal relativas ao período 1994-1998;
- Estatísticas vitais de mortalidade geral e por algumas causas para os anos de 1994, 1996, 1997, 1998 e 1999;
- Doenças transmissíveis de declaração obrigatória nos períodos 1994-1998 e 1995-1999 existindo ainda dados provisórios para 2000.

Em termos de capacidade de processamento, era visível a mudança de formato dos dados disponibilizados de HTML para PDF, traduzindo-se, portanto, numa maior dificuldade.

3.3. A Matriz SIE / Factos Económicos

A organização de uma matriz que relacione os SIE com os factos económicos que são objecto da sua actividade facilita o processo de pesquisa de informação sobre a economia é facilitado. Consideremos os SIE do quadro seguinte.

Quadro 30. Os SIE considerados

Número	*Nome_Sie*	*Endereço*
1	Instituto Nacional de Estatística	www.ine.pt
2	Banco de Portugal	www.bportugal.pt
3	Departamento de Estatística do Trabalho e Formação	www.detefp.pt
4	Direcção Geral do Comércio e Concorrência	www.dgcc.pt
5	Direcção Geral da Energia	www.dge.pt
6	Direcção Geral da Saúde	www.dgsaude.pt
7	Direcção Geral do Turismo	www.dgturismo.pt
8	Direcção Geral do Orçamento	www.dgo.pt
9	Departamento de Prospectiva e Planeamento	www.dpp.pt
10	Gabinete de Estudos e Prospectiva da Economia	www.gepe.pt
11	Investimento, Comércio e Turismo de Portugal	www.icep.pt
12	Observatório do Comércio	www.obscommin-economia.pt
13	Direcção Geral de Estudos e Previsão	www.dgep.pt
14	Gabinete de Estudos e Planeamento do Min Justiça	www.gepmj.pt (1999) http://www.gplp.mj.pt (2001)
15	Direcção Geral dos Impostos	www.dgci.pt
16	Confederação do Comércio e Serviços de Portugal	www.ccp.pt

A *Matriz SIE/Factos* económicos assume o aspecto do quadro seguinte:

Quadro 31. Matriz SIE / Factos Económicos

Tema Económico	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*
Agricultura e Pesca	✓															
Balança de Pagamentos		✓														
Comércio e Concorrência	✓			✓												
Comércio Internacional	✓									✓	✓					
Condições de Vida das Famílias	✓															
Conjuntura	✓												✓			
Contabilidade Nacional	✓	✓						✓					✓			
Demografia	✓															
Emprego e Desemprego	✓		✓						✓	✓						
Energia	✓				✓											
Estatísticas Cambiais		✓														
Estatísticas Monetárias e Financeiras		✓														
Finanças Públicas		✓											✓			
Fiscalidade															✓	
Formação	✓		✓										✓			
Habitação, Construção e Obras Públicas	✓															
Indicadores de convergência									✓				✓			
Indicadores Macroeconómicos	✓								✓	✓			✓			
Indústria	✓															
Instituições Financeiras e Seguros	✓															
Investimento Externo		✓								✓	✓					
Justiça														✓		
Preços, Inflação	✓	✓		✓						✓						
Prospectiva / Cenários Macroeconómicois									✓	✓			✓			
Protecção Social	✓															
Salários	✓		✓							✓			✓			
Saúde	✓					✓										
Segredo Estatístico	✓													✓		
Transportes e Comunicações	✓															
Turismo / Restauração	✓						✓				✓					

4. Da heterogeneidade dos utilizadores à heterogeneidade da informação dos SIE

Poder-se-á pensar que a diversidade de utilizadores dos SIE conduziu a múltiplas formas de disponibilização da informação.

No entanto, com excepção de um caso particular de analistas, não detectámos em qualquer SIE uma formatação diferente da informação em função do utilizador ou mesmo a construção de repositórios de dados diferentes. O que se sente é que haverá produtos mais vocacionados para determinado tipo de utilizador. Apesar de tudo, e considerando a excepção referida, os SIE podem ser considerados um apoio muito importante aos processos de tratamento da informação e aos processos de tomada de decisão, designadamente:

- o Governo encontra nos SIE informação essencial à análise da conjuntura (sendo um dos seus principais alimentadores em termos de informação) e à definição e redefinição da política económica;
- as organizações buscam nos SIE e nas decisões governamentais elementos fundamentais para a decisão estratégica;
- é nas organizações e no Governo que se encontra a figura do analista da economia que introduzimos inicialmente e que pode assumir um estatuto mais passivo (análise e interpretação dos produtos disponibilizados, típico de níveis elevados do processo de decisão) ou mais activo (recepção de informação para processamento posterior e eventual interpretação ou envio de resultados a níveis superiores);
- um tipo especial de analistas, *os investigadores que desenvolvam a sua actividade de investigação no quadro de instituições credenciadas pelo Ministério da Ciência e da Tecnologia como instituições de investigação científica e desenvolvimento tecnológico*, terá acesso a bases de dados anonimizadas disponibilizadas pelo INE e suportadas por uma rede própria de acordo com um protocolo estabelecido entre o Ministério da Ciência e Tecnologia e o Instituto Nacional de Estatística em 1999;
- os estudantes encontram nos SIE um amplo campo de treino em termos de cálculo de indicadores e de simulação da tomada de decisão.

Em termos muito resumidos, podemos apontar vários produtos dos SIE que possibilitam ao Governo a tomada de decisão na área de Economia. Estão neste caso:

- as séries de dados relativas ao índice de preços no consumidor que permitem a determinação de taxa de inflação e a sua utilização como ponto de referência na negociação salarial (tanto no sector público como no privado), na revisão do valor de pensões, na actualização de prémios de seguros e como deflactor na contabilidade nacional;
- a informação sobre a evolução de salários e lucros determinantes da antecipação de decisões relativas ao consumo e investimento;
- os indicadores relativos ao emprego e desemprego utilizados como indicadores do potencial da produção corrente e da capacidade de trabalho inactiva, respectivamente;
- a medição da actividade económica ao nível do crescimento do PIB, da procura em termos de consumo, investimento e comércio externo e antecipada através de indicadores de confiança associados ao consumo, comércio e indústria;
- os resultados da comércio externo e das transferências de capitais manifestados ao nível da balança de pagamentos;
- a informação sobre finanças pública onde o saldo orçamental assume uma importância crucial, pondo constantemente a questão de como diminuir a despesa ou aumentar a receita e pondo o dedo sobre uma ferida que teima em permitir que alguns agentes económicos se furtem às obrigações fiscais e consigam, por essa via, transferir parte da carga fiscal que lhes competiria satisfazer para outros.

Estas categorias foram comentadas em pontos que precederam o actual. Consideremos, agora, a tomada de decisão a nível estratégico nas organizações. Este processo engloba um conjunto de acções de gestão que relaciona a organização com o seu ambiente, guia as actividades internas e determina o seu desempenho a longo prazo. Estas decisões, essenciais para a sua vivência e sobrevivência, envolvem uma quantidade significativa de recursos e respeitam a novas áreas de interesse ou a problemas pouco

comuns, tendo repercussões no modo como decisões de mais baixo nível serão tomadas.

A gestão estratégica das organizações desdobra-se por três processos fundamentais: (1) a formulação da estratégia, (2) a implementação da estratégia e (3) a avaliação e controlo da estratégia.

Na actividade de *formulação da estratégia*, os SIE intervêm em múltiplos níveis. Podemos questionar-nos sobre o seu papel na consumação de uma análise das forças e fraquezas, das oportunidades e ameaças, vulgarmente designada por *análise SWOT*, para a tomada de decisões estratégicas sobre o produto, para a análise do grupo competitivo e do consumidor e para suportar processos de previsão. Uma explicitação destas práticas, que seguiremos nos pontos seguintes, encontra-se em Stahl et al (1992). Mais completo, embora mais superficial, Rascão (2000) apresenta-nos um panorama de múltiplas concepções associadas à prática estratégica e propõe um sistema de informação para apoio à tomada de decisão estratégica onde é possível distinguir entre um sistema de informação do meio envolvente e um sistema de informação para a gestão. Os instrumentos que descrevemos fazem, naturalmente, parte do primeiro.

4.1. Importância dos SIE para a análise do ambiente

Esta análise, que tem por objecto o ambiente interno e externo onde se processa a prática organizacional, por nela se referirem as forças e fraquezas internas e as oportunidades e ameaça externas, é fundamental para a formulação da estratégia.

Ao nível das forças e fraquezas internas são examinados a gestão de topo e o quadro de directores, a posição financeira, as operações, o *marketing*, os recursos humanos, a área de pesquisa e desenvolvimento, os sistemas de informação e a estrutura organizacional e cultural da empresa.

Mas é ao nível da análise das oportunidades e ameaças externas que mais se detecta a necessidade dos SIE, pois, nesta matéria, as preocupações respeitam às tendências económicas, tecnológicas, regulamentares, físicas, sociais, demográficas e competitivas.

A *análise SWOT* carece de dados externos obtidos de uma forma sistemática e de um processo interno sólido de avaliação de recursos humanos e financeiros, para além do julgamento adequado sobre o desempenho do marketing, operações, sistema de informação e outras áreas.

A obtenção desses dados externos pode ser confiada a organizações cuja actividade se situa na área de produção de informação sobre negócios o que levanta a questão da sua fiabilidade, se se não pretender dispor de uma área interna com a missão da prospectiva.

O DPP e o GEPE serão os SIE que melhor se posicionam para um processo de apoio à análise do ambiente, tendo em conta os indicadores e os estudos macroeconómicos que disponibilizam.

No entanto, a publicação electrónica do Orçamento Geral do Estado pela DGO e as notas de conjuntura do DGEPE são mais um instrumento de que se não pode abstrair, pois disponibilizam todo um conjunto de orientações e informações macroeconómicas que formarão o pano de fundo da actividade dos agentes económicos.

Relevantes serão ainda os estudos e indicadores de confiança elaborados pelo INE.

4.2. Os SIE e as decisões estratégicas sobre o produto

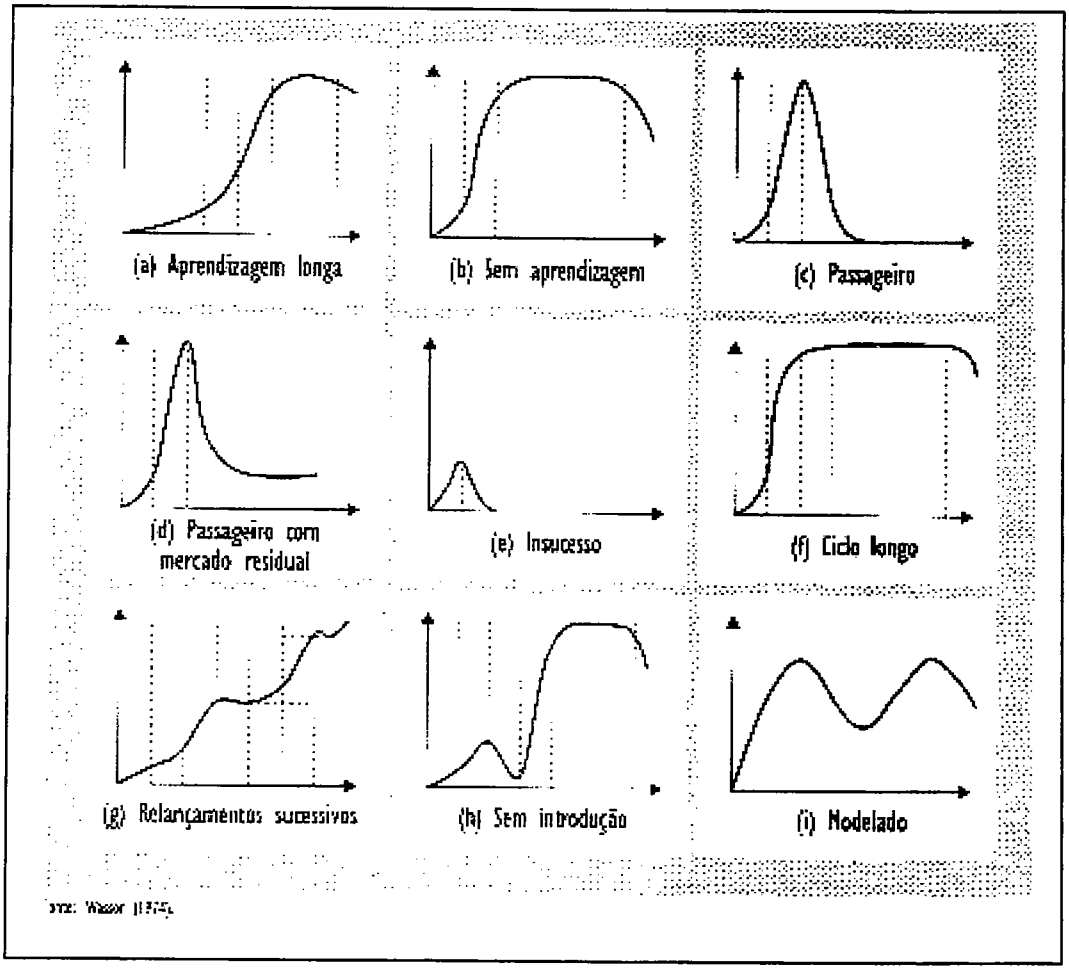
Entre as ferramentas mais importantes de apoio a todo este processo e, bem assim, ao de formulação das decisões estratégicas encontramos a análise do ciclo de vida do produto, a matriz BCG e a matriz GE. Esta referência não significa que elas esgotam as técnicas de que um analista dispõe ou que não existam outras que venham a gozar a sua preferência. É, no entanto, pacífico que são de conhecimento generalizado e, neste contexto, o facto de dependerem de dados dos SIE é determinante para as linhas que se seguem.

A análise do *ciclo de vida do produto* parte do exame no tempo da curva de vendas de determinado produto, procurando detectar nesta as fases de início, crescimento, maturidade e declínio através do seu declive.

Uma das suas consequências é a possibilidade de redefinição de estratégia para o produto antecipadamente relativamente aos momentos de declínio.

O conhecimento da fase do ciclo em que o produto se encontra é crucial para a definição da estratégia da empresa, pois permite ilações gerais sobre qual será a resposta de consumidores e concorrentes face à actuação da empresa no mercado. A figura seguinte apresenta formas diferentes da curva de ciclo de vida que permitem uma classificação do tipo de curva em função da ausência ou prolongamento de alguma fase, mostrando que esta ferramenta deve usar-se com extremo cuidado e que as conclusões não devem ser lineares.

Figura 29. Formas diferentes do ciclo de vida



Fonte: Lambis (2000)

A análise do ciclo de vida do produto carece de um registo fidedigno das vendas da organização num período mais ou menos longo. A transposição desses dados para uma base de dados e a sua referência ao tempo possibilitam a criação de séries que podem ser tratadas por *software* de produção de gráficos susceptíveis de, pelo seu exame, sugerirem o momento actual de cada produto. Os mesmos dados, complementados por métodos de previsão, permitirão antecipar tendências. A concretização desta análise parece, assim, apenas limitada pela disponibilidade de dados e de *software*

adequado. É importante fazer ressaltar, neste ponto, que muitos sistemas de informação estratégica podem ser implementados em folhas de cálculo susceptíveis de integrarem dados operacionais e este é um desses casos.

Encontrámos séries cronológicas sobre produtos no INE e no GEPE. Se esta informação crescer, poderá suportar análises do ciclo de vida do produto e possibilitar aos agentes económicos uma formulação fundamentada da sua estratégia.

A *matriz do Boston Consulting Group (BCG)* articula-se em torno de duas variáveis estratégicas: a taxa de crescimento do segmento de actividade analisado e a a quota de mercado relativa da empresa no referido segmento. A escolha da segunda variável faz referência à teoria do efeito de experiência. A quota de mercado relativa mede-se pelo *rácio*

Quota de mercado da empresa / Quota de mercado do concorrente principal (47)

Este *rácio* constitui, segundo o BCG, o indicador instantâneo da posição da empresa na escala dos custos e, portanto, da sua situação concorrencial. Quanto à taxa de crescimento, ela é, para o BCG, o factor essencial da dinâmica do segmento de actividade. Com efeito, só as actividades em crescimento conhecem baixas de custos importantes e permitem a criação de vantagens concorrenciais duradouras. Inversamente, nas actividades estáveis, os custos, tal como as quotas de mercado, estão fixos: os jogos estão feitos.

Estruturalmente, a matriz apresenta-se sob a forma de um quadrado, no qual a taxa de crescimento da actividade surge como ordenada e a quota de mercado relativa da empresa como abcissa.

A matriz BCG permite encontrar *estrelas, vacas leiteiras, dilemas* e *pesos mortos* classificações que lhe permitirão determinar a estratégia.

Tal como no caso anterior, esta análise carece do registo das transacções de uma organização repartido pelos diferentes produtos que comercializa. Por outro lado, carece de informação sobre o crescimento do mercado e sobre a quota de mercado relativa. Naturalmente, as tendências a nível de mercado provirão da análise SWOT pelo que temos aqui uma primeira articulação estabelecida ao nível dos elementos produtores de informação estratégica. A quota de mercado relativa também tem a sua fonte em dados externos

admitindo-se a possibilidade de a obter através da contabilidade nacional. Os SIE têm assim um alto valor informativo e viabilizador para esta ferramenta de apoio à decisão estratégica.

Resultados algo semelhantes são proporcionados pela *matriz da General Electric (GE)*, onde os critérios de decisão se baseiam (1) na atractibilidade da indústria, representada no eixo vertical, função de uma série de indicadores, incluindo naturalmente a taxa de crescimento do mercado que na BCG era o único indicador considerado e (2) na posição competitiva de determinado produto, representada no eixo horizontal, medida por vários indicadores incluindo a quota de mercado relativa. Cada eixo é repartido em três partes, formando assim nove sectores distintos. Cada negócio será representado por um círculo, cujo raio é proporcional ao peso que as vendas desta unidade têm na facturação total da organização e dentro deste círculo representa-se a fatia que corresponde à quota de mercado da unidade.

A atractibilidade do mercado é medida através de uma série de indicadores como a dimensão do mercado, a taxa de crescimento, as margens que se praticam, o nível de rivalidade entre os concorrentes, os requisitos tecnológicos para competir.

Por outro lado, a posição competitiva mede-se através da quota de mercado, da imagem de marca de que a unidade desfruta, do seu sistema de distribuição.

As decisões estratégicas que se podem tirar desta matriz baseiam-se no nível de investimento a concretizar em determinados mercados.

Relativamente a este tipo de decisões, o ICEP e o INE posicionam-se como os SIE que apresentarão elementos de maior interesse para os agentes económicos fundamentarem as suas decisões e estamos concretamente a lembrar-nos de um produto do primeiro: as fichas técnicas sobre países e mercados externos.

4.3. As dificuldades da análise do grupo competitivo

Através de uma *análise do grupo competitivo* uma organização pode relacionar o seu desempenho com organizações concorrentes. Nesta análise, há que seguir um conjunto de passos sistemático: (1) identificar as dimensões de

concorrência importantes, (2) elaborar gráficos bidimensionais das dimensões competitivas para a empresa e seus concorrentes e (3) analisar a posição da firma relativamente aos concorrentes e às necessidades dos clientes.

Trata-se de uma análise que carecerá do recurso a organizações de consultoria em virtude de necessitar de dados sobre a concorrência que possivelmente não estão disponíveis. O recurso aos SIE, eventualmente, terá pouco sucesso embora o contrário fosse desejável em termos de transparência. Sem estes dados, qualquer análise será subjectiva. Caso a informação exista, pode ser guardada em repositório adequado a partir do qual se poderão fazer comparações estáticas e comparações dinâmicas.

4.4. A viabilização da análise do consumidor

Através da *análise do consumidor*, uma organização procura estudar em profundidade as entidades que compõem a procura num determinado mercado.

Geralmente, essas entidades apresentam características muito diversificadas que tiram eficácia à utilização de uma mesma abordagem indiferenciada a todas elas. Por isso, é usual tentar definir grupos de consumidores com características relativamente homogéneas e de tal forma que o comportamento dentro de cada grupo seja substancialmente mais idêntico que de grupo para grupo (a homogeneidade intragrupal é superior à intergrupal).

A este processo de fragmentação do total de consumidores em grupos homogéneos dá-se o nome de *segmentação do mercado*, e a cada grupo assim definido o de segmento de mercado.

A segmentação do mercado é feita com base nas características ou critérios que permitem definir grupos homogéneos e a sua aplicabilidade a cada mercado é muito variável. É preciso ter a capacidade, e a argúcia, para definir operacionalmente os segmentos de mercado partindo dos critérios correctos para o mercado concreto em análise. Alguns dos critérios de segmentação de mercados mais utilizados são o geográfico, os tipos de produtos e as características do consumidor (rendimento, sexo, idade,...).

A informação demográfica disponibilizada pelo INE poderá funcionar como um complemento fundamental ou prévio a este tipo de análise. Não podemos

esquecer que o disponibilizado por aquele SIE é sujeito a um processo de agregação, inviabilizando uma verdadeira segmentação a qual carece de dados detalhados. No entanto, antes de uma organização enveredar pela realização de questionários e processos de recolha de dados sempre onerosos, poderá analisar determinadas regiões e mercados em função do que aquele SIE disponibiliza e procurar criar indirectamente alguns aspectos da segmentação. No mesmo sentido podemos considerar o que é disponibilizado pelo ICEP já referido anteriormente.

5. A questão da disponibilização de informação incompleta em Economia

No início deste capítulo, esboçámos uma clarificação do que designámos por *disponibilização de informação incompleta.* Estamos agora em condições de procurar estabelecer um indicador relativo a esta questão. Tal não significa que estejamos a referir um fenómeno que deriva da vontade e consciência dos que têm por função a gestão dos SIE. No entanto, a sua existência objectiva obriga a uma palavra que pode servir de alerta para uma actuação correctora: no passado, a disponibilização de informação utilizava como suporte dominantemente o papel, por vezes os apuramentos baseavam-se em amostras. Hoje, a trivialidade e o baixo custo da disponibilização interactiva, as novas formas de recolha e os algoritmos de mineração de grandes volumes de dados justificam que este seja um aspecto estratégico a repensar de forma a condicionar a manutenção dos sítios e a permitir que os analistas executem o seu processamento analítico.

Admitamos que determinado facto económico pode ser analisado em termos de x dimensões independentes. Consideramos *dimensão independente* aquela que interioriza todas as possíveis relações hierárquicas em que intervenham elementos da dimensão. Por exemplo, existindo a dimensão *Região* ela será independente de outra se o elemento *País* existir exclusivamente nela.

Considerando as x dimensões referidas existem 2^x-1 maneiras diferentes de visualizar os dados. Por exemplo, consideremos o investimento directo estrangeiro e as dimensões Região de Origem, Actividade Económica, Operação e Tempo. É possível visualizar os dados por cada dimensão isoladamente, pela sua combinação duas a duas, pela sua combinação três a

três e pela sua combinação quatro a quatro, isto é o número de listas possível será $C^4_1 + C^4_2 + C^4_3 + C^4_4 = 4 + 6 + 4 + 1 = 2^4 - 1 = 15$.

Admitamos que um SIE disponibiliza relativamente ao mesmo facto *y* vistas elementares. Se estas vistas não forem elementares, é possível proceder ao seu desdobramento e determinar quantas é possível formar. Mantenhamos que y representa esse número.

Uma medida para a *disponibilização de informação incompleta* (DII) pode ser dada por:

$$DII = 1 - y / (2^x - 1) \qquad (48)$$

Voltando ao nosso exemplo, se o SIE se limitar a disponibilizar dados por Região de Origem ou Tipo de Operação ou Actividade Económica ou Tempo, estará a proporcionar quatro modos diferentes de ver os dados em 15 possíveis. Virá $DII = 1 - 4 / 15 = 11/15$: *a informação fica lá quase toda.*

Admitamos, em contrapartida, que o SIE, optando pela simplicidade, coloca no sítio um ficheiro em formato texto com quatro colunas relativamente à intersecção de cada uma daquelas dimensão e uma quinta coluna com o valor do IDE que lhe corresponde. Nestas condições, o analista, dispondo de uma ferramenta para o processamento, poderá chegar aos apuramentos que entender. Isto é, DII assumirá o valor *0*.

6. Avaliação dos sítios da Administração Directa e Indirecta e sua relação com a maturidade dos SIE

A Resolução do Conselho de Ministros nº 16/96 de 21/3/96 criou a Missão para a Sociedade da Informação (MSI) e atribuiu-lhe um mandato para *promover um amplo debate nacional sobre o tema sociedade da informação, tendo em vista a elaboração de um Livro Verde que ... contenha propostas de medidas a curto, médio e longo prazos, a ser presente à Assembleia da República.*

Um dos primeiros resultados da actividade desta MSI foi o Livro Verde para a Sociedade da Informação em Portugal cujo objectivo essencial foi o estabelecimento de uma orientação estratégica para o desenvolvimento da *sociedade da informação*, designação cujo sentido é clarificado nos termos seguintes:

> *A expressão 'Sociedade da Informação' refere-se a um modo de desenvolvimento social e económico em que a aquisição, armazenamento, processamento, valorização, transmissão, distribuição e disseminação de informação conducente à criação de conhecimento e à satisfação das necessidades dos cidadãos e das empresas, desempenham um papel central na actividade económica, na criação de riqueza, na definição da qualidade de vida dos cidadãos e das suas práticas culturais. A sociedade da informação corresponde, por conseguinte, a uma sociedade cujo funcionamento recorre crescentemente a redes digitais de informação. Esta alteração do domínio da actividade económica e dos factores determinantes do bem-estar social é resultante do desenvolvimento das novas tecnologias da informação, do audiovisual e das comunicações, com as suas importantes ramificações e impactos no trabalho, na educação, na ciência, na saúde, no lazer, nos transportes e no ambiente, entre outras.* (MSI, 1997)

O Presidente da Missão para a Sociedade da Informação, o Prof. Doutor Dias Coelho, reconhece, no prefácio àquele documento, que *"a Sociedade da Informação pode contribuir para a melhoria do bem-estar dos cidadãos, em virtude de facilitar a construção de um Estado mais aberto, a inovação no ensino e na formação profissional, o acesso ao saber, o desenvolvimento de novas actividades económicas e o aumento da oferta de emprego com níveis de qualificação profissionais mais elevados, entre outros contributos positivos"*. Tal facto não esconde, no entanto, *"a existência de barreiras de acesso à Sociedade da Informação, de natureza económica, educacional e cultural, assim como os riscos de que importantes camadas da população fiquem excluídas dos seus benefícios, em consequência do fenómeno da info-exclusão"*.

No mesmo sentido vai a Introdução a MSI (1997), da responsabilidade de Sua Excelência o Ministro da Ciência e Tecnologia, o Dr. José Mariano Gago:

> *"Não somos tecnicistas. Ao tomarmos como nosso esse lema geral, sedutor e aparentemente neutro da Sociedade da Informação, retirámos-lhe a falsa neutralidade e tomámos, antes de mais, partido pela* ***cidadania****, contra a exclusão; pelo* ***conhecimento****, contra a manipulação do espírito; pela* ***liberdade****, contra a opressão, especialmente contra a opressão confortada tecnicamente; pela* ***inovação*** *contra os monopólios."*

Nestas condições, é natural que MSI (1997) apresente uma estrutura que, aposta na generalização da formação e da utilização das tecnologias de informação.

Quadro 32. A implementação da sociedade da informação em Portugal

1. a democraticidade da sociedade da informação	1.1 a sociedade da informação e a democracia 1.2 acesso à sociedade da informação 1.3 o combate à iniquidade 1.4 responsabilidade social para os que requerem consideração especial 1.5 medidas
2. o estado aberto	2.1 melhorar a eficiência da administração pública 2.2 um enquadramento incentivador da informatização da administração pública 2.3 acesso digital do cidadão e das empresas à informação pública 2.4 rede electrónica de interligação dos organismos públicos 2.5 arquivos electrónicos 2.6 divulgação da informação ao cidadão e às empresas 2.7 democracia electrónica 2.8 medidas
3. o saber disponível	3.1 rede electrónica de investigação científica, cultura e educação 3.2 desenvolvimento de bibliotecas digitais 3.3 digitalização do património cultural 3.4 difusão do património cultural, da língua portuguesa e a diáspora lusitana 3.5 medidas
4. a escola informada: aprender na sociedade da informação	4.1. objectivos e desafios da escola informada 4.2. dinamização estratégica 4.3 equipar os estabelecimentos escolares 4.4 qualificação do professor para a sociedade da informação 4.5 rede de serviços e comunidades educacionais 4.6 investigação e avaliação de impactos 4.7 medidas
5. a empresa na sociedade da informação	5.1 o sector empresarial da indústria da informação como sector estratégico de desenvolvimento 5.2 a competitividade das empresas no ambiente global da sociedade da informação 5.3 as indústrias tradicionais e a sociedade da informação 5.4 a sociedade da informação e a reinvenção da organização do trabalho 5.5 adequação da criação de empresas à sociedade da informação (empresas virtuais) 5.6 teletrabalho 5.7 comércio electrónico 5.8 medidas
6. o emprego na sociedade da informação	6.1 a sociedade da informação como paradigma de uma sociedade intelectualmente criativa 6.2 o mercado de trabalho na sociedade da informação 6.3 contributo para novos enquadramentos do trabalho na sociedade da informação 6.4 aprendizagem e formação profissional em tecnologias da informação 6.5 aumento da posição competitiva das pme's portuguesas 6.6 medidas

Quadro 32. A implementação da sociedade da informação em Portugal *(cont.)*

7. o mercado e a indústria da informação	7.1 a convergência das tecnologias da informação, telecomunicações e audiovisual 7.2 indústria dos conteúdos 7.3 indústria do software 7.4 indústria electrónica de suporte à sociedade da informação 7.5 indústria do audiovisual e do entretenimento 7.6 indústria das telecomunicações 7.7 medidas
8. implicações sociais da sociedade da informação	8.1 a melhoria da qualidade de vida e bem-estar social dos cidadãos 8.2 info-alfabetização e info-exclusão 8.3 privacidade e protecção dos direitos individuais 8.4 protecção dos menores 8.5 apoio a grupos socialmente desfavorecidos 8.6 medidas
9. implicações jurídicas da sociedade da informação	9.1 protecção da privacidade e dos dados individuais, das empresas e das instituições 9.2 notariado electrónico 9.3 documentos e transacções electrónicas 9.4 protecção dos direitos de propriedade intelectual e direitos de autor 9.5 medidas
10. infraestrutura nacional de informação	10.1 uma eficiente e acessível infraestrutura nacional de informação 10.2 a liberalização do sector das telecomunicações 10.3 acessibilidade na sociedade da informação 10.4 encorajar novos serviços e aplicações 10.5 privacidade e segurança da informação 10.6 medidas
11. a investigação e desenvolvimento na sociedade da informação	11.1 a i&d no contexto da sociedade da informação 11.2 um programa nacional de i&d de suporte ao desenvolvimento da sociedade da informação 11.3 a interacção com programas de âmbito internacional 11.4 medidas

Fonte: Adaptado de MSI (1997)

Uma aposta essencial da MSI foi a melhoria da relação do cidadão com a Administração Pública que passa por documentos, formulários e publicações diversas às quais, no passado, se associavam filas de espera no mínimo desagradáveis. As potencialidades dos sistemas de informação foram consideradas como detendo larga capacidade de aplicação nesta área e não surpreende o surgimento da já citada Resolução do Conselho de Ministros n.º 95/99 publicada no DR n.º 198, I Série B, de 25 de Agosto de 1999.

Quadro 33. Os termos da declaração para a disponibilização interactiva

Nos termos da alínea g) do artigo 199º da Constituição, o Conselho de Ministros resolve:

1.

1.1 As direcções gerais e serviços equiparados, bem como os institutos públicos nas suas diversas modalidades deverão disponibilizar ao público, em formato digital na Internet, as respectivas publicações, tenham ou não carácter periódico, os formulários que utilizem e ainda, genericamente, toda a informação que produzam e que seja objecto de publicação.

1.2 A regra referida no ponto anterior aplica-se às publicações e informações publicadas após a entrada em vigor do presente diploma, bem como aos formulários em uso após essa data.

1.3 A disponibilização a que se refere o ponto 1.1 pode ser feita a título gratuito ou oneroso.

1.4 A divulgação na Internet feita de acordo com o estabelecido no presente diploma, far-se-á com respeito pelos direitos de propriedade intelectual.

2.

Os organismos referidos no número anterior deverão ainda elaborar planos calendarizados de disponibilização de outros conteúdos de interesse público na Internet, que poderão abranger, não apenas textos, mas também imagens e sons, devendo submetê-los à respectiva tutela.

3.

3.1 As modalidades de aplicação do disposto no número 1 poderão ser objecto de concretização e faseamento pela tutela dos organismos nele referidos, devendo, em todo o caso, ser dada prioridade à divulgação da informação de maior interesse público.

3.2 As modalidades referidas no ponto anterior poderão excepcionar justificadamente a publicação na Internet de determinadas obras ou grupos de obras.

4.

4.1 Toda a informação disponibilizada na Internet pelas entidades referidas no número 1 deverá ser apresentada de forma clara e de molde a ser facilmente acedida e consultada devendo, nomeadamente, ser implementados mecanismos de pesquisa de fácil utilização.

4.2 Na disponibilização da informação a que se refere o ponto anterior deverão ser levadas em linha de conta as exigências específicas dos cidadãos com necessidades especiais, por forma a que lhes seja garantido o acesso efectivo à informação facultada em suporte digital na Internet.

5.

5.1 Toda a informação disponibilizada na Internet pelas entidades referidas no número 1 pode ser livremente utilizada pelo público que a ela acede, desde que se faça menção da respectiva fonte.

5.2 Exceptua-se do disposto no ponto anterior a informação que por força de instrumento juridicamente vinculativo não possa ser livremente utilizada devendo, nesse caso, disso fazer-se menção expressa.

6.

O Ministro responsável pela Administração Pública dinamizará o processo de concretização do disposto no presente diploma, sensibilizando nesse sentido os serviços e organismos por este abrangidos e acompanhará a aplicação do mesmo, devendo propor a adopção de medidas complementares que esse acompanhamento recomende.

7.

O Ministro da Ciência e da Tecnologia promoverá o acompanhamento e avaliação da execução do presente diploma, informando regularmente o Governo sobre a sua aplicação.

8.

Os organismos abrangidos pelo presente diploma deverão reportar às respectivas tutelas, no prazo máximo de seis meses, as medidas tomadas no sentido de dar cumprimento ao estabelecido no presente diploma.

A MSI esgotou, entretanto, a sua actividade e surgiu uma Comissão Interministerial para a Sociedade da Informação de cuja actividade resultou um documento relativo às boas práticas na construção de sítios nas Administração Pública cujo resumo apresentamos em *Anexos*.

A existência deste guia, SECISI (2001), e a disposição relativa à generalização da disponibilização interactiva justificam a Resolução do Conselho de Ministros N° 22/2001 que determina a *avaliação periódica dos Web Sites dos organismos da Administração Directa e Indirecta do Estado* (CM, 2001), visando nomeadamente aferir o grau de actualização da informação disponibilizada, a clareza da forma como é apresentada, a facilidade de pesquisa da informação e o cumprimento das disposições legais relativas ao conteúdo e forma de apresentação das páginas dos organismos públicos, designadamente a sua acessibilidade por cidadãos com necessidades especiais.

O objecto desta avaliação é a produção de recomendações tendentes ao aperfeiçoamento e melhoria da qualidade das páginas avaliadas, a divulgação das boas práticas e uma avaliação da maturidade dos sítios de acordo com o grau de interactividade, segundo o modelo do eEurope que assenta em quatro estádios: informação, interacção, interacção bi-direccional e transacção. Os critérios utilizados e os indicadores que lhe são associados, apresentados em CISI (2001), são resumidos nos anexos da dissertação.

Em Fevereiro de 2002, surgiu o primeiro relatório de avaliação dos organismos de Administração Directa e Indirecta, CISI (2002) que considerou treze dos dezasseis SIE objecto desta dissertação.

A partir do quadro resumo que apresentamos na página seguinte, poder-se-ia pensar que quase tudo estaria bem no quadro dos SIE. Com efeito quase 90% tinham sido tidos em conta na avaliação e, no que respeita a esta, apenas dois em seis critérios estariam longe do bom e do muito bom. No entanto, é importante não esquecer o seguinte:

1. O guia das boas práticas relativo à construção de sítios na Administração Pública não coincide com o guia das boas práticas desenvolvido pelas Nações Unidas para os SIE apresentado em 1.1.1. deste capítulo.

2. Em termos de SIE, apesar de alguns aspectos tecnológicos serem importantes o que é relevante é a informação disponibilizada em termos de conteúdo e características. Estas têm a ver com a actualidade, a multidimensionalidade, a facilidade de processamento posterior, etc.

Quadro 34. Resumo da avaliação dos sítios da Administração Pública

	Conteúdos	Actualização dos Conteúdos	Acessibilidade	Navegabilidade	Facilidades para Cidadãos com Necessidades Especiais	Serviços Online
INE	B	B	MB	Ex	I	R
BP	nd					
DETFP	R	R	Ex	MB	I	R
DGCC	B	B	MB	Ex	I	B
DGE	B	R	MB	Ex	I	I
DGS	B	B	B	B	I	I
DGT	R	I	MB	MB	I	I
DGO	B	MB	MB	Ex	Ex	R
DPP	B	R	Ex	Ex	B	R
GEPE	B	R	MB	MB	I	R
ICEP	B	MB	MB	MB	I	I
OBC	MB	R	B	Ex	I	R
DGEP	B	MB	MB	Ex	R	B
GPMJ	nd					
DGI	R	B	MB	MB	I	R
CCP	nd					
Média	<B	<B	MB	<Ex, >MB	<R	R

Legenda:
nd – não disponível, I – Insuficiente, R – Regular, B – Bom, MB – Muito Bom, Ex – Excelente

Fonte: Adaptado de CISI (2002)

Se não for tida em conta a especificidade dos SIE, a sua avaliação nunca ultrapassará o problema da gestão tecnológica e a sua maturidade mesmo em termos do eEurope é uma maturidade falseada. Assim, é preciso ir mais longe em termos de critérios de análise. Esse é o objectivo dos capítulos V e VI desta dissertação, marcando o seu conteúdo a grande diferença relativamente ao trabalho da MSI e da CISI com o qual tem naturalmente algumas afinidades.

Resumo do Capítulo III

Neste capítulo procedemos a um breve exame das organizações responsáveis pela informação para a Economia em Portugal em termos dos seus objectivos e de alguns produtos. Estas organizações constituem o objecto de pesquisa desta dissertação.

Começámos por examinar a disponibilização e utilização de informação para a Economia como um processo em que a figura do analista ocupa agora um lugar essencial. Ele é um cliente do SIE que deverá organizar-se para lhe proporcionar o melhor serviço. Este princípio permitiu-nos formular uma visão do processo de disponibilização baseado nas necessidades do analista determinadas pelos seus projectos de investigação que consultará os SIE, avaliará a informação disponibilizada, fará pedidos, sugestões e definirá requisitos, podendo extrair dados e proceder aos seus próprios apuramentos. Tal perspectiva exige da parte do SIE uma atitude de procura de satisfação de necessidades em informação e da sua disponibilização em formato trabalhável. A breve passagem que fizemos pelos SIE permitiu-nos determinar que, apesar de existir muito informação disponibilizada, estamos longe de poder considerar satisfatória a mesma em termos do que afirmámos anteriormente.

Em Portugal, existem duas organizações com uma vocação de disponibilização alargada da informação: o INE e o Banco de Portugal. Esta vocação começa a manifestar-se em outras organizações que no nosso documento associámos a questões de prospectiva como o DPP e o GEPE, mas a imensa maioria dos SIE presta informação de âmbito restrito, isto é informação associada a um conjunto limitado de factos económicos.

O apoio dos SIE relativamente ao processo de tomada de decisão estratégica nas organizações foi equacionado relativamente a quatro elementos essenciais: análise do ambiente, decisões estratégicas sobre o produto, análise do grupo competitivo e viabilização da análise do consumidor.

É aceitável que a informação dos SIE pode apoiar alguns daqueles processos, devendo os seus dados integrar-se como dados externos no sistema de informação das organizações as quais, no entanto, não deixarão de necessitar de outra informação para a viabilizar. Mas esta possibilidade de integração de dados nos SIE traz consigo desde logo o problema das suas características em termos de formato, de multidimensionalidade, de integridade temporal e de múltiplos outros aspectos que retomaremos em capítulos posteriores.

Admitindo a eventualidade de a informação ser proporcionada, mas não o ser com o nível de detalhe necessário, apresentámos um indicador que pode ser

utilizado para medir o grau da informação retida ou não apurável pelos SIE conceito que designámos por *disponibilização de informação incompleta.*

Finalmente, fizemos uma referência à avaliação dos organismos da avaliação directa e indirecta do Estado por a mesma poder confundir-se com os objectivos deste trabalho. Determinámos que a grande maioria dos SIE consta na lista dos organismos avaliados. No entanto, deixámos claro que os objectivos da avaliação são diferentes. A avaliação decorrente da CISI é válida para qualquer sítio da Web, inclusivamente de uma empresa privada. Ela é extremamente genérica, perdendo a especificidade do que é relevante em termos de informação para a Economia onde a actualidade da informação, a multidimensionalidade em termos de dimensões genéricas e temporais e o processamento posterior se revelam determinantes. Ela não segue as recomendações formuladas para os SIE de que fizemos eco no capítulo. Há todo o perigo de um SIE ter uma excelente avaliação como sítio e ser escondida a sua insuficiência em termos de apoio ao analista da economia.

Capítulo IV – Aplicações de Processamento Analítico suportadas por Sistemas de Informação para a Economia em Portugal

Neste capítulo, vamos proceder à divulgação dos resultados de uma experiência em torno da informação proporcionada pelos SIE. Trata-se, no fundo, de determinar até que ponto ela viabiliza o processamento analítico ou é o resultado de processos analíticos susceptíveis de serem desencadeados interactivamente pelo analista.

Procurámos um conjunto de processos analíticos que ultrapassa muito os que descrevemos no Capítulo II. Num primeiro momento, era nosso objectivo tudo o que poderia caracterizar-se como descoberta de conhecimento em bases de dados e que passaria, para além dos processos descritos, pela utilização de redes neuronais, algoritmos genéticos ou redes causais. A indisponibilidade de tais resultados fez-nos reduzir o âmbito da dissertação aos processo de interrogação, análise multidimensional e análise de afinidades e utilizar os de classificação e pesquisa de *clusters* nos próximos capítulos.

Mas essa redução permitiu, até pela necessidade de uma pesquisa mais detalhada, constatar a existência de processos que combinam características dos descritos. Isto é, os processos podem não se concretizar na sua forma pura mas em termos híbridos o que nos permitiu contactar com mais informação disponibilizada pelo sistema estatístico nacional.

Neste capítulo, é, também, nossa preocupação determinar se existe, nos SIE, sem ser disponibilizada interactivamente, informação que viabilize a utilização de processos analíticos pelo que descrevemos aqui resultados de contactos para a sua obtenção e posterior processamento.

Numa primeira secção do capítulo, descrevemos um processo de interrogação em torno de indicadores disponibilizados pelo INE que, pela sua eficácia, terá com certeza grande utilidade para comparações estáticas.

A ilustração de processos multidimensionais é efectuada com base num trabalho do autor apresentado ao seminário do departamento de Economia do ISEG em 2000 e com um percurso sobre a base de dados de comércio internacional (BDCI) da responsabilidade do GEPE.

Na terceira secção, apresentamos dois casos em que as facilidades de *drill-down* são integradas em processos que, sem elas, teriam de ser considerados processos de interrogação.

Na quarta secção, relatamos um processo de descoberta de conhecimento em bases de dados, viabilizado por informação do ICEP: a pesquisa de associações no âmbito do Investimento Directo Estrangeiro.

Terminamos com uma reflexão sobre a viabilidade do processamento analítico operado e proporcionado pelos SIE.

1. Processos de interrogação em informação demográfica e social

O INE disponibiliza uma extensa lista de indicadores nas áreas demográfica e social. Dessa lista extraímos alguns para ilustração deste processo, constituindo-se a própria documentação consultável na sua página electrónica como a essencial para uma melhor compreensão. Referimo-nos ao subsistema designado por *Metainformação,* responsável pelos conceitos apresentados nas linhas seguintes.

O indicador *Excedente de Vidas* corresponde à diferença entre o número de nados-vivos e o número de óbitos, num dado período de tempo.

Outro indicador demográfico que utilizaremos será o *Índice de Envelhecimento* que, de acordo com o INE, corresponde à relação existente entre o número de idosos e o de jovens, definido habitualmente como a relação entre a população com 65 e mais anos e a população com 0 - 14 anos.

Por *Taxa de Divórcio* entenderemos o número de divórcios ocorridos durante um certo período de tempo, normalmente o ano, referido à população média desse período.

A *Taxa de Mortalidade* consiste no número de óbitos ocorridos durante um certo período de tempo, normalmente o ano, referido à população média desse período.

A *Taxa de Natalidade* é definida pelo número de nados-vivos ocorrido durante um certo período de tempo, normalmente o ano, referido à população média desse período.

A *Taxa de Nupcialidade* corresponde ao número de casamentos ocorrido durante um certo período de tempo, normalmente o ano, referido à população média desse período

Habitualmente os indicadores demográficos são expressos em termos de permilagem do número de habitantes, sentido que deve ser precisado para população média do período.

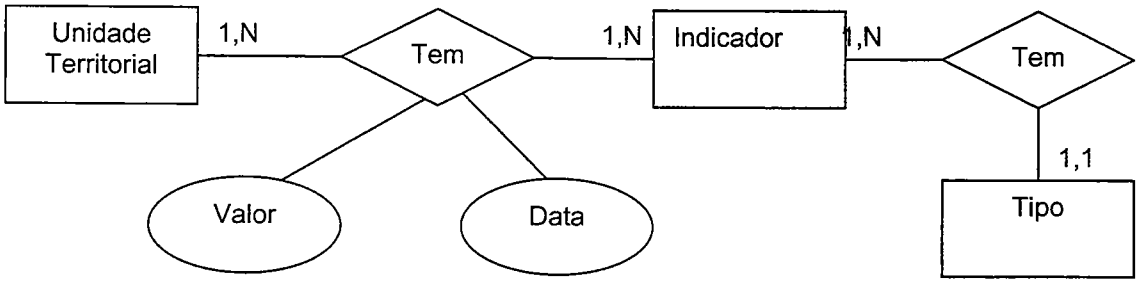
Igualmente importantes serão os indicadores relativos a condições sociais dos quais seleccionámos quatro: (1) Camas Hospitalares por 1000 Habitantes, (2) Índice *Per Capita* do Poder de Compra (Portugal = 100), (3) Médicos por 1000 Habitantes e (4) Pensão Média Anual por Pensionista Activo. A designação destes indicadores é suficiente para se compreender o seu conteúdo.

Nesta primeira apresentação de uma aplicação de processamento analítico suportada por um processo interrogacional vamos abordar especificamente os tópicos de modelação, processo analíticos suportados pela informação extraída e aspectos criticáveis no sistema disponibilizado.

1.1. Modelação

Tratando-se de um processo de interrogação, a sua modelação é legitimamente explicitada por um *Diagrama Entidade Associação* sendo opção do SIE o que se refere a modelo lógico.

Figura 30. Modelação da informação demográfica e social



De acordo com o diagrama apresentado, podemos afirmar que existem vários indicadores para uma mesma unidade territorial.

Alternativamente um indicador encontra expressão em várias unidades territoriais. Essa expressão assume a forma de um valor e uma data. Assim, no sistema existente no INE uma conjunção de um indicador e uma unidade

territorial materializa-se num valor que respeita a uma data. Isto afasta a perspectiva temporal do sistema.

Por outro lado, o diagrama mostra a existência da vários indicadores relacionados com um mesmo tipo de indicador.

A partir desta criação, conhece-se a estrutura da informação que, no INE, satisfaz o tópico referido. Em termos lógicos, há várias alternativas que, para já, nos dispensamos de comentar.

1.2. Processos analíticos suportados

A utilização de processos analíticos com base no sistema descrito assenta numa metodologia que envolve os seguintes passos: (1) extracção, (2) transferência, (3) transformação e (4) utilização.

1.2.1. A extracção

Admitindo uma implementação relacional, a extracção de informação de um tal sistema exigiria a união do resultado de tantos comandos de extracção quantos os indicadores, obedecendo à estrutura:

SELECT IND# From Ind_UNT WHERE UNT = "U1" OR UNT = "U2"...OR UNT="Un"

onde Ind_UNT seria a tabela de residência original dos dados.

É fácil detectar o incómodo a nível de escrita da cláusula *or* e a nível de duplicação do comando para todos os indicadores.

O INE resolve este problema através de um sistema de selecção visual simultânea de unidades territoriais e indicadores introduzindo um grande dinamismo no processo.

Para melhorar esse desempenho, é possível a selecção utilizando diferentes modalidades de organização das unidades territoriais, designadamente as designadas por NUTs e Distrito. Além disso, a escolha do tipo do indicador permite visualizar, na mesma página, os indicadores de vários tipos.

Em termos de experimentação, escolhemos regiões intermédias no sistema NUTs e procurámos indicadores demográficos e associados a condições sociais.

1.2.2. Transferência

Na posse da informação disponibilizada de acordo com as condições de selecção introduzidas, houve que proceder à sua passagem para a máquina de trabalho do analista.

Apesar de o sistema desenhado pelo INE admitir uma transferência a pedido, no momento de utilização esta não se encontrava activa, pelo que houve que seleccionar os dados no monitor e copiá-los para o novo ambiente através do processo normal de Windows que podemos designar por *COPY → PASTE.*

Contrariamente a outros produtos, esta passagem não ofereceu qualquer problema, pelo que os dados foram introduzidos em folha de cálculo.

1.2.3. Transformação dos dados extraídos

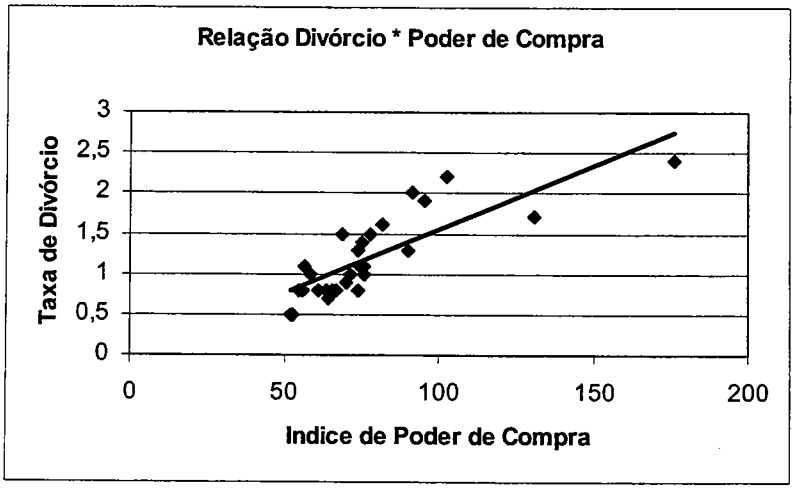
Não conseguimos utilizar os dados na forma em que foram extraídos por uma questão relacionada com o ponto decimal. O INE envia ".", a implementação de Excel que dispomos utiliza ",". Assim, tornou-se necessário proceder a essa tranformação o que foi operado através do editor de *Replace.*

1.2.4. Utilização dos dados

Os dados disponibilizados pelo INE, nesta componente, podem apoiar diversos processos analíticos, designadamente os que se apoiam em Estatística. Por exemplo, é possível determinar se existe alguma associação entre o *poder de compra* e a *taxa de divórcio* para as diferentes regiões.

A utilização das potencialidades da folha de cálculo em termos de determinação de uma recta de regressão e da qualidade do ajustamento, permite-nos chegar a um modelo que serve como um indicador de uma possível resposta para este problema. É o que podemos ver na figura seguinte.

Figura 31. A associação poder de compra vs taxa de divórcio



Efectivamente, o gráfico sugere alguma associação entre as variáveis em causa, leva a suspeitar da presença de heterocedasticidade e é possível complementar a sua informação com o ajustamento de uma recta e um indicador de qualidade. Sendo *y* a *taxa de divórcio* e *x* o *índice de poder de compra*, constata-se que há uma tendência para a taxa de divórcio subir com a subida da outra variável o que é expresso pela equação

$$y = 0{,}0159x - 0{,}0371 \qquad (49)$$

sendo determinado um $R^2 = 0{,}6633$ que nos leva a acreditar numa qualidade satisfatória.

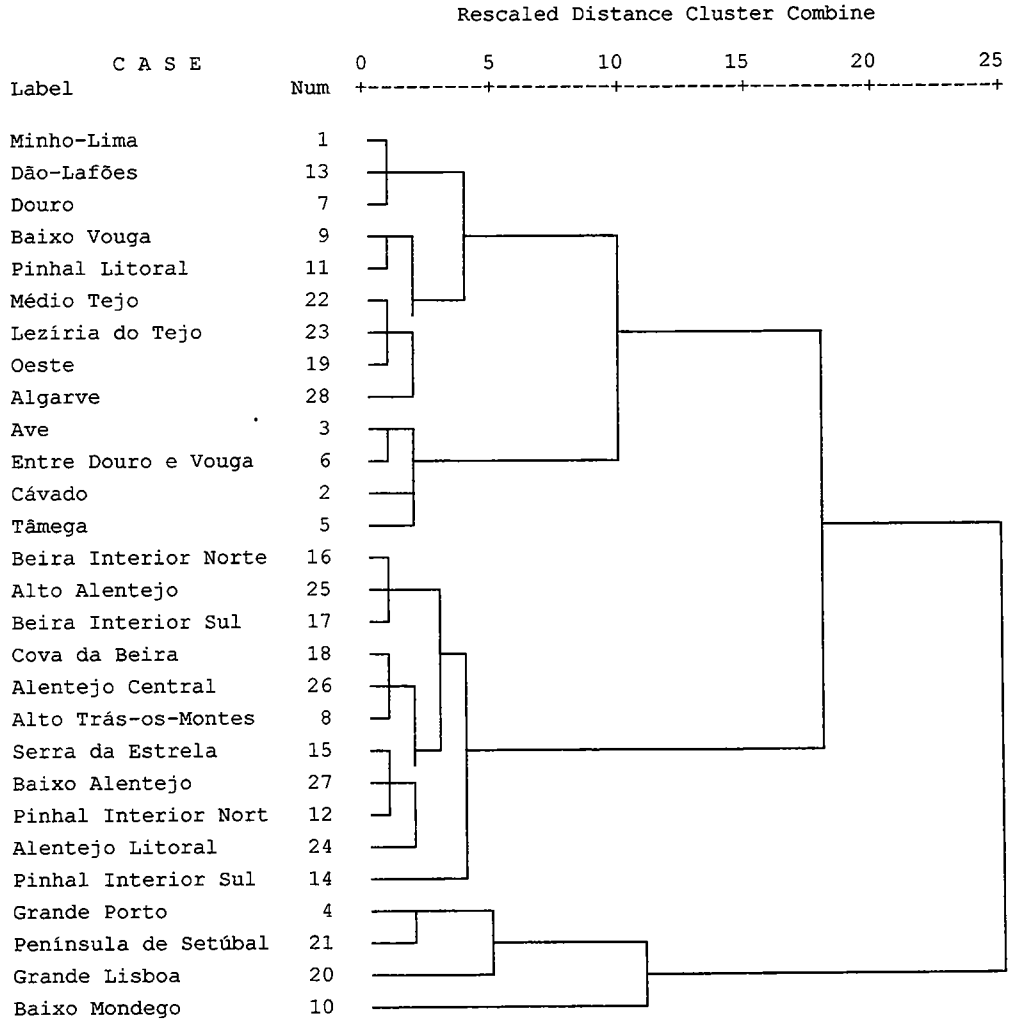
Com os dados transferidos, outro processo possível seria a determinação das regiões que apresentarão maiores semelhanças o que pode ser viabilizado pela análise de clusters. Tal objectivo obriga à transferência dos dados para o ambiente de uma ferramenta que integre essa análise, por exemplo o SPSS. Procedemos a essa transferência e, desta vez, foi-nos necessário desfazer a transformação que anteriormente tínhamos operado. A partir daí, a utilização do produto permitiu-nos elaborar um dendrograma e descobrir um certo isolamento do Baixo Mondego em relação a qualquer outra região ou grupo de regiões. Com efeito, tendo utilizado os métodos de agregação mais comuns, de acordo com todos eles esta região não se integrava em qualquer agrupamento.

Em seguida apresentamos o dendrograma e a composição dos diferentes clusters.

Figura 32. Dendrograma relativo a indicadores demográficos e sociais por NUTs

* * * * * * H I E R A R C H I C A L C L U S T E R A N A L Y S I S * * * * * *

Dendrogram using Complete Linkage



1.3. Aspectos criticáveis na informação demográfica e social providenciada pelo INE em termos de unidades territoriais

Os indicadores demográficos e sociais da responsabilidade do INE são, com certeza, um produto de utilidade elevada. No entanto, há que pensar em alguns aspectos críticos para além do mais grave de todos que é o de tratar-se de informação paga.

Em primeiro lugar, há que referir a ausência de *integridade temporal*: num processo de interrogação, deveria ser possível indicar um período relativo a uma análise da evolução de alguns dos indicadores.

Por outro lado, há que considerar a questão da actualidade: é preciso dispor de dados mais recentes embora se reconheça que, em termos de demografia, as mudanças não são tão rápidas como com outros indicadores económicos.

Finalmente, há que referir a questão da necessidade de transformação dos dados em cada ambiente para onde são transportados, condição para viabilizar o seu *processamento*. Talvez aqui a responsabilidade não seja do INE, mas dos ambientes de destino, mas é possível que a transferência em formato texto fosse mais pacífica.

Noutro contexto, poder-se-ia referir a baixa *multidimensionalidade* do sistema, mas acreditamos que aqui não seja muito importante.

2. Processos de análise multidimensional

Apresentámos no Capítulo II as características e os operadores da análise multidimensional. Nesta secção vamos tratar duas aplicações integráveis nesta área: (1) um modelo de análise multidimensional para o IDE e (2) um modelo de análise multidimensional para o comércio internacional.

2.1. Um modelo da análise multidimensional para o IDE

Nesta breve apresentação de um modelo de análise multidimensional para o IDE (MAMIDE) preocupar-nos-emos essencialmente com aspectos conceptuais e descritivos pelo que integramos nela quatro pontos essenciais: (1) uma abordagem dos elementos de análise do IDE suportada em referências pesquisadas, (2) uma especificação do modelo de acordo com os diagramas

associados à análise multidimensional, (3) alguns aspectos da aplicação ao caso português e (4) algumas conclusões e perspectivas relevantes para investigação futura (referindo-se o autor ao momento de apresentação ao seminário do departamento de Economia do ISEG). Uma apresentação completa do documento consta dos *Anexos* à dissertação.

2.1.1. Elementos da análise do IDE

Matos (1973) e Mendonça (1998) são duas referências fundamentais para a análise do Investimento Directo Estrangeiro. O primeiro tem uma abordagem extremamente diversificada, mas encontramos nela todos os elementos de uma análise multidimensional actual, nomeadamente o país investidor, o tipo de operação e a actividade económica de destino. Entre o texto de Salgado de Matos e o de António Mendonça passou mais de um quarto de século onde análises empíricas surgiram em revistas como *Economia e Socialismo* e estudos mais elaborados foram produzidos inclusivamente a nível de teses de doutoramento.

A Temática do IDE

Como se disse, Matos (1973) apresenta uma análise muito diversificada em termos do temas abordados sendo possível citar o capital social de empresas em Portugal e a participação de capitais estrangeiros, o investimento directo estrangeiro a nível mundial, o investimento directo estrangeiro em Portugal, os movimentos capital e balança de pagamentos entre centro e periferia, a rentabilidade de empresas estrangeiras e de ramos de actividade e as vendas mundiais.

Mendonça (1998) apresenta uma análise mais centrada nos movimentos internacionais de capitais, tendo como temas fundamentais o investimento directo estrangeiro em geral e o caso português, o investimento directo estrangeiro em Portugal e o investimento directo português no estrangeiro.

As Métricas e as Dimensões para Análise do IDE

As métricas utilizadas no estudo do capital social foram o capital médio em contos das empresas *portuguesas* e das empresas *estrangeiras*, o capital social total, o capital social de empresas em que existe participação estrangeira, a percentagem do capital de empresas em que existe participação estrangeira no

capital social total, as receitas, o total de sociedades com capital estrangeiro, o valor dos capitais próprios, o capital estrangeiro, o crédito, o total de sociedades registadas, o capital social criado e o número de pessoas ao serviço.

Estes indicadores foram, no todo ou em parte, considerados a nível de actividade principal ou ramo de actividade económica, ano, designação da empresa e indicador de sociedade.

Já o estudo do IDE a nível mundial utilizou um conjunto de métricas relativamente diferente, nomeadamente, o capital investido no estrangeiro por diferentes países, o valor contabilístico de investimentos no estrangeiro, o investimento líquido, os lucros totais provenientes de investimento no estrangeiro, os lucros reinvestidos, o valor, o valor acumulado em período, o valor em percentagem, o valor relativo a Portugal, o valor total, o montante e os lucros repatriados.

Neste caso, as dimensões utilizadas foram o ano ou período, a fonte de financiamento, o indicador acumulado/ano, o indicador de acumulado, o indicador Europa/EUA, o país de origem ou país investidor ou proveniência, o sector e o tipo de investimento

No estudo do Investimento Directo estrangeiro em Portugal, Matos (1973) utiliza as métricas total do investimento, percentagem do total e valor e apesar de não o fazer explicitamente pressente-se a sua necessidade de apresentar resultados em três dimensões fundamentais: país investidor, ramo de actividade económica e tipo de operação, temas a que corresponde um capítulo da obra.

Mais vaga é a abordagem do tema Movimentos Capital onde as métricas se resumem aos valor do centro para a periferia e valor da periferia para o centro e as dimensões envolvem apenas ano e tipo de movimento.

A rentabilidade dos investimentos estrangeiro em Portugal merece um capítulo em Matos (1973) e é abordada em termos de nacionalidade de capitais (embora aqui apenas considere a diferenciação entre sociedades multinacionais e americanas) e por ramos de actividade, sendo métricas fundamentais o

número de sociedades analisadas, o lucro líquido, o capital médio por sociedade, o capital social e reservas e o rendimento médio do capital próprio

O capítulo fornece ainda informação sobre a evolução dos invisíveis correntes entre 1943 e 1971 com métricas saídas, entradas e saldo e utilizando as dimensões ano, indicador empresa portuguesa/estrangeira, ramo de actividade económica e tipo de sociedade.

O último tema de Matos (1973) são as vendas mundiais onde procura estabelecer uma relação entre as vendas de participadas portuguesas com as vendas da empresa mãe o que é visto no tempo (ano) e a nível de sector/mercado.

Como se disse, Mendonça (1998) apresenta uma análise mais centrada nos movimentos internacionais de capitais, tendo como temas fundamentais o investimento directo estrangeiro em geral e o caso português, o investimento directo estrangeiro em Portugal e o investimento directo português no estrangeiro.

O primeiro tema utiliza variáveis macro-económicas exteriores ao IDE e métricas globais, a nível de ano e região, nomeadamente a formação bruta de capital fixo (FBCF), o produto interno bruto (PIB), a taxa de cobertura, o índice de Balassa, o valor do investimento directo estrangeiro em Portugal em termos brutos e líquidos (entradas), o valor do investimento português no estrangeiro em termos brutos e líquidos (saídas), o rácio entradas/FBCF, o rácio saídas/FBCF, o rácio entradas/PIB, e o rácio saídas/PIB.

A análise do IDE em Portugal utiliza as dimensões ano, indicador de operação país, região, sector e tipo de operação que se repetem na análise do investimento de empresas portuguesas no estrangeiro.

A Utilização de Hierarquias nas Dimensões

Em algumas dimensões fundamentais foi possível descortinar hierarquias, nomeadamente a nível de tempo, tipo de operação, região e sector de actividade económica.

Na dimensão *Tempo*, várias vezes se detectou a hierarquia *Período → Ano*. Na dimensão *Tipo de Operação*, é possível considerar um grupo de operação

acima do tipo de operação. Ao nível do *Sector de Actividade Económica*, é possível partir da distinção muito geral operada entre três ramos de actividade, para classificações com dois e três digitos ou mais que garantem um detalhe superior. Mais problemático é a dimensão *Região* considerada a partir do país investidor ou do país de realização do investimento. É vulgar a utilização de classificações não disjuntas e classificações que se transformam no tempo como o caso da relativa à União Europeia.

2.1.2. Especificação do modelo para análise multidimensional do IDE

Nesta secção, procuraremos detalhar o objecto central deste trabalho: o modelo para análise multidimensional do IDE. Num primeiro momento, discutiremos a adequabilidade desta variável macroeconómica à ferramenta de análise e elaboraremos um roteiro de análise.

Em seguida, determinaremos a partir das listas de elementos constituintes detectados nos estudos que referenciámos na secção anterior aqueles que, do nosso ponto de vista, melhor se ajustam ao objecto em análise.

A especificação do MAMIDE em termos multidimensionais termina a secção e torna viável a sua implementação com suporte computacional. É a altura de apresentar o que designamos por modelo conceptual e por modelo lógico e especificar o conteúdo de vários elementos nomeadamente o esquema das dimensões, o esquema da tabela de factos, o esquema das hierarquias e o esquema do cubo.

2.1.2.1. Roteiro de análise

O Investimento Directo Estrangeiro é uma questão com características adequadas à integração numa análise multidimensional, nomeadamente:

1. é susceptível de análise em múltiplas dimensões: tempo, origem, destino, actividade económica, tipo de operação;

2. encontramos em algumas dimensões níveis de agregação naturais imediatos: por exemplo, a nível de destino podemos considerar o mundo, o continente, a comunidade, o país, o distrito, a empresa, enquanto no sector de actividade podemos considerar maior ou menor agregação;

3. é relativamente fácil a construção de métricas a partir das métricas de base: o investimento líquido é uma diferença entre crédito e débito, os valores globais facilmente se transformam em percentagens, a componente temporal permite o cálculo de taxas de variação;

4. o sistema tem simetria sem necessidade de qualquer tipo de artifício: considerando o caso português é possível analisar o investimento estrangeiro em Portugal e também o investimento português no exterior.

Estas considerações levam-nos a propor um roteiro muito simples para a análise multidimensional do IDE.

Quadro 35. Roteiro de Análise para o IDE

Nível	Ordem	Tempo	Origem	Destino	CAE	Operação
0	1					
1	1	✓				
2	1	✓	✓			
2	2	✓		✓		
2	3	✓			✓	
2	4	✓				✓
3	1	✓	✓	✓		
3	2	✓	✓		✓	
3	3	✓	✓			✓
3	4	✓		✓	✓	
3	5	✓		✓		✓
3	6	✓			✓	✓

A Análise Multidimensional do IDE, relativa a um determinado período, parte do investimento externo (a crédito, débito ou líquido) global. O valor encontrado pode ser decomposto em termos de valores anuais o que é materializado no que designámos por *nível 1*. Embora possível, não vemos necessidade de integrar no modelo o investimento global no período repartido pelas diferentes dimensões de análise. Já a consideração destas no nível seguinte de análise nos parece imprescindível. Assim, ao *nível 2*, surge-nos a repartição do investimento por origem no tempo, por destino no tempo, por actividade económica no tempo e por operação no tempo. A articulação de

dimensões termina no nível seguinte onde se relaciona origem e destino no tempo, origem e actividade económica no tempo, origem e operação no tempo, destino e actividade económica no tempo, destino e operação no tempo e, finalmente, actividade económica e operação no tempo. Não vemos razão para avançar no nível de detalhe considerado: a análise multidimensional é uma ferramenta de generalização, considerar mais combinações de dimensões, apesar de possível, traz-nos praticamente para o nível operacional.

De notar ainda que a partir do primeiro nível consideramos o tempo sempre presente na análise o que permite uma visão mais dinâmica dos resultados e a constituição de elementos com características cronológicas.

2.1.2.2. Elementos considerados, não considerados e extensibilidade de análise

No modelo de análise multidimensional do IDE proposto, integramos como facto essencial os valores a débito e crédito dos movimentos de investimento directo estrangeiro a nível mundial. Estes movimentos poderão ser analisados considerando as dimensões *Origem, Tempo, Destino, Actividade e Operação.*

Na dimensão *Origem,* consideramos como elemento chave o investidor que relacionamos com o país origem e uma comunidade em que este país se integre. Uma extensão possível é a associação do continente ao país que possibilitaria uma agregação por continente. Outros agregadores poderão ser considerados de acordo com conveniências de análise devendo isso traduzir-se em atributos adicionais.

Na dimensão *Tempo* consideramos apenas o atributo *Ano,* embora seja possível uma visão do investimento a nível de mês e a consideração de períodos de agregação.

O elemento chave da dimensão *Destino* vai ser o identificador da empresa que é objecto de um movimento de investimento ou desinvestimento. A partir da empresa podemos deduzir a cidade, o país e até uma comunidade a que o país pertença sendo aqui legítimo aplicar o que dissemos antes quanto à origem.

A consideração de uma dimensão *Actividade* abre o modelo a possibilidades como a de uma empresa ter interesses espalhados por vários sectores de actividade económica. A actividade pode ser vista a um nível mais detalhado

(CAE a dois ou três dígitos) ou mais agregado (grandes ramos da actividade económica).

Na dimensão *Operação*, consideramos o tipo desta e um agregador possível semelhante ao utilizado por Mendonça (1998) no seu estudo.

Métricas relativas à questão do capital social e à participação de capitais estrangeiros não são consideradas no modelo apenas para não o tornar demasiado extenso e de difícil leitura. Note-se, no entanto, que a sua inclusão é possível através da geração de uma constelação de factos que partilhe a dimensão *Destino*. O problema do capital estrangeiro por actividade económica não ficaria resolvido neste caso pois antes considerámos não existir dependência funcional entre empresa e actividade económica. Uma visão mais rígida ultrapassaria esta questão associando o sector à empresa pelo que a deixamos em aberto.

Também não considerámos no modelo métricas relativas à rentabilidade do capital estrangeiro que podem ser incluídas através de constelação de factos.

A inclusão de variáveis macreconómicas exteriores à temática do IDE, como a FBCF e o PIB, podem ter soluções semelhantes aos elementos anteriores não considerados.

Apresentamos, de seguida, o modelo de análise multidimensional do IDE.

2.1.2.3. Modelação

Em analogia com o *framework* de Zachman (1987), vamos considerar os níveis conceptual, lógico e físico na especificação do MAMIDE. Para o primeiro nível, utilizaremos um *Diagrama Estrela* onde é representada a tabela de factos e as dimensões de análise. A nível lógico há que especificar esses elementos e as hierarquias associadas às dimensões. O nível físico, dependente da ferramenta utilizada, será referido em momento de implementação.

2.1.2.3.1. Modelo conceptual

A nível conceptual consideramos a possibilidade de uma representação mais genérica ou mais detalhada. Comecemos pela representação mais genérica

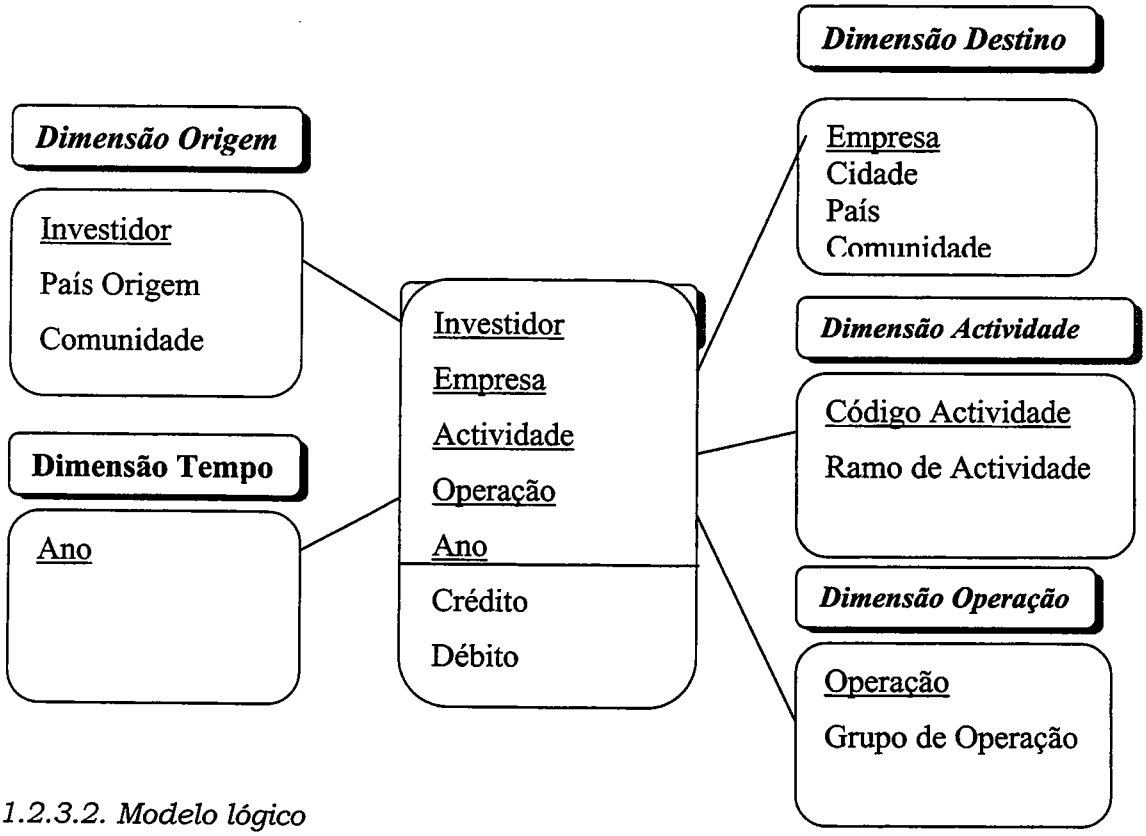
Figura 33. Representação Genérica do MAMIDE



Na figura anterior, vê-se com facilidade que vamos analisar o Investimento Directo Estrangeiro utilizando as dimensões Origem, Tempo, Destino, Actividade e Operação.

Esta representação é desenvolvida conceptualmente na figura seguinte, onde se destacam os atributos das dimensões e as métricas de análise.

Figura 34. Representação mais Detalhada do MAMIDE



2.1.2.3.2. Modelo lógico

A partir da representação mais detalhada do MAMIDE, podemos deduzir os esquemas das tabelas relativas ao facto e às dimensões. A este esquema

podemos juntar a definição de hierarquias e a partir daqui apresentar o esquema do cubo relativo à análise multidimensional do IDE.

1) Esquema das dimensões

Vamos referi-lo utilizando o estabelecido a nível da modelação relacional, mas antepondo a letra D ao nome da dimensão para a distinguir da hierarquia eventualmente existente com o mesmo nome.

D (Origem) = {Investidor, País Origem, Comunidade}

D (Tempo) = {Ano}

D (Destino) = {Empresa, Cidade, País, Comunidade}

D (Actividade) = {Código de Actividade, Ramo de Actividade}

D (Operação) = {Operação, Grupo de Operação}

Sempre que referirmos D estaremos a apontar para as várias dimensões.

2) Esquema da tabela de factos

IDE = {Investidor, Empresa, Código de Actividade, Operação, Ano, Crédito, Débito}

3) Esquema das hierarquias

No caso presente, o esquema das hierarquias coincide com o esquema das dimensões. Podemos referi-lo, antepondo a letra H ao nome da dimensão.

H (Origem) = {Investidor, País Origem, Comunidade}

H (Tempo) = {Ano}

H (Destino) = {Empresa, Cidade, País, Comunidade}

H (Actividade) = {Código de Actividade, Ramo de Actividade}

H (Operação) = {Operação, Grupo de Operação}

4) Esquema do Cubo

Um cubo é totalmente especificado através da tabela de factos, das dimensões de análise e dos caminhos de consolidação ou hierarquia, concretamente, no caso presente. O cubo relativo ao IDE será:

C(IDE) = {D, IDE, H}

2.1.3. Aspectos da aplicação ao caso português

A aplicação ao IDE em Portugal apresenta-se como um vasto campo de análise, mas o objectivo principal de testar a adequabilidade de uma ferramenta, mais do que de chegar a resultados novos, determina que se proceda a uma delimitação da aplicação. É esse o primeiro objectivo encontrado nesta secção.

Num momento seguinte, fazemos referência às questões suscitadas pelos dados utilizados (mais de 12000 linhas com o investimento de um país num sector em termos de tipo de operação por ano).

A análise dos resultados da aplicação é obviamente o momento mais importante deste texto sendo suportada exclusivamente pelo conhecimento que os dados em questão e o software utilizado suportam, procurando-se no final sistematizar algumas conclusões.

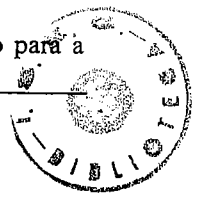
2.1.3.1. Delimitação da aplicação

Devido à extensão que uma análise global exigiria, vemo-nos confrontados com a necessidade de redução das considerações que um modelo como o elaborado possibilitaria.

Assim, apesar de os resultados, em termos de cálculo, estarem muito para além do seleccionado, vamos reduzir as considerações futuras a três elementos de análise fundamentais:

- evolução do IDE em Portugal no período 1993 a 1996;
- IDE nas dimensões Origem, Destino e Operação dado o ano;
- contribuição dos países para o investimento e desinvestimento do Mundo em Portugal.

O segundo elemento será, quando se justificar, acompanhado por uma descida a níveis de maior detalhe. Qualquer indicador produzido será visto em termos de evolução temporal.



2.1.3.2. Nota metodológica sobre os dados utilizados

Os dados utilizados tiveram como fonte estatísticas do Eurostat e foram sujeitos a uma confirmação com dados publicados pelo ICEP no documento Mendonça (1998) tendo sido detectado algum *ruído*, nomeadamente:

- ausência de valores para a Alemanha e Dinamarca em 1995;
- valores incorrectos para o Japão no subsector de Indústria Têxtil em 1993.

Decidimos desenvolver este trabalho apesar desse problema até para detectar até que ponto o produto utilizado resistia ao *ruído* e fomos levados a mais duas decisões em termos de cálculo:

- cálculo de percentagens sempre com o denominador adicionado de 0,0001 por motivos óbvios;
- consideração como *Débito* de valores classificados como *Crédito* mas tendo sinal negativo.

2.1.3.3. Implementação do Mamide

Para implementação do *Modelo de Análise Multidimensional para o IDE* vamos utilizar um produto multidimensional da empresa Speedware: o Media.

Neste produto, após a construção do dicionário que permite definir as dimensões de análise e as métricas utilizadas, existem dois processos essenciais: o carregamento de dados e o desenvolvimento / utilização da aplicação de exploração. Como se disse todos estes aspectos podem ser consultados em documento com anexos à dissertação.

2.1.4. Conclusões e perspectivas para a investigação futura: a necessidade/oportunidade de outras ferramentas para um melhor conhecimento do IDE

Dispor de um modelo e de uma ferramenta que, a partir de um conjunto de dados estruturados de acordo com determinado normativo, proporciona todo um conjunto de indicadores sobre o IDE e a sua evolução temporal em moldes reutilizáveis, permitindo, por outro lado, a análise a níveis de detalhe mais rico, é sem dúvida algo que pode melhorar a prática de análise da Economia.

As consolidações desde logo definidas no modelo através de hierarquias permitem uma navegação antes impossível e que obrigava a uma constante construção de fórmulas, modificação de repositórios e proliferação de folhas de cálculo ou bases de dados.

No entanto, este é apenas um dos aspectos que o processamento analítico pode trazer para a análise de dados económicos, um aspecto que, em certa medida, já seria possível com algumas limitações utilizando linguagens de query sobre bases de dados como por exemplo o SQL. A grande diferença entre as ferramentas tradicionais que possibilitam uma implementação relacional de estruturas de dados e as ferramentas multidimensionais tem a ver com as características dos resultados obtidos: o relacional tem uma vocação de produção em massa, enquanto o partir de uma visão geral para uma análise centrada numa instância dimensional por meios interactivos é mais característico de ferramentas multidimensionais. Trata-se, no fim de contas, de ver informação em todas as possíveis dimensões de análise com uma grande variedade de caminhos de consolidação e de detalhe. O tipo de resultado a que se chega é um somatório, uma percentagem, um indicador.

O processamento analítico pode ainda ir para além deste tipo de resultado. Com ele é possível responder a questões como:

1. Será que o investimento numa determinada empresa de um investidor é acompanhado de aplicações de outros investidores?

2. Será que um investimento numa determinada empresa de um investidor é acompanhado por um investimento do mesmo investidor noutra empresa?

3. Nos dois casos anteriores, existirá algum factor de repetição?

4. Será possível analisar transferências de investimento de Portugal para outros países?

5. Qual o impacto de recentes casos de investimento na estrutura do desemprego de uma determinada região?

Estas questões podem encontrar resposta através de modelos proporcionados pelo processamento analítico e suportados por bases de dados,

materializando-se no que podemos designar por *extracção de conhecimento de bases de dados* e traduzindo-se em resultados sob a forma de regras, árvores de decisão, associações, equações, indicadores de impacto e utilizando novos produtos de software onde as redes neuronais podem constituir um elemento importante.

O IDE não escapa a esta necessidade de um conhecimento com ferramentas cada vez mais diversificadas, mas que, ao mesmo tempo, chama a atenção para uma estruturação inicial adequada dos dados com o objectivo de se obterem os resultados mais satisfatórios, o que indicia que quem analisa a Economia deve ter cada vez mais a preocupação de analisar e desenhar correctamente sistemas de informação e, por outro lado, procurar novas ferramentas que possibilitem resultados diferentes dos habituais.

Analisar o IDE ou outra variável económica deixa assim de ser uma questão rotineira em que se procuram exclusivamente conclusões com uma dada estrutura de pensamento para um período diferente passando a integrar preocupações com novas questões a que se pode responder pelo facto de existirem bases de dados e produtos com determinadas potencialidades para a sua interrogação.

2.2. Um modelo da análise multidimensional para o comércio externo português

Nesta secção, vamos apresentar um caso de aplicação de uma ferramenta multidimensional à análise do comércio externo português feita no Gabinete de Estudos e Prospectiva Económica (GEPE) do Ministério da Economia.

Procuraremos seguir um percurso semelhante ao utilizado para o MAMIDE, embora o facto de se tratar de uma aplicação não documentada nos force a algumas adaptações.

Começaremos com a apresentação de alguns elementos de análise do Comércio Externo, designadamente a bibliografia utilizada e alguns conceitos centrais a que se seguirá uma descrição da aplicação implementada utilizando os conceitos associados à análise multidimensional, isto é procuraremos, dimensões, métricas e níveis de consolidação nas dimensões.

Num segundo ponto procederemos à especificação do sistema em termos de análise multidimensional.

No terceiro ponto, vamos proceder a uma descrição de uma utilização do produto. Tratando-se de produtos com um grau de liberdade muito elevado, necessitaremos de, previamente, descrever essa utilização a qual se pautará pela simplicidade procurando saber até que ponto é que a promessa multidimensional se materializa num caso concreto de pesquisa.

Terminaremos com alguns aspectos críticos relativos ao produto que julgamos serem bem importantes para serem tidos em conta pelos serviços que disponibilizam informação para a Economia. Não basta publicar números e fornecer uma ferramenta: há que ter um cuidado especial com as designações utilizadas ou prover um dicionário eficiente.

2.2.1. Elementos da análise do comércio externo

Existe uma longa lista de referências para análise do comércio externo. Essa extensão materializa-se em termos temporais e quantitativos. De acordo com os objectivos deste tópico, vamos limitar-nos a referências de apoio à explicitação publicadas na língua portuguesa. Neste contexto, não podemos omitir um texto publicado pelo INE para divulgação do conceitos associados ao comércio externo, INE (1996), onde, ao longo de onze capítulos, são tratados aspectos conceptuais e metodológicos. Por outro lado, ao longo da utilização da Base de Dados de Comércio Internacional apercebemo-nos da existência de um texto de apoio ao utilizador, GEPE (2001,2), que nos foi facultado por um técnico do autor institucional. Há ainda que referir as publicações do CEDIN onde Mendonça (1996) e Romão (1998) produzem formulações teóricas e aspectos associados à análise da concretização das trocas comerciais de Portugal com o exterior.

Temática do Comércio Externo

No capítulo dedicado a definições, INE (1996) apresenta os conceitos das duas operações económicas fundamentais da área do comércio externo: importação e exportação.

A *importação* consiste na entrada em território nacional de mercadorias provenientes de outros países (mercadorias entradas em regime normal). A

importação assume um carácter definitivo, após a sujeição a todos os trâmites relacionados com a sua entrada em território nacional, como sejam, a descarga das mercadorias; a apresentação da declaração de importação para despacho; a verificação da quantidade, qualidade e valor das mercadorias; o pagamento de imposições aduaneiras. Após o seu desalfandegamento as mercadorias são introduzidas no consumo fechando-se o processo de importação.

A *exportação* consiste na saída do território nacional de mercadorias com destino a outros países (mercadorias saídas em regime normal). A exportação assume um carácter definitivo, após a sujeição a todos os trâmites relacionados com a sua saída do território nacional, como sejam, a apresentação da declaração de exportação para despacho; a verificação da quantidade, qualidade e valor das mercadorias; o pagamento de imposições aduaneiras; o seu carregamento para posterior envio. Após o seu desalfandegamento as mercadorias são introduzidas no consumo no país de destino, fechando-se o processo de exportação.

Mas podem ser consideradas outras operações, de certa maneira subsidiárias conceptualmente das primeiras e que permitem uma análise empírica mais detalhada. Referimo-nos à reimportação, à reexportação e às operações relacionadas com o aperfeiçoamento.

A *reimportação* consiste na entrada no país de origem de mercadorias exportadas temporariamente, após decorrido um período de tempo permitido legalmente para a sua permanência fora do território nacional. Podemos dar, como exemplo a exportação temporária, de produtos destinados a figurarem, durante um período de tempo préviamente determinado, numa exposição internacional, os quais, finda a exposição, serão reimportados para o país de onde saíram inicialmente.

A *reexportação* consiste no envio de mercadorias para o país da sua origem ou proveniência, decorrido o prazo permitido para a sua permanência no território nacional em regime de importação temporária. O exemplo aplicado à reimportação aplica-se também à reexportação se considerarmos o movimento inverso.

As mercadorias que se podem considerar como estando abrangidas por operações sob o regime de aperfeiçoamento activo e passivo, dependem da legislação existente na matéria para cada país ou união aduaneira.

Contudo podemos definir de, um modo global, como mercadorias sujeitas a aperfeiçoamento, como sendo aquelas que estão nas seguintes condições:

a) Em complemento de fabrico, incluindo a sua montagem, reunião e adaptação a outras mercadorias.

b) A reparação de mercadorias, incluindo a sua reparação e afinação.

Considerando o aperfeiçoamento nos termos anteriores, INE (1996) aborda ainda as operações de exportação e importação para e após aperfeiçoamento activo.

Dimensões de Análise do Comércio Externo

A análise do Comércio Externo carece de três dimensões fundamentais: (1) a caracterização do agente ou dos produtos envolvidos em termos de actividade económica, (2) a caracterização do território económico que pode assumir muitos aspectos se pensarmos em alguns dos seus aspectos internos ou na sua participação em espaços mais vastos e (3) a perspectiva temporal.

A presença da dimensão tempo ao nível da análise aproxima esta análise das características definidas por Inmon

Existem múltiplos agentes económicos a actuar nas actividades do comércio externo com papéis distintos: o proprietário ou dono das mercadorias, os intermediários e o destinatário das mercadorias.

A aplicação que vamos descrever não se preocupa com aspectos relacionados com o intermediário, mas é de certa maneira importante conhecer características dos outros dois agentes ou do produto envolvido nas transacções que permitam a consideração de sectores de actividade económica para concretizar agregações.

Estando a discutir uma base de dados relativa ao comércio externo de Portugal, faz ainda sentido pensar na organização dos seus oponentes (outros países) em termos de espaço internacional. Isto permite considerar a

existência de um país origem e um país destino em termos de comércio externo. De acordo com INE (1996), por *país de origem* entende-se o país ou território estatístico onde os produtos são inteiramente obtidos ou, tratando-se de produtos em obra, onde foi realizada a última transformação ou operação substancial, economicamente justificada, efectuada numa empresa equipada para o efeito e que resulta na obtenção de um produto novo ou que representa uma fase importante do fabrico. A referência apresenta ainda cinco requisitos para a determinação de país de origem quando vários países participam sucessivamente no processo de transformação ou operação de um produto. Define-se por *país de destino* o último país ou território aduaneiro, conhecido no momento da exportação, para o qual as mercadorias devem ser exportadas. A definição de país de destino remete-nos para a necessidade, por vezes sentida, de conhecermos o país de primeiro destino, que se considera como sendo o país ou território aduaneiro em que cessa de ser utilizado o modo de transporte que levou as mercadorias à saída da fronteira nacional.

Parece assim possível a análise do comércio externo em torno das dimensões sector de actividade, mercado e tempo.

Métricas

Existem três indicadores fundamentais ao nível de análise primário: (1) o valor das importações, (2) o valor das exportações e (3) o saldo da balança comercial.

Mas é possível integrar novos indicadores associados às dimensões de análise ou situando-nos mesmo a um nível global.

O primeiro grupo de indicadores pode ser considerado como relacionado com a estrutura das importação e das exportações podendo ser considerado em peso e em valor.

A *estrutura da importação* pode ser analisada em termos sectoriais no destino e geográficos quanto à origem. Assim, a *estrutura sectorial da importação* diz-nos como contribuíram para a importação os diversos sectores de actividade em que terão sido classificados os agentes. A *estrutura geográfica das importações* informa de como a importação se terá repartido em termos de países de origem.

Os dois indicadores anteriores podem ser considerados do ponto de vista do exportador falando-se então em estrutura sectorial das expoortações (na origem) e estrutura geográfica das exportações (no destino).

Mas os indicadores a nível de comércio externo não se restringem a aspectos estruturais. GEPE (2001,2) refere a existência de um conjunto importante de indicadores presente na implementação da Base de Dados de Comércio Internacional.

O *grau de cobertura das importações* pelas exportações permite determinar quais os produtos em que o país é exportador líquido e quais aqueles em que é importador líquido. Sendo X as exportações e IM as importações pode ser expresso por

$$Gc=X/IM. \qquad (50)$$

O *Grau de Abertura* ao exterior relaciona o valor do comércio externo no seu conjunto (importações e exportações) com o do PIB sendo expresso por

$$Ga = (IM + X) / (2 * PIB)^{19} \qquad (51)$$

Sendo a abertura um processo, este indicador tem mais sentido em termos temporais ou através da comparação produto a produto. Um valor maior induz um maior grau de abertura. A técnica utilizada para o comando é idêntica à anterior considerando agora três variáveis, pelo que não a detalhamos mais.

O *Peso de défice externo no PIB* possibilita comparações internacionais e temporais permite medir quanto da produção nacional irá ficar hipotecada à salvaguarda de compromissos internacionais. Pode ser expresso por

$$PDEF = (IM - X) / PIB \qquad (52)$$

A *taxa de cobertura das importações pelas exportações* pode ser analisada a nível de produto. Neste caso, temos um comando que produz um resultado mais detalhado que o primeiro comando deste conjunto.

O indicador associado às quotas de mercado interno traduz a dependência da economia relativamente ao exterior a nível de alguns produtos e é expresso por

[19] Esta é a explicitação formulada pelo GEPE e implementada na BDCI. Em outros textos, não encontramos a duplicação do denominador

$$Qm_i = (Pi - Xi) / (Pi - Xi - IMi) \quad (53)$$

onde P é a produção interna e i o produto.

A *taxa de penetração das importações* trata-se de um indicador que reflecte o grau de satisfação da procura interna através do recurso à importação e exprime-se por

$$Tp = M / (P-X+M) \quad (54)$$

quando considerada a nível global e por

$$Tp_i = M_i / (P_i-X_i+M_i) \quad (55)$$

quando considerada a nível de produto

Cabe ainda referir que poderão existir múltiplas formas de valorização para as métricas absolutas suscitadas por diferentes condições de entrega.

Por *condições de entrega* define-se um conjunto de regras internacionalmente aceites, que estabelecem para cada tipo de contrato, o lugar onde se verifica a transferência da responsabilidade do vendedor para o comprador, no que se refere às despesas relativas ao transporte, seguro e operações de desalfandegamento das mercadorias.

A fixação do lugar em que o vendedor deverá proceder à entrega das mercadorias ao comprador é de grande importância, não só porque, a partir desse local, o comprador assume a responsabilidade por todos os riscos de desaparecimento ou danos eventuais, mas também porque o preço da mercadoria depende do local onde esta for colocada à disposição do comprador.

Com efeito, todas as despesas efectuadas desde o armazém do vendedor até ao local de entrega, são suportadas, em regra, por conta do comprador.

Das condições de entrega mais usuais no comércio internacional destacam-se as definições de valor FOB e CIF válidos para a exportação e importação respectivamente.

FOB corresponde à expressão inglesa "Free On Board", que pode traduzir-se por franco a bordo, e designa o valor da mercadoria na fronteira do país exportador, com todas as despesas incluídas.

Neste caso, o vendedor responde pelos riscos e pelas despesas da mercadorias até à sua colocação a bordo do navio, no porto de embarque indicado, competindo-lhe também, se o navio se encontrar ancorado ao largo, as despesas de fragatagem e embarque.

São responsabilidades do comprador o frete, o seguro e todas as outras despesas necessárias até à colocação das mercadorias no seu destino final.

Valor FOB = o preço das mercadorias + custos de transporte até à fronteira do país exportador + despesas de seguro relativas ao transporte das mercadorias do interior do país até à respectiva fronteira + custos de carregamento até ao navio + eventuais taxas de exportação. (56)

A notação internacional adopta a abreviatura inglesa CIF correspondente à expressão inglesa "Cost, Insurance, Freight", que se traduz por custo, seguro e frete, que significa que o vendedor estabelece o preço da mercadoria, incluindo, além do seu custo, o prémio do seguro e o frete até à colocação da mercadoria sobre o navio no porto de destino.

Valor CIF = O valor FOB da mercadoria + transporte até à fronteira do país importador + O seguro até à fronteira do país importador (57)

No entanto, é possível serem consideradas outras condições de entrega. Elas são apresentadas em INE (1996), mas, como não as vamos utilizar, não vemos razão para a sua apresentação.

Utilização de Hierarquias nas dimensões

A aplicação apresenta hierarquias a nível de actividade económica e de país. No primeiro caso, existe um nível *total* que se subdivide em 37 sectores os quais também se subdividem formando um total de 103 elementos base. É de registar a existência de outras classificações, assentes em nomenclaturas oficiais e que, em outras opções da aplicação, permitem o seu funcionamento. Nesta opção, só é possível considerar a hierarquia GEPE e nenhuma

hierarquia. Apresentamos, no quadro seguinte, o primeiro e segundo nível superiores da hierarquia GEPE.

Quadro 36. Actividades económicas segundo a hierarquia GEPE

Total
Actividades mal definidas
Agricultura, silvicultura e pesca
Indústria extractiva
Indústrias alimentares
Indústria das bebidas
Tabaco
Têxteis
Vestuário
Curtumes e artigos de couro
Calçado
Indústria da madeira
Artigos de cortiça
Mobiliário de madeira e colchoaria
Pasta para papel
Papel, cartão e embalagem
Artes gráficas e publicações
Indústrias químicas de base
Indústrias químicas ligeiras
Indústria do petróleo
Borracha
Plásticos
Cerâmica
Indústria do vidro
Cimento e artigos de cimento
Rochas ornamentais
Outros minerais não metálicos
Metalurgia ferrosa
Metalurgia não ferrosa
Produtos metálicos
Máquinas não eléctricas
Indústria eléctrica e electrónica
Construção e reparação naval
Indústria automóvel
Outras indústrias transformadoras
Electricidade, gás e água (ex)
Bancos e seguros, etc (ex)
Outros serviços (ex)

Fonte: BDCI - GEPE

Já ao nível de país, a hierarquia é estabelecida de uma forma mais equilibrada. Há o nível global (Mundo) sob o qual coabitam Europa, África, Ásia, América, Oceância e Territórios Estatísticos Diversos. No total há 272

elementos na descrição global da hierarquia, sendo ainda possível encontrar níveis intermédios até ao nível mais baixo.

Muito interessante nesta matéria é a possibilidade de considerar várias organizações para os países. Por exemplo, encontrámos agrupamentos formando a OPEP, a OCDE, os países bálticos, os Tigres e muitos outros menos conhecidos. Esta particularidade confere um grande interesse à BDCI pois, embora sem uma utilização dinâmica, acaba por implementar uma estrutura em rede e não uma estrutura hierárquica ao admitir que um elemento seja integrado em mais do que um nível superior. É verdade que em determinado momento só é utilizável um agrupamento, mas esta extensão dá uma funcionalidade acrescida ao produto e faz crescer as suas potencialidades em termos de produção de informação.

2.2.2. Especificação do modelo para análise multidimensional do comércio internacional

Nesta secção, vamos preocupar-nos essencialmente com o roteiro de análise proporcionado pela aplicação do GEPE, apresentando as suas facilidades em termos de utilização pelo analista. Terminamos especificando um modelo conceptual para a mesma.

2.2.2.1. Roteiro de análise

De acordo com GEPE (2000), a Base de Dados para o Comércio Internacional (BDCI), na sua versão para a Internet, é um subconjunto de um repositório mais vasto existente no GEPE, de informação estatística anual (valores definitivos) fornecida pelo INE, respeitante às trocas comerciais entre Portugal e os países e territórios estatísticos de todo o mundo.

A disponibilização dos dados encontra-se organizada em quatro grupos optativos: (1) Consulta Geral, (2) Consulta Específica, (3) Top 10 países e (4) Top 10 mercadorias.

A *Consulta Geral* respeita a séries de dados considerandos no período 1980-1999 ventilados por Mercados / Sectores ou Tipo de Bens / Anos.

A *Consulta Específica* utiliza informação relativa ao período de 1988-99, segundo as nomenclaturas de sectores de actividade (CAE rev.1 ou CAE rev.2),

de categorias económicas (CGCE) ou de Produtos (CTCI), consultável mediante a selecção prévia de um ano, um país ou conjunto de países e um conjunto de Mercadorias (dentro da cadeia hierárquica da nomenclatura escolhida), sendo a escolha das mercadorias é feita por código pelo que terá de consultar previamente as respectivas nomenclaturas.

A consulta designada por *Top 10 Países* lista os 10 países mais importantes em cada ano, para cada indicador (importação ou exportação), sector de actividade, categorias económicas ou tipo de produtos, respectivamente, segundo a CAE (rev.1 ou rev.2), a CGCE ou a CTCI.

A consulta *Top 10 Mercadorias* lista os 10 subsectores de actividade económica, categorias económicas ou produtos (CTCI) mais importantes em cada ano, para cada indicador (importação ou exportação) e cada mercado (país ou conjunto de países).

A forma de consulta da BDCI é dinâmica, ou seja, cada utilizador deverá interrogar directamente a base de dados para extrair a informação de acordo com os seus próprios parâmetros de análise.

Entre as quatro opções de consulta da BDCI, é ao nível da *Consulta Geral* que a possibilidade de manipulação dos dados em função do interesse específico de cada utilizador se manifesta mais rica, interessante e inovadora. Nela são disponibilizados quadros tipo, cuja transformação deverá ser operada por cada utilizador, na medida das suas necessidades através da selecção das variáveis, selecção dos mercados, dentro de uma hierarquia organizada por Continentes (designada "Mundo") ou de uma lista de agrupamentos geo-económicos pré-definidos e selecção do período de análise – ano(s).

É possível proceder à formatação de quadros multi-dimensionais, com possibilidade de cruzamento da informação, até um máximo de três elementos em linha ou em coluna.

Estamos assim perante um claro ambiente de análise multidimensional.

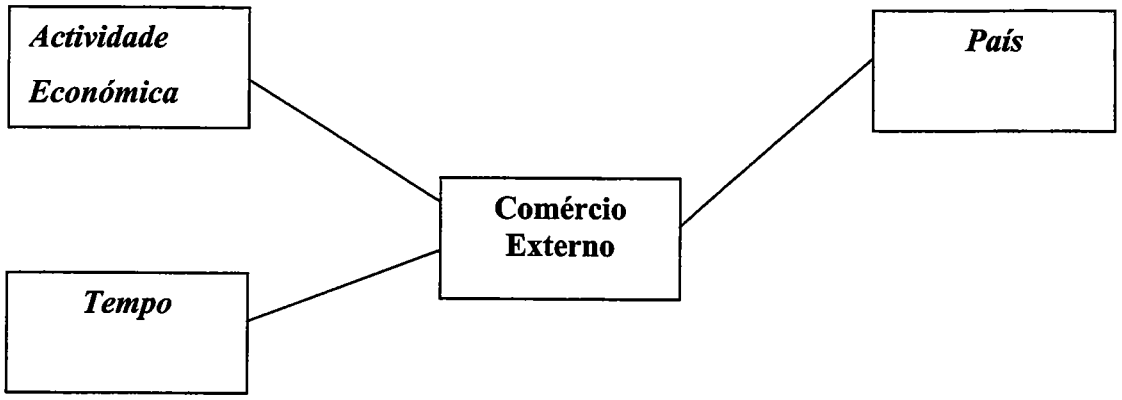
2.2.2.2. Modelação

Em termos de explicitação de um modelo multidimensional procuraremos proceder à explicitação de um modelo conceptual (uma visão do analista) e de um modelo lógico.

2.2.2.2.1. Modelo conceptual

A análise do Comércio Internacional implica três dimensões fundamentais: a associada à actividade económica, a associada ao espaço económico e a dimensão temporal o que é representado na figura seguinte.

Figura 35. Um modelo conceptual multidimensional para o Comércio Internacional



2.2.2.2.2. Modelo lógico

A BDCI não documenta o tipo de modelo lógico utilizado o que torna impossível especificar este elemento. Sabemos, no entanto, que, a nível de espaço económico, vamos encontrar uma hierarquia que nos levará do país a um espaço mais amplo e deste a Mundo. Por outro lado a partir do exame da BDCI podemos concluir sobre a existência de várias nomenclaturas associadas à caracterização da actividade económica.

2.2.3. Resultados da aplicação ao caso português

A fim de apresentarmos uma aplicação ao caso português, torna-se necessário definir o âmbito da análise. É o que fazemos num primeiro ponto. Em seguida referimos alguns aspectos metodológicos relacionados com os dados utilizados e que de certa maneira nos são impostos pela fonte estatística. Há quebras em séries, há deflatores utilizados em termos que ultrapassam completamente o

analista. Finalmente, apresentamos os resultados de análise relativamente ao problema seleccionado.

2.2.3.1. Delimitação da aplicação – o problema que vamos tratar

De acordo com GEPE (2000), os indicadores (métricas) disponibilizadas a nível da BDCI podem ser organizadas em quatro grupos:

gerais - um conjunto de indicadores cujas fórmulas são independentes da perspectiva de análise e das suas possíveis formas de leitura, dentro do quadro de multidimensionalidade e de leituras cruzadas que a base de dados oferece; estão calculados para todos os países e territórios estatísticos e respectivos agrupamentos, bem como para todas as nomenclaturas económicas, nos diferentes níveis hierárquicos considerados.

macro-económicos - um conjunto de indicadores que são expressos em relação a variáveis macro-económicas, os quais foram calculados sem desagregação sectorial e apenas relativamente a um número restrito de mercados (Mundo; UE(15) e resto do Mundo; OCDE e resto do mundo; Tríade (UE+EUA+Japão) e resto do mundo).

sectoriais - um conjunto de indicadores que têm uma leitura sectorial específica ou centrada em determinados sectores, tipo de bens ou produtos, podendo ser consultados em termos globais (para todo o mundo) ou ventilados em relação a um só país ou agrupamento de países.

geográficos - um conjunto de indicadores que têm uma leitura centrada nos mercados de origem ou de destino das correntes comerciais, podendo ser consultados em termos globais (todos os sectores, tipo de bens ou produtos) ou ventilados sector a sector.

Este conjunto de indicadores permite uma análise extremamente vasta do comércio externo. Mas, para demonstrar a sua funcionalidade e conseguir proceder a uma avaliação, não se torna necessário uma utilização total ou sequer em grande escala. É de acreditar que, se um sistema aplicacional tem um comportamento com determinadas características relativamente a uma métrica, o irá manter em outras a não ser que ocorram erros inesperados. Isso permite-nos uma delimitação da aplicação.

Sendo os indicadores relativos ao valor das importações, das exportações e saldo comercial disponibilizados por defeito em momento de acesso a produto, a utilização de um indicador diferente dos mesmos permite testar o trabalho com a escolha de indicadores e a associação ou libertação de métricas. Assim escolhemos um indicador (*grau de abertura*) como possível para um melhor conhecimento do produto.

Escolhido o indicador, elegemos duas questões como as essenciais para terem uma resposta neste documento: (1) como tem evoluído em termos globais o grau de abertura da economia portuguesa relativamente ao exterior?; (2) quais os países ou agrupamentos de países que terão sentido mais a transformação?

Isto significa que vamos tratar um indicador do comércio externo (o grau de abertura) nas dimensões tempo e região, considerando nesta dimensão a hierarquia *Mundo.→ Agrupamento → País.*

2.2.3.2. Nota metodológica sobre os dados utilizados

O documento GEPE (2000) apresenta três tipos de questões associadas oas dados utilizados: (1) quebras de séries, (2) conversão de unidades e (3) deflatores.

Regista-se uma *quebra de série* a partir do ano de 1993, que deriva da alteração do sistema de recolha de dados ao nível do comércio intra-comunitário (INTRASTAT). Este, passou a basear-se em inquéritos às empresas, sendo os resultados finais estimados pelo INE, admitindo-se que possam encontrar-se subavaliados.

Por outro lado, a actualização dos dados para 1995, introduz o apuramento do Comércio internacional segundo a nova versão da CAE (rev.2). Procurou-se minimizar os inconvenientes desta quebra de série, solicitando ao INE o prolongamento da série segundo a CAE (rev.1). Assim, a partir de 1995, a BDCI dispõe de dados segundo as duas versões da CAE.

No que respeita aos países, procurou-se ultrapassar a questão das quebras de série decorrentes do desmembramento da ex-URSS e da ex-Jugoslávia, assim como da junção do Iémen do Norte com o Iémen do Sul, através da criação da possibilidade de consulta directa dos agregados constituídos pelos territórios correspondentes. Quanto à antiga República Democrática Alemã, devido às

características específicas e ao significado reduzido do comércio com Portugal, optou-se por não criar um série conjunta com a Alemanha Ocidental (que, a partir de 1991, passa a incluir o território da ex-RDA).

As unidades de base das variáveis são, em valor, os Escudos (no momento em que redigimos este documento) e, em quantidade, o quilograma.

Incluiu-se na BDCI, por um lado, a possibilidade de *conversão dos valores* para outras unidades monetárias - ECU e Dólares EUA e, por outro lado, de arredondamentos de todas as unidades ao nível dos milhares (de Contos, de ECU, de Dólares EUA; toneladas), aligeirando a leitura dos dados aos níveis mais agregados.

Na conversão dos valores dos fluxos de importação e de exportação em Escudos para ECU e para Dólares EUA, utilizou-se, até 1994, a série de conversão de várias moedas para ECU utilizada pelo Eurostat na base de dados VISA, actualizada a partir de 1995 com os dados anuais do Banco de Portugal.

Identicamente, introduziu-se também a possibilidade de consulta das quantidades em toneladas.

A questão do cálculo das variáveis a preços constantes deparou, desde logo, com o problema da indisponibilidade de séries de *deflatores* ao nível das nomenclaturas utilizadas.

A hipótese de escolha de deflatores alternativos, o mais próximos e adequados possível às várias hipóteses de análise viáveis a partir da BDCI, sobretudo face à multiplicidade de perspectivas e de níveis de desagregação em causa, levou a deixar para o utilizador final o ónus dessas opções e das consequentes reservas metodológicas que possam surgir em cada caso, inclusive no eventual tratamento das séries por países.

Contudo, não obstante as características de multidimensionalidade da BDCI, admite-se como viável, numa próxima fase de aperfeiçoamento, introduzir o cálculo de algumas variáveis a preços constantes a nível global ou, eventualmente, por CAE (rev.1) a 2/3 dígitos e por CGCE (recorrendo, respectivamente, às Séries Longas para a Economia Portuguesa do Banco de Portugal, às Contas Nacionais do INE e aos Índices de preços do Comércio

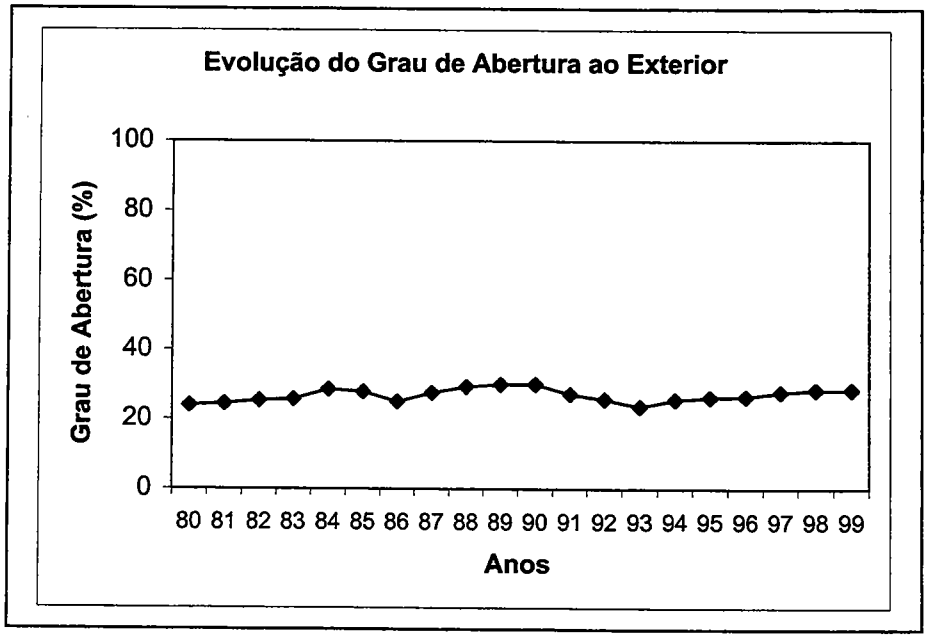
Internacional publicados pela Direcção-Geral das Relações Económicas Internacionais), eliminando, no entanto, a possibilidade de ventilação por mercados.

Entretanto, considera-se que análises de tipo estrutural, efectuadas a partir de alguns dos indicadores como os disponibilizados nesta base de dados (designadamente, os expressos em percentagem do PIB e os que traduzem posições relativas entre sectores ou países), são suficientemente significativos e válidos, permitindo identificar, no essencial, as características e tendências de evolução das trocas internacionais em períodos longos.

2.2.3.3. Apresentação do processo e resultados da análise

A figura seguinte parece indicar-nos que entre 1980 e 1999 não existe uma grande modificação ao nível do grau de abertura[20] da economia portuguesa relativamente ao exterior o qual terá passado de 23,95% para 28,47% correspondendo a uma alteração na ordem dos 4.5 pontos percentuais. Com a escala utilizada, o gráfico não permite detectar grandes flutuações no período. Esta primeira constatação é desmentida por uma análise mais profunda, mediatizada pelo operador *drill-down*.

Figura 36. Evolução global do *grau de abertura* no período 1980-1999

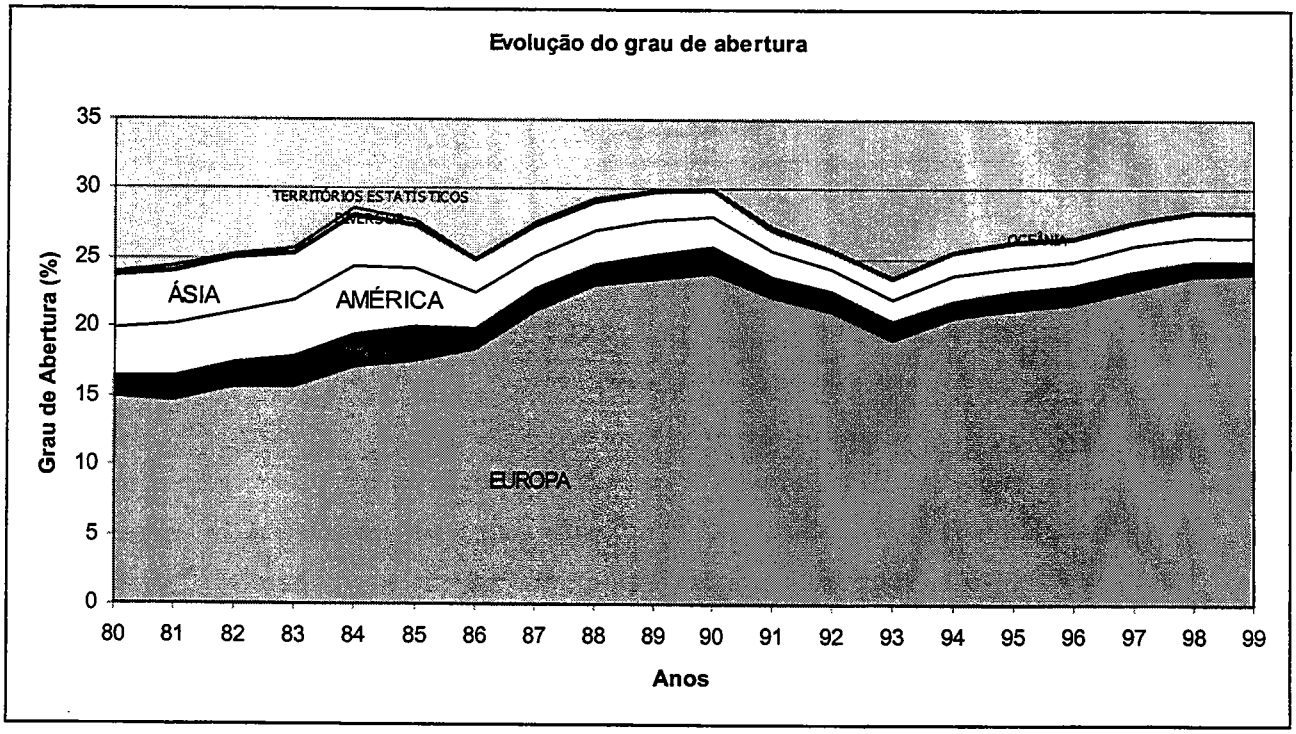


Fonte: BDCI

[20] Recordamos que a utilização do **grau de abertura** é meramente exemplificativa do tipo de análise que pode ser feita com os dados disponibilizados pelo SIE, acreditando-se que seria extensível aos outros indicadores associados à BDCI.

Um *drill-down* sobre Mundo, permite-nos conhecer o comportamento de grandes espaços perante o indicador em questão. Tal é apresentado na figura seguinte a qual indicia o crescimento do *grau de abertura* relativamente ao espaço europeu e a sua diminuição relativamente à Ásia e América. Só o valor associado ao espaço europeu mede em 1999 quase todo como o associado aos diferentes espaços. Isto é, sendo possível considerar um grau de abertura não muito diferente entre 1980 e 1999 este é, no entanto, marcado por modificações importantes a nível de espaços territoriais bem delimitados.

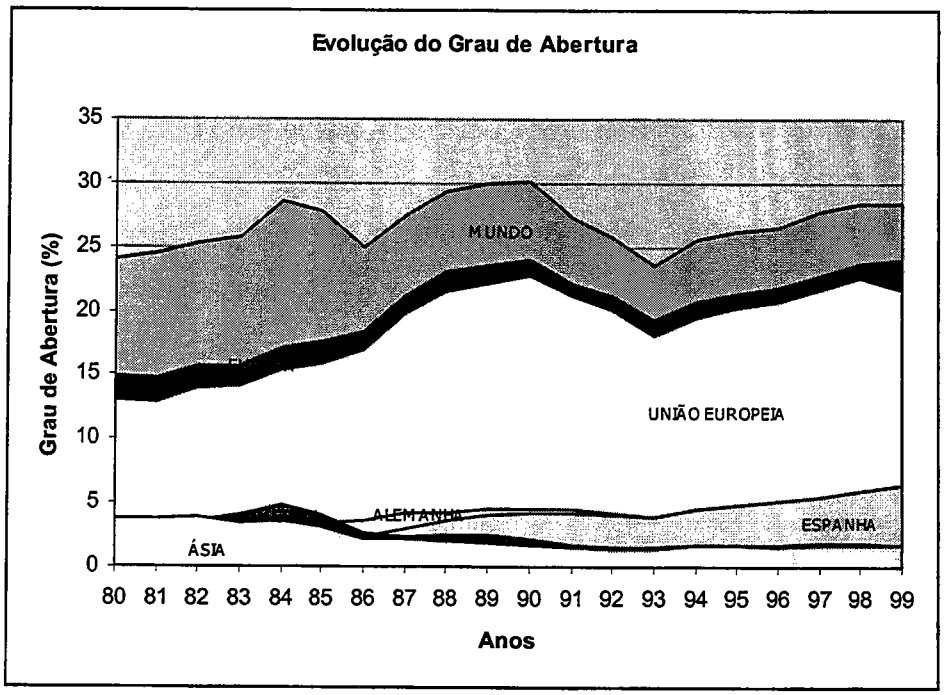
Figura 37. A evolução do *grau de abertura* por grande espaço



Fonte: BDCI

Interrogamo-nos, finalmente, se existirá algum país cujo crescimento vá influenciando decisivamente o que se passa a este nível e a figura seguinte, mais uma vez conseguida a partir de facilidades de *drill-down* mostra-nos um crescimento substancial para um país, a Espanha que se tem vindo a manifestar ao longo da década de noventa. O crescimento do grau de abertura relativamente à Alemanha é de certo modo ofuscado pelo relativo a este país.

Figura 38. Evolução do *grau de abertura* por grande espaço e país



Fonte: BDCI

2.2.4. Aspectos criticáveis da BDCI

A utilização da BDCI para resolver um pequeno problema a nível de informação para a Economia permitiu-nos detectar alguns aspectos positivos e outros mais preocupantes associados ao produto, mas que eventualmente podemos associar a produtos desta família disponibilizados no ambiente informático que caracterizou a sua disponibilização.

Entre os pontos fortes desta base de dados merecem destaque: (1) a integridade temática, (2) o esquema de rotação (3) os guias para a informação e (2) a facilidade disponibilizada na Internet

Em termos de *integridade temática* é de referir que a BDCI disponibiliza um enorme conjunto de indicadores para análise do comércio internacional, podendo considerar-se relativamente completa a lista passível de consulta.

Quanto ao *esquema de rotação* detectámos facilidades para troca de variáveis entre linha e coluna, mas também a sua integração e exclusão de dimensão de análise, correspondendo, assim, plenamente aos requisitos de operadores apresentados por Agrawal et al (1996).

No início, sentimos algumas dificuldades ao procurar utilizar a BDCI. Um contacto através de *mail* electrónico com o GEPE permitiu-nos detectar um *guia para a informação* que a curto prazo nos disponibilizou documentação e eventuais facilidades para análises mais completas.

Finalmente, é de referir que o software de análise funciona através de *facilidade disponibilizada na Internet* podendo portanto ser utilizado por analistas que não disponham dele. Este é um ponto forte no aspecto de que permite a divulgação do produto a um campo muito mais vasto do que se poderia considerar se os dados fossem disponibilizados como texto. Tal não significa que esta disponibilização não seja criticável devido a outro aspecto.

A lista de pontos fracos associados à BDCI engloba (1) a impossibilidade de criação dinâmica de indicadores, (2) a ausência de documentação sobre dimensões, (3) a ausência de documentação em termos de definição e forma de cálculo, (4) o abuso de siglas não descodificadas, (5) as dificuldades de tratamento a nível gráfico e (6) as dificuldades de articulação com outros produtos.

A *impossibilidade de criação dinâmica de indicadores* é uma característica da ferramenta multidimensional utilizada, não comum a outras ferramentas multidimensionais. Na presente perspectiva, todos os indicadores são calculados previamente e introduzidos com as variáveis de base em momento de carregamento, traduzindo-se o fornecimento ao utilizador meramente em apresentação. É possível que esta opção tenha levado em linha de conta os problemas resultantes do acesso de muitos analistas em simultâneo à BDCI a exigirem processamentos intensivos em termos de utilização de processador central, mas o certo é que cria limitações à ferramenta em termos de comparação, por exemplo, com uma folha de cálculo.

A *ausência de documentação sobre dimensões* manifesta-se sobretudo relativamente às hierarquias utilizadas. Um analista faz acesso à BDCI e encontra-se perante países ou agrupamentos de países e códigos de actividade económica sem uma explicação adicional pois a ajuda interactiva não entra nesse nível de pormenores.

O desconhecimento de métricas e do modo como foram calculadas manifesta-se na *ausência de documentação em termos de definição e forma de cálculo.*

Tais elementos deveriam ser facultados ao analista até para este fazer uma escolha mais ponderada. No entanto, esta questão foi ultrapassada através do acesso ao manual de notas metodológicas que nos foi proporcionado por um guia para a informação.

O *abuso de siglas não descodificadas* manifestou-se ao nível da nomenclatura de métricas e de descritores de actividade económica. O primeiro aspecto deriva de abreviaturas utilizadas que não permitem chegar à verdadeira designação, parecendo que as métricas só interessam a quem participou na implementação. O segundo tem a ver com a utilização de sistemas de descodificação pouco satisfatórios e com a utilização em certos casos dos códigos de actividade económica com uma remissão para uma nomenclatura que não se sabe se existe em papel ou noutro meio junto do analista. É uma solução cómoda mas não pode ser admitida em termos de informação para a economia.

As *dificuldades de tratamento a nível gráfico* fizeram com que preferíssemos utilizar as facilidades de folhas de cálculo nas três figuras anteriores. Essas dificuldades manifestavam-se ao nível do associação de designações às diferentes áreas que no levaram a uma exportação de dados.

Esta exportação mostrou a existência de uma opção que não funciona no produto que obrigou a uma solução de recurso através de uma passagem conseguida por um procedimento com vários passos. Podemos assim falar em *dificuldades de articulação com outros produtos*, embora admitamos que elas venham a ser ultrapassadas no futuro.

3. Processos híbridos

É possível encontrarmos nas diferentes concretizações de disponibilização interactiva de informação para a Economia combinações de características de produtos que caracterizámos ao tratar o processamento analítico. Tais combinações originam o que aqui designamos por processos híbridos dos quais vamos examinar e explorar dois exemplos: (1) o retrato por unidade territorial do INE disponível em INE (2001,1) e (2) a cobrança de imposto sobre o rendimento da Direcção Geral de contribuições e impostos disponível em DGI (2001).

3.1. Retrato de unidades territoriais

O fundamento deste tipo de informação consiste em disponibilizar um número de indicadores relativamente elevado em termos de unidades territoriais organizadas desde o nível de freguesia até ao de país.

Cada unidade é associada a uma bateria de indicadores e a junção de várias unidades permite comparações entre as mesmas.

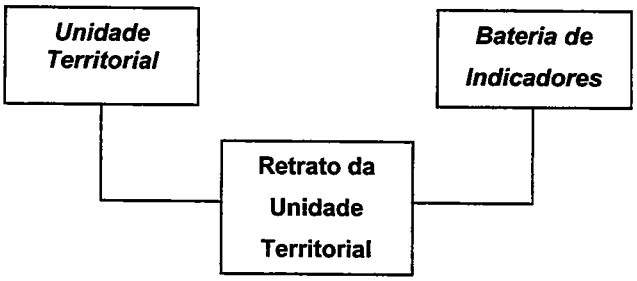
O sistema não é portador de integridade temporal, mas a organização das unidades territoriais é hierárquica. Apoiando-se nessa hierarquia, é possível realizar processos de *drill-down* e *roll-up*.

Não integramos o sistema nas ferramentas multidimensionais devido à ausência de integridade temporal.

3.1.1. Modelação

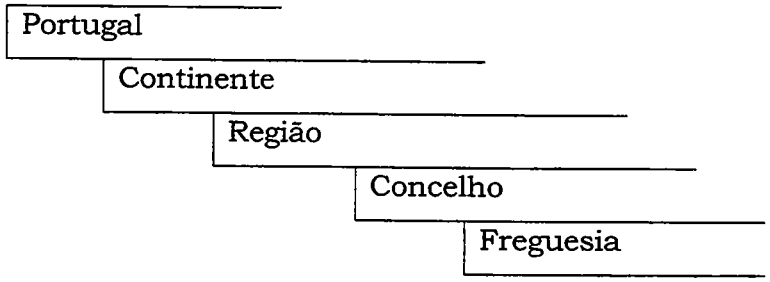
Em termos conceptuais o sistema obedece a uma estrutura muito simples documentada na figura seguinte em termos multidimensionais.

Figura 39. Modelo Conceptual do Retrato por Unidade Territorial



Ao detalharmos a análise deste sistema a um nível conceptual, constatamos a existência de uma hierarquia na dimensão *Unidade Territorial*. Esta hierarquia nasce ao nível de continente e vai até freguesia.

Figura 40. Hierarquia Presente em Retrato por Unidade Territorial



O conjunto de indicadores proporcionado a nível de freguesia é diferente do proporcionado nos outros níveis hierárquicos sendo este nível sujeito a uma tarifa.

3.1.2. Uma aplicação baseada no retrato das unidades territoriais

Quando não dispomos de uma série cronológica e os dados se referem a um mesmo momento, uma possível utilização dos mesmos é, como dissemos, a execução de comparações no espaço.

No conjunto de indicadores proporcionado por esta facilidade, encontramos os relacionados com a utilização da área territorial a qual é considerada em termos de quatro subdivisões: (1) a agrícola, (2) a florestal, (3) a urbana e (4) a afectada a outros usos.

Para ser possível ter comparações é necessário seleccionar várias unidades territoriais actividade em que utilizámos processos de *drill-down*. O quadro seguinte é o resultado de um processo de pesquisa em termos de 5 unidades territoriais utilizando a hierarquia presente na dimensão.

Quadro 37. Utilização da Área Territorial

Ano: 1998

	Portugal	Continente	Leiria	Batalha	Ourém	Santarém
Área Agrícola	39,1	39,2	30	26,4	45	70
Área Florestal	35,4	35,7	39	9,2	46,1	22,8
Área Urbana	8,7	8,5	31	3,7	1,1	7,2
Área para Outros Usos	16,7	16,5		60,7	7,9	0

Fonte: INE - INFOLINE

O quadro permite desde logo constatar uma grande diversidade de utilização da área territorial em termos das cinco unidades territoriais.

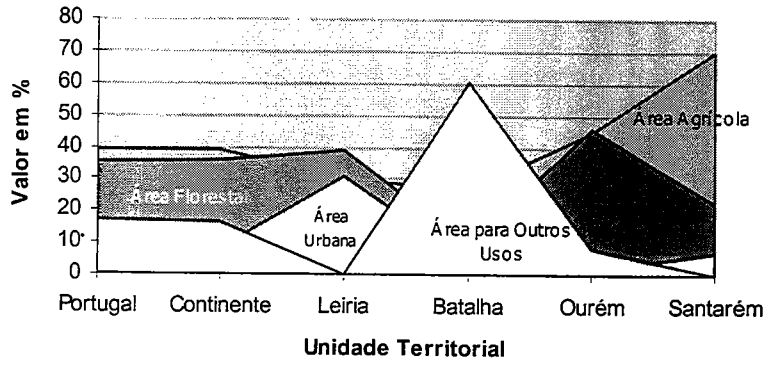
Tal diversidade ficará mais expressa se apresentada em termos gráficos. Não estando essa facilidade presente no produto que nos disponibiliza os dados, há que utilizar um produto alternativo o que implica a transferência dos dados para outro ambiente.

Ao realizarmos esta aplicação, apesar das facilidades garantidas aos utilizadores a nível de menu, o INE não possibilitava a importação dos dados o

que nos forçou a fazê-lo através de um processo de selecção e cópia presente no sistema operativo utilizado.

Na figura seguinte, apresentamos um primeiro resultado que nos mostra a dominância da área agrícola em Santarém, da área para outros usos na Batalha da área florestal em Ourém e uma certa importância da área urbana em Leiria.

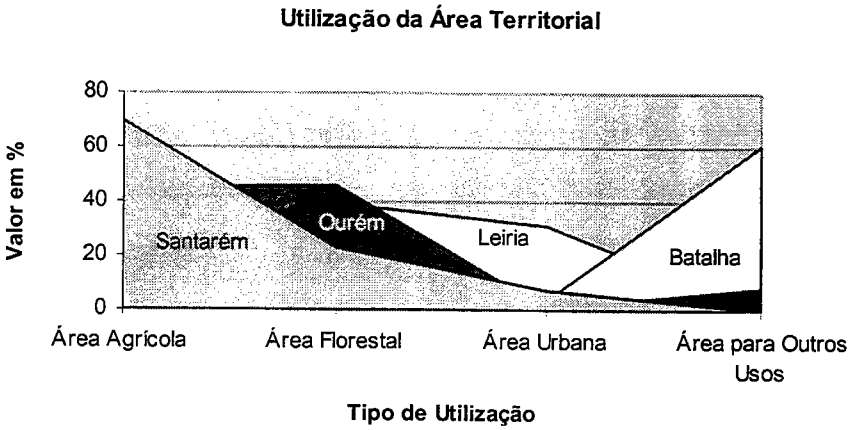
Figura 41. Utilização da àrea por Unidade Territorial



Fonte: INE - INFOLINE

O mesmo tipo de resultado pode ser obtido na figura seguinte, conseguida através de uma *rotação* nas dimensões de análise

Figura 42. Utilização da Área por tipo



Fonte: INE - INFOLINE

3.1.3. Aspectos criticáveis no retrato por unidade territorial do INE

Apesar da sua simplicidade, constatamos que um sistema como o descrito pode ser extremamente útil. O conjunto de indicadores é vasto e a estrutura de unidades é completa. Apontamos como principal ponto forte deste sistema a sua facilidade de utilização e o manancial de informação proporcionado num só acesso.

Encontrámos no entanto alguns pontos fracos, designadamente: a ausência de integridade referencial, a ausência de uma hierarquia para os indicadores, a diferença entre os indicadores proporcionados a nível de freguesia e dos outros níveis, a tarifação da informação e as dificuldades de comunicação com outras ferramentas.

3.2. Cobrança de imposto sobre o rendimento

O segundo processo híbrido que vamos referir neste documento provêm da Direcção Geral de Contribuições e Impostos e respeita à cobrança de IRS e IRC no país.

Este SIE merece ainda atenção pelo facto de a mesma informação ser disponibilizada em ambiente claramente multidimensional suportado pelo produto Excel, embora, nesse caso, seja constituída por um conjunto de ficheiros formados numa óptica anual.

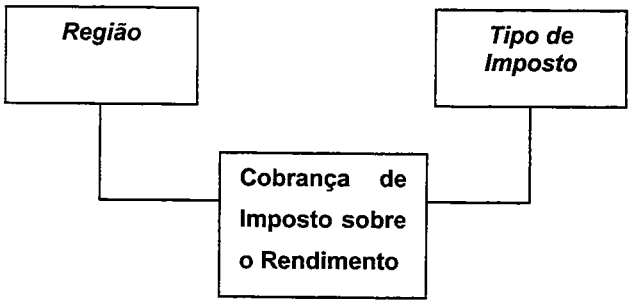
A informação disponibilizada é de natureza geral (Área, População Residente, Densidade Populacional, Freguesias, Concelhos, Repartições de Finanças) e específica (identificador de região quando for o caso, número de agregados, rendimento bruto, rendimento colectável e imposto líquido).

Esta informação é disponibilizada para o IRS e IRC devendo o utilizador escolher o imposto através de uma facilidade de interrogação.

3.2.1. Modelação

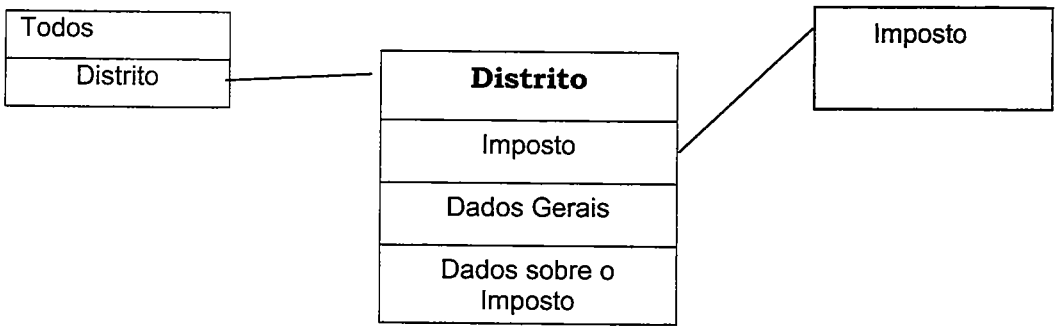
Em termos conceptuais a estrutura do sistema é bastante simples e é apresentada na figura seguinte.

Figura 43. Modelo Conceptual da Cobrança de Imposto sobre Rendimento



Mais detalhadamente, há que considerar a hierarquia presente na Região e a multiplicidade de indicadores que são proporcionados por uma instância de Região e Tipo de Imposto. O seguinte Diagrama Estrela pode ilustrar o produto.

Figura 44. Diagrama Estrela relativo à Cobrança de Imposto sobre Rendimento



3.2.2. Aplicação baseada na cobrança de imposto sobre o rendimento

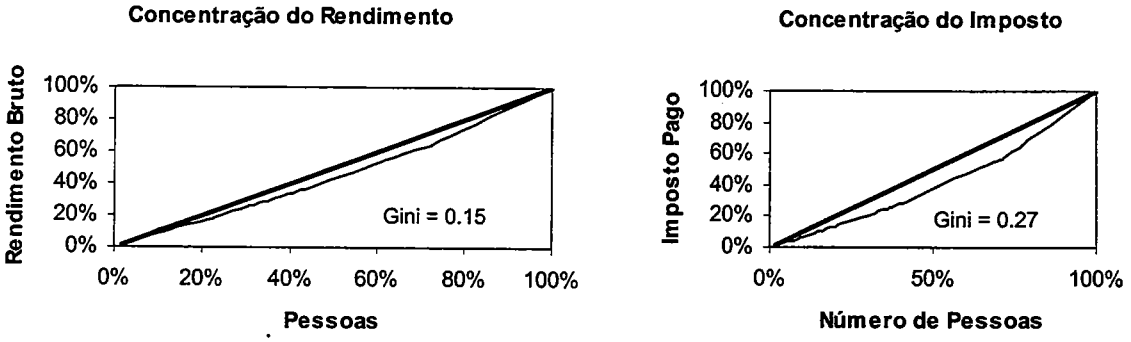
A extracção e transformação do que a DGCI disponibiliza relativamente a 1994 e 1995 permitiu a criação de uma base de dados relativa à cobrança destes impostos. Isto foi possível após um trabalho de normalização com alguma extensão que conduziu:

- à codificação de repartições, concelhos, distritos e impostos;
- à correcção de algumas designações com problemas quanto a caracteres portugueses e
- à conversão da tabela de cobranças para os códigos gerados.

O *StarJoin* construído conduziu à desnormalização da tabela convertida de cobranças pela junção da informação relativa a distrito e concelho que era dedutível de repartição. A sua obtenção torna agora possível um conjunto de apuramentos quase imediatos relativos ao *total de agregados por..., total de matéria colectável por..., total de imposto cobrado por....*

Procedemos a uma articulação dos dados assim obtidos com a exploração em folha de cálculo. Assim, desta bases de dados extraímos quadros relativos ao número total de agregados, ao rendimento bruto e ao imposto cobrado para Excel e desenvolvemos uma rotina para a determinação da concentração do imposto e do rendimento declarado, ordenando previamente os quadros pela respectiva capitação. Os resultados constam dos gráficos seguintes:

Figura 45. Concentração do Rendimento e do Imposto (1994)



A representação gráfica dos resultados sugere uma maior concentração a nível do imposto do que a nível de rendimento, já detectável pela análise da capitação, o que não refuta a hipótese de progressividade a nível de impostos sobre o rendimento.

3.2.3. Aspectos criticáveis na aplicação da DGCI

Tal como no caso anterior o ponto forte desta aplicação é a sua simplicidade. A utilização de drill-down suportado pelo mapa do país facilita a localização da área geográfica a que se pretende acesso e dá um cunho visual ao produto que deve ser tido em conta. Quanto aos pontos fracos, pensamos que não andam muito longe dos que temos apresentado para outros casos: (1) integridade temporal, (2) actualidade, (3) capacidade imediata de processamento e (4) tempo de resposta.

A *integridade temporal* é baixa: no produto os dados só existem para o ano em questão e nos ficheiros proporcionados noutro local tem que ser construída no destino pois o que é disponibilizado é um ficheiro por ano.

Também aqui é de referir o problema da *actualidade* da disponibilização. Com a tecnologia e os recursos actuais é possível ter apuramentos que respeitem a anos posteriores a 1998.

Mais grave é o que se passa a nível de *capacidade imediata de processamento*: os dados disponibilizados pelo processo híbrido em causa não são integráveis em texto ou folha de cálculo sob nenhuma forma processável, obrigando, caso não existissem os outros, a uma reintrodução no sistema do utilizador.

Finalmente é de referir que qualquer consulta a este SIE é extremamente demorada pelo *tempo de resposta* e pela sua própria organização, muitas vezes utilizando designações que criam dúvidas.

4. Processos de pesquisa de associações

A extracção de conhecimento de bases de dados integra processos de diversa natureza que vão das perspectivas lógicas às perspectivas equacionais (regressão e redes neuronais), passando por perspectivas baseadas em contagens (indicadores de impacto e redes causais). No conjunto das perspectivas lógicas, a classificação, também designada por lógica condicional, expressa-se sob a forma de regras e árvores de decisão, distinguindo-se dos processos de afinidade lógica quer em termos da linguagem de expressão quer das estruturas de dados que usa.

A Análise de Afinidades (ou Análise de Associações) é a procura de padrões e condições que descrevem como vários itens *se agrupam juntos* ou *acontecem juntos* em séries de eventos ou transacções. Uma regra de afinidade tem a forma *Quando Item1 Também Item2*. O problema considerado é encontrar regras de associação a partir de dados binários que por seu lado terão sido obtidos a partir de ficheiros de transacções organizados de acordo com os modelos usuais, nomeadamente o modelo relacional.

Esta perspectiva tem tido um considerável desenvolvimento a partir de grupos de investigação do IBM Almaden Research Center (Rakesh Agrawal), do Department of Computer Science da Universidade de Helsinquia (Heikki Mannila e Hannu Toivonnen) e, mais recentemente, do grupo de investigação formado em torno do projecto DBMiner no Canadá (Jiawei Han e outros).

Nesta secção, procuramos fazer a aplicação de algoritmos relativos à descoberta de associações em bases de dados a dados fornecidos pelo ICEP sobre o Investimento Directo Estrangeiro (IDE) em Portugal. Começaremos por apresentar os fundamentos teóricos relacionados com a questão das

associações e os da sua aplicação ao IDE. Em seguida, apresentamos os resultados da pesquisa em termos das categorias susceptíveis de integrarem associações no IDE. Terminamos com conclusões sobre algoritmos e conhecimento descoberto a que juntamos anexos para uma melhor ilustração de diferentes métodos de pesquisa de associações.

4.1. A pesquisa de associações no âmbito da análise do IDE

Na análise de dados acerca do IDE a questão da pesquisa de associações pode surgir-nos em múltiplos aspectos, nomeadamente, ao nível de associações entre sectores, transferências intersectoriais, associações entre investidores, diversificação entre empresas e transferências internacionais

Associações entre sectores. Esta associação transparece quando um investidor, feita uma aplicação numa empresa que exerce a sua actividade num determinado sector[21], a faz seguir de aplicações em empresas noutros sectores.

Transferência intersectorial. Neste caso há uma sequência de acções que se inicia por uma operação de desinvestimento a que se segue um investimento numa empresa situada noutro sector de actividade económica. A sua detecção é mais complexa que na associação entre sectores pois tem necessidade de fazer intervir o sentido da operação.

Associações entre investidores. Esta associação é detectada quando a uma aplicação de um investidor numa empresa se seguem aplicações na mesma empresa por outros investidores.

Diversificação entre empresas. Trata-se de um caso mais geral que o relativo a associação entre sectores, pois é considerado quando um investimento numa determinada empresa é acompanhado de investimentos noutras empresas independentemente dos sectores em que estas actuem.

[21] De acordo corn os dados que varnos utilizar, fornecidos pelo ICEP uma empresa exerce a sua actividade prirnordialmente num sector o que se reflecte no facto de, a nivel de modelo l6gico de dados, o sector ser urn atributo de empresa.

Transferência internacional. É o caso mais complexo e, ao mesmo tempo, impossível de detectar com os dados de que dispomos: a uma operação de desinvestimento numa empresa que opera num determinado sector num pais segue-se uma operação de investimento noutro pais. Esta associação pode ser analisada em moldes mais ou menos fechados conforme se exija ou não que a aplicação no segundo pais seja no mesmo sector de actividade económica.

4.2. Resultados da pesquisa

4.2.1. Os dados utilizados

Na ficha seguinte resumimos algumas características dos dados utilizados.

Quadro 38. Elementos sobre os Dados Utilizados

Fonte	ICEP
Ano de Referência	1997
Total de Empresas	1200
Total de Investidores[22]	1988
Total de Movimentos a débito e crédito	2414
Valor do Investimento em escudos	437.333.252.761
Valor do Desinvestimento em escudos	68.703.908.164

4.2.2. Suporte e confiança

Os *thresholds* utilizados serão diferentes de acordo com a maior ou menor granularidade da informação.

No caso de relacionamentos intersectoriais entendemos que o nível de suporte deve ultrapassar os 10% e o de confiança os 80%.

No caso de pesquisas ocorrendo em dados de base com grande variabilidade o nível do suporte tem de ser diminuído. Os padrões descobertos foram-no com um suporte na ordem de 1 por 1000 tendo-se mantido a confiança.

[22] No documento com anexos à disssertação, apresentamos a repartição destes investidores por país de origem.

4.2.3. A pesquisa de associações entre sectores

Partindo dos dados referidos, foi possível formar a tabela seguinte onde se reflectem as frequências de investimentos nos sectores considerados no ano de 1997.

Quadro 39. Frequências de Acções de Investimento por Sector

Sector	Frequência
S1	455
S2	63
S3	584
S4	131
S5	14
S6	61
S7	327

De acordo com os *threshold* utilizados, apenas os sectores S1, S3, S4 e S7 podem ser considerados frequentes pelo que a análise prosseguirá apenas com eles, procurando, agora a sua combinação, o que se reflecte no quadro seguinte.

Quadro 40. Frequências de Acções de Investimento por Combinação de Sectores

Combinação de Sectores	Frequência
{S1,S3}	25
{S1,S4}	5
{S1,S7}	8
{S3,S4}	2
{S3,S7}	10
{S4,S7}	0

Com os *threshold* utilizados e de acordo com o descrito, pode concluir-se sobre a não existência de associações entre sectores no âmbito do estudo do IDE.

4.2.4. A pesquisa de transferências intersectoriais

Com base nos mesmos dados, foi possível formar a tabela seguinte onde se reflectem as frequências de operações de investimento e desinvestimento nos sectores considerados no ano de 1997.

Quadro 41. Frequências de Acções de Investimento e Desinvestimento por Sector

Sector	Investimento		Desinvestimento	
	Referência	*Frequência*	*Referência*	*Frequência*
S1	IS1	455	DS1	93
S2	IS2	63	DS2	11
S3	IS3	584	DS3	90
S4	IS4	131	DS4	20
S5	IS5	14	DS5	1
S6	IS6	61	DS6	11
S7	IS7	327	DS7	86

De acordo com os *threshold* utilizados, apenas os sectores S1, S3 e S7 podem ser considerados frequentes em termos de desinvestimento pelo que a análise prosseguirá apenas com eles, procurando, agora, combinar acções de investimento e desinvestimento o que se reflecte no quadro seguinte.

Quadro 42. Combinação de Acções de Investimento e Desinvestimento

Combinação de Sectores	Frequência
{IS1,DS3}	1
{IS1,DS7}	2
{1S3,DS3}	19
{1S3,DS7}	1
{1S4,DS3}	0
{1S4,DS7}	0
{1S7,DS3}	2
{1S7,DS7}	15

Os resultados evidenciam que não se pode concluir sobre a existência de padrões para as transferências intersectoriais.

4.2.5. A pesquisa de associações entre investidores

É possível utilizar uma rotina em SQL para a descoberta de associações. Diferentemente dos dois casos anteriores utilizaremos o algoritmo denominado *Pesquisa com SQL* na pesquisa de associações entre investidores. Por outro lado, há que ter em conta que a granularidade dos dados é nesta pesquisa excessivamente elevada: o número de investidores é pouco menor do que o número de acções de investimento. Isto leva-nos a considerar um limite para o suporte com um valor muito baixo (2) o que não se justificava nos casos

anteriores. O algoritmo escrito em SQL parece eficiente na descoberta de associações com um nível de suporte baixo. A confiança considerada foi superior a 80%.

Quadro 43. Associações entre Investidores

Nome da Empresa	Nome do Investidor
Chavisimo-Gestão E Administração De Bens S.A.	Denidis, Sa
	Disvalor S.A.
	Haudis, Sa
	Perthois Distribution, Sa
	Plessisdis, Sa
	Romidis, Sa
	Sedan Distribution S.A.
	Sipan, Sa
	Sodibrag, Sa
	Sodimeaux, Sa
	Sodivar, Sa
	Solordis, Sa
Mecanobloc Portuguesa-Divisórias Amovíveis Lda	Alain Michel Roman
	Claude Jean Paul Boyer
	Denis Raymond Frederic Salun
	Jacques Joseph Boulard
Mecarev-Revestimentos Lda	Alain Michel Roman
	Claude Jean Paul Boyer
	Denis Raymond Frederic Salun
	Jacques Joseph Boulard
Valongodis-Sociedade De Distribuição Sa	Denidis, Sa
	Disvalor S.A.
	Haudis, Sa
	Perthois Distribution, Sa
	Plessisdis, Sa
	Romidis, Sa
	Sedan Distribution S.A.
	Sipan, Sa
	Sodibrag, Sa
	Sodimeaux, Sa
	Sodivar, Sa
	Solordis, Sa

Fonte: ICEP

O resultado evidencia a existência de dois padrões de associação a nível de investidores:

1. com confiança superior a 80%, podemos afirmar que, existindo investimento numa empresa em Portugal, de Alain Michel Roman, Claude Jean Paul Boyer e Denis Raymond Frederic Salun, existirá também investimento de Jacques Joseph Boulard;

2. com confiança superior a 80%, podemos afirmar que, existindo investimento numa empresa em Portugal, Denidis Sa, Disvalor S.A., Haudis Sa, Perthois, Distribution Sa, Plessisdis Sa, Romidis Sa, Sedan Distribution S.A., Sipan Sa, Sodibrag Sa, Sodimeaux Sa, Sodivar Sa, existirá também investimento de Solordis Sa.

4.2.6. A pesquisa de associações entre empresas

Também em relação à associação entre empresas o algoritmo *Pesquisa com SQL* teve resultados positivos, embora só permitisse chegar a conjuntos frequentes com três elementos e portanto descobrir regras tipo A,B→C.

Os resultados são evidenciados no quadro seguinte.

Quadro 44. Associações entre Empresas

Empresa 1	Empresa 2	Empresa 3
Fima-Produtos Alimentares S.A.	Iglo-Indústrias de Gelados S.A.	Indústrias Lever Portuguesa S.A.
Edificio Praça das Flores - Administração de Patrimonio Imobiliario Lda	Linha Um - Actividades Hoteleiras Lda	Tecniespaço-Mobiliario e Organização de Escritorios Lda
Belarte Ii, Publicidade Exterior Lda	Exclusivas Belarte-Mobiliario de Publicidade Lda	Produções Belarte-Som Video e Entretenimento Lda
Aransa-Sociedade de Mediação Imobiliaria Lda	Edilessa-Promoções Imobiliarias Lda	Edishora Promoções Imobiliarias Lda

Fonte: ICEP

4.2.7. Conclusões sobre a pesquisa de associações ao nível do IDE

Mesmo para o estabelecimento de conclusões pela negativa é possível afirmar-se que os processos de afinidade lógica têm aplicação na análise de variáveis

macroeconómicas nomeadamente no Investimento Directo Estrangeiro. A pesquisa efectuada demonstrou a viabilidade de descoberta de associações entre investidores e entre empresas, revelando-se eficiente a um nível muito baixo de suporte e distinguindo-as claramente associações de meros casos de repetição de acção.

Começámos por caracterizar a Afinidade Lógica de acordo com as investigações mais recentes no campo da extracção de conhecimento de bases de dados. Definimos campos de aplicação na análise do IDE onde a detecção de afinidades poderia reunir condições de interesse. Utilizando dados oficiais sobre aquela variável aplicámos algoritmos de detecção de afinidades e concluímos sobre a sua ausência ao nível sectorial quer em termos de associações positivas quer em termos de transferências e, baixando o valor do suporte, detectámos associações entre investidores e empresas.

A utilização de uma perspectiva baseada na transacção cuja repetição se despreza em vez de uma perspectiva baseada no valor é criticável, mas ela constitui a essência do método e nada obsta a que seja acompanhada por outras análises que poderão refutar ou confirmar as suas conclusões.

Mais uma vez estamos perante algo que, com a disponibilização acrescida de ferramentas na área de sistemas de informação, apoia o que temos vindo a defender: é possível analisar a realidade com múltiplos instrumentos e ver até que ponto as conclusões se reforçam em vez de se contradizer, o que significa que é dever do economista pesquisar estas ferramentas e as estruturas de dados que as suportam.

5. Breve reflexão sobre a viabilidade do processamento analítico suportado pelos SIE

No capítulo II, apresentámos os múltiplos processos que considerámos relevantes como estruturantes do processamento analítico, tendo-os subdividido em processos de análise multidimensional e processos de descoberta de conhecimento em bases de dados.

A pesquisa operada às organizações descritas no capítulo III, cujos resultados foram apresentados neste capítulo, veio demonstrar a existência, nos SIE, da primeira categoria de processos na sua forma pura ou na base de formas

híbridas. Podemos acrescentar que tal facto não é o reflexo de qualquer particularidade portuguesa, pois pode encontrar-se o mesmo tipo de processos em sítios internacionais, como é o caso do Banco Mundial, disponível em www.worldbank.org, o qual proporciona acesso livre a uma base de dados denominada *World Development Indicators*, constituída por 54 séries cronológicas de indicadores relativos a 207 países e 18 grupos, relativos aos cinco anos do período 1995-1999.

Pelo contrário, não conseguimos identificar qualquer resultado decorrente de processos de descoberta de conhecimentos em bases de dados suportado em informação dos SIE apesar de estes disporem de informação utilizável com tal objectivo. Restringir-se-á a actividade destas organizações à primeira categoria?

A questão anterior já foi abordada por pessoas responsáveis pelos destinos das organizações nacionais e internacionais cujo objectivo é a informação estatística. Saporta (2000) faz um ponto de situação, mostrando que é necessário cortar com uma perspectiva de acordo com a qual os NSI (National Statiscal Offices) deveriam limitar-se à produção de dados em vez de procederem à sua análise. Tal concepção é, neste momento, inibidora da utilização de técnicas mais poderosas de análise de dados.

Apesar de o texto referido confundir entre descoberta de conhecimento em bases de dados e mineração de dados (um dos seus passos na perspectiva de Fayadd et al (1996), não deixa de nos mostrar que um ambiente de dados com novas características - disponibilidade de grandes bases de dados, uso massivo de técnicas provenientes da comunidade dos sistemas de informação e conhecimento, presença de interesses comerciais de organizações poderosas e disponibilização de novos e mais amigáveis pacotes de software – é susceptível de ser utilizado com o objectivo de construção de modelos[23] e descoberta de padrões[24].

Saporta (2000) nota uma particularidade importante das técnicas de mineração de dados: elas utilizam registos individuais onde a

[23] Sumário global de relações entre variáveis que ensina a compreender fenómenos e permite predições (Saporta, 2000)

[24] Estrutura característica exibida por um pequeno número de pontos : por exemplo, um pequeno grupo de clientes com valor comercial elevado ou com risco elevado (Saporta, 2000)

confidencialidade é crucial. Nestas condições, o documento considera o NSI como o local indicado para a utilização destes processos o que implica novas direcções de pesquisa e novos recrutamentos associados a estas organizações.

Resumo do Capítulo IV

Não é grande a lista de processos analíticos encontrados numa pesquisa aos sistemas de informação que, em Portugal, disponibilizam informação sobre a Economia. Esta lista seria menor se não considerássemos processos potenciais a partir da informação disponibilizada explícita ou implicitamente.

Na sequência de uma pesquisa às dezasseis organizações referidas neste capítulo, encontrámos processos de interrogação, processos de análise multidimensional, processos híbridos e idealizámos processos de pesquisa de associações.

Num processo de interrogação relativo à informação demográfica e social proporcionada pelo INE, um analista faz acesso ao SIE explicitando indicadores ou regiões sobre os quais pretende informação.

Num processo de análise multidimensional, a pesquisa do analista é apoiada em operadores que permitem a obtenção de consolidações automáticas, utilizando hierarquias definidas anteriormente num sistema de *Metainformação* ou integradas na estrutura de dados. Este é o tipo de processo mais avançado, hoje disponibilizado, que foi encontrado apenas numa organização, o GEPE, em termos de análise do Comércio Internacional. Ilustrámos, ainda, no capítulo, a aplicação deste tipo de processo à análise do Investimento Directo Estrangeiro.

Tal como o seu nome indica, os processo híbridos combinam características de outros processos e não esgotam as características que permitiriam situá-los em processos mais elaborados. É o caso de ser possível a obtenção de uma visão multidimensional com uma folha de cálculo ou o uso de facilidades de *drill-down* em sistemas claramente interrogacionais. Há clara melhoria nos processos, mas não podem ser classificados na categoria mais elevada. Encontrámos este tipo de processos numa pesquisa de unidades territoriais do INE e na disponibilização de informação sobre a cobrança fiscal feita pela Direcção Geral de Impostos.

Finalmente, um processo de pesquisa de associações procura padrões e condições que descrevem como vários itens *se agrupam juntos* ou *acontecem juntos* em séries de eventos ou transacções. A ausência de qualquer aplicação nesta área e o acesso a dados do ICEP, permitiu-nos chegar a uma aplicação da pesquisa de associações ao IDE onde procurámos associações entre sectores, transferências intersectoriais, associações entre investidores, diversificação entre empresas e transferências internacionais.

Apesar de tudo, numa lista tão pequena de processos, detectámos um lote de problemas que convém recordar nas linhas seguintes.

Na informação demográfica e social do INE é patente:

- integridade temporal não satisfatória;
- actualidade deficiente;
- necessidade de transformação dos dados no ambiente de recepção.

Na Base de Dados de Comércio Internacional do GEPE, detectámos:

- impossibilidade de criação dinâmica de indicadores;
- ausência de documentação;
- siglas não descodificadas;
- dificuldades de tratamento a nível gráfico;
- dificuldades de articulação com outros produtos.

No retrato de unidades territoriais do INE, encontrámos problemas como:

- ausência de integridade referencial;
- incoerência na bateria de indicadores entre unidades territoriais distintas;
- dificuldade de comunicação com outras ferramentas;
- tarifação.

Na informação sobre cobrança fiscal da Direcção Geral de Impostos, é visível:

- integridade temporal fraca;
- actualidade deficiente;

- dificuldade na capacidade imediata de processamento;
- tempo de resposta pouco satisfatório.

Esta lista demonstra, considerando apenas o que é evidente, explícito, que há ainda muito trabalho a fazer pelos SIE para melhorar o processo de disponibilização.

Por outro lado, questões como a confidencialidade dos dados e o facto de a sua mineração fazer acesso a registos individuais mostram a necessidade de os SIE se repensarem em termos estratégicos.

Capítulo V – Métricas para análise da maturidade dos sistemas de informação para a Economia

Neste capítulo, vamos apresentar medidas para a análise de maturidade dos sistemas de informação para a Economia em termos de um conjunto de variáveis de análise, procurando na parte final determinar a sua capacidade de predição.

Na primeira secção, propomo-nos esboçar os fundamentos de um sistema que permita analisar os Sistemas de Informação para a Economia em termos do seu grau de desenvolvimento num contexto que aponta para as concepções mais recentes de Nolan, apresentadas em Nolan e Croson (1995) e em McFarlan e Nolan (1995), as quais os situam na era das redes. O modelo desenvolvido permitirá classificar os SIE em termos de um conjunto de critérios de análise.

Em seguida, vamos apresentar a concepção e os resultados da implementação de uma base de dados relativa à maturidade dos SIE em Portugal Será utilizada uma perspectiva temporal que permitirá analisar a evolução dos diversos organismos produtores de informação para a Economia em termos das variáveis centradas no analista decorrentes da secção anterior. Tratando-se de uma visão externa aos SIE, serão utilizados apenas os critérios que focam o SIE nos termos da sua relação com o analista o que será operado pela redução às respectivas variáveis de análise. Esta restrição manter-se-á até ao final do documento de dissertação.

Dispondo de uma base de dados, de um modelo de análise da maturidade e de observações das variáveis, vamos proceder a uma análise dos SIE numa perspectiva difusa utilizando a informação recolhida sobre 16 organismos produtores de informação para a Economia, tendo os dados originais sido recolhidos por observação directa do *site* durante o mês de Dezembro de 1999. A extensão da dissertação no tempo permitiu-nos operar novo exame dos *sites* em Dezembro de 2000 e testar a capacidade do sistema difuso em termos de predição. Esta secção permitir-nos-á propor um conjunto de passos para análise da maturidade em termos difusos, aplicável a outros estudos desta natureza na área de economia.

1. A Maturidade dos Sistemas de Informação no contexto da Análise da Economia

Nesta secção, de acordo com o que afirmámos anteriormente, vamos esboçar os fundamentos de um sistema que permita analisar os SIE em termos do seu grau de desenvolvimento, seguindo as concepções mais recentes de Richard Nolan.

Começamos por formular um conceito de maturidade resultante de somas ponderadas, baseadas num conjunto de critérios cuja definição será da responsabilidade do interessado na análise, mas que deverá ter em conta as grandes formulações teóricas relativa ao processo de desenvolvimento de sistemas de informação.

As descrições relativas a redes neuronais, designadamente as que se encontram em Kasabov (1996), chamaram-nos a atenção para o facto de que o resultado anterior poderia ser considerado meramente como um índice e que seria possível a concepção de uma função de maturidade com alguma analogia às funções de activação. Esse exercício é objecto de tratamento nesta mesma secção.

O terceiro ponto tem a ver com a evolução. Como se modificam os SIE no tempo? Como é que essa modificação se articula com as modificações a nível de variáveis determinantes na análise? A consideração de um ciclo de vida para as variáveis de análise é a chave para utilizar o mesmo sistema de análise de maturidade dos SI na era das redes e na era do processamento de dados. A mudança de paradigma, com a consequente passagem a outra era, não afecta esta formulação: basta determinar as variáveis de análise determinantes numa era.

A secção termina com a aplicação destes elementos teórico aos Sistemas de Informação para a Economia focando as características importantes a nível de tecnologia e a nível do Analista da Economia. O instrumento gerado pode assim ser usado no interior dos SIE para determinar as áreas cuja modificação tornarão a sua prática mais coerente com os seus objectivos.

1.1. Primeira abordagem à maturidade dos SI

Admitamos uma escala de 1 a 5 para representar o grau de adequação (A_i) de um sistema de informação concreto relativamente às características que associámos ao processo de maturidade dos SI. Consideremos um conjunto de ponderadores (P_i) para essas características. Admitamos que i representa um índice para a característica e para o ponderador e que varia entre 1 e n. O valor (V) do processo de maturidade do SI será então determinado pela fórmula:

$$V = \sum_{i=1}^{n} P_i A_i \qquad (58)$$

Se, na escala utilizada, *1* representar a desadequação total e 5 o seu inverso (a maturidade no SI), poderemos introduzir o conceito de ***DISTÂNCIA*** como a diferença entre a situação real e o limite superior daquele intervalo, sendo possível definir um vector de distâncias da seguinte forma:

$$D_i = 5 - A_i \qquad (59)$$

e uma distância global (Dg) por

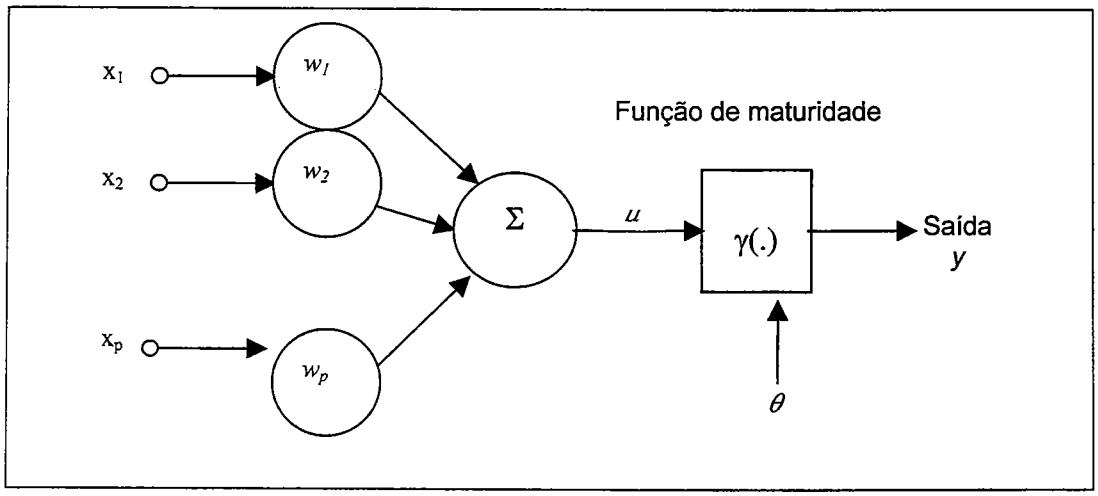
$$Dg = \sum_{i=1}^{n} P_i D_i \qquad (60)$$

As maiores distâncias positivas representarão aquelas áreas de caracterização do SI onde se devem concentrar esforços para desenvolver a sua maturidade. A distância global será um indicador de ordem geral que pode ajudar a medir o resultado das acções ou a avaliar a necessidade de esforço.

1.2 - A maturidade dos SI como resultado de uma analogia com o neurónio de McCulloch e Pitts

Os termos práticos em que a descrição anterior assentou podem ser apoiados e desenvolvidos num modelo de análise da maturidade dos SI que ultrapasse as meras somas ponderadas, baseado no modelo de neurónio artificial apresentado em Kasabov (1996). Considere-se a figura seguinte:

Figura 46. Modelo para Análise da Maturidade dos SI



Suponha-se que temos p variáveis de análise $x_1,x_2,...x_p$ com pesos w_1, w_2, ..., w_p e um limiar θ. As variáveis de análise poderão assumir valores reais ou valores booleanos (0 ou 1), enquanto os pesos e o limiar assumirão valores reais.

Neste modelo, podemos definir

o índice de maturidade μ dado por:

$$\mu = w_1x_1 + w_2x_2 + ... + w_px_p \qquad (61)$$

o ***valor da maturidade*** y dado por:

$$y = 1, \text{ se } \mu >= \theta \text{ ou}$$
$$y = 0, \text{ se } \mu < \theta. \qquad (62)$$

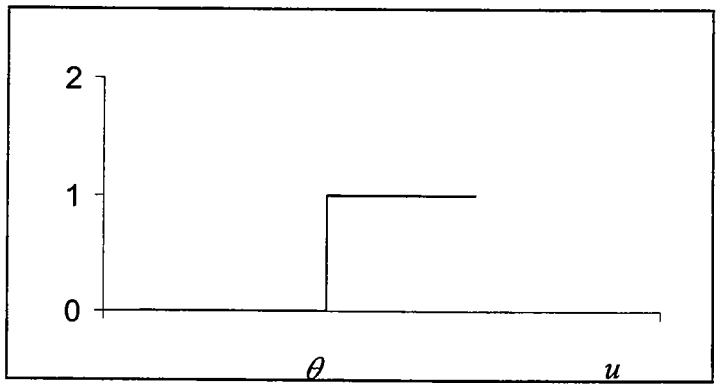
Entre as funções de maturidade susceptíveis de utilização podemos considerar:

- a função limiar de delimitadores fixos;
- a função linear;
- a função *Sigmoid.*

A função limiar de delimitador fixo implementa a regra: se o valor de input u (índice de maturidade) estiver acima de um dado limiar θ, o valor de maturidade será 1 caso contrário permanece em 0. A sua aplicação à questão da maturidade significa que um SIE, relativamente a este conceito, tem

apenas dois estados possíveis: *sim* ou *não*. Naturalmente este resultado é pouco satisfatório.

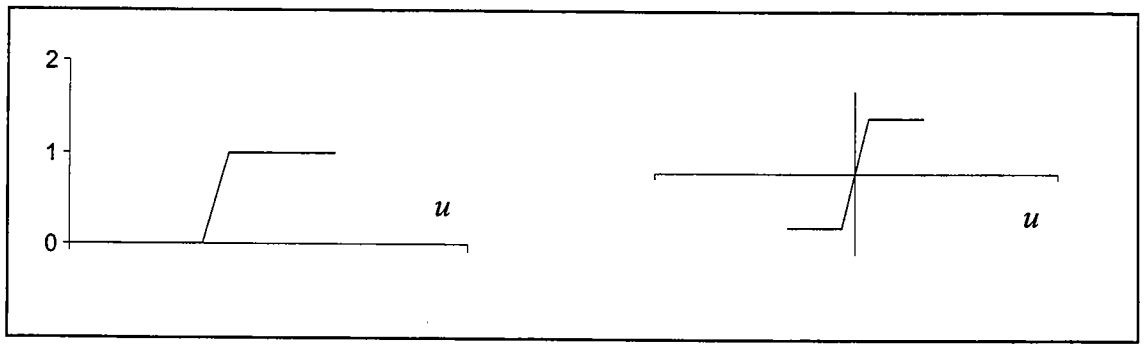
Figura 47. Função limiar com delimitador fixo



A função limiar linear permite que o valor de maturidade cresça linearmente com o índice de maturidade *u*, mas após um determinado limiar fica saturada (por exemplo, toma o valor 1).

Admitem-se diferentes variantes para esta função em função dos valores de output.

Figura 48. Variantes da função limiar linear

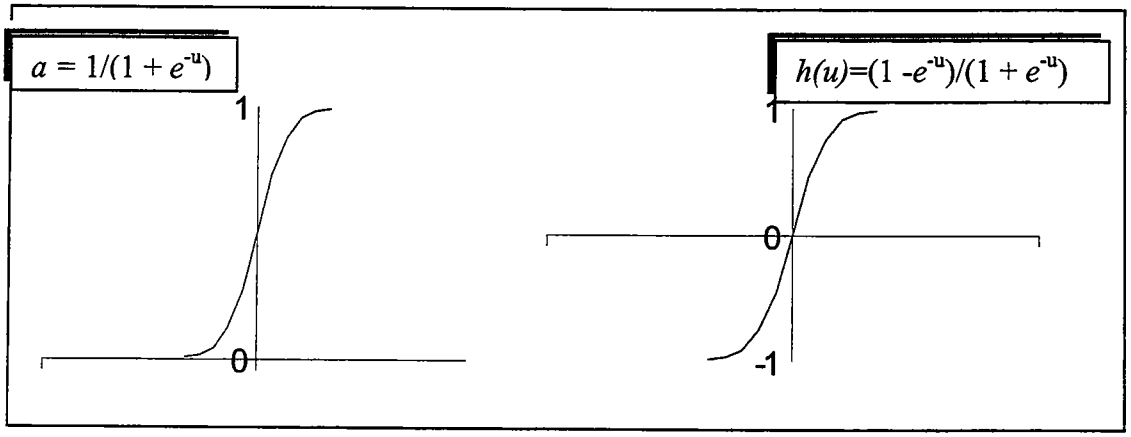


De acordo com esta função haverá um momento em que um SI é não maduro. Passada essa fase haverá um processo linear de maturação que culminará num ponto de saturação. Esta abordagem parece mais satisfatória, tendo, apenas, o inconveniente da sua linearidade.

A função *Sigmoid* é uma função de transformação não linear *g(u)* em forma de *S* caracterizada por os seus valores estarem compreendidos entre dois limites, por ser monótona crescente e contínua, diferenciável em qualquer ponto do domínio.

Na prática, diferentes tipos desta função têm sido usados, nomeadamente a função logística de expressão $a = 1/(1 + e^{-u})$, a função logística bipolar de expressão $h(u) = (1 - e^{-u})/(1 + e^{-u})$ e a tangente hiperbólica $tanh(u) = (e^{u} - e^{-u})/(e^{u} + e^{-u})$.

Figura 49. Variantes da função logística



Curiosamente, a função *Sigmoid* lembra-nos as concepções originais de Nolan com a sua famosa curva em *S* relativamente às despesas em tecnologias de informação.

1.3. A maturidade no tempo

Até este momento, considerámos a maturidade apenas em função de um conjunto de critérios de análise. No entanto, o problema deve ser visto em termos de um processo. Efectivamente, a maturidade deriva de um processo de maturação que pode ser influenciado pelos agentes.

Assim, a análise da evolução de um SI em termos de maturidade parece justa quando utiliza um conjunto de critérios de análise, mas carece da introdução da variável tempo.

Por outro lado, um modelo de análise deve admitir a possibilidade de um sistema alternar entre posicionamentos favoráveis e outros menos favoráveis.

Estes requisitos levantam uma questão que não será discutida neste documento: *como interiorizar nos agentes da gestão de um SI a necessidade de mudança se este estiver bem classificado?*

Nesta secção, vamos introduzir a questão do ciclo de vida das variáveis de análise e aplicá-la a uma função de determinação do índice de maturidade no tempo, terminando com um exemplo muito simples..

1.3.1. Ciclo de vida das variáveis de análise

Em determinado momento, uma recomendação quanto às características desejáveis para um SI pode deter uma ponderação elevada, noutro pode considerar-se que o seu valor é baixo.

A hipótese fundamental deste trabalho é que os ponderadores das variáveis de análise percorrem um ciclo de vida do tipo: surgimento, crescimento, maturidade, declínio, desaparecimento. Este ciclo pode ser simétrico, pode ser enviesado à esquerda ou à direita, mas a sua consideração permite fugir à linearidade.

Apesar de não se utilizar esta facilidade neste trabalho, devido a dificuldades em fases de declínio, é possível gerar ponderadores matematicamente utilizando uma função da forma:

$$Y = e^{\beta_1 + \beta_2 \frac{1}{X}} \qquad (63)$$

onde β_1 = lim E (lnY) quando X→ ∞, e $\beta_2 < 0$.

Com estes parâmetros Y cresce a taxas crescentes até ao valor em que X=-β_2/2 e a partir desse valor cresce a taxas decrescentes. O valor médio de *lnY* estabiliza no valor assintótico β_1.

1.3.2. O índice de maturidade no tempo

Consideremos o número de ponderadores igual a *n*. A sua não existência num determinado período é comunicada ao sistema por um peso a zero. O índice de maturidade no tempo, *f(t)*, assume a forma:

$$f_t = \frac{\sum_{i=1}^{n} Wti \ * \ Xti}{\sum_{i=1}^{n} Wti} \qquad (64)$$

onde *t* é o período de análise *(1,2,...,m)* e *i* a ordem do ponderador *(1,2,...n)*.

Exemplo 14. Consideremos o quadro seguinte com um conjunto de ponderadores hipotéticos para a análise de maturidade de um SI.

Quadro 45. Valores para os Ponderadores no Tempo

Tempo	*Ponderadores*									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
1	5	5	4	4	3	2	1	1	0	0
2	5	5	5	4	4	2	1	1	0	0
3	4	5	5	4	4	2	1	1	0	0
4	4	4	5	5	4	3	1	2	1	0
5	4	4	5	5	4	3	2	2	2	1
6	4	3	4	4	4	4	2	3	2	1
7	3	3	4	4	5	5	2	3	3	1
8	3	3	3	4	5	5	2	3	3	2
9	2	3	3	3	4	5	3	4	3	2
10	2	3	3	3	4	4	3	4	4	2
11	2	2	3	2	4	4	3	4	4	2
12	1	2	2	2	3	3	3	5	4	3
13	1	2	2	2	3	3	2	5	5	3
14	1	1	2	2	3	2	2	5	5	3
15	0	0	1	2	2	2	2	4	5	4
16	0	0	1	2	1	2	1	4	5	4
17	0	0	0	1	0	1	1	4	5	4
18	0	0	0	0	0	1	1	4	4	5

Estes ponderadores evidenciam os pressupostos enunciados: alguns detêm o valor máximo no início do período de análise o qual vai decrescendo até que no final atingem o valor 0 (é, por exemplo, o caso da existência, no interior do centro de informática, de recursos humanos para o controlo da exploração). Outros, inicialmente, não têm qualquer significado, possivelmente até serão desconhecidos e a sua afirmação temporal leva-os a deter uma posição significativa no fim do período de análise (percurso seguido pela filosofia cliente-servidor).

No quadro seguinte, apresentamos valores para as variáveis de análise para o período considerado os quais são considerados constantes reflectindo a posição: *o SI não mudou.* A geração deste valores teve em consideração o facto de os valores observados serem mais elevados para as variáveis de análise cujo ciclo de vida esteja na fase de maturidade e na sua vizinhança, reflectindo o pressuposto de que houve um momento em que o SI atingiu a fase de maturidade. Na última coluna, é apresentado o *índice de maturidade*, calculado de acordo com a última

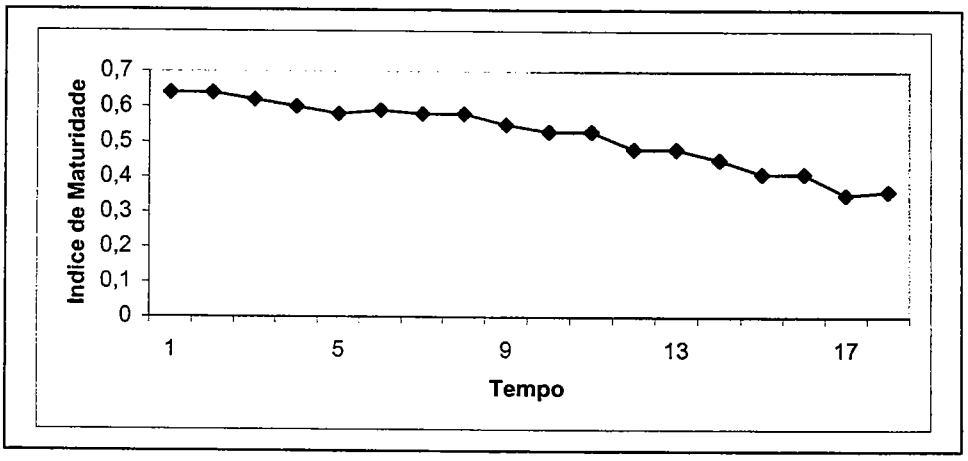
expressão apresentada, fundamental para avaliar o valor do SI em termos de maturidade, o qual revela o envelhecimento do sistema.

Quadro 46. Valores para as variáveis de análise e índice de maturidade

Tempo	*Valores Observados*										*Índice de*
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	***maturidade***
1	1	0.6	0.5	0.4	0.6	1	0.4	0.3	0.2	0.4	0.64
2	1	0.6	0.5	0.4	0.6	1	0.4	0.3	0.2	0.4	0.64
3	1	0.6	0.5	0.4	0.6	1	0.4	0.3	0.2	0.4	0.62
4	1	0.6	0.5	0.4	0.6	1	0.4	0.3	0.2	0.4	0.60
5	1	0.6	0.5	0.4	0.6	1	0.4	0.3	0.2	0.4	0.58
6	1	0.6	0.5	0.4	0.6	1	0.4	0.3	0.2	0.4	0.59
7	1	0.6	0.5	0.4	0.6	1	0.4	0.3	0.2	0.4	0.58
8	1	0.6	0.5	0.4	0.6	1	0.4	0.3	0.2	0.4	0.58
9	1	0.6	0.5	0.4	0.6	1	0.4	0.3	0.2	0.4	0.55
10	1	0.6	0.5	0.4	0.6	1	0.4	0.3	0.2	0.4	0.53
11	1	0.6	0.5	0.4	0.6	1	0.4	0.3	0.2	0.4	0.53
12	1	0.6	0.5	0.4	0.6	1	0.4	0.3	0.2	0.4	0.48
13	1	0.6	0.5	0.4	0.6	1	0.4	0.3	0.2	0.4	0.48
14	1	0.6	0.5	0.4	0.6	1	0.4	0.3	0.2	0.4	0.45
15	1	0.6	0.5	0.4	0.6	1	0.4	0.3	0.2	0.4	0.41
16	1	0.6	0.5	0.4	0.6	1	0.4	0.3	0.2	0.4	0.41
17	1	0.6	0.5	0.4	0.6	1	0.4	0.3	0.2	0.4	0.35
18	1	0.6	0.5	0.4	0.6	1	0.4	0.3	0.2	0.4	0.36

É possível concluir que, apesar de um SI não se modificar no tempo, o facto de as variáveis relevantes para análise de maturidade serem sujeitas a um ciclo de vida tipo *surgimento, crescimento, declínio e desaparecimento* conduz ao envelhecimento do SI. A figura seguinte ilustra este facto.

Figura 50. Função de Maturidade no Tempo sem Modificações das Observações



Mas é possível simular uma posição *mais realista* de acordo com a qual um SI que vai evoluindo com modificações nas observações feitas. Algumas destas modificações podem corresponder a acções isoladas, não sustentadas, o que faz com que o seu impacto deixe de se fazer sentir em períodos seguintes.

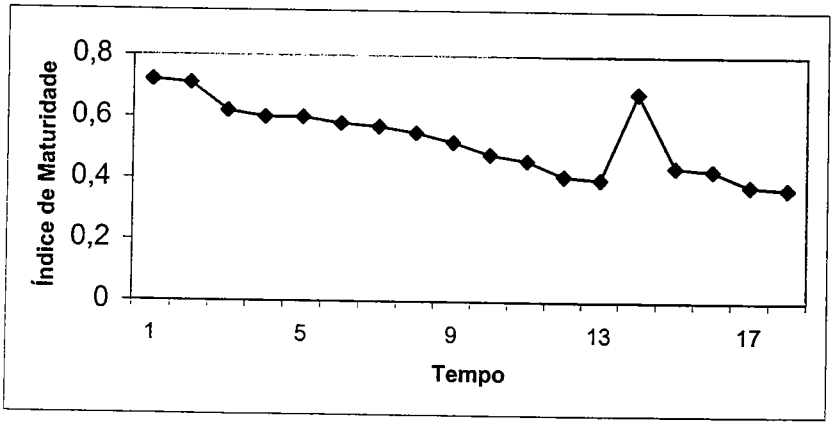
Exemplo 15. O quadro seguinte reflecte a existência na sua elaboração da consideração de um conjunto de valores que denunciam a existências de acções sem continuidade que poderão traduzir-se em reflexos imediatos positivos de curto prazo, mas cuja avaliação para o período conjunto torna imediatamente transparente.

Quadro 47. Valores para as variáveis de análise e índice de maturidade

Tempo	*Valores Observados*										*Índice de*
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*maturidade*
1	1	1	0.5	0.4	0.6	1	0.4	0.3	0.2	0	0.72
2	1	1	0.5	0.4	0.6	1	0.4	0.3	0.2	0	0.71
3	1	0.6	0.5	0.4	0.6	1	0.4	0.3	0.2	0	0.62
4	1	0.6	0.5	0.4	0.6	1	0.4	0.3	0.2	0	0.60
5	1	0.7	0.6	0.6	0.4	1	0.4	0.3	0.2	0	0.60
6	1	0.6	0.5	0.4	0.6	1	0.4	0.3	0.2	0	0.58
7	1	0.6	0.5	0.4	0.6	1	0.4	0.3	0.2	0	0.57
8	1	0.6	0.5	0.4	0.6	1	0.4	0.3	0.2	0	0.55
9	0.8	0.6	0.5	0.4	0.6	1	0.4	0.3	0.2	0	0.52
10	0.7	0.6	0.5	0.4	0.6	1	0.5	0.2	0.2	0	0.48
11	0.4	0.6	0.5	0.4	0.6	1	0.5	0.2	0.2	0	0.46
12	0.3	0.6	0.5	0.4	0.6	1	0.5	0.2	0.2	0	0.41
13	0.3	0.6	0.5	0.4	0.6	1	0.5	0.2	0.2	0	0.40
14	0.3	0.6	0.5	0.4	0.6	1	0.5	0.8	0.8	0.7	0.68
15	0.3	0.6	0.5	0.4	0.6	1	0.5	0.8	0.2	0	0.44
16	0.3	0.6	0.5	0.4	0.6	1	0.5	0.8	0.2	0	0.43
17	0.2	0.6	0.5	0.4	0.6	1	0.5	0.8	0.2	0	0.38
18	0.1	0.6	0.5	0.4	0.6	1	0.5	0.8	0.2	0	0.37

A figura seguinte reflecte um conjunto de observações onde ressalta este tipo de acções. Com essas observações, a representação da maturidade no tempo evidencia uma tendência para a descida embora momentaneamente possa mostrar picos associados à modernidade não sustentada.

Figura 51. Função de Maturidade do SI mais realista



1.4. A maturidade no contexto da análise da Economia

A abordagem da questão da maturidade no contexto da análise da Economia traduz-se nos ponderadores a considerar como relevantes de acordo com a sua localização temporal.

Podemos considerar estes ponderadores como estando ligados à tecnologia ou sendo determinados por uma perspectiva do SI centrada no analista económico.

Muitos dos critérios que vamos referir têm a ver com as regras de decorrentes de Codd (1993) para ferramentas OLAP, tecnologias cujos objectos, a nível de informação, detêm elevado grau de parentesco com os problemas de Economia.

1.4.1. Ponderadores relacionados com a tecnologia

Multidimensionalidade. Os SI que apoiam a análise poderão disponibilizar ferramentas para a análise multidimensional dos dados, embora, como servidores, a sua vocação se relacione mais com o fornecimento de dados.

Extracção de conhecimento de bases de dados. Pela mesma razão, poderão viabilizar processos de extracção de conhecimento de bases de dados como a classificação, a pesquisa de associações, o tratamento de redes causais.

Acessibilidade. Deverão, ainda, conhecer a localização física dos dados. As análises especificadas pelos utilizadores, com base nos esquemas conceptuais de dados, deverão ser dirigidas para a fonte mais apropriada ao suporte da pesquisa. As ferramentas deverão ainda proceder à conversão e reformatação automática dos dados, para a visualização da informação ser realizada da forma mais adequada ao modelo de análise utilizado.

Arquitectura Cliente/Servidor. Este tipo de arquitectura do SI separa radicalmente os ambientes de suporte dos dados dos ambientes computacionais que se destinam à sua apresentação ao utilizador.

Suporte multiutilizador. A gestão da concorrência é essencial no suporte às actividades de tomada de decisão em grupo e, por outro lado, à resposta às solicitações provenientes dos mais diversos analistas.

Manipulação de dados intuitiva. O analista deverá poder efectuar pesquisas intuitivas sobre os dados, como, por exemplo, a navegação ao detalhe dos dados, sem necessitar de utilizar linguagens de comandos ou menus de acções.

Flexibilidade na criação de relatórios. O analista deverá poder apresentar, em diversos formatos, os dados pela perspectiva, ou conjunto de perspectivas, mais conveniente para a análise pretendida

Transparência. Os analistas não terão de conhecer a complexidade técnica da estrutura de dados ou das ferramentas de exploração que suportam as suas análises, devendo apenas concentrar-se na sua utilização

Mecanismo de Segurança. O SI deve ser protegido relativamente a acessos indevidos e deve ter possibilidade de seleccionar a informação que disponibiliza aos analistas de acordo com o seu perfil de acesso.

A questão da multidimensionalidade pode ser analisada como um subproduto de um conjunto de critérios importantes como é o caso de:

1) o aumento do número de dimensões de análise, ou do número de níveis de agregação, não dever prejudicar o desempenho do sistema *(Desempenho consistente);*

2) todas as dimensões de análise deverem ser equivalentes, tanto ao nível estrutural como funcional *(Dimensionalidade genérica)*;

3) os modelos multidimensionais poderem conter milhões de células, muitas das quais não terão, em determinadas circunstâncias, valor atribuído, as quais deverão ser armazenadas eficientemente, de forma a não prejudicarem o desempenho do sistema e a optimizarem a gestão da memória *(Gestão de matrizes esparsas)*;

4) as ferramentas permitirem a manipulação de dados e a realização de cálculos entre quaisquer dimensões de análise, bem como não limitar as possibilidades de relacionamento entre células *(Operações interdimensionais ilimitadas)*;

5) não ser imposto um limite ao número de dimensões de um modelo multidimensional *(Dimensões e níveis de agregação ilimitados)*.

1.4.2. Ponderadores centrados no analista da Economia

São, fundamentalmente, derivados das preocupações de Davenport e das concepções que defendem uma finalidade social para a Economia e para os instrumentos de produção.

Facilidades de Navegação. A existência de ajudas à pesquisa e de facilidades de extracção são elementos importantíssimos no apoio ao analista o qual pode ser reforçado por informação acerca do site e ajudas em sentido genérico.

Suporte informacional à multidimensionalidade de análise. A Economia é uma área de conhecimento onde algumas das variáveis já referidas carecem de uma expressão a nível de região origem e região destino, sector de actividade económica, tipo de transacção, tempo e, mesmo sem se ser exaustivo, onde é de admitir a existência de analistas interessados em apenas algumas destas particularidades. Assim, os sistemas de informação que a apoiam deverão suportar a análise multidimensional dos dados pela existência de séries com nível de detalhe satisfatório ao trabalho de análise.

Informação ao analista. Esta característica manifesta-se num conjunto de elementos onde destacamos a existência de guias para a informação. Se ao nível dos SI se encontrarem elementos (pessoas, documentos, repositórios) que possibilitem uma determinação da informação existente e a forma de lhe ter acesso esse factor é um facilitador de acesso e pesa positivamente na análise de maturidade.

Abertura ao Analista. Tal como nas organizações os sistemas de informação interiorizam as necessidades dos seus utilizadores é de admitir que no campo da Economia estas sejam manifestadas pelos analistas e que este constitua um vector importante do seu desenvolvimento, Existem outras características importantes como a dinamização de fóruns de discussão e a universalidade de expressão. O terreno da Economia Internacional faz surgir imediatamente este critério, de acordo com o qual é conveniente a utilização de uma linguagem de expressão o mais conhecida possível ou a utilização potencial de várias linguagens de forma a que o analista possa determinar a que lhe interessa. Esta facilidade já se encontra ao nível de estatísticas do Eurostat, podendo o utilizador escolher a linguagem de expressão. Outro elemento importante tem a ver com as questões confidencialidade e transparência. Também, quanto a este critério, é vulgar encontrarem-se situações em que os dados não são disponibilizados por serem confidenciais. Naturalmente, tudo o que envolve a identificação de pessoas ou organizações pode traduzir-se numa violação da sua privacidade. No entanto, isso deixa de passar-se caso a informação de identificação seja sujeita a um *processo de anonimização*. Por exemplo, não permitir o acesso a um ficheiro de desempregados por questão de privacidade gera desconhecimento, enquanto permitir o acesso a todos os dados de caracterização menos a identificação da pessoa possibilitará todo o tipo de estatísticas sem perda de privacidade.

A disponibilização gratuita de dados e publicações. O princípio de um fornecimento dos dados mediante uma reciprocidade monetária, chegando ao extremo de valorizar células de folhas de cálculo objecto de *download* como é o caso do Instituto Nacional de Estatística no produto ***Infoline***, não parece permitir que esta seja uma questão na sociedade

actual. No entanto, é defensável que um SI centrado nas pessoas disponibilizará a informação gratuitamente.

Integração da informação no tempo. Os dados disponíveis devem referir-se a vários períodos económicos, claramente delimitados e garantir a sua coerência relativamente aos critérios de obtenção de forma a permitir a análise de evoluções e a efectivação de comparações.

Capacidade de processamento posterior. A extracção de informação de um SIE pode ser para publicação imediata num documento, mas o seu destino mais natural é a integração em processos de obtenção de nova informação e conhecimento. O grau de automatismo que este processo pode ter que, no nível mínimo, exige uma total reintrodução dos dados é, sem dúvida, um elemento que permite medir até que ponto um SIE está centrado no analista.

Actualidade. Trata-se de outra importantíssima característica da informação extraída, eminentemente relacionada com o analista. Este, com certeza, tem necessidade dos dados mais recentes sem o que o seu estudo perde valor.

Diversidade. Se um analista conseguir encontrar num SIE informação que lhe possibilite ganhar o tempo de acesso a outro que inclusivamente poderá desconhecer e portanto que teria de ser sujeito a um processo de pesquisa por palavra-chave ou alternativo, terá um importante ganho que se reflectirá positivamente numa diminuição do tempo de conclusão ao seu projecto.

1.4.3. O portfolio dos SI no contexto da análise da Economia

Tendo chegado a dois grupos de ponderadores, é possível explicitar alguns resultados considerando para cada grande grupo três subdivisões fundamentais.

Em termos tecnológicos, é possível descortinar a existência de:

1. Sistemas de Informação de valor tecnológico baixo.
2. Sistemas de Informação tecnologicamente evoluídos em termos de processamento e recolha.

3. Sistemas de Informação tecnologicamente evoluídos em termos de processamento, recolha, extracção, pesquisa, descoberta e disponibilização.

Em termos do analista, as subdivisões consideradas são:

1. Sistemas de informação sem valor informacional.
2. Sistemas de informação com valor informacional fechado aos analistas.
3. Sistemas de informação com valor informacional virados para os analistas.

Considerando as formulações anteriores das somas ponderadas, agora separadas para cada um dos grupos de ponderadores, é possível chegarmos a uma representação do valor dos SI no contexto da análise da Economia e organizá-la em termos de três vectores essenciais: supressão, reeducação e modelo.

Quadro 48. Caracterização de Maturidade dos SI

		Supressão	Reeducação	Modelo
Valor Tecnológico	Elevado	(3)	(6)	(9)
	Médio	(2)	(5)	(8)
	Baixo	(1)	(4)	(7)
		Baixo	Médio	Elevado

Valor para o Analista

Os sistemas de baixo valor para o analista, apesar da sua eventual riqueza tecnológica, representam práticas de utilização dos SI assentes em objectivos definidos no interior do órgão de controlo do SI sem uma preocupação de alinhamento com os interesses dos que carecem de elementos informacionais e justificam a existência do serviço. O SI, neste caso, surge como um fim em si mesmo e o facto de estar desligado dos interesses dos que procuram servir-se dele leva à eventualidade da sua supressão independentemente da capacidade tecnológica que pode ser transferida para áreas de maior valor informacional.

Os sistemas com valor informacional fechados ao analista carecem de um processo de reeducação, redefinição e desburocratização que, de acordo com o seu estado em termos tecnológicos, pode ser acompanhado por instalação e utilização de ferramentas para processamento, recolha, extracção, pesquisa, descoberta e disponibilização

Os sistemas de elevado valor para o analista funcionam como modelos em termos da sua relação com utentes e podem deter desde logo uma estrutura computacional adequada ou carecer de investimentos que levem a uma melhor adequação da sua capacidade técnica à sua finalidade económico-social.

2. Uma base de dados relativa à maturidade dos Sistemas de Informação para a Economia

Nesta secção, vamos descrever a concepção e os resultados da implementação de uma base de dados relativa à maturidade dos SIE em Portugal Será utilizada uma perspectiva temporal que permitirá analisar a evolução dos diversos organismos produtores de informação para a Economia.

Tratando-se de uma visão externa aos SIE, serão utilizados, como se referiu inicialmente, apenas os critérios que focam o SIE nos termos da sua relação com o analista o que será operado pela redução às respectivas variáveis de análise. Esta restrição manter-se-á até ao final do documento de dissertação.

Reportando-nos exclusivamente à secção, começaremos por apresentar o modelo conceptual de um sistema vocacionado para a análise de maturidade dos sistemas de informação para a Economia utilizando para tal efeito o que se denomina por *diagrama entidade associação* proveniente da escola de Chen (1978) e apresentado com desenvolvimentos em McFadden (1994).

Esta perspectiva será complementada pela utilização da derivação do *modelo relacional* como elemento determinante da constituição da base de dados matéria em que se seguirão as regras vigentes ao nível das formulações da área de sistemas de informação. É possível encontrar uma articulação completa entre o diagrama entidade associação, a derivação relacional e o processamento analítico na obra Ramakrishnan e Gehrke (2000) que tem por objectivo a apresentação de sistemas de gestão de bases de dados. A sua existência mostra que a utilização da *modelação entidade associação*, apesar de ter completado um quarto de século, permanece válida no momento actual.

Num terceiro ponto, apresentamos algumas indicações sobre o processo de consolidação. Na base deste sistema está a ausência ou presença de uma característica da relação do SIE com o analista. Essa característica agrupa-se ao nível de uma variável de análise proveniente da secção anterior. Ao todo são

consideradas 66 características para a descrição da maturidade pelo que esta consolidação é extremamente importante em termos de redução.

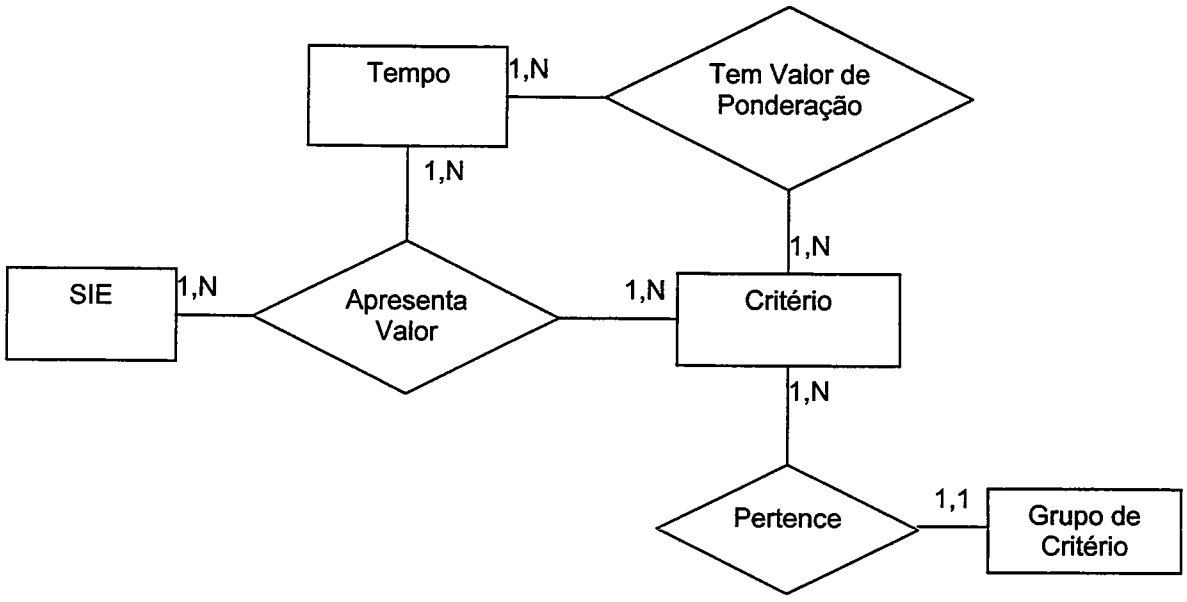
Num último ponto, examinamos o estado dos SIE em termos das características consideradas o que permite desde logo algumas ilações sobre a disponibilização interactiva de informação para a Economia à entrada do terceiro milénio.

2.1 Modelo conceptual

Na figura seguinte, apresentamos o modelo conceptual de uma bases de dados destinada ao tratamento da maturidade dos Sistemas de Informação para a Economia, utilizando a simbologia e os conceitos constantes em McFadden (1994). São consideradas as entidades SIE, Critério, Grupo de Critério e Tempo e as associações Apresenta Valor e Tem Valor de Ponderação.

Pela simplicidade da estrutura e com o objectivo de não carregar este *diagrama Entidade-Associação* fizemos a sua representação sem atributos. Nas entidades SIE, Critério e Grupo de critério os atributos considerados foram código e designação, sendo o primeiro identificador. Em tempo integrou-se o atributo Ano também identificador. As associações possuem um atributo numérico relativo à concretização da propriedade que estão a ilustrar para o conjunto de entidades que as suporta.

Figura 52. Modelo Entidade Associação relativo à Maturidade dos SIE



Na descrição que se segue formulamos uma legenda para este modelo a qual representa os limites da sua utilização. A partir dela, deduz-se que a descrição da secção anterior relativa à maturidade dos sistemas de informação no tempo é susceptível de ser integrada e tratada.

Relacionamento *Apresenta Valor*

1. Um SIE apresenta valor relativamente a um critério em um ou vários momentos no tempo.

2. Em determinado momento, um SIE apresenta valor relativamente a um ou vários critérios.

3. Em determinado momento, um critério apresenta valores relativamente a um ou vários SIE.

Relacionamento *Critério Tem Valor de Ponderação*

1. Em determinado momento, existem um ou vários critérios com valor de ponderação.

2. Um critério apresenta valores de ponderação em um ou vários momentos no tempo.

Relacionamento *Pertence*

1. Um critério pertence a um e só um grupo de critério.

2. A um grupo de critério pertencem um ou vários critérios.

2.2. Modelo lógico

Para apresentação do modelo lógico relativo à maturidade dos Sistemas de informação para a Economia, utilizaremos as regras de derivação do modelo relacional constantes em McFadden (1994), designadamente as relativas a associações de cardinalidade tipo N:N e 1:N. Assim:

1. cada entidade origina uma relação ou, em linguagem menos rigorosa, uma tabela;

2. uma associação tipo N:N origina uma nova tabela cujo identificador provém dos identificadores das entidades participantes na associação;

3. uma associação tipo 1:N é representada no modelo relacional pela adição à tabela resultante da entidade do lado N do atributo identificador da entidade do lado 1.

Como resultado da aplicação destas regras ao modelo especificado na figura anterior, surge o seguinte *esquema relacional*:

SIE = {Código, Designação}
Critério = {Código, Designação, Código de Grupo}
Tempo = {Ano}
Grupo = {Código, Designação}
Valor do Critério no SIE = {Código do SIE, Código do Critério, Ano, Valor}
Ponderação do Critério = {Código de Critério, Ano, Valor}

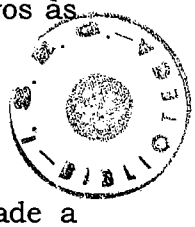
2.3. Implementação

A base de dados descrita pode ser implementada em qualquer sistema de gestão de base de dados da família relacional utilizando as facilidades que esse sistema proporcionar para a criação de tabelas, eventuais relacionamentos entre tabelas e controlo de integridade referencial e entrada de dados.

Utilizámos o Microsoft Access, perspectiva que tem em linha de conta as potencialidades a nível de elaboração de consultas simples e de acção e a exportação de valores para comunicação com outros produtos de software mais vocacionados para análises de natureza diferente ou mais complexas. Na parte final desta secção, mostramos para cada Grupo de Critério a pontuação considerada para os Critérios que o integram em termos dos SIE considerados.

Convém deixar aqui uma palavra em torno de aspectos da consolidação, matéria em que foi utilizado o valor de ponderação do critério no tempo. As variáveis de análise associadas aos SIE materializam-se na ausência ou presença de certas características materializadas nos critérios. Assim, ao nível da recolha original dos dados, a ausência traduz-se em 0 e a presença em 1 cabendo ao valor de ponderação do critério no tempo operar uma consolidação do valor a nível de grupo de critério o que se traduz num processo que

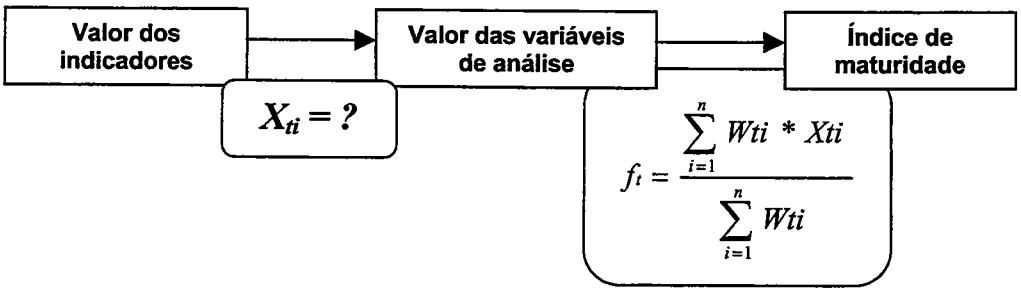
transforma os valores originais em valores numa escala de 1 a 5 relativos às variáveis de análise que são representadas pelos grupos de critério.



2.5. Do valor dos indicadores ao índice de maturidade

É necessário operacionalizar a determinação dos índices de maturidade a partir dos dados recolhidos o que passa, naturalmente, pela expressão do índice de maturidade no tempo. Mas a questão está em que ainda não se dispõe do valor das variáveis de análise. Esta secção tem por objectivo contribuir para esse passo. Esquematicamente, o processo será de acordo com a figura seguinte.

Figura 53. Esquema de determinação dos índices de maturidade



A resolução do problema deve levar em linha de conta que, neste caso concreto, o valor das variáveis de análise pode ser determinado de dois modos diferentes conforme o critério em causa.

No que respeita aos critérios associados às características da informação disponibilizada, há que proceder à distribuição do *portfolio* de dados e publicações disponibilizados por um SIE pelo intervalo real [0;5] em função de um conjunto de marcos predefinidos. O mesmo tipo de distribuição será utilizado no caso de *Diversidade*. Nos outros casos, o valor das variáveis de análise provém da contagem das características presentes no critério e da sua relação com as características possíveis no critério.

Para cada variável de análise associada às características da informação disponibilizada, há que considerar os p produtos (dados e publicações) de um um SIE e n marcos, onde $n \geq 2$, sendo o marco inicial 0 e o marco final o limite superior da escala, no nosso caso 5. Cada produto deve ser associado ao marco correspondente o que permite a criação de uma tabela de frequências. Para um SIE determinado, o valor de uma variável de análise associada às

características da informação poderá assim ser determinada de acordo com a expressão seguinte

$$X_i = \sum_{j=1}^{n} f_{ij} * M_j \qquad (65)$$

onde

i – ordem da variável de análise;

j – ordem do marco;

f_{ij} - representa a frequência relativa de produtos associados ao marco j no SIE para a variável de análise i;

Mj - um número real entre 0 e 5 associado ao marco.

Exemplo 16. Consideremos a linha relativa ao banco de Portugal no critério *Actualidade*. O vector tem o valor (1,0,0,0,0). Neste caso, M_1 é 5. Virá

Xi = 5*1 + 0 + 0 + 0 + 0 = 5

Quanto às variáveis de análise relacionadas com o sítio, há algumas particularidades que obrigam a uma explicitação suplementar. Assim:

1. Há que gerar valores compreendidos entre 0 e 5 onde o zero absoluto representa a não disponibilização interactiva de informação.
2. Há que não sobrevalorizar uma variável de análise pelo simples facto de corresponder à síntese de mais indicadores de existência.
3. O início da disponibilização interactiva pode ser valorizado por um factor arbitrário β cujo valor concreto dependerá das intenções do responsável pela análise, podendo considerar-se como um prémio para as organizações que optam por esta via de disponibilização de dados.

A expressão seguinte obedece a esses requisitos.

$$X_i = \frac{sc_{sup}((1-\beta)\sum_{j=1}^{n} V_{ij} + n\beta)}{n} \qquad (66)$$

onde

sc_{sup} corresponde ao limite superior da escala e

β corresponde ao peso da concretização da disponibilização interactiva no valor da variável de análise.

Exemplo 17. Consideremos a linha relativa ao Banco de Portugal no critério *Informação ao Analista*. O vector tem o valor (1,1,1,1,1,1,0,0,0,0). Neste caso, n é 10. Se utilizarmos um $\beta = 0.2$ e um limite superior igual a 5 virá

$$X_i = 5\ ((1\text{-}0{,}2)*6+10*0{,}2)/10=3.4$$

Passámos assim a dispor de um processo de medição da maturidade dos SIE.

2.4. Indicadores relativos à disponibilização interactiva de informação pelos SIE

Nesta secção, vamos proceder à avaliação dos SIE em termos de disponibilização de informação de acordo com um conjunto de critérios que têm a ver com a amigabilidade do serviço e as características da informação fornecida.

Os indicadores criados assentam em cadeias binárias onde 1 significa a presença da característica e 0 a sua ausência. Foram avaliados os dezasseis *sites* já referidos na secção anterior. Os resultados dessa avaliação relativamente a cada uma das características consideradas encontram-se em *Anexos* a este documento. Utilizando essas observações e os algoritmos descritos em pontos anteriores, foi-nos possível chegar a valores para a maturidade dos SIE relativamente aos anos de 1999, 2000 e 2001. O quadro seguinte resume esse cálculo efectuado para três grupos de ponderadores.

O grupo de ponderadores α significa uma opção relativamente ao que no momento actual os SIE poderiam privilegiar em termos de disponibilização, designadamente a actualidade, a multidimensionalidade, a disponibilização de dados e publicações, a capacidade de processamento e a integridade temporal.

O grupo de ponderadores β não pondera mais qualquer critério de análise, entendendo que todos têm o mesmo valor.

O grupo de ponderadores δ reflecte uma maior valorização da Abertura ao Analista e de critérios associados ao processamento analítico. Não é uma substancial modificação relativamente ao primeiro, mas inclui uma modificação importante para o critério referido.

Como se pode ver os resultados não são substancialmente diferentes. Em 1999, apenas o GEPE é prejudicado pela adopção do primeiro grupo de ponderadores. Logo que envereda pela disponibilização de dados em condições que o inserem melhor na era das redes, isso deixa de acontecer.

Por outro lado, a mudança no sentido da Abertura ao Analista ainda não foi tão grande que tivesse força para provocar uma alteração de posições, mesmo suportada por uma maior ponderação. Os SIE que ganharam vantagem no início do processo vão mantendo a sua posição.

Repare-se finalmente que qualquer dos grupos de ponderadores conseguiu captar alguma obsolescência a nível da DGO.

Quadro 49. Maturidade dos SIE (1999-2001)

Designação do SIE	*Ponderadores α*			*Ponderadores β*			*Ponderadores δ*		
	1999	2000	2001	1999	2000	2001	1999	2000	2001
BP	3.96	3.96	3.97	3.81	3.81	3.90	3.56	3.56	3.67
CCP	1.72	1.72	0.37	1.71	1.71	0.44	1.77	1.77	0.44
DETEFP	2.62	2.71	3.10	2.52	2.59	3.04	2.34	2.37	2.87
DGCC	2.42	2.42	3.13	2.26	2.26	3.01	2.26	2.26	2.94
DGCI	3.07	3.30	3.06	2.97	3.15	3.00	3.08	3.34	3.22
DGE	2.68	2.73	3.32	2.50	2.58	3.22	2.58	2.69	3.35
DGEP	4.14	4.20	3.80	3.75	3.84	3.52	3.80	3.92	3.51
DGO	3.66	3.23	3.73	3.39	3.05	3.51	3.40	2.82	3.47
DGS	2.25	2.25	2.50	2.11	2.11	2.42	2.34	2.34	2.61
DGT	2.00	2.73	3.36	1.92	2.52	3.16	1.82	2.41	3.15
DPP	2.66	2.95	3.22	2.51	2.88	3.16	2.40	2.75	2.88
GEPE	1.69	4.42	4.24	1.96	4.02	3.94	1.57	4.15	3.91
GEPMJ	2.85	2.82	2.77	2.80	2.76	2.62	2.68	2.62	2.63
ICEP	3.16	3.29	3.71	2.94	3.24	3.62	2.74	3.06	3.52
INE	4.05	4.10	3.83	3.99	4.03	3.96	3.92	3.94	3.68
OBSCOM	2.99	3.27	2.93	2.94	3.22	3.00	2.84	3.07	2.85

Os ponderadores utilizados são os que constam do quadro seguinte. Manteremos na parte final deste documento a utilização dos ponderadores α por os considerarmos os mais adequados ao momento.

Quadro 50.Ponderadores utilizados

Critério_Grupo	α	β	δ
Abertura ao Analista	2	1	5
Actualidade	5	1	3
Dados e Publicações	4	1	2
Diversidade	1	1	0
Informação ao Analista	2	1	5
Integridade Temporal	3	1	5
Multidimensionalidade	4	1	5
Navegação	2	1	2
Processamento	4	1	5

No quadro seguinte, detalhamos os valores da maturidade, em termos de critérios de análise, para o período 1999-2000.

Em termos de SIEs, é possível dizer que a grande transformação se operou ao nível do GEPE através da disponibilização de uma base de dados sobre o comércio externo e de algumas estatísticas adicionais que garantem alguma diversidade ao site. O DPP também apresentou modificações positivas em vários critérios. A prática destes dois SIEs contrasta com o monolitismo do INE e do Banco de Portugal onde não se detectou qualquer melhoria.

Em termos dos critérios de análise, a maior instabilidade parece associada à abertura ao analista, estando o que respeita a dados e publicações aparentemente saturado.

Quadro 51. A maturidade dos SIE por critério de análise (1999-2001)

	Abertura ao Analista			Actualidade			Dados e Publicações			Diversidade			Informação ao Analista			Integridade Temporal			Multidimensionalidade			Navegação			Processamento		
	1999	2000	2001	1999	2000	2001	1999	2000	2001	1999	2000	2001	1999	2000	2001	1999	2000	2001	1999	2000	2001	1999	2000	2001	1999	2000	2001
BP	1.73	1.73	2.09	5.00	5.00	4.00	4.69	4.69	5.00	4.00	4.00	4.00	3.40	3.40	3.00	4.00	4.00	5.00	3.00	3.00	3.00	4.43	4.43	5.00	4.00	4.00	4.00
CCP	1.36	1.36	nd	1.00	1.00	nd	2.85	2.85	nd	1.00	1.00	nd	2.60	2.60	nd	2.00	2.00	nd	2.00	2.00	nd	1.57	1.57	nd	1.00	1.00	nd
DETEFP	1.00	1.00	2.45	3.00	3.00	4.00	3.15	3.77	3.46	3.00	3.00	3.00	1.80	1.80	2.60	3.00	3.00	3.00	2.00	2.00	3.00	2.71	2.71	3.86	3.00	3.00	2.00
DGCC	1.73	1.73	1.73	3.00	3.00	3.00	2.85	2.85	4.08	2.00	2.00	3.00	2.20	2.20	2.60	2.00	2.00	4.00	3.00	3.00	4.00	1.57	1.57	2.71	2.00	2.00	2.00
DGCI	1.73	1.36	1.73	2.00	2.00	1.00	3.15	3.15	2.85	2.00	2.00	2.00	3.00	3.00	2.60	3.00	4.00	4.00	4.00	4.00	4.00	3.86	3.86	3.86	4.00	5.00	5.00
DGE	1.36	2.09	1.73	2.00	2.00	3.00	3.15	3.15	2.85	2.00	2.00	2.00	1.80	1.80	3.00	3.00	3.00	5.00	3.00	3.00	3.00	2.14	2.14	4.43	4.00	4.00	4.00
DGEP	1.73	1.73	2.09	5.00	5.00	5.00	4.69	4.69	3.15	3.00	3.00	3.00	2.60	3.40	2.60	5.00	5.00	3.00	4.00	4.00	2.00	2.71	2.71	3.86	5.00	5.00	7.00
DGO	2.09	2.45	2.09	5.00	5.00	5.00	3.46	4.69	2.85	3.00	3.00	3.00	2.20	2.60	2.20	5.00	3.00	4.00	2.00	2.00	4.00	2.71	2.71	4.43	5.00	2.00	4.00
DGS	1.00	1.00	1.73	1.00	1.00	1.00	2.85	2.85	3.46	1.00	1.00	2.00	2.60	2.60	2.60	3.00	3.00	4.00	3.00	3.00	3.00	1.57	1.57	1.00	3.00	3.00	3.00
DGT	1.36	1.73	1.73	2.00	3.00	4.00	3.15	4.08	2.85	2.00	2.00	3.00	2.20	2.20	2.60	1.00	1.00	3.00	3.00	3.00	3.00	1.57	2.71	3.29	1.00	3.00	5.00
DPP	1.00	1.73	1.73	4.00	4.00	5.00	2.85	3.46	3.46	3.00	3.00	4.00	2.20	3.00	3.00	4.00	4.00	4.00	1.00	1.00	2.00	1.57	2.71	3.29	3.00	3.00	2.00
GEPE	1.73	2.45	2.09	1.00	5.00	5.00	3.15	5.00	4.69	4.00	3.00	4.00	2.60	3.00	3.00	1.00	5.00	5.00	1.00	5.00	4.00	2.14	2.71	2.71	1.00	5.00	5.00
GEPMJ	1.73	1.36	2.09	3.00	3.00	3.00	3.77	3.77	3.77	3.00	3.00	2.00	3.00	3.00	2.60	3.00	3.00	3.00	3.00	3.00	3.00	2.71	2.71	2.14	2.00	2.00	2.00
ICEP	1.36	3.55	3.55	5.00	5.00	5.00	3.77	3.77	3.46	3.00	3.00	3.00	2.60	3.00	2.60	3.00	3.00	4.00	2.00	2.00	2.00	2.71	3.86	5.00	3.00	2.00	4.00
INE	2.09	2.09	3.55	4.00	4.00	4.00	3.77	4.08	4.08	5.00	5.00	5.00	3.80	3.80	3.00	5.00	5.00	5.00	5.00	5.00	2.00	3.29	3.29	5.00	4.00	4.00	4.00
OBSCOM	2.09	2.45	3.18	3.00	3.00	3.00	3.46	4.69	3.15	3.00	3.00	3.00	2.60	3.00	2.20	3.00	3.00	3.00	2.00	2.00	3.00	3.29	3.86	4.43	4.00	4.00	2.00

Nota: nd – valor não disponível (O site do CCP não estava activo na altura da consulta)

3. A maturidade dos SIE numa perspectiva difusa

Nesta secção vamos proceder a uma análise dos SIE numa perspectiva difusa utilizando a informação recolhida sobre 16 organismos produtores de informação para a Economia, tendo os dados sido recolhidos por observação directa do *site* durante o mês de Dezembro de 1999.

Os dados dizem respeito às variáveis *Navegação*, *Abertura ao Analista*, *Informação ao Analista*, *Multidimensionalidade*, *Disponibilização de Dados e Publicações*, *Integridade Temporal*, *Actualidade*, *Diversidade Temática* e *Processamento* que medem a *Maturidade* do SIE em cada um dos aspectos referidos.

Estas variáveis resultam da agregação de um conjunto de 66 critérios do tipo ausência/presença derivados das recomendações de autores relacionados com o processamento analítico interactivo: Nolan, Bhabuta, Codd, Inmon e Davenport, conforme explicitado em Vieira (1997).

A agregação gera a transformação dos dados binários nas variáveis referidas que passam a expressar-se numa escala com zero absoluto (não disponibilização) e marcos reais com significado intuitivo: por exemplo, 1 (nada importante, valor baixo), 5 (muito importante).

A informação recolhida vai agora ser sujeita ao tratamento seguinte:

1. Obtenção de uma árvore de decisão relativa à maturidade dos SIE a partir dos 9 atributos através processos analíticos lógicos.
2. Obtenção de uma classificação dos SIE em termos de maturidade a partir da árvore de decisão integrada num processo analítico difuso.
3. Comparação das classificações em termos de maturidade através de testes de diferenciação de médias.

Os produtos de software utilizados envolvem a utilização de Microsoft Access, KnowledgeSeeker da Angoss e FuzzyTech, representado na Europa por Inform GmbH.

Em *Anexos à dissertação*, descrevemos a extracção de um ficheiro em termos lógicos a partir da base de dados relacional com informação sobre os SIE, o qual é a base de todo o processamento.

3.1. Obtenção de árvore de decisão relativa à maturidade dos SIE

Nesta primeira fase foram utilizadas técnicas de classificação que permitem transformar uma base de dados num conjunto de regras derivadas dos seus registos a partir dum momento em que seja definida uma variável como dependente.

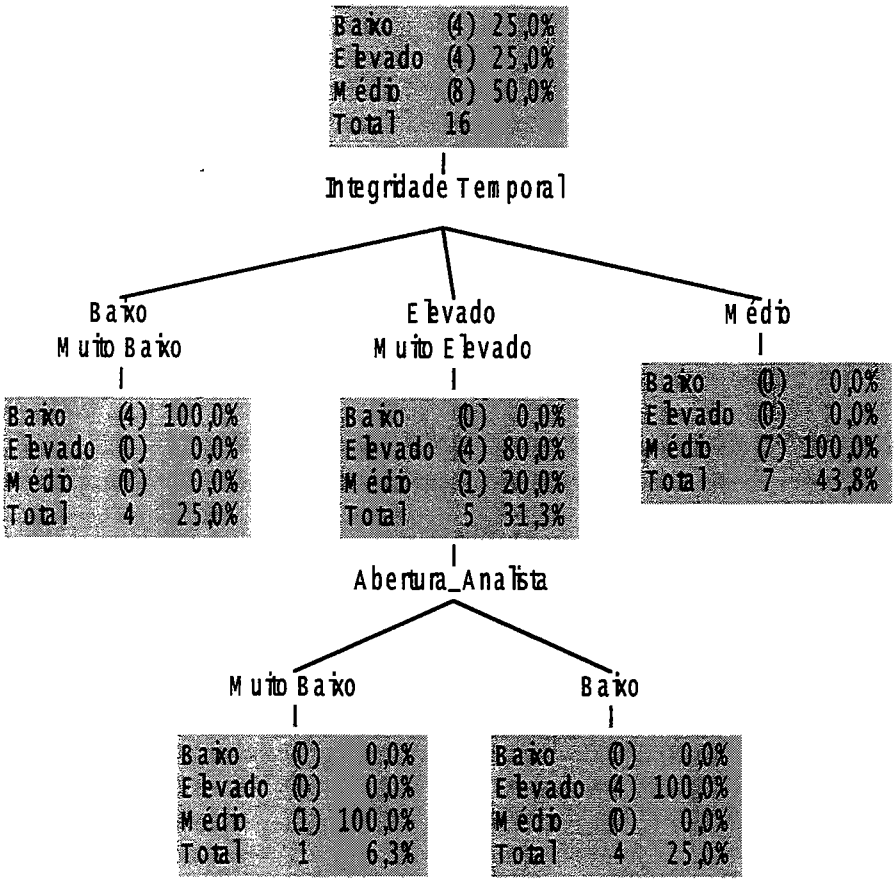
Os dados de entrada respeitantes às variáveis *Navegação, Abertura ao Analista, Informação ao Analista, Multidimensionalidade, Disponibilização de Dados e Publicações, Integridade Temporal, Actualidade, Diversidade Temática, Processamento e Maturidade* foram sujeitos a uma transformação qualitativa. Esta transformação e os dados que a originaram são descritos em *Anexos à dissertação.*

A extracção manual de uma árvore de decisão de uma base de dados descrita em Parsaye (1996) pode considerar-se uma actividade relativamente simples.

A utilização de produtos de software tem interesse em termos de aceleração do processo e pela capacidade de optimização e escolha dos caminhos que polarizam a distinção intrínseca ao sistema em análise.

Na figura seguinte apresentamos a árvore de decisão resultante do estudo em desenvolvimento.

Figura 54. Árvore de Decisão relativa à Maturidade dos SIE



De acordo com esta árvore, a ***Maturidade*** dos SIE pode ser considerada baixa em 25% de casos, elevada em 25% também e média nos restantes 50%. Um primeiro *split* foi considerado através da *Integridade Temporal* critério que associa *Maturidade Baixa* a 100% dos casos de *Integridade Temporal Baixa* ou *Média* e *Maturidade Média* a 100% dos casos de *Integridade Temporal Média*. O critério não permite a totalidade da classificação pois, no caso de a *Integridade Temporal* ser *elevada* ou *muito elevada* a *Maturidade* será em 80% dos casos *elevada* e *média* em 20% dos casos. Para continuar a classificação, houve que recorrer ao critério *Abertura ao Analista* que, se for *muito baixa* conduzirá a uma *Maturidade Média* e se for *baixa* conduzirá a uma *Maturidade Elevada*.

Isto é, utilizando dois critérios de análise (*Integridade Temporal* e *Abertura ao Analista*), é possível descrever a *Maturidade* dos SIE em Portugal no final de 1999.

Esta foi a classificação encontrada pelo *software* utilizado e considerada a mais interessante dentre todas as possibilidades que descobriu.

Teria sido possível forçar a construção da árvore de decisão através de outras variáveis. Um reflexo da não escolha da árvore óptima é o aumento do número de variáveis necessárias para uma classificação total como se vê em *Anexos à dissertação.*

3.2. A classificação de maturidade dos SIE numa perspectiva difusa

Para obtermos uma classificação suportada numa perspectiva difusa procedemos à derivação das regras a partir da árvore de decisão óptima obtida no ponto anterior as quais se resumem ao quadro seguinte:

Quadro 52. Regras Relativas à Maturidade dos SIE

RULE_1	
IF	Integridade Temporal = Baixo or Muito Baixo
THEN	Maturidade = Baixo 100,0%
	Maturidade = Elevado 0,0%
	Maturidade = Médio 0,0%
RULE_2	
IF	Abertura_Analista = Muito Baixo
	Integridade Temporal = Elevado or Muito Elevado
THEN	Maturidade = Baixo 0,0%
	Maturidade = Elevado 0,0%
	Maturidade = Médio 100,0%
RULE_3	
IF	Abertura_Analista = Baixo
	Integridade Temporal = Elevado or Muito Elevado
THEN	Maturidade = Baixo 0,0%
	Maturidade = Elevado 100,0%
	Maturidade = Médio 0,0%
RULE_4	
IF	Integridade Temporal = Médio
THEN	Maturidade = Baixo 0,0%
	Maturidade = Elevado 0,0%
	Maturidade = Médio 100,0%

Em seguida, integrámos as regras num programa de lógica difusa escrito em formato FTL (Fuzzy Technology Language) cuja listagem apresentamos em *Anexos à dissertação.* Uma descrição completa de FTL encontra-se em VonAltrock (1995).

Uma consequência imediata destas regras é que conseguiremos chegar a um valor relativo à maturidade de um SIE considerando no máximo aquelas duas variáveis de análise pelo que no programa desenvolvido não será necessário integrar outras variáveis de análise.

O modelo difuso poderá ser especificado pelo seguinte diagrama.

Figura 55. Modelo Conceptual do Sistema Difuso de maturidade dos SIE



O sistema considera duas variáveis linguísticas de input, *Integridade Temporal* e *Abertura ao Analista*, e uma variável de output, *Maturidade* do SIE, derivada das *Regras de Determinação de Maturidade* integradas no sistema difuso.

Estas variáveis assumem na origem valores contínuos entre 1 e 5, tendo como valores típicos os do quadro seguinte:

Quadro 53. Valores Típicos dos Termos

Nome	Valores típicos
Muito Baixa	1
Baixa	2
Média	3
Elevada	4
Muito Elevada	5

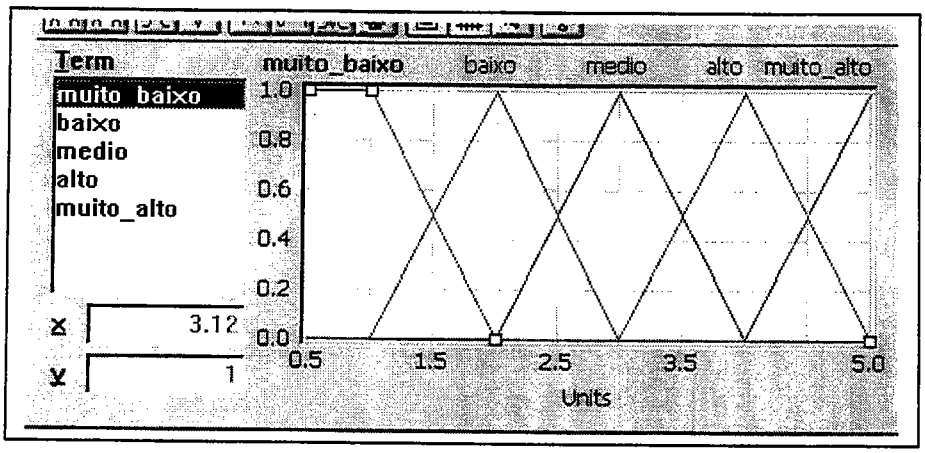
As funções de pertença consideradas serão do tipo triangular e são ilustradas na figura seguinte em termos gerais, isto é, considerámos o mesmo tipo para todas as variáveis. Apresentamos, de seguida, uma explicitação válida para a variável *Maturidade* e extensível às restantes.

- A *Maturidade* será considerada *muito baixa* com um grau de pertença igual a 1 para o valor 1, momento em que a pertença ao termo *baixa* passará a diminuir até se anular para um valor de 2.

- A *Maturidade* será considerada *baixa* com um grau de pertença de 1 para o valor 2, momento em que o grau de pertença começa a baixar, anulando-se a sua pertença ao termo para os valores inferiores a 1 e superiores 3, momento em que será considerada *média* com um grau de pertença de 1.

- De acordo com o descrito ela será *alta* e *muito alta* para os valores 4 e 5 podendo a derivação do grau de pertença ser feita como nos casos anteriores e de acordo com o gráfico apresentado e os valores típicos.

Quadro 54. Definição dos Termos



Os dados considerados para processamento são, apenas, os relativos às variáveis relevantes em termos de regras sendo apresentados no quadro seguinte.

Quadro 55. Dados Relativos aos SIE

	Abertura	Int_Temporal	Maturidade
BP	1.72	4	3.95
ICEP	1.36	2	1.72
DETEFP	1.00	3	2.61
DGCC	1.72	2	2.42
DGCI	1.72	3	3.06
DGE	1.36	3	2.67
DGEP	1.72	5	4.14
DGO	2.09	5	3.66
DGS	1.00	3	2.54
DGT	1.36	1	1.99
DPP	1.00	4	2.66
GEPE	1.72	1	1.68
GEPMJ	1.72	3	2.85
ICEP	1.36	3	3.16
INE	2.09	5	4.05
OBSCOM	2.09	3	2.99

Os resultados do processamento, neste caso a classificação difusa, serão apresentados em termos dos três métodos referidos para explicitação de resultados.

O primeiro quadro é o referente à *Explicitação Difusa* dos resultados. Neste caso, é visível a utilização dos termos *baixa, média* e *alta* para a *Maturidade.* Assim, o Banco de Portugal tem valores de pertença ao conjunto dos SIE de *Maturidade Media* na ordem dos 0.28 e de 0.72 ao conjunto dos SIE de *Maturidade Alta.* Estes valores derivam das observações referentes a *Abertura* e *Integridade Temporal.* Relativamente à *Abertura,* o grau de pertença associada a 1.72 é inferior a 1 levando a considerar a *Abertura muito baixa* com um grau de pertença na ordem dos 0.28 e *baixa* com um grau de pertença na ordem dos 0.72. A integridade temporal, com um valor de 4, associa um grau de pertença de 1 aos SIE de *Integridade Temporal Elevada.*

A Direcção Geral do Orçamento é um tanto penalizada pelo facto de o valor para a *Abertura* (2.09) já estar na fase descendente e não existir uma regra para a *Abertura Média* o que leva a perder algum conteúdo em termos de avaliação.

Quadro 56. Explicitação *Difusa* de Resultados

	Abertura	Int_Temporal	Maturidade baixa	Maturidade media	Maturidade alta	__flags__
BP	1.72	4	0	0.28002	0.71998	0
CCP	1.36	2	1	0	0	0
DETEFP	1.00	3	0	1	0	0
DGCC	1.72	2	1	0	0	0
DGCI	1.72	3	0	1	0	0
DGE	1.36	3	0	1	0	0
DGEP	1.72	5	0	0.28002	0.71998	0
DGO	2.09	5	0	0	0.90998	0
DGS	1.00	3	0	1	0	0
DGT	1.36	1	1	0	0	0
DPP	1.00	4	0	1	0	0
GEPE	1.72	1	1	0	0	0
GEPMJ	1.72	3	0	1	0	0
ICEP	1.36	3	0	1	0	0
INE	2.09	5	0	0	0.90998	0
OBSCOM	2.09	3	0	1	0	0

Os resultados relativos ao método de explicitação denominado *Melhor Compromisso* são evidenciados no quadro seguinte.

Os SIE que obtiveram um grau de pertença com o valor de 1 a um termo de maturidade recebem uma maturidade com valor típico desse termo. Os restantes são avaliados por uma ponderação entre o valor típico do termo e o grau de pertença. Assim o valor do Banco de Portugal explica-se por 0.28002 * 3 + 0.71998 * 4 o que conduz a um valor entre 3 e 4 mais perto do último.

O valor 0 para o campo *flag* significa que todos os casos foram avaliados utilizando um caminho fornecido pelas regras.

Quadro 57. Explicitação dos Resultados *Melhor Compromisso*

	Abertura	Integridade Temporal	Maturidade	Flags
BP	1.72	4	3.72	0
ICEP	1.36	2	2	0
DETEFP	1	3	3	0
DGCC	1.72	2	2	0
DGCI	1.72	3	3	0
DGE	1.36	3	3	0
DGEP	1.72	5	3.72	0
DGO	2.09	5	4	0
DGS	1	3	3	0
DGT	1.36	1	2	0
DPP	1	4	3	0
GEPE	1.72	1	2	0
GEPMJ	1.72	3	3	0
ICEP	1.36	3	3	0
INE	2.09	5	4	0
OBSCOM	2.09	3	3	0

Por fim, os resultados associados à *Solução Mais Plausível* são apresentados no quadro seguinte onde se torna evidente que resultam da atribuição do valor típico do termo com maior grau de pertença.

Quadro 58. Explicitação dos Resultados *Solução Mais Plausível*

	Abertura	Integridade Temporal	Maturidade	Flags
BP	1.72	4	4	0
ICEP	1.36	2	2	0
DETEFP	1	3	3	0
DGCC	1.72	2	2	0
DGCI	1.72	3	3	0
DGE	1.36	3	3	0
DGEP	1.72	5	4	0
DGO	2.09	5	4	0
DGS	1	3	3	0
DGT	1.36	1	2	0
DPP	1	4	3	0
GEPE	1.72	1	2	0
GEPMJ	1.72	3	3	0
ICEP	1.36	3	3	0
INE	2.09	5	4	0
OBSCOM	2.09	3	3	0

Admitindo-se que cada um dos métodos de explicitação terá os casos em que se revelará melhor a sua utilidade, pensamos que no caso presente os dois primeiros são os mais adequados.

Um resultado que informe que a maturidade de um SIE terá um determinado grau de pertença a determinada característica de maturidade é inteiramente aceitável. Mas qual será a legitimidade do segundo resultado? Será ele significativamente diferente do valor de maturidade de que partimos? Na próxima secção tentamos uma resposta a esta questão.

3.3. Valor do melhor compromisso para avaliação da maturidade

Será que os resultados a que chegamos com a explicitação tipo *Melhor Compromisso* são significativamente diferentes dos valores iniciais da *Maturidade*?

Se isso acontecer temos de concluir que a árvore de decisão gerada não terá qualquer utilidade para avaliação dos SIE pois conduz-nos a resultados incorrectos. Mas, se os resultados não forem significativamente diferentes dos originais no curto prazo, enquanto não houver substanciais modificações nos SIE poderemos analisar a sua evolução através do exame do que acontece com

as variáveis *Integridade Temporal* e *Abertura*, isto é temos uma redução no número de variáveis intervenientes na análise.

Procedemos a um teste de diferença de médias de valores emparelhados que nos conduziu aos resultados do quadro seguinte.

Quadro 59. Explicitação dos Resultados Melhor Compromisso

	Variable 1	*Variable 2*
Mean	2.965	2.884375
Variance	0.467813333	0.593572917
Observations	16	16
Pearson Correlation	0.928612281	
Hypothesized Mean Difference	0	
df	15	
t Stat	1.121354813	
P(T<=t) one-tail	0.139884159	
t Critical one-tail	1.753051038	
P(T<=t) two-tail	0.279768318	
t Critical two-tail	2.131450856	

Considerando um grau de confiança na ordem dos 95%, vemos que o valor para a *estatística t* cai numa zona de aceitação, pelo que aceitamos a hipótese de não haver diferenças significativas entre os valores de que dispúnhamos para a *Maturidade*. Isto significa que, *no curto prazo, é possível utilizar apenas as variáveis associadas ao corpo de regras para avaliação de Maturidade dos SIE.*

3.4. A maturidade dos SIE no ano 2000 com base no modelo difuso desenvolvido para 1999

Nesta secção, procedemos à utilização do modelo gerado e implementado em sistema difuso, considerando os dados relativos aos SIE em 2000. Esta é uma utilização legítima dos modelos de classificação: são desenvolvidos para um conjunto de exemplos (em que considerámos os dados dos SIE para 1999) e são aplicados a nova informação para a qual os valores a predizer se apresentam desconhecidos.

Quadro 60. Predição Difusa de Maturidade dos SIE em 2000

Designação do SIE	Abertura ao Analista	Integridade Temporal	Maturidade Difusa
BP	1.73	4.00	4.16
CCP	1.36	2.00	1.75
DETEFP	1.00	3.00	3.25
DGCC	1.73	2.00	1.75
DGCI	1.36	4.00	3.70
DGE	2.09	3.00	3.25
DGEP	1.73	5.00	4.16
DGO	2.45	3.00	3.25
DGS	1.00	3.00	3.25
DGT	1.73	1.00	1.75
DPP	1.73	4.00	4.16
GEPE	2.45	5.00	4.50
GEPMJ	1.36	3.00	3.25
ICEP	3.55	3.00	3.25
INE	2.09	5.00	4.50
OBSCOM	2.45	3.00	3.25

Comparando os valores relativos a 1999 com os agora calculados para a maturidade, constata-se a enorme subida do GEPE e do DPP em termos da qualidade da informação disponibilizada para a Economia. Este facto está de acordo com a realidade o que significa que o modelo desenvolvido conseguiu percepcionar o fenómeno, pelo menos, relativamente ao GEPE.

Podemos interrogar-nos sobre até que ponto é que o modelo desenvolvido captou a realidade de 2000. Para responder a essa questão, é possível compararmos os valores de maturidade apurados através da métrica de maturidade definida inicialmente com o actual. Os resultados dessa comparação surgem no quadro seguinte.

Quadro 61. A Maturidade dos SIE em 2000 de acordo com Métrica e Modelo

SIE	Maturidade	Maturidade Difusa	Correcção no Termo	Variação Absoluta	Variação Relativa
BP	3.96	4.16	S	0.20	5.00%
CCP	1.72	1.75	S	0.03	0.75%
DETEFP	2.71	3.25	S	0.54	13.50%
DGCC	2.42	1.75	S	-0.67	-16.75%
DGCI	3.30	3.70	N	0.40	10.00%
DGE	2.73	3.25	S	0.52	13.00%
DGEP	4.20	4.16	S	-0.04	-1.00%
DGO	3.23	3.25	S	0.02	0.50%
DGS	2.25	3.25	N	1.00	25.00%
DGT	2.73	1.75	N	-0.98	-24.50%
DPP	2.95	4.16	N	1.21	30.25%
GEPE	4.42	4.50	S*	0.08	2.00%
GEPMJ	2.82	3.25	S	0.43	10.75%
ICEP	3.29	3.25	S	-0.04	-1.00%
INE	4.10	4.50	S*	0.40	10.00%
OBSCOM	3.27	3.25	S	-0.02	-0.50%

Constata-se a existência de diferenças que, em termos relativos ao intervalo de variação, são importantes, o que é expresso na última coluna. Uma comparação quanto aos termos originados pelo modelo difuso, levaria à detecção de quatro casos em que os termos provenientes da predição não estariam de acordo com os decorrentes da métrica, havendo ainda que considerar dois casos em que a proximidade entre valores recomendaria uma consideração de igualdade, mas onde os termos considerados poderiam ser tidos com diferentes por estarem numa zona de fronteira

Apesar destas diferenças, o teste de médias emparelhadas, cujos resultados surgem no quadro seguinte, sugere a não existência de diferenças significativas a um nível de 95% entre os valores para a maturidade de acordo com a expressão original e os decorrentes do modelo difuso.

Quadro 62. Resultados de teste de diferença de médias

	Variable 1	*Variable 2*
Mean	3.131656847	3.3242125
Variance	0.561059451	0.832483922
Observations	16	16
Pearson Correlation	0.805057621	
Hypothesized Mean Difference	0	
df	15	
t Stat	-1.422570745	
P(T<=t) one-tail	0.087661691	
t Critical one-tail	1.753051038	
P(T<=t) two-tail	0.175323382	
t Critical two-tail	2.131450856	

3.5. Avaliação da correcção do modelo

Apesar do resultado anterior, é possível argumentar que a igualdade de médias pode ser obtida a partir de valores que, não correspondendo a diferenças significativas em termos estatísticos, parecem diferentes ao observador comum. Um teste mais profundo é o que deriva do que descrevemos inicialmente sobre avaliação. Vamos procurar, assim, determinar a medida de correcção global e por classe.

Em termos globais, partimos do quadro seguinte. Na diagonal principal, estão os acertos; os outros elementos correspondem ao total de elementos de determinada classe, avaliada pela métrica de maturidade, que foram considerados em outra classe. Adicionando os elementos da diagonal principal e relacionando-os com o total, chegamos a um valor global para a medida de acertos na ordem dos 10/16 ou 62.5%.

Quadro 63. Frequências relativas a Maturidade e Predição

Maturidade	Maturidade Difusa (predição)				
	2	3	4	5	Total
2	2	1			3
3	1	6	2		9
4			2	2	4
Total	3	7	4	2	16

Assim, se se procura uma predição muito rigorosa, superior a 90%, é de rejeitar o modelo e propor a sua reelaboração ou reavaliação. Se não for tão importante o rigor e for satisfatória uma taxa de acertos superior a 50%, conclui-se que o modelo elaborado para 1999 é aplicável em 2000, os seus resultados são válidos nesse contexto e a avaliação para 2001 iniciar-se-á ainda com esse resultado.

Podemos, em seguida, examinar o que se passou relativamente a cada classe.

Na classe 1, a taxa de positivos falsos é 0 o que significa que, não existindo a classe nos dados concretos em análise, ela nunca foi utilizada na predição.

Quadro 64. Matriz de Acertos e Erros para a Classe 1

	Previsão	
Classe	**1**	**^1**
1	0	0
^1	0	16

True Positive Rate	não definido (nd)
False Positive Rate	0
Precision	(nd)
Recall	(nd)

Na classe 2, há uma aproximação aos valores globais apresentados no início desta secção. A precisão é apenas de 66,6% e as considerações que fizemos inicialmente permanecem válidas.

Quadro 65. Matriz de Acertos e Erros para a Classe 2

	Previsão	
Classe	**2**	**^2**
2	2	1
^2	1	12

True Positive Rate	66.666667
False Positive Rate	7.6923077
Precision	66.666667
Recall	66.666667

A classe 3 vê a precisão subir para os 85,7% o que mostra o grau de satisfação do modelo para uma classe intermédia.

Quadro 66. Matriz de Acertos e Erros para a Classe 3

	Previsão	
Classe	**3**	**^3**
3	6	3
^3	1	6

True Positive Rate	66.666667
False Positive Rate	14.285714
Precision	85.714286
Recall	66.666667

É ao nível das classes 4 e 5 que o modelo se mostra mais desajustado. A precisão baixa até se anular completamente ao nível da classe 5.

Quadro 67. Matriz de Acertos e Erros para a Classe 4

	Previsão	
Classe	**4**	**^4**
4	2	2
^4	2	10

True Positive Rate	50
False Positive Rate	16.666667
Precision	50
Recall	50

O próximo quadro mostra os resultados para a classe 5, classe para a qual não existiam valores derivados da métrica de avaliação, mas que foi utilizada efectivamente na predição difusa.

Quadro 68. Matriz de Acertos e Erros para a Classe 5

	Previsão	
Classe	**5**	**^5**
5	0	0
^5	2	14

True Positive Rate	(nd)
False Positive Rate	12.5
Precision	0
Recall	(nd)

É possível pensar em mudar de árvore quando a precisão for inferior a um λ previamente especificado.

3.6. Proposta de um processo de análise de maturidade dos SIE em perspectiva difusa

O caminho percorrido ao longo deste documento permite-nos, demonstrada que está a validade dos resultados, propor um processo de análise de maturidade dos SIE em termos difusos que integra os elementos anteriores e garante a produção de resultados com um mínimo de variáveis explicativas.

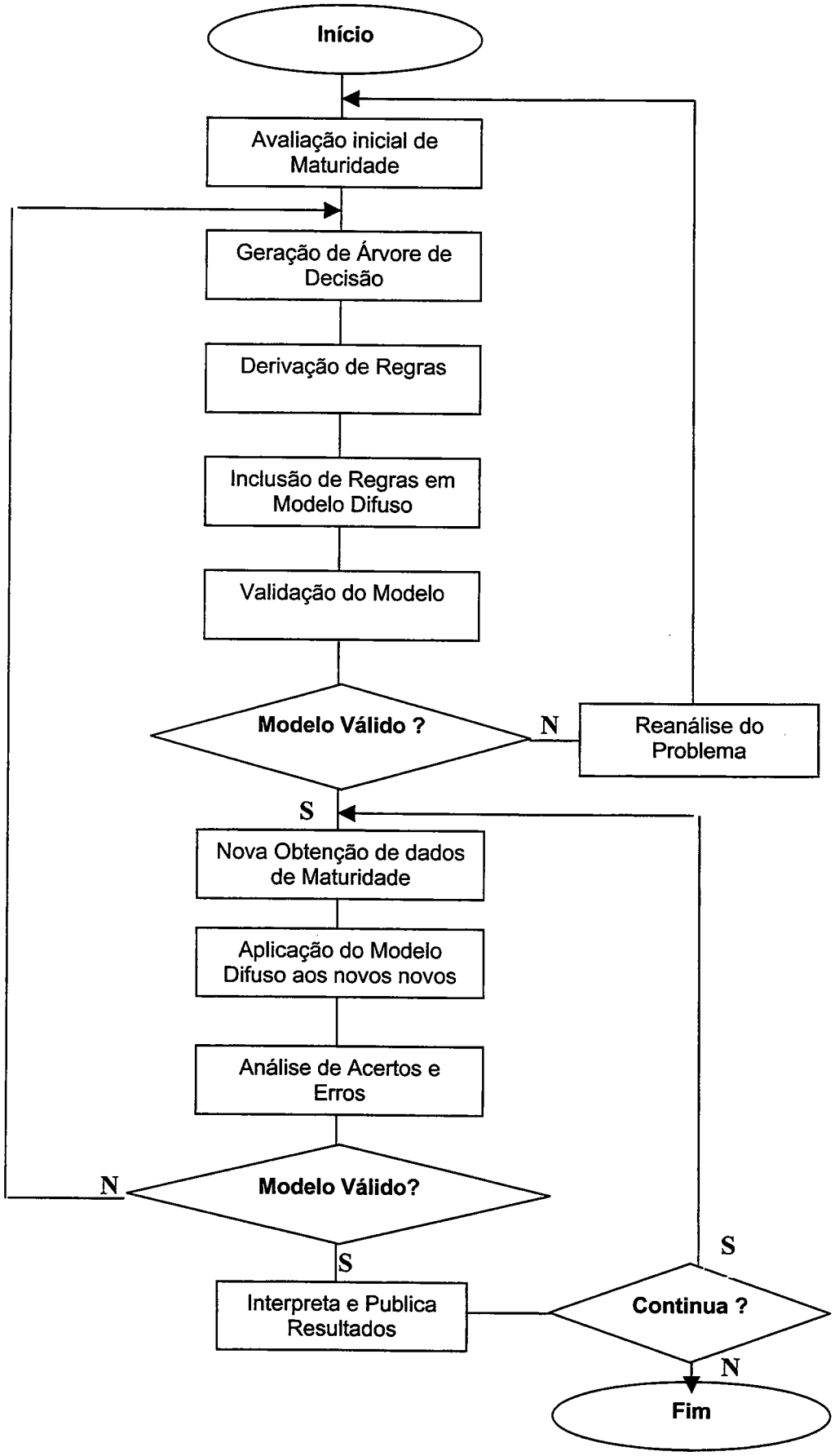
Este processo é apresentado no diagrama constante na figura seguinte. Tudo se inicia por uma primeira avaliação de maturidade que tem por base os valores observados para as variáveis de análise e uma métrica de maturidade derivada fundamentalmente da sensibilidade do analista ao problema. Com o conjunto de valores assim gerados e a sua tradução em variáveis qualitativas, é possível gerar uma árvore de decisão que, a existir, e a serem utilizados os optimizadores associados aos algoritmos disponíveis permitirá encontrar o

melhor conjunto de atributos que pelo assumir de determinados valores explicará o valor existente para a maturidade.

A partir da árvore, podem ser deduzidas regras que conferem ao modelo a possibilidade de ser integrado com linguagens de programação ou com comandos de linguagens de interrogação. No caso presente, estas regras são incluídas num programa de lógica difusa podendo ser utilizadas como processo analítico difuso nesse contexto. Neste momento, torna-se importante a validação do modelo difuso gerado pois a modificação operada pode distorcer resultados. Assim, com teste conveniente é possível determinar se o modelo permanece válido ou não. Em caso negativo, há que reanalisar o problema em ordem a operar qualquer modificação que possa conduzir a um modelo válido. Em caso afirmativo, o modelo pode ser aplicado a dados que não contribuíram para ele. Por exemplo, se o modelo foi elaborado a partir de dados de 1999 pode ser aplicado aos dados das variáveis determinantes em 2000. Há assim que esperar por novos dados e, quando estes existirem proceder à aplicação do modelo difuso.

Esta aplicação conduzirá a resultados em termos de maturidade que, se se dispuser da totalidade de observações para a nova informação, permitirá uma análise de acertos e erros a qual vem de novo pôr em causa a validade do modelo agora através de um confronto entre realidade e predição. Se o modelo já não for válido, há que proceder à geração de nova árvore de decisão e o processo repete-se a partir daí. Pelo contrário, se o for, há que interpretar e publicar resultados e decidir sobre a continuação de aplicação do modelo com novos dados. Em caso negativo, cessa a utilização deste processo, em caso afirmativo há que voltar à recolha de dados, reavaliação e análise de acertos, durante tantos anos quantos aqueles em que se considere o processo como sendo útil.

Figura 11 – Conjunto de passos para aplicação de uma perspectiva difusa a dados sobre a maturidade dos SIE



Resumo do Capítulo V

Neste capítulo, apresentámos várias métricas para análise da maturidade e procedemos à sua utilização nos sistemas de informação para a Economia em Portugal no período compreendido entre 1999 e 2001.

A primeira opção ao nível de métricas é a consideração da maturidade como o resultado de somas ponderadas. A simplicidade associada a este processo pode ser alterada pela consideração de uma interrogação do tipo: *com a soma ponderada encontramos um número, como sabemos se estaremos ou não perante um estádio de maturidade?*. Esta interrogação levou-nos a avançar para uma métrica de segundo tipo que pode utilizar várias funções para chegar a uma conclusão: a função limiar de delimitadores fixos, a função linear e a função *Sigmoid*, sendo esta última uma função de transformação não linear.

Mas a métrica que nesta fase mais nos interessou foi a esboçada para o tratamento da maturidade no tempo: sendo a maturidade um processo de evolução, há que ter meios para a seguir temporalmente. A consideração de somas ponderadas no tempo conduziu-nos rapidamente ao problema da transformação dos ponderadores: a inexistência desta transformação levaria a que SIEs bem classificados em determinado momento mantivessem eternamente esse estatuto. Pusemos assim a hipótese de os ponderadores seguirem um ciclo de vida do tipo: surgimento, crescimento, maturidade, declínio e desaparecimento. Esta hipótese parece-nos de acordo com as contribuições de Nolan e Earl apresentadas no Capítulo I deste documento. Uma simulação da evolução de um SIE sem modificação em variáveis de análise e com modificação de ponderadores gerados de acordo com a hipótese anterior permitiu visualizar graficamente o declínio da sua maturidade no tempo, mesmo em situações em que havia acções de rejuvenescimento não sustentadas.

Assumindo que, na era das redes, o grau zero de maturidade é aplicável aos SIE que não disponibilizam interactivamente informação, considerou-se que este processo seria correctamente acompanhado através das seguintes variáveis de análise: actualidade, informação ao analista, características da navegação, características do processo de disponibilização de dados e publicações, abertura ao analista, multidimensionalidade dos dados,

integridade temporal, capacidade de processamento pelo analista e diversidade. Foi assim constituída uma base de dados relativa à maturidade dos SIE, baseada na satisfação de determinados requisitos associadas às variáveis de análise referidas e definida a transformação da presença ou ausência desses requisitos em valores das variáveis de análise do processo de maturidade.

O capítulo terminou com a apresentação de uma metodologia destinada à implementação de uma métrica de predição. Foi elaborado um modelo difuso de predição com os dados de 1999 que, em 2000, foi testado com resultados satisfatórios. Neste ano, o modelo conseguiu captar importantes modificações no processo de disponibilização interactiva de informação sobre a economia, designadamente, as protagonizadas pelo GEPE, com o início da disponibilização interactiva de dados, acompanhado pela utilização de uma ferramenta multidimensional, e pela expansão da informação proveniente do DPP.

Capítulo VI - Estádios de desenvolvimento dos Sistemas de Informação para a Economia na era das redes

Chegamos assim ao capítulo decisivo deste documento, aquele que nos vai permitir apresentar uma visão sobre o que consideramos relevante em termos de modelo de análise dos SIE.

Não sendo muito conclusiva a análise anterior em termos de maturidade dos SIE, propomo-nos procurar estádios de evolução neste objecto de estudo, analisando os SIE numa perspectiva multivariada e utilizando a informação recolhida nos capítulos anteriores.

Este estudo terá como universo os 16 organismos anteriormente referidos, produtores de informação para a Economia, tendo os dados sido recolhidos por observação directa do sítio durante o mês de Dezembro de 1999 utilizando uma escala com zero absoluto e valores reais típicos entre 1, de significado *nada importante* ou *valor muito baixo,* e 5, de significado *muito importante* ou *valor muito alto.* Observações relativas a 2000 e 2001 servirão para refinar conclusões.

A informação recolhida, já descrita, vai ser sujeita a um tratamento que envolve:

- a identificação das dimensões de análise dos SIE a partir dos 9 atributos através de técnicas de análise factorial;
- a identificação dos segmentos de informação para a Economia através de técnicas de análise de clusters;
- a diferenciação dos segmentos de informação para a Economia através de testes de diferenciação de médias.

O software utilizado para a produção dos resultados é o produto SPSS para Windows (versões 9.0 e 10.0).

O capítulo integra ainda uma discussão acerca da viabilidade de concretização de um estádio de desenvolvimento, procurando, através de um paralelismo

com contribuições para a epistemologia da ciência fundadas em Gaston Bachelard, formular uma recomendação para o alcançar.

1. Identificação da dimensões de análise da maturidade dos SIE

A identificação das dimensões de análise de maturidade dos SIE foi realizada utilizando a Análise Factorial pelo método das Componentes Principais (AFCP), método estatístico multivariado que permite transformar um conjunto inicial de variáveis, correlacionadas entre si, num outro conjunto de variáveis não correlacionadas, as chamadas componentes principais que resultam de combinações lineares do conjunto inicial. Como esta análise nem sempre é uma boa opção, há que realizar alguns testes para julgar da sua adequabilidade ao problema em causa.

1.1. Adequação da Análise Factorial

Num primeiro momento, para a percepção da viabilidade de uma análise factorial, torna-se conveniente proceder a um teste à normalidade das nove variáveis. Trata-se de uma soma de testes de normalidade univariada que não implica normalidade multivariada, mas, só por si, dá-nos algumas indicações sobre a eventualidade de transformação dos dados. Os resultados apresentam-se no quadro seguinte.

Quadro 69. Testes de Normalidade

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Abertura	.227	16	.027	.888	16	.051
Actualidade	.154	16	.200	.894	16	.068
Dados e Publicações	.238	16	.016	.828	16	.010
Diversidade	.220	16	.038	.919	16	.209
Informação ao Analista	.231	16	.022	.921	16	.226
Integridade Temporal	.227	16	.027	.902	16	.089
Multidimensionalidade	.198	16	.092	.923	16	.246
Navegação	.176	16	.198	.896	16	.073
Processamento	.216	16	.045	.891	16	.061

Utilizando um nível de significância de 0.05, apenas *Actualidade*, *Multidimensionalidade* e *Navegação* apresentam uma distribuição normal.

Note-se, no entanto, que, em termos de análise factorial, o pressuposto de normalidade não inviabiliza a execução da análise quando não se verifica. Isto significa que, se outros testes forem satisfatórios poderemos pensar em prosseguir com os dados tal como estão. Esta atitude pode ser reforçada pela seguinte constatação: se baixássemos o nível de significância do teste para 0.01, já poderíamos considerar normais as distribuições das variáveis.

O quadro seguinte apresenta algumas estatísticas descritivas associadas às variáveis, sendo de reter o valor obtido para a *Abertura ao Analista* excessivamente baixo na escala utilizada. O baixo valor para o desvio padrão induz que esta realidade se estende a muitos SIE o que traduz desde logo a necessidade de acções que conduzam à transformação dos elementos que contribuem para esta variável de análise.

Quadro 70. Estatísticas Descritivas

	Média	Desvio Padrão	Coeficiente de variação
Abertura	1.5682	.3748	0.24
Actualidade	3.0625	1.4818	0.48
Dados e Publicações	3.4231	.5932	0.17
Diversidade	2.7500	1.0646	0.39
Informação ao Analista	2.5750	.5360	0.21
Integridade Temporal	3.1250	1.2583	0.40
Multidimensionalidade	2.6875	1.0782	0.40
Navegação	2.5357	.8783	0.35
Processamento	3.0625	1.3401	0.44

Nota: número de sites observados: 16

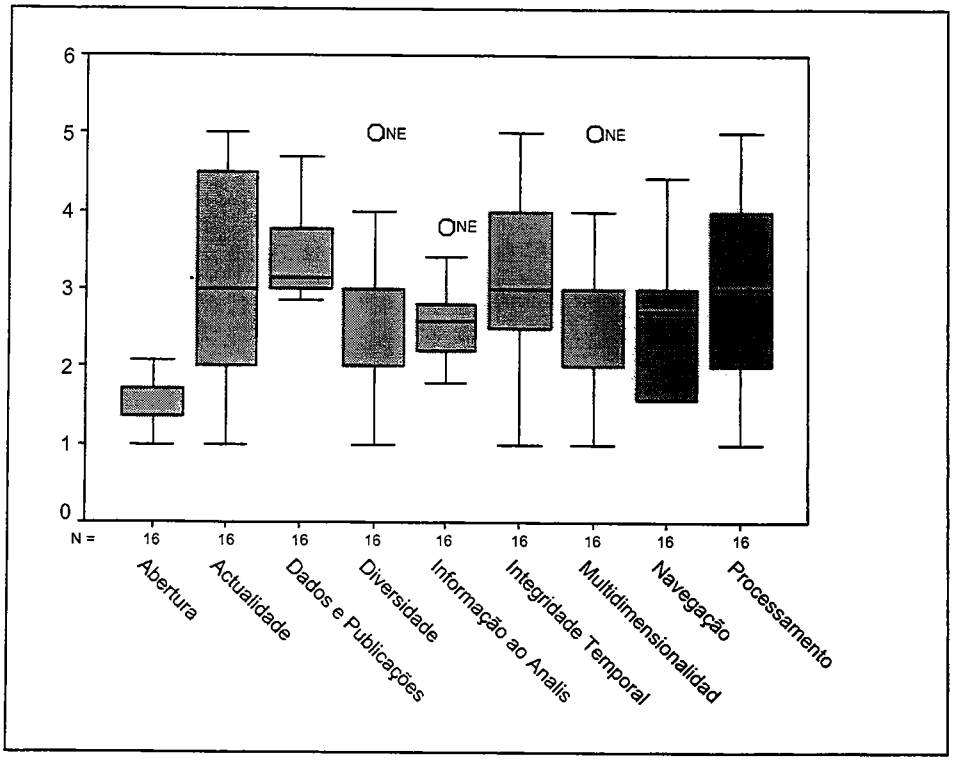
Também a análise das *caixas de bigodes* permite constatar o valor excessivamente baixo da *Abertura ao Analista* e a pouca variabilidade no conjunto dos SIE. Pelo contrário, a variável *Actualidade* cujo valor médio se situa sobre o ponto médio da variação a nível de escala utilizada apresenta uma dispersão relativamente elevada no que é acompanhada por *Processamento* e *Integridade Temporal.* Este facto já era observável no quadro anterior, manifestando-se num coeficiente de variação (relação desvio padrão / média) mais elevado.

As *caixas de bigodes* evidenciam ainda a existência de *outliers* com valores elevados do INE em termos de *Diversidade, Informação ao Analista* e

Multidimensionalidade. Apesar disso, manteremos estas variáveis na análise pois o seu valor não nos parece excessivamente elevado: aqui, o problema não estará no INE, mas nos outros SIE.

É de realçar, finalmente, o baixo nível de simetria associado às diferentes variáveis o que leva mais uma vez a interrogarmo-nos se deveremos avançar com uma AFCP.

Figura 56. Caixas de Bigodes Relativas a Variáveis de Análise



A matriz de correlações, apresentada no quadro seguinte, mede a associação linear entre as variáveis através do coeficiente de correlação de Pearson.

A aplicação do modelo factorial exige uma elevada correlação. Não se pode afirmar que esse é o caso, mas, dada a existência de múltiplas correlações na ordem dos 0.5 ou superior, como é o caso da correlação entre *Processamento* e *Integridade Temporal* que atinge o valor de 0.8, também não é possível concluir sobre a sua ausência pelo que nos basearemos noutros critérios para julgarmos a justeza da análise factorial.

Na parte inferior da matriz de correlações, são apresentados valores para o seu nível de significância. A existência de múltiplos valores inferiores a 0.05 em correlações superiores a 0.4 leva-nos a considerá-las significativas.

Quadro 71. Matriz de Correlação

		Abertura	Actualidade	Dados e Publicações	Diversidade	Informação ao Analista	Integridade Temporal	Multidimensionalidade	Navegação	Processamento
Correlation	Abertura	1.000	.325	.440	.501	.510	.302	.349	.539	.196
	Actualidade	.325	1.000	.703	.560	.204	.746	.138	.477	.468
	Dados e Publicações	.440	.703	1.000	.568	.513	.557	.397	.683	.398
	Diversidade	.501	.560	.568	1.000	.502	.423	.102	.560	.121
	Informação ao Analista	.510	.204	.513	.502	1.000	.321	.539	.605	.146
	Integridade Temporal	.302	.746	.557	.423	.321	1.000	.375	.470	.804
	Multidimensionalidade	.349	.138	.397	.102	.539	.375	1.000	.380	.388
	Navegação	.539	.477	.683	.560	.605	.470	.380	1.000	.455
	Processamento	.196	.468	.398	.121	.146	.804	.388	.455	1.000
Sig. (1-tailed)	Abertura		.110	.044	.024	.022	.128	.093	.016	.234
	Actualidade	.110		.001	.012	.225	.000	.305	.031	.034
	Dados e Publicações	.044	.001		.011	.021	.013	.064	.002	.064
	Diversidade	.024	.012	.011		.024	.051	.354	.012	.328
	Informação ao Analista	.022	.225	.021	.024		.113	.016	.007	.294
	Integridade Temporal	.128	.000	.013	.051	.113		.076	.033	.000
	Multidimensionalidade	.093	.305	.064	.354	.016	.076		.073	.069
	Navegação	.016	.031	.002	.012	.007	.033	.073		.038
	Processamento	.234	.034	.064	.328	.294	.000	.069	.038	

O teste de esfericidade de Bartlett pode ser usado para testar a hipótese da matriz das correlações ser a matriz identidade com determinante igual a 1. De acordo com o quadro seguinte, este teste tem associado um nível de significância de 0.000 que, por ser inferior a 0.05, leva à rejeição da hipótese da matriz das correlações na população ser a identidade, mostrando que existe correlação entre as variáveis que, se não se verificasse, deveria conduzir a uma reanálise da utilização do modelo factorial.

Outro indicador, o Kaiser-Meier-Olkin (KMO), compara as correlações simples com as parciais observadas entre as variáveis. O KMO perto de 1 indica coeficientes de correlação parciais pequenos. Estes devem aproximar-se de zero quando as suposições de análise factorial se verificam. O valor encontrado, 0.709, é um valor considerado razoável para prosseguir com a análise factorial.

Quadro 72. KMO e Teste de Bartlett

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.709
Bartlett's Test of Sphericity	Approx. Chi-Square	69.465
	df	36
	Sig.	.001

A matriz anti-imagem é uma medida da adequação de cada variável para uso da análise factorial, onde pequenos valores na diagonal principal levam a considerar a eliminação da variável. As medidas de adequação das dados encontram-se no quadro seguinte na diagonal da matriz e quanto maiores forem melhor resulta a análise factorial. Como se pode observar, nenhuma variável tem, na diagonal principal, um valor perto de zero pelo que é de considerar todo o conjunto de variáveis para a ACP.

Quadro 73. Matriz Anti-imagem

Anti-image Matrices

		Abertura	Actualidade	Dados e Publicações	Diversidade	Informação ao Analista	Integridade Temporal	Multidimensionalidade	Navegação	Processamento
Anti-image Correlation	Abertura	**.731**	-.156	6.362E-02	-.308	-.352	.358	-5.509E-02	.139	-.366
	Actualidade	-.156	**.717**	-.537	-.181	.378	-.515	.143	-.148	.241
	Dados e Publicações	6.362E-02	-.537	**.852**	-7.096E-02	-.148	.120	-.232	-.207	-3.370E-02
	Diversidade	-.308	-.181	-7.096E-02	**.821**	-4.501E-02	-.225	.222	-.268	.264
	Informação ao Analista	-.352	.378	-.148	-4.501E-02	**.457**	-.675	-.439	-.653	.734
	Integridade Temporal	.358	-.515	.120	-.225	-.675	**.531**	9.019E-02	.618	-.880
	Multidimensionalidade	-5.509E-02	.143	-.232	.222	-.439	9.019E-02	**.693**	.223	-.281
	Navegação	.139	-.148	-.207	-.268	-.653	.618	.223	**.603**	-.744
	Processamento	-.366	.241	-3.370E-02	.264	.734	-.880	-.281	-.744	**.494**

a Measures of Sampling Adequacy(MSA)

Comunalidade é o termo utilizado para designar a variância de cada variável explicada pelos factores comuns. As comunalidades costumam ser apresentadas antes e depois de extracção do número de factores considerado correcto, sendo as iniciais iguais a 1. As comunalidades finais apresentam valores entre 0 e 1 o que está associado ao grau de explicação das variáveis pelos factores. No caso presente, vê-se que apenas a variável *Abertura ao Analista* apresenta um valor relativamente baixo o qual não é, no entanto, suficiente para a excluirmos da análise.

Quadro 74. Comunalidades

	Initial	Extraction
Abertura	1.000	.580
Actualidade	1.000	.856
Dados e Publicações	1.000	.717
Diversidade	1.000	.827
Informação ao Analista	1.000	.819
Integridade Temporal	1.000	.878
Multidimensionalidade	1.000	.884
Navegação	1.000	.713
Processamento	1.000	.891

Extraction Method: Principal Component Analysis.

Em resumo, podemos afirmar que, apesar de existirem alguns elementos que levem a desconfiar sobre a qualidade dos dados em termos de uma adequada Análise Factorial, indicadores mais específicos como o teste de esfericidade de Bartlett, o Keiser-Meier-Olkin, a matriz anti-imagem e o valor das comunalidades levam-nos a considerar que será correcto avançar com a análise que nos propusemos.

É, no entanto, de realçar desde já algum défice explicativo em termos da variável *Abertura* que nos leva a uma primeira conclusão: sem excepção, os SIE, em Portugal, devem encetar um conjunto de acções de investimento na área de *Abertura ao Analista* de forma a transformarem-se em moldes que facilitem o Processamento Analítico.

1.2. Dimensões de análise



No quadro seguinte, são apresentados os valores próprios ordenados por tamanho na coluna *Total.*

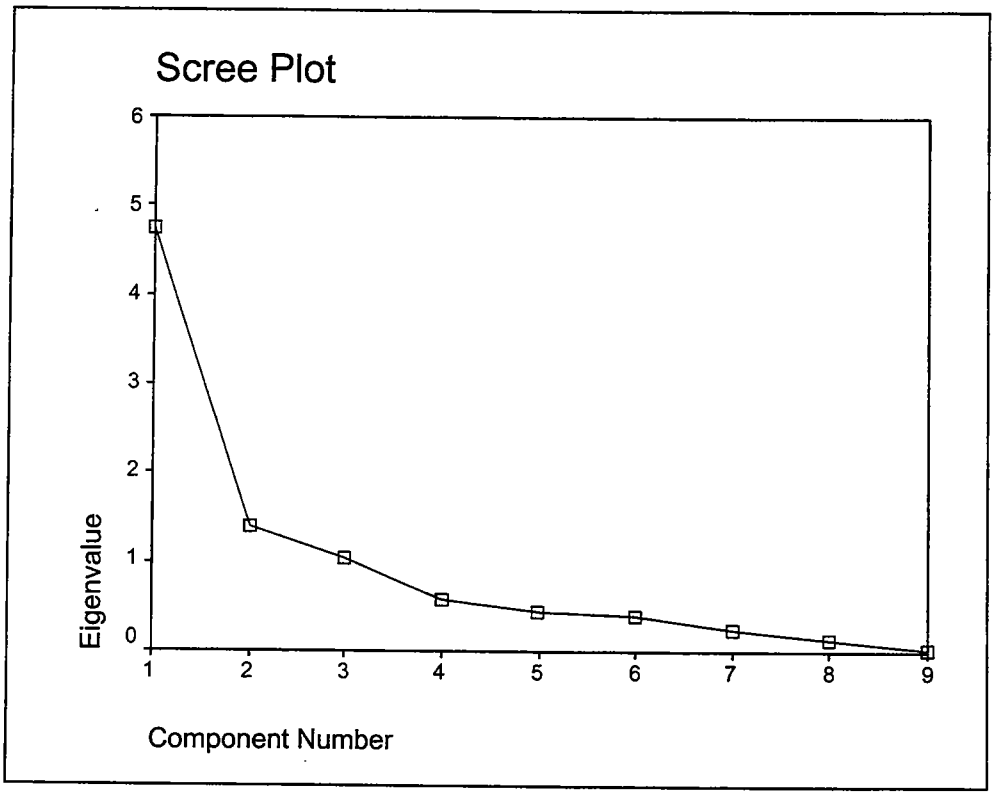
De acordo com os critérios de Kaiser, devem ser retidas os factores com valores próprios superiores a 1. Isto acontece com três: o primeiro, sem dúvida o mais importante, explica quase 53% da variância total enquanto, no conjunto, os três factores explicam 80% da variância total o que pode ser considerado uma solução satisfatória.

Quadro 75. Variância Total Explicada

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Factor	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.753	52.809	52.809	4.753	52.809	52.809	2.914	32.373	32.373
2	1.385	15.394	68.203	1.385	15.394	68.203	2.825	31.392	63.765
3	1.027	11.413	79.616	1.027	11.413	79.616	1.427	15.852	79.616
4	.580	6.439	86.056						
5	.440	4.887	90.942						
6	.390	4.333	95.276						
7	.251	2.791	98.066						
8	.143	1.589	99.655						
9	3.106E-02	.345	100.000						

Igualmente, o *Scree Plot*, gráfico em que os valores próprios são representados em relação ao número de factores a reter, corrobora a retenção de três factores correspondentes à maior inclinação da recta.

Figura 57. Scree Plot



No quadro seguinte, apresentamos a matriz de pesos dos factores (*loadings*). Esta permite concluir da existência de algumas variáveis que não podem ser claramente associadas a um factor o que é frequente na primeira aproximação à AFCP.

Quadro 76. Matriz de pesos dos factores

	Factores		
	1	2	3
Dados e Publicações	.839	-4.234E-03	-.112
Navegação	.822	.184	-5.764E-02
Integridade Temporal	.791	-.455	.212
Actualidade	.761	-.476	-.223
Diversidade	.753	-.475	.312
Informação ao Analista	.698	.127	-.569
Abertura	.648	.381	-.125
Processamento	.644	.635	3.367E-02
Multidimensionalidade	.528	.357	.692

Com a execução de uma *Rotação Varimax* é possível extremar a associação das variáveis aos factores. O resultado desta operação surge no quadro seguinte que nos permite associar o primeiro factor a *Características da Informação sem Multidimensionalidade*, o segundo a *Características do sítio* e o terceiro à *Multidimensionalidade*.

Com efeito, o primeiro factor resulta de contribuições elevadas das variáveis Processamento, Integridade Temporal e Actualidade. O segundo tem a contribuição mais elevada na Diversidade, mas apresenta contribuições importantes também por parte das variáveis *Dados e Publicações, Navegação, Abertura e Informação ao Analista.*

Quadro 77. Matriz de pesos dos factores após rotação

	Factores		
	1	2	3
Processamento	**.900**	9.899E-02	.266
Integridade Temporal	**.896**	.187	.202
Actualidade	**.822**	.384	-.181
Diversidade	.291	**.835**	-.213
Informação ao Analista	-1.739E-02	**.748**	.510
Abertura	.144	**.701**	.261
Navegação	.414	**.681**	.278
Dados e Publicações	.553	**.623**	.152
Multidimensionalidade	.208	.181	**.899**

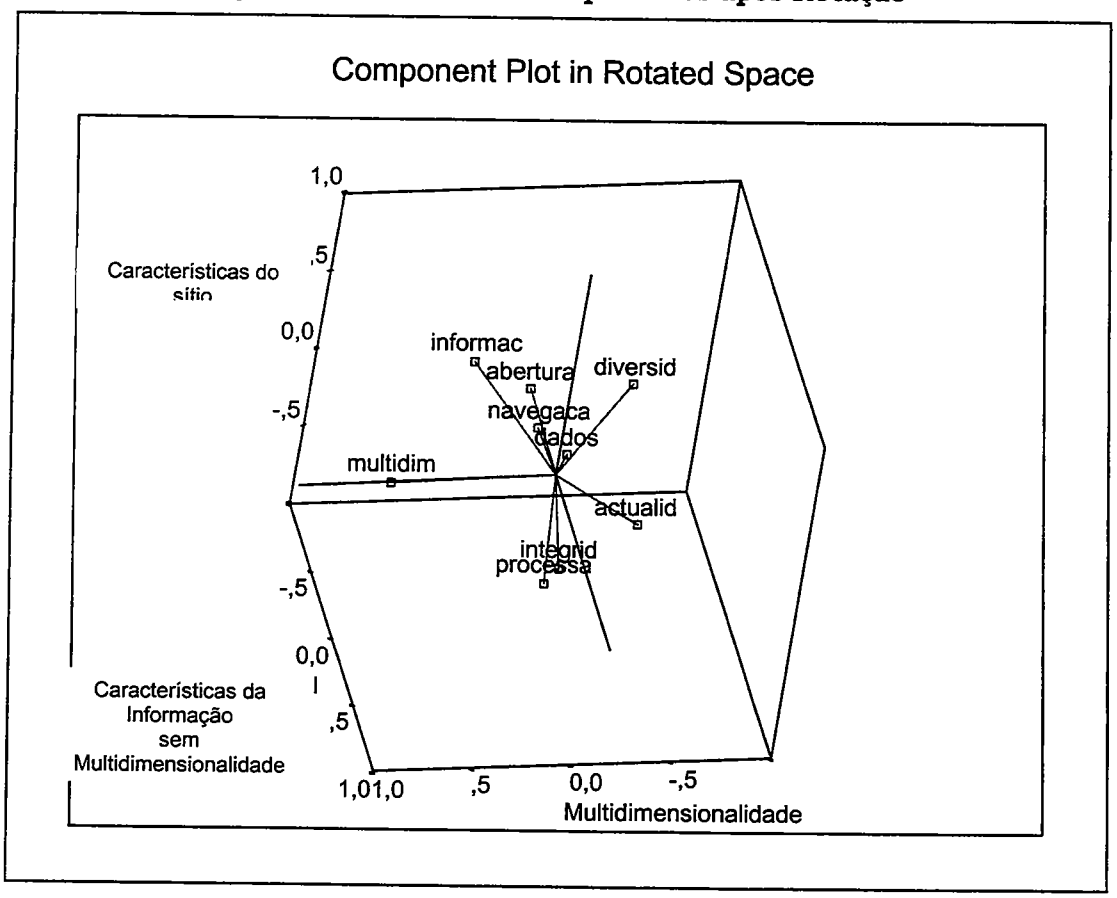
No gráfico seguinte, os factores após rotação são representados relativamente às variáveis de análise.

A designação dos factores resulta da combinação do significado das variáveis com maior contribuição. Nos resultados precedentes, a interpretação foi, de certo modo, facilitada por ser possível referi-los ao primeiro nível de agrupamento de variáveis de que é exemplo *Características do sítio.*

Nos três factores, é visível a sua associação às variáveis com uma representação mais clara em termos de peso.

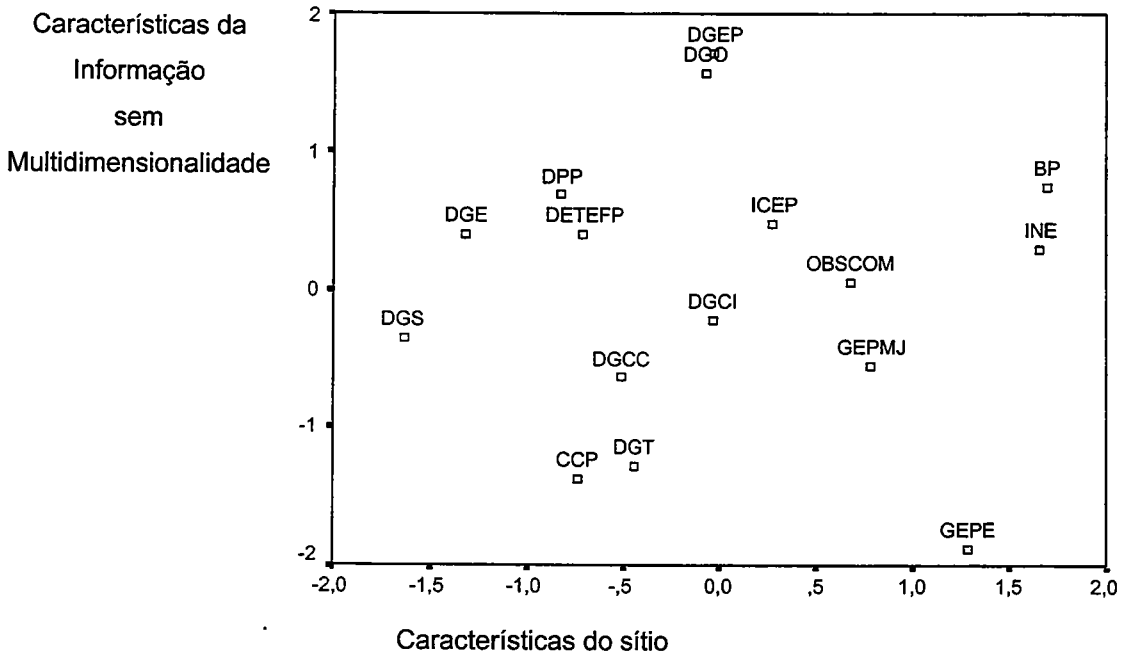
No caso particular do factor *Multidimensionalidade,* é possível notar a colinearidade com a variável do mesmo nome, pois esta é a única com uma contribuição significativa. Na parte inferior da representação gráfica, pode apreciar-se a relação do factor que designámos por *Características da Informação sem Multidimensionalidade* com as variáveis que contribuem para o mesmo, destacando-se o maior afastamento relativamente à origem de *Integridade Temporal* e *Processamento.* Finalmente, a componente que designámos por *Diversidade e Informação ao Analista* é a que é representada graficamente em termos da contribuição de mais variáveis o que é legítimo de acordo com o quadro anterior.

Figura 58. Gráfico das Componentes após Rotação



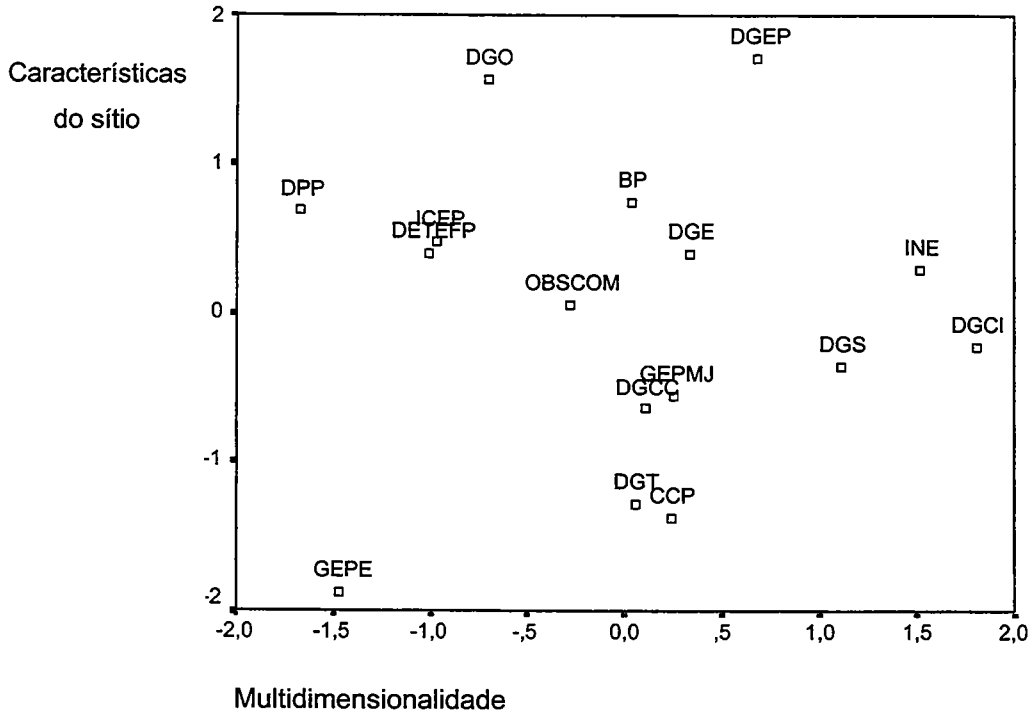
A representação gráfica dos SIE em termos dos factores encontrados é o objecto dos três gráficos seguintes. No primeiro detecta-se que as *Características do Sítio* opõem o Banco de Portugal (BP), o INE e o Gabinete de Estudos e Prospectiva da Economia (GEPE) relativamente à Direcção Geral de Saúde. Por outro lado, as *Características da Informação sem Multidimensionalidade* opõem o GEPE à Direcção Geral do Orçamento (DGO) e à Direcção Geral de Estudos e Previsão do Ministério das Finanças (DGEP).

Figura 59. Os SIE nos Factores Retidos (1)



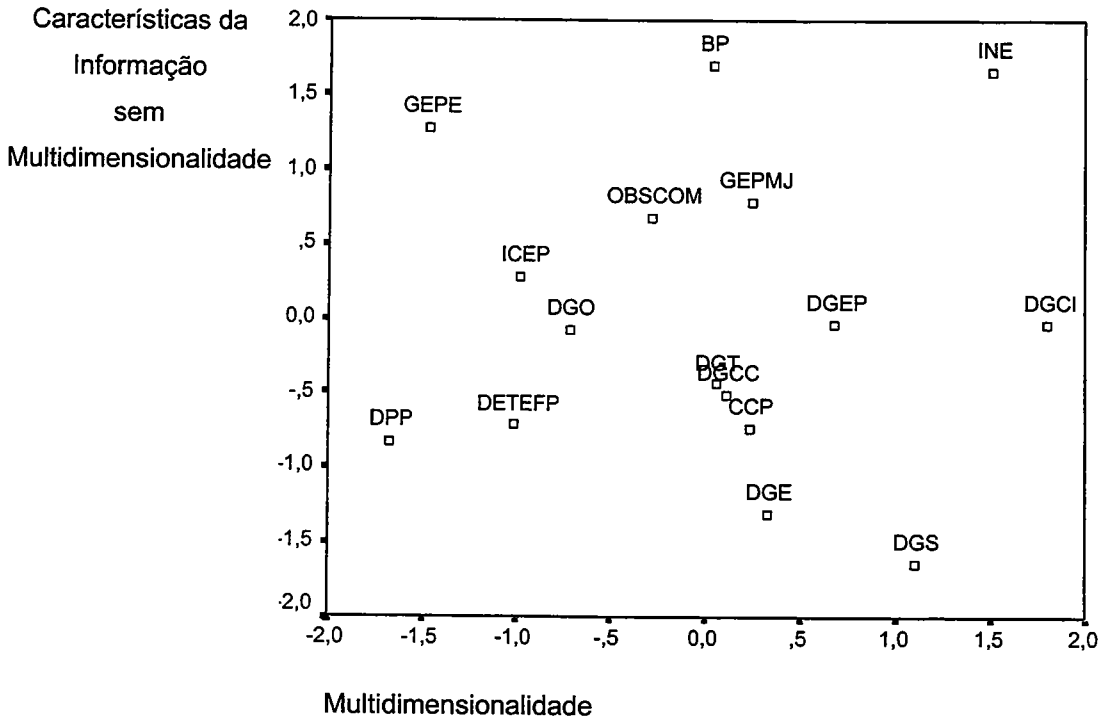
O segundo gráfico torna clara uma oposição entre o INE e a Direcção Geral de Impostos (DGCI), por um lado, e o Departamento de Prospectiva e Planeamento (DPP) e o Departamento de Estatísticas do Trabalho e Formação Profissional (DETEFP), por outro, em termos de *Multidimensionalidade.*

Figura 60. Os SIE nos Factores Retidos (2)



Finalmente, é de referir a oposição entre a Direcção Geral do Orçamento (DGO) e o DGEP relativamente ao GEPE em termos de *Características da Informação sem Multidimensionalidade*, motivado fundamentalmente pela não disponibilização de dados por esta organização à data de avaliação (1999).

Figura 61. Os SIE nos Factores Retidos (3)



2. Identificação dos segmentos de informação para a Economia

A identificação dos segmentos de informação foi realizada através da Análise de Clusters. Esta consiste num conjunto de procedimentos estatísticos multivariados que associa em grupos de elevada homogeneidade um conjunto de entidades.

Inicialmente, considerámos como variáveis os factores resultantes da AFCP, mas os resultados levaram-nos a prosseguir com as variáveis originais, embora em momentos chave possamos intervir com o conhecimento sobre factores.

Foram utilizados cinco processos de aglomeração, os designados por *average linkage between groups, vizinho mais afastado, Ward, vizinho mais próximo e centróide.* Os dados foram sujeitos a uma transformação estandardizada denominada *zscores.*

Em termos de processo de aglomeração constatou-se a enorme proximidade das distâncias até ao último estádio de junção. Assim, num primeiro momento, considerámos dois *clusters*. Esta perspectiva levou-nos à conclusão que os resultados pelos três primeiros métodos eram semelhantes enquanto os associados aos dois últimos eram pouco satisfatórios pelo facto de a forma de aglomeração conduzir a elementos isolados como *clusters* o que levou à sua rejeição.

A pesquisa de diferenças significativas entre *clusters* baseadas nos factores levou-nos à constatação da impossibilidade de realizar testes *post-hoc* com apenas dois *clusters* pelo que se considerou a utilização de três. Esta decisão levou-nos à rejeição dos clusters de acordo com o método *vizinho mais afastado* devido a problema semelhante aos dois outros métodos afastados.

Constatou-se, então, que a composição dos três *clusters* era idêntica considerando o *average linkage between groups* e o *critério de Ward* pelo que foi esse o número considerado para prosseguirmos a análise.

Nas linhas seguintes, reportamos a composição dos *clusters* em 1999.

Grupo 1 - formado pelo Instituto Nacional de Estatística (INE), Banco de Portugal (BP) e Direcção Geral de Estudos e Previsão (DGEP);

Grupo 2 - formado pela Confederação do Comércio de Portugal (CCP), Direcção Geral do Comércio e Concorrência (DGCC), Departamento de Estatística do Trabalho e Formação Profissional (DETEFP), Direcção Geral da Energia (DGE), Direcção Geral do Turismo (DGT), Direcção Geral da Saúde (DGS), Departamento de Prospectiva e Planeamento (DPP) e Gabinete de Estudos e Prospectiva da Economia (GEPE);

Grupo 3 - formado por Gabinete de Estudos e Prospectiva do Ministério da Justiça (GEPMJ), Investimento, Comércio e Turismo de Portugal (ICEP), Direcção Geral do Orçamento (DGO), Observatório do Comércio (OBSCOM) e Direcção Geral de Impostos (DGCI).

No quadro seguinte, é apresentado o processo de aglomeração de acordo com o primeiro dos critérios retido.

Uma das indicações importantes dada por este quadro é o valor do *coeficiente de fusão.*

Um coeficiente de fusão significativamente mais elevado que o anterior é uma indicação para uma possível paragem no processo de aglomeração pois pode testemunhar a junção a um *cluster* de um elemento com uma distância significativa relativa aos outros elementos do *cluster.*

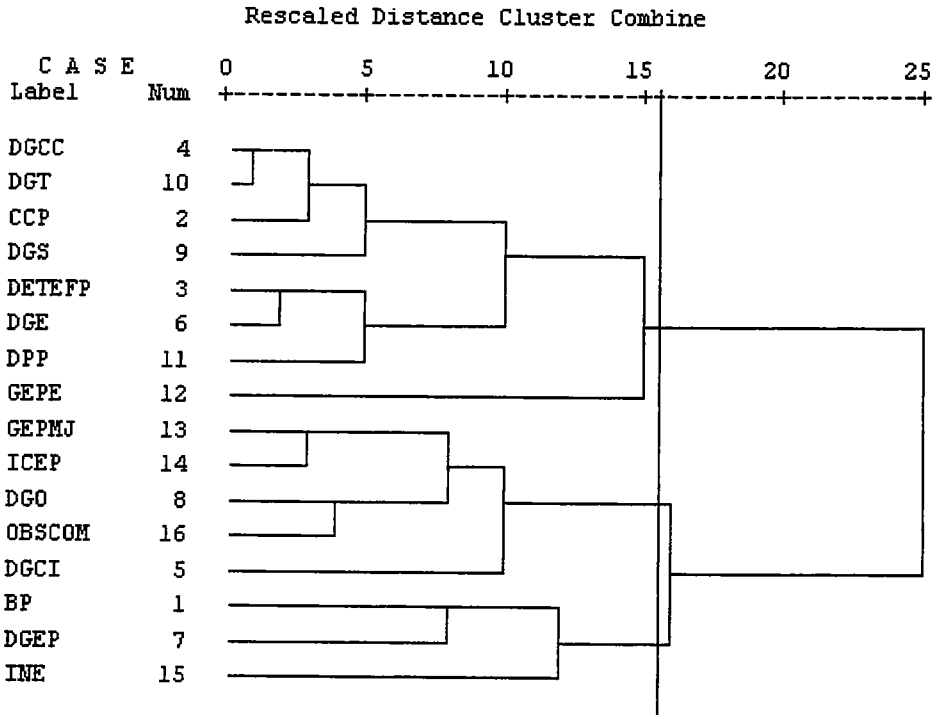
Como se pode constatar, o coeficiente de fusão tem ligeiras subidas até ao penúltimo estádio, concretizando, na última aglomeração, um aumento substancial quando comparado com anteriores o que indicia que, não fosse o que dissemos atrás sobre a pesquisa de diferenças significativas baseadas nos factores, não seria incorrecto considerar dois *clusters.*

Quadro 78. Processo de Aglomeração

	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
Stage	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	4	10	2.854	0	0	3
2	3	6	4.119	0	0	7
3	2	4	4.637	0	1	6
4	13	14	4.737	0	0	8
5	8	16	5.885	0	0	8
6	2	9	6.911	3	0	10
7	3	11	6.995	2	0	10
8	8	13	8.779	5	4	11
9	1	7	8.968	0	0	12
10	2	3	10.627	6	7	13
11	5	8	11.011	0	8	14
12	1	15	12.611	9	0	14
13	2	12	14.715	10	0	15
14	1	5	15.729	12	11	15
15	1	2	23.728	14	13	0

A figura seguinte mostra o dendrograma associado ao processo utilizado onde se torna patente à escala refeita que o corte utilizado garante elevada proximidade no interior de cada grupo.

Figura 62. Dendrograma



Dispondo destes resultados, é possível estudar o comportamento das variáveis no interior dos três *clusters* o que vamos fazer recorrendo a três estatísticas simples e a *caixas de bigodes.*

Quadro 79. Estatísticas Descritivas para os Clusters

	1			2			3			Total		
	Média	**N**	**Desvio Padrão**	**Média**	**N**	**Desvio Padrão**	**Média**	**N**	**Desvio Padrão**	**Média**	**N**	**Desvio Padrão**
Abertura	1.85	3	0.21	1.32	8	0.30	1.80	5	0.30	1.57	16	0.37
Actualidade	4.67	3	0.58	2.13	8	1.13	3.60	5	1.34	3.06	16	1.48
Dados e Publicações	4.38	3	0.53	3.00	8	0.16	3.52	5	0.26	3.42	16	0.59
Diversidade	4.00	3	1.00	2.25	8	1.04	2.80	5	0.45	2.75	16	1.06
Informação ao Analista	3.27	3	0.61	2.25	8	0.33	2.68	5	0.33	2.58	16	0.54
Integridade Temporal	4.67	3	0.58	2.38	8	1.06	3.40	5	0.89	3.13	16	1.26
Multidimensionalidade	4.00	3	1.00	2.25	8	0.89	2.60	5	0.89	2.69	16	1.08
Navegação	3.48	3	0.87	1.86	8	0.43	3.06	5	0.51	2.54	16	0.88
Processamento	4.33	3	0.58	2.25	8	1.17	3.60	5	1.14	3.06	16	1.34

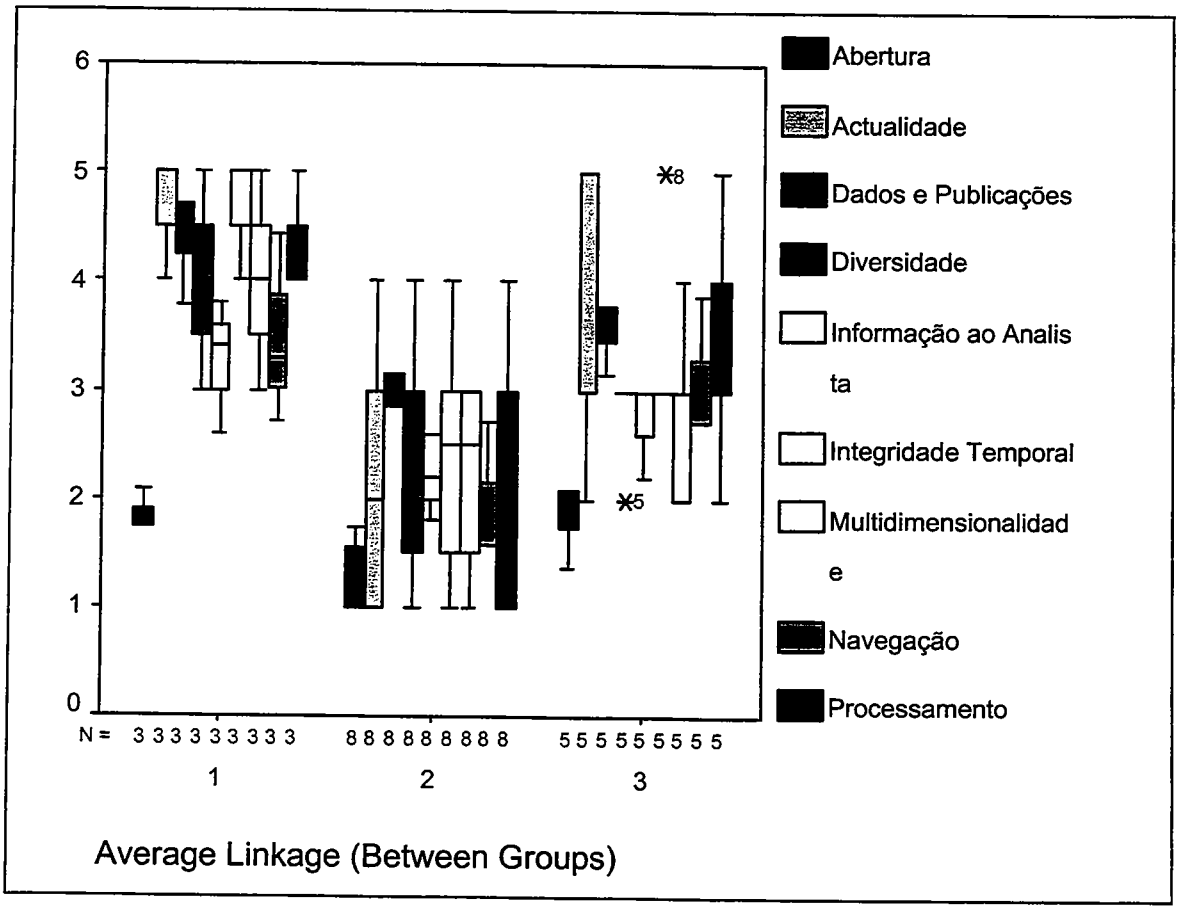
Os valores e os gráficos apresentados evidenciam diferenças entre os três grupos.

O primeiro grupo (BP, DGEP, INE) caracteriza-se por um conjunto de valores superiores à média em quase todas as variáveis de análise, referindo-se a excepção à *Abertura ao Analista.* Os seus pontos fortes situam-se na *Actualidade* e *Integridade Temporal.*

O segundo grupo (CCP, DETEFP, DGCC, DGE, DGS, DGT, DPP) apresenta valores abaixo da média em praticamente todas as variáveis de análise excluindo-se a referente a *Dados e Publicações* onde haverá SIE com um desempenho satisfatório.

O terceiro grupo (DGCI, DGO, GEPMJ, ICEP, OBSCOM) apresenta valores satisfatórios para a *Actualidade, Dados e Publicações* e *Processamento* mas manifesta deficiências na *Abertura, Informação ao Analista* e *Multidimensionalidade,* estando a *Navegação* sobre o ponto médio.

Figura 63. Descrição Gráfica dos Clusters



Mas a diferenciação entre os três grupos pode ser ainda mais significativa se considerarmos os três factores decorrentes da AFCP. É esse o objectivo da secção seguinte.

3. Diferenciação dos segmentos de informação para a Economia

Neste momento da análise, importa caracterizar cada grupo (variável qualitativa) relativamente a cada componente (variável quantitativa) para o que vai ser usada a análise da variância *One-Way Anova.*

O quadro seguinte apresenta os três factores e os três *clusters* obtidos através da análise factorial e da análise de *clusters.*

Recordamos que o primeiro factor engloba as variáveis associadas às *Características da Informação sem Multidimensionalidade.* O segundo factor engloba as variáveis relacionadas com as *Características do sítio.* O terceiro factor foi designado por *Multidimensionalidade.*

Quadro 80. Expressão dos SIE nos Factores Retidos

SIE	Factor 1	Factor 2	Factor 3	Cluster
BP	0.74013	1.69621	0.03862	1
CCP	-1.37343	-0.7383	0.23442	2
DETEFP	0.40292	-0.72151	-1.01409	2
DGCC	-0.64177	-0.51766	0.10952	2
DGCI	-0.22276	-0.03744	1.80495	3
DGE	0.40189	-1.3148	0.33039	2
DGEP	1.71307	-0.03652	0.68078	1
DGO	1.55883	-0.07714	-0.70795	3
DGS	-0.35411	-1.63592	1.10752	2
DGT	-1.28628	-0.44245	0.05983	2
DPP	0.68313	-0.82908	-1.67476	2
GEPE	-1.88233	1.27646	-1.46982	2
GEPMJ	-0.55888	0.77844	0.24661	3
ICEP	0.47542	0.27437	-0.974	3
INE	0.2918	1.65168	1.51197	1
OBSCOM	0.05236	0.67366	-0.28398	3

Estes dados vão ser sujeitos a testes de normalidade e de homogeneidade de variâncias. Em seguida procurar-se-á determinar se existirá comportamento diferenciado dos factores nos quatro grupos, terminando com uma análise da diferenciação de cada grupo operada através dos factores.

Quadro 81. Testes de Normalidade para os Factores

	Kolmogorov-Smirnov Statistic	df	Sig.	Shapiro-Wilk Statistic	df	Sig.
Características da Informação sem Multidimensionalidade	.115	16	.200	.972	16	.831
Características do sítio	.140	16	.200	.959	16	.622
Multidimensionalidade	.140	16	.200	.969	16	.787

Os testes de normalidade levam-nos a considerar a sua existência na distribuição dos três factores. O quadro seguinte permite que concluamos pela positiva a hipótese de homogeneidade de variâncias

Quadro 82. Teste de Homogeneidade de Variâncias

	Levene Statistic	df1	df2	Sig.
Características da Informação sem Multidimensionalidade	.508	2	13	.613
Características do sítio	.751	2	13	.491
Multidimensionalidade	.458	2	13	.642

O quadro seguinte indica que os factores 1 (significância 0.076) e 3 (significância 0.333) não têm um comportamento diferenciado (pela significância > 0.05) nos três grupos decorrentes da análise de clusters o que já não acontece com o factor 2.

Quadro 83. Resultados do ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Factor 1 Características da Informação sem Multidimensionalidade	Between Groups	4.903	2	2.451	3.156	.076
	Within Groups	10.097	13	.777		
	Total	15.000	15			
Factor 2 Características do sítio	Between Groups	7.205	2	3.602	6.007	.014
	Within Groups	7.795	13	.600		
	Total	15.000	15			
Factor 3 Multidimensionalidade	Between Groups	2.332	2	1.166	1.197	.333
	Within Groups	12.668	13	.974		
	Total	15.000	15			

No quadro seguinte apresentamos os resultados associados ao teste *post-hoc* de Bonferroni essencial para determinar os grupos que se diferenciam em cada factor.

Quadro 84. Multiplas Comparações de Bonferroni

Dependent Variable	(I) Average Linkage (Between Groups)	(J) Average Linkage (Between Groups)	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Características da Informação sem Multidimensionalidade	1	2	1,4212474	,5966571	,100	-,2171366	3,0596313
		3	,6540070	,6436259	,984	-1,1133502	2,4213641
	2	1	-1,4212474	,5966571	,100	-3,0596313	,2171366
		3	-,7672404	,5024303	,452	-2,1468831	,6124023
	3	1	-,6540070	,6436259	,984	-2,4213641	1,1133502
		2	,7672404	,5024303	,452	-,6124023	2,1468831
Características do sítio	1	2	1,7192001*	,5242523	,018	,2796353	3,1587649
		3	,7814144	,5655214	,571	-,7714726	2,3343014
	2	1	-1,7192001*	,5242523	,018	-3,1587649	-,2796353
		3	-,9377857	,4414600	,160	-2,1500077	,2744363
	3	1	-,7814144	,5655214	,571	-2,3343014	,7714726
		2	,9377857	,4414600	,160	-,2744363	2,1500077
Multidimensionalidade	1	2	1,0334154	,6682972	,438	-,8016878	2,8685186
		3	,7266667	,7209054	,996	-1,2528955	2,7062288
	2	1	-1,0334154	,6682972	,438	-2,8685186	,8016878
		3	-,3067487	,5627566	1,000	-1,8520439	1,2385465
	3	1	-,7266667	,7209054	,996	-2,7062288	1,2528955
		2	,3067487	,5627566	1,000	-1,2385465	1,8520439

* The mean difference is significant at the .05 level.

De acordo com o teste *post-hoc* de Bonferroni, relativamente ao factor 2, os clusters 1 e 2 são significativamente diferentes entre si a um nível de significância de 0.018. A um nível de significância de 0.571 os clusters 1 e 3 serão semelhantes entre si o que acontecerá também entre os clusters 2 e 3 a um nível de significância de 0.18. Este resultado corrobora a nossa conclusão da secção anterior e junta-lhe que a maior diferenciação provém de factor 2, nomeadamente da *Diversidade* e em menor escala da *Informação ao Analista, Navegação, Abertura* e *Dados e Publicações.*

A análise das médias permite caracterizar os SIE pertencentes a cada grupo. O *cluster* 1 de que fazem parte Instituto Nacional de Estatística, Banco de Portugal e Direcção Geral de Estudos e Previsão de acordo com o critério de Ward ou os mesmos acompanhados pelo ICEP e GEPMJ, tem valores estandardizados positivos em todos os factores.

O *cluster 2* (CCP, DGCC, DGT, GEPE) tem valores estandardizados negativos em todos os factores o que significa valores estandardizados negativos para estes SIE a nível de integridade temporal, processamento imediato, actualidade da informação, dados e publicações e multidimensionalidade.

O *cluster 3* tem valores estandardizados positivos em todos os factores, não apresentando diferenças significativas em relação aos restantes, assumindo-se como cluster de transição.

Quadro 85. Descrição dos Clusters nos Factores Retidos

		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
						Lower Bound	Upper Bound		
Factor1 Características da Informação sem Multidimensionalidade	1	3	.9150009	.7265918	.4194980	-.8899533	2.7199550	.29180	1.71307
	2	8	-.5062465	.9533005	.3370426	-1.3032257	.2907327	-1.88233	.68313
	3	5	.2609939	.8185509	.3660671	-.7553713	1.2773591	-.55888	1.55883
	Total	16	6.938894E-17	1.0000000	.2500000	-.5328624	.5328624	-1.88233	1.71307
Factor 2 Características do sítio	1	3	1.1037920	.9877906	.5703011	-1.3500157	3.5575998	-.03652	1.69621
	2	8	-.6154080	.8636494	.3053462	-1.3374370	.1066209	-1.63592	1.27646
	3	5	.3223777	.3945833	.1764630	-.1675622	.8123175	-.07714	.77844
	Total	16	3.885781E-16	1.0000000	.2500000	-.5328624	.5328624	-1.63592	1.69621
Factor 3 Multidimensionalidade	1	3	.7437910	.7386897	.4264827	-1.0912160	2.5787981	.03862	1.51197
	2	8	-.2896244	.9804358	.3466364	-1.1092892	.5300404	-1.67476	1.10752
	3	5	1.712437E-02	1.1008741	.4923259	-1.3497914	1.3840401	-.97400	1.80495
	Total	16	-2.0816682E-17	1.0000000	.2500000	-.5328624	.5328624	-1.67476	1.80495

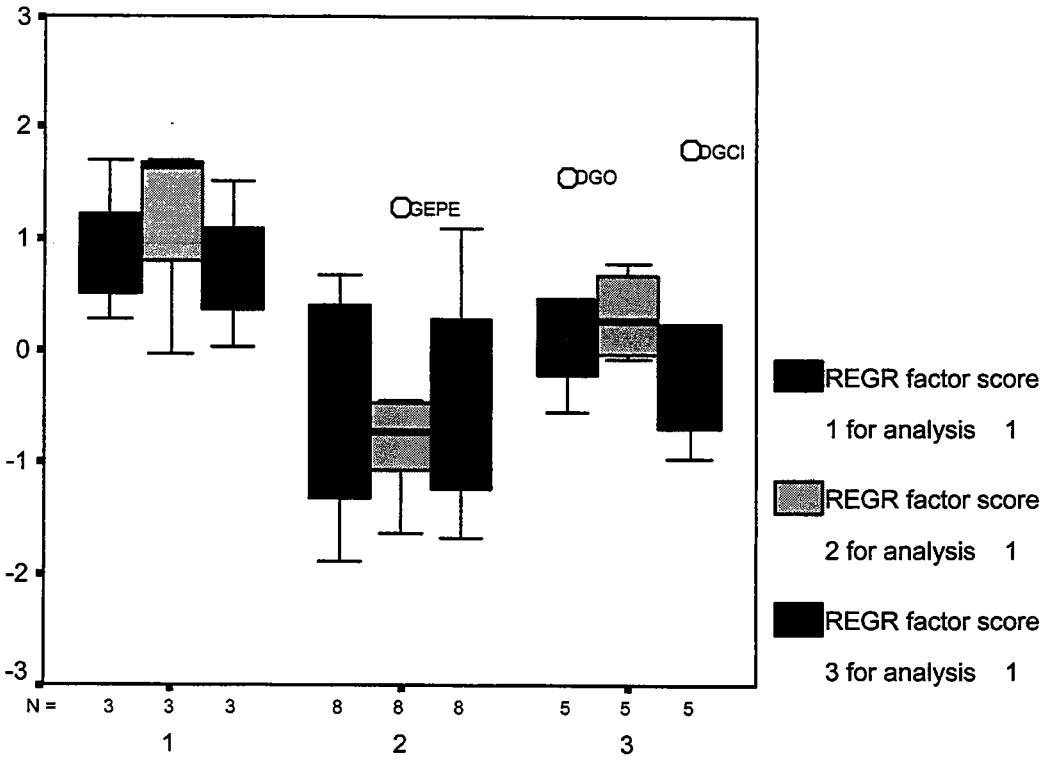
No quadro seguinte, resumimos a análise realizada mostrando os resultados positivos e negativos nos *clusters* relativamente aos factores retidos onde P significa valores estandardizados positivos e N valores estandardizados negativos. Relembra-se, no entanto, que as diferenças associadas aos factores 1 e 3 não são estatisticamente significativas.

Quadro 86. Resumo de Análise

		Factor		
		1	**2***	**3**
Cluster	**1**	P	P	P
	2	N	N	N
	3	P	P	P

Algumas das conclusões anteriores são visíveis na seguinte representação gráfica que relaciona os factores com os *clusters*.

Figura 64. Os *clusters* em termos de factores



4. Hipótese sobre os estádios de desenvolvimento dos Sistemas de Informação para a Economia

A maturidade dos Sistemas de Informação para a Economia pode ser vista como um processo de evolução onde, no momento da análise são relevantes três dimensões:

1) combinação derivada das variáveis de análise *Características da Informação sem Multidimensionalidade*;
2) combinação derivada das variáveis de análise *Características do sítio* e
3) *Multidimensionalidade.*

Estas dimensões, identificadas a partir de um conjunto de 9 atributos que resultaram de um processo de consolidação de 68 variáveis de análise de tipo binário conduziram à identificação de três grandes grupos de Sistemas de Informação para a Economia:

Grupo I: maioritariamente constituído pelos sistemas de informação tradicionalmente ligados à informação para a Economia que apresentam valores globalmente positivos nas dimensões de análise embora se reconheça que em termos de relação com o analista da Economia têm um processo longo a percorrer.

Grupo II: constituído por um grupo diversificado vocacionado para o comércio, prospectiva e outra informação (turismo, saúde, trabalho e formação profissional) com valores negativos ao nível daquelas dimensões de análise que deverão por isso encetar um processo de revisão da sua forma de produção de informação e relação com o analista.

Grupo III: claramente um grupo de transição entre o segundo e o primeiro, evidenciando as dimensões com um valor positivo estandardizado, mas não escondendo aspectos negativos em variáveis de análise como a *Informação ao Analista* e a *Multidimensionalidade de Análise* e aspectos positivos como a *Actualidade* dos dados e a constituição de ficheiros para facilitar o *Processamento* pelo analista.

Na análise ficou claro o atraso de todos os grupos em termos da variável *Abertura ao Analista.* Este critério define a aposta estratégica para os SIE que tiverem como objectivo cumprir melhor a sua missão.

A principal conclusão a reter deste estudo é que a ideia de uma evolução da globalidade dos SIE através de um conjunto de fases associadas às variáveis de análise é consubstanciada nos resultados aqui alcançados, tornando difícil uma análise estatística suportada pela existência de diferentes segmentos. Os SIE diferenciam-se mais pelo estádio de evolução do que por diferentes características.

Nesta perspectiva, avançamos com a identificação de quatro estádios na disponibilização interactiva de informação para a Economia. No primeiro estádio, que denominamos ***Disponibilização de Dados e Publicações***, o SIE concentra-se naquilo que identifica como seu principal objectivo e torna disponível os dados que produz, informação sobre as suas publicações e inclusivamente artigos elaborados pelos seus departamentos técnicos. Trata-se de uma perspectiva de fornecimento um tanto despida de outras preocupações, podendo coexistir com elementos desactualizados, volatilidade nos dados e fraca multidimensionalidade. Inclusivamente, por vezes, os dados são disponibilizados em formato que obriga à sua reintrodução nos sistemas informáticos por aqueles que os pretendem tratar. Apesar de tudo, esta faceta do SIE é saudada pelos seus utilizadores que deixam de ter necessidade de recorrer aos antigos meios de acesso aos dados. Neste contexto, o SIE fecha-se mais em si mesmo, vai fornecendo os dados que julga necessários, mas não facilita o relacionamento nem atende a pedidos dos seus utilizadores.

Com o tempo, as preocupações com a actualidade dos dados vão ganhando forma e desencadeiam todo um conjunto de acções que vão permitir um hiato cada vez menor entre o momento da consulta e aquele a que os dados ou a informação se referem. Ganha assim corpo, um novo estádio que denominamos de ***Procura da Actualidade***, onde, para além desta variável, se dá maior atenção à utilidade que os dados podem ter para quem os usa e se passa a disponibilizá-los num formato que facilita a sua recepção e processamento nos computadores de destino. Este processo é acompanhado pelo aumento de dados e publicações disponibilizados, gerando algum entusiasmo no SIE e nos seus utilizadores.

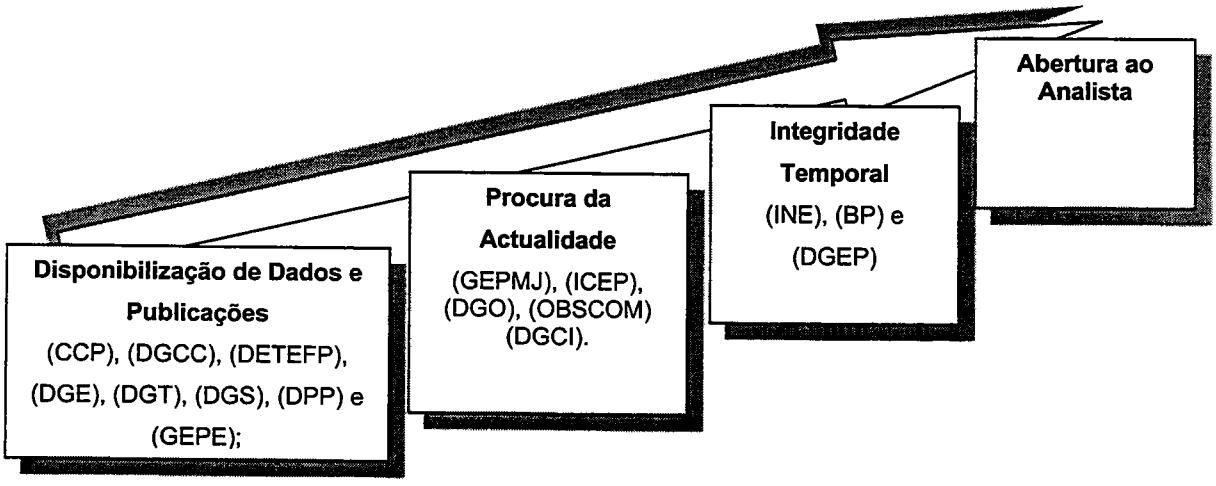
No estádio de ***Integridade Temporal***, desenvolve-se a disponibilização de séries cronológicas que irão facilitar a análise de tendências económicas e libertar ainda mais outros circuitos da relação do analista com o SIE. Este

facto consubstancia uma preocupação fundamental com a não volatilidade dos dados. Desenvolve-se ainda o que se pode designar por facilitação do multidimensional através da disponibilização de dados com uma granularidade que ultrapassa em muito os tradicionais agregados por variável de análise. Mantém-se, apesar de tudo, uma baixa abertura ao analista a qual é acompanhada, muitas vezes, por informação medíocre.

O estádio de ***Abertura ao Analista*** é considerado por nós como o estádio objectivo deste processo de evolução, podendo ser caracterizado como aquele em que continua a disponibilização de dados e publicações com elevada integridade temporal e multidimensionalidade, mas estes processos são acompanhados pela possibilidade de o analista, no contexto dos dados disponibilizados pelo SIE, poder definir requisitos de informação que originem processos internos de desenvolvimento. Admite-se que este estádio seja acompanhado por uma maior granularidade da informação disponibilizada eventualmente acompanhada por ferramentas de extracção que evitarão longas ocupações do sistema de comunicações. Em muitos SIE será recolhido o perfil do visitante e o próprio interface modificar-se-á de uma forma automática em função das características do visitante.

Na figura seguinte, ilustramos este processo e de acordo com ele podemos concluir que os SIE mais evoluídos mostram o caminho aos restantes. O que se seguirá em relação ao estádio de *Abertura ao Analista* será o futuro que o dirá e está de acordo com a nossa formulação sobre o ciclo de vida dos ponderadores.

Figura 65. Representação de Estádios de Desenvolvimento dos SIE



Há, ainda, que chamar a atenção para a extrema sensibilidade dos resultados

associados às técnicas utilizadas.

Pequenas modificações nos dados traduzem-se na necessidade de uma revisão global de análise. Isto induz que o conhecimento que hoje se detém nesta área está sujeito a grandes modificações após a passagem de algum período que os SIE tenham aproveitado para se transformar.

O futuro poderá assim vir a confirmar a ideia que lançámos associada aos estádios de transformação, mas por outro lado carece de um indicador mais robusto como o que apresentámos no início do capítulo e que designámos como índice de maturidade que pela sua estrutura estará menos sujeito a flutuações por este tipo de factos apesar de as poder reflectir.

5. Será a abertura ao analista uma utopia?

A captação do desenvolvimento da maturidade dos SIE na era das redes, num momento relativamente inicial do processo, permitiu a percepção de diferentes estádios de desenvolvimento que descrevemos nos pontos anteriores concluindo que o estádio de *abertura ao analista* se posicionava como uma próxima etapa.

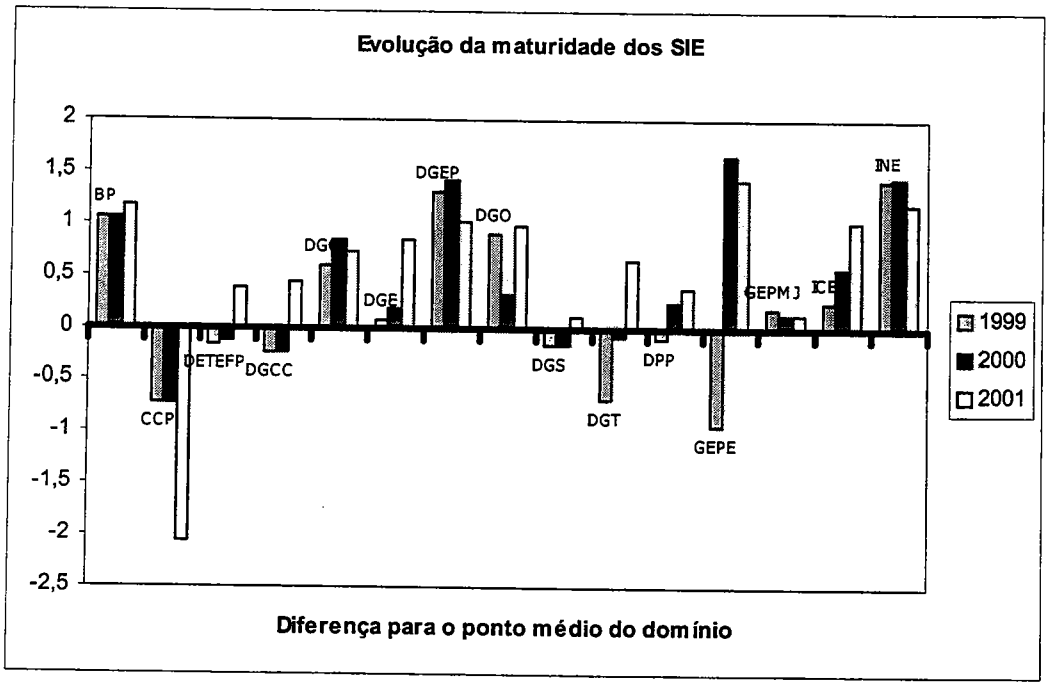
Esta ideia não é determinada pelo facto de acreditarmos que esta sucessão de estádios é determinística. Na nossa perspectiva, para o estádio final estar garantido, não basta um SIE situar-se no estádio precedente e prosseguir o desenvolvimento da maturidade, podendo existir desenvolvimento em sentidos que contrariem o que sentimos que seria o que melhor serviria os interesses dos analistas da economia.

Nesta secção vamos mostrar, num primeiro momento, que nos anos se seguiram a 1999 houve uma tendência para a suavização das diferenças entre os SIE o que justifica, em seguida, a concretização de uma análise discriminante para associar os SIE aos clusters respectivos de acordo com a definição em 1999. Formularemos uma conclusão sobre os estádios de desenvolvimento dos SIE que nos levará a rejeitar a ideia de sequência entre os estádios e terminaremos com a referência à necessidade de um corte epistemológico para apoiar o estádio objectivo de maturidade.

5.1. A suavização das diferenças entre os SIE nos anos de 2000 e 2001

É de admitir que o desenvolvimento dos SIE que estão nos estádios iniciais do processo seja mais rápido levando ao apagar da imagem inicial tão bem definida que conseguimos no capítulo anterior. Por outras palavras, pode haver uma convergência dos SIE para um estádio anterior ao de abertura ao analista. Esta convergência é anunciada pela representação gráfica seguinte.

Figura 66. Maturidade dos SIE no período 1999-2001



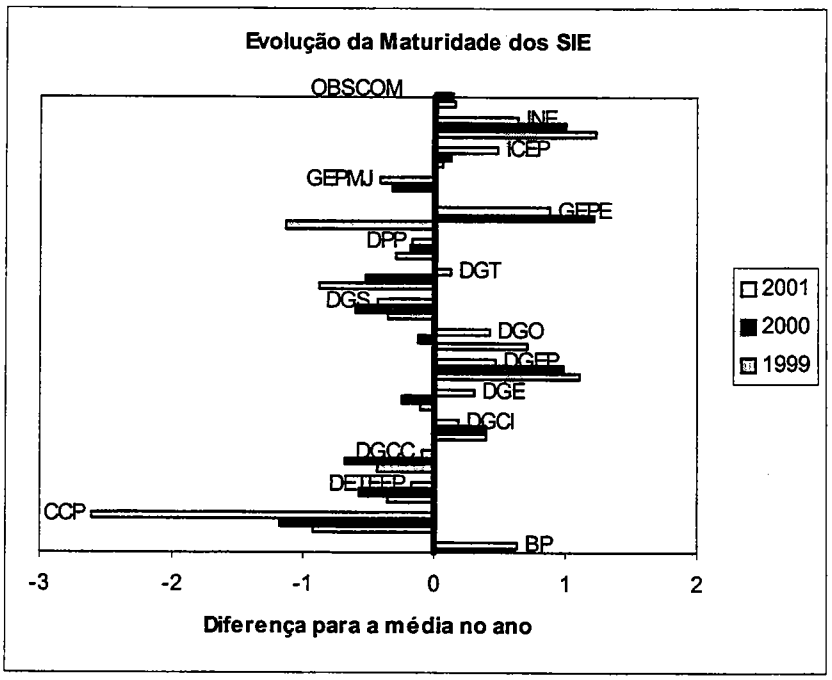
A figura anterior mostra um progressivo afastamento dos vários SIE relativamente a um ponto médio na escala de maturidade. Entre 1999 e 2001 diminuiu substancialmente o número de SIEs com valores de maturidade inferiores a esse ponto.

Um caso que parece afastar-se do padrão definido, o do CCP, justifica-se pela não disponibilização interactiva de informação no momento de observação no ano de 2001. Nesta data, com excepção deste SIE, todos apresentavam um valor global para o índice de maturidade superior ao ponto médio da escala.

Se formos mais exigentes e considerarmos que o ponto médio não é o definido pela escala, mas pelos valores de maturidade, a perspectiva já não será tão optimista, mas permitirá visualizar grandes transformações num sentido positivo como o GEPE e num sentido negativo como o do INE e do ICEP que

começam a aproximar-se desse ponto médio não pelo facto de diminuir a qualidade dos seus serviços, mas pelo facto de não evoluir a uma taxa semelhante ou superior ao dos SIE em expansão em termos de maturidade.

Figura 67. Maturidade dos SIE no período 1999-2001 no ambiente de disponibilização

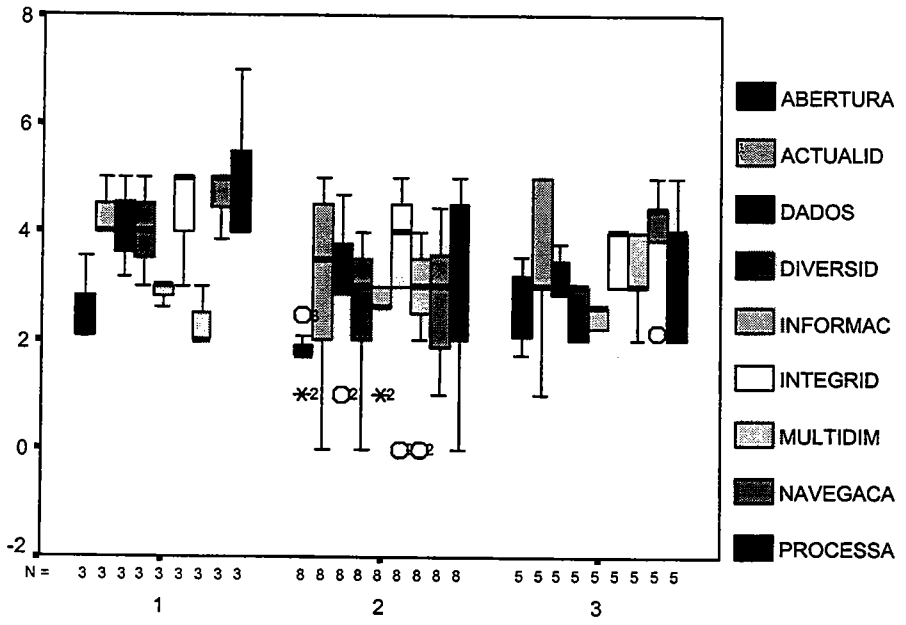


Perante esta constatação a um nível geral, somos forçados a interrogar-nos se, algum dia, será atingido o estádio de *abertura ao analista* ou se o mesmo configura mais uma utopia. Vamos procurar, neste ponto, detectar a capacidade de, sem alterações significativas, haver um alcançar do estádio e, em caso negativo, delinear algumas ideias baseadas em outras áreas do conhecimento que possam apoiar um desenvolvimento nesse sentido.

Uma primeira análise, suportada pelas *caixas de bigodes* com a anterior estrutura de clusters, relativamente aos valores observados em 2001, mostra uma suavização das diferenças entre os mesmos ao nível das variáveis de análise. Os *clusters* 1 e 3 repartem os melhores valores para algumas variáveis, mas a proximidade do segundo é evidente.

A representação consegue no entanto manifestar a manutenção de valores baixos para a abertura ao analista. Por outras palavras, não se configura a evolução para um estádio em que essa variável de análise assuma um valor determinante como diferenciadora e simultaneamente um valor elevado.

Figura 68. Caixas de bigodes com anterior estrutura de clusters



CL3_1

A partir deste ponto, procurámos novos clusters e novas visualizações que nos anunciassem a concretização deste estádio. Fizemos múltiplas simulações. Numa primeira, procurámos novos *clusters*, novas componentes e novas representações, mas a visualização de resultados nada juntou ao que esboçámos antes.

A pesquisa de novos clusters e novas componentes sem CCP, por o mesmo não disponibilizar dados em 2001, melhorou os indicadores da viabilidade de uma análise de componentes principais, deu maior definição aos componentes, mas não permitiu uma melhor definição de estádios. Idêntica impossibilidade surgiu com a não consideração de alguns critérios: *o que é que a diversidade tem a ver com a maturidade?*, interrogámo-nos.

5.2. A associação dos SIE aos clusters através de uma análise discriminante

Neste ponto, entendemos que a utilização de uma análise discriminante poderia contribuir para uma melhor análise sobre a evolução em termos de estádios o que se baseou em vários pressupostos:

- a imagem dos estádios obtida em 1999 é significativa e só terá sido conseguida pele captação do processo num momento inicial;

- com dados diferentes, os de 2000 e 2001, as análises de componentes principais e de *clusters* conduzirão a resultados também diferentes tornando difícil a comparação com o que se passava no momento inicial;
- a análise discriminante, baseada em dados de 1999 e aplicada a dados de 2000 e 2001 conduzirá a *clusters* que reflectem o conhecimento relevante naquele (1999) momento.

Em local adequado, referimos os pressupostos da análise discriminante, designadamente os que se referem à normalidade dos dados e à homogeneidade das variâncias.

Já vimos que só podemos considerar a distribuição normal para a globalidade das variáveis com um nível de significância excessivamente baixo: 0.01. No quadro seguinte, apresentamos resultados quanto ao teste de homogeneidade de variâncias: esta, a um nível de significância de 0.05, não se verifica quanto à variável *Dados e publicações*.

Quadro 87. Teste de homogeneidade de variâncias

	Levene Statistic	df1	df2	Sig.
Abertura	.299	2	13	.746
Actualidade	1.694	2	13	.222
Dados e publicações	7.062	2	13	**.008**
Diversidade	1.538	2	13	.251
Informação	1.073	2	13	.370
Integridade	.987	2	13	.399
Multidimensionalidade	.044	2	13	.957
Navegação	1.383	2	13	.285
Processamento	1.516	2	13	.256

O teste aos pressupostos da análise deixou-nos um pouco apreensivos quanto aos seus resultados. Reis (1997) é claro quanto à sua importância: (1) se a homogeneidade de variâncias não se verificar, as funções discriminantes poderão não assegurar uma separação máxima entre os grupos e as probabilidades de pertença a cada grupo poderão ser distorcidas e (2) se o pressuposto de normalidade multivariada não se verificar, tal poderá levar à tomada de decisões erradas sobretudo quando se tratar de pequenas amostras.

Mas o mesmo autor admite a questão: *é de aplicar a análise discriminante quando os pressupostos de normalidade variada e igualdade de variâncias e covariâncias não se verificam?* (Reis, 1997: página 245). O próprio texto avança uma resposta:

> *a percentagem de classificações correctas poderá servir sempre de guia para aceitar como bom qualquer modelo de descrição da situação real. Quando essa percentagem é elevada, a violação dos pressupostos não é muito grave; no entanto, se a percentagem de classificações correctas for muito pequena, não é possível distinguir se isso se deve ao facto de não terem sido respeitados os pressupostos de análise ou se estamos perante um conjunto de variáveis com fraco poder para discriminar os grupos de indivíduos.*

A verdade é que, com os dados utilizados, apesar de não podermos garantir os pressupostos da análise, a classificação teve uma percentagem de acerto de 100% pelo que considerámos a análise legítima.

Foi utilizado um método discriminante *stepwise* e como critério de selecção das variáveis o valor da estatística Λ *de Wilks* com aproximação à *F*. Os resultados permitiram constatar que existe um conjunto de variáveis sem poder discriminatório significativo para os três grupos de análise o que leva a que a hipótese de igualdade de médias dos três grupos não seja rejeitada. O quadro seguinte apresenta as variáveis que apresentam valores de Λ correspondentes a valores de *F* inferiores a 4,77 ($F_{(9,7)}$ para $\alpha=0,05$)[25].

Quadro 88. Variáveis excluídas da análise

Passo	Variável	Tolerância	Minima Tolerância	F para inclusão	Wilks' Lambda
3	Abertura	.813	.264	2.683	.018
	Actualidade	.588	.216	1.476	.021
	Diversidade	.977	.287	.198	.026
	Integridade	.370	.260	.839	.023
	Multidimensionalidade	.815	.286	1.151	.022
	Navegação	.813	.238	1.190	.022

[25] A estatística Λ de Wilks não está tabelada, mas pode ser aproximada por uma F (v,n-p-k+1) onde v é o número de variáveis, n o número de observações e k o número de grupos. A aproximação feita pode, no entanto, ser criticada pelo número de observações ser excessivamente baixo. Conscientes disso, procedemos à aproximação daquela estatística ao Rao's V a qual é válida em qualquer contexto associado à cardinalidade de dados, sendo no entanto os seus parâmetros de cálculo mais complexo. Não houve alteração nos resultados da análise discriminante pelo que mantemos a formulação.

Nestas condições, existem apenas três variáveis com poder efectivo para discriminar entre os três grupos. O quadro seguinte sumaria elementos relacionados com a sua obtenção.

Quadro 89. Sumário da análise discriminante

Passo	Variável introduzida	Número de variáveis introduzidas	Wilks' Lambda Statistic	Sig.
1	Dados e Publicações	1	.194	.000
2	Informação ao analista	2	.091	.000
3	Processamento	3	.027	.000

At each step, the variable that minimizes the overall Wilks' Lambda is entered.
a Maximum number of steps is 18.
b Minimum partial F to enter is 4,77.
c Maximum partial F to remove is 2.71.
d F level, tolerance, or VIN insufficient for further computation.

A variável *Dados e Publicações* é a que tem mais poder para discriminar os três grupos e, por isso, será a primeira a ser introduzida no modelo. Embora o poder discriminatório de *Informação ao Analista* e *Processamento* seja mais reduzido, é ainda significativo pelo que são também incluídas na análise. Com estas três variáveis a estatística *Λ de Wilks* tem um valor de 0,091, altamente significativo (<0,001).

Considerando três grupos, é possível definir duas funções discriminantes, mas, removido o efeito discriminante da primeira função, o da segunda já não é significativo.

Quadro 90. Poder discriminante das funções discriminantes

Função	Valor Próprio	% of Variância	Cumulativo %	Correlação Canónica	Teste de Funções	Wilks' Lambda	Qui-quadrado	df	Sig.
1	33.646	99.8	99.8	.985	1 through 2	.027	43.231	6	.000
2	.059	.2	100.0	.236	2	.944	.689	2	.709

a First 2 canonical discriminant functions were used in the analysis.

A capacidade discriminante da segunda função é ainda testemunhada pelo quadro seguinte através da análise dos centróides dos grupos que apresentam valores muito semelhantes para tal função. No mesmo quadro, apresentamos uma correspondência dos grupos aos estádios que decorre da análise em pontos anteriores – recorde-se que considerámos o grupo 3 como um estádio de transição.

Quadro 91. Centróides dos grupos

Average Linkage (Between Groups)	Estádio Correspondente	Função	
		1	2
1	3 – Procura da Integridade Temporal	9.328	-.235
2	1 – Disponibilização de dados e publicações	-4.515	-.110
3	2 – Procura de actualidade	1.628	.318

Unstandardized canonical discriminant functions evaluated at group means

As equações classificatórias para os três grupos são as seguintes:

Quadro 92. Coeficientes das funções de classificação

Variável de análise	Average Linkage (Between Groups)		
	1	2	3
Dados e Publicações	203.814	136.163	165.375
Informação ao analista	180.561	119.825	147.020
Processamento	50.225	32.830	40.937
(Constant)	-851.658	-377.079	-563.107

Fisher's linear discriminant functions

A utilização destas equações na classificação dos casos existente tem 100% de acerto o que significa que passámos a dispor de um modelo que pode prever o estádio a que determinada organização pertence, a partir do conhecimento de três variáveis de análise: *Dados e Publicações*, *Informação ao Analista* e *Processamento*.

O problema aqui abordado é relativamente diferente do que foi tratado no final do capítulo anterior e onde foi usada uma perspectiva difusa: ali, pretendia-se um valor para a maturidade, aqui pretende-se determinar um estádio de desenvolvimento, não sendo claro que um estádio mais avançado corresponda a um valor de maturidade mais elevado, embora esse seja o caso normal.

Dispondo destas três funções de classificação, podemos tentar determinar o movimento entre clusters operado entre 2000 e 2001, anos para os quais são conhecidos os valores daquelas variáveis que se apresentam como melhor discriminadoras. Tal movimento é resumido no quadro seguinte.

Quadro 93. Transição entre grupos com base nas equações classificatórias

SIE	Grupo		
	1999	2000	2001
BP	1	1	1
CCP	2	2	2
DETEFP	2	2	3
DGCC	2	2	3
DGCI	3	3	3
DGE	2	2	3
DGEP	1	1	3
DGO	3	3	2
DGS	2	2	3
DGT	2	3	3
DPP	2	3	3
GEPE	2	1	1
GEPMJ	3	3	3
ICEP	3	3	3
INE	1	1	1
OBSCOM	3	1	2

O quadro evidencia uma clara melhoria ao nível da disponibilização de informação para a economia nos dois anos considerados.

Quadro 94. Resumo da Transição

Transição	Estádio em 1999	Estádio em 2001	Frequência	Melhoria
2→3	Disponibilização de dados e publicações	Procura de actualidade	6	S
2→1	Disponibilização de dados e publicações	Procura de integridade temporal	1	S
3→1	Procura de actualidade	Procura de integridade temporal	1	S
3→2	Procura de actualidade	Disponibilização de dados e publicações	2	N
1→2	Procura de integridade temporal	Disponibilização de dados e publicações	1	N

Nestas condições o composição em número de cada estádio transformou-se:

Quadro 95. Evolução da distribuição por estádios

Estádio	Total de SIEs		Variação
	1999	2001	%
Disponibilização de dados e publicações	8	3	-62,5
Procura de actualidade	5	10	100
Procura de integridade temporal	3	3	0
Abertura ao analista	0	0	0

O que o quadro anterior significa é que diminuiu o número de organizações que se encontrava no estádio inicial de *Disponibilização de Dados e Publicações*, aumentou substancialmente o número de organizações no estádio de *Procura de Actualidade* e se manteve o número de organizações em termos de *Procura da Integridade Temporal.*

Existem casos de retrocesso, alguns dos quais já tinham sido captados em termos de discurso – por exemplo, o facto de o DGEP ter deixado de publicar dados em formato folha de cálculo, passando a fazê-lo em PDF prejudica *Processamento* e levou a uma descida em termos de classificação em estádios.

Surgiu-nos, entretanto, novo problema: em 1999, no conjunto das organizações existente, nenhuma se encontrava no estádio de *Abertura ao Analista* pelo que se tornou impossível gerar qualquer equação classificatória associada a este estádio.

Para o tornar possível, associámos ao conjunto de organizações real inicial duas organizações *fictícias* com valores elevados em termos de *Abertura ao Analista* e variáveis relacionadas (*Informação, Processamento*) e com valores mais elevados que a média das outras em variáveis menos importantes para o estádio.

Uma nova utilização da análise discriminante, levou a novas equações classificatórias, considerando a associada ao estádio de *Abertura ao Analista,* mas nenhuma organização foi classificada nesse estádio em 2000 e 2001.

Quadro 96. Novos coeficientes das funções de classificação

	Average Linkage (Between Groups)			
	1	2	3	4
Abertura	116.452	75.659	97.107	158.445
Actualidade	63.213	40.029	52.408	82.822
Dados e publicações	422.837	277.961	338.724	492.716
Diversidade	3.125	1.982	.376	-4.785
Informação ao analista	435.062	285.965	355.852	540.247
Integridade temporal	14.529	7.672	9.325	10.827
Multidimensionalidade	58.802	36.868	45.411	65.978
Navegação	-17.820	-15.198	-14.221	-20.090
Processamento	85.977	58.039	71.535	108.427
(Constant)	-2207.168	-936.432	-1439.070	-3220.291

Fisher's linear discriminant functions

Podemos assim concluir que existe movimento e progressão no interior dos estádios anteriores, mas que não parece haver uma aproximação do estádio final de maturidade dos SIE na era das redes.

5.3. A necessidade de um 'corte epistemológico'

Perante esta dificuldade em detectar avanços na direcção de um estádio que na nossa perspectiva é o que melhor consubstancia a articulação dos SIE com a sociedade em que se integram, interrogámo-nos como teria tal processo sido ultrapassado em outras áreas do conhecimento.

Antes de formularmos uma resposta, recordámos que o desenvolvimento dos SIE é objecto de processos de planeamento, de processos de definição estratégica. Isto significa que os seus resultados são condicionados pela visão estratégica da sua gestão.

- Será essa uma visão orientada para as novas necessidades de informação em termos das nossas variáveis de análise, manifestadas pelos agentes económicos?
- Irá nesse sentido uma disponibilização pelo ICEP em que os ficheiros processáveis exigem a satisfação de um pagamento e os ficheiros não processáveis são os disponibilizados ao grande público?
- Irá nesse sentido a transformação do formato de disponibilização do DGEP de ficheiros de folha de cálculo para ficheiros PDF?

- Irá nesse sentido a manutenção da *disponibilização de informação incompleta* num quadro totalmente modificado de recolha, capacidade de memória e algoritmos de processamento?
- Irá nesse sentido o restringir o acesso a bases de dados sujeitas a processos de anonimização a investigadores que desenvolvam a sua actividade de investigação no quadro de instituições credenciadas pelo MCT? A anonimização não é suficiente para garantir acesso a microdados?

Tentemos responder a uma pergunta distinta mas relacionada: o que aconteceria a determinados centros de poder baseados na informação se esta fosse disponibilizada em ficheiros processáveis com granularidade elevada? No nosso entender não perderiam esse poder, pois nem todos têm a capacidade de análise que os analistas da economia possuem e, portanto, manter-se-ia a necessidade do seu trabalho para múltiplas áreas. Mas há receio que, de certo modo, é um grave obstáculo para o desenvolvimento do espírito científico.

Ao romper com os pressupostos filosóficos das epistemologias clássicas que permaneciam fiéis aos esquemas mecanicistas da ciência (saídos da geometria euclidiana, do método cartesiano e da física newtoniana), Bachelard descreveu um *Novo Espírito Científico* (título do seu livro publicado em 1943), baseado na análise das geometrias não euclidianas e da relatividade de Einstein.

A epistemologia bachelardiana, apresentada em Bachelard (1986), que pretende ser um estudo científico da história das ciências, empenha-se em mostrar que a ciência procede por descontinuidade. Cada progresso é um *corte* em relação a um saber anterior que se pode revelar inteiramente ultrapassado.

Dum modo mais global, o espírito científico procede por cortes relativamente aos *obstáculos epistemológicos* que constituem as imagens e os símbolos do pensamento ingénuo e da tradição cultural, assim como as aparências sensíveis fornecidas pela experiência espontânea. Mas a própria herança científica serve de exemplo ao novo espírito científico. Eis porque Gaston Bachelard preconiza uma verdadeira estratégia do corte epistemológico, o que ele chama uma *Filosofia do não.* O investigador é o que trabalha na rectificação do saber.

A ideia de corte epistemológico não se reduziu à epistemologia das ciências.

Em Carvalho (1976), Barradas de Carvalho utilizou-a para classificar Herculano como o grande responsável pela transformação do que designou por História Crónica em História Ciência.

Ela ressurgiu na inovadora proposta althusseriana de «corte epistemológico» aplicado ao marxismo. A ideia de que uma ciência se constitui só quando produz um corte no panorama epistemológico existente, e de que nesse caminho se encontrarão sempre as «pedras» ou obstáculos promovidos por elementos interessados de carácter ideológico – a reacção, para Althusser –, era a peça chave do *pensamento althusseriano.*

Althusser (1967), obra fundamental para se conhecer o pensamento deste autor, colocou a ciência como um «continente» oculto pelas nuvens metafísicas. Quando estas nuvens se rasgam aparece por trás delas um novo continente, uma ciência nova. Assim, tal como Newton rasgou as nuvens teológicas do seu tempo e fez emergir o novo continente da física, também, Marx teria descoberto o continente da ciência histórica rasgando as nuvens da teología agustiniana e, sobretudo, do pensamento de Hegel. Sobre esta base, o filósofo francês reconstruiu a biografía de Marx e dividiu-a em duas etapas: um *Marx juvenil*, anterior a 1844, e um novo *Marx posterior*, o da maturidade. Este último descobriu o *materialismo científico.*

Também as múltiplas tentativas falhadas de transição para o socialismo parecem exigir a presença de um factor transformador adicional sem o qual o processo parece ser inviável: a mudança cultural. O problema foi levantado pela revolução chinesa e parece-nos claro que a transformação democrática em Portugal no pós 25 de Abril foi tão limitada pela ausência de uma ampla transformação cultural no seio das massas populares.

Esta necessidade de mudança, esta necessidade de desenvolver a capacidade de corte com a tradição, manifesta-se também ao nível da gestão das organizações.

Entende-se que para ocorrer uma mudança nos modelos mentais existentes em muitas organizações, se faz urgente a ascensão de pessoas construtoras da organização com visão de mundo que propiciem um corte epistemológico, com relação as características da burocracia que se encontram cristalizadas como:

- a padronização por regras ou normas legais;

- uma divisão do trabalho e uma organização dos cargos segundo princípios da hierarquia;
- a conduta regulada por regras ou normas técnicas;
- a separação entre os membros do corpo administrativo e a propriedade dos meios de produção e administração;
- a ausência de controle externo sobre recursos organizacionais;
- a formalização de regras, decisões e actos administrativos e
- o exercício de autoridade legal

Estas características da burocracia que estão sedimentadas em algumas organizações levam o comportamento organizacional a apresentar alguns tipos de reflexos como conservadorismo e resistência à mudança, alienação, modelos mentais estagnados.

Procurando eliminar as dimensões burocráticas existentes nas organizações, torna-se necessário enfatizar a relevância do processo de aprendizagem organizacional, dentro de uma abordagem holística de gestão, que se entende que pode ocorrer através de três níveis:

- o cognitivo, no qual a pessoa exposta a novas ideias, adquire conhecimento e muda os seus padrões de pensamento;
- o comportamental, no qual ocorre a interiorização de novos códigos de acção e mudança de comportamento e
- o incremento do desempenho, quando a mudança de comportamento é reflectida na melhoria dos resultados.

Voltando, agora, ao nosso problema específico, podemos resumir quatro obstáculos epistemológicos que travam a transição para o estádio de *Abertura ao Analista.*

O primeiro, relacionado com o nível comportamental, manifesta-se no receio de perder o poder que a posse da informação, ilusoriamente, traz. Este problema já se terá manifestado nas organizações em geral quando a microinformática invadiu a área antes exclusiva dos grandes sistemas computacionais levando a que os utilizadores se transformassem nos efectivos

possuidores da sua informação. De certa forma, é necessário operar igual transformação ao nível dos chamados SIE, transformando os trabalhadores dos mesmos em verdadeiros agentes da informação para os agentes económicos.

O segundo obstáculo epistemológico identificado relaciona-se com o incremento do desempenho e deriva da mentalidade que confere à questão da privacidade um sentido absoluto impedindo aos analistas o acesso aos dados de detalhe indispensáveis ao processo de mineração dos mesmos e levando com isso a que, por esta via, esta actividade não possa ser executada. É possível, como afirmámos, fazer pequenas transformações aos dados disponibilizados de modo a apagar vestígios que levem a identificar as unidades estatísticas a que respeitam e, deste modo, garantir privacidade e garantir acesso.

O terceiro obstáculo epistemológico, também ao nível do incremento do desempenho, tem a ver com um preconceito de acordo com o qual os SIE deverão, pura e simplesmente, proceder a apuramentos estatísticos, sendo sua missão publicar apenas os resultados dos mesmos, não lhes cabendo, portanto, proceder à mineração de dados e descoberta de conhecimento decorrente.

Há que notar que a junção do segundo e terceiro obstáculo se transforma num efectivo bloqueio ao processamento analítico em Economia: este ou se concretiza na origem o que é contrariado pelo preconceito *apuramento*, ou se concretiza no domínio do analista o que é contrariado pelo preconceito *privacidade*. Parece tornar-se necessário o repensar estrategicamente a missão dos SIE tendo em conta as novas potencialidades a nível de tratamento dos dados e da sua comunicação.

Finalmente, um quarto obstáculo, ao nível do cognitivo, materializa-se na relutância, por parte dos responsáveis pela formação em Economia, em reconhecer valor de uso às matérias de sistemas de informação e em particular às que temos vindo a tratar. O mais grave desta posição manifesta-se no facto de, ao recusar esta área de actividade, o economista estar a deixá-la para pessoas de outras áreas de conhecimento com menor sensibilidade à Economia e portadores de uma visão fechada por vezes limitada aos instrumentos informáticos e sua utilização. A verdade é que parece-nos

necessário criar um profissional diferente que alie um elevado conhecimento das matérias de Economia e Estatística à utilização de conceitos, métodos e técnicas oriundos da área de Sistemas de Informação. Tal seria enriquecer em termos do proveito associado à utilização dos instrumentos, mas também pela criação de adicionais oportunidades de exercício de actividade aos economistas.

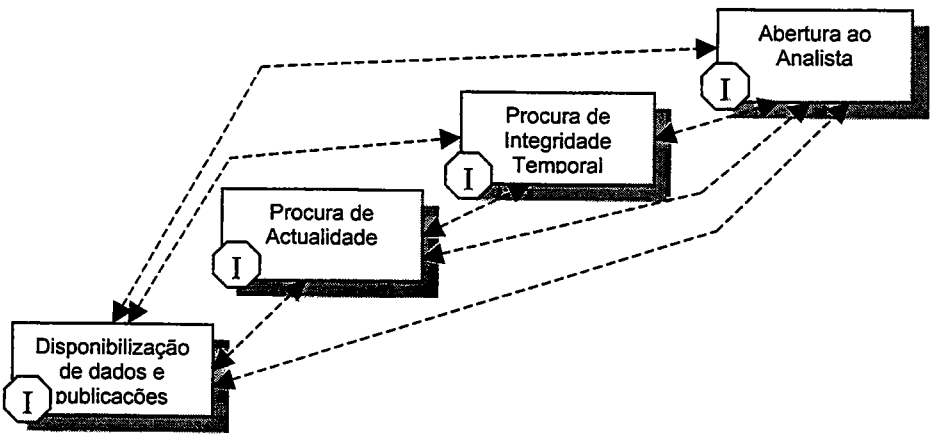
Em conclusão, a actividade dos SIE (no contexto dos princípios reguladores e da gestão), os trabalhadores de conhecimento no seu interior e o padrão de formação do economista devem ser sujeitos a um processo de transformação cultural que permita que o estádio de *Abertura ao analista* não seja mais uma utopia mas algo que terá uma concretização fácil num futuro próximo. Assim, a solução do problema está no interior das próprias organizações embora possa existir necessidade de ser catalisada a partir do exterior.

5.4. Conclusão sobre os estádios de desenvolvimento dos SIE na era das redes

Os resultados anteriores levam-nos a rejeitar a hipótese de uma sequência determinística entre os estádios identificados: disponibilização de dados e publicações, procura de actualidade, procura de integridade temporal e abertura ao analista. Isto significa que, partindo do conhecimento organizado em torno do desenvolvimento anterior, um novo SIE pode, desde logo situar-se num estádio de *Abertura ao Analista* ou em qualquer outro estádio diferente do inicial.

Graficamente podemos visualizar um SIE desde a sua génese nos moldes seguintes.

Figura 69. Evolução de um SIE em estádios



A figura anterior tem as seguintes implicações relativamente à maturidade dos SIE:

- as características da disponibilização podem ser vistas em termos gradativos onde as manchas marcantes, equivalentes aos estádios, poderão ser definidas de acordo com o quadro seguinte.

Quadro 97. Caracterização dos estádios em termos de variáveis de análise

	Disponibilização de dados e publicações	Procura de Actualidade	Procura de Integridade Temporal	Abertura ao Analista
Dados e Publicações	E	E	E	E
Actualidade	B	E	E	E
Integridade Temporal	B	B	E	E
Multidimensionalidade	B	B	M	E
Informação	B	B	M	E
Processamento Posterior	M	M	M	E
Abertura	B	B	B	E
Navegação	M	M	M	E
Diversidade	I	I	I	I

Legenda: I – indefinido; B – baixo; M – Médio; E - Elevado

- o processo de disponibilização interactiva pode iniciar-se em qualquer estádio, o que é marcado pelo I rodeado por um círculo, sendo as suas características que permitirão associá-lo a um em particular;
- a transição entre estádios, marcada em termos de linhas tracejadas com duas direcções, é livre, isto é
    - o beneficiando do conhecimento acumulado em processos anteriores por outras organizações, é possível passar de qualquer estádio para outro que lhe seja superior, mas também
    - o pode existir retorno de um estádio para outro menos evoluído na cadeia considerada o que é suportado empíricamente e pelo modelo de análise utilizado.

Podemos assim concluir que a vontade dos agentes com o objectivo de transformar os SIE e as acções decorrentes da mesma podem ter uma importância elevada para a evolução de maturidade dos SIE. Por outro lado,

detecta-se que as características associadas ao processamento analítico, pelo seu impacto na informação disponibilizada, são cruciais no processo de análise de maturidade.

6. Contribuição para o desenvolvimento do estádio de *abertura ao analista*

A constituição de uma secção de Sistemas de Informação para a Economia no Observatório das Ciências e Tecnologias (OCT) tendo por objectivo a maturidade dos SIE, aparece-nos como detendo uma concretização viável e naturalmente enriquecedora das potencialidades dos que carecem de informação sobre a Economia. Tal constituição, se conseguir deter um apoio político forte, pode ser a maior ajuda ao desenvolvimento do estádio de *abertura ao analista.*

Entre as múltiplas responsabilidades da secção poderíamos integrar (1) a manutenção actualizada de roteiros de informação disponibilizada pelos SIE, (2) a operacionalização de um índice geral de maturidade e (3) a concretização mais forte da abertura ao analista.

A primeira das atribuições referidas, manutenção de roteiros de informação, justifica-se porque é possível inventariar e classificar a informação disponibilizada pelos SIE. Nós esboçámos esse exercício ao elaborarmos a matriz SIE / factos económicos, mas é possível ir a um nível bem mais detalhado que associe a um facto / tema económico todo um conjunto de textos e dados que existam nos diferentes SIE.

Os resultados da elaboração de roteiros de informação, disponibilizados em sítios da Web seriam elementos fundamentais para a pesquisa que todos os interessados utilizariam para acesso àquela informação.

Por outro lado, se é possível, nos termos já descritos, ter uma métrica acerca da maturidade, também o é proporcionar uma visão do desenvolvimento no tempo desta maturidade através de um *índice geral de maturidade.* Tal desenvolvimento pode concretizar-se através de um número índice cujo conteúdo não deve confundir-se com o índice de maturidade descrito no Capítulo V.

O índice de desenvolvimento da maturidade (IDM) tem uma forte conotação temporal e pode ser definido pela seguinte expressão:

$$IDM_t = IDM_{t-1} * \frac{VI_t}{VI_{t-1}} \qquad (67)$$

onde

IDM: índice de desenvolvimento de maturidade

VI: Valor Informacional dos SIE

t: momento

O índice de desenvolvimento de maturidade é assim o resultado de uma transformação do índice de desenvolvimento de maturidade no período anterior através do crescimento do valor informacional do que é disponibilizado pelos SIE.

Por seu turno, o cálculo do Valor Informacional (*VI*) deve ter em conta as transformações existentes no processo de disponibilização. Através de inventário operado pelo observatório de maturidade, sabe-se que, em determinado momento, existe um certo número de organizações que disponibilizam informação para a Economia. Representemos esse número por *z*. Para determinação do índice de desenvolvimento de maturidade serão considerados os *y* SIE de maior valor informacional num determinado momento. Assim,

$$VI_t = \sum_{i=1}^{y} VI_{t,i} \qquad (68)$$

onde *i* representa um número de ordem do SIE considerado.

O cálculo dos valores informacionais individuais no momento t para os z SIE deve considerar a existência de:

- *p* critérios para caracterizar globalmente um sítio, neste caso, um SIE;
- *m* critérios para caracterizar os dados de um SIE;

- n critérios para caracterizar as publicações de um SIE;
- r elementos informacionais materializados em dados (# de ficheiros de dados);
- u publicações (# de publicações).

Seja s um índice para o SIE. Nestas condições podemos utilizar a expressão seguinte para o valor informacional de um SIE.

$$VI_t^s = \sum_{i=1}^{p} CS_{t,i}^s * PCS_{t,i} + \sum_{j=1}^{r}\sum_{i=1}^{m} CD_{t,j,i}^s * PD_{t,i} + \sum_{j=1}^{u}\sum_{i=1}^{n} CP_{t,j,i}^s * PP_{t,i} \quad (69)$$

onde

CS – critérios de análise da maturidade do sítio,

PCS – ponderação de critérios de análise de maturidade do sítio,

CD – critérios de análise de maturidade dos dados,

PD – ponderação dos critérios de análise de maturidade dos dados,

CP – critérios de análise de maturidade das publicações,

PP – ponderação de critérios de análise de maturidade das publicações,

e estando os ponderadores sujeitos à restrição:

$$\sum_{i=1}^{p} PCS_{t,i} + \sum_{i=1}^{m} PD_{t,i} + \sum_{i=1}^{n} PP_{t,i} = 1 \quad (70)$$

Chegámos assim a uma expressão que considera relativamente a um SIE um conjunto de critérios de análise gerais, relativos ao serviço de disponibilização, um conjunto de critérios que é aplicado a cada elemento disponibilizado sob a forma de dados e novo conjunto de critérios aplicados individualmente às publicações.

Por exemplo, se um SIE disponibilizar 5 ficheiros de dados e 3 publicações e p=3, m=5 e n=4 um resultado possível de análise é o seguinte.

Quadro 98. Caracterização do valor informacional de um SIE

SIE			Dados					Publicações			
1	0	1	1	0	0	0	0	0	1	1	1
			1	1	1	0	1	1	1	0	0
			1	1	1	0	0	1	1	1	1
			0	0	0	0	1				
			0	1	1	0	0				

Esta estrutura assimétrica conduz a uma expressão relativamente complexa para o valor informacional do SIE que pode ser simplificada. Com efeito, pode admitir-se uma representação matricial para os dados anteriores sujeita à anulação dos valores dos critérios de análise dos SIE para a linha 2 e seguintes e, igualmente, para as linhas de publicações que não existam. Tal facilidade não tem qualquer consequência a nível do *VI*, mas a expressão deste é extremamente simplificada.

$$VI_t^s = \sum_{i=1}^{p+n+m} C_{t,i}^s * P_{t,i} \qquad (71)$$

onde C e P representam agora critérios de análise e ponderadores genéricos e

$$\sum_{i=1}^{n+m+p} P_{t,i} = 1 \qquad (72)$$

Finalmente, a criação da secção relativa à maturidade dos SIE no OCT, permitiria uma forte abertura ao analista através da recepção de sugestões para a disponibilização de informação que se traduziria em encomendas aos SIE e que poderia traduzir-se em novos conteúdos ou em anteriores conteúdos utilizando diferentes formatos a que, adicionalmente, a realização de fóruns centrados em temas económicos possibilitaria um enriquecimento substancial.

Resumo do Capítulo VI

Neste capítulo procedemos à pesquisa dos estádios de desenvolvimento dos Sistemas de Informação para a Economia na era das redes para o que considerámos os passos seguintes:

- identificação das dimensões de análise da maturidade dos SIE;
- identificação dos segmentos de informação para a Economia;

- diferenciação dos segmentos de informação para a Economia e

- desenvolvimento de uma hipótese sobre os estádios de desenvolvimento dos Sistemas de Informação para a Economia.

O primeiro dos passos enunciados materializou-se no desenvolvimento de uma análise factorial pelo método das componentes principais que, partindo das variáveis inicialmente consideradas, procurou outras não correlacionadas entre si e que explicassem o máxima da variação. Tal obrigou a que, num primeiro momento, nos interrogássemos sobre a legitimidade de tal análise o que nos levou a examinar mais rigorosamente os dados em termos estatísticos sujeitando-os a um conjunto de testes, designadamente:

- um *teste à normalidade* das nove variáveis iniciais de resultado positivo;

- a análise do *diagrama de extremos e quartis* (*caixa de bigodes*) que permitiu a detecção de alguns *outliers* que decidimos manter na análise exactamente por considerarmos que os mesmos poderiam indiciar a localização em estádios diferentes;

- o exame da *matriz de correlações* pois a análise factorial exige uma elevada correlação, procedimento que se revelou pouco conclusivo;

- o *teste de esfericidade de Bartlett*, de acordo com o qual não existirá correlação entre as variáveis se a matriz de correlações for a matriz identidade, conduziu à rejeição da hipótese;

- a comparação das correlações simples com as parciais observadas entre as variáveis através do indicador designado por *Kaiser-Meier-Olkin* (KMO) cujo valor, não muito afastado de 1, permitiu considerar que se poderia prosseguir a análise factorial;

- A *matriz anti-imagem* que mede a adequação de cada variável para uso da análise factorial que mostrou que nenhuma variável tinha na diagonal principal um valor perto de zero levando à não exclusão de nenhuma da análise e

- o exame das *comunalidades* ou variância de cada variável explicada pelos factores comuns que não levou à exclusão de qualquer variável da

análise situação que se consideraria se apresentasse um valor entre 0 e 1 mais perto de 0.

Foram assim identificadas três dimensões da análise de maturidade dos SIE: a primeira, associada às *Características da Informação sem Multidimensionalidade* (*Processamento, Integridade Temporal e Actualidade*), a segunda, associada às *Características do Sítio* (*Diversidade* mas com contribuições importantes de *Dados e Publicações, Navegação, Abertura* e *Informação ao Analista*) e a terceira claramente relacionada com a *Multidimensionalidade.*

A identificação dos segmentos de informação para a Economia foi consumada através de uma *análise de clusters* que permitiu identificar três grandes grupos: o primeiro, com pontos fortes em *Actualidade* e *Integridade Temporal,* o segundo com valores abaixo da média em praticamente todas as variáveis de análise com excepção em *Dados e Publicações* e o terceiro com valores satisfatórios em *Actualidade, Dados e Publicações* e *Processamento.*

A utilização de técnicas estatísticas de diferenciação mais rigorosas considerando os resultados da primeira análise e da segunda permitiu associar o primeiro grupo a valores mais avançados, o segundo aos SIE em estádios mais atrasados e o terceiro à caracterização como *cluster de transição.*

Uma reflexão sobre os resultados anteriores levou-nos a admitir estar perante três estádios de desenvolvimento consumados no processo de desenvolvimento da disponibilização da informação para a Economia em Portugal no ano de 1999:

- uma inicialização no processo através de um estádio de *Disponibilização de dados e publicações*;
- a manifestação de preocupações de *Procura de actualidade* e
- a *Procura de integridade temporal.*

A manifestação comum de valores muito baixos nessa área de análise, levou-nos a considerar um quarto estádio, o da *Abertura ao Analista,* como aquele que se perfila no horizonte como estádio objectivo para os SIE.

Na posse de uma hipótese sobre o desenvolvimento do processo de disponibilização da informação para a Economia, examinámos a sua evolução em dois anos posteriores (2000 e 2001), tendo concluído que não se desenhava uma progressão para o estádio considerado como o desejável e que havia uma tendência para a diferenciação se esbater. Isso revelou-nos que tínhamos captado o fenómeno numa fase inicial que, pela sua riqueza de manifestações, teria exibido os vários estádios permitindo a formulação da hipótese avançada neste documento, mas permitindo-nos concluir que estes estádios não tinham associado um sentido de evolução determinística. Por outro lado, concluímos, em consonância com o operado em outras áreas do conhecimento, que é preciso uma mudança social e política muito forte para motivar um corte comportamental que leve ao estádio de *Abertura ao Analista.* Terminámos assim com o desenvolvimento de uma contribuição traduzida numa proposta de actuação para se atingir tal estádio: a criação e operacionalização de uma secção de sistemas de informação no Observatório das Ciências e Tecnologias, funcionando como observatório da maturidade dos SIE.

Conclusões

Ao longo deste documento, fomos estabelecendo um conjunto de conclusões importantes que chegou o momento de especificar. Recordamos que, na introdução a este documento, formulámos duas questões como as essenciais:

i. *Será possível identificar um conjunto de estádios na era das redes?*

ii. *Atingido um estádio será possível um retorno a um estádio anterior?*

As respostas que se pretendia deveriam ser suportadas pela informação disponibilizada interactivamente pelas organizações que disponibilizam informação sobre a economia.

O percurso seguido obrigou à concretização de um extenso plano de actividades que passou pela revisita aos critérios de análise de desenvolvimento da maturidade dos sistemas de informação, pelo relacionamento desses critérios com a problemática da sociedade do conhecimento e das novas necessidade em informação pelos agentes económicos, pelo exame do processo de disponibilização interactiva de informação para economia, pela pesquisa nos organismos que procedem a esta disponibilização de informação resultante de processos analíticos ou que possa ser integrada em processos analíticos, pela criação de indicadores relativos à disponibilização interactiva de informação para a economia e pela análise deste processo com estes indicadores e com instrumentos susceptíveis de percepcionarem uma evolução em estádios.

Resumimos os resultados da investigação em nove conclusões essenciais.

Da teoria

Os critérios herdados da teoria de estádios de desenvolvimento dos sistemas de informação não são suficientes para avaliação dos sistemas de informação para a economia no afirmar da sociedade do conhecimento suportada por novas tecnologias de informação e de comunicação

A teoria dos estádios de desenvolvimento dos sistemas de informação é responsável por uma lista de critérios destinados a avaliar a maturidade de um SI. Tal lista revela-se relativamente curta à entrada do século XXI.

As transformações económicas e sociais que conduziram a uma economia em rede, baseada no conhecimento, trouxeram múltiplas implicações em relação ao acesso, processamento e disponibilização de informação.

Em termos de acesso, é de referir a necessidade de comunicação entre sistemas diferentes levando à partilha de informação entre elementos que têm alguma relação no processo de transformação do valor. As empresas ligam-se aos seus clientes, aos seus fornecedores o que se traduz em requisitos importantes de acessibilidade e capacidade de processamento dos dados disponibilizados.

Em termos de processamento, é de realçar que, resolvidos os problemas da informação operacional, as organizações procuram reforçar o apoio aos níveis táctico e estratégico os quais se constituem em grandes preocupações em termos do seu esforço informacional. A integração e a abstracção implicam processos bem determinados de transformação de dados em conhecimento. A necessidade de prever o futuro carece da determinação de tendências a partir de dados que devem respeitar uma integridade temporal relativamente longa. Multidimensionalidade e descoberta de conhecimento tornam-se palavras cada vez mais pronunciadas no final da década de noventa.

Mas, se é importante toda esta capacidade de processamento interno, as organizações já não podem estar fechadas em si próprias. Elas participam numa extensa rede que não se compadece com a utilização do papel no suporte à comunicação. Assim, a convergência da informática e das telecomunicações terá possibilitado a transferência de conteúdos resultantes de processos de utilização de sistemas de informação interna ou externa, pondo-se de imediato o problema do seu processamento posterior nos sistemas de destino onde, conceptualmente, terão ainda necessidade de uma integração pacífica a nível de modelo de dados.

Assim, é de admitir a necessidade de considerar outros critérios para avaliar a maturidade dos sistemas de informação, critérios que poderão ser utilizados em conjugação com alguns dos provenientes da teoria de estádios, mas que se relacionam eminentemente com a nova problemática, designadamente, a viabilidade do processamento dos dados no destino, uma multidimensionalidade satisfatória e a garantia de integridade temporal.

Da consciência

Apesar da consciência do poder político relativamente ao papel da informação sobre a economia na sociedade do conhecimento, não existem medidas de fundo eficazes para garantir a sua utilização eficiente pelos agentes económicos e pelos analistas da economia

Neste documento, procedemos a uma descrição de alguns aspectos da disponibilização interactiva de informação para a Economia por um conjunto de fontes consideradas como as mais adequadas para tal efeito. Tal descrição foi suportada na compreensão de um processo de acesso aos dados e à informação que relaciona os que carecem da informação com os que a disponibilizam.

Uma primeira constatação é a de que, em termos do analista da Economia, as formas mais simples de disponibilização são as que melhor servem os seus interesses: isto é, informação detalhada em texto. Tal não invalida que, para outros utilizadores, se façam apuramentos imediatos e que os mesmos guiem as páginas dos SIE, mas não deverá ser esquecida a facilidade de extracção a um nível de detalhe satisfatório e com garantias de processamento posterior.

Por outro lado, existe plena consciência, a nível governamental, da importância que o recurso informação tem nas sociedades actuais e no papel determinante que os processos examinados assumem perante a sociedade do conhecimento. Apesar de tudo, constatámos que tal consciência não é acompanhada com medidas de fundo capazes de atingirem uma optimização dos benefícios para os agentes envolvidos. Tal como não basta encher um relatório com a palavra *inovação* para existir inovação a um nível satisfatório envolvendo o tecido empresarial, também uma simples ordem de transformação de publicações em papel em publicações digitais é insuficiente.

A ausência de medidas de fundo, que denuncia a inexistência de um planeamento para esse edifício que será a informação sobre a Economia, tem múltiplas manifestações:

- na redundância em termos de apuramento e difusão da informação: a matriz SIE / factos económicos revela que um mesmo facto económico é

tratado, por vezes, desde a instância apuramento em múltiplos departamentos, existindo, frequentemente, informação contraditória;

- na ausência de uma política quanto ao formato dos dados disponibilizados o que conduz a custos acrescidos a vários níveis: (1) no aspecto da disponibilização em detrimento da sua utilidade, (2) na necessidade de os que carecem dos dados os reintroduzirem em computador;

- na manutenção do fenómeno de disponibilização de informação incompleta pois os SIE, na sua quase generalidade, em vez de constituírem informação detalhada associado a um facto económico constituído em objecto de conhecimento, traduzida numa granularidade que suporte as dimensões associadas ao espaço, ao tempo, à operação, ao agente, fazem apuramentos incompletos por uma ou outra dimensão e ornamentam-nos como se fossem o único resultado possível.

Da contradição

É possível a existência de uma contradição entre, por um lado, os interesses de analistas da economia e agentes económicos e, por outro, os interesses da gestão de um sistema de informação para a economia.

Tal contradição, com íntimas relações com o que foi reportado anteriormente, pode manifestar-se nos moldes seguintes:

- o analista tem interesse em informação multidimensional, processável, com integridade temporal, extraída eficientemente, gratuita;

- a gestão do SIE afirma-se por obter a vanguarda em formas de apresentação e pode ceder a tentação de adivinhar as necessidades dos seus utilizadores, disponibilizando-lhe pequenas fatias de informação.

No entanto, em SIEs que privilegiam mais o interface com os seus utilizadores, esta contradição pode coexistir com a consciência da necessidade de proporcionar informação processável. Por exemplo, o ICEP entrega informação não processável ao utilizador comum, e aos utilizadores registados, a quem confere um *userid* e uma *password* após um pagamento, proporciona informação em formato folha de cálculo.

Da prática

Apesar da disponibilização de múltiplos elementos informacionais, não é significativa a existência de informação / conhecimento extraídos por processamento analítico em sistemas de informação para a economia ou a possibilidade da sua concretização nos agentes económicos ou pelos analistas da economia a partir dos dados disponibilizados interactivamente pelas mesmas organizações

Procedemos a uma pesquisa cujo objecto foi constituído pelos Sistemas de Informação para a Economia em termos de determinar a existência de informação ou aplicações que suportem o processamento analítico o que pode traduzir-se nas três modalidades seguintes:

1. informação em formato de texto ou folha de cálculo detentora de características de multidimensionalidade, capacidade de processamento, integridade temporal susceptível de ser transferida para a máquina do analista da Economia e tratada localmente,

2. aplicações de processamento analítico susceptíveis de disponibilizar resultados integráveis no que se designa por interrogação, análise multidimensional e descoberta de conhecimento em bases de dados e

3. informação por contacto directo com o SIE que viabilize o processamento analítico.

Neste momento, podemos afirmar que são escassos os elementos que se podem considerar como respondendo positivamente à busca operada apesar de, pela informação resultante da terceira modalidade considerada, eles poderem existir.

Num total de dezasseis organizações que elegemos como as mais indicadas para a disponibilização de informação para a Economia, num conjunto de mais de duas dezenas de áreas temáticas e de múltiplas facetas do processamento analítico, identificámos um caso de interrogação, dois casos de análise multidimensional, dois processos híbridos e um caso de pesquisa de associações.

O caso que classificámos como de interrogação tem a sua sede no INE ao nível de informação rápida sobre unidades territoriais e é detentor de uma grande versatilidade a nível da informação propiciada devido ao elevado número de indicadores que integra. É possível identificar alguns processos semelhantes no ICEP ou noutros SIE, mas não os considerámos muito relevantes.

Os dois casos de análise multidimensional tiveram uma origem bem diferente entre si. Um deles, o modelo de análise multidimensional para o IDE, resulta de uma pesquisa directa junto do fornecedor de informação e não tem réplica possível a nível de disponibilização interactiva. Tanto o GEPE como o Banco de Portugal publicam a informação sobre o IDE como um produto acabado, sem características de multidimensionalidade. O segundo caso, o modelo de análise multidimensional para o comércio externo, é uma feliz incursão do GEPE na área da análise multidimensional suportada através de recursos da Internet. São enormes as potencialidades de pesquisa que fornece. Apresenta, contudo, algumas limitações: não podemos aqui esquecer a impossibilidade de definir novos indicadores e até alguma degradação do tempo de resposta ou, pelo menos, uma instabilidade neste suscitada, eventualmente, por acessos estranhos ao produto. Isto permite-nos pôr a questão: apesar de se tratar de uma utilização muito positiva não seria melhor fornecer os dados em texto com a multidimensionalidade necessária e deixar que fosse o analista a tratá-los? A resposta parece-nos óbvia, mas a solução é um tanto perversa para o GEPE. Na situação actual, ele surge com a glória de ser o primeiro SIE a disponibilizar um produto multidimensional na Internet em Portugal - parece a vanguarda do processo. Nas condições que referimos, não passaria de mais um SIE a fornecer dados nada o distinguindo de outros.

Os dois processos híbridos que registámos combinam características de processos de interrogação com processos de análise multidimensional, mas falta-lhes algum elemento que nos permita integrá-los nesta. Mais uma vez a informação sobre unidades territoriais é considerada agora por apresentar facilidades de *drill-down* viável numa perspectiva diferente da referida para o primeiro caso. O segundo processo híbrido foi encontrado na Direcção Geral de Contribuições e Impostos a propósito da cobrança do IRS. Também a este processo faltam algumas características para o integrar em processo de análise multidimensional.

O último caso, pesquisa de associações na análise do IDE, foi possível a partir de dados solicitados ao ICEP e disponibilizados por esta organização. Há plena consciência de que os mesmos não poderiam ser encontrados na Internet.

Naturalmente, ao longo da pesquisa tomámos contacto com outros dados susceptíveis de serem transferidos para a máquina do analista e tratados localmente, embora sem características de multidimensionalidade.

Nestas condições podemos desde já formular uma primeira conclusão para este capítulo: em Portugal a informação disponibilizada interactivamente através dos SIE ainda não tem condições para apoiar uma analista da Economia em termos de análise multidimensional e descoberta de conhecimento em bases de dados. Falta à informação disponibilizada multidimensionalidade, capacidade de processamento, nível de detalhe e integridade temporal.

Apesar de tudo, aquela conclusão foi acompanhada de uma outra que deve ser registada: existe muita informação disponibilizada interactivamente em Portugal. Ela procura informar imediatamente, mas não transmite capacidade de processamento ao analista. Receamos assim que o que esteja em causa seja a questão do poder associado à informação. Ele *cai na rua* se os dados forem disponibilizados nos moldes que preconizamos.

Da descoberta de conhecimento

A informação sobre a economia constitui-se num enorme armazém de dados (data warehouse), mas tal facto não é suficiente para a existência de processos de descoberta de conhecimento suportados por algoritmos de mineração de dados (data mining).

Esta conclusão não era visível para nós ao inicializar esta dissertação começando a ganhar forma na elaboração do Capítulo III e a constituir-se como uma conclusão essencial da mesma.

A *descoberta de conhecimento em bases de dados, ao procurar novas e não antecipadas relações entre as variáveis, carece de informação detalhada. Não é essa a natureza da informação sobre a Economia, designadamente a respeitante ao nível macroeconómico. A agregação esconde diferenças, cria relações que podem não ser reais. Então, a descoberta de conhecimento em*

bases de dados, materializada através dos algoritmos de data mining, não é viável em termos de uma grande fatia da informação sobre a Economia.

Ultrapassar este problema implica uma ou algumas modificações suscitadas por uma visão estratégica diferente relativamente aos SIE: a utilização massiva de processos de anonimização a nível de microdados com a sua consequente disponibilização à generalidade dos analistas e/ou o abandono do preconceito *exclusivamente apuramento estatístico* com a possibilidade de tais processos serem efectuados no seu interior, o que implica a existência de técnicos com uma formação rica em Economia e em Sistemas de Informação, designadamente na área de mineração de dados.

Do avanço e do retrocesso

O processo de desenvolvimento da disponibilização interactiva de informação sobre a economia pode ser medido considerando transformações temporais que indiciem retrocesso, ultrapassando-se, assim, a contribuição dos autores da teoria de estádios

É possível partir da presença ou ausência de um conjunto de características nos sistemas que disponibilizam informação sobre a Economia com o objectivo de gerar modelos de avaliação e predição associados à maturidade dos SIE.

Neste documento, ficou demonstrada a viabilidade de construção de uma métrica que, baseando-se em dados tipo booleano, chegasse a um valor relativo ao estado de um SIE em termos de maturidade. Mas, se esse resultado já é importante, o facto de essa métrica integrar a possibilidade de retrocesso confere-lhe um interesse substancialmente superior.

Parte do que descrevemos pertence a autores associados à teoria de estádios dos sistemas de informação. No limite, todo o gestor de um sistema de informação pode definir um conjunto de características para o levar a determinar o grau de concretização de determinada variável.

Aquilo que entendemos que é substancialmente novo neste documento é a possibilidade de retrocesso. Embora o senso comum permitisse concluir da existência de sistemas de informação que terão passado de uma classificação muito elevada para um valor inferior, em termos teóricos não existia qualquer instrumento que permitisse compreender a razão de tal facto. Nós

introduzimos o conceito de ciclo de vida de variável de análise e, a partir desse momento, tornou-se possível seguir a evolução histórica de um SIE em termos de maturidade.

Da predição

Existindo informação relativamente ao desenvolvimento do processo em termos de maturidade, é possível conceber processos analíticos de predição que avaliem a maturidade dos SIE a partir de um número reduzido de critérios de análise.

Partimos de dados relativos à maturidade de dezasseis organizações que disponibilizam informação para a Economia relativos a *Navegação, Abertura ao Analista, Informação ao Analista, Multidimensionalidade, Disponibilização de Dados e Publicações, Integridade Temporal, Actualidade, Diversidade Temática, Processamento e Maturidade.* O valor relativo a Maturidade era uma combinação linear dos restantes.

O problema que se punha era o de chegar a um valor sem diferença significativa para a Maturidade das diferentes organizações, utilizando uma redução no número de variáveis de análise.

Através de uma associação dos valores originais a uma escala ordinal gerámos o input para a obtenção de uma árvore de decisão (utilizando um conjunto de parâmetros na ferramenta Knowledge Seeker conforme descrevemos no volume Anexo)

A árvore gerada tinha capacidade para predizer o valor de maturidade das organizações a partir de apenas duas variáveis originais. Procurámos determinar se esta característica se mantinha se as regras derivadas da árvore fossem inseridas num modelo de lógica difusa. Relativamente aos dados originais concluímos por uma resposta afirmativa. O modelo podia ser aplicado em termos difusos, não existindo uma diferença significativa entre os resultado originais e os gerados pelo modelo difuso o que foi comprovado por um teste de diferença de médias em variáveis emparelhadas com um nível de confiança de 95%.

Mas o verdadeiro teste a este procedimento foi proporcionado pela observação dos valores dos SIE considerados no final de 2000 cujos resultados

integravam uma mudança de situação abrupta pelo menos em um dos casos: a protagonizada pelo Gabinete de Estudos e Prospectiva para a Economia (GEPE) com uma transformação notável na disponibilização interactiva de dados estatísticos que chegou a integrar possibilidade de interrogação com utilização de uma ferramenta multidimensional uma base de dados sobre o Comércio Externo garantindo a integridade temporal para um período entre 1980 e 1998, facto fora do comum em todo o processo de disponibilização interactiva de informação estatística em Portugal. Este facto foi captado pelo modelo elaborado relativo a 1999 e mais uma vez foi possível predizer a maturidade dos SIE em 2000 utilizando apenas duas variáveis seleccionadas com os dados relativos a 1999.

Apesar de captar uma mudança abrupta nos valores observados, pusemos o problema da real capacidade de predição do modelo e analisámo-la utilizando uma matriz de acertos e erros. Os resultados encontrados são por nós classificados estarem numa zona de transição: o modelo não pode ser recusado para uma exigência de acerto inferior a 65% mas, se esta for superior, necessita ser refeito.

Parece-nos, assim, legítima a proposta de uma metodologia para a análise em termos difusos da maturidade dos SIE que integre todas estas preocupações e fizemo-lo considerando as seguintes etapas: (1) avaliação inicial de maturidade, (2) geração de árvore de decisão, (3) derivação de regras de decisão, (4) inclusão de regras em modelo difuso, (5) validação do modelo difuso perante os dados originais, (6) obtenção de dados novos, (7) utilização do modelo difuso com os dados novos, (8) análise de acertos e erros e conclusão sobre a validade de aplicação do modelo e (9) interpretação e publicação de resultados.

Esta metodologia, com as devidas adaptações, pode ser aplicada em outras áreas de análise de dados sobre a Economia em que se pretenda experimentar modelos difusos com a respectiva adaptação, demonstrando-se mais uma vez a viabilidade de aplicação de resultados provenientes da área de Inteligência Artificial à área da Economia.

Dos estádios

A evolução dos sistemas de informação para a economia, na era das redes, numa perspectiva de suporte à sociedade do conhecimento, segue um conjunto de quatro estádios que se iniciam com a disponibilização interactiva de dados e publicações e culminam na abertura ao analista.

Eis a principal conclusão a reter deste estudo: a ideia de uma evolução da globalidade dos SIE através de um conjunto de fases associadas às variáveis de análise.

Tal ideia é suportada pelos resultados aqui alcançados, tornando difícil uma análise estatística suportada pela existência de diferentes segmentos. Em 1999, os SIE diferenciam-se mais pelo estádio de evolução do que por diferentes características.

Identificámos assim quatro estádios para caracterizar em termos de maturidade o desenvolvimento dos sistemas de informação para a economia na era das redes. Era das redes que muitos designam por sociedade do conhecimento.

Esses estádios passam pela disponibilização de dados e publicações, procura de actualidade, integridade temporal e abertura ao analista.

A análise efectuada permitiu-nos concluir que esta transição não era linear, mas antes marcada por avanços e recuos que tinham a sua razão de ser em múltiplos factores um dos quais se relaciona obviamente com a gestão da organização que disponibiliza informação para a economia.

Assim, o grande passo para uma modificação futura do processo de disponibilização tem a ver com uma transformação cultural naqueles que coordenam o processo de disponibilização: é preciso operar um corte radical com as concepções do passado. É preciso pôr a informação ao serviço dos analistas, é preciso pô-la ao serviço da sociedade do conhecimento.

E neste ponto retomamos algumas das conclusões para as quais já se apontava em pontos anteriores: esta disponibilização acrescida de informação sobre a economia deverá passar pela revisão de padrões associados à

actualidade, à capacidade do seu processamento, ao formato, à integridade temporal.

A utilização de formas de acesso facilitadoras é decerto importante, mas as exigências da nova sociedade exigem ainda uma captação em linha das necessidades dos analistas e a manutenção de um sistema de diálogo que possibilite um controlo efectivo da sua satisfação.

Da destruição criadora

Uma condição necessária para os SIE atingirem o estádio de abertura ao analista é a sua concretização da destruição criadora.

Apesar de uma sucessão de estádios ser algo de natural em termos do processo de análise de maturidade, tal não significa que o alcançar de qualquer estádio esteja garantido num quadro de actuações que se limite à gestão corrente, a algum investimento e a esperar que o estádio se concretize.

O estádio de abertura ao analista representa uma ruptura com o passado recente. Ele implica que o utilizador do SIE seja posto no centro das preocupações de quem coordena o processo de disponibilização, ele implica um certo foco no cliente que não foi trivial no passado.

Há que criar novos produtos, novas formas de diálogo, novas formas de captação das necessidades. Há que compreender a extrema dependência dos agentes económicos e analistas relativamente à informação económica. Há que entender que o processo de extracção de informação e descoberta de conhecimento em bases de dados pode não suportar um plano prévio, pode exigir multifacetadas formas de interrogação e a sua dialéctica com novas questões.

É possível que a informação de hoje não seja a mais necessária e que tenha de ser destruído o seu processo de produção para outro ocupar o seu lugar. As organizações que disponibilizam informação sobre a economia têm de saber inserir-se em processos de destruição criativa de processos de produção de informação e de produtos de informação para garantirem um melhor serviço aos seus utilizadores.

Referências

Agrawal, Rakesh e Srikant, Ramakrishnan (1994). "Fast algorithms for mining association rules in large data bases". In *Proceedings of the Twentieth International Conference on Very Large Data Bases (VLDB'94)*, pp. 487-499. September. URL: http://www.acm.org/sigmod/vldb/conf/1994/P487.PDF.

Agrawal, Rakesh, Gupta, Asish e Sarawagi, Sunita (1996). *Modeling Multidimensional Databases*, Research report, IBM Almaden Research Center, San Jose, California. URL:

http://infolab.usc.edu/csci599/Fall2002/paper/I3_agrawal95modeling.pdf

Agrawal, Rakesh, Imielinski, Tomasz e Swami, Arun (1993). "Database mining: A performance perspective". In *IEEE Transactions on Knowledge and Data Engineering*, vol. 5, no. 6, December, pp. 914-925. Special Issue on Learning and Discovery in Knowledge-Based Databases. URL:
http://dlib.computer.org/tk/books/tk1993/pdf/k0914.pdf
ou
http://www.almaden.ibm.com/cs/people/ragrawal/papers/tkde93.ps

Agrawal, Rakesh, Imielinski, Tomasz e Swami, Arun (1993). "Mining Association Rules Between Sets of Items in Large Databases". In *Proceedings of the 1993 ACM SIGMOD Conference*, Washington DC, USA, May. URL:
http://www.almaden.ibm.com/cs/people/ragrawal/papers/sigmod93.ps

Agrawal, Rakesh, Mehta, Manish, Shafer, John, Srikant, Ramakrishnan, Arning, Andreas e Bollinger, Toni (1996). *The Quest Data Mining System*, IBM Almaden Research Group. URL:
http://www.almaden.ibm.com/cs/people/ragrawal/papers/kdd96_quest.ps

Ait-Kaci, Hassau e Nasr, Roger (1986). "LOGIN: A logic programming

language with built-in inheritance". In *Journal of Logic Programming*, no. 3, pp. 185-215.

Ali, Kamal, Manganaris, Stefanos e Srikant, Ramakrishnan (1997). *Partial Classification using Association Rules*, American Association for Artificial Intelligence.

Alter, Steven (1996). *Information Systems: a Management Perspective*, second edition, The Benjamin/Cummings Publishing Company Inc.

Althusser, L. (1967). *Sobre o Trabalho Teórico*, Biblioteca de Ciências Humanas, Editorial Presença, Lisboa.

Anderson, Scott e Walker, Edward (1995). *Building a SAS Data Warehouse*, SAS Institute Inc.

Anthony, R.N. (1965). *Planning and Control Systems: a Framework for Analysis*, Harvard University Press.

Bachelard, Gaston (1986). *O novo espírito científico*, Edições 70, Lisboa.

Barreto, António (1996). *A Situação Social em Portugal, 1960-1995*, pp. 21-33, Instituto de Ciências Sociais, Universidade de Lisboa.

Benbasat, IZAK e Zmud, Robert W. (1999). "Empirical research in information systems: The practice of relevance". In *MIS Quarterly*, March, Minneapolis.

Benbasat, Izak, Dexter, Albert S., Drury, Donald H. e Goldstein, Robert C. (1984). "A Critique of the Stage Hypothesis: Theory and Empirical Evidence". In *Communications of ACM*, vol. 27, n. 5, May.

Bhabuta, L. (1988). "Sustaing productivity and competitiveness by marshalling IT". In: *Proceedings: Information Technology Management for Productivity and Strategic Advantage*, IFIP TC-8 Open Conference, Singapore.

Bohn, Kathy (1997). "Converting Data for Warehouses". In *DBMS Online*. June. URL: http://www.dbmsmag.com/9706d15.html

Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984). Classification and Regression Trees, Wadsworth: Belmont, CA.

Buckingham, R.A., Hirshheim, F., Land, F. e Tally, C. J. (1987). *Information Systems Curriculum: A basis for course design in Information Systems Education: Recommendation and Implementation,* Cambridge University Press.

Cai, Yandong, Cercone, Nick e Han, Jiawei (1991). "Attribute-oriented induction in relational databases". In *Piatetsky-Shapiro et al (ed.),* pp. 213-228.

Carvalho, Joaquim Barradas de (1976). *Da História-crónica à história-ciência,* Livros Horizonte, Lisboa.

Cash, James I. (Jr.), McFarlan, F. Warren e McKenney, James L. (1992). *Corporate Information Systems Management - The Issues Facing Senior Executives,* Dow Jones-Irwin, Homewood, Illinois, USA.

Cavique, Luís e Themido, Isabel (2001). "SIMILIS um Novo Algoritmo para Análise do Cabaz de Compras", em *Apresentações de CAUS 2001 - 7ª Conferência Anual de Utilizadores SAS,* Lisboa.

Chamberlin, Donald D. e Boyce, Raymond F. (1974). "SEQUEL: A structured English Query Language.". In *Proceedings of ACM SIGMOD Workshop on Data Description, Access and Control,* pp. 249-264, Ann Arbor, Michigan.

Chen, P. P-S. (1976). "The Entity-Relationship Model - Toward a Unified View of Data". *In ACM Transactions on Database Systems* 1, March, pp. 9-36.

CISI (2001). *Método de Avaliação dos "web sites" da Administração Directa e Indirecta do Estado,* Anexo ao Caderno de Encargos, Observatório das Ciências e das Tecnologias, Julho de 2001. URL: http://www.cisi.mct.pt/ficheiros/ficheiros/concurso/conc004.rtf.zip

CISI (2002). *Avaliação Externa de Web Sites dos Organismos da Administração Directa e Indirecta do Estado - Relatório Final,*

Observatório das Ciências e das Tecnologias, Fevereiro de 2002. URL: http://www.si.mct.pt/site

CM (2001). *Resolução do Conselho de Ministros que prevê a avaliação dos sites na Internet de organismos integrados na administração directa e indirecta do Estado*, Resolução do Conselho de Ministros 22/2001. URL: http://www.mct.pt/novo/legislacao/despachos/aval.htm

Codd, E.F. (1970) "A Relational Model of data for Large Shared Data Banks". In *Communications of ACM*, vol. 13, no. 6, pp. 377-387. URL: http://www.acm.org/classics/nov95/toc.html

Codd, E.F., Codd, S.B. e Saley, C.T., (1993). *Providing OLAP (On-line Analytical Processing) to User Analysts: An IT Mandate*, White Paper from E.F.Codd & Associates. URL: http://www.essbase.com/whitepaper/olap/olap.pdf

Damas, Luís (1999). *SQL Structurd Query Language*, FCA - Editora de Informática, Lisboa.

Date, C. J. (1985). *An introduction to Database Systems*, sixth edition, Addison Wesley.

Davenport, Thomas e Prusak, Laurence (1998). *Working Knowledge: How organization manage what they know*, Harvard Business School Press, Boston, Massachusetts.

Davenport, Thomas H. (1994). "Saving IT's Soul: Human-Centered Information Management". In *Harvard Business Review*, vol. 72, no. 2, March-April, pp. 119-131.

DGCI (2001). *Apresentação, Morada e Estatísticas*, disponível em http://www.dgci.min-financas.pt/dgciappl/informacaoDGCI.nsf/Portugal?OpenNavigator, Outubro de 2001.

Earl, M. J. (1989). *Management Strategies for Information Technology*, Prentice Hall, Hemel Hempstead.

Emmerich, Thomas e Walker, Edward (1996). *SAS Institute's Rapid*

Warehousing Methodology, White paper from SAS Institute Inc.

Evans, Lisa (1995). *Information Delivery with the Sas System Part 2*, White paper from SAS Institute Inc.

Fayyad, Usama M., Piatetsky-Shapiro, Gregory, Smyth, Padhraic e Uthurusamy, Ramasamy (1996). *Advances in Knowledge Discovery and Data Mining*, AAAI Press / The MIT Press.

Fisher (1936). "The use of multiple measurements in taxonomic problems", In *Annals of Eugenics*, vol. 7, pp.179-188, The University of Adelaide Library, Australia. URL: http://www.library.adelaide.edu.au/digitised/fisher/138.pdf

Fontoura, Maria Paula (1997). *Factores Determinantes do Comércio Internacional: A Abordagem Empírica*, documentos de trabalho do CEDIN, 3/97.

Galliers, R. D. e Sutherland , A. R. (1994). "Information Systems management and Strategy formulation: applying and extending the 'stages of growth' concept". In: *Strategic Information Management*, Galliers, R. D. e Baker, B. S. H., Management Readers series, pp. 91-117, Butherworth/Heineman.

Galliers, R.D. (1987). *Information Analysis: Selected Readings*, Addison Wesley.

GEPE (2000). *BDCI – Notas Metodológicas*, manual disponibilizado pela Divisão de Bases de Dados, Outubro de 2001.

GEPE (2001). *Bases de Dados / Estatísticas*, disponível em http://www.gepe.pt/cgi-bin/redirect.asp, Outubro de 2001.

Gibson, C e Nolan, R. L. (1974). "Managing the four stages of EDP growth", *Harvard Business Review*, vol. 52, no. 1, January-February.

Gluck, F. W., Kaufman, S. P. e Walleck, A. S. (1980). "Strategic Management for Competitive Advantage", *Harvard Business Review*, vol. 58, no. 4, July-August.

Gray, Jim, Chaudhuri, Surajit, Bosworth, Adam, Layman, Andrew, Reichart, Don, Venkatrao, Murali, Pellow, Frank e Pirahesh, Hamid (1997). *Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab and Sub-totals*, Kluwer Academic Publishers. URL: http://research.microsoft.com/~gray/DataCube.doc

Grover, Varun, Goslar, Martin D. (1993). "The initiation, adoption, and implementation of telecommunications technologies in U.S. organizations". In *Journal of Management Information Systems,* vol. 10, no. 1, Summer, pp. 141-164, Armonk.

Gyssens, Marc e Lakshmanan, Laks V.S. (1997). "A Foundation for Multi-Dimensional Databases". In *Proceedings of the 23nd VLDB Conference.* Atenas, Grécia. URL: http://www.acm.org/sigmod/vldb/conf/1997/P106.PDF

Han, Jiawei, Fu, Yongjian, Wang, Wei, Koperski, Krzysztof e Zaiane, Osmar (1997). *DMQL: A Data Mining Query Language for Relational Databases.* Database Systems Research Laboratory, School of Computing Science, Simon Fraser University, Canada. URL: ftp://ftp.fas.sfu.ca/pub/cs/han/kdd/dmql96.ps

Hawkins, D.M. & Kass, G.V. (1982). "Automatic Interaction Detection". In *Topics in Applied Multivariate Analysis,* Hawkins, D.M. (ed.), pp. 267-302, Cambridge University Press, Cambridge.

Haykin, Simon (1994). *Neural Networks: A Comprehensive Foundation,* Macmillan College Publishing Company.

Hirschheim, R., Earl, M., Feeny, D. e Lockett, M. (1988). "An exploration into the management of the information systems function: key issues and an evolutionary model". In *Proceedings of the TC 8 Open Conference on Information Technology Management for Productivity and Strategic Advantage,* C. Yuen and G. Davis (eds.), pp. 4.15-4.38, Singapore.

Holsheimer, M. e Sieber, Arno (1994). *Data Mining: the search for knowledge in databases,* Computer Science / Department of Algorithmic and Architecture, CWI, Amsterdam. URL: http://www.cwi.nl/ftp/CWIreports/AA/CS-R9406.ps.Z

Houtsma, M e Swami, A. (1993). *Set-oriented mining of association rules.* Research Report RJ 9567, IBM Almaden Research Center, San Jose, California.

INE (1996). *Informação Estatística - Comércio Externo,* manual elaborado para utilização no âmbito de um programa de formação de quadros médios dos Institutos de Estatística dos PALOP.

INE (2001). *Metainformação,* disponível em http://www.ine.pt/prodserv/Meta/Metainf.html, Outubro de 2001.

INE (2001). *Pesquisa por Unidade Territorial,* disponível em http://infoline.ine.pt/inf/Prodserv/Pesqut/pesqut.asp, Outubro de 2001.

INE (2001). *Retrato de Unidades Territoriais,* disponível em http://infoline.ine.pt/inf/Prodserv/Retrato/retrato.asp, Outubro de 2001.

Information Discovery, Inc.(1996). *Data Mining with OLAP affinities.* URL: http://www.patternwarehouse.com/papers.htm

Inmon, W.H. e Hackathorn, Richard D. (1994.) *Using the Data Warehouse,* John Wiley & Sons, Inc.

Karimi, Jahangir, Gupta, Yash P. e Somers, Toni M. (1996). "Impact of competitive strategy and information technology maturity on firms' strategic response to globalisation". In *Journal of Management Information Systems,* vol. 12, no. 4, Spring, pp. 55-88, Armonk.

Kasabov, Nikola K. (1996). *Foundations of Neural Networks, Fuzzy Systems and Knowledge Engineering,* The MIT Press.

Keim, Daniel A., Kriegel, Hans-Peter e Seidl, Thomas (1993). *Supporting Data Mining of Large Databases by Visual Feedback Queries.* Institute for Computer Science, University of Munich. URL: http://www.dbs.informatik.unimuenchen.de/dbs/projekt/papers/datamining.ps

Kelly, Sean (1996). *Data Warehousing: The Route to Mass Customization,*

John Wiley and Sons.

Kenan Technologies (1995). *An Introduction to Multidimensional Database Technology*, Kenan Systems Corporation.

Kimball, Ralph (1996). "Data Warehouse Architect". In *Dbms Online*, August. URL: http://www.dbmsmag.com/9608d05.html

Kimball, Ralph (1997). "A Dimensional Modeling Manifesto". In *DBMS Online*, September. URL: http://www.dbmsmag.com/9708d15.html

Klemettinen, Mika, Mannila, Heikki, Ronkainen, Pirjo, Toivonen, Hannu, Verkamo, A. Inkeri (1994). "Finding Interesting Rules from Large Sets of Discovered Association Rules". In *the Third International Conference on Information and Knowledge Management*, ACM Press. URL: http://www.cs.helsinki.fi/research/fdk/datamining/pubs/cikm94.ps.gz

Korn, Flip, Labrinidis, Alexandros, Kotidis, Yannis, Faloutsos, Christos, Kaplunovich, Alex e Perkovié, Dejan (1996). *Quantifiable Data Mining Using Principal Component Analysis*. Department of Computer Science, University of Maryland at College Park. URL: http://www.isr.umd.edu/TechReports/ISR/1997/TR_97-25/TR_97-25.pdf

Lambin, Jean-Jacques (2000). *Marketing Estratégico*, Quarta edição, Mc-Graw Hill, Lisboa.

Lederer, A.L. e Mendelow, A.L. (1988). "Convincing top management of the strategic potential of information systems". In *MIS Quarterly*, December, Society for Information Management and The Management Information Systems Research Center, Minneapolis, USA.

Lendrevie, Jacques, Lindon, Denis, Dionísio, Pedro e Rodrigues, Vicente (1996). *Mercator - Teoria e prática do Marketing*, Publicações D. Quixote, Lisboa.

Li, Chang e Wang, X. Sean (1996). *A Data Model for Supporting On-Line Analytical Processing*. George Mason University. URL:

http://www.isse.gmu.edu/faculty/xywang/Papers/cikm96.ps

Mannila, Heikki (1996). *Data Mining: machine learning, statistics and databases*. Department of Computer Science, University of Helsinki. URL: http://www.cs.helsinki.fi/~mannila/postscripts/ssdbm.ps

Mannila, Heikki (1997). "Methods and problems in data mining". In *Proceedings of International Conference On Database Theory*, January, Greece. URL: http://www.cs.helsinki.fi/u/mannila/postscripts/icdt-tutorial.ps

Matos, Luís Salgado (1973). *Investimentos Estrangeiros em Portugal*, Seara Nova, Lisboa.

Mauritti, R., Martins, S., Antunes, A. e Costa, A. (2002). "Utilização de bases de microdados na investigação em ciências sociais". In *Revista de Estatística*, no. 27, Instituto Nacional de Estatística, Lisboa. URL: http://www.ine.pt/prodserv/estudos/debito.asp?x_estudoid=254

McFadden, Fred R. e Hoffer, Jeffrey A. (1994). *Modern Database Management*, fourth edition, Addison Wesley, World Student Series, The Benjamin/Cummings Publishing Company, Inc., Redwood City, Canada.

McFarlan, F. Warren e Nolan, Richard L. (1995). "How to Manage an IT Outsourcing Alliance". In *Sloan Management Review*, vol. 36, no. 2, Winter, pp. 9-23.

McIntyre, John, Moorman, Mark, King, Richard e Walker, Edward (1996). *Business Intelligence Systems and OLAP*, White paper from SAS Institute Inc.

Mendonça, António (1990). *Aspectos Teóricos do Comércio Internacional*, documentos de trabalho do CEDIN, 2/90.

Mendonça, António (1998). "O investimento directo estrangeiro em/de Portugal (1980/1996)". In *Comércio e Investimento Internacional*, edição ICEP, Portugal.

Mintzberg, Henry (1987). "The Strategy Concept I: Five Ps For Strategy". In

California Management Review , vol. 30, no. 1, Autumn, pp. 11-32.

Mintzberg, Henry (1994). "The Fall and Rise of Strategic Planning". In *Harvard Business Review*, vol. 72, no. 1, January-February, pp.107-114.

Morton, Steve e Anderson, Scott (1996). *Requirements-driven Data Modelling for the SAS*, White Paper from SAS Institute Inc.

MSI (1997). *Livro Verde para a Sociedade da Informação em Portugal*, Ministério das Ciências e das Tecnologias. URL: http://www.missao-si.mct.pt

Nolan, R.L. (1973). "Managing the computer resource: A stage hypothesis". In *Communications of ACM*, vol. 16, no. 7, pp. 399-405.

Nolan, R.L. (1973). "Plight of the EDP manager". In *Harvard Business Review*, vol. 51, no. 3, pp. 143-152.

Nolan, R.L. (1975). "Thoughts about the fifth stage". In *DATABASE*, vol. 7, no. 2, pp. 4-10.

Nolan, R.L. (1977). "Controlling the cost of data services". In *Harvard Business Review*, vol. 55, no. 4, July-August, pp. 114-124.

Nolan, R.L. (1978). "Organizational response and information technology". In *AFIPS National Computer Conference Proceedings*, vol. 47, pp. 517-524, Anaheim, California.

Nolan, R.L. (1979). "Managing the crises in data processing". In *Harvard Business Review*, vol. 57, no. 2, March-April.

Nolan, Richard L. e Croson, David C. (1995). *Creative Destruction - A Six-Stage Process for Transforming the Organization*, Harvard Business School Press, Boston, Massachusetts, USA.

Parsaye, K. (1996). "Rules are Much More Than Decision Trees". In *The Journal of Data Warehousing*, December. URL: http://www.patternwarehouse.com/trees.htm

Parsaye, Kamran (1996). *Olap and Data Mining – Bridging the Gap.* In *Database Programming and Design, February 1997.* URL: http://www.patternwarehouse.com/bridge.htm

Piatetsky-Shapiro, Gregory e Fawley, William J. (ed.). (1991) *Knowledge Discovery in Databases*, AAAI Press, Menlo Park, California.

Podehel, M. (1998). *Data Base Publishing on the Internet*, comunicação apresentada ao Seminário sobre *New Techniques & Technologies for Statistics*, URL: http://europa.eu.int/en/comm/eurostat/research/conferences/ntts-98/papers/cp/066c.pdf

Porter, M. e Millar, V. (1985). "How Information Gives You Competitive Advantage". In *Harvard Business Review*, vol. 63, no. 4, July-August, pp. 149-160.

Pyle, Dorian (1999). *Data Preparation for Data Mining*, Morgan Kaufmann Publishers, Inc.

Quinlan, J. Ross (1987). "Generating Production rules from decision trees". In *Proceedings of 10th International Joint Conference on Artificial Intelligence*, pp. 304-307, Milan.

Quinlan, J. Ross (1994). *Comparing connectionist and symbolic learning methods*, MIT Press.

Quinlan, J. Ross (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufman, San Mateo, CA.

Raden, Neil (1995). "Data, Data Everywhere". In *Information Week*, October. URL: http://home.aol.com/nraden/iw_mct01.htm

Raden, Neil (1995). *Multidimensional Modeling*, Archer Decision Sciences, Inc., Canada.

Ramakrishnan, Raghu e Gehrke, Johannes (2000). *Database Management Systems*, second edition, McGraw Hill.

Rascão, José (2000). *A análise estratégica e o sistema de informação para a*

tomada de decisão estratégica, Edições Sílabo, Lisboa.

Reis, Elizabeth e Moreira, Raul (1993). *Pesquisa de Mercados*. Edições Sílabo, Lisboa.

Reis, Elizabeth (1997). *Estatística Multivariada Aplicada*. Edições Sílabo, Lisboa.

Rodrigues, Eduardo (1997). *Data Warehouse: uma oportunidade de repensar a informação*, dissertação de Mestrado, ISEG.

Romão, António (1991). *O Comércio Externo Português*, documentos de trabalho do CEDIN, 2/91.

Ryssevik, J. e Musgrave S. (1998). *Virtual Data Libraries - the example of NESSTAR*, comunicação apresentada ao Seminário sobre *New Techniques & Technologies for Statistics*, URL: http://europa.eu.int/en/comm/eurostat/research/conferences/ntts-98/papers/cp/073c.pdf

Sakaguchi, Toru e Frolick, Mark N. (1997). *A Review of the Data Warehousing Literature*, University of Memphis. URL: http://www.nku.edu/~sakaguch/dw-web.htm

Saporta, Gilbert (2000). *Data Mining and Official Statistics*, comunicação apresentada a *Quinta Conferenza Nationale di Statistica*, ISTAT, Roma, 15 Novembro, URL: http://www.istat.it/Primpag/quintaconf/workshop/interventi/saporta.pdf ou http://cedric.cnam.fr/PUBLIS/RC184.pdf

Sarawagi, Sunita, Thomas, Shiby e Agrawal, Rakesh (1997). *Integration Association Rule Mining with Relational Database Systems: Alternatives and Implications*, IBM Research Division, Almaden Research Center, San Jose, California.

Schwartz, Peter, Kelly, Eammon e Boyer, Nicole (2001). "O despertar da Economia Global do Conhecimento". In *O futuro da Economia Global. Rumo a uma expansão duradoura?*, OCDE, edição e tradução do GEPE, Junho.

SECISI (2001). *Guia de Boas Práticas na Construção de Web Sites da Administração Directa e Indirecta do Estado*, Versão 2.0 de 22 de Julho de 2001, OCT – Observatório das Ciências e das Tecnologias do Ministério das Ciências e das Tecnologias. URL: http://www.cisi.mct.pt/ficheiros/ficheiros/concurso/conc007.doc.zip

Silvestre, António Luís (2000). *Análise de Dados: Estatística Descritiva*, editora Vulgata, Lisboa.

Stahl, Michael J. e Grigsby, David W. (1992). *Strategic Management - Formulation and Implementation*, PWS-KENT Publishing Company.

Storey, V.C. (1991). "Relational Database design based on the Entity Relationship Model", In *Data and Knowledge Engineering* vol. 7, pp. 47-83.

Toivonen, Hannu (1996). "Sampling Large Databases for Association Rules". In *Proceedings of the 22nd International Conference on Very Large Data Bases*, pp. 134-145, India. URL: http://www.acm.org/sigmod/vldb/conf/1996/P134.PDF

Turban, Efraim e Aronson, Jay E. (1998). *Decision Support Systems and Intelligent Systems*, fifth edition, Prentice-Hall International, Inc.

UNSC/ECE (2000). *Guidelines for Statistical Metadata on the Internet*, Conference of European Statisticians – Statistical Standards and Studies, n° 52, Geneva. URL: http://www.unece.org/stats/publications/metadata.pdf

UNSC/ECE (2001). *A guide to the websites of national and international statistical organizations*, Conference of European Statisticians - Methodological Material, Geneva. URL: http://www.unece.org/stats/publications/Webguide.pdf

UNSC/ECE (2001). *Best practices in designing websites for dissemination of statistics*, Conference of European Statisticians - Methodological Material, Geneva. URL: http://www.unece.org/stats/publications/websitebestpractice.pdf

UNSC/ECE (2001). *Recommendations on formats relevant to the downloading of statistical data from the Internet*, Conference of European Statisticians - Methodological Material, Geneva. URL: http://www.unece.org/stats/publications/downloadingformats.pdf

Vários (1993). *Strategor - Política Global da Empresa*, Publicações D. Quixote, Lisboa.

Vidette, Poe e Reeves, Laura (1996). *Building a Data Warehouse for Decision Support*, Prentice Hall.

Vieira, Luís M. Ribeiro (1997). *Sistema de apoio à decisão aplicável à determinação do estádio actual de um sistema de informação de acordo com as teses de Richard Nolan*, dissertação de Mestrado, ISEG.

Von Altrock, Constantin (1995). *Fuzzy Logic and NeuroFuzzy applications explained*, Prentice Hall PTR, Upper Saddle River, New Jersey, USA.

Von Altrock, Constantin (1997). *Fuzzy Logic and NeuroFuzzy applications in business and finance*, Prentice Hall PTR, Upper Saddle River, New Jersey, USA.

Ward, J. (1995). *Principles of Informations Systems Management*, Routledge, London.

Ward, J. Griffiths, P e Whitmore, P. (1990). *Strategic Planning for Information Systems*, John Wiley & Sons.

Zachman, J. (1987). "A Framework for Information System Architecture". In *IBM Systems Journal*, vol. 26, no. 3.

Zadeh L. (1965). "Fuzzy Sets". In *Information and Control*, no. 8, pp. 338-353.

Zadeh L. (1984). *Fuzzy Sets, Fuzzy Logic and Fuzzy Systems*, editado por George J. Klir e Bo Yuan, World Scientific Pub Co.

Zaki, M.J., Ogihara, M., Parthasarathy, S. e Li, W. (1996) *Parallel Data Mining for Association Rules on Shared-memory Multi-processors*, Technical Report 618, Department of Computer Science, University of Rochester. URL:

http://www.bib.informatik.tu-darmstadt.de/sc96/ZAKI/INDEX.HTM

Zimmermann, H.-J. and Zysno, (1983). "Decision Analysis and Evaluation by Hierarchical Aggregation of Informational". In *Fuzzy Sets and Systems*, vol. 10, pp. 243-266.

Zorrinho, Carlos (1991). *Gestão da Informação.* Biblioteca de Gestão Moderna, Editorial Presença, Lisboa.

Siglas utilizadas nas referências

ACM	Association for Computing Machinery
CISI	Comissão Interministerial para a Sociedade da Informação
CM	Conselho de Ministros
DGCI	Direcção Geral de Impostos
DPP	Departamento de Prospectiva e Planeamento
GEPE	Gabinete de Estudos e Prospectiva Económica
INE	Instituto Nacional de Estatística
MSI	Missão para a Sociedade da Informação
SECISI	Secretariado Executivo da Comissão Interministerial para Sociedade da Informação
UNSC/ECE	United Nations Statistical Comission and Economic Comission for Europe
URL	Uniform Resource Locator