

Análise Simbólica de Dados e a sua Aplicação na Extração
de Informação de Estatísticas Oficiais:
Análise do Inquérito à Ocupação do Tempo

por

Vera Maria Fernández de Campos

Dissertação apresentada como requisito
parcial para a obtenção do grau de

Mestre em Estatística e Gestão da Informação

pelo

Instituto Superior de Estatística e Gestão da Informação
da
Universidade Nova de Lisboa

Sob a orientação de
Maria Paula Brito
Faculdade de Economia da Universidade do Porto

Sob a co-orientação de
Carlos Marcelo
Instituto Nacional de Estatística

Dezembro / 2007

Agradecimentos

Ao finalizar a elaboração desta dissertação gostaria de agradecer a todos os que contribuíram para que este trabalho fosse concluído.

Gostaria de agradecer especialmente à Professora Paula Brito a forma como orientou o meu trabalho e a forma como sempre me recebeu e estimulou.

Finalmente, um agradecimento ao Dr. Carlos Marcelo pela orientação que me deu na parte prática deste trabalho.

Resumo

Pretende-se com o presente trabalho estudar e aplicar as técnicas da Análise de Dados Simbólicos. O tema tem como objectivo descrever e estudar métodos de Análise de Dados Simbólicos e a sua aplicação nas Estatísticas Oficiais. Como aplicação prática é analisado o estudo do Inquérito à Ocupação do Tempo, realizado pelo INE. Neste trabalho faz-se uma exposição dos conceitos gerais da Análise de Dados Simbólicos, nomeadamente os conceitos básicos, os tipos de variáveis simbólicas, a geração dos objectos simbólicos, as estatísticas descritivas e alguns métodos de classificação hierárquica e não-hierárquica. A análise dos dados parte de uma base de micro-dados que será trabalhada agregando os dados segundo diferentes níveis (macro-dados), de forma a caracterizar as regiões de Portugal, segmentadas por sexo e grupo etário, quanto à forma como ocupam o seu tempo diário, associando ainda o sentimento de pressa vivido nestas áreas.

Abstract

With this work we intend to explore and study the main methods of Symbolic Data Analysis. The objective is to describe and study the theory of Symbolic Data Analysis, having as practical application to the analysis of Time Use Survey, collected by the Portuguese National Institute of Statistics. In this work we make a detailed presentation of the main concepts of Symbolic Data Analysis, particularly the basic concepts, types of symbolic variables, generation of symbolic data, descriptive statistics, and some methods of hierarchical and non-hierarchical clustering. In the practical application, the statistical data set (micro-data) is aggregated in different levels (macro-data). The aim of this work is to study the occupation of time in the different regions in Portugal separated by the gender and age, associated with the individual's perception of feeling "pressed".

Conteúdo

Índice de Tabelas	10
Índice de Figuras	15
1 Introdução	17
1.1 Estrutura da Dissertação	17
1.2 Enquadramento	18
1.3 Objectivos e Metodologia	18
2 História e Evolução	21
2.1 Definições e Conceitos Gerais da Análise de Dados	22
3 Dados Simbólicos	25
3.1 Variáveis Simbólicas	25
3.2 Objectos Simbólicos	27
3.3 Geração de Objectos Simbólicos	32
3.3.1 Interacção com a Base de Dados	33
3.3.2 Operador Generalizador	36
4 Métodos Estatísticos de Análise de Dados para Dados Simbólicos	39
4.1 Estatísticas Descritivas	39
4.1.1 Variáveis Intervalares	40
4.1.2 Variáveis Modais	43

4.2	Análise Classificatória	43
4.2.1	Classificação Hierárquica pelo método Divisivo	44
4.2.2	Hierarquias e Pirâmides	48
4.2.3	Classificação Simbólica Não-Hierárquica	56
5	Descrição e Geração dos Dados	67
5.1	Inquérito à Ocupação do Tempo	67
5.2	Variáveis Clássicas	68
5.3	<i>SODAS</i> Software	72
5.4	Geração dos Dados Simbólicos	73
6	Análises Simbólicas	77
6.1	Estatísticas Descritivas	77
6.2	Análise Classificatória	104
6.2.1	Método Divisivo	104
6.2.2	Método em Hierarquia e Pirâmide	113
6.2.3	Classificação Não-Hierárquica: Método <i>SCLUST</i>	136
6.2.4	Comparação dos resultados obtidos métodos de classificação . .	148
7	Considerações finais	151
	Bibliografia	153
A	Notação	159

Lista de Tabelas

4.1	Matriz de frequências associadas às categorias	46
4.2	Operadores de generalização e medidas correspondentes	54
4.3	Matriz de frequências	62
6.1	Resumo das classes formadas pelo método DIV - <i>Região/Sexo</i> , variáveis intervalares	107
6.2	Resumo das classes formadas pelo método DIV - <i>Região/Sexo/Grupo Etário</i> , variáveis intervalares	108
6.3	Resumo das classes formadas pelo método DIV - <i>Região/Sexo</i> , variáveis modais . .	111
6.4	Resumo das classes formadas pelo método DIV - <i>Região/Sexo/Grupo Etário</i> , variáveis modais	112
6.5	Comparação da classificação em Hierarquia e em Pirâmide - <i>Região/Sexo</i> , variáveis intervalares	120
6.6	Comparação da classificação em Hierarquia e em Pirâmide - <i>Região/Sexo/Grupo Etário</i> , variáveis intervalares	120
6.7	Comparação da classificação em Hierarquia com a classificação em Pirâmide - <i>Região/Sexo</i> , variáveis modais	126
6.8	Comparação da classificação em Hierarquia e em Pirâmide - <i>Região/Sexo/Grupo Etário</i> , variáveis modais	127
6.9	Comparação da classificação em Hierarquia e em Pirâmide - <i>Região/Sexo/Grupo Etário</i> , faixas etárias mais jovens com variáveis intervalares	132
6.10	Comparação da classificação em Hierarquia e em Pirâmide - <i>Região/Sexo/Grupo Etário</i> , faixas etárias mais jovens com variáveis modais	135
6.11	Resumo das classes formadas pelo método <i>SCLUST</i> - <i>Região/Sexo</i> , variáveis intervalares	143

6.12	Resumo das classes formadas pelo método <i>SCLUST- Região/Sexo/Grupo Etário</i> , variáveis intervalares	143
6.13	Resumo das classes formadas pelo método <i>SCLUST- Região/Sexo</i> , variáveis modais	147
6.14	Resumo das classes formadas pelo método <i>SCLUST- Região/Sexo/Grupo Etário</i> , variáveis modais	148

Lista de Figuras

4.1	Simplificação de uma Pirâmide	55
5.1	<i>SODAS - chaining</i>	72
5.2	Excerto Base Dados Access - <i>Região/Sexo</i>	74
5.3	Excerto ficheiro <i>SODAS - Região/Sexo/Grupo Etário</i>	75
6.1	Distribuições de frequências das variáveis intervalares - <i>Região/Sexo</i>	78
6.2	Distribuições de frequências das variáveis intervalares - <i>Região/Sexo/Grupo Etário</i> .	79
6.3	Estrela 2D RN/F	79
6.4	Estrela 2D RLVT/F	80
6.5	Estrela 2D RA/F	80
6.6	Estrela 2D - Objecto Simbólico RLVT/M	81
6.7	Estrela 2D - Objecto Simbólico RN/M	81
6.8	Estrela 2D - Objecto Simbólico RA/M	82
6.9	Estrela 2D <i>Superimpose</i> - Objectos Simbólicos RN/F/15-24 e RN/F/25-34	83
6.10	Estrela 2D <i>Superimpose</i> - Objectos Simbólicos RN/M/15-24 e RN/M/25-34	83
6.11	Estrela 2D <i>Superimpose</i> - Objectos Simbólicos RLVT/F/15-24 e RLVT/F/25-34 . .	84
6.12	Estrela 2D <i>Superimpose</i> - Objectos Simbólicos RLVT/M/15-24 e RLVT/M/25-34 .	84
6.13	Estrela 2D <i>Superimpose</i> - Objectos Simbólicos RA/F/15-24 e RA/F/25-34	85
6.14	Estrela 2D <i>Superimpose</i> - Objectos Simbólicos RA/M/15-24 e RA/M/25-34	85
6.15	Estatísticas básicas das variáveis intervalares - <i>Região/Sexo</i>	86
6.16	Variável intervalar <i>Segunda Actividade - Região/Sexo</i>	87

6.17	Estatísticas descritivas clássicas das variáveis intervalares - <i>Região/Sexo</i>	88
6.18	Comparação das Médias das variáveis intervalares - <i>Região/Sexo</i>	88
6.19	Bi-plot das variáveis intervalares <i>Roupa e Calçado</i> e <i>Cuidar de crianças</i> - <i>Região/Sexo</i>	89
6.20	Bi-plot das variáveis intervalares <i>Lazer passivo</i> e <i>Estudo</i> - <i>Região/Sexo</i>	90
6.21	Estatísticas básicas das variáveis intervalares - <i>Região/Sexo/Grupo Etário</i>	90
6.22	Bi-plot das variáveis intervalares <i>Arranjo-casa</i> e <i>Prep-alimentos</i> - <i>Região/Sexo/Grupo Etário</i>	91
6.23	Bi-plot das variáveis intervalares <i>Lazer passivo</i> e <i>Dormir</i> - <i>Região/Sexo/Grupo Etário</i>	92
6.24	Capacidades das variáveis <i>Apressado</i> e <i>Freq-apressado</i> - <i>Região/Sexo</i>	93
6.25	Capacidades das variáveis <i>Apressado-5anos</i> , <i>Apressado-3anos</i> , <i>Apressado-1ano</i> - <i>Região/Sexo</i>	93
6.26	Capacidades das variáveis <i>vp-ntft</i> , <i>tem-fds</i> e <i>tempo-disp</i> - <i>Região/Sexo</i>	94
6.27	Min/Max/Med variáveis <i>Apressado</i> e <i>Freq-apressado</i> - <i>Região/Sexo</i>	94
6.28	Min/Max/Med variáveis <i>Apressado-5anos</i> , <i>Apressado-3anos</i> e <i>Apressado-1ano</i> - <i>Região/Sexo</i>	95
6.29	Min/Max/Med variáveis <i>vp-ntft</i> , <i>tem-fds</i> , <i>tempo-disp</i> - <i>Região/Sexo</i>	96
6.30	Capacidades das variáveis <i>Apressado</i> e <i>Freq-apressado</i> - <i>Região/Sexo/Grupo Etário</i>	96
6.31	Capacidades das variáveis <i>Apressado-5anos</i> , <i>Apressado-3anos</i> e <i>Apressado-1ano</i> - <i>Região/Sexo/Grupo Etário</i>	97
6.32	Capacidades das variáveis <i>vp-ntft</i> , <i>tem-fds</i> , <i>tempo-disp</i> - <i>Região/Sexo/Grupo Etário</i>	97
6.33	Min/Max/Med variáveis <i>Apressado</i> e <i>Freq-apressado</i> - <i>Região/Sexo/Grupo Etário</i> .	98
6.34	Min/Max/Med variáveis <i>Apressado-5anos</i> , <i>Apressado-3anos</i> e <i>Apressado-1ano</i> - <i>Região/Sexo/Grupo Etário</i>	98
6.35	Min/Max/Med variáveis <i>vp-ntft</i> , <i>tem-fds</i> , <i>tempo-disp</i> - <i>Região/Sexo/Grupo Etário</i> .	99
6.36	Estrela 3D <i>Superimpose</i> - Objectos Simbólicos RN/F, RC/F, RLVT/F	100
6.37	Estrela 3D <i>Superimpose</i> - Objectos Simbólicos RA/F, RALG/F	100
6.38	Estrela 3D <i>Superimpose</i> - Objectos Simbólicos RN/M, RC/M, RLVT/M	101
6.39	Estrela 3D <i>Superimpose</i> - Objectos Simbólicos RA/M, RALG/M	101

6.40	Estrela 3D <i>Superimpose</i> - Objectos Simbólicos - RALF/F/15-24 RA/F/15-24 RLVT/F/15-24 RC/F/15-24 RN/F/15-24	102
6.41	Estrela 3D <i>Superimpose</i> - Objectos Simbólicos - RALF/M/15-24 RA/M/15-24 RLVT/M/15-24 RC/M/15-24 RN/M/15-24	102
6.42	Estrela 3D <i>Superimpose</i> - Objectos Simbólicos - RALF/F/25-34 RA/F/25-34 RLVT/F/25-34 RC/F/25-34 RN/F/25-34	103
6.43	Estrela 3D <i>Superimpose</i> - Objectos Simbólicos - RALG/M/25-34 RA/M/25-34 RLVT/M/25-34 RC/M/25-34 RN/M/25-34	103
6.44	Classificação Divisiva - <i>Região/Sexo</i> , variáveis intervalares	105
6.45	Classificação Divisiva - <i>Região/Sexo/Grupo Etário</i> , variáveis intervalares	106
6.46	Classificação Divisiva - <i>Região/Sexo</i> , variáveis modais	109
6.47	Classificação Divisiva - <i>Região/Sexo/Grupo Etário</i> , variáveis modais	110
6.48	Classificação em Hierarquia - <i>Região/Sexo</i> , variáveis intervalares	113
6.49	Classificação em Hierarquia - <i>Região/Sexo</i> , classes 7/9 e 8/9	114
6.50	Classificação em Pirâmide - <i>Região/Sexo</i> , variáveis intervalares	115
6.51	Classificação em Pirâmide - <i>Região/Sexo</i> , classes 40/45 e 41/45	116
6.52	Classificação em Hierarquia - <i>Região/Sexo/Grupo Etário</i> , classes 31/39 e 38/39	117
6.53	Classificação em Hierarquia - <i>Região/Sexo/Grupo Etário</i> , classes 25/39 e 36/39	118
6.54	Classificação em Pirâmide - <i>Região/Sexo/Grupo Etário</i> , variáveis intervalares	119
6.55	Classificação em Hierarquia - <i>Região/Sexo</i> , variáveis modais	121
6.56	Estrela 3D <i>Superimpose</i> - classes 1/9 e 8/9	121
6.57	Pirâmide - <i>Região/Sexo</i> , variáveis modais	122
6.58	Estrela 3D <i>Superimpose</i> - classes 35/45 e 36/45	122
6.59	Classificação em Hierarquia - <i>Região/Sexo/Grupo Etário</i> , variáveis modais	123
6.60	Classificação em Hierarquia - <i>Região/Sexo/Grupo Etário</i> , classe 36/49	124
6.61	Classificação em Hierarquia - <i>Região/Sexo/Grupo Etário</i> , classes 46/39 e 47/49	124
6.62	Classificação em Pirâmide - <i>Região/Sexo/Grupo Etário</i> , variáveis modais	125
6.63	Classificação em Pirâmide - <i>Região/Sexo/Grupo Etário</i> , classes 355/460, 422/460, 451/460	126

6.64	Classificação em Hierarquia - <i>Região/Sexo/Grupo Etário</i> , variáveis intervalares . . .	128
6.65	Estrela - <i>Região/Sexo/Grupo Etário</i> , classes 17/19 e 18/19	128
6.66	Classificação em Pirâmide - <i>Região/Sexo/Grupo Etário</i> , variáveis intervalares . . .	129
6.67	“Pruning” da Pirâmide - <i>Região/Sexo/Grupo Etário</i> , variáveis intervalares	129
6.68	Classificação em Pirâmide - <i>Região/Sexo/Grupo Etário</i> , classes 17/114, 34/114 e 104/114	131
6.69	Hierarquia - <i>Região/Sexo/Grupo Etário</i> , variáveis modais	133
6.70	Hierarquia - classes 15/19 e 18/19	133
6.71	Pirâmide - <i>Região/Sexo/Grupo Etário</i> , variáveis modais	134
6.72	Pirâmide - classe 80/174	134
6.73	Pirâmide - classe 142/174	135
6.74	Pirâmide - classe 112/174	135
6.75	<i>SCLUST</i> - <i>Região/Sexo</i> - métodos de Calinski and Harabasz, C-index e Γ -index . .	136
6.76	<i>SCLUST</i> - <i>Região/Sexo</i> - variáveis intervalares	137
6.77	<i>SCLUST</i> - <i>Região/Sexo</i> - Índice de qualidade e contribuição das classes	137
6.78	<i>SCLUST</i> - <i>Região/Sexo</i> - Índice de qualidade das variáveis intervalares	138
6.79	<i>SCLUST</i> - <i>Região/Sexo</i> - Caracterização das classes pelas variáveis intervalares . .	138
6.80	<i>SCLUST</i> - <i>Região/Sexo/Grupo Etário</i> - métodos de Calinski and Harabasz, C-index e Γ -index	139
6.81	<i>SCLUST</i> - <i>Região/Sexo/Grupo Etário</i> - variáveis intervalares	140
6.82	<i>SCLUST</i> - <i>Região/Sexo/Grupo Etário</i> - Índice de qualidade e contribuição das classes	140
6.83	<i>SCLUST</i> - <i>Região/Sexo/Grupo Etário</i> - Índice de qualidade das variáveis intervalares	141
6.84	<i>SCLUST</i> - <i>Região/Sexo/Grupo Etário</i> - Caracterização das classes pelas variáveis intervalares	142
6.85	<i>SCLUST</i> - <i>Região/Sexo</i> - variáveis modais	144
6.86	<i>SCLUST</i> - <i>Região/Sexo</i> - Índice de qualidade e contribuição das classes	144
6.87	<i>SCLUST</i> - <i>Região/Sexo</i> - Índice de qualidade das variáveis modais	145
6.88	<i>SCLUST</i> - <i>Região/Sexo</i> - Caracterização das classes pelas variáveis modais	145
6.89	<i>SCLUST</i> - <i>Região/Sexo/Grupo Etário</i> - variáveis modais	146

6.90	<i>SCLUST - Região/Sexo/Grupo Etário - Índice de qualidade e contribuição das classes</i>	146
6.91	<i>SCLUST - Região/Sexo/Grupo Etário - Índice de qualidade das variáveis modais</i>	147
6.92	<i>SCLUST - Região/Sexo/Grupo Etário - Caracterização das classes pelas variáveis modais</i>	147

Capítulo 1

Introdução

1.1 Estrutura da Dissertação

No Capítulo 1 desta dissertação é feita uma introdução ao tema em estudo, a Análise de Dados Simbólicos, descrevendo-se os objectivos e metodologia utilizados neste trabalho. No Capítulo 2 fazemos um resumo histórico da teoria da Análise de Dados Simbólicos, onde são evidenciados alguns trabalhos importantes para o avanço desta área e se descrevem algumas aplicações práticas relevantes. O estudo teórico dos dados simbólicos e variáveis simbólicas está presente no Capítulo 3, assim como do processo de construção das classes e generalização dos objectos simbólicos. No Capítulo 4 são expostos os métodos estatísticos descritivos e de classificação para dados simbólicos, utilizados na análise dos dados. As estatísticas descritivas univariadas e bivariadas apresentadas na Secção 4.1, referem-se aos casos das variáveis intervalares e modais, de que nos ocuparemos neste trabalho. Os métodos de classificação discutidos na Secção 4.2, são a classificação hierárquica pelo Método Divisivo, a classificação hierárquica com representações em Pirâmide e Hierarquia e a classificação não hierárquica pelo Método das Nuvens Dinâmicas.

A aplicação prática das técnicas de Análise Simbólica ao Inquérito à Ocupação do Tempo é efectuada nos Capítulos 5 e 6. Na Secção 5.1 descrevem-se os questionários utilizados e na Secção 5.2 as variáveis clássicas escolhidas para análise. Na Secção 5.3 descreve-se o software *SODAS* utilizado no tratamento dos dados. A geração dos dados simbólicos a partir dos micro-dados é feita na Secção 5.4. Na Secção 6.1 são apresentados os resultados obtidos com as estatísticas descritivas e representações em estrela das descrições dos objectos simbólicos. Na Secção 6.2 são apresentados e comparados os resultados obtidos com a aplicação dos três métodos de classificação.

Por último os resultados são discutidos globalmente no Capítulo 7. O Capítulo 8 conclui a dissertação, apresentando perspectivas de trabalho futuro.

1.2 Enquadramento

Os dados simbólicos permitem generalizar os métodos de análise exploratória, estatística e gráfica a dados complexos, que ultrapassam o modelo tabular usual - em que a cada indivíduo, em linha, está associado um único valor de cada variável, em coluna. Estão neste caso os dados que descrevem classes ou grupos de indivíduos mais ou menos homogêneos - objectos de segunda ordem ou macro-dados - em vez de casos individuais - objectos de primeira ordem ou micro-dados. A Análise Simbólica de Dados (ASD) estende o modelo tabular usual permitindo valores múltiplos, eventualmente ponderados, para cada variável descritora, o que permite representar a variabilidade e/ou incerteza contida nos dados. Surgem assim as ditas “matrizes de dados simbólicos”, em que cada célula não contém necessariamente um valor único quantitativo ou uma categoria; uma célula pode conter, por exemplo, uma distribuição sobre um conjunto de categorias possíveis ou um intervalo de valores, ou ainda um conjunto de valores ligados por uma taxonomia ou por regras. Desta forma permite-se a descrição de unidades complexas ou conceitos, a obtenção de informação mais precisa ou a sumarização de extensos conjuntos de dados armazenados em grandes bases de dados.

Dados simbólicos ocorrem em muitas situações, por exemplo, ao sumariar grandes conjuntos de dados ou ao descrever conceitos subjacentes a uma base de dados - uma cidade, um grupo sócio-demográfico, um cenário de acidentes. Um conjunto de dados simbólicos pode ser estruturado de forma inerente ou resultar da agregação de um conjunto de dados clássicos.

1.3 Objectivos e Metodologia

Os métodos de Análise Simbólica de Dados são uma extensão da Análise de Dados tradicional, adequados para analisar “matrizes de dados simbólicos”. É uma área em pleno desenvolvimento que obteve um grande avanço com o Projecto Europeu “Symbolic Official Data Analysis System (*SODAS*)”, seguido do projecto “Analysis System of Symbolic Official data (*ASSO*)”.

O tema proposto vem no seguimento destes projectos, e pretende mostrar o interesse

da utilização das técnicas de Análise Simbólica e do software desenvolvido, na análise de informação produzida no âmbito das Estatísticas Oficiais. A Análise de Dados Simbólicos tem uma importante aplicação na área das Estatísticas Oficiais: como, por lei, os Institutos Nacionais de Estatística estão proibidos de divulgar informações individuais, os dados são previamente agregados por razões de confidencialidade (segredo estatístico) antes de serem distribuídos a qualquer outra agência governamental, indivíduo ou empresa; a ASD fornece instrumentos adequados para analisar estes dados agregados. De um modo geral, com este trabalho, pretende-se mostrar a grande utilidade e mais-valia da utilização destas técnicas na análise de Estatísticas Oficiais.

Em particular, será analisado o Inquérito à Ocupação do Tempo, realizado pelo Instituto Nacional de Estatística (INE). O estudo do inquérito pretende caracterizar as regiões de Portugal, segmentadas por sexo e grupo etário, quanto à forma como ocupam o seu tempo diário, associando ainda o sentimento de pressa vivido nestas áreas. Para este efeito são utilizadas medidas estatísticas de sumarização e visualização e vários métodos de classificação simbólica.

Para a elaboração deste trabalho foi necessária a cooperação do Instituto Nacional de Estatística, no que diz respeito ao acesso aos dados do Inquérito à Ocupação do Tempo. Para esse efeito, foi necessário efectuar a credenciação como investigadora junto do Observatório da Ciência e Ensino Superior (OCES) do Ministério da Ciência e Ensino Superior (MCES), que permitiu ao abrigo do protocolo INE/MCES ter acesso aos dados estatísticos pretendidos.

Capítulo 2

História e Evolução

Desde os primeiros artigos descrevendo os princípios básicos da Análise Simbólica de Dados, Diday em 1981, 1988 e 1989, até ao livro que irá ser agora publicado por [Diday e Noirhomme-Fraiture (2008)], foram desenvolvidas muitas técnicas nesta área. Os métodos de Análise Simbólica sofreram um grande avanço com projecto Europeu *SODAS* desenvolvido de 1997 a 1999, por 17 grupos de investigadores incluindo três Institutos Nacionais de Estatística (EUSTAT - Espanha, INE - Portugal e ONS - Inglaterra). O resultado deste projecto foi o desenvolvimento do software *SODAS* e a publicação de um livro [Bock e Diday (2000)]. O projecto ASSO, foi desenvolvido de 2001 a 2003 através da “*Information Society Technologies Program*” com o objectivo de explorar métodos e ferramentas de Software para resolver problemas de análise simbólica em bases de dados de Estatísticas Oficiais.

Métodos de Análise Simbólica de Dados incluem métodos de descrição univariada, métodos de visualização, classificação, análise discriminante, árvores de decisão, regressão, análise factorial, etc., que permitem analisar matrizes de dados simbólicos. No desenvolvimento teórico, trabalhos importantes impulsionaram o avanço de algumas técnicas da ASD, a título de exemplo poderemos citar: Cazes, P., Chouakria, A., Diday, E. e Schecktmann, Y. na análise factorial [Cazes *et al.* (1990)]; Gowda, C. e Diday, E. [Gowda e Diday (1992)] e De Carvalho [De Carvalho (1998)] no cálculo de dissimilaridades entre objectos simbólicos; Brito, P. [Brito (1991)], Chavent, M. [Chavent (1997)] e Rodriguez, O. [Rodriguez (2000)], nas técnicas de análise classificatória simbólica; Billard, L. e Diday, E. [Billard e Diday (2003)] publicam uma síntese da ASD no “*Journal of the American Statistical Association*”.

Aplicações efectuadas dizem respeito, por exemplo, à descrição de cenários de acidentes rodoviários [Diday (1993)]; aplicação de objectos simbólicos nas estatísticas oficiais

[Calvo (2000)]; estudo sobre a ocupação do tempo no País Basco [Mas e Olaeta (2002)]; o estudo de uma base de dados de 156 países agrupados por um certo número de conceitos segundo o nível de riqueza [Soussi (2003)]; o estudo na análise da gestão administrativa de uma cidade através das opiniões acerca dos serviços, escolas, tráfico [Lima e De Carvalho (2003)]. Foram já efectuados trabalhos em Portugal no âmbito das Estatísticas Oficiais, nomeadamente, ao nível da caracterização do emprego [Marcelo (2001)], e aplicação da ASD nas Estatísticas Oficiais [Marcelo (2002)].

Como ponto de partida, foi utilizado o estudo realizado por [Mas e Olaeta (2002)] no País Basco, a partir do Inquérito à Ocupação do Tempo realizado pelo Eustat (Instituto de Estatística do país Basco). A aplicação prática deste trabalho é feita com recurso ao software *SODAS*.

2.1 Definições e Conceitos Gerais da Análise de Dados

Os métodos da análise clássica foram desenhados para estudar dados sem complexidade e que descrevam uma situação simples. Os dados são obtidos a partir de indivíduos singulares, os quais são descritos por variáveis que tomam um único valor.

A análise clássica parte de um conjunto $\Omega = \{\omega_1, \dots, \omega_n\}$ de objectos ou indivíduos que são caracterizados por p variáveis $\tilde{Y}_1, \dots, \tilde{Y}_p$. Cada variável $\tilde{Y}_j$ toma valores no seu domínio ou espaço de observação $\mathcal{Y}_j$. Para cada indivíduo ω_i a variável realiza um único valor $\tilde{Y}_j(\omega_i)$. Uma variável clássica é definida por uma aplicação de Ω para $\mathcal{Y}_j$, tal que $x_{ij} = \tilde{Y}_j(\omega_i)$ é o valor observado para o indivíduo ω_i . A matriz $X = (x_{ij})$ reúne toda a informação, sendo o vector das p variáveis definido por $X = (\tilde{Y}_1, \dots, \tilde{Y}_p)'$ e o vector das p observações, para o indivíduo ω_i , definido por $X(i) = (x_{i1}, \dots, x_{ip})'$. Atendendo ao tipo de valores que $\mathcal{Y}_j$ pode tomar, as variáveis dividem-se em dois grandes grupos: quantitativas ou qualitativas.

Definição 2.1 *Variáveis quantitativas* são variáveis numéricas tal que o conjunto dos seus valores possíveis é idêntico ou está contido em $\mathbb{R}$. Se o conjunto dos valores possíveis é contínuo não-numerável a variável diz-se contínua, caso contrário diz-se discreta e toma valores num conjunto numerável (finito ou não).

Definição 2.2 *Variáveis qualitativas* são variáveis onde o conjunto dos seus valores possíveis é finito, sem significado numérico e com ou sem estrutura interna. Estas

variáveis subdividem-se em nominais (sem estrutura interna) e ordinais (existe uma ordem linear entre os valores possíveis), consideram-se ainda neste último caso as variáveis ordinais generalizadas, como as taxonomias (ver [Bock e Diday (2000a)]).

Sobre as variáveis quantitativas e qualitativas podem ser aplicadas relações lógicas dando origem às variáveis dependentes. Esta dependência pode ser lógica, hierárquica ou estocástica. No âmbito deste trabalho serão consideradas apenas as variáveis dependentes logicamente e hierarquicamente. Sobre a dependência estocástica consulte-se [Bock e Diday (2000a)].

Definição 2.3 *Dadas duas variáveis Y e Z diz-se que Z **depende logicamente** de Y se os valores que Z pode tomar dependem logicamente dos valores tomados por Y .*

Exemplo 2.1 *Se Y = tipo-família e Z = Bebés, tem-se uma possível regra de dependência,*

Se $Y(\omega)$ = Casal c / filhos então $Z(\omega) = \{1, 2, 3, \dots\}$

Definição 2.4 *Uma variável Z **depende hierarquicamente** de uma variável Y se Z não faz sentido para certos valores de Y . Considerem-se $\mathcal{Y}$ e $\mathcal{Z}$ os conjuntos onde as respectivas variáveis tomam valores e seja $\mathcal{Y}' \subseteq \mathcal{Y}$ o conjunto dos valores de $\mathcal{Y}$ para os quais Z não tem sentido. A dependência hierárquica, [Brito e De Carvalho (2002)], traduz-se pela regra:*

$$Y \text{ toma valores em } \mathcal{Y}' \Leftrightarrow Z \text{ é não aplicável}$$

Exemplo 2.2 *Se Y = “sente que anda apressado” com $\mathcal{Y} = \{\text{Sim}, \text{Não}\}$ e Z = “apressado relativamente ao ano passado” com $\mathcal{Z} = \{\text{na mesma}, \text{mais apressado}, \text{menos apressado}\}$ vem, para $\mathcal{Y}' = \{\text{não}\}$ que $Z = \text{NA}$.*

Capítulo 3

Dados Simbólicos

3.1 Variáveis Simbólicas

Como já foi referido anteriormente, na análise de dados clássica, as variáveis tomam um único valor ou categoria para cada indivíduo, enquanto que para os dados simbólicos as variáveis podem também assumir conjuntos de valores ou categorias, contendo uma variação interna. Se considerarmos um conjunto de dados de clássicos, cada observação corresponde a uma única realização da variável, ou seja, a representação no espaço p -dimensional corresponde a um único ponto, enquanto que para dados simbólicos cada realização será um hipercubo no espaço p -dimensional ou no produto cartesiano de distribuições.

Um conjunto de dados simbólicos é constituído pelos indivíduos que são denominados por unidades ou objectos, e pelas variáveis simbólicas que descrevem as unidades. As unidades ou objectos, podem ser por natureza dados estruturados ou podem ser obtidos a partir de dados clássicos através da agregação. A construção de classes a partir de unidades singulares é um conceito, “**paradigma dois níveis**”, que pode ser estendido ao terceiro nível ou mais. Inicialmente, considera-se o conjunto $\Omega = \{\omega_1, \dots, \omega_n\}$ formado pelos indivíduos ou objectos de primeira ordem descritos por variáveis clássicas ou “*single*”. O conjunto $E = \{u_1, \dots, u_k\}$ será o conjunto dos elementos ou objectos de segunda ordem, que neste caso são classes formadas pelos indivíduos de Ω . O objectivo é caracterizar como se comportam as classes definidas com respeito às variáveis clássicas. Para dar resposta a este problema definem-se as variáveis simbólicas que permitem descrever este tipo de objectos.

As variáveis podem assumir diversos tipos: quantitativa “*single*” e categórica “*single*”,

como na análise de dados clássica e ainda intervalar, categórica multivaluada ou modal, no caso da análise simbólica.

As definições que se seguem têm suporte em [Bock e Diday (2000a)].

As variáveis multivaluadas podem assumir todos os subconjuntos finitos do seu domínio, ou seja, assumem um conjunto de valores para cada unidade.

Uma variável simbólica Y diz-se intervalar se os seus valores são intervalos de $\mathbb{R}$, e diz-se categórica multivaluada se os seus valores são conjuntos finitos de categorias. A variável modal é o caso mais complexo, pois faz corresponder uma medida ou uma distribuição de probabilidades a cada objecto ou unidade.

Definição 3.1 *Uma variável simbólica Y_j com domínio subjacente $\mathcal{Y}_j$, é uma aplicação,*

$$\begin{aligned} Y_j : E &\longrightarrow B_j \\ u &\longrightarrow Y_j(u) \end{aligned} \tag{3.1}$$

com $u \in E$ (indivíduos ou classes) e B_j o conjunto onde a variável toma valores.

Conforme a especificação de B_j a variável pode assumir qualquer um dos tipos referidos anteriormente.

Definição 3.2 *Uma variável Y é quantitativa “single”, se $B_j = \mathcal{Y}_j$ é um subconjunto de $\mathbb{R}$.*

São consideradas variáveis quantitativas “single” as variáveis clássicas quantitativas.

Definição 3.3 *Uma variável Y é categórica “single”, se $B_j = \mathcal{Y}_j$ é um conjunto finito de categorias.*

São consideradas variáveis categóricas “single” as variáveis clássicas qualitativas.

Definição 3.4 *Uma variável Y é categórica multivaluada, se $\mathcal{Y}_j$ é um conjunto finito de categorias, $B_j = P(\mathcal{Y}_j)$ o conjunto das partes de $\mathcal{Y}_j$ não vazias e $Y_j(u) \subseteq \mathcal{Y}_j$.*

Exemplo 3.1 *Se $E = \{\text{Região Norte}, \text{Região Centro}, \dots\}$ com $Y = \text{“sente-se apressado”}$ e $\mathcal{Y} = \{\text{Sim}, \text{Não}\}$, $B = \{\{\text{Sim}\}, \{\text{Não}\}, \{\text{Sim}, \text{Não}\}\}$*

Definição 3.5 Uma variável Y_j é **intervalar**, se $\mathcal{Y}_j$ é $\mathbb{R}$ ou um intervalo em $\mathbb{R}$ e B_j é o conjunto dos intervalos de $\mathcal{Y}_j$, tal que para $u \in E$, $Y_j(u) = [\alpha, \beta]$ com $\alpha, \beta \in \mathcal{Y}_j$ e $\alpha \leq \beta$.

Exemplo 3.2 Seja $E = \{\text{Casalc/filhos}, \text{individual}, \text{monoparentalc/filhos}\}$
 $Y_1 = \text{“número de minutos a dormir por dia”}$ e $\mathcal{Y}_1 = [0; +\infty[$, pode-se obter para $u = \text{individual}$
 $Y_1(\text{individual}) = [300, 500]$

Definição 3.6 Uma variável Y_j é **modal** se $B_j = M(\mathcal{Y}_j)$ é uma família de distribuições de frequências ou pesos sobre $\mathcal{Y}_j$. Para qualquer $u \in E$, Y_j é uma aplicação,

$$Y_j(u) = (U(u), \pi_u)$$

com $U(u) \subseteq \mathcal{Y}_j$ o conjunto das categorias que o elemento pode tomar e π_u a medida ou a distribuição de frequências, probabilidades ou pesos, dos valores possíveis observados em $\mathcal{Y}_j$.

Exemplo 3.3 Seja $E = \{\text{Casalc/filhos}, \text{individual}, \text{monoparentalc/filhos}\}$
 $Y_1 = \text{“sente-se apressado”}$ e $\mathcal{Y}_1 = \{\text{Sim}, \text{Não}\}$, pode-se obter para $u = \text{Casalc/filhos}$
 $Y_1(\text{Casalc/filhos}) = (\text{Sim}(0.7), \text{Não}(0.3))$

Um caso bastante importante dentro das variáveis modais são as '**variáveis histograma**'. Consideremos Ω o conjunto dos indivíduos singulares e $\tilde{Y}$ uma variável quantitativa “single” com domínio $\mathcal{Y} = \mathbb{R}$. Seja E o conjunto de todos os subconjuntos não vazios de indivíduos e $C \in E$ uma classe de objectos singulares agregados. C pode assim ser descrita pela variável simbólica histograma, cujos valores $Y(C)$ são especificados por um *histograma* descrevendo a distribuição empírica de $\tilde{Y}$ em C .

3.2 Objectos Simbólicos

Uma **matriz de dados simbólicos**, por oposição à matriz clássica rectangular, pode conter assim em cada célula conjuntos, intervalos, histogramas, distribuições de frequências, etc. A matriz dos dados simbólicos reúne toda a informação e é representada por $\Xi = (\xi_{ij})$ com dimensão $k \times p$:

$$\Xi := \begin{pmatrix} \xi_{11} & \xi_{12} & \cdots & \xi_{1p} \\ \xi_{21} & \xi_{22} & \cdots & \xi_{2p} \\ \vdots & \vdots & & \vdots \\ \xi_{k1} & \xi_{k2} & \cdots & \xi_{kp} \end{pmatrix}$$

com $\xi_{ij} = Y_j(u_i)$ o valor observado da j -ésima variável para o objecto u_i .

Cada linha da matriz de dados clássica $X = (x_{ij})$ descreve um elemento $\omega_i \in \Omega$ com valores $\tilde{Y}_j(\omega_i) = x_{ij}$ do domínio $\mathcal{Y}_j$. De forma semelhante, para uma matriz de dados simbólicos $\Xi = (\xi_{ij})$, a linha $\xi_i = (\xi_{i1}, \dots, \xi_{ip})'$ descreve o elemento $u_i \in E$ pelos valores simbólicos $Y_j(u_i) = \xi_{ij}$ de um conjunto B_j que pode ser, $\mathcal{Y}_j$, $P(\mathcal{Y}_j)$ ou $M(\mathcal{Y}_j)$.

Se o vector $\Xi(u_i) := (Y_1(u_i), \dots, Y_p(u_i))'$ engloba todas as variáveis descritoras do objecto u_i então:

- cada objecto $u_i \in E$ é descrito pelo vector simbólico $(\xi_{i1}, \dots, \xi_{ip})'$;
- cada linha da matriz Ξ é uma **descrição simbólica** do objecto u_i dada pelas p variáveis.

Definição 3.7 *Seja Y_j , com $j = 1, \dots, p$ uma variável simbólica de domínio $\mathcal{Y}_j$ e B_j o conjunto onde a variável toma os seus valores.*

- *O **espaço das descrições** ou conjunto de todas as descrições é representado por $B = B_1 \times \dots \times B_p$.*
- *Considerando que $d_j \in B_j$ tem-se que $d = (d_1, \dots, d_p) \in B$ é uma **descrição** ou **vector descrição**.*

Para qualquer $u_i \in E$ a descrição $d_{u_i} = (d_{u_i1}, \dots, d_{u_ip})$ corresponde à linha da matriz simbólica $(Y_1(u_i), \dots, Y_p(u_i))' = (\xi_{i1}, \dots, \xi_{ip})'$. Cada valor observado por uma variável para determinado objecto corresponde a uma descrição parcial desse objecto.

Um objecto simbólico é caracterizado pelo vector de variáveis Y , uma descrição $d \in B$ e uma relação R que compara Y com d . O resultado dessa comparação é dado por uma função binária ou booleana a . Atendendo à definição dada por Diday (ver por exemplo [Diday e Noirhomme-Fraiture (2008)]), um objecto simbólico s_u está associado a um conceito $u \in E$, que se define pelas suas propriedades características (intenção de u) e pelo conjunto dos indivíduos de Ω que satisfazem essas propriedades (extensão de

u). Enquanto que R e d devolvem uma descrição intensional, ou seja, especificam as propriedades desejadas, a aplicação a e a sua extensão indicam o conjunto dos elementos que verificam essas propriedades.

Definição 3.8 Um *objecto simbólico*, [Bock e Diday (2000a)], é um triplete $s = (a, R, d)$ em que $d \in B$ é uma descrição, R é uma relação entre descrições e a é uma aplicação de $E \rightarrow L$ com $L=[0;1]$, que associa aos elementos $u \in E$ um valor de L dependendo da relação R .

A relação R permite determinar quais os elementos $u \in E$ que verificam a descrição d do objecto simbólico s .

Sejam $\mathcal{Y}, \mathcal{Z}$ dois conjuntos quaisquer e $\mathcal{Y} \times \mathcal{Z}$ o produto cartesiano contendo todos os pares (y, z) com $y \in \mathcal{Y}$ e $z \in \mathcal{Z}$.

Definição 3.9 Uma *relação* R está associada a uma função binária, definida no produto cartesiano $\mathcal{Y} \times \mathcal{Z}$, com $\phi(y, z) := [yRz]$ tomando valores no conjunto $\{0, 1\}$. Para o par (y, z) a relação é verdadeira se $\phi(y, z) = 1$ e, por conseguinte, é falsa se $\phi(y, z) = 0$.

As relações podem assumir diversas formas tais como, $\leq, \geq, =, \subseteq, \subset, \in$.

Se considerarmos mais do que uma variável teremos obviamente mais do que uma relação definida. A conjunção das várias relações remete para a definição de “relação produto”.

Definição 3.10 Seja $(R_1, \dots, R_p)$ um conjunto de relações com R_j definida pelo produto cartesiano $\mathcal{Y}_j \times \mathcal{Z}_j$ de dois conjuntos $\mathcal{Y}_j, \mathcal{Z}_j$ onde $[y_j R_j z_j]$ está associada a uma função binária definida para todos os pares $(y_j, z_j) \in \mathcal{Y}_j \times \mathcal{Z}_j$.

Sejam $\mathcal{Y} := \times_{j=1}^p \mathcal{Y}_j$ e $\mathcal{Z} := \times_{j=1}^p \mathcal{Z}_j$ os correspondentes produtos cartesianos compreendendo todos os p -uplos $y = (y_1, \dots, y_p)'$ e $z = (z_1, \dots, z_p)'$ respectivamente.

A *relação produto* $\mathcal{R} := \times_{j=1}^p R_j$ é definida por:

$$[yRz] := \wedge_{j=1}^p [y_j R_j z_j] = [y_1 R_1 z_1] \wedge \dots \wedge [y_p R_p z_p]. \quad (3.2)$$

Se ϕ e ϕ_j , com $j = 1, \dots, p$ são as respectivas funções binárias tem-se: $\phi(y, z) = \prod_{j=1}^p \phi_j(y_j, z_j)$.

Exemplo 3.4 Considere-se $Y_1 = \text{“Número de minutos dormidos por dia”}$, $Y_2 = \text{“Número de minutos a estudar por dia”}$ e $Y_3 = \text{“Número de minutos a ler por dia”}$

Se considerarmos o conjunto das relações $(\in, \in, \in)$, pode ser definida uma condição do tipo

$$[Y_1 \in [200, 400]] \wedge [Y_2 \in [50, 100]] \wedge [Y_3 \in [20, 60]].$$

Formalize-se agora o conceito do conjunto dos indivíduos que verificam condições da forma $[Y_j R_j d_j]$ com d_j uma descrição. Nas definições seguintes considere-se $Y = (Y_1, \dots, Y_p)'$ um vector com p variáveis, clássicas ou simbólicas, com domínios ou espaços de observação $\mathcal{Y}_1, \dots, \mathcal{Y}_p$ respectivamente.

Definição 3.11 Um *evento* é uma condição do tipo $[Y_j R_j d_j]$. É uma função binária $a : E \longrightarrow \{0, 1\}$ que toma o valor 1 ou 0 caso a relação seja verdadeira ou falsa, respectivamente.

$$[Y_j(u) R_j d_j] = \begin{cases} 1 & \text{se } Y_j(u) R_j d_j \text{ é verdadeira} \\ 0 & \text{se } Y_j(u) R_j d_j \text{ é falsa} \end{cases} \quad (3.3)$$

O objecto simbólico mais simples é designado de asserção e pode ser também interpretado como uma *query*.

Definição 3.12 Uma *asserção* é um objecto simbólico definido pela conjunção de eventos:

$$s = \bigwedge_{j=1}^p [Y_j R_j d_j]. \quad (3.4)$$

Definição 3.13 A *aplicação extensão* do objecto asserção s é a função binária $a : E \longrightarrow \{0, 1\}$ definida por:

$$a(u) = \bigwedge_{j=1}^p [Y_j(u) R_j d_j]. \quad (3.5)$$

Definição 3.14 A *extensão* da asserção é constituída por todos os indivíduos para os quais a asserção é verdadeira. O conjunto de objectos $u \in E$ que verificam as condições especificadas por s , é dado por:

$$Ext_E(s) := \{u \in E \mid a(u) = 1\} = \{u \in E \mid [Y_j(u) R_j d_j] = 1, j = 1, \dots, p\} \quad (3.6)$$

Este conjunto é chamado **extensão** da asserção do objecto s . Uma classe poderá ser assim identificada pelo objecto simbólico que a descreve. A extensão do objecto constrói a classe dos indivíduos que têm determinada descrição, ou seja, que verificam o conceito em estudo. De uma forma mais geral, uma asserção pode assumir valores intermédios, como por exemplo, $0 \leq a(u) \leq 1$, que representará o grau de ligação ou de correspondência de um elemento u com a asserção a .

Exemplo 3.5 *Sejam Y_1 ="Número de minutos dormidos por dia", Y_2 =Grupo Etário e Y_3 = Sexo*

$$s = [Y_1 < 500] \wedge [Y_2 \in [15, 24]] \wedge [Y_3 \subseteq \{feminino\}]$$

- s é a asserção;
- $d = d_1 \times d_2 \times d_3 = [0, 500) \times [15, 24] \times \{feminino\}$;
- $R_1 = R_2 = \in$ e $R_3 = \subseteq$;
- $a(u)$ é a extensão da aplicação.

O conjunto de indivíduos obtidos através da extensão são descritos pelo objecto simbólico booleano, "pessoa que dorme menos que 500 minutos por dia, tem entre 15 e 24 anos e é do sexo feminino".

Os objectos simbólicos podem ser do tipo booleanos ou modais e distinguem-se segundo o espaço L .

Definição 3.15 *Um **objecto simbólico booleano** é um triplete $s = (a, R, d)$ com d uma dada descrição do espaço B , R a relação que compara Y e d , e a a função binária definida por $a : E \rightarrow \{0, 1\}$, tal que:*

$$a(u) := \begin{cases} 1 & \text{se } Y(u) \text{ verifica a relação} \\ 0 & \text{se } Y(u) \text{ não verifica a relação} \end{cases} \quad (3.7)$$

Neste caso a extensão é dada por $Ext_E(s) = \{u \in E : a(u) = 1\}$ e devolve todos os elementos que partilham as propriedades indicadas em s .

É de salientar que R e d fornecem uma descrição "intensional" do objecto simbólico, ou seja, especificam as propriedades desejadas, enquanto que a extensão da asserção

$Ext_E(s) \subseteq E$ fornece uma descrição “extensional”, ou seja, indica quais os indivíduos de E que partilham essas propriedades.

Definição 3.16 *Um **objecto simbólico não booleano**¹ é um triplete $s = (a, \Phi, d)$ com d uma dada descrição do espaço B , Φ uma relação entre descrições e a uma aplicação $a : E \rightarrow [0, 1]$, tal que:*

$$a(u) := \Phi(Y(u), d) \quad \text{para } u \in E \quad (3.8)$$

Para cada $u \in E$ o valor $a(u)$ mede o grau de “ligação” do objecto u com a descrição d . Neste caso não há a verificação ou não da propriedade como nos objectos booleanos, mas sim estados intermédios.

Para um dado nível $\alpha \in [0, 1]$ a extensão da aplicação é dada por:

$$Ext_E^\alpha(s) = \{u \in E \mid a(u) \geq \alpha\}. \quad (3.9)$$

Na realidade a definição é análoga à anterior, apenas a função não é binária e passa a tomar valores no intervalo $[0, 1]$.

3.3 Geração de Objectos Simbólicos

Neste trabalho os objectos simbólicos são gerados pela agregação dos indivíduos singulares da matriz de dados clássica em classes, que remete para o “paradigma dois níveis” descrito na Secção 3.1. Por essa razão expõe-se agora como se precede à construção dos dados simbólicos através do *processo de generalização* de uma matriz de dados clássica extraída de uma base de dados relacional.

A base de dados relacional é uma base de dados formada por uma ou várias tabelas, que têm como nome “relações”. Neste tipo de tabelas ou “relações” as linhas são chamadas t-uplos e as colunas atributos. Entre as tabelas ou “relações” podem ser efectuadas operações que dão origem a novas tabelas. As *queries*, operações utilizadas para relacionar tabelas, produzem novas tabelas com a estrutura de uma “relação”. A matriz de dados simbólicos será construída tendo como base uma tabela relacional.

¹Em [Bock e Diday (2000a)] o objecto é designado “modal”, mas decidiu-se utilizar a designação “não booleano” de forma a não causar dúvidas com as variáveis do tipo modal.

3.3.1 Interacção com a Base de Dados

Neste capítulo, será descrito como se geram dados simbólicos a partir de uma tabela relacional.

Defina-se novamente Ω como o conjunto dos indivíduos, com $\Omega = \{\omega_1, \dots, \omega_n\}$ gerando os t-uplos ou linhas da tabela, obtida através de uma *query*. As propriedades de cada indivíduo são dadas pelas p variáveis clássicas $\tilde{Y}_1, \dots, \tilde{Y}_p$, que correspondem às p colunas da tabela. Cada indivíduo $\omega \in \Omega$ corresponde a um vector e a sua descrição é dada por $\tilde{Y}(\omega) = (\tilde{Y}_1(\omega), \dots, \tilde{Y}_p(\omega)) \in \mathcal{Y}_1 \times \dots \times \mathcal{Y}_p$.

A matriz $\tilde{X}$ é o resultado da *query* aplicada e contém informação de como os membros estão reunidos nas classes de acordo com a descrição que queremos obter. Considera-se uma variável nominal G que classifica o indivíduo relativamente à classe a que pertence, ou seja, corresponde à propriedade Y que queremos estudar; os valores observados $G(\omega_1), \dots, G(\omega_n)$ classificam os n indivíduos em uma das k classes. Cada $\omega \in \Omega$ corresponde a um vector $\tilde{Y}_1(\omega), \dots, \tilde{Y}_p(\omega)$ e ao valor observado $G(\omega)$. Para $j = 1, \dots, p$ cada variável clássica $\tilde{Y}_j$ corresponde ao $(j + 2)$ º atributo da *query*, sendo $\mathcal{Y}_j$ o espaço das observações de $\tilde{Y}_j$ deduzido pela respectiva coluna da extensão da *query*.

Cada indivíduo pertence assim a uma certa classe de acordo com o valor observado para a variável G . Suponhamos que os indivíduos singulares se agrupam em k classes $C_1, \dots, C_k$ (às quais serão associados os elementos de $E = \{u_1, \dots, u_k\}$). As classes formam uma partição que cobrem Ω e seguem a seguinte regra:

$$\text{Se } G(\omega) = t \text{ então } \omega \in C_t \text{ para } \forall \omega \in \Omega$$

A matriz de dados $\tilde{X} = (x_{ij})$ resultante da *query* tem então a seguinte estrutura:

Ω	G	$\tilde{Y}_1$	$\dots$	$\tilde{Y}_j$	$\dots$	$\tilde{Y}_p$
ω_1						
$\vdots$						
ω_i	$G(\omega_i)$	$\dots$	$\dots$	$\tilde{Y}_j(\omega_i) = x_{ij}$	$\dots$	
$\vdots$						
ω_n						

As variáveis simbólicas devem ser definidas antes de construirmos os objectos simbóli-

cos. Cada variável simbólica $Y_j : E \rightarrow B_j$ é definida a partir de uma variável clássica $\tilde{Y}_j$ tendo em consideração o seu tipo:

- Se a variável clássica é quantitativa define-se uma variável intervalar e cada classe C_t é caracterizada por um intervalo descrevendo os valores observados $\tilde{Y}_j$ nessa classe.
- Se a variável clássica é qualitativa existem duas opções, ou seja, a correspondente variável simbólica pode ser:
 - categórica multivaluada, em que cada classe é representada pelo conjunto de categorias observadas na classe;
 - modal, em que a variável assume valores na forma de uma distribuição sobre as categorias observadas na classe.

Todos os valores observados pelas p variáveis simbólicas são agrupados numa matriz $k \times p$ representada por $\Xi = (\xi_{ij})$, a matriz de dados simbólicos. O valor observado pela j -ésima variável simbólica para um elemento genérico u_i , é dado por $\xi_{ij} = Y_j(u_i) \in B_j$. A matriz Ξ vem então:

E	Y_1	$\dots$	Y_j	$\dots$	Y_p
u_1					
$\vdots$					
u_i	$\dots$	$\dots$	$Y_j(u_i) = \xi_{ij}$	$\dots$	
$\vdots$					
u_k					

Construção de um Objecto Simbólico

O primeiro passo para a construção de um objecto simbólico é a escolha da variável ou das variáveis clássicas qualitativas que dão origem às classes ou níveis de agregação dos indivíduos. Seja $E = \{u_1, \dots, u_k\}$ o conjunto dos k elementos, correspondendo às k categorias de uma variável clássica qualitativa, ou seja, às classes $C_1, \dots, C_k$. As classes formadas correspondem ao produto cartesiano das categorias da variável (ou variáveis) qualitativas escolhidas. Por exemplo, se escolhermos duas variáveis clássicas *Sexo* e *Região*, obtemos pelo produto cartesiano das categorias as seguintes classes: RN/F, RLVT/M, RA/F,...

Para cada $u_i \in E$ com $i = 1, \dots, k$, definimos um objecto simbólico $S_i = (a_{u_i}, R, d_{u_i})$ que irá corresponder à generalização das características dos indivíduos correspondentes à i -ésima categoria, ou seja, pertencendo à classe C_i .

O vector $d_{u_i} = (\xi_{i1}, \dots, \xi_{ip})'$ é a descrição simbólica de u_i , ou seja, é o vector que corresponde à i -ésima linha da matriz simbólica Ξ e que descreve a classe C_i através do vector de variáveis $(Y_1, \dots, Y_p)$.

- Se a variável clássica $\tilde{Y}_j$ é quantitativa, a correspondente variável simbólica é intervalar, e a descrição simbólica $Y_j(u_i) = \xi_{ij}$ da classe C_i (a que corresponde o elemento u_i) é o intervalo $[\alpha; \beta]$ definido por,

$$\alpha := \text{Min}_{\omega \in C_i} \tilde{Y}_j(\omega) \quad e \quad \beta := \text{Max}_{\omega \in C_i} \tilde{Y}_j(\omega) \quad (3.10)$$

- Se a variável clássica $\tilde{Y}_j$ é qualitativa de domínio $\mathcal{Y}_j = \{m_1, \dots, m_r\}$, a variável simbólica pode ser:

- multivaluada e neste caso a descrição simbólica $Y_j(u_i)$ vai ser um elemento de $B_j = P(\mathcal{Y}_j)$;
- modal e neste caso para cada classe C_i , tem-se um vector $(U(u_i), \pi(u_i)) = (\tau_1, \dots, \tau_r)$ que representa a medida discreta ou a distribuição de probabilidade para a variável Y_j na classe C_i . Cada elemento τ_l é um par (m_l, w_l) onde $m_l \in \mathcal{Y}_j$ é uma categoria da variável $\tilde{Y}_j$ observada na classe C_i e w_l é um valor real não negativo.

Se considerarmos,

$$w_l = \# \left\{ \omega \in C_i \mid \tilde{Y}_j(\omega) = m_l \right\}, \quad (3.11)$$

a medida discreta $Y_j(u_i)$ será uma distribuição de “pesos”. Se assumirmos um modelo probabilístico e estimarmos a probabilidade $P(\tilde{Y}_j = m_l \mid C_i)$ por:

$$w_l = \frac{\# \left\{ \omega \in C_i \mid \tilde{Y}_j(\omega) = m_l \right\}}{\# \left\{ \omega \in C_i \mid \tilde{Y}_j(\omega) \in \mathcal{Y}_j \right\}}, \quad (3.12)$$

temos que $Y_j(u_i)$ é uma distribuição de frequências ou uma função de probabilidade.

3.3.2 Operador Generalizador

O processo de generalização é utilizado quando o objectivo é caracterizar as propriedades das classes através das variáveis conhecidas para os indivíduos. O operador generalizador g fornece a descrição para cada classe $C_1, \dots, C_k$ na forma de um objecto simbólico s_i com $i = 1, \dots, k$.

Seja $S = \{s_1, \dots, s_k\}$ o conjunto dos objectos simbólicos resultante da aplicação da generalização a C_i . O objecto simbólico s_i associado ao elemento $u_i \in E$ é (a_{u_i}, R, d_{u_i}) com a extensão da aplicação a_{u_i} em E e um vector descritor $d_{u_i} = (\xi_{i1}, \dots, \xi_{ip})$.

Definição 3.17 *O operador generalizador g , [Stéphan et al. (2000)], baseado nas variáveis $\tilde{Y}_1, \dots, \tilde{Y}_p$ em Ω é definido por $g = (g_1, \dots, g_p)$ com operadores generalizadores g_j para cada variável que expressam as propriedades comuns da variável na classe C_i :*

$$d_{u_i} = g(C_i) = Y(u_i) \text{ onde } d_{u_i} = (\xi_{i1}, \dots, \xi_{ip})' \text{ com } \xi_{ij} = g_j(C_i) = Y_j(u_i) \quad (3.13)$$

A partir da matriz original $X = (x_{ij})_{n \times p}$ definimos g_j para a classe C_i com $i = 1, \dots, k$ e $j = 1, \dots, p$, como um operador “união”:

$$\xi_{ij} := \begin{cases} [\text{Min}_{\omega_\ell \in C_i} \{x_{\ell j}\}, \text{Max}_{\omega_\ell \in C_i} \{x_{\ell j}\}] & \text{se } \tilde{Y}_j \text{ é quantitativa} \\ \{v \in \mathcal{Y}_j | \exists \omega_\ell, \omega_{\ell'} \in C_i, x_{\ell j} \leq v \leq x_{\ell' j}\} & \text{se } \tilde{Y}_j \text{ é ordinal} \\ \{v \in \mathcal{Y}_j | \exists \omega_\ell \in C_i, x_{\ell j} = v\} & \text{se } \tilde{Y}_j \text{ é qualitativa} \\ \text{ou} & \\ ((m_l, w_l) : \exists \omega_\ell \in C_i, x_{\ell j} = m_l); w_l \text{ dado por (3.11) ou (3.12)} & \text{se } \tilde{Y}_j \text{ é qualitativa} \end{cases}$$

Ao aplicar-se o operador generalizador pode ser originado um problema que passa pela inclusão de indivíduos atípicos a uma classe, ou seja, inclusão de valores que raramente são observados e que podem influenciar os resultados criando descrições de classes que não reflectem as propriedades mais características. Se esses indivíduos não forem considerados obtém-se uma descrição mais homogênea do grupo. A este processo é chamado o **passo de especialização**. Através do algoritmo de especialização a generalização inicial é reduzida ao retirarem-se os indivíduos atípicos.

Crítério de qualidade para avaliar a descrição generalizada

A qualidade de um conjunto descritor d_{u_i} de uma generalização g_i dentro da classe C_i é medida por uma boa “negociação” entre a homegeneidade da distribuição dos indivíduos e a cobertura de d_{u_i} em C_i . O **conjunto cobertura** de d_{u_i} em C_i corresponde à extensão do objecto simbólico. Para variáveis não modais pode ser definido por:

$$Ext(s_i|C_i) := \{\omega \in C_i | \tilde{Y}_j(\omega) \in \xi_{ij}, j = 1, \dots, p\} \quad (3.14)$$

Neste caso para medir a “parsimonia” de uma generalização pode-se utilizar o **critério volume** (ver [Brito (1994)]), que produz o índice de generalização de d_{u_i} :

$$vol(d_{u_i}) = \prod_{j=1}^p \mu(\xi_{ij}) \quad (3.15)$$

onde,

$$\mu(\xi_{ij}) := \begin{cases} \#\xi_{ij} & \text{se } \tilde{Y}_j \text{ é nominal} \\ Max(\xi_{ij}) - Min(\xi_{ij}) & \text{se } \tilde{Y}_j \text{ é quantitativa ou ordinal} \end{cases}$$

Define-se a densidade de um objecto simbólico s_i como:

$$den(s_i) = \frac{\#Ext(s_i|C_i)}{vol(d_{u_i})} \quad (3.16)$$

A base do passo de especialização é reduzir o conjunto cobertura de uma generalização e testar quando essa redução é suficiente em comparação com o resultado da redução do volume. É baseado na noção de uma α -generalização para um patamar $0 < \alpha < 1$.

Definição 3.18 *Uma α -generalização é um objecto simbólico $(a_i^\alpha, \in, d_{u_i}^\alpha)$ chamado s_i^α , tal que:*

- $\#Ext(s_i^\alpha|C_i) \geq \alpha\#C_i$
- $d_{u_i}^\alpha$ tal que:

$$vol(d_{u_i}^\alpha) = Min_{d \in B_1 \times \dots \times B_p} \{vol(d) | \#Ext(s|C_i) > \alpha\#C_i\} \quad (3.17)$$

O objectivo do processo de especialização é encontrar um nível α que corresponda à melhor “negociação” entre a redução do volume e a diminuição da cobertura.

Para uma dada sequência $\alpha_1 > \alpha_2 > \dots > \alpha_t > \dots > \alpha_l > 0$ pode ser definida a α_t -generalização $s_i^{\alpha_t}$. Cada $s_i^{\alpha_t}$ é avaliado de acordo com a sua densidade relativa $\frac{dens(s_i^{\alpha_t})}{dens(s_i)}$, onde s_i é a primeira asserção obtida pela generalização. A melhor α -generalização corresponde ao melhor rácio de crescimento da curva de densidade relativa, ou seja, o primeiro salto de crescimento corresponde à melhor generalização.

O critério do volume não pode ser aplicado com variáveis qualitativas e quantitativas em simultâneo devido a problemas de escala. Para contornar esta situação as variáveis intervalares devem ser recodificadas em ordinais. Esta codificação é feita através de uma distribuição uniforme pelos intervalos, dividindo o intervalo inicial em sub-intervalos tais que para cada um deles a distribuição subjacente seja o mais uniforme possível. Este processo é feito através do particionamento usando um algoritmo recursivo que segue uma regra de paragem baseada no teste da uniformidade, partindo da hipótese nula: “os valores observados seguem uma distribuição uniforme no intervalo”. Aplicando a discretização, a descrição $Y_j(u_i) = \xi_{ij}$ para uma variável intervalar Y_j , torna-se num conjunto de valores ordinais. Para um aprofundamento deste ponto consulte-se [Stéphan *et al.* (2000)].

Capítulo 4

Métodos Estatísticos de Análise de Dados para Dados Simbólicos

4.1 Estatísticas Descritivas

Considere-se que a tabela de dados simbólicos observados $\Xi = (\xi_{ij})$ é resultante da observação de p variáveis simbólicas $Y_1, \dots, Y_p$ para cada elemento u_i do conjunto E . A linha i da matriz Ξ fornece a descrição simbólica (ver Secção 3.2 Definição 3.7) do elemento $u_i \in E$, denominada por vector descrição d_{u_i} :

$$d_{u_i} = (\xi_{i1}, \dots, \xi_{ip})', \text{ com } u_i \in E$$

Seja $d = (d_1, \dots, d_p)$ um vector descrição pertencente ao conjunto $B := B_1 \times \dots \times B_p$.

Definição 4.1 *Um vector descrição $x \equiv (d_1, \dots, d_p)'$ é um vector **descrição individual** (ver [Bertrand e Goupil (2000)]), se cada conjunto $d_j \subseteq \mathcal{Y}_j$ é singular: $d_j = \{y_l\}$ com $y_l \in \mathcal{Y}_j$ para $j=1, \dots, p$.*

Para além do vector descritor individual x interessa definir também a dependência lógica que pode existir no espaço do domínio $\mathcal{X} = \times_{j=1}^p P(\mathcal{Y}_j)$.

Se considerarmos A_1 e A_2 dois conjuntos de $\mathcal{X}$ e x um vector descritor individual qualquer, pode-se formular a regra:

$$v : [x \in A_1] \Rightarrow [x \in A_2] \tag{4.1}$$

Note-se que o vector x também satisfaz a regra v se $x \notin A_1$.

Definição 4.2 Um vector descriptor individual $x \in B$ **satisfaz a regra** $v: [x \in A_1] \Rightarrow [x \in A_2]$ se e só se $x \in A_1 \cap A_2$ ou $x \notin A_1$. Se x satisfaz v então $v(x) = 1$, caso contrário, $v(x) = 0$ e x não satisfaz a regra.

Representamos por $V_{\mathcal{X}}$ o conjunto de todas as regras que descrevem todas as dependências lógicas definidas no espaço $\mathcal{X}$.

Definição 4.3 A **extensão virtual** $vir(d)$ de um vector descriptor d , é constituída por todos os vectores descriptors individuais de B que satisfazem todas as regras em $V_{\mathcal{X}}$,

$$vir(d) := \{x \in B | v(x) = 1 \ \forall v \in V_{\mathcal{X}}\} \quad (4.2)$$

4.1.1 Variáveis Intervalares

Neste caso serão consideradas $Y_1, \dots, Y_p$ variáveis intervalares e Y será uma das p variáveis. Para $u \in E = \{u_1, \dots, u_k\}$ tem-se o intervalo $Y(u)$ dado por $[\underline{y}_u, \bar{y}_u] \subseteq \mathbb{R}$. Assuma-se que a variável Y não depende logicamente de nenhuma outra variável e que são garantidas as seguintes hipóteses:

- Cada observação $u \in E$ é seleccionada com a mesma probabilidade $\frac{1}{k}$;
- todos os vectores descriptors individuais $x \in vir(d_u)$ são distribuídos uniformemente no intervalo $Y(u) = [\underline{y}_u, \bar{y}_u]$.

Definição 4.4 Seja Y uma variável intervalar definida em E . A **função distribuição empírica** de Y ([Bertrand e Goupil (2000)]) denotada por F_y , é a função distribuição conjunta de n distribuições uniformes definidas nos intervalos $Y(u)$ para $u \in E$.

Para um dado número real $t \in \mathbb{R}$ deduz-se que:

$$F_Y(t) = \frac{1}{k} \sum_{u \in E} Pr(x_Y \leq t \mid x \in vir(d_u)) \quad (4.3)$$

com,

$$Pr(X_Y \leq t \mid x \in vir(d_u)) = \begin{cases} 0 & \text{se } t < \underline{y}_u \\ \frac{t - \underline{y}_u}{\bar{y}_u - \underline{y}_u} & \text{se } \underline{y}_u \leq t \leq \bar{y}_u \\ 1 & \text{se } \bar{y}_u \leq t \end{cases} \quad (4.4)$$

As definições seguintes têm suporte em [Billard e Diday (2003)].

Definição 4.5 Considerando que $I_u(\cdot)$ é a função indicatriz do intervalo $Y(u)$ e $\|Y(u)\|$ o comprimento do intervalo, a **função densidade empírica** é dada por:

$$f(t) = \frac{1}{k} \sum_{u \in E} \frac{I_u(t)}{\|Y(u)\|} = \frac{1}{k} \sum_{u \in E} \frac{I_u(t)}{\bar{y}_u - \underline{y}_u}, \quad t \in \mathbb{R}. \quad (4.5)$$

Para variáveis intervalares pode definir-se o histograma, que permite uma visualização da distribuição das frequências dos valores. Consideremos I o intervalo contendo todos os valores observados para a variável simbólica intervalar Y :

$$I = [\text{Min} \{ \underline{y}_u | u \in E \}, \text{Max} \{ \bar{y}_u | u \in E \}].$$

Consideremos também uma partição de I em m intervalos $I_j = [h_{j-1}, h_j]$ para $j = 1, \dots, m$. O histograma associado à variável Y com a partição do intervalo $\{I_1, \dots, I_m\}$ é a representação gráfica da distribuição de frequências $\{(I_j, p_j), j = 1, \dots, m\}$ onde p_j é a probabilidade do valor x_Y pertencer ao intervalo I_j para uma dado vector descritor individual.

$$p_j = \frac{1}{k} \sum_{u \in E} \frac{\|Y(u) \cap I_j\|}{\|Y(u)\|} \quad (4.6)$$

Definição 4.6 A **média amostral** é dada por:

$$\bar{Y} = \frac{1}{k} \sum_{u \in E} \int_{-\infty}^{+\infty} \frac{I_u(t)}{\|Y(u)\|} t dt = \frac{1}{k} \sum_{u \in E} \int_{\underline{y}_u}^{\bar{y}_u} \frac{t}{\|Y(u)\|} dt = \frac{1}{k} \sum_{u \in E} \frac{\underline{y}_u + \bar{y}_u}{2} \quad (4.7)$$

isto é, a média amostral de Y é média dos pontos médios dos intervalos.

Definição 4.7 A **variância amostral** é dada por:

$$s_Y^2 = \frac{1}{3k} \sum_{u \in E} (\bar{y}_u^2 + \underline{y}_u \bar{y}_u + \underline{y}_u^2) - \frac{1}{4k^2} [\sum_{u \in E} (\underline{y}_u + \bar{y}_u)]^2, \quad (4.8)$$

uma vez que

$$s_Y = \sqrt{\int_{-\infty}^{+\infty} t^2 f_Y(t) dt - \bar{Y}^2},$$

e

$$\int_{-\infty}^{+\infty} t^2 f_Y(t) dt = \frac{1}{k} \sum_{u \in E} \int_{\underline{y}_u}^{\bar{y}_u} \frac{t^2}{\|Y(u)\|} dt = \frac{1}{k} \sum_{u \in E} \frac{\bar{y}_u^3 - \underline{y}_u^3}{3 \cdot \|Y(u)\|} = \frac{1}{3k} \sum_{u \in E} (\bar{y}_u^2 + \bar{y}_u \underline{y}_u + \underline{y}_u^2).$$

Considerem-se agora Y_1 e Y_2 duas variáveis simbólicas intervalares com observações no rectângulo $D(u) = Y_1(u) \times Y_2(u) = ([\underline{y}_{1u}, \bar{y}_{1u}], [\underline{y}_{2u}, \bar{y}_{2u}])$ para cada $u \in E$.

Definição 4.8 A *função densidade conjunta* para (Y_1, Y_2) é dada por:

$$f(y_1, y_2) = \frac{1}{k} \sum_{u \in E} \frac{I_u(y_1, y_2)}{\|D(u)\|} \quad (4.9)$$

com $I_u(y_1, y_2)$ a função indicatriz do rectângulo $D(u)$ de área $\|D(u)\|$.

Definição 4.9 A *função covariância amostral simbólica* é dada por:

$$\begin{aligned} s_{Y_1, Y_2} &= \int_{-\infty}^{+\infty} (y_1 - \bar{Y}_1)(y_2 - \bar{Y}_2) f(y_1, y_2) dy_1 dy_2 = \\ &= \frac{1}{k} \sum_{u \in E} \frac{1}{(\bar{y}_{1u} - \underline{y}_{1u})(\bar{y}_{2u} - \underline{y}_{2u})} \times \int \int_{(y_1, y_2) \in Y(u)} y_1 y_2 dy_1 dy_2 - \bar{Y}_1 \bar{Y}_2 = \\ &= \frac{1}{4k} \sum_{u \in E} (\bar{y}_{1u} + \underline{y}_{1u})(\bar{y}_{2u} + \underline{y}_{2u}) - \frac{1}{4k^2} \left[\sum_{u \in E} (\bar{y}_{1u} + \underline{y}_{1u}) \right] \left[\sum_{u \in E} (\bar{y}_{2u} + \underline{y}_{2u}) \right] \end{aligned} \quad (4.10)$$

Definição 4.10 A *função correlação amostral simbólica* entre duas variáveis intervalares Y_1 e Y_2 é dada por:

$$r(Y_1, Y_2) = \frac{s_{Y_1 Y_2}}{\sqrt{(s_{Y_1^2})(s_{Y_2^2})}} \quad (4.11)$$

onde $s_{Y_1 Y_2}$ é dada por (4.10) e $s_{Y_1^2}$ e $s_{Y_2^2}$ são dados por (4.8).

Note-se que com as definições supra, a variância amostral não é um caso particular da covariância, isto é, S_{YY} é diferente de S_Y^2 . Isto resulta de a definição de covariância envolver a integração num rectângulo, enquanto que no cálculo da variância a integração é efectuada num segmento de recta.

Para ultrapassar este problema, Billard ([Billard (2007)]) propôs recentemente uma definição alternativa para a covariância, dada por

$$s_{Y_1 Y_2} = \frac{1}{3m} \sum_{u \in E} G_1 G_2 [Q_1 Q_2]^{1/2} \quad (4.12)$$

com

$$\begin{aligned} Q_j &= (\underline{y}_{ju} - \bar{Y}_j)^2 + (\underline{y}_{ju} - \bar{Y}_j)(\bar{y}_{ju} - \bar{Y}_j) + (\bar{y}_{ju} - \bar{Y}_j)^2 \\ G_j &= \begin{cases} -1, & \text{if } \bar{Y}_{ju} \leq \bar{Y}_j, \\ 1, & \text{if } \bar{Y}_{ju} > \bar{Y}_j, \end{cases} \end{aligned}$$

para $j = 1, 2$ onde a média amostral $\bar{Y}_j$ é dada por (4.7) e

$$\bar{Y}_{ju} = (\underline{y}_{ju} + \overline{y}_{ju})/2. \quad (4.13)$$

4.1.2 Variáveis Modais

Nesta secção serão referidas as medidas de sumarização para variáveis modais, como as capacidades, mínimos, máximos e médias ([Goupil (2000)]).

Para variáveis modais Y considera-se uma nova medida de sumarização chamada **capacidade**. A capacidade de uma categoria de uma variável modal Y é a estimativa da probabilidade de observar a categoria em pelo menos um dos objectos $\{u_1, \dots, u_k\}$. Seja $Y(u) = (m_1(w_1), \dots, m_r(w_r))$ o valor de uma variável modal com categorias $m_1, \dots, m_r$ e $w_1, \dots, w_r$ probabilidades ou frequências associadas a cada categoria, admitindo a independência.

Definição 4.11 *Sejam u_1 e $u_2 \in E$ com $Y(u_1) = (m_1(w_{11}), \dots, m_r(w_{r1}))$ e $Y(u_2) = (m_1(w_{12}), \dots, m_r(w_{r2}))$. A **capacidade** da categoria m_l é dada por:*

$$w_{l1} + w_{l2} - w_{l1} \times w_{l2}, \quad l = 1, \dots, r \quad (4.14)$$

O cálculo da capacidade de uma categoria para $E = \{u_1, \dots, u_k\}$ é efectuada pela fórmula de Poincaré.

Definição 4.12 *mínimo, máximo e média associados à categoria m_l são dados por:*

- o mínimo é o $\text{Min}\{w_{l1}, w_{l2}, \dots, w_{lk}\}$;
- a média é a média aritmética de $\{w_{l1}, w_{l2}, \dots, w_{lk}\}$;
- o máximo é o $\text{Max}\{w_{l1}, w_{l2}, \dots, w_{lk}\}$.

4.2 Análise Classificatória

Na análise de dados simbólicos as técnicas de classificação têm o mesmo objectivo que na análise de dados clássica, agrupar indivíduos em classes homogêneas com base nos

valores observados para as variáveis. A partir de um conjunto $E = \{u_1, \dots, u_k\}$ faz-se um agrupamento de objectos homogêneos em classes $C_1, \dots, C_K$ ¹, de forma a que os elementos de uma mesma classe apresentem grande semelhança. Existem vários métodos de classificação que variam consoante o tipo dos dados, o tipo de critério de aglomeração, o tipo de estrutura, etc. Distinguem-se, genericamente, os métodos de classificação não-hierárquicos, que determinam directamente uma partição do conjunto a classificar, optimizando (em geral localmente) uma medida de homogeneidade das classes, e os modelos de classificação do tipo hierárquico, que constroem estruturas de classes encaixadas. De uma forma geral, as técnicas de classificação hierárquica dividem-se em dois grandes tipos: divisiva e aglomerativa. Nos métodos de tipo divisivo, as classes são construídas sucessivamente de forma descendente, ou seja, “de cima para baixo”. Nos métodos do tipo aglomerativo as classes vão sendo sucessivamente agrupadas de forma crescente, ou seja, “de baixo para cima”. Consoante o tipo de estrutura considerada, obtém-se representações em **Hierarquia** ou **Pirâmide**.

4.2.1 Classificação Hierárquica pelo método Divisivo

A classificação divisiva é um método que produz uma classificação hierárquica descendente partindo de uma única classe. São efectuadas divisões ou cortes sucessivos dessa classe, cada um baseado numa variável, utilizando um critério de variância intra-classe. O resultado final é uma árvore. O método de classificação divisiva apresentado neste estudo é o método desenvolvido por Marie Chavent [Chavent (2000)], mas pode ser consultado também para dados clássicos e simbólicos em [Chavent (1997)] e [Chavent (1998)]. O algoritmo segue a filosofia do método CART [Breiman *et al.* (1984)], mas agora num contexto de análise classificatória e não de análise discriminante, isto é, pretende encontrar-se partições em classes homogêneas, mas não há classes à priori a explicar.

A exposição do método será restringida ao caso de variáveis simbólicas intervalares ou modais.

Seja $P_K = \{C_1, \dots, C_K\}$ uma K-partição de E . A cada nível é construída uma nova partição K+1 dividindo uma das classes $C_i \in P_K$ em duas novas classes C_i^1 e C_i^2 . A escolha de C_i deve ser feita tal que a nova partição, $P_{K+1} = P_K \cup \{C_i^1, C_i^2\} - \{C_i\}$, tenha uma dispersão interna mínima.

A divisão das classes é feita utilizando o **critério da variância intra-classe** gener-

¹Note-se que C_i representa agora um subconjunto de $E = \{u_1, \dots, u_k\}$, conjunto a analisar.

alizada a dados simbólicos, que avalia a qualidade da partição e é uma extensão do critério da soma dos quadrados intra-classe clássico à matriz de distâncias. A qualidade de uma classe é dada pela sua inércia, calculada a partir da matriz de distâncias D .

Definição 4.13 A *inércia* de uma classe C_i define-se por:

$$I(C_i) := \frac{1}{2\mu_i} \sum_{u_h \in C_i} \sum_{u_l \in C_i} \lambda_u \lambda_l d(u_h, u_l)^2 \quad (4.15)$$

com $\mu_i := \sum_{u \in C_i} \lambda_u$ e $\lambda_u \geq 0$ o peso associado a cada objecto $u \in E$.

No caso particular de todos os indivíduos terem o mesmo peso, $\lambda_u = \frac{1}{k}$, tem-se:

$$I(C_i) := \frac{1}{2 \times k \times n_i} \sum_{u_h \in C_i} \sum_{u_l \in C_i} d(u_h, u_l)^2 \quad (4.16)$$

com $n_i := \#C_i$ o cardinal da classe C_i .

Usando um critério aditivo sobre as classes obtém-se a generalização do critério da variância:

Definição 4.14 A *generalização do critério da variância*, de uma K -partição $P_K = \{C_1, \dots, C_K\}$, para o caso de uma matriz de distâncias é dado por:

$$W(P_K) := \sum_{i=1}^K I(C_i) \quad (4.17)$$

com a medida de heterogeneidade $I(C_i)$ dada por (4.15) ou (4.16).

Se $W(P_{K+1}) = W(P_K) - I(C_i) + I(C_i^1) + I(C_i^2)$, então minimizar $W(P_{K+1})$ é equivalente a escolher a classe $C_i \in P_K$, tal que a diferença entre a inércia de C_i e a inércia intra classes da sua bipartição (C_i^1, C_i^2) seja maximizada.

O critério para seleccionar a classe que será dividida é então dado por maximização de:

$$\Delta(C_i) := I(C_i) - I(C_i^1) - I(C_i^2) \quad (4.18)$$

onde $I(C)$ é dado por (4.15) ou (4.16).

Considere-se o conjunto $E = \{u_1, \dots, u_k\}$ e as p variáveis $Y_1, \dots, Y_p$. Seja d uma medida de distância no conjunto E que permite calcular uma matriz $k \times k$ de distâncias $D = (d(u_h, u_l))$ a partir da matriz de dados simbólicos Ξ . Neste método são propostas duas medidas, uma para dados quantitativos e outra para dados categóricos.

Medida de distância para uma matriz de dados simbólicos intervalares

Sejam $([\alpha_{h1}, \beta_{h1}], [\alpha_{h2}, \beta_{h2}], \dots, [\alpha_{hp}, \beta_{hp}])$ e $([\alpha_{l1}, \beta_{l1}], [\alpha_{l2}, \beta_{l2}], \dots, [\alpha_{lp}, \beta_{lp}])$ as descrições associadas aos objectos simbólicos, u_h e u_l .

A distância utilizada para este tipo de dados é a distância de **Hausdorff** e é definida por:

$$d(u_h, u_l) = \left(\sum_{j=1}^p \text{Max} \{ |\alpha_{hj} - \alpha_{lj}|, |\beta_{hj} - \beta_{lj}| \}^2 \right)^{\frac{1}{2}} \quad (4.19)$$

Se em vez de intervalos estiverem a ser considerados pontos individuais em $\mathbb{R}$ obtém-se a distância **Euclideana**.

Medida de distância para uma matriz de dados simbólicos categóricos

Neste caso consideramos as variáveis $Y_1, \dots, Y_p$ do tipo modais em que o valor que cada variável toma é uma distribuição de probabilidades discreta ou uma distribuição de frequências. A matriz de dados simbólicos Ξ pode ser transformada numa matriz de frequências $F = (f_{ij})_{k \times r}$ onde $r = \#Y_1 + \dots + \#Y_p$ é o número total de categorias e $j = 1, \dots, r$ são os índices de todas as categorias na matriz de frequências (ver Tabela 4.1).

	1	...	j	...	r	Σ
u_1	f_{11}		f_{1j}		f_{1r}	p
$\vdots$						
u_i	f_{i1}		f_{ij}		f_{ir}	p
$\vdots$						
u_k	f_{k1}		f_{kj}		f_{kr}	p
Σ	$f_{.1}$		$f_{.j}$		$f_{.r}$	kp

Tabela 4.1: Matriz de frequências associadas às categorias

Neste caso, para comparar dois objectos deve ser utilizada a **distância do χ^2** :

$$d^2(u_h, u_l) = \sum_{j=1}^r \frac{p_{..}}{p_{.j}} \left(\frac{p_{hj}}{p_{h.}} - \frac{p_{lj}}{p_{l.}} \right)^2 \quad (4.20)$$

onde $h, l = 1, \dots, k$ e $j = 1, \dots, r$, com

$$p_{ij} := \frac{f_{ij}}{kp}; \quad p_{i.} := \sum_{j=1}^r p_{ij} = \frac{p_i}{kp} = \frac{1}{k}; \quad p_{.j} := \sum_{i=1}^k p_{ij}$$

sendo,

$$p_{..} := \sum_{i=1}^k \sum_{j=1}^r p_{ij} = 1$$

Em cada etapa a escolha da bipartição $C_i = (C_i^1, C_i^2)$ é feita através da determinação do corte que minimiza $W(P_{K+1})$, ou seja, que tenha o menor valor para a inércia intra-classe. São efectuados sucessivos cortes para cada variável, em que cada corte corresponde a uma questão binária colocada a cada variável. Este conceito foi desenvolvido por Breiman, Friedman, Olshen & Stone ([Breiman *et al.* (1984)]). Em [Chavent (2000)] é feita a extensão deste conceito para o caso das variáveis simbólicas intervalares, ordinais multivaluadas e modais, a qual será aqui apresentada.

Cada classe C é dividida conforme a resposta à questão binária da forma “É $Y_j \leq c$?”, com $c \in \mathcal{Y}_j$, o **valor de corte**. Um objecto $u \in C$ responde “sim” ou “não” à questão segundo a função binária $q_c : E \rightarrow \{\text{verdadeiro}, \text{falso}\}$ e a bipartição (C_1, C_2) de C vem:

$$C_1 = \{u \in C | q_c(u) = \text{verdadeiro}\}$$

$$C_2 = \{u \in C | q_c(u) = \text{falso}\}$$

Variáveis Intervalares Para o caso de variáveis intervalares com $Y_j(u) = [\alpha, \beta]$ a função q_c é definida da seguinte forma:

$$q_c(u) = \text{verdadeiro se } \nu_u \leq c$$

$$q_c(u) = \text{falso se } \nu_u > c$$

com $\nu_u = \frac{\alpha + \beta}{2}$ o ponto médio do intervalo. Os valores de corte c são definidos pelos pontos centrais entre dois valores de ν_u consecutivos.

Variáveis Modais Para o caso de variáveis modais, $Y_j(u) = \pi_u$ é uma distribuição de probabilidades discreta ou uma distribuição de frequências e consideram-se os cortes definidos por todas as bi-partições (M_1, M_2) do conjunto das modalidades. A função

q_c é definida da seguinte forma:

$$\begin{aligned} q_c(u) &= \text{verdadeiro se } \sum_{m \in M_1} \pi_u(m) \geq \frac{1}{2} \\ q_c(u) &= \text{falso se } \sum_{m \in M_1} \pi_u(m) < \frac{1}{2} \end{aligned}$$

Finalmente, de todas as bipartições (C_i^1, C_i^2) obtidas, escolhe-se a que tenha menor valor da inércia intra classe.

A paragem da divisão é feita após L iterações onde L é dado como input pelo utilizador. O problema na escolha de L está em garantir que todas as partições interessantes com as menores variâncias dentro das classes estejam a ser consideradas para $L < k$.

O output deste método é uma Hierarquia H cujas “folhas” são as $L+1$ classes da partição obtida na última iteração.

4.2.2 Hierarquias e Pirâmides

Os métodos de classificação em Hierarquia e Pirâmide são métodos aglomerativos, em que as classes formadas ficam organizadas segundo uma estrutura encaixada. Estas estruturas podem ser criadas a partir da matriz de dados ou a partir de uma matriz de dissimilaridades.

O principal objectivo deste método de classificação é agrupar os objectos do conjunto E em classes homogêneas e posteriormente agrupar as classes formadas pelos objectos, até que se forme uma única classe contendo todos os elementos de E .

A classificação hierárquica, segundo uma estrutura em Pirâmide, é uma extensão do modelo hierárquico [Diday (1984), Diday (1986), Bertrand e Diday (1985)]. No caso de uma Hierarquia, cada nível de agregação corresponde a uma partição, e no caso de uma Pirâmide cada nível corresponde a uma família de classes sobrepostas que são intervalos de uma ordem linear em E . A estrutura em Pirâmide é mais rica do que a Hierarquia, uma vez que, permite obter simultaneamente uma classificação e uma seriação do conjunto E a classificar.

O método de classificação simbólica apresentado foi proposto por [Brito e Diday (1989), Brito (1991), Brito (1994)], usando as representações em Hierarquia e Pirâmide, e permitindo a classificação com variáveis multivaluadas. Posteriormente o método foi desenvolvido para variáveis modais [Brito (1998)], definindo operadores adequados à classificação deste tipo de dados. Mais tarde, Brito e De Carvalho, generalizam o método ao caso das variáveis categóricas multivaluadas [Brito e De Carvalho (1999)] e modais [Brito e De Carvalho (2002)].

Definição 4.15 Uma **Hierarquia** é uma família finita $H = \{A, B, \dots\}$ de subconjuntos não-vazios $A, B, \dots \subseteq E = \{u_1, \dots, u_k\}$, chamados classes, tais que:

- $E \in H$
- Todos os singletões $\{u_1\}, \{u_2\}, \dots, \{u_k\}$ pertencem a H
- $\forall A, B \in H : A \cap B = \phi$ ou $A \subseteq B$ ou $B \subseteq A$

Definição 4.16 Uma **Hierarquia indexada** ou um dendrograma é um par (H, f) , onde H é uma Hierarquia e f é uma aplicação $f : H \rightarrow \mathbb{R}^+$ tal que:

- $f(A) = 0 \Leftrightarrow \#(A) = 1$
- $A \subseteq B \Rightarrow f(A) \leq f(B)$

Uma classe $A \in H$ diz-se um sucessor de uma classe $B \in H$ se $A \subseteq B$ e não existe nenhuma classe $C \in H$, $C \neq A$, $C \neq B$ tal que $A \subset C \subset B$. A classe B diz-se predecessor de A . Numa Hierarquia cada classe tem no máximo um predecessor.

Definição 4.17 Uma **Pirâmide** é uma família finita $P = \{A, B, \dots\}$ de subconjuntos não-vazios $A, B, \dots \subseteq E = \{u_1, \dots, u_k\}$, chamados classes, tais que:

- $E \in P$
- Todos os singletões $\{u_1\}, \{u_2\}, \dots, \{u_k\}$ pertencem a P
- A intersecção de quaisquer duas classes $A, B \in P$ ou é vazia ou pertence a P :

$$A \cap B = \phi \text{ ou } A \cap B \in P$$

- Existe uma ordem linear $\leq$ em E tal que cada classe A de P é um intervalo de $(E, \leq)$:

$$A = [\alpha, \beta] := \{u | u \in E, \alpha \leq u \leq \beta\} \text{ para algum } \alpha, \beta \in E$$

Definição 4.18 Uma **Pirâmide indexada** é um par (P, f) , onde P é uma Pirâmide e f é uma aplicação $f : P \rightarrow \mathbb{R}^+$ tal que:

- $f(A) = 0 \Leftrightarrow \#(A) = 1$

- $A \subseteq B \Rightarrow f(A) \leq f(B)$
- $f(A) = f(B)$ com $A \subset B$ e $A \neq B \Rightarrow \exists C \neq A, D \neq B$ tais que $A = C \cap D$

As noções de sucessor e predecessor também se aplicam no caso da Pirâmide, mas neste caso cada classe pode ter dois predecessores.

Neste trabalho interessar-nos-emos apenas pelo método de classificação hierárquico dito “simbólico”, em que as classes são formadas a partir das descrições dos elementos do conjunto a classificar E de objectos, caracterizados por p variáveis simbólicas $Y_1, \dots, Y_p$. A principal característica deste método assenta na construção de classes que têm que ser associadas a objectos simbólicos completos [Brito (1991)].

Neste caso, dados dois objectos simbólicos s e s' , diz-se que s é *mais geral* que s' se $Ext(s') \subseteq Ext(s)$, s' diz-se *mais específico* que s . A intenção de um conjunto $C \subseteq E$, $int(C)$, é o objecto simbólico s mais específico tal que $\forall u \in C, u \in Ext_E(s)$.

Um objecto simbólico s é *completo* se a intenção da sua extensão é o próprio objecto simbólico, $int(Ext_E(s)) = s$, ou seja, é o objecto simbólico mais específico que descreve exhaustivamente a sua extensão.

Um conceito é um par (C, s) onde $C \subseteq E$ e s é um objecto simbólico completo, em que $C = Ext_E(s)$ com $s = int(C)$.

Neste método, o critério utilizado na formação das classes é a dualidade intensão-extensão: cada classe da Hierarquia ou Pirâmide deve corresponder a um conceito, ou seja, é representada por um objecto simbólico completo cuja extensão é a própria classe. A classe é descrita extensionalmente pelo conjunto dos seus membros e é descrito intensionalmente por um objecto simbólico que expressa as suas propriedades. Uma classe é associada a um objecto simbólico que generaliza os seus membros e nenhum elemento fora da classe deverá verificar a descrição. A agregação das classes é feita tendo em conta também um critério numérico que permite escolher qual a melhor agregação em cada etapa.

Processo de generalização

Quando dois conceitos são agregados é necessário determinar a sua intenção na forma de objecto simbólico. A generalização de dois objectos simbólicos s e s' , corresponde a determinar um objecto simbólico s'' que seja mais geral que s e s' . Este procedimento difere consoante o tipo das variáveis.

Variáveis Intervalares Se Y é uma variável intervalar, os objectos simbólicos $s_h = (a_h, R_h, d_h)$ com $h = 1, \dots, q$, têm eventos da forma $e_h = [Y \subseteq I_h]$, com $I_h = [\alpha_h, \beta_h]$. No objecto generalizado o evento vem:

$$e = [Y \subseteq [\text{Min} \{\alpha_h\}, \text{Max} \{\beta_h\}]],$$

um vez que $[\text{Min} \{\alpha_h\}, \text{Max} \{\beta_h\}]$ é o menor intervalo que contém todos os intervalos $I_h = [\alpha_h, \beta_h]$, $h = 1, \dots, q$.

Exemplo 4.1 Considerem-se os objectos simbólicos RA/F e RA/M descritos pela variável simbólica intervalar $Y = \text{Dormir}$, associados às seguintes asserções:

$$RA/F = [Y \subseteq [230; 1330]] \text{ e } RA/M = [Y \subseteq [170; 1010]]$$

O objecto generalizado é associado à asserção $e = [Y \subseteq [170; 1330]]$.

Variáveis Modais Neste tipo de descrições os eventos tomam a seguinte forma:

$$e_j = [Y_j = (m_1(w_1), \dots, m_r(w_r))]$$

onde $\mathcal{Y}_j = \{m_1, \dots, m_r\}$ é o conjunto das categorias da variável Y_j e $w_1, \dots, w_r$ são as frequências, probabilidades ou pesos associados às categorias. A condição $w_1 + \dots + w_r = 1$ não é imposta.

Existem duas formas alternativas de determinar uma descrição generalizada de dois eventos, usando a generalização pelo máximo ou a generalização pelo mínimo.

Generalização pelo Máximo

Neste caso, a generalização de dois eventos e_1, e_2 é feita tomando o supremo das suas frequências para cada categoria w_v . Formalmente tem-se,

$$e_1 = [Y_j = (m_1(w_1^1), \dots, m_r(w_r^1))], \quad e_2 = [Y_j = (m_1(w_1^2), \dots, m_r(w_r^2))]$$

Se $w_v = \text{Max} \{w_v^1, w_v^2\}$ para $v = 1, \dots, r$ a generalização vem,

$$e := e_1 \cup e_2 = [Y_j = (m_1(w_1^1), \dots, m_r(w_r^1))] \cup [Y_j = (m_1(w_1^2), \dots, m_r(w_r^2))] \quad (4.21)$$

$$:= [Y_j = (m_1(w_1), \dots, m_r(w_r))]$$

A extensão deste evento é dada por:

$$Ext(e) := \{s | w_v^s \leq w_v, v = 1, \dots, r\}$$

Pode-se concluir que a $Ext(e)$ é o conjunto de todos os objectos simbólicos s , que contêm para a variável Y_j um evento $[Y_j = (m_1(w_1^s), \dots, w_r(w_r^s))]$, tal que para cada categoria m_v de Y_j a frequência não ultrapassa w_v .

Exemplo 4.2 Considerem-se os objectos simbólicos RA/F e RA/M descritos pela variável simbólica modal $Y = \text{Apressado}$, associados às seguintes asserções:

$RA/F = [Y = (\text{Não}(0, 508); \text{Sim}(0, 492))]$ e $RA/M = [Y = (\text{Não}(0, 683); \text{Sim}(0, 317))]$
A generalização pelo máximo é $e = [Y = (\text{Não}(0, 683); \text{Sim}(0, 492))]$.

Generalização pelo Mínimo

Neste caso para cada categoria m_v tomamos o ínfimo das suas frequências:

$$\begin{aligned} e := e_1 \cup e_2 &= [Y_j = (m_1(w_1^1), \dots, m_r(w_r^1))] \cup [Y_j = (m_1(w_1^2), \dots, m_r(w_r^2))] \\ &:= [Y_j = (m_1(w_1), \dots, m_r(w_r))] \end{aligned} \quad (4.22)$$

Tal como no caso do Máximo a extensão é dada por:

$$Ext(e) := \{s | w_v^s \geq w_v, v = 1, \dots, r\} \quad \text{com } w_v = \text{Min} \{w_v^1, w_v^2\} \quad \text{para } v = 1, \dots, r.$$

Exemplo 4.3 Considerem-se os objectos simbólicos RA/F e RA/M descritos pela variável simbólica modal $Y = \text{Apressado}$, associados às seguintes asserções:

$RA/F = [Y = (\text{Não}(0, 508); \text{Sim}(0, 492))]$ e $RA/M = [Y = (\text{Não}(0, 683); \text{Sim}(0, 317)) =]$
A generalização pelo mínimo é $e = [Y = (\text{Não}(0, 508); \text{Sim}(0, 317))]$.

Para avaliar a generalidade de um objecto simbólico é utilizado um critério numérico, designado *grau de generalidade*. Este critério permite escolher qual a melhor agregação entre todas as agregações possíveis numa dada etapa. Para variáveis intervalares e categóricas multivaluadas, o critério avalia a proporção do espaço de descrição que é coberto pelo objecto simbólico. Para variáveis modais avalia em quanto a distribuição dada para cada variável se aproxima da distribuição uniforme.

Definição 4.19 *Seja $s = \bigwedge_{j=1}^p e_j$ com $e_j = [Y_j \in D_j]$, $D_j \subseteq \mathcal{Y}_j$ ($j = 1, \dots, p$) os eventos. Assuma-se que todos os domínios $\mathcal{Y}_j$ são conjuntos finitos ou intervalos limitados em $\mathbb{R}$. O grau de generalidade de s é definido por:*

$$G(s) := \prod_{j=1}^p \frac{c(D_j)}{c(\mathcal{Y}_j)} = \prod_{j=1}^p G(e_j), \quad (4.23)$$

onde $c(D_j)$ representa a cardinalidade de D_j se Y_j é categórica multivaluada e o comprimento do intervalo se Y_j é intervalar.

Para cada variável Y_j que corresponde a cada evento e_j , é determinado um factor $G(e_j) \in [0, 1]$; os seus valores são combinados de forma a determinar uma medida de variabilidade do objecto simbólico. A definição do operador $c(\cdot)$ depende do tipo de variáveis e é uma medida do tamanho dos conjuntos subjacentes.

Se as variáveis são intervalares, $c(\cdot)$ é definido como o comprimento do intervalo correspondente e se as variáveis são categóricas “single” ou multivaluadas ou ordinais $c(\cdot)$ é definido como a cardinalidade.

Para variáveis modais com r categorias $m_1, \dots, m_r$ onde se tem associada a frequência de distribuições ou probabilidade $(w_1, \dots, w_r)$ e $e = [Y R \{m_1(w_1), \dots, m_r(w_r)\}]$, com $R \in \{=, \leq, \geq\}$, de acordo com o tipo de generalização tem-se:

Se a generalização é feita pelo máximo com $R \in \{=, \leq\}$, o grau de generalidade de e é definido por:

$$G_1(e) = \sum_{l=1}^r \sqrt{w_l \times \frac{1}{r}} = \frac{\sum_{l=1}^r \sqrt{w_l}}{\sqrt{r}} \quad (4.24)$$

Se a generalização é feita mínimo com $R \in \{=, \geq\}$, o grau de generalidade de e é definido por:

$$G_2(e) = \sum_{l=1}^r \sqrt{(1 - w_l) \times \frac{1}{r(r-1)}} = \frac{\sum_{l=1}^r \sqrt{1 - w_l}}{\sqrt{r(r-1)}} \quad (4.25)$$

Estas medidas avaliam, em cada caso, a semelhança entre a distribuição dada e a distribuição uniforme. $G_1(e)$ é o coeficiente de afinidade [Matusita (1951)] entre $(w_1, \dots, w_r)$ e a distribuição uniforme, sendo este valor máximo quando a distribuição dada é uniforme: $w_l = \frac{1}{r}$ com $l = 1, \dots, r$. $G_2(e)$ é o coeficiente de afinidade entre $(1 - w_1, \dots, 1 - w_r)$ e $(\frac{1}{r}, \frac{1}{r}, \dots, \frac{1}{r})$. Se para a variável modal $w_1 + \dots + w_r = 1$, $G_2(e)$ é máximo, $G_2(e) = 1$, quando a distribuição dada é uniforme. Isto significa que quando consideramos um objecto mais geral que outro as distribuições correspondentes são uniformes.

Na Tabela 4.2 [Brito e De Carvalho (2008)] apresenta-se um resumo dos procedimentos de generalização e os correspondentes operadores para cada tipo de variáveis:

Variáveis	Generalização	$c(.)$
Intervalares	intervalo cobertura	comprimento do intervalo
Catégoricas “single” e multivaluadas	conjunto união	cardinal
Modal	Máximo	G_1
Modal	Máximo	G_2

Tabela 4.2: Operadores de generalização e medidas correspondentes

Classificação em variáveis com regras

Neste trabalho serão utilizadas variáveis modais com regras, pelo que referiremos o procedimento neste caso. Para um aprofundamento deste tema consulte-se [Brito e De Carvalho (1999), Brito e De Carvalho (2008)].

O processo de generalização para variáveis modais com regras pode também ser feito de duas formas, pelo máximo ou pelo mínimo.

Sejam Y e Z duas variáveis simbólicas em que Z depende hierarquicamente de Y (ver Definição 2.4).

Generalização pelo Máximo Neste caso toma-se para cada categoria o máximo das frequências. Se existem regras hierárquicas da forma:

$$Y(u) \in B'_Y \subset B_Y \Rightarrow Z(u) = NA$$

então o máximo é calculado para todas as categorias excepto NA e o peso de NA no objecto generalizado é dado por $\text{Min} \{1, \sum_{m_i \in B'_Y} w_i\}$.

Generalização pelo Mínimo Neste caso toma-se para cada categoria o mínimo das frequências. O mínimo é calculado para todas as categorias excepto NA e o peso de NA no objecto generalizado é dado por $\sum_{m_i \in B'_Y} w_i$.

Grau de generalidade Para variáveis modais com t regras de dependência as medidas de generalidade G_1 (4.24) e G_2 (4.25) são utilizadas. Contudo, na presença de regras, é subtraída a generalidade das descrições não coerentes. Seja $G(s|r_1 \wedge \dots \wedge r_t)$ a generalidade de um objecto simbólico s tendo em consideração as regras $r_1, \dots, r_t$ vem,

$$G(s|r_1 \wedge \dots \wedge r_t) = G(s) - G(s|\neg G(s|r_1 \wedge \dots \wedge r_t))$$

Exemplo 4.4 Sejam $Y = \text{Apressado}$ e $Z = \text{FreqApressado}$ com $\mathcal{Y} = \{\text{Sim}, \text{Não}\}$ e $\mathcal{Z} = \{\text{NA}, \text{com frequência}, \text{todos os dias}, \text{algumas vezes}\}$. Se $Y(u) = \text{Não}$ para algum

u então $Z(u) = NA$.

Considere-se o objectos simbólico associado à asserção:

$$a = [Y = (N\tilde{a}o(0,508), Sim(0,492))] \wedge$$

$$[Z = (NA(0,508), com\ frequ\ência(0,135), todos\ os\ dias(0,138), algumas\ vezes(0,219))]$$

Neste exemplo a regra é, se $Y(u) = N\tilde{a}o$ então $Z(u) = NA$, ou seja, as descrições não coerentes são:

$$Y(u) = \{Sim\} \text{ e } Z(u) = \{NA\},$$

$$Y(u) = \{N\tilde{a}o\} \text{ e } Z(u) = \{com\ frequ\ência, todos\ os\ dias, algumas\ vezes\}$$

Temos então:

$$G_1(s) = \left(\frac{\sqrt{0,508} + \sqrt{0,492}}{\sqrt{2}}\right) \left(\frac{\sqrt{0,508} + \sqrt{0,135} + \sqrt{0,138} + \sqrt{0,219}}{\sqrt{4}}\right) = 0,95878$$

$$G_1(s | \neg r) = \left(\frac{\sqrt{0,492}}{\sqrt{2}}\right) \left(\frac{\sqrt{0,508}}{\sqrt{4}}\right) + \left(\frac{0,508}{\sqrt{2}}\right) \left(\frac{\sqrt{0,135} + \sqrt{0,138} + \sqrt{0,219}}{\sqrt{4}}\right) = 0,48088$$

$$G_1(s|r) = G_1(s) - G_1(s | \neg r) = 0,4779$$

Simplificação da Pirâmide

A simplificação da Pirâmide é designada “*pruning*” e consiste em suprimir algumas classes com o objectivo de tornar a estrutura mais simples de interpretar.

Seja C uma classe com um predecessor único C' (no caso das Pirâmides uma classe pode ter até dois predecessores), e seja f a função de indexação. Suponhamos que $f(C) - f(C') < \gamma$ com $\gamma > 0$ um valor fixo. Neste caso pode-se suprimir C sem grande perda de informação, como se pode observar na Figura 4.1.



Figura 4.1: Simplificação de uma Pirâmide

A escolha de $\gamma > 0$ prende-se com o grau de simplificação que se pretende: se $\gamma > 0$ é muito baixo a estrutura da Pirâmide quase não é alterada e se é muito elevado há uma grande perda de informação. Regra geral é escolhida para $\gamma > 0$ uma pequena percentagem do índice máximo.

4.2.3 Classificação Simbólica Não-Hierárquica

O método de classificação simbólica não-hierárquica implementado no software *SODAS* com o módulo *SCLUST*, tem como objectivo agrupar elementos ou dados complexos, definidos como objectos simbólicos, em classes homogéneas. Este é um método iterativo e não-hierárquico, que agrupa os elementos do conjunto E em K classes homogéneas. O algoritmo é baseado na escolha de protótipos de classes e funções de proximidade que permitem afectar os elementos ou objectos simbólicos de E às classes. O protótipo é um modelo da classe que generaliza as características dos elementos pertencentes à classe e a sua representação pode ser um elemento do mesmo espaço que os elementos ou conceitos que vão ser agrupados.

O método segue a lógica do método das “nuvens dinâmicas” [Diday (1971)]: partindo de um conjunto inicial de “centros” ou de uma partição inicial (representação inicial das classes), o método aplica sucessivamente uma função de alocação seguida de uma função de representação, até à convergência (ou até ao número de iterações permitido). Dado um conjunto de representantes das classes, a função de alocação afecta cada elemento do conjunto a classificar à classe de cujo representante está mais próximo e depende da natureza das variáveis que descrevem os objectos simbólicos. Obtida uma partição, a função de representação determina um novo representante para cada classe. Em cada etapa, o método faz diminuir o valor do critério que avalia a adequação entre as classes e os respectivos representantes. A particularidade deste método assenta na interpretação das classes como conceitos (ver [De Carvalho *et al.* (2008)]).

Seja $u \in E$ um objecto qualquer pertencente ao conjunto dos k objectos descritos por p variáveis simbólicas e seja $\xi_i = (\xi_{i1}, \dots, \xi_{ip})^2$ o vector que descreve o objecto simbólico u_i . Seja P_K o conjunto das partições de E em K classes, $P = \{C_1, \dots, C_K\}$ um elemento de P_K e seja G_t o protótipo da classe C_t . Genericamente os protótipos que representam as classes podem ser pontos médios, grupos de objectos, funções de probabilidade, etc.

Princípio básicos do método:

- o número de classes K é fixado pelo utilizador, mas pode ser redefinido tendo em conta a melhor partição em K classes proposta;
- *Passo de representação*, descrição das K classes $\{C_1, \dots, C_t, \dots, C_K\}$ que formam a partição P do conjunto E , pelo vector de protótipos $L = (G_1, \dots, G_t, \dots, G_K)$;

²Corresponde ao vector ξ_i descrito na Secção 3.2

- *Passo de alocação*, definição de uma função de alocação que afecta os elementos do conjunto E às classes da partição P , conforme a sua proximidade aos protótipos. A proximidade é dada por uma função $\psi(\cdot)$, definida de $E \times \Lambda$, onde Λ é o espaço dos protótipos, tais que $G_t \in \Lambda$ e $L \in \Lambda^K$.

O algoritmo procura uma partição $P^* \in P_K$ de E em K classes, de entre todas as partições possíveis de P_K , e o vector $L^* \in \Lambda^K$ dos K protótipos que representam as classes de P^* , tais que o critério Δ seja minimizado:

$$\Delta(P^*, L^*) = \text{Min} \left\{ \Delta(P, L) \mid P \in P_K, L \in \Lambda^K \right\} \quad (4.26)$$

A convergência do algoritmo para um valor óptimo local da função Δ é garantida pela consistência entre a representação das classes pelos seus protótipos e a função de alocação $\psi(\cdot)$. O critério Δ mede a adequação entre a partição P e o vector L dos K protótipos. O critério é definido pela soma em todas as K classes e em todos os objectos $u_i \in E$ da função de alocação:

$$\Delta(P, L) = \sum_{t=1}^K \sum_{u_i \in C_t} \psi(\xi_i, G_t) \quad (4.27)$$

Uma vez que o critério é aditivo, a convergência fica garantida pela existência e unicidade de $G_t \in \Lambda$. Dado um índice arbitrário no vector $L = (G_1, \dots, G_K)$, definimos a função *argmin* em L por $t = \text{argmin} \{ \psi(\xi_i, G_z) \mid z = 1, \dots, K \}$ para $u_i \in E$, como o índice do elemento de G_t do vector L que verifica as seguintes propriedades: $\psi(\xi_i, G_t) = \min \{ \psi(\xi_i, G_z) \mid z = 1, \dots, K \}$ e o índice de G_t é mínimo.

A definição da função *argmin* garante o decrescimento do critério durante o passo de alocação e a existência e unicidade de $G_t \in \Lambda$ para todas as K classes em P .

O algoritmo proposto neste trabalho baseia-se na minimização do critério Δ e é do tipo *k-means* ([Macqueen (1967)]).

Algoritmo

- *Inicialização*

Seja $G_0 = \{G_1^0, \dots, G_K^0\}$ o conjunto dos centros iniciais formado por K protótipos.

- *Passo de alocação*

Cada objecto $u_i \in E$ é associado a uma classe C_t se e só se $\psi(\xi_i, G_t)$ é mínimo, $C_t^{(h+1)} = \{u_i \in E \mid t = \text{argmin} \{ \psi(\xi_i, G_z) \mid z = 1, \dots, K \} \}$

- *Passo de representação*

Para $t = 1, \dots, K$ é determinado um protótipo $G_t^{(h+1)}$ que representa a classe $C_t \in P^{(h)}$

- Regra de paragem

Se $P^{(h+1)} = P^{(h)}$ pára, caso contrário volta para *Passo de alocação*.

O algoritmo pode ser inicializado com uma partição ou com um conjunto de centros “protótipos”. Neste caso estamos a supor que a inicialização é feita com os protótipos, logo o primeiro passo é a alocação.

Medidas de proximidade

De acordo com a natureza das variáveis que descrevem os objectos simbólicos e os protótipos, são propostas diferentes medidas de proximidade ou distâncias para alocar os elementos de E às classes. Neste trabalho é utilizada a distância de *Hausdorff* para as variáveis intervalares e a distância *Euclideana* para as variáveis modais.

Distância Euclideana Seja $X = (x_{ij})$ uma matriz de dados clássica com as observações para p variáveis quantitativas e sejam $x_i, x_l \in \mathbb{R}^p$. A distância euclidiana entre x_i e x_l em $\mathbb{R}^p$ é dada por:

$$d(x_i, x_l) = \sqrt{\sum_{j=1}^p (x_{ij} - x_{lj})^2} \quad (4.28)$$

Distância de Hausdorff Sejam $([\alpha_{h1}, \beta_{h1}], \dots, [\alpha_{hp}, \beta_{hp}])$ e $([\alpha_{l1}, \beta_{l1}], \dots, [\alpha_{lp}, \beta_{lp}])$ as descrições associadas aos objectos simbólicos, u_h e u_l .

$$d(u_h, u_l) = \sqrt{\sum_{j=1}^p \max\{|\alpha_{hj} - \alpha_{lj}|, |\beta_{hj} - \beta_{lj}|\}^2} \quad (4.29)$$

Se em vez de intervalos estiverem a ser considerados pontos individuais em $\mathbb{R}$ obtém-se a distância *Euclideana*.

Interpretação do método

As ajudas à interpretação do método são índices que medem a qualidade da partição ou a qualidade das classes da partição, tendo como suporte o critério da inércia generalizado aos dados simbólicos, onde os centróides são substituídos pelos protótipos das classes [De Carvalho *et al.* (2008)].

Na análise clássica em que as variáveis tomam valores reais [Celeux *et al.* (1988)], os índices propostos baseiam-se na decomposição da soma total dos quadrados (TSS), que é dividida na soma dos quadrados intra-classe (WSS) e na soma dos quadrados inter-classe (BSS). Supõe-se que os elementos de E têm todos o mesmo peso, igual a 1.

No caso clássico, a decomposição de TSS é dada por:

$$\underbrace{\sum_{i=1}^k d^2(\xi_i, g)}_{TSS} = \underbrace{\sum_{t=1}^K \sum_{u_i \in C_t} d^2(\xi_i, g_t)}_{WSS} + \underbrace{\sum_{t=1}^K n_t d^2(g_t, g)}_{BSS} \quad (4.30)$$

onde d é a distância Euclideana, g é o ponto médio dos K pontos $u_i \in E$, g_t é o ponto médio dos pontos $u_i \in C_t$ e $n_t = \#C_t$.

O ponto médio g_t da classe C_t é o ponto de $\mathbb{R}^p$ que minimiza o seguinte critério de adequação:

$$f_{C_t}(g_t) = \sum_{u_i \in C_t} d^2(\xi_i, g_t)$$

No método proposto, a generalização do ponto médio g_t de uma classe C_t ao protótipo G_t é efectuada por minimização do critério de adequação:

$$f_{C_t}(G_t) = \sum_{u_i \in C_t} \psi(\xi_i, G_t)$$

A inércia total (TSS) e a inércia intra-classe (WSS) são generalizadas utilizando os protótipos definidos e as funções de comparação dadas.

- a generalização de WSS, é o critério $\Delta(P, L) = \sum_{t=1}^K \sum_{u_i \in C_t} \psi(\xi_i, G_t)$;
- a generalização da inércia total, TSS, é o critério de adequação $f_E(G_E) = \sum_{t=1}^K f_{C_t}(G_E)$ onde G_E é o protótipo do conjunto global E .

Assumindo a generalização, o ganho obtido por substituir o protótipo G_E de E pelos K protótipos $(G_1, \dots, G_K)$ da partição P , corresponde à inércia inter-classe (BSS). Pela construção dos protótipos temos:

$$\underbrace{\sum_{t=1}^K f_{C_t}(G_t)}_{\Delta(P, L)} \leq \underbrace{\sum_{t=1}^K f_{C_t}(G_E)}_{f_E(G_E)}, \quad (4.31)$$

logo o ganho obtido na homogeneidade é a diferença entre $\Delta(P, L)$ e $f_E(G_E)$. O ganho pode ser interpretado como a “diferença” entre a hipótese nula “não há estrutura=partição numa classe” e partição em K classes obtida pela minimização do critério Δ .

Os índices utilizados na interpretação da partição e das classes, são:

- $f_{C_t}(G_t)$ que é uma medida de homogeneidade da classe C_t , dada pela função de alocação;

- $\Delta(P, L)$ que é uma medida da homogeneidade intra-classe da partição P ;
- $f_E(E)$ que é uma medida da homogeneidade global do conjunto E .

O problema da otimização, utilizando a distância de *Hausdorff* proposta por [Chavent *et al.* (2002,2003)] na classificação de objectos simbólicos para variáveis intervalares, fica resolvido pela minimização da distância média de *Hausdorff*, d_H , entre todos os intervalos que descrevem a classe e o melhor intervalo $G^j = [\underline{\theta}^j; \bar{\theta}^j]$ com $j = 1, \dots, p$, que represente a classe. O protótipo G_t de uma classe C_t é um vector de intervalos que minimiza o critério de adequação:

$$f_{C_t}(G_t) = \sum_{u_i \in C_t} \psi(\xi_i, G_t) = \sum_{u_i \in C_t} \sum_{j=1}^p d_H(\xi_{ij}, G_t^j) = \sum_{j=1}^p \underbrace{\sum_{u_i \in C_t} d_H(\xi_{ij}, G_t^j)}_{\tilde{f}_{C_t}(G_t^j)}$$

com ξ_i para $i = 1, \dots, k$ a i -ésima linha da matriz Ξ que contém a descrição d_{u_i} .

O solução para minimizar o critério passa por encontrar o intervalo $G^j = [\underline{\theta}^j; \bar{\theta}^j]$ que minimiza:

$$\tilde{f}_{C_t}(G_t^j) = \sum_{u_i \in C_t} \text{Max} \left\{ \left| \underline{\theta}^j - \alpha_i^j \right|, \left| \bar{\theta}^j - \beta_i^j \right| \right\} \quad (4.32)$$

Seja m_i^j o ponto médio do intervalo $\xi_{ij} = [\alpha_i^j; \beta_i^j]$ e l_i^j a metade do seu comprimento:

$$m_i^j = \frac{\alpha_i^j + \beta_i^j}{2} \quad \text{e} \quad l_i^j = \frac{\alpha_i^j - \beta_i^j}{2},$$

e seja μ^j o ponto médio e λ^j a metade do comprimento do intervalo $G^j = [\underline{\theta}^j; \bar{\theta}^j]$. Seguindo a propriedade para x e y em $\mathbb{R}^p$:

$$\text{max} \{ |x - y|, |x + y| \} = |x| + |y|,$$

a equação 4.32 vem:

$$\begin{aligned} \tilde{f}_{C_t}(G_t^j) = \sum_{u_i \in C_t} \text{max} \left\{ \left| (\mu^j - \lambda^j) - (m_i^j - l_i^j) \right|, \left| (\mu^j + \lambda^j) - (m_i^j + l_i^j) \right| \right\} = \\ \sum_{u_i \in C_t} \left| \mu^j - m_i^j \right| + \sum_{u_i \in C_t} \left| \lambda^j - l_i^j \right| \quad (4.33) \end{aligned}$$

Para cada variável Y_j a solução é o intervalo,

$$G_j = [\hat{\mu}^j - \hat{\lambda}^j, \hat{\mu}^j + \hat{\lambda}^j]$$

com

$$\hat{\mu}^j = \text{mediana} \{m_i^j, u_i \in C_t\} \quad \text{e} \quad \hat{\lambda}^j = \text{mediana} \{l_i^j, u_i \in C_t\},$$

e o protótipo da classe C_t é dado por $G_t = (G_1, \dots, G_p)$.

No caso das variáveis modais, o objecto u_i é descrito por p variáveis modais (ver Secção 3.1 Definição 3.6). Seja r_j o número de categorias que a variável Y_j pode tomar e $q_{j,u_i}(m_{jl})$ a frequência associada à categoria m_{jl} com $j = 1, \dots, p$ e $l = 1, \dots, r_j$ da variável Y_j .

A representação simbólica do objecto $u_i \in E$ será agora um vector $(r_1 + \dots + r_p)$ -dimensional dada por:

$$d_{u_i} = (((q_{1,u_i}(m_{11}), \dots, q_{1,u_i}(m_{1r_1})), \dots, (q_{p,u_i}(m_{p1}), \dots, q_{p,u_i}(m_{pr_p}))))$$

A matriz simbólica original $\Xi = (Y_j(u_i))$ é transformada numa matriz de frequências, onde para qualquer $u_i \in E$ e para qualquer $j = 1, \dots, p$, $\sum_{l=1}^{r_j} q_{j,u_i}(m_{jl}) = 1$ (ver Tabela 4.3).

	Y_1			$\dots$	Y_p		
	1	$\dots$	r_1	$\dots$	1	$\dots$	r_p
u_1	$q_{1,u_1}(m_{11})$	$\dots$	$q_{1,u_1}(m_{1r_1})$	$\dots$	$q_{p,u_1}(m_{p1})$	$\dots$	$q_{p,u_1}(m_{pr_p})$
$\vdots$	$\vdots$		$\vdots$		$\vdots$		$\vdots$
u_i	$q_{1,u_i}(m_{11})$	$\dots$	$q_{1,u_i}(m_{1r_1})$	$\dots$	$q_{p,u_i}(m_{p1})$	$\dots$	$q_{p,u_i}(m_{pr_p})$
$\vdots$	$\vdots$		$\vdots$		$\vdots$		$\vdots$
u_k	$q_{1,u_k}(m_{11})$	$\dots$	$q_{1,u_k}(m_{1r_1})$	$\dots$	$q_{p,u_k}(m_{p1})$	$\dots$	$q_{p,u_k}(m_{pr_p})$

Tabela 4.3: Matriz de frequências

O protótipo G_t da classe C_t é então definido por:

$$G_t = (\frac{1}{n_t} \sum_{u_i \in C_t} q_{1,u_i}(m_{11}), \dots, \frac{1}{n_t} \sum_{u_i \in C_t} q_{1,u_i}(m_{1r_1}), \dots, (\frac{1}{n_t} \sum_{u_i \in C_t} q_{p,u_i}(m_{p1}), \dots, \frac{1}{n_t} \sum_{u_i \in C_t} q_{p,u_i}(m_{pr_p}))$$

onde n_t é o número de elementos da classe C_t .

Para as variáveis modais utilizando a distância *Euclidiana*, a função critério a minimizar é:

$$f_{C_t}(G_t) = \sum_{u_i \in C_t} \psi(\xi_i, G_t) = \sum_{u_i \in C_t} \sum_{j=1}^p \sum_{l=1}^{r_j} (q_{lj}^i - q_{lj}^t)^2 \quad (4.34)$$

Interpretação da partição

O índice de qualidade $Q(P)$ da partição P é interpretado como o ganho entre a hipótese nula “não há estrutura=partição numa classe” e partição em K classes obtida pela minimização do critério Δ :

$$Q(P) = 1 - \frac{\Delta(P, L)}{f_E(E)} \quad (4.35)$$

O índice toma valores entre 0 e 1, e é igual a 1 quando todas as classes são reduzidas a um único elemento e é igual a 0 quando temos uma partição de E em uma única classe. Como o critério decresce com o número de classes, só pode ser utilizado quando as partições têm o mesmo número de classes. Como uma partição é melhor que outra se o critério $\Delta(P, L)$ for menor, então uma partição é melhor se $Q(P)$ for maior. Neste contexto, $Q(P)$ mede a homogeneidade de E explicada por P .

Para cada variável Y_j , define-se de forma semelhante a qualidade da partição na variável Y_j :

$$Q_j(P) = 1 - \frac{\sum_{t=1}^K f_{C_t}(G_t^j)}{f_E(G_E^j)}. \quad (4.36)$$

Este critério devolve a parte da homogeneidade das classes C_t em P dada pela variável Y_j , ou seja, mede o “poder de discriminação” da variável na partição. A qualidade de P dada por $Q(P)$ é uma média ponderada dos valores de $Q_j(P)$:

$$Q(P) = \sum_{j=1}^p \frac{f_E(G_E^j)}{f_E(G_E)} Q_j(P). \quad (4.37)$$

Este índice mede a importância da variável Y_j na construção da partição P .

Interpretação das classes

A qualidade de uma classe C_t é dada por:

$$Q(C_t) = 1 - \frac{f_{C_t}(G_t)}{f_{C_t}(G_E)} \quad (4.38)$$

Se o valor $Q(C_t)$ é próximo de 1 então os elementos da classe são muito diferentes do protótipo G_E .

Por outro lado, uma medida de contribuição da classe C_t para a variabilidade total é dada pelo rácio entre a homogeneidade da classe e o critério da homogeneidade global $\Delta(P, L)$:

$$K(C_t) = \frac{f_{C_t}(G_t)}{\Delta(P, L)} \quad (4.39)$$

A soma de todas as contribuições é igual a 1.

Um indicador para caracterizar a classe em relação à variável Y_j é:

$$Q_j(C_t) = 1 - \frac{f_{C_t}(G_t^j)}{f_{C_t}(G_E^j)}. \quad (4.40)$$

Quando este valor é próximo de 1 então a descrição do protótipo G_t é muito diferente de G_E . Este critério é útil para encontrar variáveis que caracterizem a classe C_t . Como a qualidade da classe é dada por $Q(C_t)$, que é uma média ponderada pelos valores de $Q_j(C_t)$:

$$Q(C_t) = \sum_{j=1}^p \frac{f_{C_t}(G_E^j)}{f_{C_t}(G_E)} Q_j(C_t), \quad (4.41)$$

os valores do critério $Q_j(C_t)$ têm que ser interpretados pela comparação com o valor de $Q(C_t)$. A variável Y_j caracteriza a classe C_t se $Q_j(C_t) > Q(C_t)$.

Métodos para determinar qual o melhor número de classes

No módulo *SCLUST* está disponível um procedimento que permite determinar qual o melhor número de classes a formar, de entre várias soluções. Existem vários métodos que fornecem uma série de valores associados a cada partição e que permitem escolher qual o melhor número de classes.

O problema da classificação está em encontrar uma partição $P = \{C_1, \dots, C_K\}$ de E em K classes. O método para determinar qual o melhor número de classes, é aplicado ao conjunto de partições produzidas pelo *SCLUST*, e funciona como uma regra de paragem. Três métodos podem ser aplicados, o método de Calinski and

Harabasz ([Calínski e Harabasz (1974)]), o C-index ([Hubert e Levin (1976)]) e o Γ -index ([Baker e Hubert (1976)]).

- **Calinski and Harabasz (C-H)**: este índice é dado por,

$$CH = [B/(K - 1)]/[W/(k - K)] \quad (4.42)$$

onde k é o número total de objectos e K o número de classes em cada partição. W e B representam, respectivamente, a soma dos quadrados das distâncias intra-classes (aos protótipos) e a soma dos quadrados das distâncias inter-classes.

O valor máximo do índice é utilizado para indicar o número ideal de classes.

- **C-index**: este índice requer a soma V de todos os pares de dissemelhanças intra-classes. Se a partição tiver r dissemelhanças, V_{min} (respectivamente V_{max}), será a soma dos r mais pequenos (respectivamente maiores) pares de dissemelhanças. O índice é dado por:

$$C = \frac{V - V_{min}}{V_{max} - V_{min}} \quad (4.43)$$

O valor mínimo do índice para todas as partições será considerado para determinar o melhor número de classes. O valor mínimo óptimo é 0.

- **Γ -index**: neste método as comparações são feitas com todos os pares de dissemelhanças intra-classes e todos pares de dissemelhanças inter-classes. A comparação entre os pares de dissemelhanças é considerada consistente (respectivamente inconsistente), se a dissemelhança intra-classe é estritamente menor (respectivamente maior) que a dissemelhança inter-classe. O índice é dado por:

$$\Gamma = \frac{\Gamma_+ - \Gamma_-}{\Gamma_+ + \Gamma_-} \quad (4.44)$$

onde Γ_+ (respectivamente Γ_-) representa o número de comparações consistentes (respectivamente inconsistentes). O valor máximo do índice indica o melhor número de classes, sendo o valor máximo óptimo igual 1.

Capítulo 5

Descrição e Geração dos Dados

5.1 Inquérito à Ocupação do Tempo

O Inquérito à Ocupação do Tempo (IOT) foi realizado pelo INE em 1999 e abrangeu o território do Continente e das Regiões Autónomas dos Açores e da Madeira, sendo estabelecido que os resultados deveriam ser representativos ao nível de NUTS I e NUTS II e das Áreas Metropolitanas de Lisboa e do Porto. A população em estudo é constituída pelos indivíduos residentes em unidades de alojamento não colectivas, com 6 ou mais anos de idade. A amostra foi dimensionada de modo independente para cada uma das sete regiões a nível NUTS II e ainda para cada uma das Áreas Metropolitanas de Lisboa e do Porto. Foram consideradas duas unidades amostrais, o indivíduo e o alojamento. As dimensões amostrais obtidas foram 5439 alojamentos e 10013 indivíduos para o total do país.

A base de amostragem é uma amostra de unidades de alojamento destinada a servir de base de sondagem dos inquéritos a realizar pelo INE junto das famílias. É uma amostra probabilística (a probabilidade de selecção de cada indivíduo é conhecida) areolar (basea-se nas áreas) e bietápica (feita em duas etapas), construída a partir dos dados do Recenseamento da População e Habitação de 1991. Na primeira etapa foram seleccionadas Freguesias e na segunda etapa foram seleccionadas Secções Estatísticas, sendo que em ambas as etapas a selecção processou-se com probabilidade proporcional à dimensão. Em cada uma das secções seleccionadas, os alojamentos foram estratificados de acordo com o número e tipo de indivíduos: alojamentos com um único indivíduo (adulto); alojamentos com pelo menos dois indivíduos, sendo um adulto (15 ou mais anos) e o outro criança (6 - 14 anos); alojamentos com pelo menos dois indivíduos, ambos adultos. A amostra também foi equitativamente distribuída

por todos os dias da semana.

Foram realizados três questionários, podendo a informação obtida através de cada um deles ser cruzada com a restante: Questionário de Família; Questionário Individual; Diário Individual. Para este estudo apenas foram utilizados o Questionário de Famílias e o Questionário Individual.

Questionários

- **Questionário de Família**

- Composição da família
- Características dos residentes (idade, sexo, parentesco, situação familiar, nível de instrução)
- Apoio à família
- Caracterização da utilização de alguns equipamentos (televisor, computador pessoal)
- Rendimento

- **Questionário Individual**

- Actividade principal
- Profissão
- Actividade secundária
- Ocupação do tempo
- Lazer/recreação
- Televisão
- Trabalho doméstico
- Cuidados à família
- Apoio dado a outras famílias
- Trabalho voluntário

5.2 Variáveis Clássicas

Na primeira fase do estudo escolheram-se as variáveis clássicas de interesse de cada um dos questionários.

As variáveis qualitativas, retiradas do questionário às famílias, permitem considerar vários níveis de agregação dos dados e avaliar o sentimento de pressa. As variáveis quantitativas, retiradas do questionário individual, medem a forma como o indivíduo distribui o seu tempo diário. Foram consideradas 11 variáveis qualitativas e 29 quantitativas.

- **Variáveis Qualitativas**

- ***Sexo***: Sexo
F=Feminino
M=Masculino
- ***Grupo-Etário***: Grupo Etário
15-24=15 - 24 anos
25-34=25 - 34 anos
35-54=35 - 54 anos
55-64=55 - 64 anos
65 ou mais=65 ou mais anos
- ***Região***: NUTSII
RN=Norte
RC=Centro
RLVT=Lisboa e Vale do Tejo
RA=Alentejo
RALG=Algarve
- ***Apressado***: Sente que anda apressado
NS/NR
Sim
Não
- ***Freq-apressado***: Com que frequência anda apressado
NS/NR
Todos os dias
Com frequência
Algumas vezes
- ***Apressado-5anos***: Sente-se apressado em comparação com a situação há 5 anos atrás
NS/NR
Anda mais apressado

- Na mesma
Menos apressado
- ***Apressado-3anos***: Sente-se apressado em comparação com a situação há três anos atrás
NS/NR
Anda mais apressado
Na mesma
Menos apressado
- ***Apressado-1ano***: Relativamente ao ano passado
NS/NR
Anda mais apressado
Na mesma
Menos apressado
- ***vp-ntft*** Vida particular - tem tempo para realizar todas as tarefas
NS/NR
Sim, todos os dias
Sim, com frequência
Sim, algumas vezes
Sim, mas raramente
Nunca
- ***tem-fds***: No fim de semana - tem tempo para realizar tudo o que queria
NS/NR
Sim, todos os dias
Sim, com frequência
Sim, algumas vezes
Sim, mas raramente
Nunca
- ***tempo-disp***: Tempo disponível sem saber o que fazer com ele
NS/NR
Todos os dias
Com frequência
Algumas vezes
Raramente
Nunca

• Variáveis Quantitativas

Variáveis medidas em minutos que registam o tempo usado em cada uma das seguintes actividades:

- *Dormir*: Sono
- *Alimentação*: Comer e beber
- *Cuidados pessoais*: Outros cuidados pessoais
- *Trabalho*: Trabalho profissional principal
- *2ª Actividade*
- *Tempo-emprego*: Tempo relacionado com o emprego
- *Escola*: Escola/universidade
- *Estudo*: Cursos complementares e estudo realizado nas horas livres
- *Prep-alimentos*: Preparação de alimentos
- *Arranjo da casa*: Arranjo da casa
- *Roupa e Calçado*: Confecção e tratamento de vestuário e calçado
- *Jardim*: Jardinagem, cultivo de hortas e quintais e cuidados a animais
- *Construção*: Construção e reparação
- *Compras*: Compra de bens e serviços
- *Gestão da casa*
- *Cuidar crianças*: Cuidados às crianças.
- *Cuidar adultos*: Cuidados e assistência a adultos da própria família
- *Convívio*: Convívio e actividades sociais
- *Cultura*: Entretenimento e cultura
- *Lazer passivo*
- *Exercício físico*
- *Exercício produtivo*
- *Artes*
- *Jogos*
- *Outros passatempos*
- *Ler*: Leitura
- *Televisão*: TV e vídeo
- *Música*: Rádio / música
- *Trajecto casa*

5.3 SODAS Software

O software utilizado para efectuar o tratamento dos dados foi o *SODAS*.

O *SODAS* é um software criado especialmente para analisar dados simbólicos. É composto por módulos, em que cada módulo corresponde a um método estatístico. Cada método é manipulado como um ícone e todos os ícones ficam agregados numa “*chaining*” (ver Figura 5.1). O primeiro ícone corresponde ao ficheiro *SODAS*, tabela de dados simbólicos, e os restantes aos diversos métodos. A tabela de dados simbólicos que serve de base de estudo neste software, contém as variáveis simbólicas nas colunas e nas linhas as descrições simbólicas.

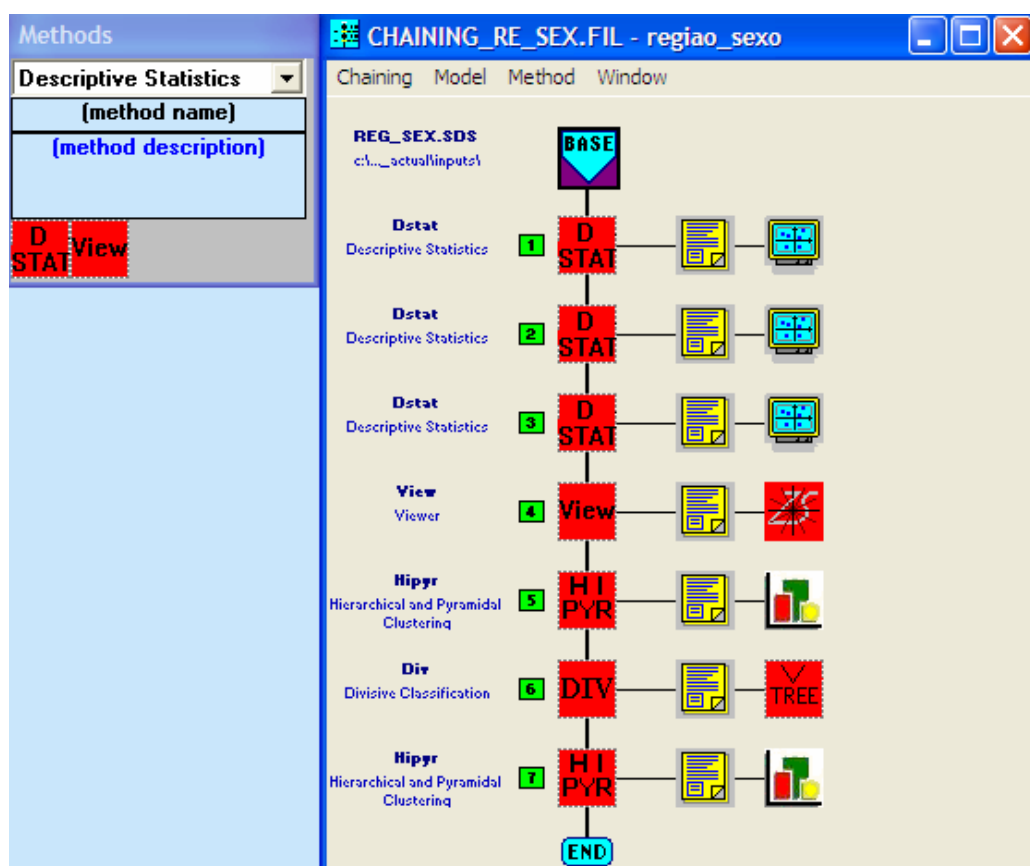


Figura 5.1: *SODAS - chaining*

Resumo de alguns dos métodos do software *SODAS*

- **DB2SO**, “da base de dados para os objectos simbólicos”: é um módulo que gera a matriz de dados simbólicos, a partir de bases de dados relacionais. O resultado da aplicação deste módulo é um ficheiro *SODAS*. Cada objecto

simbólico gerado é uma asserção descrevendo as propriedades de um grupo de indivíduos. As variáveis com regras também podem ser definidas neste módulo. Os dados são gerados através de *queries* em SQL.

- **D-STAT, estatísticas básicas:** calcula estatísticas descritivas e representações gráficas, histogramas, frequências relativas, capacidades e biplots.
- **VIEW, visualização:** permite a visualização das descrições dos grupos através da representação em estrelas.
- **DIV, classificação pelo método divisivo:** devolve uma Hierarquia de objectos simbólicos pelo método de classificação divisivo. O output é uma Hierarquia de partições de duas até K classes, numa árvore de classificação.
- **HIPYR, classificação pelo método em Hierarquia ou Pirâmide:** tem como output uma Hierarquia ou Pirâmide de acordo com os parâmetros escolhidos.
- **SCLUST, classificação simbólica:** tem como output uma partição em classes de objectos simbólicos obtidas de acordo com os parâmetros introduzidos. O método também fornece ajudas de interpretação, como a consistência da classe e a contribuição de cada variável. O utilizador pode escolher o número de classes que pretende ou pode escolher que o método determine qual o melhor número de classes.

5.4 Geração dos Dados Simbólicos

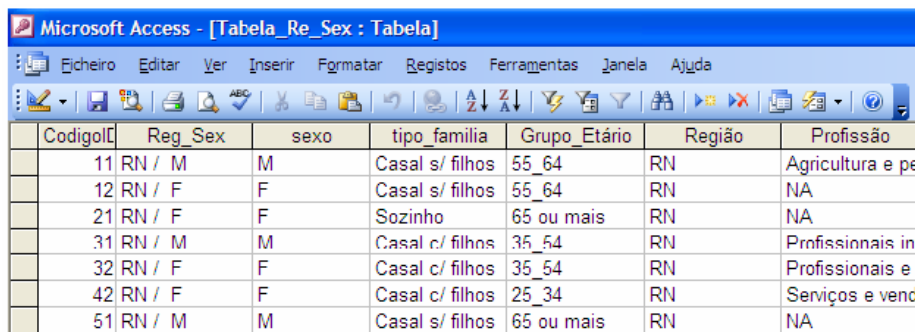
A partir dos micro-dados fornecidos foram criadas duas bases de dados, uma a partir de cada questionário e de seguida foi feita uma ligação entre as duas bases de dados obtendo-se uma tabela única com os micro-dados. Após uma recodificação das variáveis com *labels* mais adequados, foram escolhidas três variáveis clássicas qualitativas (ver Secção 3.3.1) para produzir dois níveis de agregação dos dados, a que correspondem os objectos de segunda-ordem.

As variáveis escolhidas foram: *Região*, com cinco categorias, *Sexo*, com duas categorias e *Grupo Etário*, com 5 categorias.

O objectivo é estudar os objectos simbólicos, criados através do cruzamento das *Regiões*, por *Sexo* e por *Grupo Etário*.

- Primeiro nível de agregação: *Região/Sexo*
- Segundo nível de agregação: *Região/Sexo/Grupo Etário*

Os dados foram agregados através do produto cartesiano das modalidades de cada variável escolhida para formar as classes e cada indivíduo ficou assim associado às várias classes formadas. Numa última fase foram criadas tantas tabelas quantas os cruzamentos efectuados (ver Figura 5.2), ou seja, obtemos duas tabelas que correspondem à matriz $\tilde{X}$ (ver Secção 3.3.1), estando na primeira coluna a classe, nas restantes colunas as variáveis em análise e por último o ponderador associado a cada indivíduo.



Codigol	Reg_Sex	sexo	tipo_familia	Grupo_Etário	Região	Profissão
11	RN / M	M	Casal s/ filhos	55_64	RN	Agricultura e pe
12	RN / F	F	Casal s/ filhos	55_64	RN	NA
21	RN / F	F	Sozinho	65 ou mais	RN	NA
31	RN / M	M	Casal c/ filhos	35_54	RN	Profissionais in
32	RN / F	F	Casal c/ filhos	35_54	RN	Profissionais e
42	RN / F	F	Casal c/ filhos	25_34	RN	Serviços e vend
51	RN / M	M	Casal s/ filhos	65 ou mais	RN	NA

Figura 5.2: Excerto Base Dados Access - *Região/Sexo*

O primeiro nível de cruzamento efectuado, *Região/Sexo*, tem para a variável *Região* 5 modalidades possíveis e para a variável *Sexo* duas modalidades, logo são possíveis $5 \times 2 = 10$ objectos simbólicos. No segundo nível temos mais cinco modalidades da variável *Grupo Etário*, logo são possíveis $5 \times 2 \times 5 = 50$ objectos simbólicos.

Após a escolha das variáveis para construção das classes e a criação da matriz ou das matrizes $\tilde{X}$ a partir da matriz de dados clássica X , procedemos à generalização das classes, através da descrição de cada uma sob a forma de um objecto simbólico. A construção dos objectos simbólicos e o processo de generalização são efectuados com o módulo DB2SO do *SODAS*.

As 29 variáveis clássicas quantitativas associadas à ocupação do tempo produzem 29 variáveis simbólicas intervalares, através do método dado por 3.10 na Secção 3.3.1: *Dormir, Alimentação, Cuidados pessoais, Trabalho, 2ª Actividade, Tempo-emprego, Escola, Estudo, Prep-alimentos, Arranjo casa, Roupas e Calçado, Jardim, Construção, Compras, Gestão da casa, Cuidar crianças, Cuidar adultos, Convívio, Cultura, Lazer passivo, Exercício físico, Exercício produtivo, Artes, Jogos, Outros passatempos, Ler, Televisão, Música e Trajecto casa.*

As variáveis intervalares descrevem cada elemento ou classe indicando como se distribui o tempo em intervalos de minutos, pelas várias actividades durante um dia.

As 8 variáveis qualitativas clássicas resultam em 8 variáveis simbólicas modais, através do método dado por 3.11, indicado na Secção 3.3.1: *Apressado*, *Freq-apressado*, *Apressado-5anos*, *Apressado-3anos*, *Apressado-1ano*, *vp-ntft*, *tem-fds*, *tempo-disp*. As modalidades das variáveis modais são as modalidades das variáveis clássicas.

As variáveis modais traduzem a distribuição de frequências que cada um dos grupos tem relativamente às categorias da variável clássica.

Através do módulo DB2SO formam-se os objectos simbólicos e a matriz de dados simbólicos. Cada objecto simbólico generaliza os elementos pertencentes a cada classe. Na Figura 5.3 pode observar-se uma parte de uma das matrizes de dados simbólicos, resultante do processo de generalização, para o cruzamento *Região/Sexo/Grupo Etário*, cujas linhas são as descrições simbólicas de cada objecto criado.

	Tempo_disp	Dormir
RALG / M / 15_2	Raram (0.141), NA (0.744), Algum (0.076), Nunca (0.039)	[260.00 : 770.00]
RALG / F / 15_2	Raram (0.164), NA (0.533), Algum (0.113), Nunca (0.141), Com f (0.049)	[360.00 : 800.00]
RALG / M / 25_3	Raram (0.095), NA (0.759), Nunca (0.146)	[60.00 : 1060.00]
RALG / F / 25_3	Raram (0.198), NA (0.318), Algum (0.090), Nunca (0.395)	[330.00 : 780.00]
RA / F / 25_34	Raram (0.308), NA (0.272), Algum (0.066), Nunca (0.354)	[340.00 : 640.00]
RA / F / 15_24	Raram (0.257), NA (0.443), Algum (0.134), Nunca (0.124), Com f (0.018), Todos (0.023)	[300.00 : 830.00]
RC / F / 15_24	Raram (0.252), NA (0.421), Algum (0.108), Nunca (0.134), NS/NR (0.010), Com f (0.065), Todos (0.011)	[240.00 : 890.00]
RC / F / 25_34	Raram (0.315), NA (0.190), Algum (0.124), Nunca (0.338), Com f (0.014), Todos (0.019)	[320.00 : 990.00]
RC / M / 15_24	Raram (0.168), NA (0.514), Algum (0.177), Nunca (0.124), Com f (0.016)	[170.00 : 1020.00]
RC / M / 25_34	Raram (0.169), NA (0.205), Algum (0.181), Nunca (0.422), Com f (0.023)	[280.00 : 800.00]
RLVT / F / 25_3	Raram (0.202), NA (0.251), Algum (0.048), Nunca (0.489), Com f (0.005), Todos (0.004)	[220.00 : 770.00]
RLVT / M / 25_3	Raram (0.309), NA (0.330), Algum (0.113), Nunca (0.223), Com f (0.026)	[210.00 : 720.00]
RLVT / F / 15_2	Raram (0.345), NA (0.266), Algum (0.162), Nunca (0.183), NS/NR (0.006), Com f (0.039)	[140.00 : 710.00]
RLVT / M / 15_2	Raram (0.211), NA (0.401), Algum (0.136), Nunca (0.148), Com f (0.091), Todos (0.013)	[150.00 : 1250.00]
RN / F / 25_34	Raram (0.316), NA (0.255), Algum (0.089), Nunca (0.303), Com f (0.028), Todos (0.008)	[190.00 : 980.00]
RN / F / 15_24	Raram (0.272), NA (0.427), Algum (0.134), Nunca (0.138), Com f (0.029)	[80.00 : 770.00]

Figura 5.3: Excerto ficheiro *SODAS - Região/Sexo/Grupo Etário*

Também foram aqui definidas as variáveis com regras de dependência.

As variáveis *Freq-apressado*, *Apressado-5anos*, *Apressado-3anos*, *Apressado-1ano*, *vp-ntft*, *tem-fds*, *tempo-disp* foram consideradas variáveis dependentes hierarquicamente (ver Definição 2.4) da variável *Apressado*. As regras são as seguintes:

- Se $Apressado = \{Não\} \Rightarrow Freq-apressado = \{NA\}$
- Se $Apressado = \{Não\} \Rightarrow Apressado-5anos = \{NA\}$

- Se $Apressado=\{N\tilde{a}o\} \Rightarrow Apressado-3anos=\{NA\}$
- Se $Apressado=\{N\tilde{a}o\} \Rightarrow Apressado-1ano=\{NA\}$
- Se $Apressado=\{N\tilde{a}o\} \Rightarrow vp-ntft=\{NA\}$
- Se $Apressado=\{N\tilde{a}o\} \Rightarrow tem-fds=\{NA\}$
- Se $Apressado=\{N\tilde{a}o\} \Rightarrow tempo-disp=\{NA\}$

Capítulo 6

Análises Simbólicas

6.1 Estatísticas Descritivas

As análises seguintes foram efectuadas com os módulos D-STAT e VIEW do software *SODAS* que permitem efectuar análises de estatísticas descritivas e visualizar os dados. Através da ferramenta D-STAT calculam-se as estatísticas descritivas básicas, histogramas e bi-plots. A funcionalidade VIEW permite visualizar os objectos simbólicos em estrelas, com máximos e mínimos se a variável é intervalar, e a distribuição de frequências se a variável é modal. A visualização pode ser feita em 2D ou 3D. Para variáveis intervalares as opções 2D ou 3D, têm a escala dos eixos feita com base no mínimo e no máximo globais, atingidos na variável, independentemente do objecto simbólico, mas cada variável tem o seu máximo e o seu mínimo representado no eixo. Para variáveis modais a visualização em 2D não é tão rica como em 3D, pois só permite observar os máximos obtidos em cada variável, enquanto que, em 3D obtém-se o diagrama de barras. Com a ferramenta VIEW, no primeiro nível de agregação, *Região/Sexo*, são analisados todos os objectos simbólicos e no segundo nível, *Região/Sexo/Grupo Etário*, são analisados apenas os objectos respectivos às classes etárias dos 15 aos 24 anos e dos 25 aos 34 anos, de forma a não tornar a exposição da análise dos dados demasiado extensa.

Utilizando o 1º cruzamento *Região/Sexo* e as variáveis intervalares relativas à ocupação do tempo, observa-se nos histogramas (ver fórmula 4.6 na Secção 4.1.1) na Figura 6.1, que a distribuição de frequências tende a concentrar-se nos valores mais baixos em quase todas as variáveis. Apenas a variável relativa a *Dormir*, tem maior frequência nos valores centrais. Verifica-se também que as actividades com mais tempo ocupado

são *Dormir*, a *Alimentação*, o *Trabalho*, o *Arranjo da casa* e o *Jardim*.

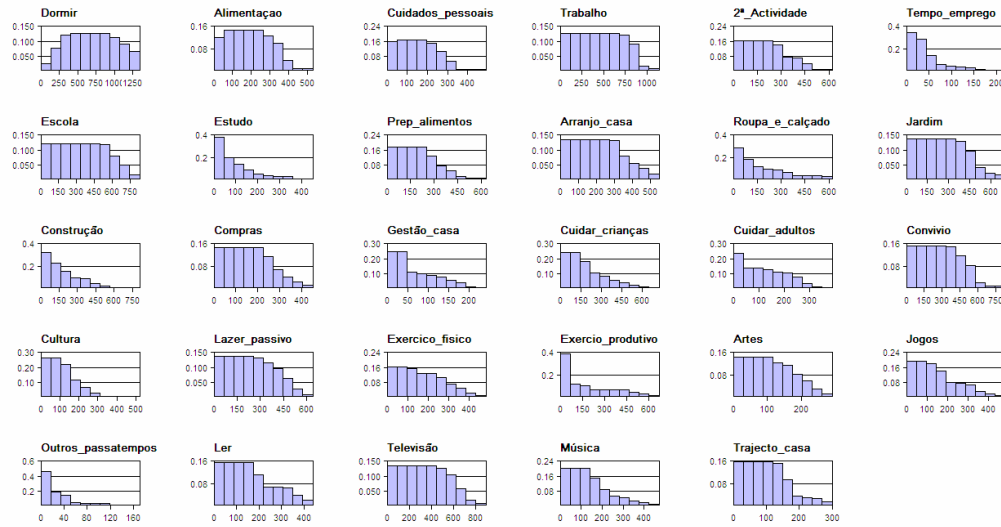


Figura 6.1: Distribuições de frequências das variáveis intervalares - *Região/Sexo*

Descendo ao nível da *Região/Sexo/Grupo Etário* (ver Figura 6.2), para as mesmas variáveis, podemos verificar que as distribuições de frequências revelam algumas diferenças. Com a introdução do *Grupo Etário* as variáveis têm os valores mais concentrados em algumas classes, as primeiras, com excepção da actividade *Dormir* que se concentra agora ainda mais nas classes dos 250 minutos aos 750 minutos. Todas as actividades estão agora menos dispersas pelas classes, sendo de salientar que a *Escola*, *Estudo*, *Gestão da casa* e *Cuidar de adultos* têm peso superior a 60 % das observações na primeira classe.

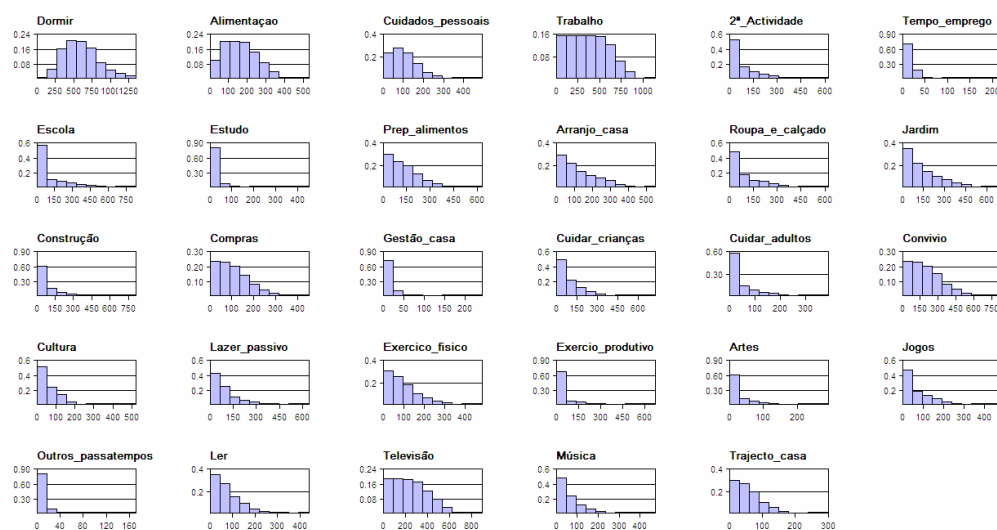


Figura 6.2: Distribuições de frequências das variáveis intervalares - *Região/Sexo/Grupo Etário*

Analisando a representação em estrela dos objectos simbólicos criados no 1º nível de cruzamento (ver Figuras 6.3, 6.4), onde para cada variável é indicado o intervalo de variação do número de minutos dedicados à actividade correspondente, tem-se para as mulheres da região Norte, as actividades relacionadas com a *Gestão da casa*, *Cuidar de crianças*, *Roupa e Calçado* e *Arranjo da casa*, a atingir os seus máximos. Se compararmos com a região de Lisboa e Vale do Tejo, as actividades com valores mais altos são a *Escola*, *Roupa e Calçado*, *Preparação de alimentos*, *Trabalho*, *Compras* e *Ler*.

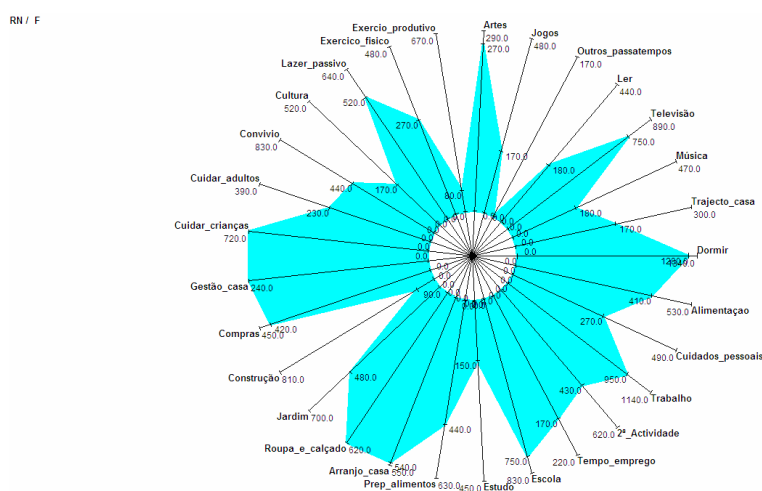


Figura 6.3: Estrela 2D RN/F

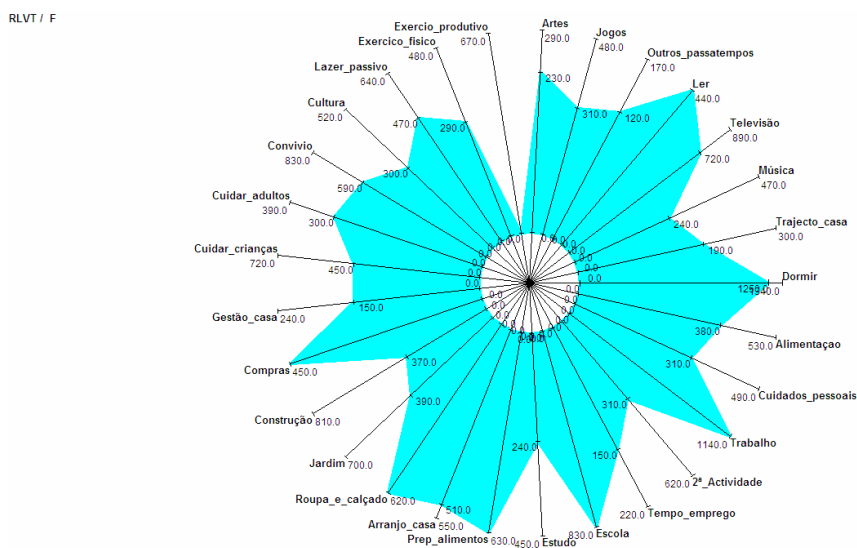


Figura 6.4: Estrela 2D RLVT/F

Já na região do Alentejo (ver Figura 6.5) as variáveis que atingem os seus máximos para as mulheres são a *2ª Actividade*, *Dormir* e *Tempo-emprego*.

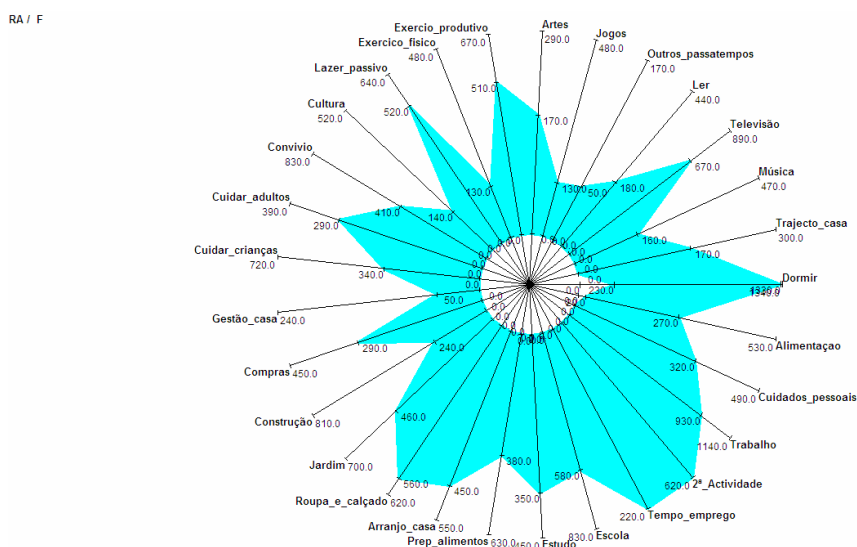


Figura 6.5: Estrela 2D RA/F

Na região de Lisboa e Vale do Tejo para o sexo masculino (ver Figura 6.6), temos várias actividades que atingem os seus máximos, como as *Artes*, *Outros passatempos*, *Música*, *Trajecto para casa*, *Cultura* e *Cuidar de adultos*.

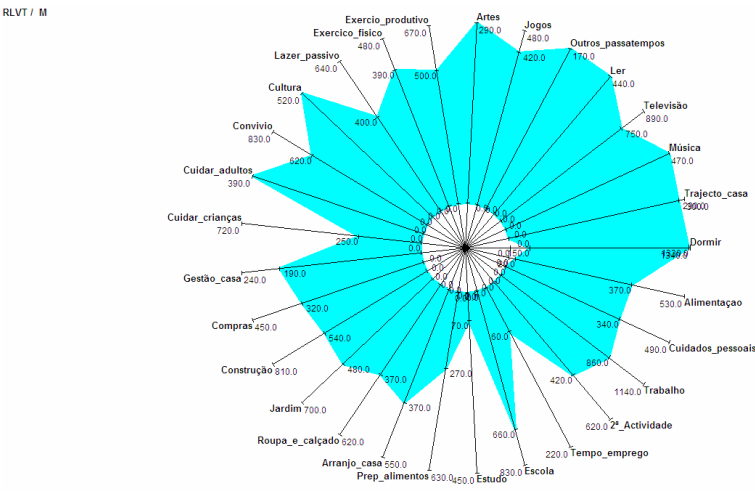


Figura 6.6: Estrela 2D - Objecto Simbólico RLVT/M

Na região Norte os homens (ver Figura 6.7), ocupam mais tempo com a *Televisão*, *Jogos*, *Convívio* e *Escola*. Os homens no Alentejo (ver Figura 6.8), além do *Exercício produtivo* que atinge o máximo, também dão grande importância a *Ler*, *Construção* e *Jardim*.

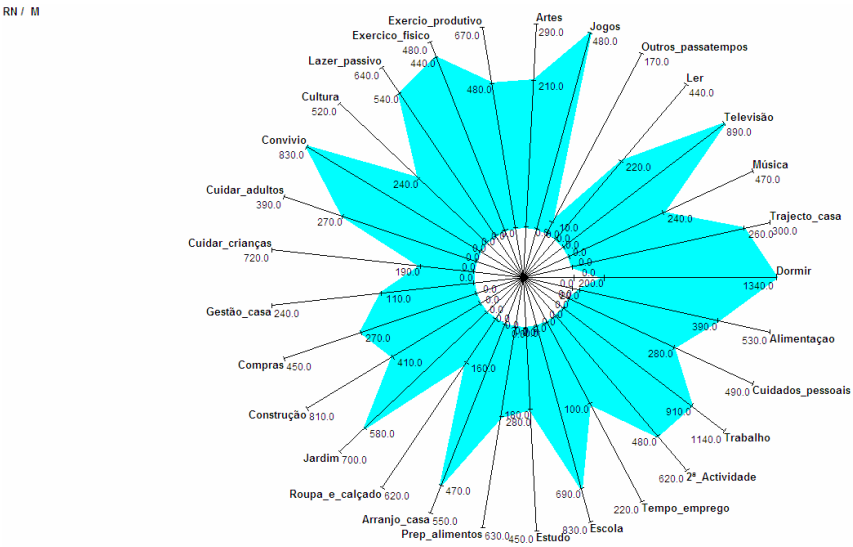


Figura 6.7: Estrela 2D - Objecto Simbólico RN/M

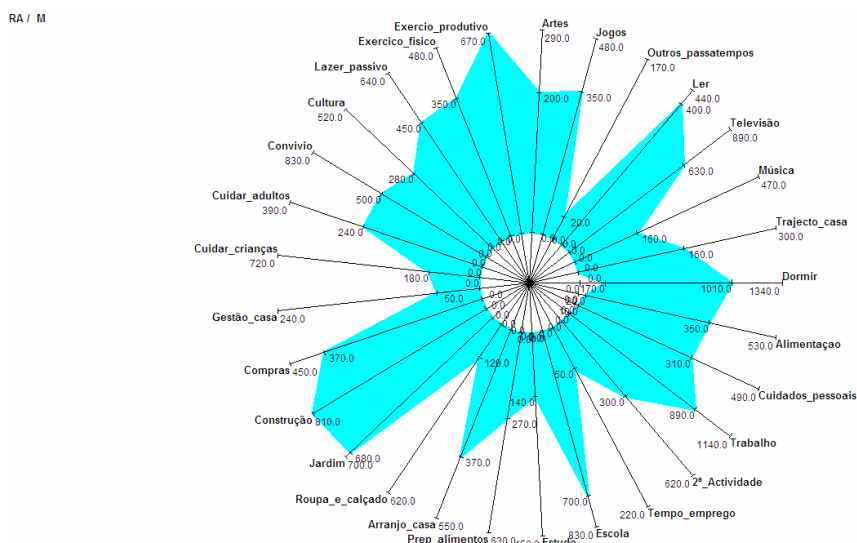


Figura 6.8: Estrela 2D - Objecto Simbólico RA/M

Descendo agora ao nível da classe etária, as Figuras 6.9 e 6.10 apresentam, em sobreposição, estrelas representando a descrição de dois grupos, permitindo assim verificar directamente, quais as variáveis para as quais as descrições são análogas e aquelas onde há diferenças importantes. Cada uma destas figuras compara as descrições para um sexo e região, dos indivíduos mais jovens (15-24 anos) e com idades compreendidas entre os 25 e os 34 anos.

Observamos pela Figura 6.9, que na região Norte para o sexo feminino em ambos os grupos etários, verifica-se uma grande ocupação do tempo com *Cuidar de crianças*, *Compras* e *Arranjo da casa*. As *Artes* adquirem grande importância no *Grupo Etário* dos 25 aos 34 anos e a *Escola* no *Grupo Etário* mais novo. Quanto ao sexo masculino (ver Figura 6.10) observa-se que o *Convívio* e os *Jogos* são as actividades que mais tempo ocupam na classe etária dos 25 aos 34 anos. É de salientar que o *Arranjo da casa* também tem grande importância neste *Grupo Etário*. Dos 15 aos 24 mantém-se as actividades da *Escola* e *Convívio* e verifica-se que o *Trajecto casa* adquire mais importância.

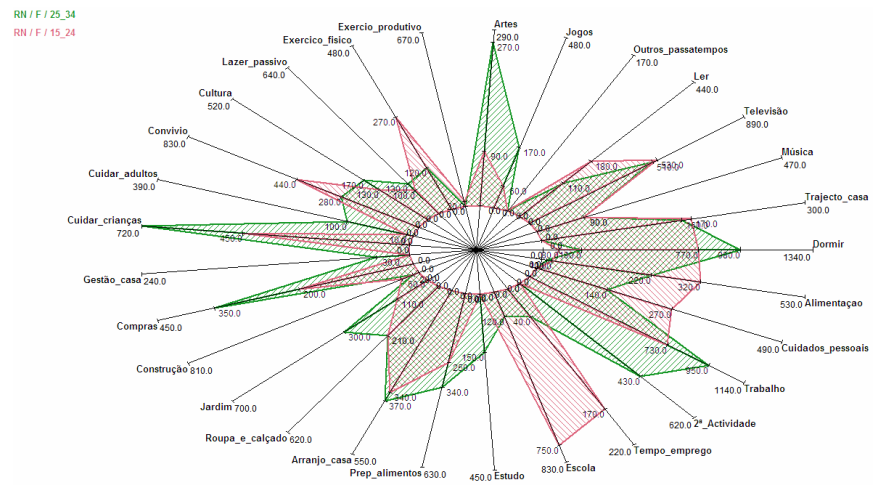


Figura 6.9: Estrela 2D *Superimpose* - Objectos Simbólicos RN/F/15-24 e RN/F/25-34

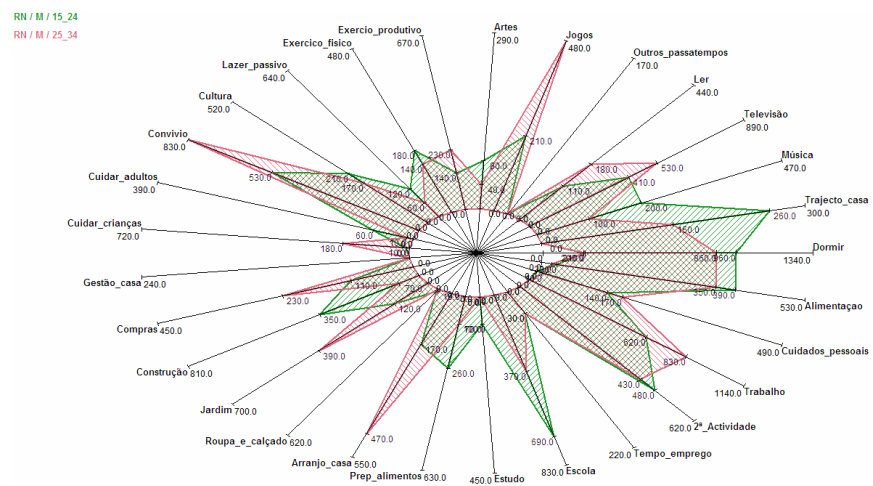


Figura 6.10: Estrela 2D *Superimpose* - Objectos Simbólicos RN/M/15-24 e RN/M/25-34

Na região de Lisboa e Vale do Tejo, como podemos observar na Figura 6.11, a *Escola*, a *Televisão* e os *Jogos* têm mais relevância nas mulheres dos 15 aos 24 anos, e as *Compras* e *Arranjo da casa* nas mulheres dos 25 aos 34 anos. A leitura que antes atingia o seu máximo deixa de ter tanta importância principalmente no *Grupo Etário* mais novo. Se observarmos o sexo masculino (ver Figura 6.12) nesta região entre os 15 e os 24, quanto a actividades de lazer, o *Convívio*, a *Televisão*, os *Jogos* e o *Exercício físico*, são as actividades que mais tempo ocupam. No *Grupo Etário* dos 25 aos 34 anos, as actividades de lazer mais relevantes continuam a ser a *Televisão* e o *Convívio*, este com menos importância. O *Trabalho* e a *Escola* são agora as actividades que ocupam mais tempo.

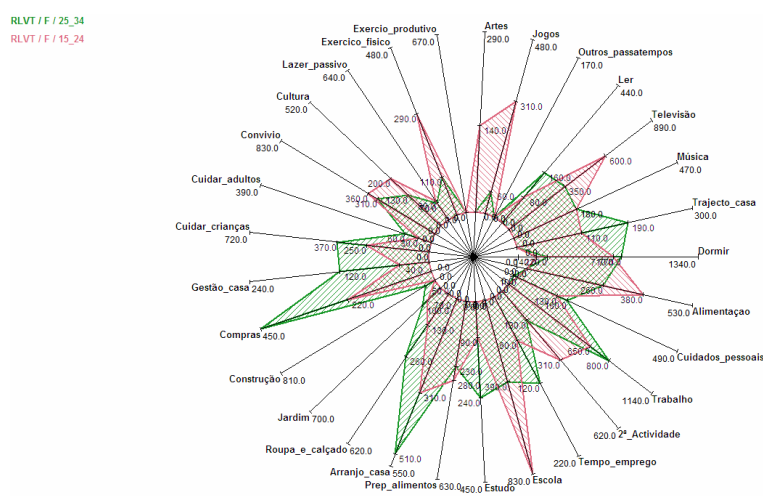


Figura 6.11: Estrela 2D *Superimpose* - Objectos Simbólicos RLVT/F/15-24 e RLVT/F/25-34

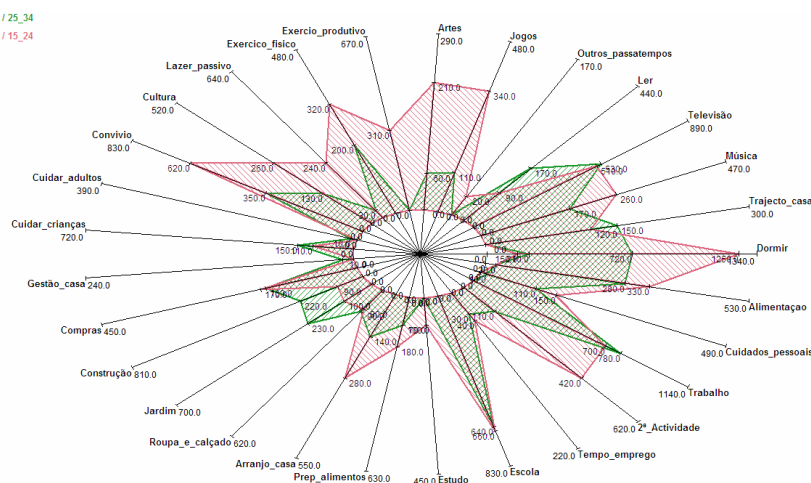


Figura 6.12: Estrela 2D *Superimpose* - Objectos Simbólicos RLVT/M/15-24 e RLVT/M/25-34

No Alentejo, para o sexo feminino (ver Figura 6.13) verifica-se que a 2ª Actividade e o tempo relacionado com o emprego que antes atingiam os seus máximos, não são nada relevantes nestes dois grupos etários. A *Escola* e o *Estudo* passam a ser as actividades mais importantes para o *Grupo Etário* dos 15 aos 24 anos e o *Arranjo da casa* a actividade mais importante a seguir ao *Trabalho*, no *Grupo Etário* dos 25 aos 34 anos. As actividades de lazer como *Televisão*, *Jogos* e leitura não têm muita relevância nestas classes etárias.

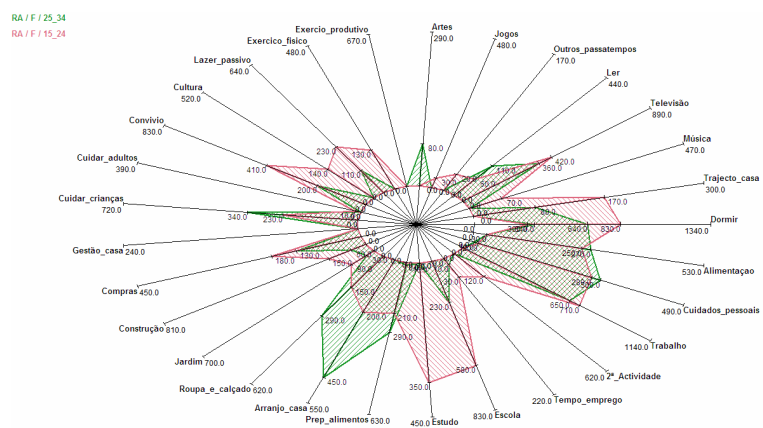


Figura 6.13: Estrela 2D *Superimpose* - Objectos Simbólicos RA/F/15-24 e RA/F/25-34

No sexo masculino para o *Grupo Etário* 15-24 (ver Figura 6.14), as actividades da *Construção*, *Escola*, *Exercício produtivo* e *Trabalho*, são as actividades mais importantes. Já no grupo dos 25 aos 34 anos temos como actividades mais importantes, para além de *Dormir*, os *Jogos*, *Trabalho* e *Exercício produtivo*.

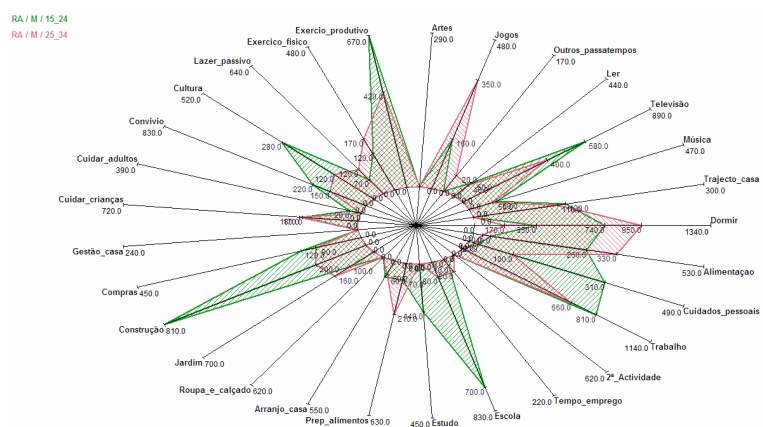


Figura 6.14: Estrela 2D *Superimpose* - Objectos Simbólicos RA/M/15-24 e RA/M/25-34

Na Figura 6.15 apresentam-se as médias calculadas como descrito na Secção 4.1.1 (ver Definição 4.7), desvio-padrão (deriva da Definição 4.8) e correlações (ver Definição 4.11) entre as variáveis intervalares correspondentes à agregação *Região/Sexo*.

NUMERIC CHARACTERISTICS				
	mean	std dev.	best correlated	
Dormir	696.50	66.22	Estudo	(0.504)
Alimentação	194.00	32.39	Escola	(0.673)
Cuidados_pessoais	156.50	34.64	Outros_passatempos	(0.624)
Trabalho	455.50	45.80	Compras	(0.776)
2ª_Actividade	205.00	53.62	Tempo_emprego	(0.544)
Tempo_emprego	47.00	30.59	Trabalho	(0.561)
Escola	346.50	43.88	Trabalho	(0.737)
Estudo	101.50	66.64	Lazer_passivo	(0.785)
Prep_alimentos	193.50	55.45	Roupa_e_calçado	(0.758)
Arranjo_casa	212.50	42.44	Roupa_e_calçado	(0.750)
Roupa_e_calçado	187.50	104.84	Cuidar_crianças	(0.804)
Jardim	262.00	47.97	Exercício_produtivo	(0.743)
Construção	187.00	102.99	Exercício_produtivo	(0.729)
Compras	159.50	35.81	Trabalho	(0.776)
Gestão_casa	68.50	32.79	Cuidar_crianças	(0.748)
Cuidar_crianças	188.50	89.47	Roupa_e_calçado	(0.804)
Cuidar_adultos	124.00	47.53	Televisão	(0.535)
Convívio	283.00	55.33	Jogos	(0.719)
Cultura	117.00	55.33	Jogos	(0.643)
Lazer_passivo	241.50	45.77	Estudo	(0.785)
Exercício_físico	165.00	48.17	Convívio	(0.612)
Exercício_produtivo	188.00	125.60	Jardim	(0.743)
Artes	107.50	24.52	Cultura	(0.634)
Jogos	147.50	55.42	Convívio	(0.719)
Outros_passatempos	34.00	28.88	Cultura	(0.627)
Ler	158.50	50.10	Cultura	(0.519)
Televisão	337.00	52.83	Cuidar_adultos	(0.535)
Música	135.50	53.73	Trajecto_casa	(0.659)
Trajecto_casa	100.50	27.97	Música	(0.659)

Figura 6.15: Estatísticas básicas das variáveis intervalares - *Região/Sexo*

Lembramos que a média de uma variável intervalar é definida como a média aritmética dos centros dos intervalos observados, isto é:

$$\bar{Y} = \frac{1}{k} \sum_{u \in E} \frac{y_u + \bar{y}_u}{2}.$$

Tomando como exemplo a variável $\mathcal{A}$ *Actividade* (ver Figura 6.16), a média calcula-se:

$$\bar{Y}_{\mathcal{A} \text{ Actividade}} = (1/10)[(0 + 620)/2 + (0 + 500)/2 + (0 + 270)/2 + (0 + 300)/2 + (0 + 300)/2 + (0 + 470)/2 + (0 + 310)/2 + (0 + 420)/2 + (0 + 480)/2 + (0 + 430)/2] = 205$$

O desvio-padrão de uma variável intervalar é dado por:

$$s_y = \sqrt{\frac{1}{3k} \sum_{u \in E} (\bar{y}_u^2 + y_u \bar{y}_u + y_u^2) - \frac{1}{4k^2} [\sum_{u \in E} (y_u + \bar{y}_u)]^2}.$$

A actividade *Dormir* é a que tem a média mais elevada, 696 minutos (cerca de 11 horas), seguida do *Trabalho* com 455 minutos. A *Escola*, *Televisão*, o *Jardim*, o *Convívio* e o *Lazer passivo* também têm médias elevadas. As variáveis com médias mais baixas são *Outros passatempos*, *Tempo-emprego* e *Gestão da casa*. O desvio-padrão mais elevado diz respeito à actividade *Exercício produtivo*, seguida da *Roupa e Calçado* e *Construção*. São as variáveis que têm maior variabilidade tendo como base

	2ª_Actividade
RA / F	[0.00 : 620.00]
RALG / M	[0.00 : 500.00]
RALG / F	[0.00 : 270.00]
RA / M	[0.00 : 300.00]
RC / F	[0.00 : 300.00]
RC / M	[0.00 : 470.00]
RLVT / F	[0.00 : 310.00]
RLVT / M	[0.00 : 420.00]
RN / M	[0.00 : 480.00]
RN / F	[0.00 : 430.00]

Figura 6.16: Variável intervalar *Segunda Actividade* - *Região/Sexo*

a *Região* e o *Sexo*, provavelmente porque o sexo é uma variável que influencia bastante os resultados obtidos nestas variáveis.

As variáveis mais correlacionadas são *Roupa e Calçado* com *Cuidar de crianças*, *Estudo* com *Lazer passivo*, e *Trabalho* com *Compras*. Das variáveis com correlação mais elevada, as mais baixas são, *Dormir* com *Estudo* e *Cuidar de adultos* com *Televisão*.

Comparando estes resultados com as Estatísticas calculadas através dos dados originais (ver Figura 6.17) verificam-se algumas diferenças, como por exemplo, a *Escola*, que tem uma média de 21 minutos com os dados originais, mas quando se consideram os dados simbólicos passa a ter uma média de 346 minutos, a *2ª Actividade* que agora tem um média de 3,99 minutos e com os dados simbólicos tem uma média de 205 minutos, os *Jogos* que com os dados clássicos tem uma média de 5,4 minutos e com os dados simbólicos tem uma média de 147 minutos. Se observarmos as actividades mais comuns em todos os grupos, as diferenças já não são tão acentuadas, como na actividade *Dormir* e *Alimentação*.

Estas diferenças devem-se ao facto de se ter assumido uma distribuição uniforme em cada intervalo, o que em muitos casos não se verifica. Uma forma de ultrapassar este problema consiste em generalizar as variáveis clássicas numéricas, não para variáveis intervalares, mas como variáveis histograma, de forma a ter em conta a distribuição dos valores observados.

Um modo alternativo à medida dos pontos médios para resumir as observações de uma variável intervalar, será considerar um intervalo “médio” cujos limites são respectivamente a média dos limites inferiores e superiores dos intervalos observados. Para o caso obter-se-iam os valores indicados na Figura 6.18.

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
Dormir	5.638	0	1.340	531,82	127,715
Alimentação	5.638	0	530	116,76	48,531
Cuidados_pessoais	5.638	0	490	50,64	34,477
Trabalho	5.638	0	1.140	177,29	236,267
2_Actividade	5.638	0	620	3,99	32,543
Tempo_emprego	5.638	0	220	0,72	6,217
Escola	5.638	0	830	21,11	93,829
Estudo	5.638	0	450	0,67	11,715
Prep_alimentos	5.638	0	630	68,15	80,915
Arranjo casa	5.638	0	550	46,09	67,270
Compras	5.638	0	450	20,04	41,796
Gestão casa	5.638	0	240	0,88	9,121
Cuidar crianças	5.638	0	720	11,34	41,945
Cuidar adultos	5.638	0	390	3,32	21,061
Convívio	5.638	0	830	46,87	79,140
Cultura	5.638	0	520	2,42	19,418
Lazer passivo	5.638	0	640	18,50	54,376
Exercício físico	5.638	0	480	11,38	38,926
Exercício produtivo	5.638	0	670	2,30	29,869
Artes	5.638	0	290	0,98	12,354
Jogos	5.638	0	480	5,04	27,905
Outros passatempos	5.638	0	170	0,25	4,382
Ler	5.638	0	440	9,44	32,880
Televisão	5.638	0	890	120,03	113,474
Música	5.638	0	470	4,34	23,564
Trajecto casa	5.638	0	300	13,74	23,561
Roupa e calçado	5.638	0	620	29,78	64,059
Jardim	5.638	0	700	24,50	66,331
Construção	5.638	0	810	3,81	30,263

Figura 6.17: Estatísticas descritivas clássicas das variáveis intervalares - *Região/Sexo*

Comparação das Médias			
	Média Clássica	Média Simbólica	Média Simbólica Alternativa
Dormir	531,82	696	[148;1245]
Alimentação	116,76	194	[9;379]
Cuidados_pessoais	50,64	156	[2;311]
Trabalho	177,29	455	[0;911]
2_Actividade	3,99	205	[0;410]
Tempo_emprego	0,72	47	[0;94]
Escola	21,11	346	[0;693]
Estudo	0,67	101	[0;203]
Prep_alimentos	68,15	193	[0;387]
Arranjo casa	46,09	212	[0;425]
Compras	20,04	159	[0;319]
Gestão casa	0,88	68	[0;137]
Cuidar crianças	11,34	188	[0;377]
Cuidar adultos	3,32	124	[0;248]
Convívio	46,87	283	[0;566]
Cultura	2,42	117	[0;234]
Lazer passivo	18,50	241	[0;483]
Exercício físico	11,38	165	[0;330]
Exercício produtivo	2,30	188	[0;376]
Artes	0,98	107	[0;215]
Jogos	5,04	147	[0;295]
Outros passatempos	0,25	34	[0;68]
Ler	9,44	158	[0;317]
Televisão	120,03	337	[0;674]
Música	4,34	135	[0;271]
Trajecto casa	13,74	100	[0;201]
Roupa e calçado	29,78	187	[0;375]
Jardim	24,50	262	[0;524]
Construção	3,81	187	[0;374]

Figura 6.18: Comparação das Médias das variáveis intervalares - *Região/Sexo*

O bi-plot na Figura 6.19 representa as coordenadas dos objectos simbólicos para as variáveis *Roupa e Calçado* e *Cuidar de crianças*. As mulheres da região Centro e região Norte são as que ocupam mais tempo com estas tarefas. Na região do Algarve as mulheres dedicam muito menos tempo a estas actividades do que as restantes, principalmente na *Roupa e Calçado*. É de salientar que os homens da região de Lisboa e Vale do Tejo também dedicam grande parte do seu tempo com *Roupa e Calçado*. Os homens da região do Algarve são os indivíduos que menos tempo ocupam com estas duas actividades.

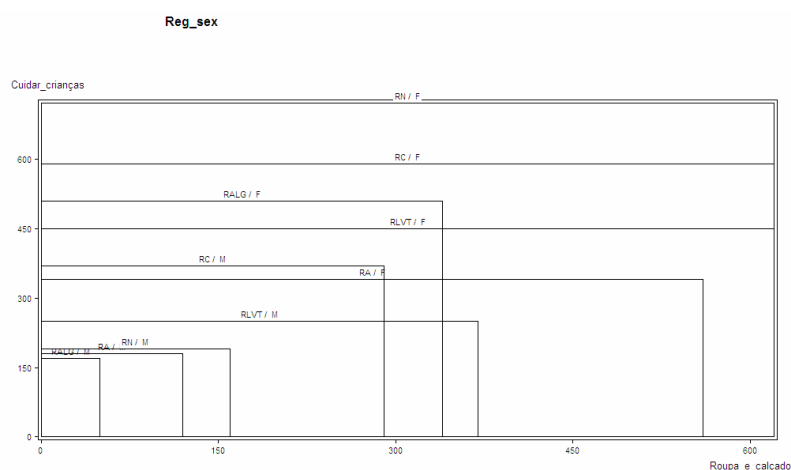


Figura 6.19: Bi-plot das variáveis intervalares *Roupa e Calçado* e *Cuidar de crianças - Região/Sexo*

Na Figura 6.20 temos as coordenadas dos objectos para as actividades de *Estudo* e *Lazer passivo*. Tanto os homens como as mulheres na região Centro são os que gastam mais tempo com estas actividades comparando com o resto dos grupos. No entanto, as mulheres e os homens da região Norte também ocupam algum tempo com o *Lazer passivo*. No *Estudo*, para além da região Centro, o sexo feminino no Alentejo também dedica algum tempo. Por sua vez as mulheres da região do Algarve são as que menos tempo ocupam com o *Estudo* e o *Lazer passivo*.

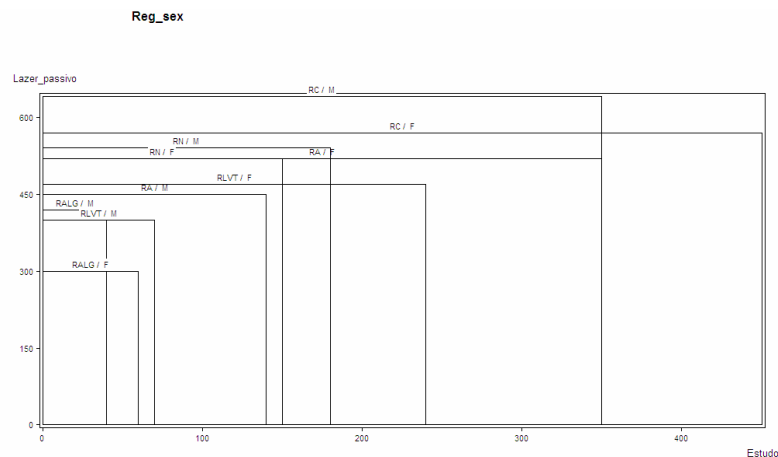


Figura 6.20: Bi-plot das variáveis intervalares *Lazer passivo* e *Estudo - Região/Sexo*

Analisando as médias para o cruzamento *Região/Sexo/Grupo Etário*, na Figura 6.21, verifica-se que as actividades de *Dormir* e *Trabalho* continuam com as médias mais elevadas.

NUMERIC CHARACTERISTICS				
	mean	std dev.	best correlated	
Dormir	600.40	109.33	Lazer_passivo	(0.644)
Alimentação	163.70	34.52	Escola	(0.406)
Cuidados_pessoais	106.00	37.96	Outros_passatempos	(0.466)
Trabalho	374.60	60.46	Tempo_emprego	(0.474)
2ª_Actividade	96.70	87.76	Trajecto_casa	(0.521)
Tempo_emprego	19.80	21.63	Trabalho	(0.474)
Escola	135.20	138.36	Trajecto_casa	(0.416)
Estudo	27.10	50.50	Escola	(0.402)
Prep_alimentos	134.30	60.11	Roupa_e_calçado	(0.720)
Arranjo_casa	136.00	67.91	Prep_alimentos	(0.653)
Roupa_e_calçado	109.00	88.81	Prep_alimentos	(0.720)
Jardim	153.50	89.33	Dormir	(0.564)
construção	98.50	88.92	Exercio_produtivo	(0.541)
Compras	110.60	42.29	Arranjo_casa	(0.473)
Gestão_casa	23.40	32.12	Televisão	(0.463)
Cuidar_crianças	107.00	80.21	Arranjo_casa	(0.410)
Cuidar_adultos	55.40	54.02	Prep_alimentos	(0.600)
Convívio	199.80	68.23	Música	(0.347)
Cultura	64.10	50.56	Jogos	(0.545)
Lazer_passivo	116.10	82.14	Dormir	(0.644)
Exercico_fisico	102.30	51.56	Lazer_passivo	(0.467)
Exercio_produtivo	80.00	104.39	construção	(0.541)
Artes	38.80	44.23	Outros_passatempos	(0.410)
Jogos	80.40	61.63	Cultura	(0.545)
Outros_passatempos	10.50	18.93	Cuidados_pessoais	(0.466)
Ler	88.00	53.08	Lazer_passivo	(0.480)
Televisão	251.90	66.57	Cuidar_adultos	(0.571)
Música	69.40	50.81	Dormir	(0.447)
Trajecto_casa	61.80	29.76	2ª_Actividade	(0.521)

Figura 6.21: Estatísticas básicas das variáveis intervalares - *Região/Sexo/Grupo Etário*

A *Escola*, *Jardim* e *Lazer passivo* que antes tinham médias também elevadas passam agora para actividades com médias pouco significativas. A *Televisão* que antes não tinha demasiada importância, agora tem uma média bastante elevada, 250 minutos, cerca de 4 horas. Num nível de cruzamento mais “fino”, variáveis que produzam intervalos com menor amplitude ficam com as médias mais reduzidas. Por exemplo, a

Escola é característica de grupos com idades mais jovens, logo só nestas faixas etárias irá produzir intervalos com maior amplitude, do que resulta no geral uma média menor.

Com a introdução do *Grupo Etário*, relativamente à variabilidade, temos três variáveis de destaque, *Dormir*, *Escola* e *Exercício produtivo*, enquanto que com o cruzamento anterior tínhamos *Roupa e Calçado*, *Construção* e *Exercício produtivo*. Com a nova variável de destaque introduzida no cruzamento, temos uma maior variabilidade nas variáveis *Dormir*, *Escola* e *Exercício produtivo*, o que significa que a idade produz grandes variações na ocupação do tempo com estas actividades.

As variáveis que têm maior correlação são *Dormir* com *Lazer passivo*, *Preparação de alimentos* com *Arranjo da casa*, e *Preparação de alimentos* com *Roupa e Calçado*. Se observarmos o Bi-plot da Figura 6.22 relativamente às variáveis *Preparação de alimentos* com *Arranjo da casa*, verifica-se que as mulheres de Lisboa e Vale do Tejo na faixa etária entre os 24 e os 35 anos são as que mais tempo dedicam ao *Arranjo da casa* e que as mulheres da região Norte da mesma faixa etária são as que dedicam mais tempo à preparação dos alimentos. É de salientar que os homens da região Norte entre os 25 e os 34 anos ocupam algum tempo em *Arranjo da casa*.

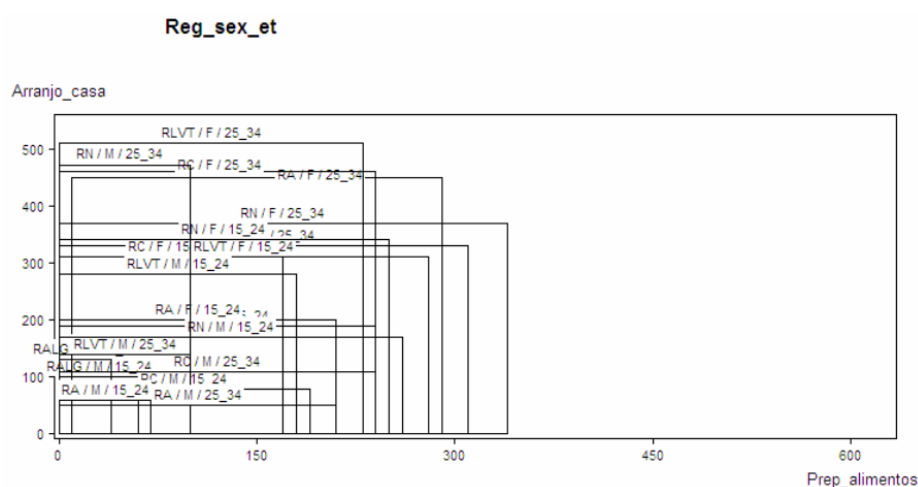


Figura 6.22: Bi-plot das variáveis intervalares *Arranjo-casa* e *Prep-alimentos* - *Região/Sexo/Grupo Etário*

Para as variáveis *Dormir* e *Lazer passivo*, temos a região Centro do sexo masculino com idades entre os 15 aos 24 anos, o principal responsável pela elevada correlação entre estas duas variáveis. Observa-se também que na região do Algarve os homens entre 25 e os 34 anos também ocupam bastante tempo em *Dormir* e *Lazer passivo*, principalmente em *Lazer passivo*. Na variável *Dormir* as mulheres da região do Alentejo entre os 15 e os 24 anos também têm grande importância.

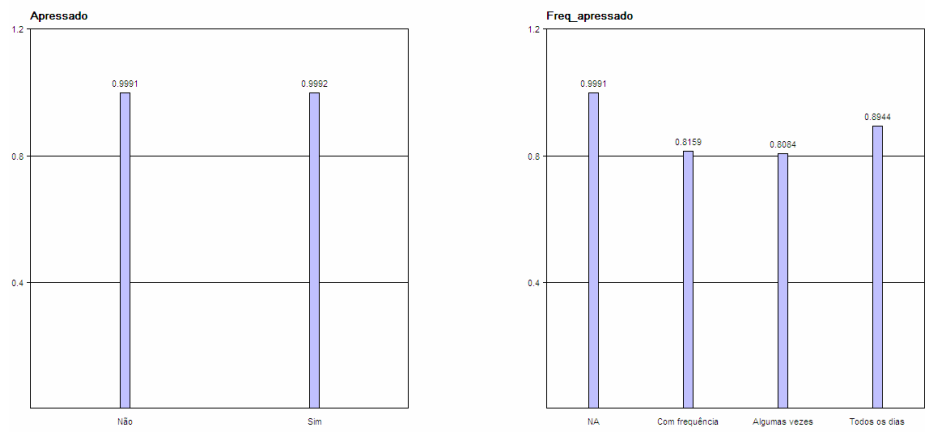


Figura 6.24: Capacidades das variáveis *Apressado* e *Freq-apressado* - *Região/Sexo*

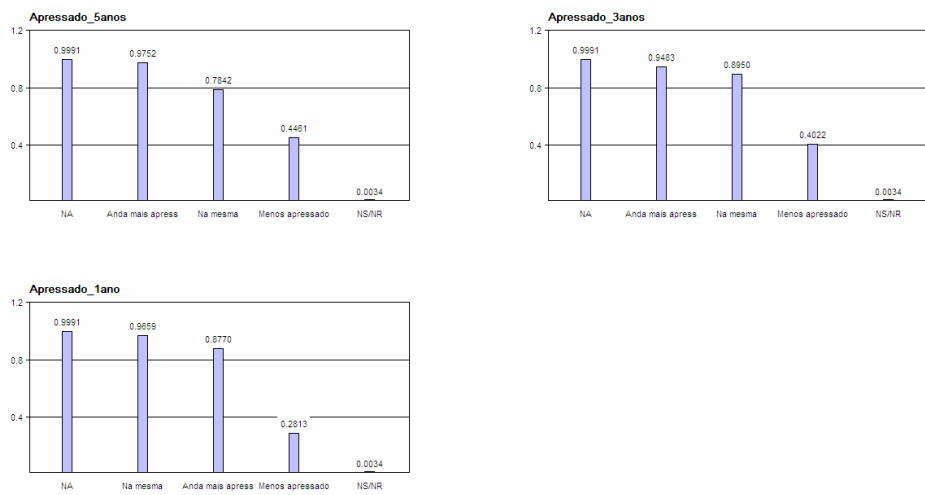


Figura 6.25: Capacidades das variáveis *Apressado-5anos*, *Apressado-3anos*, *Apressado-1ano* - *Região/Sexo*

Quanto ao tempo disponível (exceptuando a categoria *NA*), temos na vida particular e no fim-de-semana a maior capacidade em *Sim, algumas vezes* (ver Figura 6.26). Quanto ao tempo disponível sem saber o que fazer com ele, capacidade mais elevada observa-se em *Nunca*.

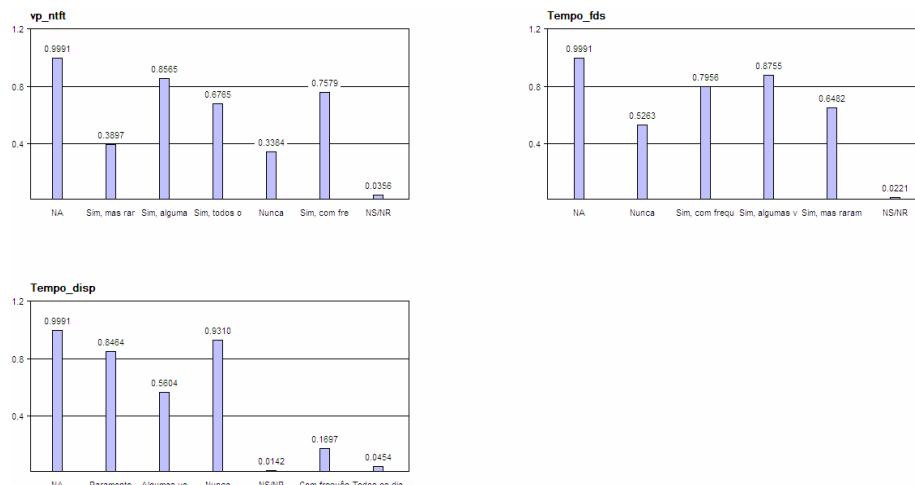


Figura 6.26: Capacidades das variáveis *vp-ntft*, *tem-fds* e *tempo-disp* - *Região/Sexo*

Na Figura 6.27 apresentam-se os mínimos, máximos e médias para a variável *Apressado* que são praticamente iguais, o que significa que não há grandes diferenças entre os objectos simbólicos para esta variável. Verifica-se no entanto que para a resposta *Sim* temos valores mais baixos. Para a variável *Freq-apressado*, ou seja, frequência com que o indivíduo anda apressado, já se observa uma média mais elevada para a resposta *Todos os dias*.

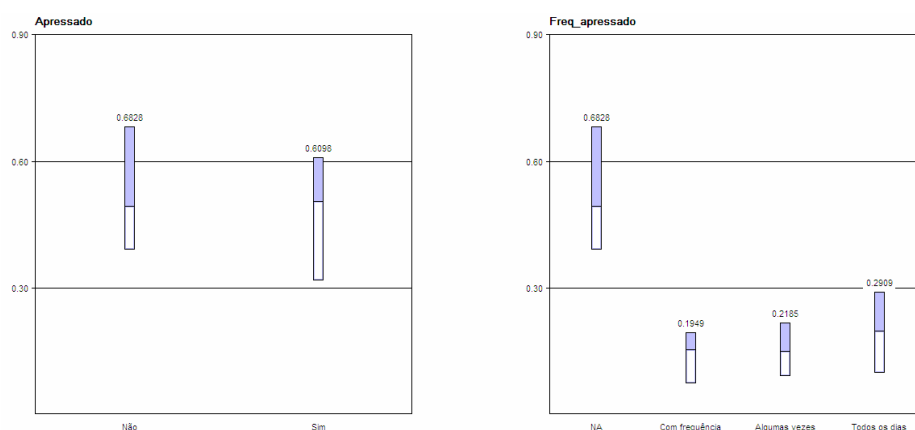


Figura 6.27: Min/Max/Med variáveis *Apressado* e *Freq-apressado* - *Região/Sexo*

O cenário para as variáveis que medem o que o indivíduo sente relativamente ao passado, é semelhante para 5 e 3 anos atrás. Obtêm-se médias, como se pode observar na Figura 6.28, entre os 50% e os 60% para a resposta *Anda mais apressado*, o que leva a concluir que a maioria dos indivíduos se sente mais apressado do que há 5 anos atrás. Se tivermos em conta a situação face há 1 ano, podemos concluir que a maioria dos indivíduos que se dizem apressados se sente *Na mesma*.

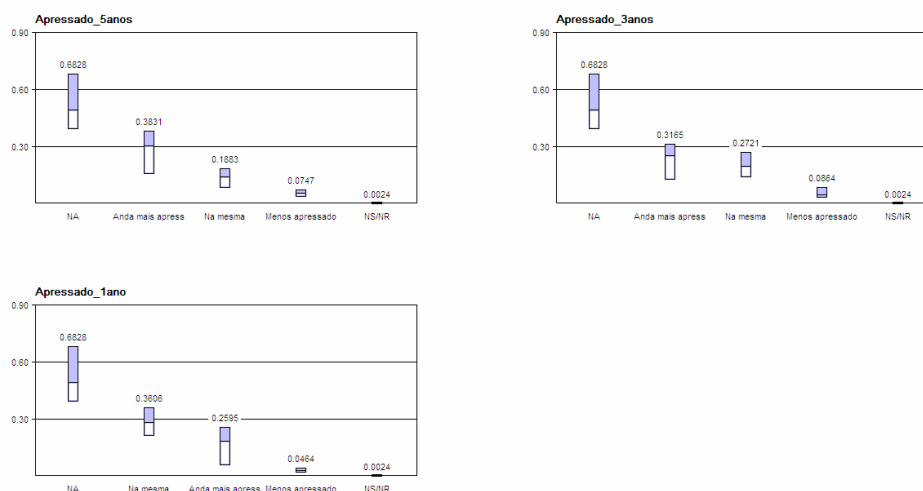


Figura 6.28: Min/Max/Med variáveis *Apressado-5anos*, *Apressado-3anos* e *Apressado-1ano* - *Região/Sexo*

Quanto ao tempo disponível para realizar todas as tarefas na vida particular (ver Figura 6.29) e ao tempo disponível no fim-de-semana, a média mais elevada está nas respostas que indicam que existe algum tempo disponível. Quanto ao tempo disponível sem saber os que fazer com ele, a média mais elevada está na categoria *Nunca*.

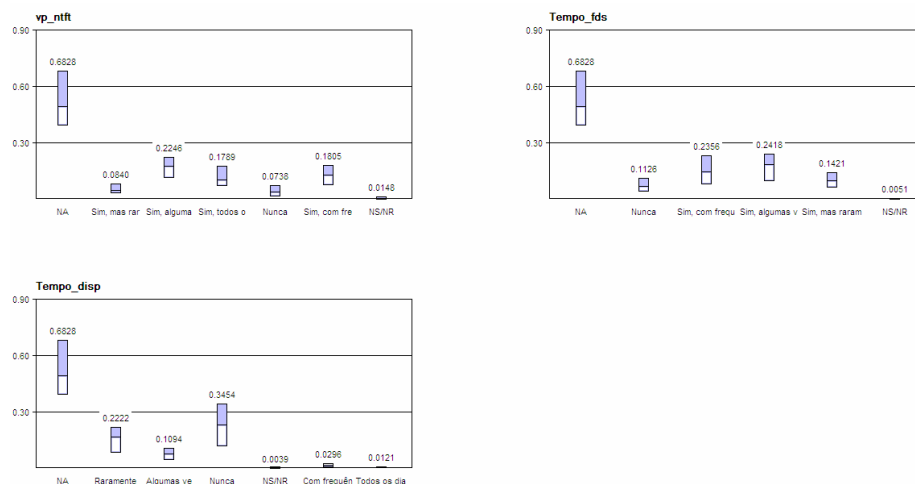


Figura 6.29: Min/Max/Med variáveis *vp-ntft*, *tem-fds*, *tempo-disp* - *Região/Sexo*

Descendo ao segundo nível, com o cruzamento *Região/Sexo/Grupo Etário*, se analisarmos as capacidades das variáveis *Apressado* e *Freq-apressado* (ver Figura 6.30) verificamos que todas as categorias têm capacidades próximas de 100%, ou seja, todas as respostas têm grande probabilidade de se observar. O mesmo se pode concluir relativamente ao passado (ver Figura 6.31), pois todas as categorias têm capacidades próximas de 100% excepto as respostas *Menos Apressado*. Não há uma redução do sentimento de pressa face ao passado.

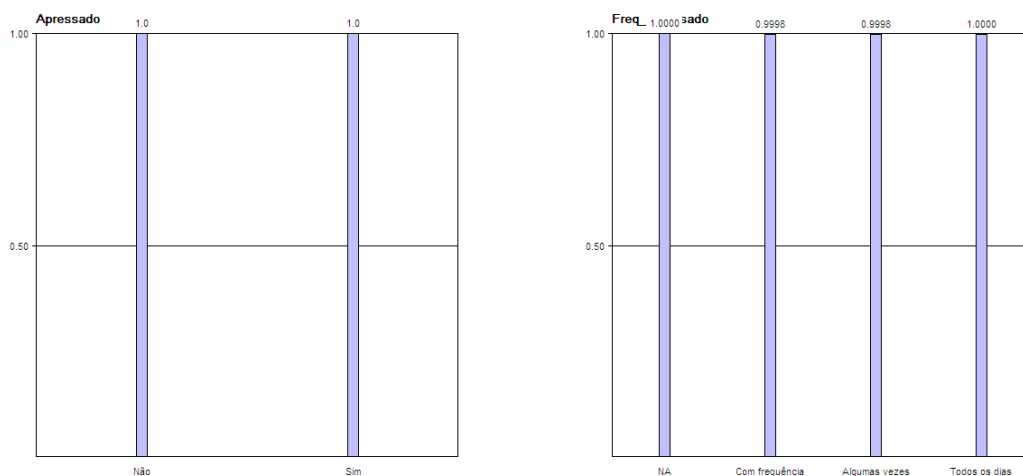


Figura 6.30: Capacidades das variáveis *Apressado* e *Freq-apressado* - *Região/Sexo/Grupo Etário*

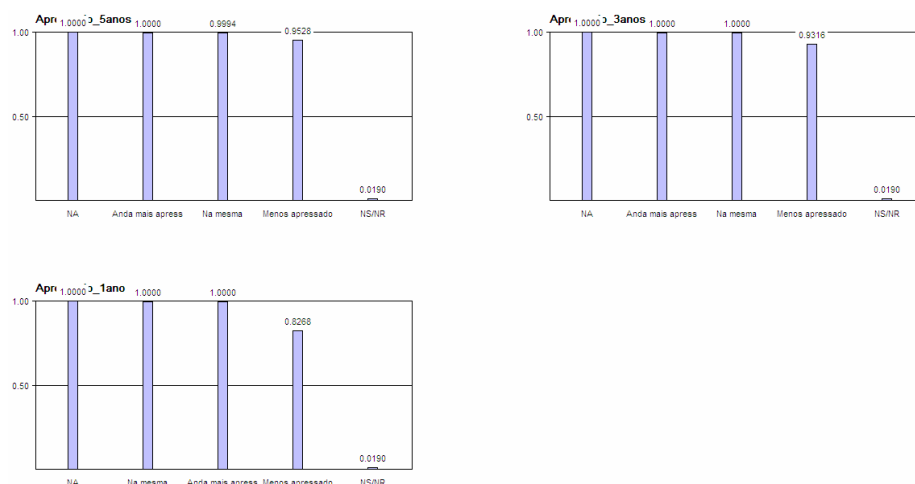


Figura 6.31: Capacidades das variáveis *Apressado-5anos*, *Apressado-3anos* e *Apressado-1ano* - Região/Sexo/Grupo Etário

Relativamente às restantes variáveis *vp-ntft*, *tem-fds* e *tempo-disp*, (ver Figura 6.32), que medem o sentimento de “pressa” face ao passado, há que salientar o facto de no tempo disponível sem saber o que fazer com ele as repostas *Com frequência* e *Todos os dias* são as que têm as capacidades mais baixas.

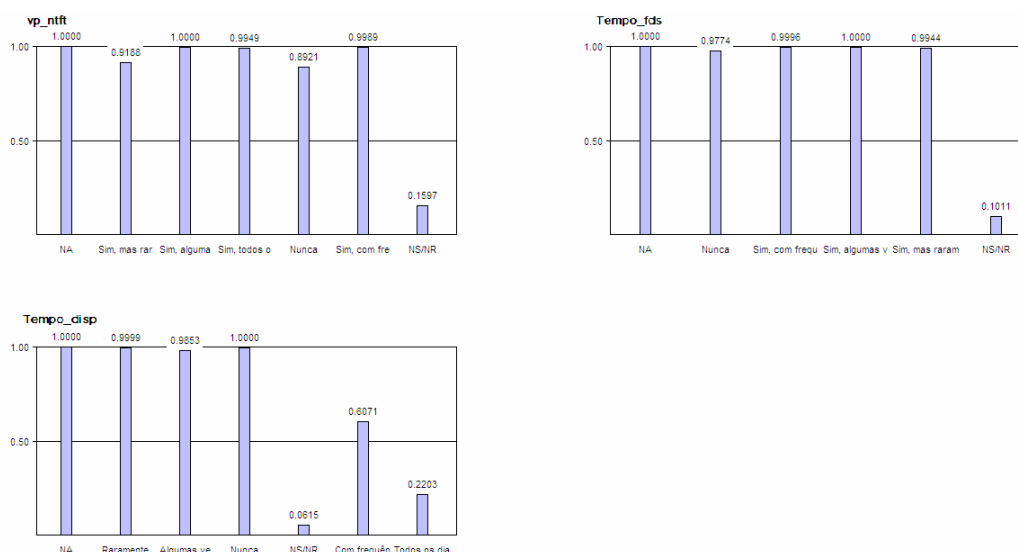


Figura 6.32: Capacidades das variáveis *vp-ntft*, *tem-fds*, *tempo-disp* - Região/Sexo/Grupo Etário

Analisando as médias, mínimos e máximos para as variáveis *Apressado* e *Freq-apressado* (ver Figura 6.33) verificamos que o máximo atingido para a resposta *Não* é superior ao máximo atingido para a resposta *Sim*, o que indica que é na categoria *Não* que o máximo das frequências é atingido. É de referir que dos que sentem apressados, o

máximo mais elevado e a média mais elevada, relativamente à frequência com que se sentem apressados está na categoria *Todos os dias*. Comparando com o cruzamento por *Região* e *Sexo* (ver Figura 6.27) verificamos que existem diferenças mínimas relativamente à média, que no primeiro nível de agregação era mais elevada para a categoria *Sim*, e no segundo nível é um pouco mais elevada na categorias *Não*. Também se verifica que os máximos e mínimos tendem a ser mais extremos no segundo cruzamento.

Se observarmos o sentimento relativamente ao passado (ver Figura 6.34) o comportamento destas variáveis é muito semelhante ao do primeiro cruzamento. As médias são praticamente iguais, apenas os máximos e mínimos são mais extremos, como já acontecia anteriormente.

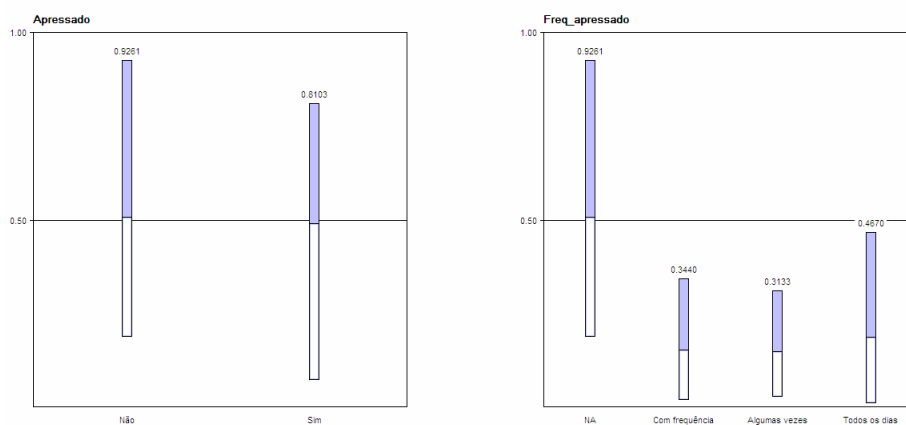


Figura 6.33: Min/Max/Med variáveis *Apressado* e *Freq-apressado* - *Região*/*Sexo*/*Grupo Etário*

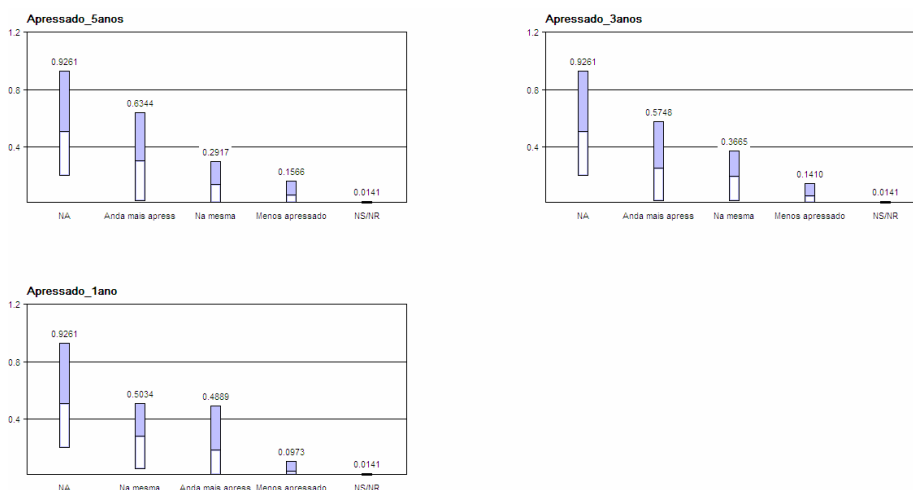


Figura 6.34: Min/Max/Med variáveis *Apressado-5anos*, *Apressado-3anos* e *Apressado-1ano* - *Região*/*Sexo*/*Grupo Etário*

Para o tempo disponível também não se verificam diferenças quanto ao comportamento das variáveis face ao primeiro nível, a não ser a variação dos máximos e mínimos.

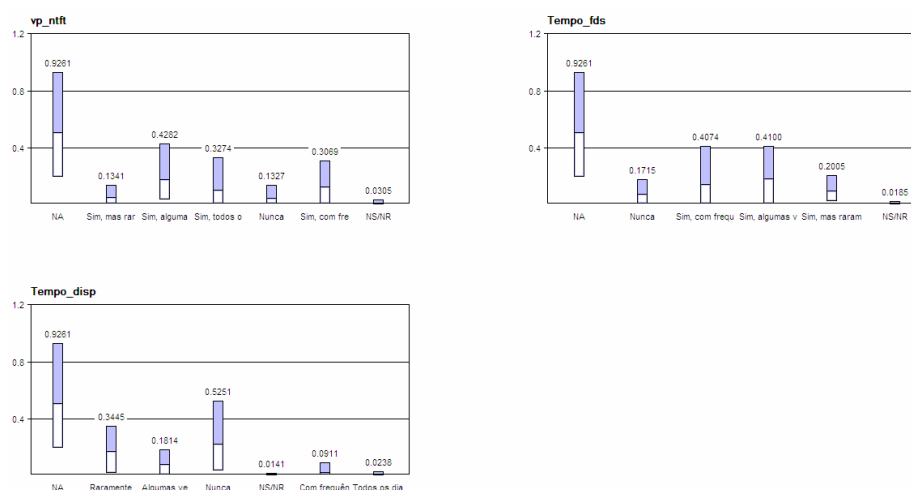


Figura 6.35: Min/Max/Med variáveis *vp-ntft*, *tem-fds*, *tempo-disp* - *Região/Sexo/Grupo Etário*

Para analisar as diferenças quanto ao sentimento de pressa através da visualização dos objectos simbólicos, com o cruzamento *Região/Sexo*, optou-se por ignorar a categoria *NA* nas variáveis dependentes. A estrela com sobreposição permite representar simultaneamente as descrições de vários grupos e assim comparar as distribuições de frequências das diferentes variáveis nos grupos considerados.

Para o sexo feminino (ver Figura 6.36), as regiões do Norte, Centro e Lisboa e Vale do Tejo, têm praticamente um comportamento igual, verificando-se um sentimento geral de pressa. Relativamente ao passado, observa-se que as mulheres que se sentem apressadas, comparando com a situação há 5 anos atrás, sentem-se mais apressadas na sua maioria, mas relativamente há um 1 ano atrás, a maioria sente-se na mesma.

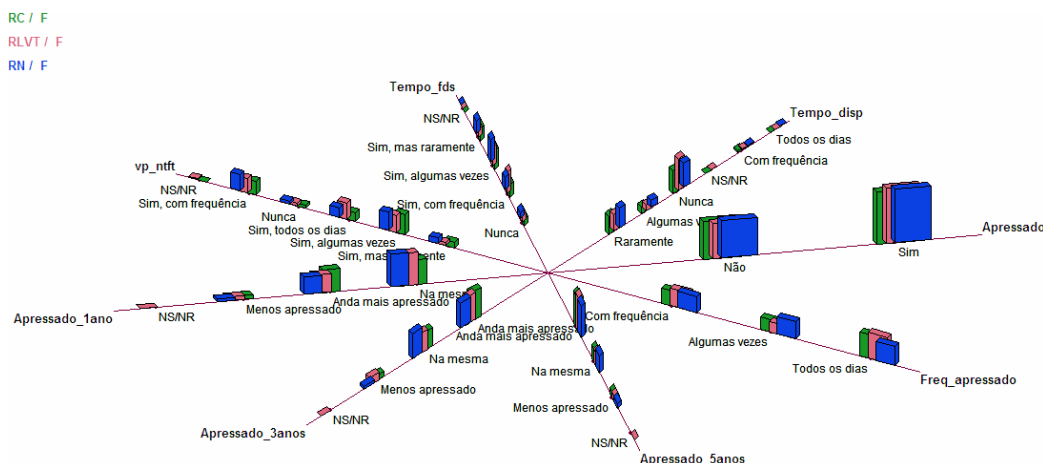


Figura 6.36: Estrela 3D *Superimpose* - Objectos Simbólicos RN/F, RC/F, RLVT/F

Por outro lado as regiões mais a Sul do país (ver Figura 6.37), Alentejo e Algarve, já têm o seu máximo para a variável *Apressado* na resposta *Não*. O comportamento relativamente ao passado, das mulheres que se sentem apressadas, é semelhante às regiões mais a Norte do país.

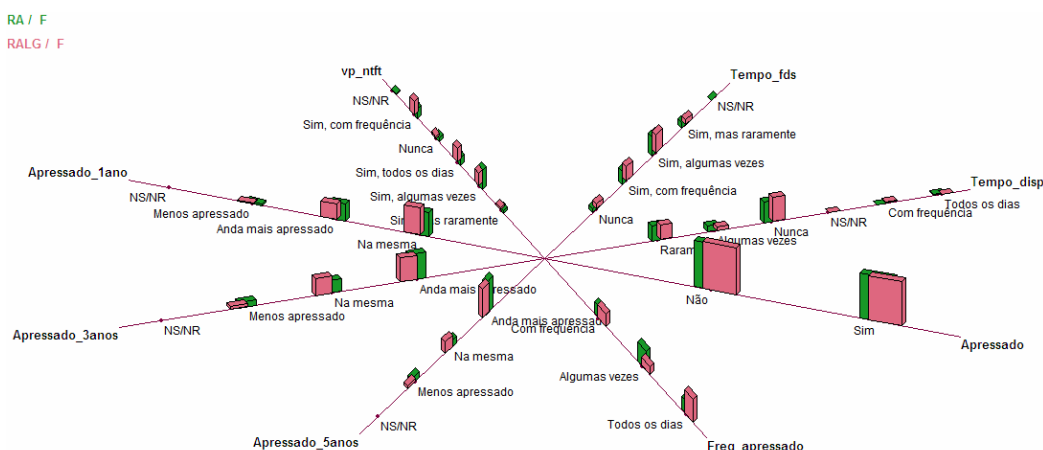


Figura 6.37: Estrela 3D *Superimpose* - Objectos Simbólicos RA/F, RALG/F

Para os homens, o comportamento é semelhante ao das mulheres, existindo a mesma divisão Sul/Norte, como se verifica através das Figuras 6.38, 6.39. Na região Norte temos um comportamento igual, para o sexo feminino e masculino, mas comparando a frequência de homens apressados com as mulheres apressadas, constatamos que o sexo feminino apresenta valores mais elevados. Nas regiões do Alentejo e Algarve, já se verifica uma diferença mais acentuada entre os que sentem pressa e os que não sentem. De uma forma geral pode dizer-se que as regiões do Sul para o sexo masculino, é onde menos se regista um sentimento de pressa.

Para o nível da *Região/Sexo/Grupo Etário* (ver Figura 6.40) verifica-se que na faixa etária dos 15 aos 24 anos a maioria sente-se apressado e mais apressado relativamente ao passado. A região de Lisboa e Vale do Tejo é a que tem valores mais elevados para a resposta *Sim* na variável *Apressado*. Regista-se novamente uma tendência para na região Norte do país se sentir mais pressa. Para o sexo masculino, na mesma faixa etária, já temos uma maioria de respostas *Não* na variável *Apressado*, excepto na região Norte.

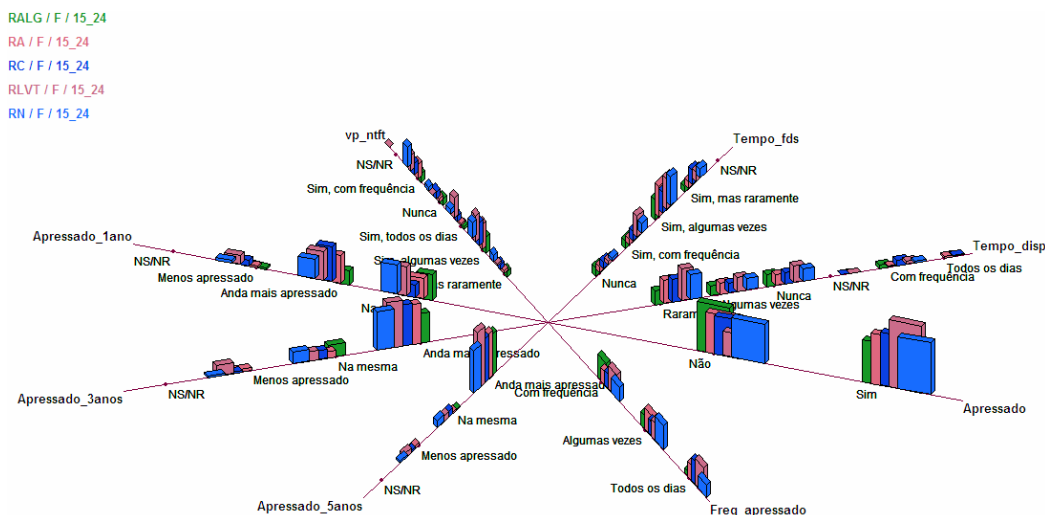


Figura 6.40: Estrela 3D *Superimpose* - Objectos Simbólicos - RALF/F/15-24 RA/F/15-24 RLVT/F/15-24 RC/F/15-24 RN/F/15-24

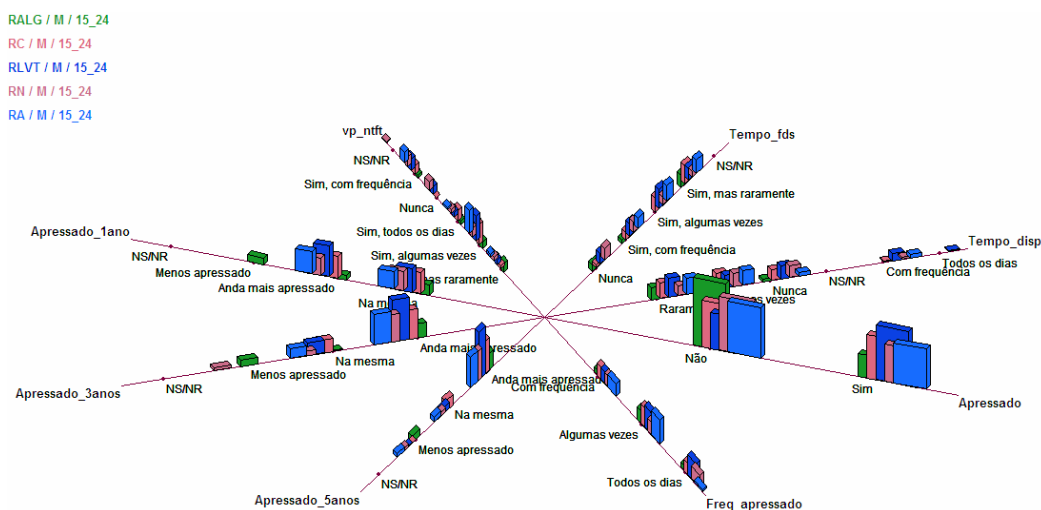


Figura 6.41: Estrela 3D *Superimpose* - Objectos Simbólicos - RALF/M/15-24 RA/M/15-24 RLVT/M/15-24 RC/M/15-24 RN/M/15-24

Para o *Grupo Etário* dos 25 aos 34 anos (ver Figura 6.42) e para todas as regiões, as mulheres sentem-se na maioria apressadas e na mesma ou mais relativamente a passado. Para o sexo masculino (ver Figura 6.43) verifica-se o mesmo comportamento que para o sexo feminino, com excepção da região do Algarve, única região onde os homens entre os 25 e os 34 anos não se sentem na maioria apressados.

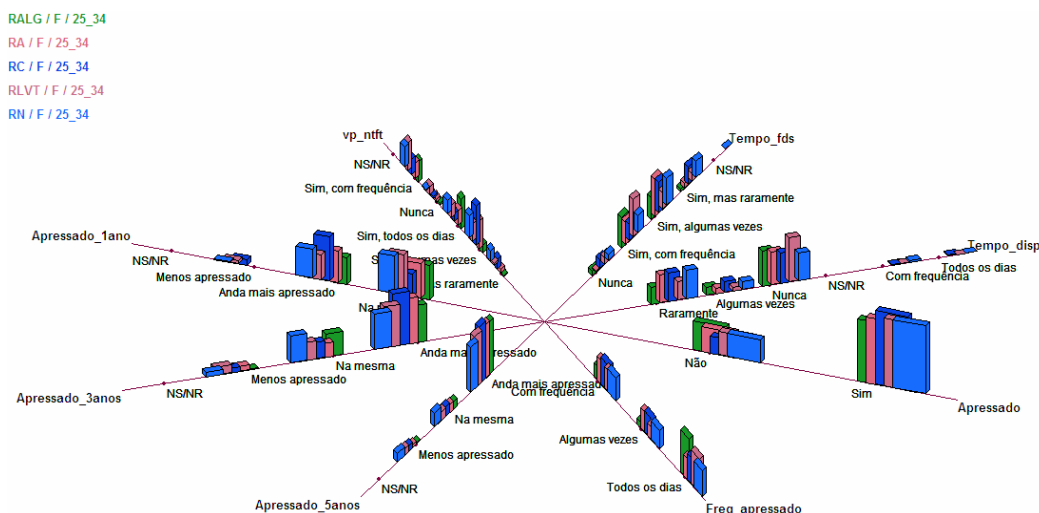


Figura 6.42: Estrela 3D *Superimpose* - Objectos Simbólicos - RALG/F/25-34 RA/F/25-34 RLVT/F/25-34 RC/F/25-34 RN/F/25-34

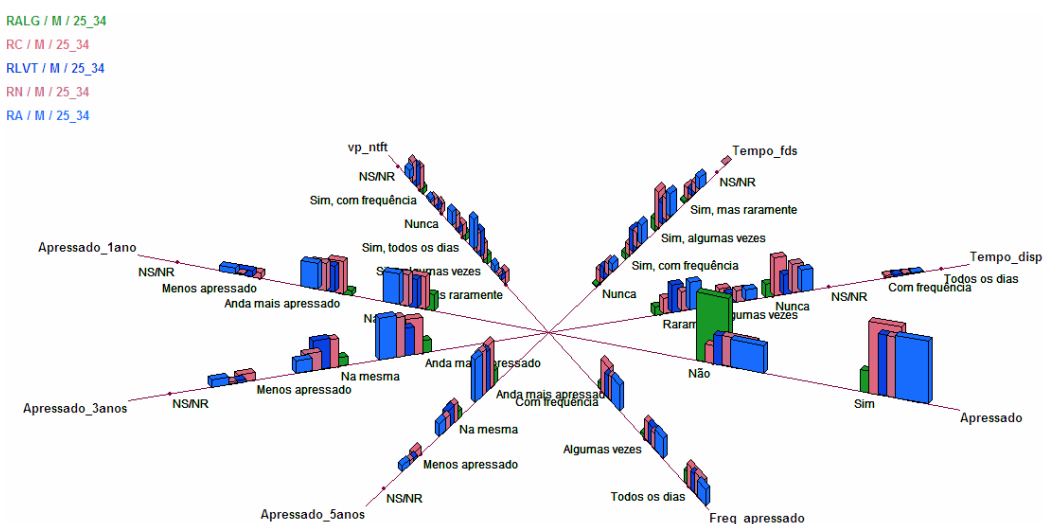


Figura 6.43: Estrela 3D *Superimpose* - Objectos Simbólicos - RALG/M/25-34 RA/M/25-34 RLVT/M/25-34 RC/M/25-34 RN/M/25-34

6.2 Análise Classificatória

Os métodos de Análise Classificatória serão aplicados sobre os dois níveis de agregação de forma salientar as diferenças que existem entre os dois níveis, apresentando-se um quadro resumo no final de cada análise, por nível de agregação e por tipo de variáveis.

6.2.1 Método Divisivo

A classificação através do método Divisivo (apresentado na Secção 4.2.1) tem como objectivo formar classes de elementos do conjunto a classificar E que revelem uma certa semelhança entre eles. O método parte de uma única classe, sobre a qual são efectuados sucessivos cortes, cada um com base numa variável e num valor de corte c , seguindo o critério da inércia intra classe mínima.

Aplicando o método divisivo para o cruzamento *Região/Sexo* (ver Figura 6.44) sobre as variáveis intervalares, obtém-se uma inércia explicada (ver Definição 4.14) de cerca de 52,94% para a partição em 3 classes. A variável mais importante para discriminar as regiões por sexo relativamente à ocupação do tempo são os *Jogos* com um valor de corte $c = 160$ minutos, seguida do *Exercício produtivo* com um valor de corte $c = 177,5$ minutos. A classe 1 agrupa 4 elementos e respondeu "sim" à questão binária "[Exercício-produtivo $\leq 177,5$]?". Os restantes 6 elementos subdivididos posteriormente, responderam "não" a esta questão e fazem parte do cluster 2 para uma partição em duas classes. Na segunda divisão a questão binária colocada foi "[Jogos ≤ 160]?" e resultou na formação das classes 2 e 3 para uma partição em três classes.

A primeira classe é formada pelas mulheres de todo o país excepto o Alentejo, associadas ao pouco tempo gasto em *Jogos* e em *Exercício produtivo*. Na segunda classe estão todas as regiões para o sexo masculino, que estão associados ao elevado tempo com os *Jogos*. Na terceira classe está o Alentejo com o sexo feminino associado ao *Exercício produtivo*, com mais de 177 minutos e ao pouco tempo dedicado aos *Jogos*.

THE CLUSTERING TREE:

- the number noted at each node indicates the order of the division
- Ng <-> yes and Nd <-> no

```

      +---- Classe 1 (Ng=4)
      !
    !----2- [Exercicio_produtivo <= 177.500000]
      !
      !      +---- Classe 3 (Nd=1)
      !
    !----1- [Jogos <= 160.000000]
      !
      +---- Classe 2 (Nd=5)

```

DESCRIPTION OF THE CLUSTERS:

```

Cluster 1:
  IF    2- [Exercicio_produtivo <= 177.500000] IS TRUE
  AND    1- [Jogos <= 160.000000] IS TRUE

Cluster 2:
  IF    1- [Jogos <= 160.000000] IS FALSE

Cluster 3:
  IF    2- [Exercicio_produtivo <= 177.500000] IS FALSE
  AND    1- [Jogos <= 160.000000] IS TRUE

```

PARTITION IN 3 CLUSTERS:
-----;

```

Cluster 1 (n=4): RALG/F  RC/F  RLVT/F  RN/F
Cluster 2 (n=5): RALG/M  RA/M  RC/M  RLVT/M  RN/M
Cluster 3 (n=1): RA/F
Explicated inertia: 52.940613

```

Figura 6.44: Classificação Divisiva - *Região/Sexo*, variáveis intervalares

Aplicando o mesmo método para 4 classes, para o cruzamento *Região/Sexo/Grupo Etário* (ver Figura 6.45), obtém-se uma inércia explicada de 42%. As variáveis mais importantes para a formação dos grupos são a *Escola*, *Dormir*, *Preparação Alimentos* e *Roupa e Calçado*. Com o primeiro nível de agregação as actividades mais importantes eram os *Jogos* e o *Exercício produtivo*.

```

THE CLUSTERING TREE:
-----
- the number noted at each node indicates the order of the division
- Ng <-> yes and Nd <-> no

      +---- Classe 1 (Ng=10)
      !
      !----2- [Dormir <= 587.500000]
      !
      !      +---- Classe 3 (Ng=8)
      !      !
      !      !----4- [Prep_alimentos <= 180.000000]
      !      !
      !      !      +---- Classe 5 (Nd=8)
      !      !
      !----1- [Escola <= 110.000000]
      !
      !      +---- Classe 2 (Ng=12)
      !      !
      !      !----3- [Roupa_e_calçado <= 52.500000]
      !      !
      !      +---- Classe 4 (Nd=12)

DESCRIPTION OF THE CLUSTERS:
-----
Cluster 1: IF      2- [Dormir <= 587.500000] IS TRUE
            AND      1- [Escola <= 110.000000] IS TRUE

Cluster 2: IF      3- [Roupa_e_calçado <= 52.500000] IS TRUE
            AND      1- [Escola <= 110.000000] IS FALSE

Cluster 3: IF      4- [Prep_alimentos <= 180.000000] IS TRUE
            AND      2- [Dormir <= 587.500000] IS FALSE
            AND      1- [Escola <= 110.000000] IS TRUE

Cluster 4: IF      3- [Roupa_e_calçado <= 52.500000] IS FALSE
            AND      1- [Escola <= 110.000000] IS FALSE

Cluster 5: IF      4- [Prep_alimentos <= 180.000000] IS FALSE
            AND      2- [Dormir <= 587.500000] IS FALSE
            AND      1- [Escola <= 110.000000] IS TRUE

PARTITION IN 5 CLUSTERS:
-----:

Cluster 1 (n=10): RA/F/55_64  RA/M/55_64  RALG/F/55_64  RALG/M/55_64  RC/F/35_54
                  RC/M/35_54  RALG/M/35_54  RALG/F/25_34  RN/F/25_34  RA/M/25_34

Cluster 2 (n=12): RALG/M/15_24  RA/M/35_54  RALG/M/25_34  RC/M/15_24  RC/M/25_34
                  RLVT/M/25_34  RLVT/M/35_54  RLVT/M/15_24  RN/M/35_54  RN/M/15_24
                  RN/M /25_34  RA/M/15_24

Cluster 3 (n=8):  RALG/M/65 ou mais  RA/M/65 ou mais  RC/M/65 ou mais  RC/M/55_64
                  RLV/M/65 ou mais  RN/M/65 ou mais  RN/M/55_64  RLVT/M/55_64

Cluster 4 (n=12): RA/F/35_54  RA/F/25_34  RA/F/15_24  RALG/F/35_54  RALG/F/15_24
                  RC/F/25_34  RC/F/15_24  RLVT/F/25_34  RLVT/F/35_54  RLVT/F/15_24
                  RN/F/35_54  RN/F/15_24

Cluster 5 (n=8):  RA/F/65 ou mais  RALG/F/65 ou mais  RC/F/65 ou mais  RLVT/F/65 ou mais
                  RN/F/55_64  RN/F/65 ou mais  RC/F/55_64  RLVT/F/55_64

Explicated inertia: 42.018304

```

Figura 6.45: Classificação Divisiva - *Região/Sexo/Grupo Etário*, variáveis intervalares

A classe 1 é formada por 10 objectos que representam, de um modo geral, as faixas etárias dos 25 aos 64 anos das regiões Centro, Algarve e Alentejo. É de salientar que a região de Lisboa e Vale do Tejo não está presente nesta classe. A região do Algarve está presente com 4 grupos e a região do Alentejo com 3. Quanto ao sexo não se verifica uma tendência nesta classe. Esta classe é caracterizada pelo tempo mais reduzido dedicado a *Dormir* e à *Escola*. A segunda classe formada representa, de um modo geral, o sexo masculino com idades entre os 15 e os 34 anos, para todas as regiões. Estes grupos dedicam menos tempo com *Roupa e Calçado* e mais tempo com a *Escola*. Na classe 3 também só estão presentes grupos do sexo masculino e caracterizam as idades mais avançadas, principalmente com 65 anos ou mais, em todas as regiões. Nesta classe há um menor tempo gasto com *Preparação de alimentos* e *Escola*, e maior tempo gasto com *Dormir*. Na quarta classe formada temos representado o sexo feminino, com idades maioritariamente entre os 15 e os 34 anos, com todas as regiões presentes, que dedicam mais tempo com *Roupa e Calçado* e *Escola*. Na classe 5 está o sexo feminino, na faixa etária mais avançada, com maior tempo dedicado a *Preparação de alimentos* e *Dormir* e menor tempo gasto com *Escola*.

Através das Tabelas 6.1 e 6.2 podemos retirar algumas conclusões mais gerais.

Classe	Composição	Características
1	Mulheres de Portugal excepto Alentejo.	Menos tempo com <i>Jogos</i> e <i>Exercício produtivo</i> .
2	Homens de Portugal.	Mais tempo com <i>Jogos</i> .
3	Mulheres do Alentejo.	Mais tempo com <i>Exercício produtivo</i> , menos tempo com <i>Jogos</i> .

Tabela 6.1: Resumo das classes formadas pelo método DIV - *Região/Sexo*, variáveis intervalares

Classe	Composição	Características
1	Faixas etárias dos 25 aos 64 anos das regiões Centro, Algarve e Alentejo.	Tempo mais reduzido dedicado a <i>Dormir</i> e à <i>Escola</i> .
2	Sexo masculino com idades entre os 15 e os 34 anos, para todas as regiões.	Menos tempo com <i>Roupa e Calçado</i> e mais tempo com a <i>Escola</i> .
3	Sexo masculino com idades mais avançadas em todas as regiões.	Menor tempo gasto com <i>Preparação de alimentos</i> e <i>Escola</i> , e maior tempo gasto com <i>Dormir</i> .
4	Sexo feminino com idades entre os 15 e os 34 anos, para todas as regiões.	Dedicam mais tempo com <i>Roupa e Calçado</i> e <i>Escola</i> .
5	Sexo feminino com idades mais avançadas em todas as regiões.	Maior tempo dedicado a <i>Preparação de alimentos</i> e <i>Dormir</i> e menor tempo gasto com <i>Escola</i> .

Tabela 6.2: Resumo das classes formadas pelo método DIV - *Região/Sexo/Grupo Etário*, variáveis intervalares

Quando consideramos a agregação *Região/Sexo*, verificamos que existe uma divisão entre as mulheres e os homens, no modo como ocupam o tempo, contudo as mulheres da região do Alentejo distinguem-se das restantes mulheres do país. A actividade responsável por esta divisão são os *Jogos*, onde os homens dedicam mais tempo do que as mulheres. As mulheres do Alentejo distinguem-se das restantes pelo tempo dedicado ao *Exercício produtivo*. Quando descemos ao nível do *Grupo Etário* a divisão é caracterizada pelo *Sexo* e *Grupo Etário*, não pela *Região*, pois verificamos a existência de duas classes femininas e duas masculinas, com idades mais jovens e mais idosas e outra classe com o resto dos grupos, com idades intermédias. O sexo feminino mais jovem dedica mais tempo com *Roupa e Calçado* e *Escola* enquanto que o masculino distingue-se pelo contrário. Na classe etária mais idosa temos a divisão do sexo masculino e feminino devido à actividade *Preparação de alimentos*, que tem mais tempo dedicado nas mulheres. A classe intermédia, distingue-se de todas as outras, pois dedica menos tempo em *Dormir* e *Escola*. Conclui-se que as variáveis *Sexo* e *Grupo Etário* são mais discriminantes que a *Região*.

Aplicando o método divisivo no primeiro nível para as variáveis modais (ver Figura 6.46), *Apressado*, *Freq-apressado*, *Apressado-5anos*, *Apressado-3anos*, *Apressado-1ano*, *vp-ntft*, *tem-fds*, *tempo-disp*, obtém-se para a formação de 3 classes uma inércia explicada de 51,7%.

```

THE CLUSTERING TREE :
-----
- the number noted at each node indicates
  the order of the division
- Ng <-> yes and Nd <-> no

+---- Classe 1 (Ng=4)
!
!----1- [Apressado = 01]
!
!      +---- Classe 2 (Ng=3)
!      !
!      !----2- [vp_ntft = 0000010]
!      !
!      +---- Classe 3 (Nd=3)

PARTITION IN 3 CLUSTERS :
-----;

Cluster 1 (n=4): RA/F  RALG/M  RALG/F  RA/M
Cluster 2 (n=3): RC/M  RLVT/M  RN/M
Cluster 3 (n=3): RC/F  RLVT/F  RN/F

Explicated inertia : 51.702333

DESCRIPTION OF THE CLUSTERS :
-----

Cluster 1: IF 1-[Apressado = Não ]

Cluster 2: IF 2-[vp_ntft = Sim, mas raramente] AND 1-[Apressado = Sim]

Cluster 3: IF 2-[vp_ntft = NS/NR OR  Sim, com frequência OR  Nunca
OR  Sim, todos os dias OR  Sim, algumas vezes OR  NA ]
AND 1-[Apressado = Sim ]

```

Figura 6.46: Classificação Divisiva - *Região/Sexo*, variáveis modais

A variável que mais se destaca é *Apressado*. Na primeira classe temos as regiões do Sul, Alentejo e Algarve, que tanto para o sexo masculino como para o feminino não se sentem apressados na maioria. Na segunda classe temos o sexo masculino nas regiões do Centro, Lisboa e Vale do Tejo e Norte, com sentimento de pressa e com tempo raro na realização de todas as tarefas na vida particular. Na última classe as mulheres das regiões do Centro, Lisboa e Vale do Tejo e Norte estão associadas ao sentimento de pressa e quanto ao tempo na vida particular, não revelam a característica da classe 2.

Para o segundo nível de agregação, com uma partição em 4 classes, o método Divisivo produz os resultados da Figura 6.47.

Comparando os resultados obtidos no primeiro nível de cruzamento com o segundo nível, (ver Figura 6.46), temos em ambos os casos a variável *Apressado* como a mais importante. Neste nível a variável *vp-ntft* deixa de ser discriminante e passamos a ter as variáveis *tem-fds* e *Freq-apressado*.

No primeiro grupo que é o maior, estão as classes etárias mais novas e as mais idosas, sendo que nas mais novas o sexo masculino predomina sem a região de Lisboa e Vale do Tejo. Associada a este grupo temos a variável *Apressado* na categoria não,

```

THE CLUSTERING TREE :
-----
- the number noted at each node indicates
  the order of the division
- Ng <-> yes and Nd <-> no

+---- Classe 1 (Ng=24)
!
!----1- [Apressado = 01]
!
!      +---- Classe 2 (Ng=6)
!      !
!      !----2- [Tempo_fds = 000010]
!      !
!      !      +---- Classe 3 (Ng=14)
!      !      !
!      !      !----3- [Freq_apressado = 0010]
!      !      !
!      !      +---- Classe 4 (Nd=6)

PARTITION IN 4 CLUSTERS :
-----;

Cluster 1 (n=24): RALG/M/15_24  RN/M/15_24  RA/M/15_24  RC/M/15_24  RALG/F/15_24
                  RALG/M/25_34  RA/M/35_54  RALG/F/55_64  RA/M/55_64  RALG/M/55_64
                  RN/F/55_64  RC/F/55_64  RC/M/55_64  RN/M/55_64
                  RA/F/65 ou mais  RALG/M/65 ou mais  RALG/F/65 ou mais  RA/M/65 ou mais
                  RC/F/65 ou mais  RC/M/65 ou mais  RLVT/F/65 ou mais
                  RLVT/M/65 ou mais  RN/M/65 ou mais  RN/F/65 ou mais

Cluster 2 (n=6):  RA/F/15_24  RLVT/M/15_24  RALG/M/35_54
                  RA/F/55_64  RLVT/F/55_64  RLVT/M/55_64

Cluster 3 (n=14): RLVT/F/15_24  RC/F/15_24  RN/F/15_24
                  RC/M/25_34  RLVT/M/25_34  RN/F/25_34  RN/M/25_34  RA/M/25_34  RA/F/25_34
                  RA/F/35_54  RC/M/35_54  RN/M/35_54  RLVT/M/35_54  RN/F/35_54

Cluster 4 (n=6):  RALG/F/25_34  RC/F/25_34  RLVT/F/25_34
                  RALG/F/35_54  RC/F/35_54  RLVT/F/35_54

Explicated inertia : 58.226856

DESCRIPTION OF THE CLUSTERS :
-----

Cluster 1 : IF 1- [Apressado = Não]

Cluster 2 : IF 2- [Tempo_fds = Nunca] AND 1- [Apressado = Sim]

Cluster 3 : IF 3- [Freq_apressado = Com frequência] AND
                2- [Tempo_fds = NS/NR OR Sim, mas raramente
                   OR Sim, algumas vezes OR Sim, com frequência OR NA] AND
                1- [Apressado = Sim]

Cluster 4 : IF 3- [Freq_apressado = Todos os dias OR Algumas vezes OR NA] AND
                2- [Tempo_fds = NS/NR OR Sim, mas raramente OR Sim, algumas vezes OR
                   Sim, com frequência OR NA ] AND
                1- [Apressado = Sim]

```

Figura 6.47: Classificação Divisiva - *Região/Sexo/Grupo Etário*, variáveis modais

ou seja, estes grupos não se sentem apressados na generalidade. Na segunda classe predomina a região de Lisboa e Vale do Tejo, com os homens na classe etária mais nova e ambos os sexos na faixa etária entre os 55 e os 64 anos. A região do Alentejo também está representada com sexo feminino nas mesmas faixas etárias. Estes grupos são caracterizados por nunca terem tempo disponível ao fim-de-semana e sentirem-se apressados. Na terceira classe estão na maioria as faixas etárias mais jovens do sexo feminino, para as regiões mais a Norte do país, e os homens, excepto do Algarve, nas faixas etárias entre os 25 e os 54 anos, que se caracterizam por sentirem pressa com frequência e com algum tempo disponível no fim-de-semana. Na quarta classe temos o sexo feminino com idades entre os 25 e os 54 anos, para as regiões do Algarve, Centro e Lisboa e Vale do Tejo. Esta classe tem um comportamento muito semelhante ao da terceira classe, mas difere quanto à frequência do sentimento de pressa, que vai desde todos os dias a algumas vezes. Não há uma grande diferença quanto ao sentimento de pressa entre as classes, a não ser entre a primeira e as restantes.

O resumo das classes obtidas, pelo método Divisivo, para ambos os níveis de agregação apresetam-se nas Tabelas 6.3 e 6.4.

Classe	Composição	Características
1	Regiões a Sul do país, Alentejo e Algarve.	Na maioria não há sentimento de pressa.
2	Homens da região Norte, Centro e Lisboa e Vale do Tejo.	Sentem-se na maioria apressados e raramente têm tempo disponível na vida particular para realizar todas as tarefas.
3	Mulheres da região Norte, Centro e Lisboa e Vale do Tejo.	Sentem-se na maioria apressadas e podem ter ou não tempo disponível na vida particular para realizar todas as tarefas.

Tabela 6.3: Resumo das classes formadas pelo método DIV - *Região/Sexo*, variáveis modais

Classe	Composição	Características
1	Classes etárias mais novas e mais idosas, com o sexo masculino predominante nas mais jovens excepto na região de Lisboa e Vale do Tejo.	Na maioria não há sentimento de pressa.
2	Predomina a região de Lisboa e Vale do Tejo, com os homens jovens e ambos os sexos na faixa etária entre os 55 e os 64 anos, e Alentejo com mulheres jovens e com idades entre os 55 e os 64 anos.	Sentem-se na maioria apressados e nunca têm tempo disponível ao fim-de-semana.
3	Na maioria, mulheres mais jovens das regiões mais a Norte do país, e os homens excepto do Algarve, nas faixas etárias entre os 25 e os 54 anos.	Sentem-se na maioria apressados com frequência e têm algum tempo disponível ao fim-semana.
4	Mulheres com idades entre os 25 e os 54 anos, para as regiões do Algarve, Centro e Lisboa e Vale do Tejo.	Sentem-se na maioria apressados todos os dias ou algumas vezes e têm algum tempo disponível ao fim-semana.

Tabela 6.4: Resumo das classes formadas pelo método DIV - *Região/Sexo/Grupo Etário*, variáveis modais

No primeiro nível de agregação temos uma divisão por regiões do país, podendo-se dizer que no Sul à menos sentimento de pressa que no Norte do País. Quando passamos ao nível da segunda agregação, a divisão ao nível das regiões não é tão clara, mas concluiu-se que os únicos grupos da Região de Lisboa e Vale do Tejo que não se sentem na maioria apressados são as faixas etárias mais idosas e que o Algarve só tem as mulheres com idades entre os 25 e os 34 com sentimento de pressa. Nos grupos etários, os mais jovens do sexo masculino e as faixas etárias mais idosas de ambos os sexos são os que revelam não se sentirem apressados.

6.2.2 Método em Hierarquia e Pirâmide

O estudo dos métodos de classificação em Hierarquia e Pirâmide foi feito na Secção 4.2.2. Relembrando, o principal objectivo deste método é agrupar os elementos do conjunto E em classes homogêneas e posteriormente agrupar as classes formadas até que se forme uma única classe contendo todos os elementos. Quando temos uma estrutura em Hierarquia, cada nível de agregação corresponde a uma partição, quando estamos no caso de uma estrutura em Pirâmide, cada nível de agregação corresponde a uma cobertura e obtemos em simultâneo uma classificação e uma seriação do conjunto E . Neste método, o critério utilizado na formação das classes é a dualidade intensão-extensão: cada classe da Hierarquia ou Pirâmide deve corresponder a um conceito, ou seja, é representada por um objecto simbólico completo cuja extensão é a própria classe.

Variáveis Intervalares

Para as variáveis intervalares e para o primeiro nível de cruzamento, a representação em Hierarquia (ver Figura 6.48) sugere duas classes distintas. Verifica-se uma clara oposição entre o sexo feminino e o sexo masculino.

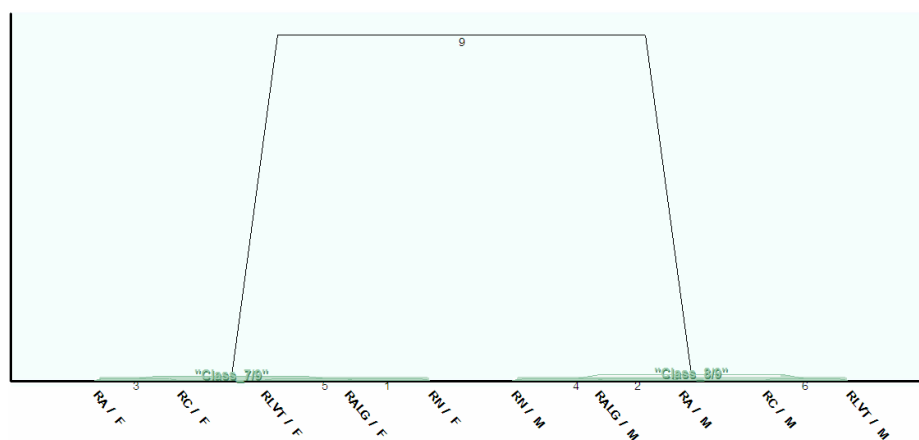


Figura 6.48: Classificação em Hierarquia - *Região/Sexo*, variáveis intervalares

Se observarmos (ver Figura 6.49) em pormenor qual a descrição de cada uma destas classes e salientando algumas variáveis, podemos dizer que a classe “feminina” (7/9) dedica mais tempo que os homens a actividades como o *Trabalho*, *Escola* e *Estudo*, e actividades relacionadas com a casa como *Preparação de alimentos* e *Roupa e Calçado*. A classe “masculina” (8/9) dedica mais tempo do que as mulheres com actividades como *Cuidados pessoais*, *Construção*, *Jardim*, *Jogos*, *Cultura* e *Convívio*. Podemos concluir que há uma generalização do conceito da classe “feminina” e “masculina”.

```

THE CLASS - "C_7/9"

AGGREGATION HEIGHT - 0.00827653

THE CLASS EXTENSION (SYMBOLIC OBJECTS)

[RA/F, RC/F, RLVT/F, RALG/F, RN/F]

LONG SYMBOLIC OBJECT DESCRIBING THE CLASS

[Dormir = [0,1330]] ^ [Alimentação = [0,530]] ^
[Cuidados_pessoais = [0,320]] ^ [Trabalho = [0,1140]] ^ [2ª_Actividade = [0,620]] ^
[Tempo_emprego = [0, 220]] ^ [Escola = [0, 830]] ^ [Estudo = [0,450]] ^
[Prep_alimentos = [0,630]] ^ [Arranjo_casa = [0, 550]] ^ [Roupa_e_calçado = [0,620]]
^ [Jardim = [0,500]] ^ [Construção = [0,370]] ^ [Compras = [0,450]] ^
[Gestão_casa = [0, 240]] ^ [Cuidar_crianças = [0,720]] ^ [Cuidar_adultos = [0,320]]
^ [Convívio = [0,600]] ^ [Cultura = [0,300]] ^ [Lazer_passivo = [0,570]] ^
[Exercício_físico = [0,480]] ^ [Exercício_produtivo = [0,510]] ^ [Artes = [0,270]] ^
[Jogos = [0,310]] ^ [Outros_passatempos = [0,120]] ^ [Ler = [0,440]] ^
[Televisão = [0,750]] ^ [Música = [0,350]] ^ [Trajecto_casa = [0,190]]

=====

THE CLASS - "C_8/9"

AGGREGATION HEIGHT - 0.0138379

THE CLASS EXTENSION (SYMBOLIC OBJECTS)

[RN/M, RALG/M, RA/M, RC/M, RLVT/M]

LONG SYMBOLIC OBJECT DESCRIBING THE CLASS

[Dormir = [60,1340]] ^ [Alimentação = [0,420]] ^
[Cuidados_pessoais = [0,490]] ^ [Trabalho = [0,910]] ^ [2ª_Actividade = [0,500]] ^
[Tempo_emprego = [0,100]] ^ [Escola = [0,790]] ^ [Estudo = [0,350]] ^
[Prep_alimentos = [0,340]] ^ [Arranjo_casa = [0,470]] ^ [Roupa_e_calçado = [0,370]]
^ [Jardim = [0,700]] ^ [Construção = [0,810]] ^ [Compras = [0,370]] ^
[Gestão_casa = [0,190]] ^ [Cuidar_crianças = [0,370]] ^ [Cuidar_adultos = [0,390]]
^ [Convívio = [0,830]] ^ [Cultura = [0,520]] ^ [Lazer_passivo = [0,640]] ^
[Exercício_físico = [0,440]] ^ [Exercício_produtivo = [0,670]] ^ [Artes = [0,290]] ^
[Jogos = [0,480]] ^ [Outros_passatempos = [0,170]] ^ [Ler = [0,440]] ^
[Televisão = [0,890]] ^ [Música = [0,470]] ^ [Trajecto_casa = [0,300]]

```

Figura 6.49: Classificação em Hierarquia - *Região/Sexo*, classes 7/9 e 8/9

Utilizando agora o método em Pirâmide para o primeiro nível de agregação com as variáveis intervalares também se verifica uma oposição entre a classe “masculina” e “feminina” (ver Figura 6.50), mas não tão acentuada como na Hierarquia. Com esta representação existe uma “aglutinação” das classes o que faz com que alguns dos objectos simbólicos pertençam a várias classes. Se observarmos as classes 40/45 e 41/45 na Figura 6.51 verificamos que as mulheres do Algarve, Região de Lisboa e Vale do Tejo e Centro, pertencem apenas à classe 40/45 e os homens da região de Lisboa e Vale do Tejo e do Centro pertencem apenas à classe 41/45, e os restantes grupos pertencem a ambas as classes. Na classe 40/45 estão representadas todas as mulheres juntamente com os homens da região do Alentejo e região Norte. Na classe 41/45 temos todos os homens juntamente com as mulheres da região do Alentejo e região Norte. Existe claramente uma oposição “feminino”/“masculino”, mas que é atenuada pela região do Norte e Alentejo.

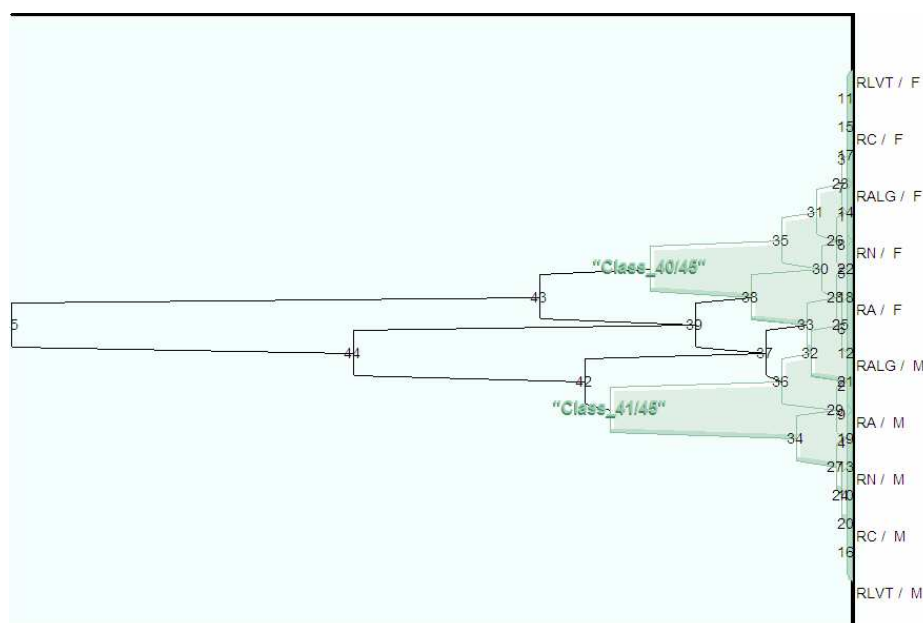


Figura 6.50: Classificação em Pirâmide - *Região/Sexo*, variáveis intervalares

Quanto às descrições das classes seleccionadas (Figuras 6.50 e 6.51), são muito semelhantes às descrições obtidas para representação em Hierarquia. A classe “feminina” e homens do Norte e Alentejo dedicam mais tempo em actividades como o *Trabalho*, *Escola e Estudo* e *Preparação de alimentos* do que a outra classe. A classe “masculina” juntamente com as mulheres do Norte e Alentejo, dedicam mais tempo que a outra classe a actividades como *Construção*, *Cultura e Música*.

Conclui-se, que para estas variáveis, o método em Hierarquia é mais rígido do que o

```

THE CLASS - "C_40/45"

AGGREGATION HEIGHT - 0.238428

THE CLASS EXTENSION (SYMBOLIC OBJECTS)

[RALG/F, RN/F, RC/F, RA/F, RALG/M, RA/M, RLVT/F, RN/M]

LONG SYMBOLIC OBJECT DESCRIBING THE CLASS

[Dormir = [0,1340]] ^ [Alimentação = [0,530]] ^
[Cuidados_pessoais = [0,490]] ^ [Trabalho = [0,1140]] ^ [2ª_Actividade = [0,620]] ^
[Tempo_emprego = [0,220]] ^ [Escola = [0,830]] ^ [Estudo = [0,450]] ^
[Prep_alimentos = [0,630]] ^ [Arranjo_casa = [0,550]] ^ [Roupa_e_calçado = [0,620]]
^ [Jardim = [0,700]] ^ [Construção = [0,810]] ^ [Compras = [0,450]] ^
[Gestão_casa = [0,240]] ^ [Cuidar_crianças = [0,720]] ^ [Cuidar_adultos = [0,320]]
^ [Convívio = [0,830]] ^ [Cultura = [0,300]] ^ [Lazer_passivo = [0,570]] ^
[Exercicio_fisico = [0,480]] ^ [Exercicio_produtivo = [0,670]] ^ [Artes = [0,270]] ^
[Jogos = [0,480]] ^ [Outros_passatempos = [0,160]] ^ [Ler = [0,440]] ^
[Televisão = [0,890]] ^ [Música = [0,350]] ^ [Trajecto_casa = [0,260]]

=====

THE CLASS - "C_41/45"

AGGREGATION HEIGHT - 0.286806

THE CLASS EXTENSION (SYMBOLIC OBJECTS)

[RALG /M, RA/M, RN/M, RC/M, RLVT/M, RA/F, RN/F]

LONG SYMBOLIC OBJECT DESCRIBING THE CLASS

[Dormir = [0,1340]] ^ [Alimentação = [0,420]] ^
[Cuidados_pessoais = [0,490]] ^ [Trabalho = [0,950]] ^ [2ª_Actividade = [0,620]] ^
[Tempo_emprego = [0,220]] ^ [Escola = [0,790]] ^ [Estudo = [0,350]] ^
[Prep_alimentos = [0,440]] ^ [Arranjo_casa = [0,540]] ^ [Roupa_e_calçado = [0,620]]
^ [Jardim = [0,700]] ^ [Construção = [0,810]] ^ [Compras = [0,420]] ^
[Gestão_casa = [0,240]] ^ [Cuidar_crianças = [0,720]] ^ [Cuidar_adultos = [0,390]]
^ [Convívio = [0,830]] ^ [Cultura = [0,520]] ^ [Lazer_passivo = [0,640]] ^
[Exercicio_fisico = [0,440]] ^ [Exercicio_produtivo = [0,670]] ^ [Artes = [0,290]] ^
[Jogos = [0,480]] ^ [Outros_passatempos = [0,170]] ^ [Ler = [0,440]] ^
[Televisão = [0,890]] ^ [Música = [0,470]] ^ [Trajecto_casa = [0,300]]

=====

```

Figura 6.51: Classificação em Pirâmide - *Região/Sexo*, classes 40/45 e 41/45

método em Pirâmide, havendo contudo semelhança nos resultados. A representação em Hierarquia fornece uma informação mais geral e o método em Pirâmide uma informação mais detalhada, principalmente se queremos averiguar a descrição dos objectos simbólicos que estão na “fronteira”.

Quando descemos ao segundo nível de agregação, com as variáveis intervalares, as representações em Pirâmide e Hierarquia, não são muito claras, uma vez que este é um método aglomerativo e quando os níveis de agregação têm valores muito baixos, como é o caso, a visualização não é muito explícita. Sendo assim proceder-se-á à interpretação de todos os objectos simbólicos para as variáveis intervalares, apenas com base nas listagens fornecidas.

Considerando em primeiro lugar a representação em Hierarquia, para o segundo nível de agregação e para as variáveis intervalares, verificamos a existência de algumas classes interessantes (ver Figuras 6.52 e 6.53).

Nas classes 31/39 e 38/39 (ver Figura 6.52) temos uma maioria de grupos jovens ou em idade activa de várias regiões, sendo a classe 38/39 formada apenas por elementos do sexo masculino, excepto da região de Lisboa e Vale do Tejo. Na classe 31/39 temos

em maioria elementos do sexo feminino nas faixas etárias jovens, aos quais se junta o grupo dos homens de Lisboa e Vale do Tejo com idades entre os 25 e os 34 anos.

Comparando as duas classes, temos por exemplo, maior tempo dedicado em actividades como *Dormir*, *2ª Actividade*, *Jardim*, *Construção* e *Lazer passivo* na classe masculina jovem e activa, sem Lisboa e Vale do Tejo, do que na classe feminina jovem de todo o País e masculina activa de Lisboa e Vale do Tejo. Nesta classe temos maior tempo dedicado em *Arranjo da casa*, *Roupa e Calçado* e *Compras*.

```

THE CLASS - "C_38/39"

AGGREGATION HEIGHT - 2.92719e-007

THE CLASS EXTENSION (SYMBOLIC OBJECTS)

[RC/M/15_24, RA/M/15_24, RN/M/15_24, RALG/M/15_24,
RALG/M/25_34, RA/M/25_34, RALG/M/35_54, RA/M/35_54, RALG/M/55_64, RC/M/25_34]

LONG SYMBOLIC OBJECT DESCRIBING THE CLASS

[Dormir = [60, 1060]] ^ [Cuidados_pessoais = [0, 310]] ^ [Trabalho = [0, 880]] ^
[2ª_Actividade = [0, 500]] ^ [Tempo_emprego = [0, 70]] ^ [Escola = [0, 790]] ^
[Estudo = [0, 350]] ^ [Arranjo_casa = [0, 170]] ^ [Roupa_e_calçado = [0, 70]] ^
[Jardim = [0, 700]] ^ [Construção = [0, 810]] ^ [Compras = [0, 370]] ^
[Gestão_casa = [0, 10]] ^ [Cuidar_crianças = [0, 370]] ^ [Convívio = [0, 580]] ^
[Cultura = [0, 280]] ^ [Lazer_passivo = [0, 640]] ^ [Exercicio_fisico = [0, 290]] ^
[Exercicio_produtivo = [0, 670]] ^ [Artes = [0, 240]] ^ [Jogos = [0, 370]] ^
[Outros_passatempos = [0, 20]] ^ [Ler = [0, 400]] ^ [Televisão = [0, 580]] ^
[Música = [0, 320]] ^ [Trajecto_casa = [0, 300]]

=====

THE CLASS - "C_31/39"

AGGREGATION HEIGHT - 2.32585e-009

THE CLASS EXTENSION (SYMBOLIC OBJECTS)

[RLVT/F/25_34, RN/F/15_24, RALG/F/35_54, RALG/F/25_34, RA/F/15_24, RLVT/M/25_34]

LONG SYMBOLIC OBJECT DESCRIBING THE CLASS

[Dormir = [80, 930]] ^ [Cuidados_pessoais = [0, 280]] ^ [Trabalho = [0, 800]] ^
[2ª_Actividade = [0, 270]] ^ [Tempo_emprego = [0, 170]] ^ [Escola = [0, 750]] ^
[Estudo = [0, 350]] ^ [Arranjo_casa = [0, 510]] ^ [Roupa_e_calçado = [0, 340]] ^
[Jardim = [0, 230]] ^ [Construção = [0, 220]] ^ [Compras = [0, 450]] ^
[Gestão_casa = [0, 120]] ^ [Cuidar_crianças = [0, 510]] ^ [Convívio = [0, 580]] ^
[Cultura = [0, 150]] ^ [Lazer_passivo = [0, 230]] ^ [Exercicio_fisico = [0, 270]] ^
[Exercicio_produtivo = [0, 20]] ^ [Artes = [0, 120]] ^ [Jogos = [0, 110]] ^
[Outros_passatempos = [0, 60]] ^ [Ler = [0, 180]] ^ [Televisão = [0, 530]] ^
[Música = [0, 180]] ^ [Trajecto_casa = [0, 190]]

```

Figura 6.52: Classificação em Hierarquia - *Região/Sexo/Grupo Etário*, classes 31/39 e 38/39

Nas classes 25/39 e 36/39 (ver Figura 6.53), temos faixas etárias mais elevadas, constituídas principalmente pelo sexo feminino. Na classe 25/39 temos as regiões a Sul do País maioritariamente do sexo feminino e na classe 36/39 temos elementos do sexo feminino das regiões mais a Norte do país e dois grupos masculinos também de regiões mais a Norte do país. As descrições destas duas classes são muito semelhantes, contudo verifica-se que para a classe que representa as regiões mais a Norte do País, observamos mais tempo dedicado em actividades como *Trabalho*, *Arranjo da casa*, *Construção* e *Exercício produtivo*.

```

THE CLASS - "C_36/39"

AGGREGATION HEIGHT - 1.26171e-007

THE CLASS EXTENSION (SYMBOLIC OBJECTS)

[RC/F/35_54, RC/F/55_64, RN/F/55_64,
RN/M/65 ou mais, RC/M/65 ou mais, RC/F/65 ou mais, RN/F/65 ou mais,]

LONG SYMBOLIC OBJECT DESCRIBING THE CLASS

[Dormir = [0, 1340]] ^ [Cuidados_pessoais = [0, 280]] ^ [Trabalho = [0, 920]] ^
[2ª_Actividade = [0, 300]] ^ [Tempo_emprego = [0, 60]] ^ [Escola = [0, 120]] ^
[Estudo = [0, 20]] ^ [Arranjo_casa = [0, 550]] ^ [Roupa_e_calçado = [0, 620]] ^
[Jardim = [0, 500]] ^ [Construção = [0, 510]] ^ [Compras = [0, 420]] ^
[Gestão_casa = [0, 240]] ^ [Cuidar_crianças = [0, 390]] ^ [Convívio = [0, 600]] ^
[Cultura = [0, 240]] ^ [Lazer_passivo = [0, 570]] ^ [Exercício_físico = [0, 440]] ^
[Exercício_produtivo = [0, 480]] ^ [Artes = [0, 210]] ^ [Jogos = [0, 270]] ^
[Outros_passatempos = [0, 10]] ^ [Ler = [0, 380]] ^ [Televisão = [0, 750]] ^
[Música = [0, 420]] ^ [Trajecto_casa = [0, 140]]

```

```

THE CLASS - "C_25/39"

AGGREGATION HEIGHT - 1.82658e-010

THE CLASS EXTENSION (SYMBOLIC OBJECTS)

[RA/M/65 ou mais, RA/F/65 ou mais, RALG/F/65 ou mais, RLVT/F/65 ou mais]

LONG SYMBOLIC OBJECT DESCRIBING THE CLASS

[Dormir = [0, 1330]] ^ [Cuidados_pessoais = [0, 320]] ^ [Trabalho = [0, 890]] ^
[2ª_Actividade = [0, 70]] ^ [Tempo_emprego = [0, 50]] ^ [Escola = [0, 60]] ^
[Estudo = [0, 10]] ^ [Arranjo_casa = [0, 370]] ^ [Roupa_e_calçado = [0, 620]] ^
[Jardim = [0, 680]] ^ [Construção = [0, 390]] ^ [Compras = [0, 290]] ^
[Gestão_casa = [0, 150]] ^ [Cuidar_crianças = [0, 160]] ^ [Convívio = [0, 590]] ^
[Cultura = [0, 130]] ^ [Lazer_passivo = [0, 520]] ^ [Exercício_físico = [0, 480]] ^
[Exercício_produtivo = [0, 50]] ^ [Artes = [0, 230]] ^ [Jogos = [0, 280]] ^
[Outros_passatempos = [0, 70]] ^ [Ler = [0, 440]] ^ [Televisão = [0, 720]] ^
[Música = [0, 240]] ^ [Trajecto_casa = [0, 110]]

```

Figura 6.53: Classificação em Hierarquia - *Região/Sexo/Grupo Etário*, classes 25/39 e 36/39

Com a estrutura em Pirâmide (ver Figura 6.54), o segundo nível de agregação com as variáveis intervalares, sugere três classes mais distintas, mas que não identificam grupos com características tão acentuadas com na representação em Hierarquia. Temos uma classe (211/212) formada por muitos grupos com idades entre os 15 e os 34 anos e alguns elementos com idades entre os 55 e os 64 anos. Na classe 203/212 e 188/212 temos maioritariamente elementos com mais de 65 anos e dois grupos jovens masculinos do Norte e Centro. Comparando o tempo dedicado pelas várias actividades, a classe 211/212 dedica menos tempo em actividades como *Roupa e Calçado*, *Lazer passivo* e *Exercício físico* e mais tempo a *Jogos* e *Tempo emprego*, do que as outras duas classes. A classe 203/212, formada por mais elementos do sexo masculino, tem uma descrição muito semelhante à da classe 188/212, que é formada por mais elementos do sexo feminino. Contudo na classe “masculina” temos mais tempo dedicado a *Música*, *Construção*, *Exercício produtivo* e *Estudo*.

```

THE CLASS - "C_211/212"

AGGREGATION HEIGHT - 1.01058e-005

THE CLASS EXTENSION (SYMBOLIC OBJECTS)

[RA/F/15_24, RA/M/15_24, RN/F/15_24, RALG/M/15_24, RALG/F/15_24, RLVT/F/15_24,
RC/F/15_24, RLVT/M/25_34, RA/M/25_34, RA/F/25_34, RC/M/25_34, RALG/M/25_34,
RN/M/25_34, RA/M/35_54, RA/F/55_64, RC/M/55_64, RALG/M/55_64]

LONG SYMBOLIC OBJECT DESCRIBING THE CLASS

[Dormir = [60, 1230]] ^ [Alimentação = [0, 530]] ^ [Cuidados_pessoais = [0, 330]] ^
[Trabalho = [0, 880]] ^ [2ª_Actividade = [0, 460]] ^ [Tempo_emprego = [0, 170]] ^
[Escola = [0, 830]] ^ [Estudo = [0, 350]] ^ [Prep_alimentos = [0, 370]] ^
[Arranjo_casa = [0, 470]] ^ [Roupa_e_calçado = [0, 380]] ^ [Jardim = [0, 530]] ^
[Construção = [0, 810]] ^ [Compras = [0, 370]] ^ [Gestão_casa = [0, 50]] ^
[Cuidar_crianças = [0, 450]] ^ [Cuidar_adultos = [0, 130]] ^ [Convívio = [0, 830]] ^
[Cultura = [0, 280]] ^ [Lazer_passivo = [0, 330]] ^ [Exercicio_fisico = [0, 290]] ^
[Exercicio_produtivo = [0, 670]] ^ [Artes = [0, 140]] ^ [Jogos = [0, 480]] ^
[Outros_passatempos = [0, 50]] ^ [Ler = [0, 400]] ^ [Televisão = [0, 640]] ^
[Música = [0, 320]] ^ [Trajecto_casa = [0, 170]]

=====

THE CLASS - "C_203/212"

AGGREGATION HEIGHT - 2.55392e-006

THE CLASS EXTENSION (SYMBOLIC OBJECTS)

[RC/F/65 ou mais, RN/M/65 ou mais, RC/M/15_24, RC/M/65 ou mais]

LONG SYMBOLIC OBJECT DESCRIBING THE CLASS

[Dormir = [170, 1340]] ^ [Alimentação = [10, 360]] ^ [Cuidados_pessoais = [0, 280]] ^
[Trabalho = [0, 710]] ^ [2ª_Actividade = [0, 470]] ^ [Tempo_emprego = [0, 60]] ^
[Escola = [0, 790]] ^ [Estudo = [0, 350]] ^ [Prep_alimentos = [0, 400]] ^
[Arranjo_casa = [0, 340]] ^ [Roupa_e_calçado = [0, 620]] ^ [Jardim = [0, 500]] ^
[Construção = [0, 510]] ^ [Compras = [0, 270]] ^ [Gestão_casa = [0, 180]] ^
[Cuidar_crianças = [0, 340]] ^ [Cuidar_adultos = [0, 250]] ^ [Convívio = [0, 600]] ^
[Cultura = [0, 240]] ^ [Lazer_passivo = [0, 640]] ^ [Exercicio_fisico = [0, 440]] ^
[Exercicio_produtivo = [0, 480]] ^ [Artes = [0, 240]] ^ [Jogos = [0, 270]] ^
[Outros_passatempos = [0, 10]] ^ [Ler = [0, 380]] ^ [Televisão = [0, 650]] ^
[Música = [0, 420]] ^ [Trajecto_casa = [0, 300]]

=====

THE CLASS - "C_188/212"

AGGREGATION HEIGHT - 2.98397e-007

THE CLASS EXTENSION (SYMBOLIC OBJECTS)

[RA/F/65 ou mais, RALG/F/65 ou mais, RLVT/F/65 ou mais, RN/F/65 ou mais,
RA/M/55_64, RN/M/15_24]

LONG SYMBOLIC OBJECT DESCRIBING THE CLASS

[Dormir = [0, 1330]] ^ [Alimentação = [0, 390]] ^ [Cuidados_pessoais = [0, 320]] ^
[Trabalho = [0, 750]] ^ [2ª_Actividade = [0, 480]] ^ [Tempo_emprego = [0, 40]] ^
[Escola = [0, 690]] ^ [Estudo = [0, 70]] ^ [Prep_alimentos = [0, 440]] ^
[Arranjo_casa = [0, 390]] ^ [Roupa_e_calçado = [0, 620]] ^ [Jardim = [0, 560]] ^
[Construção = [0, 350]] ^ [Compras = [0, 290]] ^ [Gestão_casa = [0, 240]] ^
[Cuidar_crianças = [0, 160]] ^ [Cuidar_adultos = [0, 320]] ^ [Convívio = [0, 590]] ^
[Cultura = [0, 210]] ^ [Lazer_passivo = [0, 520]] ^ [Exercicio_fisico = [0, 480]] ^
[Exercicio_produtivo = [0, 140]] ^ [Artes = [0, 230]] ^ [Jogos = [0, 210]] ^
[Outros_passatempos = [0, 70]] ^ [Ler = [0, 440]] ^ [Televisão = [0, 750]] ^
[Música = [0, 240]] ^ [Trajecto_casa = [0, 260]]

```

Figura 6.54: Classificação em Pirâmide - *Região/Sexo/Grupo Etário*, variáveis intervalares

Apresentam-se nas Tabelas, 6.5 e 6.6, a comparação entre os resultados obtidos pelos dois métodos aplicados para os dois níveis de agregação:

Hierarquia	Pirâmide
Oposição entre classe “feminina” e “masculina”, com as mulheres a dispenderem mais tempo com actividades como o <i>Trabalho</i> e <i>Escola</i> e actividades relacionadas com a casa, e os homens a dispenderem mais tempo com <i>Cuidados pessoais</i> , <i>Construção</i> e <i>Convívio</i> .	Oposição entre classe “feminina” e “masculina” que é atenuada pelas regiões Norte e Alentejo.

Tabela 6.5: Comparação da classificação em Hierarquia e em Pirâmide - *Região/Sexo*, variáveis intervalares

Hierarquia	Pirâmide
Classe formada por homens jovens e em idade activa de todas as regiões com mais tempo dedicado a <i>Dormir</i> , <i>2ª Actividade</i> , <i>Jardim</i> , <i>Construção</i> e <i>Lazer passivo</i> . Classe feminina jovem de todo o país e masculina de Lisboa e Vale do Tejo com mais tempo dedicado a <i>Arranjo da casa</i> , <i>Roupa e Calçado</i> e <i>Compras</i> . Classe maioritariamente formada pelo sexo feminino com idades mais avançadas, de regiões mais a Sul do país, que se opõe à classe feminina e masculina mais idosa do Norte do país, pelo menor tempo gasto em <i>Trabalho</i> , <i>Arranjo da casa</i> , <i>Construção</i> e <i>Exercício produtivo</i> .	Classe formada por grupos mais jovens e com idades entre os 55 e 54 anos; classe formada por mais elementos do sexo masculino em idade avançada e dois grupos jovens masculinos; classe formada por mais elementos do sexo feminino em idade mais avançada. A classe mais jovem dedica menos tempo a actividades como <i>Roupa e Calçado</i> , <i>Lazer passivo</i> e mais tempo a <i>Jogos</i> e <i>Tempo emprego</i> . A diferença entre as classe mais idosas, feminina e masculina, é que a masculina dedicada por exemplo mais tempo a <i>Estudo</i> , <i>Música</i> e <i>Construção</i> .

Tabela 6.6: Comparação da classificação em Hierarquia e em Pirâmide - *Região/Sexo/Grupo Etário*, variáveis intervalares

Conclui-se que tendo em conta apenas a *Região* e o *Sexo*, a variável sexo é mais importante para discriminar o modo como o tempo é ocupado. Quando consideramos o *Grupo Etário* as diferenças não são tão acentuadas, mas o *Grupo Etário* e o *Sexo*, são mais discriminantes que a *Região*.

Variáveis Modais

Analisando agora as variáveis modais no primeiro nível de cruzamento, através da estrutura em Hierarquia e com a generalização pelo máximo (ver na Secção 4.2.2. a fórmula 4.21), obtém-se a representação da Figura 6.55.

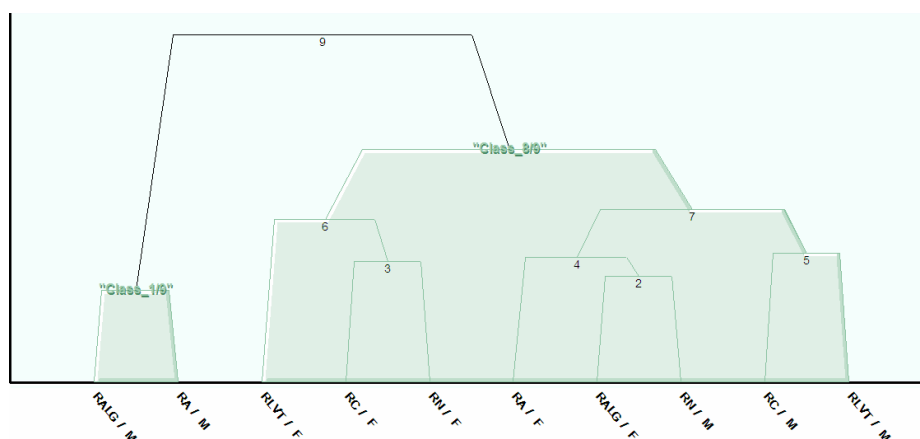


Figura 6.55: Classificação em Hierarquia - *Região/Sexo*, variáveis modais

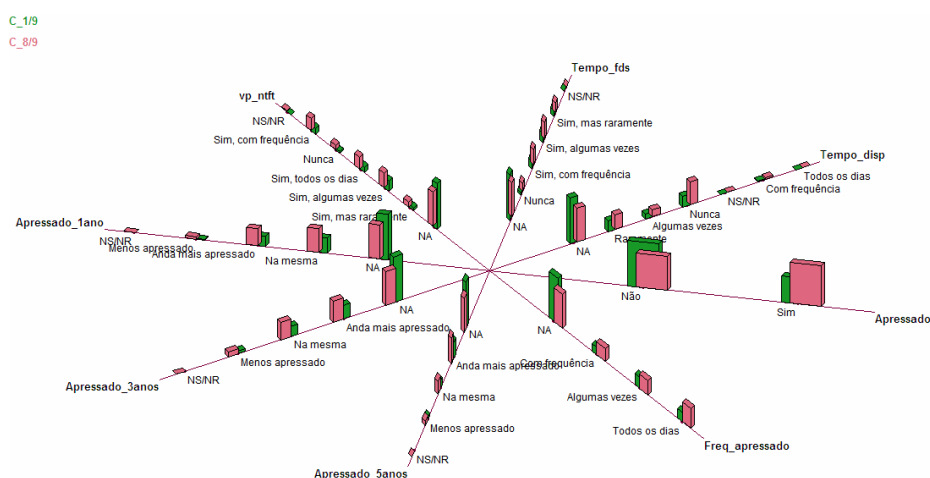


Figura 6.56: Estrela 3D *Superimpose* - classes 1/9 e 8/9

Observa-se que há uma divisão entre as regiões do Algarve e Alentejo para o sexo masculino (ver Figura 6.56), classe 1/9, e as restantes na classe 8/9. Nestes dois grupos não há um sentimento de pressa, contrapondo com o resto do país. Relativamente às outras variáveis não se verificam grandes diferenças entre as duas classes, a não ser no *tempo-disp* (tempo disponível para realizar todas as tarefas) que tem maior frequência na categoria *nunca* na classe 8/9 do que na classe 1/9.

Utilizando o método de classificação em Pirâmide, com a agregação pelo máximo para

o primeiro nível (ver Figura 6.57), verificam-se algumas diferenças comparando com a Hierarquia. Temos uma oposição entre o sexo feminino na região de Lisboa e Vale do Tejo e o sexo masculino na região do Algarve. As regiões do Alentejo e Algarve no sexo masculino são as classes mais próximas. Agregados às classes da região do Algarve e Alentejo em ambos os sexos, estão agora os homens da região Norte.

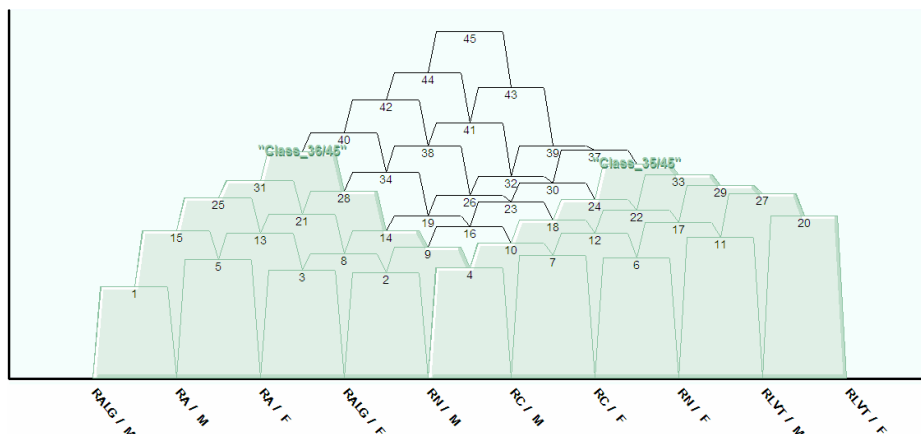


Figura 6.57: Pirâmide - *Região/Sexo*, variáveis modais

Na classe 36/45 (ver Figura 6.58), que é originada pela agregação das regiões do Algarve e Alentejo em ambos os sexos e pelo sexo masculino na região Norte, a maioria dos indivíduos não sente pressa, mas dos que se sentem apressados verifica-se que, em grande parte, não têm nunca tempo disponível. Ainda sobre os indivíduos que se sentem apressados, a maioria estão na mesma ou mais apressados, relativamente há 5, 3 e 1 ano atrás. Na classe 35/45, existem mais indivíduos com sentimento de pressa relativamente aos que não se sentem apressados, apesar da diferença não ser muito elevada.

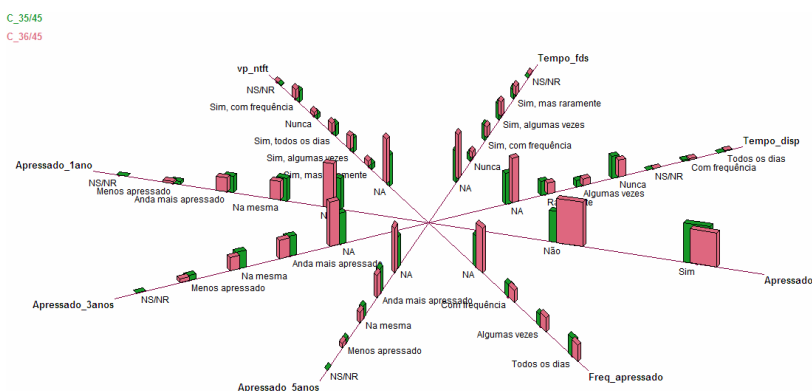


Figura 6.58: Estrela 3D *Superimpose* - classes 35/45 e 36/45

Para as variáveis modais no segundo nível de agregação, a representação em Hierarquia (ver Figura 6.59) com a generalização pelo máximo, sugere a existência de três grandes classes, que diferem entre si principalmente pelo *Grupo Etário*. Na classe 36/49 temos a maioria dos grupos etários com idades mais elevadas aos quais se juntam o grupo do sexo masculino do Algarve com idades entre os 15 e os 34 anos. Na classe 46/49 temos a faixa etária mais jovem de ambos os sexos, excepto os homens do Algarve, juntamente com os grupos com idades entre os 55 e os 64 anos. Na classe 47/49 temos a maioria dos grupos com idades entre os 25 e os 54 anos.

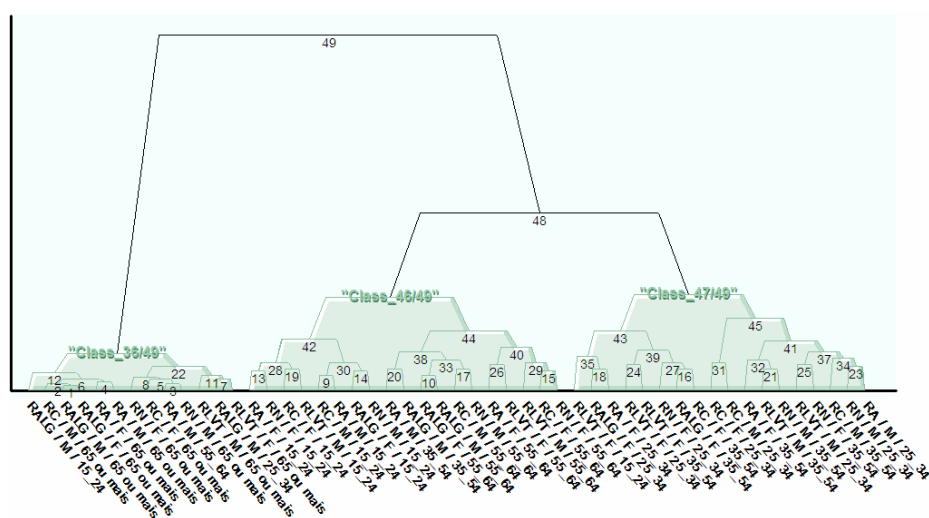


Figura 6.59: Classificação em Hierarquia - *Região/Sexo/Grupo Etário*, variáveis modais

A classe com os grupos de idades mais avançadas e jovens do sexo masculino do Algarve, não se sentem na maioria apressados (ver Figura 6.60). Na classe 46/49 não há uma tendência para o sentimento de pressa muito acentuada, já na classe 47/49 temos um sentimento de pressa generalizado, que relativamente ao passado tem vindo a aumentar ou mantém-se, de uma forma em geral (ver Figura 6.61).

Através da estrutura em Hierarquia observa-se uma divisão de Portugal quanto ao sentimento de pressa por *Grupo Etário*, que é mais acentuado nas faixas etárias dos 25 aos 54 anos, médio nas faixas etárias dos 15 aos 24 anos e dos 35 aos 54 anos, e muito baixo nos grupos etários com mais de 65 anos e nos homens dos 15 aos 34 anos do Algarve.

```

THE CLASS - "C_36/49"

AGGREGATION HEIGHT - 0.791459

THE CLASS EXTENSION (SYMBOLIC OBJECTS)

[RALG/M/15_24, RC/M/65 ou mais, RALG/M/65 ou mais, RALG/F/65 ou mais, RA/F/65 ou mais,
RA/M/65 ou mais, RN/F/65 ou mais, RC/F/65 ou mais, RA/M/55_64,
RN/M/65 ou mais, RLVT/M/65 ou mais, RALG/M/25_34, RLVT/F/65 ou mais]

LONG SYMBOLIC OBJECT DESCRIBING THE CLASS

[Apressado = (Não(0.926059), Sim(0.272232))] ^

[Freq_apressado = (NA(0.926059), Com frequência(0.102071), Algumas vezes(0.139736),
Todos os dias (0.138081))] ^

[Apressado_5anos = (NA(0.926059), Anda mais apressado(0.178031), Na mesma(0.129544),
Menos apressado(0.0925505), NS/NR(0.0140963))] ^

[Apressado_3anos = (NA(0.926059), Anda mais apressado(0.161309), Na mesma(0.174789),
Menos apressado(0.0925505), NS/NR(0.0140963))] ^

[Apressado_1ano = (NA(0.926059), Na mesma(0.180717), Anda mais apressado(0.0601596),
Menos apressado(0.0778208), NS/NR(0.0140963))] ^

[vp_ntft = (NA(0.926059), Sim, mas raramente(0.0970755), Sim, algumas vezes(0.131515),
Sim, todos os dias(0.0762321), Nunca(0.0591551), Sim, com frequência(0.0697594),
NS/NR(0.0140963))] ^

[Tempo_fds = (NA(0.926059), Nunca(0.0840422), Sim, com frequência(0.0696734), Sim,
algumas vezes(0.134087), Sim, mas raramente(0.137996), NS/NR(0.0185124))] ^

[Tempo_disp = (NA(0.926059), Raramente(0.140546), Algumas vezes(0.0763786),
Nunca(0.166015), NS/NR(0.0140963), Com frequência(0.0179513), Todos os
dias(0.0223321))]

```

Figura 6.60: Classificação em Hierarquia - *Região/Sexo/Grupo Etário*, classe 36/49

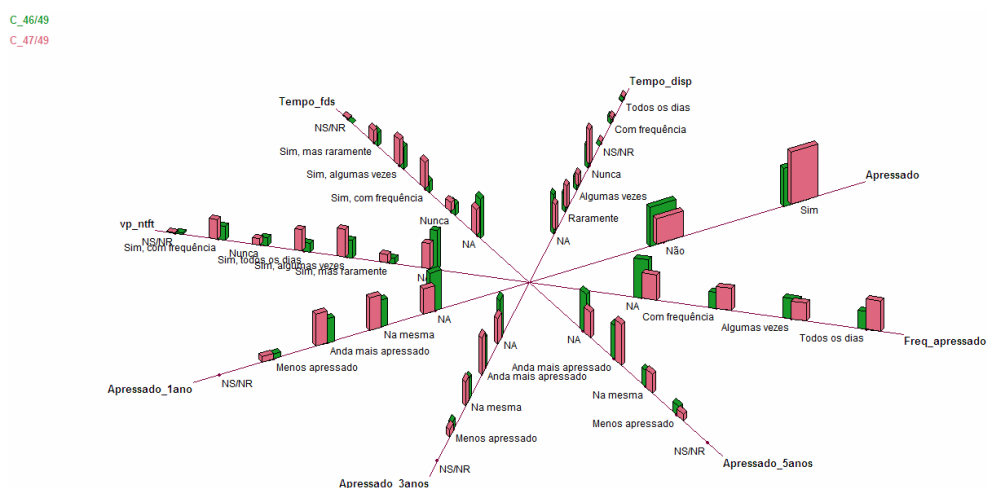


Figura 6.61: Classificação em Hierarquia - *Região/Sexo/Grupo Etário*, classes 46/39 e 47/49

Com a estrutura em Pirâmide e a generalização pelo máximo (ver Figura 6.62), no segundo nível, obtemos uma representação muito semelhante à Hierarquia. Temos uma divisão por *Grupo Etário* em três classes.

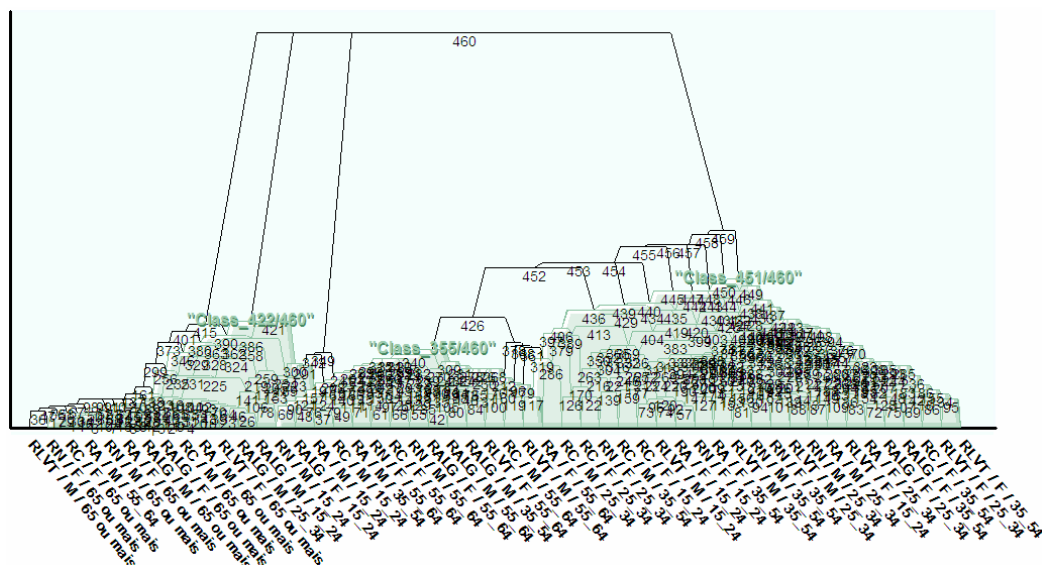


Figura 6.62: Classificação em Pirâmide - *Região/Sexo/Grupo Etário*, variáveis modais

Verifica-se que na classe constituída maioritariamente por idosos, juntam-se agora mais grupos jovens do sexo masculino e feminino do Algarve e do sexo masculino da região Norte e Alentejo. Se observarmos a Figura 6.63 o comportamento é muito semelhante ao que se verificava com a estrutura em Hierarquia. A classe 422/460 formada pelos grupos mais idosos e masculinos mais jovens da região do Algarve, Norte e Alentejo, não se sentem na maioria apressados. A classe com idades mais avançadas (classe 355/460), mas menores que 64 anos estão agora mais associadas ao não sentimento de pressa, e os grupos etários com idades entre os 25 e os 54 anos (classe 451/460) revelam um sentimento de pressa generalizado.

Hierarquia	Pirâmide
Classe com grupos etários com idades mais elevadas aos quais se juntam os grupos do sexo masculino do Algarve com idades entre os 15 e os 34 anos, que não se sentem na maioria apressados; Classe com faixas etárias mais jovens de ambos sexos (excepto Algarve) e grupos com idades entre os 55 e os 64 anos, com tendência para o sentimento de pressa muito acentuada; Classe com maioria dos grupos com idades entre os 25 e os 54 anos, com sentimento de pressa generalizado.	Classe formada pelos grupos mais idosos e masculinos mais jovens da região do Algarve, Norte e Alentejo, não se sentem na maioria apressados; Classe com idades mais avançadas, mas menores que 64 anos, estão agora mais associadas ao não sentimento de pressa; Classe com grupos etários com idades entre os 25 e os 54 anos revelam um sentimento de pressa generalizado.

Tabela 6.8: Comparação da classificação em Hierarquia e em Pirâmide - *Região/Sexo/Grupo Etário*, variáveis modais

Quando descemos ao segundo nível, as conclusões retiradas com representação em Hierarquia e Pirâmide são muito semelhantes, contudo no grupo dos mais idosos e homens mais jovens do Algarve, juntam-se agora os homens mais jovens da região Norte e Alentejo.

A título ilustrativo, apresentam-se os resultados obtidos com as estruturas em Hierarquia e Pirâmide, para as variáveis intervalares e modais, apenas para as faixas etárias mais jovens (dos 15 aos 24 anos e dos 25 aos 34 anos). De forma a facilitar a interpretação das classes, denomina-se grupo jovem os grupos com idades entre 15 aos 24 anos e grupo activo os grupos com idades entre 25 aos 34 anos.

Com as variáveis intervalares nas classes etárias mais jovens, a Hierarquia (ver Figura 6.64) produz níveis de agregação muito baixos, tendo a última classe um nível mais elevado, não permitindo assim uma visualização clara das classes formadas. Contudo, observa-se que o sexo feminino pertence quase na sua totalidade a uma classe, opondo-se ao sexo masculino. As mulheres da região do Alentejo, tanto em idade activa como jovens e as mulheres jovens de Lisboa e Vale do Tejo, juntam-se à classe masculina.

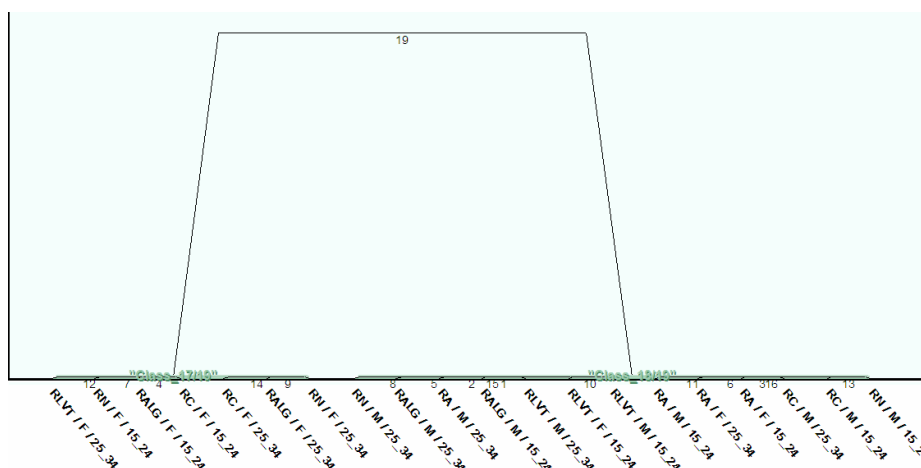


Figura 6.64: Classificação em Hierarquia - *Região/Sexo/Grupo Etário*, variáveis intervalares

Observando as diferenças relativamente à ocupação do tempo nas duas classes (ver Figura 6.65), verifica-se que na classe 17/19, constituída apenas pelo sexo feminino excepto a região do Alentejo e as jovens de Lisboa e Vale do Tejo, a ocupação do tempo incide mais sobre actividades como *Cuidar crianças*, *Compras*, *Alimentação*, *Estudo* e *Tempo-emprego*. Na classe 18/19, formada maioritariamente pelo sexo masculino ao qual se juntam as mulheres do Alentejo e as jovens de Lisboa e Vale do Tejo, o tempo já é mais ocupado com actividades como *Convívio*, *Lazer passivo*, *Exercício produtivo*, *Jogos* e *Construção*.

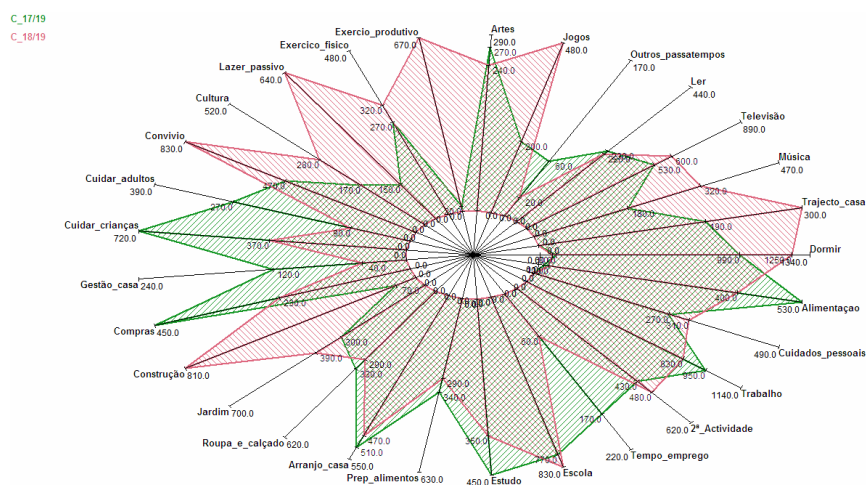


Figura 6.65: Estrela - *Região/Sexo/Grupo Etário*, classes 17/19 e 18/19

Conclui-se que existe uma separação feminino/masculino, mas as mulheres do Alentejo e as jovens de Lisboa e Vale do Tejo, tendem a ser mais parecidas com os homens quanto à forma como ocupam o tempo. O sexo continua a ser a variável que mais discrimina a forma como a distribuição do tempo é feita.

Através do método em Pirâmide (ver Figura 6.66), obtemos uma estrutura que reúne quase todos os grupos nos primeiros níveis.

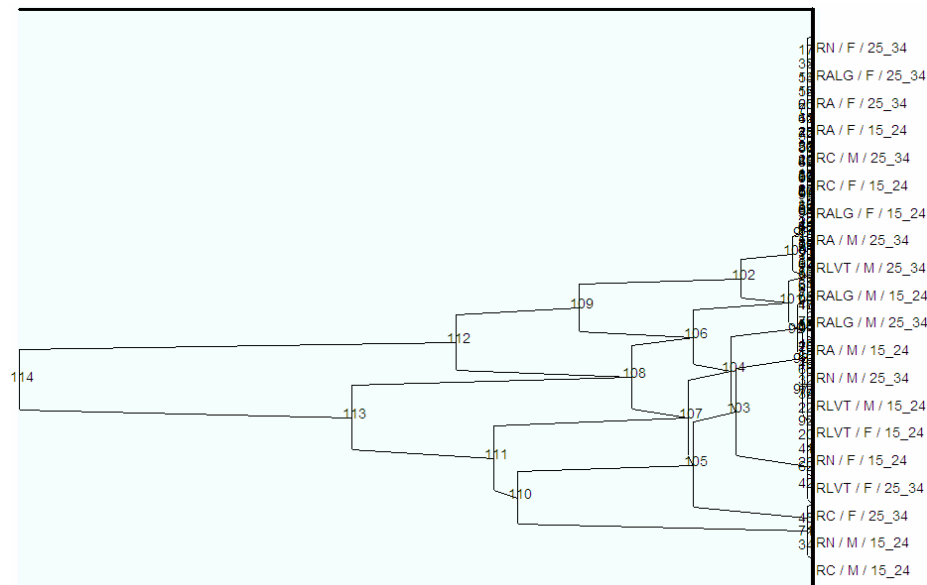


Figura 6.66: Classificação em Pirâmide - *Região/Sexo/Grupo Etário*, variáveis intervalares

Com esta estrutura torna-se difícil retirar algumas conclusões, pois a maioria das classes estão nos níveis mais baixos. Para se obter uma estrutura mais fácil de interpretar efectuou-se um “*pruning*” da Pirâmide, com um nível de 31% (ver Secção 4.2.2) obtendo-se a representação da Figura 6.67.

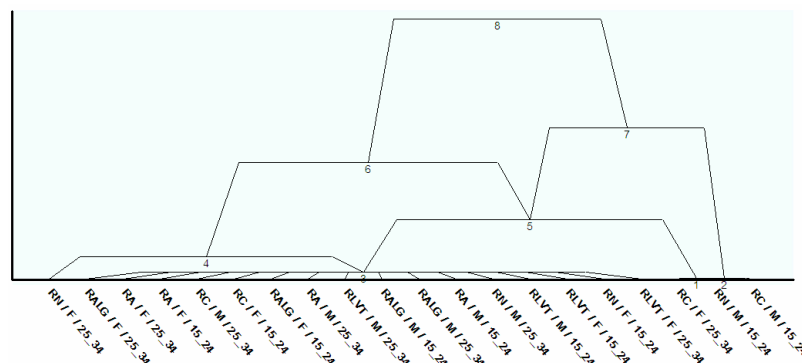


Figura 6.67: “*Pruning*” da Pirâmide - *Região/Sexo/Grupo Etário*, variáveis intervalares

Com esta estrutura já se observa que há um grande grupo formado por quase todos os objectos excepto os homens com idades entre os 15 e os 24 anos da região Centro e Norte, aos quais se juntam as mulheres da região Centro com idades entre os 25 e os 34 anos, e por outro lado as mulheres da região Norte na faixa etária dos 25 aos 34 anos.

Se observarmos em particular estas classes na Pirâmide original (ver Figura 6.68), a classe 17/114 que era constituída pelas mulheres activas do Algarve e Norte, em comparação com o grande grupo (classe 104/114), tem diferenças significativas quanto ao tempo ocupado. As actividades de *Dormir*, *Alimentação*, *Escola*, *Estudo* e *Exercício produtivo*, têm menos tempo gasto do que na classe 104/114, enquanto que *Cuidar de crianças* tem maior tempo consumido. Quanto à classe 34/114 formada pelos jovens do sexo masculino da região Norte e Centro, quando comparada com o grande grupo, tem menos tempo gasto em *Cuidados pessoais*, *Trabalho*, *Arranjo da casa*, *Roupa e calçado*, *Cuidar de crianças* e *Cuidar de adultos*, e mais tempo consumido com *Lazer passivo* e *Trajecto casa*.

```

THE CLASS - "C_17/114"

AGGREGATION HEIGHT - 3.36216e-013

THE CLASS EXTENSION (SYMBOLIC OBJECTS)

[RALG/F/25_34, RN/F/25_34]

LONG SYMBOLIC OBJECT DESCRIBING THE CLASS

[Dormir = [190,980]] ^ [Alimentação = [10,220]] ^
[Cuidados_pessoais = [0,140]] ^ [Trabalho = [0,950]] ^ [2ª_Actividade = [0,430]] ^
[Tempo_emprego = [0,40]] ^ [Escola = [0,130]] ^ [Estudo = [0,150]] ^
[Prep_alimentos = [0,340]] ^ [Arranjo_casa = [0,370]] ^ [Roupa_e_calçado = [0,330]] ^
[Jardim = [0,300]] ^ [Construção = [0,20]] ^ [Compras = [0,350]] ^
[Gestão_casa = [0, 30]] ^ [Cuidar_crianças = [0,720]] ^ [Cuidar_adultos = [0,100]] ^
[Convívio = [0, 280]] ^ [Cultura = [0,170]] ^ [Lazer_passivo = [0,130]] ^
[Exercicio_fisico = [0,120]] ^ [Exercicio_produtivo = [0,0]] ^ [Artes = [0,270]] ^
[Jogos =[0,170]] ^ [Outros_passatempos = [0,60]] ^ [Ler = [0,110]] ^
[Televisão = [0,510]] ^ [Música = [0,170]] ^ [Trajecto_casa = [0,160]]

=====

THE CLASS - "C_104/114"

AGGREGATION HEIGHT - 4.06536e-005

THE CLASS EXTENSION (SYMBOLIC OBJECTS)

[RALG/F/15_24, RC/F/15_24, RC/M/25_34, RLVT/M/25_34, RA/M/25_34,
RALG/M/15_24, RALG /M/25_34, RA/F/15_24, RA/F/25_34,
RA/M/15_24, RN/M/25_34, RLVT/M/15_24, RLVT/F/15_24,
RN/F/15_24, RLVT/F/25_34, RC/F/25_34]

LONG SYMBOLIC OBJECT DESCRIBING THE CLASS

[Dormir = [60,1250]] ^ [Alimentação = [10,530]] ^
[Cuidados_pessoais = [0,310]] ^ [Trabalho = [0,830]] ^ [2ª_Actividade = [0,430]] ^
[Tempo_emprego = [0,170]] ^ [Escola = [0,830]] ^ [Estudo = [0,450]] ^
[Prep_alimentos = [0,290]] ^ [Arranjo_casa = [0,510]] ^ [Roupa_e_calçado = [0,290]] ^
[Jardim = [0,390]] ^ [Construção = [0,810]] ^ [Compras = [0,450]] ^
[Gestão_casa = [0,120]] ^ [Cuidar_crianças = [0,590]] ^ [Cuidar_adultos = [0,270]] ^
[Convívio = [0,830]] ^ [Cultura = [0,280]] ^ [Lazer_passivo = [0,330]] ^
[Exercicio_fisico = [0,320]] ^ [Exercicio_produtivo = [0,670]] ^ [Artes = [0,210]] ^
[Jogos = [0,480]] ^ [Outros_passatempos = [0,50]] ^ [Ler = [0,230]] ^
[Televisão = [0,600]] ^ [Música = [0,320]] ^ [Trajecto_casa = [0,190]]

=====

THE CLASS - "C_34/114"

AGGREGATION HEIGHT - 6.9374e-012

THE CLASS EXTENSION (SYMBOLIC OBJECTS)

[RC/M/15_24, RN/M/15_24]

LONG SYMBOLIC OBJECT DESCRIBING THE CLASS

[Dormir = [170,1020]] ^ [Alimentação = [30,390]] ^
[Cuidados_pessoais = [0,140]] ^ [Trabalho = [0,660]] ^ [2ª_Actividade = [0,480]] ^
[Tempo_emprego = [0,30]] ^ [Escola = [0,790]] ^ [Estudo = [0,350]] ^
[Prep_alimentos = [0,260]] ^ [Arranjo_casa = [0,170]] ^ [Roupa_e_calçado = [0,60]] ^
[Jardim = [0,120]] ^ [Construção = [0,410]] ^ [Compras = [0,120]] ^
[Gestão_casa = [0,0]] ^ [Cuidar_crianças = [0,10]] ^ [Cuidar_adultos = [0,90]] ^
[Convívio = [0,530]] ^ [Cultura = [0,210]] ^ [Lazer_passivo = [0,640]] ^
[Exercicio_fisico = [0,280]] ^ [Exercicio_produtivo = [0,140]] ^ [Artes = [0,240]] ^
[Jogos = [0,210]] ^ [Outros_passatempos = [0,0]] ^ [Ler = [0,200]] ^
[Televisão = [0,550]] ^ [Música = [0,200]] ^ [Trajecto_casa = [0,300]]

```

Figura 6.68: Classificação em Pirâmide - *Região/Sexo/Grupo Etário*, classes 17/114, 34/114 e 104/114

Para o nível de agregação *Região/Sexo/Grupo Etário*, mas apenas para as duas faixas etárias mais jovens, apresenta-se o quadro resumo (ver Tabela 6.9) dos resultados obtidos com os métodos aplicados:

Hierarquia	Pirâmide
oposição entre o sexo masculino e feminino, com a região do Alentejo e jovens de Lisboa e Vale do Tejo mais próximas do comportamento masculino. Classe feminina com mais tempo em actividades como <i>Estudo</i> , <i>Cuidar de crianças</i> e <i>Compras</i> , e classe masculina com mais tempo em <i>Jogos</i> , <i>Lazer passivo</i> , <i>Exercício produtivo</i> , <i>Construção</i> e <i>Jogos</i>	mulheres activas do Algarve com menos tempo em <i>Dormir</i> , <i>Alimentação</i> , <i>Escola</i> , <i>Estudo</i> e <i>Exercício produtivo</i> e mais tempo com <i>Cuidar de crianças</i> , e homens jovens da região Centro e Norte, com menor tempo gasto em actividades relacionadas com a casa e trabalho e mais tempo gasto com <i>Trajecto casa</i> e <i>Lazer passivo</i>

Tabela 6.9: Comparação da classificação em Hierarquia e em Pirâmide - *Região/Sexo/Grupo Etário*, faixas etárias mais jovens com variáveis intervalares

A idade não é muito importante para diferenciar a ocupação do tempo nos grupos etários dos 15 aos 24 e dos 25 aos 34 anos. As duas classes identificadas não são sobretudo diferenciadas pela idade, contudo revelam algumas excepções no comportamento para as mulheres activas do Algarve e para os homens jovens da região Centro e Norte que apresentam algumas diferenças no modo como ocupam o seu tempo face aos restantes grupos.

Com a classificação hierárquica e a descrição dada pelas variáveis modais, para as classes etárias dos 15 aos 24 anos, e dos 25 aos 34 anos, podemos observar também dois grandes grupos (ver Figura 6.69).

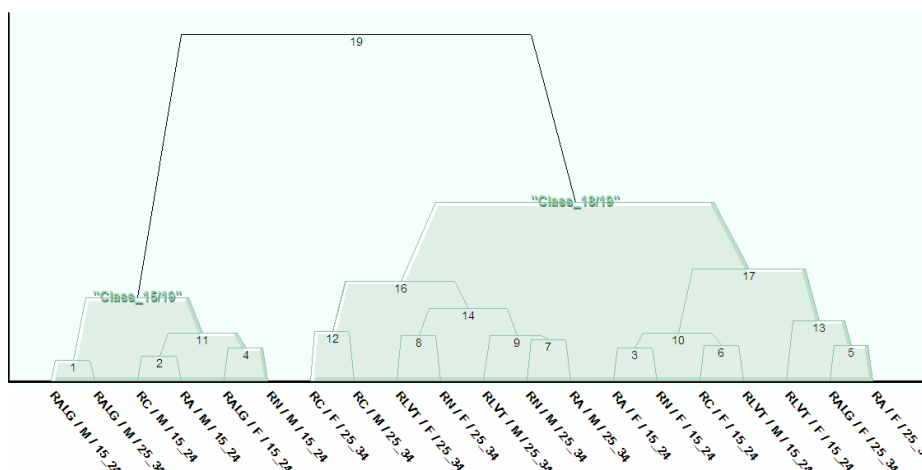


Figura 6.69: Hierarquia - *Região/Sexo/Grupo Etário*, variáveis modais

Na classe 15/19,(ver Figura 6.70), temos os homens da região do Algarve e as mulheres na faixa etária mais nova, juntamente com os homens com idades entre os 15 e os 24 anos para o resto das regiões, excepto Lisboa e Vale do Tejo. Na classe 18/19 temos os restantes grupos, que descrevem na sua maioria as mulheres, com idades entre os 15 e os 34 anos, excepto para região do Algarve na faixa etária mais nova, e os homens com idades entre os 25 e os 34 anos exceptuando também a região do Algarve. A classe 15/19, maioritariamente representada pelo sexo masculino, na faixa etária mais nova, excepto Lisboa e Vale do Tejo, não está associada à pressa, como se pode verificar pela frequência da resposta *Não* para a variável *Apressado* (AS00), enquanto que a classe 18/19 está associada à pressa (AS00) e ao pouco tempo disponível para realizar todas as tarefas (BA00).

name	"Class_15/19"
label	"C_15/19"
height	1.06669
symbolic object description	variable list (conjunction of)
name	value
AS00	(Não(0.759123), Sim(0.485567))
AT00	(NA(0.759123), Com frequência(0.258593), Algumas vezes(0.270399), Todos os dias(0.141494))
AJ00	(NA(0.759123), Andá mais apressado(0.452733), Na mesma(0.115387), Menos apressado(0.0778208))
AV00	(NA(0.759123), Andá mais apressado(0.335373), Na mesma(0.154578), Menos apressado(0.0778208))
AI00	(NA(0.759123), Na mesma(0.282742), Andá mais apressado(0.24747), Menos apressado(0.0778208))
AY00	(NA(0.759123), Sim, mas raramente(0.0970755), Sim, algumas vezes(0.253783), Sim, todos os dias(0.116379), Nunca(0.0969427), Sim, com frequência(0.122854), NS/NR(0.01...)
AZ00	(NA(0.759123), Nunca(0.130707), Sim, com frequência(0.119597), Sim, algumas vezes(0.21938), Sim, mas raramente(0.200502))
BA00	(NA(0.759123), Raramente(0.182705), Algumas vezes(0.177184), Nunca(0.146301), Com frequência(0.0494124))
base object list	RALG / M / 15_24, RALG / M / 25_34, RC / M / 15_24, RA / M / 15_24, RALG / F / 15_24, RN / M / 15_24
name	"Class_18/19"
label	"C_18/19"
height	2.32176
symbolic object description	variable list (conjunction of)
name	value
AS00	(Não(0.442992), Sim(0.810336))
AT00	(NA(0.442992), Com frequência(0.343962), Algumas vezes(0.262636), Todos os dias(0.466846))
AJ00	(NA(0.442992), Andá mais apressado(0.634414), Na mesma(0.203541), Menos apressado(0.0991183))
AV00	(NA(0.442992), Andá mais apressado(0.574788), Na mesma(0.326233), Menos apressado(0.120352))
AI00	(NA(0.442992), Na mesma(0.438394), Andá mais apressado(0.488914), Menos apressado(0.0972943))
AY00	(NA(0.442992), Sim, mas raramente(0.134112), Sim, algumas vezes(0.428156), Sim, todos os dias(0.327427), Nunca(0.103055), Sim, com frequência(0.306939), NS/NR(0.00851...)
AZ00	(NA(0.442992), Nunca(0.141266), Sim, com frequência(0.407377), Sim, algumas vezes(0.410044), Sim, mas raramente(0.190041), NS/NR(0.00795169))
BA00	(NA(0.442992), Raramente(0.344514), Algumas vezes(0.181448), Nunca(0.48936), NS/NR(0.0101584), Com frequência(0.0911316), Todos os dias(0.0231449))
base object list	RC / F / 25_34, RC / M / 25_34, RLVT / F / 25_34, RLVT / M / 25_34, RN / F / 25_34, RA / F / 15_24, RA / F / 25_34, RC / F / 15_24, RLVT / M / 15_24, RLVT / F / 15_24, RALG / F / 25_34, RA / F / 25_34

Figura 6.70: Hierarquia - classes 15/19 e 18/19

Para o mesmo nível de cruzamento e para as mesmas variáveis, mas com a classificação em Pirâmide, podemos observar pela Figura 6.71, que se formam dois grandes núcleos, opondo as duas faixas etárias, excepto para a região de Lisboa e Vale do Tejo. Dentro da classe que representa a faixa etária mais nova, existe um pequeno núcleo, a classe 80/174. Na classe mais pequena (ver Figura 6.72), está representada a faixa etária dos 15 aos 24 anos, das regiões do Norte, Centro e Alentejo em ambos os sexos, o sexo masculino da região de Lisboa e Vale do Tejo e o sexo feminino da região do Algarve. Nesta última classe os grupos dividem-se de forma equilibrada quanto ao sentimento de pressa.

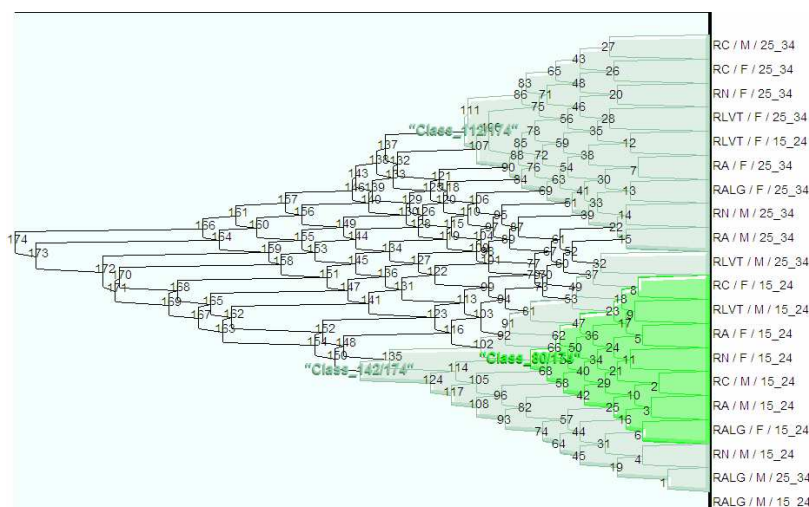


Figura 6.71: Pirâmide - Região/Sexo/Grupo Etário, variáveis modais

name	"Class_80/174"
label	"C_80/174"
height	1.14506
symbolic object description	variable list (conjunction of)
name	value
AS00	(NÃO(0.592755), Sim(0.59906))
AT00	(NÃO(0.592755), Com frequência(0.258593), Algumas vezes(0.270399), Todos os dias(0.257438))
AU00	(NÃO(0.592755), Anda mais apressado(0.524731), Na mesma(0.115387), Menos apressado(0.0432161))
AV00	(NÃO(0.592755), Anda mais apressado(0.463491), Na mesma(0.154578), Menos apressado(0.0391719))
AV00	(NÃO(0.592755), Na mesma(0.322124), Anda mais apressado(0.376686), Menos apressado(0.0642976))
AY00	(NÃO(0.592755), Sim, mas raramente(0.0826246), Sim, algumas vezes(0.280589), Sim, todos os dias(0.116379), Nunca(0.0969427), Sim, com frequência(0.220778), NS/NR(0.010))
AZ00	(NÃO(0.592755), Nunca(0.130707), Sim, com frequência(0.143091), Sim, algumas vezes(0.342849), Sim, mas raramente(0.200502))
BA00	(NÃO(0.592755), Raramente(0.271877), Algumas vezes(0.177184), Nunca(0.148419), NS/NR(0.0101584), Com frequência(0.0911316), Todos os dias(0.0231449))
base object list	RC / F / 15_24, RLV / M / 15_24, RA / F / 15_24, RN / F / 15_24, RC / M / 15_24, RA / M / 15_24, RALG / F / 15_24, RN / M / 15_24

Figura 6.72: Pirâmide - classe 80/174

Se observarmos a classe 142/174 (ver Figura 6.73), que é mais abrangente do que a classe 80/174, já temos uma maioria dos grupos que não se sentem apressados, ou seja, devido à entrada dos homens da região do Algarve em ambas as faixas etárias esta classe ficou mais associada a um não sentimento de pressa.

Na classe 112/174 (ver Figura 6.74), temos na maioria um sentimento de pressa, que é originado por ambos os sexos, em todas as regiões com idades entre os 25 e os 34

anos, excepto os homens da região do Algarve. Nesta classe também se encontra a região de Lisboa e Vale do Tejo para o sexo feminino entre os 15 e os 24 anos.

name	"Class_142/174"	
label	"C_142/174"	
height	2.2647	
symbolic object description	variable list (conjunction of)	
	name	value
	AS00	(Não(0.759123), Sim(0.670465))
	AT00	(NA(0.759123), Com frequência(0.258593), Algumas vezes(0.270399), Todos os dias(0.257438))
	AU00	(NA(0.759123), Anda mais apressado(0.524731), Na mesma(0.19766), Menos apressado(0.0778208))
	AV00	(NA(0.759123), Anda mais apressado(0.463491), Na mesma(0.326233), Menos apressado(0.0778208))
	AW00	(NA(0.759123), Na mesma(0.322124), Anda mais apressado(0.376686), Menos apressado(0.0778208))
	AY00	(NA(0.759123), Sim, mas raramente(0.0970755), Sim, algumas vezes(0.280589), Sim, todos os dias(0.116379), Nunca(0.0969427), Sim, com frequência(0.220778), NS/NR(0.010...)
	AZ00	(NA(0.759123), Nunca(0.130707), Sim, com frequência(0.234648), Sim, algumas vezes(0.342849), Sim, mas raramente(0.200502))
	BA00	(NA(0.759123), Raramente(0.309197), Algumas vezes(0.177184), Nunca(0.222593), NS/NR(0.0101584), Com frequência(0.0911316), Todos os dias(0.0231449))
base object list	RLVT/M/25_34, RC/F/15_24, RLVT/M/15_24, RA/F/15_24, RN/F/15_24, RC/M/15_24, RA/M/15_24, RALG/F/15_24, RN/M/15_24, RALG/M/25_34, RALG/M/15_24	

Figura 6.73: Pirâmide - classe 142/174

name	"Class_112/174"	
label	"C_112/174"	
height	1.59019	
symbolic object description	variable list (conjunction of)	
	name	value
	AS00	(Não(0.329535), Sim(0.610336))
	AT00	(NA(0.329535), Com frequência(0.343982), Algumas vezes(0.232732), Todos os dias(0.466846))
	AU00	(NA(0.329535), Anda mais apressado(0.634414), Na mesma(0.203541), Menos apressado(0.0991183))
	AV00	(NA(0.329535), Anda mais apressado(0.574788), Na mesma(0.326233), Menos apressado(0.120352))
	AW00	(NA(0.329535), Na mesma(0.438394), Anda mais apressado(0.488914), Menos apressado(0.0972943))
	AY00	(NA(0.329535), Sim, mas raramente(0.134112), Sim, algumas vezes(0.428156), Sim, todos os dias(0.327427), Nunca(0.103055), Sim, com frequência(0.306939), NS/NR(0.0085...)
	AZ00	(NA(0.329535), Nunca(0.141266), Sim, com frequência(0.407377), Sim, algumas vezes(0.410044), Sim, mas raramente(0.190041), NS/NR(0.00795169))
	BA00	(NA(0.329535), Raramente(0.344514), Algumas vezes(0.181448), Nunca(0.48936), NS/NR(0.00581811), Com frequência(0.038714), Todos os dias(0.018709))
base object list	RC/M/25_34, RC/F/25_34, RN/F/25_34, RLVT/F/25_34, RLVT/F/15_24, RA/F/25_34, RALG/F/25_34, RN/M/25_34, RA/M/25_34, RLVT/M/25_34	

Figura 6.74: Pirâmide - classe 112/174

Para as classes etárias mais jovens, podemos observar as diferenças entre os resultados obtidos com os dois métodos para as variáveis modais, através da Tabela 6.10.

Hierarquia	Pirâmide
Classe formada maioritariamente pelos jovens do sexo masculino, excepto Lisboa e Vale do Tejo, juntamente com os homens activos do Algarve e as mulheres jovens do Algarve, que não se sentem na generalidade apressados, em oposição aos restantes grupos que se sentem na maioria apressados.	Mesma oposição que existe na Hierarquia, mas com a particularidade de se poder verificar que os homens da região do Algarve fazem com que esse grupo esteja mais associado ao facto de não se sentirem apressados.

Tabela 6.10: Comparação da classificação em Hierarquia e em Pirâmide - *Região/Sexo/Grupo Etário*, faixas etárias mais jovens com variáveis modais

6.2.3 Classificação Não-Hierárquica: Método *SCLUST*

O método de classificação não-hierárquico *SCLUST* foi estudado na Secção 4.2.3 tem como objectivo determinar partições em classes homogéneas que são generalizadas por protótipos. O output deste método fornece:

- Lista das classes;
- Índices de qualidade das classes e da partição;
- Índices de qualidade da variável na classe e na partição;
- Vector descrição associado aos protótipos;
- A melhor partição identificada por 3 métodos para diferentes números de classes, se tal for pedido pelo utilizador.

Os índices dados para avaliar as variáveis, classes e partições são em regra geral valores entre 0 e 1, no caso do Software SODAS são indicados sob a forma de percentagem.

No primeiro nível de agregação com as variáveis intervalares, aplicou-se o método com a distância de *Hausdorff* (dada pela fórmula 4.29), para a formação no máximo de 4 classes. Com estes parâmetros os métodos *C-H* (dado pela fórmula 4.42) e o Γ -index (dado pela fórmula 4.44), induzem na escolha da partição formada por 2 classes (ver Figura 6.75).

----- ----- Best number of classes by ----- ----- MILLIGAN and COOPER indicators ----- -----			
Ngps	C-H	C-index	Gamma
4	3.18860	0.03904	0.87981
3	3.34342	0.11073	0.90800
2	5.78726	0.05024	0.90800
1	-	-	-

Figura 6.75: *SCLUST* - *Região/Sexo* - métodos de Calinski and Harabasz, C-index e Γ -index

Aplicando o método novamente, agora para a formação de uma partição com 3 classes (ver Figura 6.76), obtemos por um índice de qualidade da partição (ver fórmula 4.35) de cerca de 22% e um critério final mínimo (ver fórmula 4.27) $\Delta = 218,36$. Se a qualidade da partição for 100% as classes correspondem aos objectos singulares, se o valor do índice for 0% temos a partição formada por uma única classe.

Com esta partição obtemos uma divisão do país mediante o *Sexo*, independentemente da *Região*. Há uma clara oposição entre o sexo feminino e o masculino.


```

EDITION OPTIMAL PARTITION
=====
classe :   1   Cardinal :       5
=====
(1) RALG/M   (3) RA/M   (5) RC/M   (7) RLVT/M   (8) RN/M
classe :   2   Cardinal :       5
=====
(0) RA/F   (2) RALG/F   (4) RC/F   (6) RLVT/F   (9) RN/F
PARTITION DESCRIPTION
=====
INITIAL CRITERION      : 280.000000
FINAL CRITERION       : 218.362846
Percentage of the explained criterion :   22.01

```

Figura 6.76: *SCLUST* - *Região/Sexo* - variáveis intervalares

Na Figura 6.77 regista-se a qualidade de cada classe dada pelo índice $Q(C_t) = 1 - \frac{f_{C_t}(G_t)}{f_{C_t}(G_E)}$ (ver fórmula 4.38) e a contribuição de cada classe para a variabilidade total dada pelo índice $K(C_t) = \frac{f_{C_t}(G_t)}{\Delta(P,L)}$ (ver fórmula 4.39). A classe 2 tem maior qualidade e maior contribuição para a variabilidade global.

Cluster	Size(n _i)	Q(C _i)	K(C _i)
1	5	18.96	48.13
2	5	24.65	51.87

Figura 6.77: *SCLUST* - *Região/Sexo* - Índice de qualidade e contribuição das classes

A qualidade da partição em cada variável é a parte da homogeneidade dada por cada variável percorrendo todas as classes da partição (ver fórmula 4.36). As variáveis mais importantes na construção da partição são as que apresentam índices mais elevados do que a qualidade da partição ou inércia total, que neste caso é de 22%. Pela Figura 6.78 verificamos que as variáveis mais importantes são *Preparação de alimentos*, *Arranjo da casa*, *Roupa e calçado*, *Jardim*, *Construção*, *Cuidar de crianças*, *Exercício físico*, *Exercício produtivo* e *Jogos*.

Analisando em pormenor as variáveis mais importantes na construção da partição, mas agora tendo em consideração o critério que permite interpretar as classes de acordo com a variável (ver fórmula 4.40), verificamos que todas estas variáveis têm índices elevados, superiores ao índice de qualidade de cada classe, $Q(C_1) = 18,96$ e $Q(C_2) = 24,65$ (ver Figura 6.79). Se o índice de caracterização da classe, tendo em conta a variável for elevado, pode-se dizer que essa variável tem menor dispersão nas classes face à sua própria dispersão, logo caracteriza a classe.

Variáveis	Qj(P)
Dormir	5.51
Alimentação	3.12
Cuidados pessoais	14.58
Trabalho	10.17
2ª Actividade	17.78
Tempo emprego	18.75
Escola	8.00
Estudo	9.01
Prep alimentos	52.75
Arranjo casa	37.97
Roupa e calçado	54.64
Jardim	29.73
Construção	46.39
Compras	11.48
Gestão casa	7.27
Cuidar crianças	47.68
Cuidar adultos	7.58
Cultura	23.75
Lazer passivo	9.33
Exercicio físico	29.49
Exercio produtivo	53.74
Artes	17.07
Jogos	54.74
Outros passatempos	2.27
Ler	3.45
Televisão	2.44
Música	15.73
Trajecto casa	21.95

Figura 6.78: *SCLUST - Região/Sexo* - Índice de qualidade das variáveis intervalares

	Cluster	Minimum	Maximum	Qj(Ct)
Preparação alimentos		0.00	340.00	
	1	0.00	280.00	54.55
	2	0.00	440.00	44.93
Arranjo da casa		0.00	370.00	
	1	0.00	370.00	100.00
	2	0.00	510.00	50.82
Roupa e Calçado		0.00	340.00	
	1	0.00	160.00	63.64
	2	0.00	620.00	32.08
Jardim		0.00	480.00	
	1	0.00	580.00	64.91
	2	0.00	460.00	88.24
Construção		0.00	370.00	
	1	0.00	510.00	60.00
	2	0.00	200.00	46.91
Cuidar de crianças		0.00	340.00	
	1	0.00	190.00	45.00
	2	0.00	510.00	57.14
Exercício físico		0.00	320.00	
	1	0.00	380.00	57.14
	2	0.00	270.00	78.00
Exercício produtivo		0.00	480.00	
	1	0.00	630.00	63.46
	2	0.00	80.00	40.74
Jogos		0.00	310.00	
	1	0.00	370.00	55.00
	2	0.00	190.00	38.18

Figura 6.79: *SCLUST - Região/Sexo* - Caracterização das classes pelas variáveis intervalares

A classe 1, que representa o sexo masculino de todo o país, caracteriza-se pelo maior tempo dedicado com o *Jardim*, *Exercício produtivo*, *Construção* e *Jogos*, e menor tempo dedicado com *Preparação de alimentos*, *Arranjo da casa*, *Roupa e Calçado* e *Cuidar de crianças*. Inversamente para a classe 2, onde está representado o sexo feminino de todo o país.

Para o cruzamento *Região/Sexo/Grupo Etário*, utilizando a distância de *Hausdorf* e fixando o número máximo de classes em 5, os métodos de método C-H e Γ -index induzem para a formação de uma partição em 2 classes, enquanto que o método *C*-index (ver fórmula 4.43) indica a formação de 3 classes (ver Figura 6.80). De forma a não tornar as classes demasiado abrangentes optou-se considerar 3 classes.

----- Best number of classes by ----- ----- MILLIGAN and COOPER indicators ----- -----			
Ngps	C-H	C-index	Gamma
5	7.68013	0.16609	0.60922
4	8.77949	0.17420	0.62316
3	12.13439	0.15391	0.67478
2	14.03927	0.20820	0.69779
1	-	-	-

Figura 6.80: *SCLUST* - *Região/Sexo/Grupo Etário* - métodos de Calinski and Harabasz, C-index e Γ -index

Na classificação através do método *SCLUST* têm que ser retiradas todas as variáveis, cujo índice de qualidade da classe seja nulo, já que nessa classe os objectos para essas variáveis têm todas o mesmo valor. Nesta análise tiveram que ser ignoradas as variáveis *Alimentação*, *2ª Actividade*, *Estudo* e *Jogos*.

Para a formação de 3 classes (ver Figura 6.81) obtém-se um critério final de $\Delta = 1019$ e um índice de qualidade da partição de 18,5%.

As três classes formadas distinguem-se mais pelo *Grupo Etário* e pelo *Sexo*, do que pela *Região*. Na primeira classe temos a população masculina com idades entre os 15 e os 64 anos, mas com a maioria nas faixas etárias mais jovens. Nesta classe não estão presentes os homens da região Centro com idades entre os 25 e os 34 anos. Na segunda classe temos a população mais idosa, de ambos os sexos. Na terceira classe temos o sexo feminino com idades entre os 15 e os 64 anos, mas com a maioria nas faixas etárias mais jovens. Nesta última classe temos incluídos também os homens da região Centro com idades entre os 25 e os 34 anos.

A classe com o índice de qualidade mais elevado é a segunda e a que tem maior contribuição para a variabilidade total é a terceira (ver Figura 6.82).

```

=====
EDITION OPTIMAL PARTITION
=====

Classe : 1 Cardinal : 17
=====
(1) RALG / M / 15_24      (5) RA / M / 35_54
(10) RALG / M / 35_54    (11) RALG / M / 25_34
(15) RALG / M / 55_64    (17) RA / M / 55_64
(20) RC / M / 35_54      (26) RC / M / 15_24
(32) RLVT / M / 25_34    (33) RLVT / M / 55_64
(36) RLVT / M / 35_54    (37) RLVT / M / 15_24
(43) RN / M / 35_54      (46) RN / M / 15_24
(47) RN / M / 25_34      (48) RA / M / 15_24
(49) RA / M / 25_34

Classe : 2 Cardinal : 14
=====
(0) RA / F / 65 ou mais  (6) RALG / M / 65 ou mais
(7) RALG / F / 65 ou mais (14) RA / M / 65 ou mais
(19) RC / F / 65 ou mais  (22) RC / M / 65 ou mais
(23) RC / F / 55_64      (25) RC / M / 55_64
(30) RLVT / F / 55_64    (31) RLVT / F / 65 ou mais
(34) RLVT / M / 65 ou mais (38) RN / M / 65 ou mais
(44) RN / M / 55_64      (45) RN / F / 65 ou mais

| Classe : 3 Cardinal : 19
=====
(2) RA / F / 55_64      (3) RALG / F / 55_64
(4) RA / F / 35_54      (8) RALG / F / 35_54
(9) RALG / F / 15_24    (12) RALG / F / 25_34
(13) RA / F / 25_34     (16) RA / F / 15_24
(18) RC / F / 35_54     (21) RC / F / 15_24
(24) RC / F / 25_34     (27) RC / M / 25_34
(28) RLVT / F / 25_34   (29) RLVT / F / 35_54
(35) RLVT / F / 15_24   (39) RN / F / 35_54
(40) RN / F / 25_34     (41) RN / F / 55_64
(42) RN / F / 15_24

PARTITION DESCRIPTION
=====

INITIAL CRITERION : 1250.000000
FINAL CRITERION : 1019.021486
Percentage of the explained criterion : 18.48

```

Figura 6.81: *SCLUST* - *Região/Sexo/Grupo Etário* - variáveis intervalares

Cluster	Size(n_i)	Q(Ct)	K(Ct)
1	17	18.17	33.45
2	14	23.69	31.39
3	19	13.51	35.16

Figura 6.82: *SCLUST* - *Região/Sexo/Grupo Etário* - Índice de qualidade e contribuição das classes

Tendo agora em consideração o índice da qualidade da partição em cada variável (ver Figura 6.83), verificamos que as variáveis *Dormir*, *Escola*, *Preparação de alimentos*, *Arranjo da casa*, *Roupa e Calçado*, *Jardim*, *Cuidar de crianças*, *Cuidar de adultos*, *Lazer passivo*, *Exercício produtivo* e *Televisão* são as variáveis mais importantes na construção da partição.

Variáveis	Qj(P)
Dormir	31.96
Cuidados pessoais	11.07
Trabalho	16.73
Tempo emprego	3.17
Escola	28.30
Prep alimentos	27.27
Arranjo casa	32.52
Roupa e calçado	29.68
Jardim	32.93
Construção	18.81
Compras	6.89
Gestão casa	6.36
Cuidar crianças	29.55
Cuidar adultos	29.96
Convívio	12.09
Cultura	4.68
Lazer passivo	37.33
Exercício físico	12.35
Exercício produtivo	27.10
Artes	3.61
Outros passatempos	1.90
Ler	13.68
Televisão	26.17
Música	9.57
Trajecto casa	8.26

Figura 6.83: *SCLUST* - *Região/Sexo/Grupo Etário* - Índice de qualidade das variáveis intervalares

Podemos observarmos estas variáveis em cada classe na Figura 6.84 tendo em consideração o índice de caracterização da classe pela variável.

A classe 1, que caracteriza o sexo masculino até aos 64 anos excepto para os homens da região Centro com idades entre os 25 e os 34 anos, está associada ao maior tempo dedicado com *Escola* e *Exercício produtivo*, algum tempo dedicado ao *Jardim*, e pouco tempo em *Roupa e calçado*, *Arranjo da casa* e *Cuidar de adultos*. A segunda classe, que é a que têm um índice de qualidade mais elevado, caracteriza ambos os sexos nas faixas etárias mais idosas, e está associada a mais tempo dedicado a *Dormir*, *Preparação de alimentos*, *Arranjo da casa*, *Roupa e calçado*, *Jardim*, *Lazer passivo* e *Televisão* e pouco tempo com *Escola* e *Exercício produtivo*. Na classe 3, temos o sexo feminino com idades até aos 64 anos juntamente com os homens da região Centro com idades entre os 25 e os 34 anos, que são caracterizados por dedicarem mais tempo com *Arranjo da casa*, *Roupa e calçado*, *Cuidar de crianças*, algum tempo com *Cuidar de adultos* e menos tempo dedicado em *Jardim*, *Lazer passivo* e *Exercício produtivo*.

	Cluster	Minimum	Maximum	Qj(Ct)
Dormir		225.00	905.00	
	1	215.00	895.00	98.64
	2	275.00	1255.00	43.89
	3	255.00	825.00	85.92
Escola		0.00	130.00	
	1	0.00	480.00	77.48
	2	0.00	0.00	27.74
	3	0.00	330.00	78.88
Preparação alimentos		5.00	245.00	
	1	0.00	170.00	58.25
	2	0.00	340.00	71.51
	3	10.00	290.00	90.32
Arranjo da casa		0.00	290.00	
	1	0.00	130.00	43.28
	2	0.00	320.00	90.40
	3	0.00	330.00	87.71
Roupa e Calçado		0.00	210.00	
	1	0.00	20.00	38.65
	2	0.00	290.00	90.64
	3	0.00	280.00	88.31
Jardim		0.00	300.00	
	1	0.00	230.00	89.07
	2	0.00	470.00	36.26
	3	0.00	190.00	79.83
Cuidar de crianças		0.00	180.00	
	1	0.00	150.00	91.20
	2	0.00	120.00	89.15
	3	0.00	370.00	56.63
Cuidar de adultos		0.00	80.00	
	1	0.00	20.00	60.42
	2	0.00	200.00	51.27
	3	0.00	60.00	98.76
Lazer passivo		0.00	170.00	
	1	0.00	130.00	93.26
	2	0.00	420.00	40.29
	3	0.00	110.00	79.41
Exercício produtivo		0.00	50.00	
	1	0.00	310.00	66.32
	2	0.00	50.00	100.00
	3	0.00	0.00	63.31
Televisão		0.00	500.00	
	1	0.00	480.00	96.77
	2	0.00	630.00	50.23
	3	0.00	440.00	86.32

Figura 6.84: *SCLUST* - *Região/Sexo/Grupo Etário* - Caracterização das classes pelas variáveis intervalares

Nas Tabelas 6.11 e 6.12 temos um resumo das conclusões retiradas para cada um dos níveis de agregação.

Classe	Composição	Características
1	Sexo masculino de todo o país.	Maior tempo dedicado em <i>Jardim</i> , <i>Exercício produtivo</i> e <i>Jogos</i> , e menor tempo dedicado com <i>Preparação de alimentos</i> , <i>Arranjo da casa</i> , <i>Roupa e Calçado</i> .
2	Sexo feminino de todo o país.	Maior tempo dedicado com <i>Preparação de alimentos</i> , <i>Cuidar de crianças</i> , e menor tempo dedicado com <i>Exercício produtivo</i> e <i>Jardim</i> .

Tabela 6.11: Resumo das classes formadas pelo método *SCLUST- Região/Sexo*, variáveis intervalares

Classe	Composição	Características
1	Sexo masculino até aos 64 anos (excepto homens do Centro na faixa etária dos 25 aos 34 anos).	Maior tempo dedicado com <i>Escola</i> , <i>Televisão</i> e <i>Exercício produtivo</i> e pouco tempo em <i>Cuidar de crianças</i> e <i>Lazer passivo</i> .
2	Ambos os sexos na faixa etária mais idosa.	Mais tempo gasto com <i>Dormir</i> , <i>Preparação de alimentos</i> , <i>Arranjo da casa</i> , <i>Roupa e calçado</i> , <i>Jardim</i> , <i>Lazer passivo</i> e <i>Televisão</i> e pouco tempo com <i>Escola</i> e <i>Exercício produtivo</i> .
3	Sexo feminino até aos 64 anos (incluindo homens da região Centro na faixa etária dos 25 aos 34 anos).	Mais tempo gasto com <i>Arranjo da casa</i> , <i>Roupa e calçado</i> , <i>Cuidar de crianças</i> e menos tempo dedicado em <i>Jardim</i> , <i>Lazer passivo</i> e <i>Exercício produtivo</i> .

Tabela 6.12: Resumo das classes formadas pelo método *SCLUST- Região/Sexo/Grupo Etário*, variáveis intervalares

No primeiro nível de agregação temos uma representação do país por sexo, no que diz respeito à ocupação do tempo, já no segundo nível de agregação temos uma representação baseada em grupos etários e sexo.

Na aplicação do método *SCLUST* com as descrições dadas pelas variáveis modais, utilizou-se a distância *Euclidean*, aplicada à matriz de frequências (ver fórmula 4.34).

Fixando o número de classes para uma partição em 3 classes obtém-se um critério final $\Delta = 0,186$ e a qualidade da partição de cerca de 82% (ver Figura 6.85).

A primeira classe formada é composta pelos homens das regiões do Algarve e Alentejo, a segunda classe caracteriza as mulheres do Alentejo e Algarve e os homens das regiões Norte, Centro e Lisboa e Vale do Tejo, na terceira classe temos as mulheres do Norte, Centro, e Lisboa e vale do Tejo.

```

EDITION OPTIMAL PARTITION
=====
Classe : 1 Cardinal :      2
=====
(1) RALG/M    (3) RA/M
Classe : 2 Cardinal :      5
=====
(0) RA/F    (2) RALG/F    (5) RC/M    (7) RLVT/M    (8) RN/M
Classe : 3 Cardinal :      3
=====
(4) RC/F    (6) RLVT/F    (9) RN/F

PARTITION DESCRIPTION
=====
INITIAL CRITERION      : 0.985913
FINAL CRITERION       : 0.185695
Percentage of the explained criterion :    81.17

```

Figura 6.85: *SCLUST* - *Região/Sexo* - variáveis modais

A classe com maior qualidade é a classe 1 seguida da classe 3, no entanto a que maior contribuição para a variabilidade global é a classe 2 (ver Figura 6.86).

Cluster	Size(n_i)	Q(Ct)	K(Ct)
1	2	90.73	29.08
2	5	7.77	45.30
3	3	84.75	25.62

Figura 6.86: *SCLUST* - *Região/Sexo* - Índice de qualidade e contribuição das classes

O índice de qualidade da partição em cada variável indica que as variáveis mais importantes na construção da partição são *Apressado* e *Apressado há 5anos*, (ver Figura 6.87).

Variáveis	Qj(P)
Apressado	91.70
Freq_ apressado	73.49
Apressado_5anos	84.77
Apressado_3anos	80.66
Apressado_1ano	81.68
vp_ntft	77.98
Tempo_fds	75.50
Tempo_disp	81.64

Figura 6.87: *SCLUST* - *Região/Sexo* - Índice de qualidade das variáveis modais

Se observarmos o “peso” destas variáveis em cada classe (ver Figura 6.88) verificamos que na classe 1, que representa as regiões do Algarve e Alentejo para o sexo masculino, não apresenta um sentimento de pressa. A classe 2, que caracteriza o sexo feminino das regiões do Algarve e Alentejo e o sexo masculino das regiões de Lisboa e Vale do Tejo, Centro e Norte, tanto revelam um sentimento de pressa como não revelam, já que o peso é cerca de 51% nesta categoria e dos que se sentem apressados, andam mais apressados do que há 5 anos atrás. A classe 3, formada pelos elementos do sexo feminino das regiões de Lisboa e Vale do Tejo, Centro e Norte, sentem-se na maioria apressados e mais apressados do que há 5 anos atrás.

Variáveis	Cluster				
Apressado	Set	1	2	3	<= Cluster
		0.49	0.64	0.49	0.41 Não
		0.51	0.36	0.51	0.59 Sim
Apressado-5anos	Set	1	2	3	<= Cluster
		0.49	0.64	0.49	0.41 NA
		0.31	0.20	0.32	0.36 Anda mais apressado
		0.14	0.12	0.14	0.16 Na mesma
		0.06	0.04	0.06	0.06 Menos apressado
		0.00	0.00	0.00	0.00 NS/NR

Figura 6.88: *SCLUST* - *Região/Sexo* - Caracterização das classes pelas variáveis modais

Para o segundo nível de agregação com as variáveis modais, fixou-se o número de classes em 3 e obteve-se um critério final de $\Delta = 74,57$ e uma qualidade da partição de 82% (ver Figura 6.89). Verifica-se uma divisão baseada mais no *Grupo Etário*, do que no *Sexo* ou *Região*.

Na classe 1 temos a maioria dos grupos com idades entre os 25 e os 54 anos, de ambos os sexos, juntamente com a faixa etária mais jovem de Lisboa e Vale do Tejo. Na segunda classe juntam-se as faixas etárias dos 15 aos 24 anos de ambos os sexos, excepto Lisboa e Vale do Tejo, com a faixa etária dos 55 aos 64 anos. Nesta classe também estão presentes os homens do Algarve e Alentejo com idades entre os 35 e os 54 anos. Na terceira classe estão os grupos mais idosos de ambos os sexos.

```

EDITION OPTIMAL PARTITION
=====
Classe : 1 Cardinal : 19
=====
(4) RA / F / 35_54      (8) RALG / F / 35_54
(12) RALG / F / 25_34  (13) RA / F / 25_34
(18) RC / F / 35_54    (20) RC / M / 35_54
(24) RC / F / 25_34    (27) RC / M / 25_34
(28) RLVT / F / 25_34  (29) RLVT / F / 35_54
(32) RLVT / M / 25_34  (35) RLVT / F / 15_24
(36) RLVT / M / 35_54  (37) RLVT / M / 15_24
(39) RN / F / 35_54    (40) RN / F / 25_34
(43) RN / M / 35_54    (47) RN / M / 25_34
(49) RA / M / 25_34

Classe : 2 Cardinal : 18
=====
(2) RA / F / 55_64      (3) RALG / F / 55_64
(5) RA / M / 35_54      (9) RALG / F / 15_24
(10) RALG / M / 35_54   (15) RALG / M / 55_64
(16) RA / F / 15_24     (21) RC / F / 15_24
(23) RC / F / 55_64     (25) RC / M / 55_64
(26) RC / M / 15_24     (30) RLVT / F / 55_64
(33) RLVT / M / 55_64   (41) RN / F / 55_64
(42) RN / F / 15_24     (44) RN / M / 55_64
(46) RN / M / 15_24     (48) RA / M / 15_24

Classe : 3 Cardinal : 13
=====
(0) RA / F / 65 ou mais (1) RALG / M / 15_24
(6) RALG / M / 65 ou mais (7) RALG / F / 65 ou mais
(11) RALG / M / 25_34    (14) RA / M / 65 ou mais
(17) RA / M / 55_64      (19) RC / F / 65 ou mais
(22) RC / M / 65 ou mais (31) RLVT / F / 65 ou mais
(34) RLVT / M / 65 ou mais (38) RN / M / 65 ou mais
(45) RN / F / 65 ou mais ]

PARTITION DESCRIPTION
=====
INITIAL CRITERION : 400.000000
FINAL CRITERION : 74.575873
Percentage of the explained criterion : 81.36

```

Figura 6.89: *SCLUST* - *Região/Sexo/Grupo Etário* - variáveis modais

A classe com maior qualidade é a classe 3 a classe com maior contribuição para a variabilidade global é a classe 1 (ver Figura 6.90).

Cluster	Size(n _i)	Q(Ct)	K(Ct)
1	19	80.35	45.26
2	18	4.69	41.77
3	13	95.05	12.97

Figura 6.90: *SCLUST* - *Região/Sexo/Grupo Etário* - Índice de qualidade e contribuição das classes

Tomando em consideração o índice de qualidade da partição em cada variável (ver Figura 6.91), temos agora as variáveis *Apressado* e *Apressado há 1ano* como as variáveis mais importantes na partição.

Analisando em pormenor o “peso” de cada categoria para as variáveis com maior importância na construção da partição (ver Figura 6.92), verificamos que a classe 1, que caracteriza a faixa etária dos 25 aos 54 anos para ambos os sexos e a faixa etária mais jovem de Lisboa e Vale do Tejo, sente-se na maioria apressada e na mesma comparando com a situação face ao ano anterior. A classe 2, que representa as faixas etárias mais jovens (excepto Lisboa e Vale do Tejo) e as faixas etárias dos 55 aos 64 anos, estão numa situação “média”, ou seja, são tantos os elementos que sentem

Variáveis	Qj(P)
Apressado	91.98
Freq_ apressado	80.02
Apressado_5anos	77.46
Apressado_3anos	76.06
Apressado_1ano	83.48
vp_ntft	80.77
Tempo_fds	80.82
Tempo_disp	80.25

Figura 6.91: *SCLUST* - *Região/Sexo/Grupo Etário* - Índice de qualidade das variáveis modais

apressados como os que não se sentem e comparando com a situação face ao ano anterior sentem-se em geral na mesma. A classe 3, que representa a faixa etária mais idosa não se sente em geral apressada.

Variáveis	Set	Cluster				
Apressado		1	2	3		<= Cluster
	0.51	0.30	0.51	0.80	Não	
	0.49	0.70	0.49	0.20	Sim	
Apressado-1ano		1	2	3		<= Cluster
	0.51	0.30	0.51	0.80	NA	
	0.28	0.37	0.29	0.12	Na mesma	
	0.18	0.29	0.17	0.04	Anda mais apressado	
	0.03	0.04	0.03	0.03	Menos apressado	
	0.00	0.00	0.00	0.00	NS/NR	

Figura 6.92: *SCLUST* - *Região/Sexo/Grupo Etário* - Caracterização das classes pelas variáveis modais

Nas Tabelas 6.13 e 6.14 apresenta-se um resumo das conclusões retiradas para cada um dos níveis de agregação com as descrições dadas pelas variáveis modais.

Classe	Composição	Características
1	Sexo masculino das regiões do Algarve e Alentejo.	Não se sentem na maioria apressados.
2	Sexo feminino das regiões do Algarve e Alentejo e sexo masculino das regiões de Lisboa e Vale do Tejo, Centro e Norte.	São tantos os elementos que sentem apressados como os que não se sentem, e dos que sentem, andam mais apressados do que há cinco anos atrás.
3	Sexo feminino das regiões de Lisboa e Vale do Tejo, Centro e Norte.	Sentem-se na maioria apressados e mais apressados do que há 5 anos atrás.

Tabela 6.13: Resumo das classes formadas pelo método *SCLUST*- *Região/Sexo*, variáveis modais

Classe	Composição	Características
1	Maioria dos grupos com idades entre os 25 e os 54 anos com a faixa mais jovem de Lisboa e Vale do Tejo.	Sente-se na maioria apressada, e na mesma comparando com a situação face ao ano anterior.
2	Faixas etárias mais jovens (excepto Lisboa e Vale do Tejo) e faixas etárias dos 55 aos 64 anos.	São tantos os elementos que sentem apressados como os que não se sentem e comparando com a situação face ao ano anterior sentem-se na mesma.
3	Faixa etária mais idosa.	Não se sentem em geral apressados.

Tabela 6.14: Resumo das classes formadas pelo método *SCLUST- Região/Sexo/Grupo Etário*, variáveis modais

No primeiro nível de agregação obtém-se uma classificação do país por grupos de regiões, no que diz respeito ao sentimento de pressa, já no segundo nível de agregação temos uma representação baseada em grupos etários.

6.2.4 Comparação dos resultados obtidos métodos de classificação

Através do método DIV obtemos uma divisão em classes, segundo as variáveis que identificam como discriminantes. A caracterização do país baseada na região e sexo, quanto à forma como o tempo é ocupado, leva a concluir que as mulheres de Portugal dedicam menos tempo ao *Exercício produtivo* e *Jogos*, excepto no Alentejo, região onde as mulheres dedicam um pouco mais de tempo ao *Exercício produtivo*. Pelo contrário, os homens dedicam bastante tempo aos *Jogos*. Quanto ao sentimento de pressa, as regiões mais a Sul do país não se sentem na maioria apressadas, contrariamente às regiões mais a Norte.

Se tivermos em consideração também o *Grupo Etário*, observa-se uma divisão do país em 5 classes, que representam os grupos etários femininos mais jovens, os grupos etários femininos mais idosos, os grupos etários masculinos mais jovens, os grupos etários masculinos mais idosos, e os restantes grupos. As classes masculinas e femininas distinguem-se no geral, pela dedicação de mais tempo que as classes femininas, a actividades como *Roupa e Calçado* e *Preparação de alimentos*, a outra classe distingue-se das restantes pelo menor tempo gasto com *Dormir* e *Escola*. No sentimento de pressa as classes não produzem conclusões tão evidentes, mas mostram que os únicos

grupos da região de Lisboa e Vale do Tejo que não se sentem apressados, são os mais idosos e que no Algarve só as mulheres com idades entre os 25 e os 34 anos se sentem apressadas.

Com os métodos de classificação em Pirâmide e Hierarquia, não temos uma divisão segundo as variáveis, mas sim uma generalização das classes, que produzem novos conceitos. Para a ocupação do tempo no primeiro nível, a Hierarquia sugere uma oposição entre homens e mulheres, com as mulheres a despenderem mais tempo com *Trabalho* e *Escola* e actividades relacionadas com a casa e os homens a dispenderem mais tempo com *Jogos*, *Jardim*, *Construção* e *Convívio*, quando comparados com as mulheres. Com a Pirâmide obtemos a mesma generalização, mas que é atenuada pelas regiões do Norte e Alentejo, onde os homens e mulheres são mais “próximos”. Para o sentimento de pressa, a Hierarquia sugere uma oposição dos homens do Algarve e Alentejo que não se sentem apressados, com o resto dos grupos que se sentem na maioria apressados. O método em Pirâmide reúne os homens da região Norte aos do Algarve e Alentejo, verificando-se o mesmo comportamento do que com o método em Hierarquia.

No segundo nível, de uma forma muito geral, as conclusões retiradas com o método em Pirâmide, sugerem um sentimento de pressa associado a quase todo o país, excepto os grupos etários mais elevados e os grupos mais jovens masculinos do Algarve. Com a representação em Pirâmide, as conclusões são muito semelhantes, apenas há que salientar que os homens mais jovens do Alentejo e Norte também não se sentem na maioria apressados e que as faixas etárias mais idosas estão mais associadas a um não sentimento de pressa.

O método *SCLUST* também visa o agrupamento dos elementos em classes, mas tem como objectivo determinar directamente partições em classes homogéneas. Este método fornece ainda índices que ajudam na interpretação das classes formadas. Com a agregação por *Região/Sexo*, verificamos que o país se divide em sexo masculino e feminino, quando queremos caracterizar a forma como ocupa o seu tempo. Às mulheres estão associadas as actividades como *Preparação de alimentos* e *Cuidar de crianças* e menos tempo dedicado com o *Exercício produtivo* e *Jardim*, já os homens são mais identificados pelas actividades como *Exercício produtivo* e *Jardim*. Quando analisamos o sentimento de pressa, temos uma divisão do país em 3 classes, em que os homens do Algarve e Alentejo não se sentem na maioria apressados, as mulheres das regiões a norte do país revelam um sentimento de pressa generalizado e os restantes grupos dividem-se de forma equilibrada quanto ao sentimento de pressa.

Com a introdução do *Grupo Etário*, observa-se uma classe masculina com idades até

aos 64 anos excepto na região Centro com idades entre os 25 e os 34 anos, que se caracteriza pelas actividades *Escola*, *Televisão* e *Exercício produtivo*, uma classe feminina que inclui os homens da região Centro com idades entre os 25 e os 34 anos, que dedica mais tempo com *Arranjo da casa*, *Roupa e Calçado*, *Cuidar de crianças*, e por último uma classe mais idosa que dedica mais tempo com *Dormir* e actividades como *Preparação de alimentos*, *Arranjo da casa*, *Jardim*, *Televisão* e *Lazer passivo*. Para o sentimento de pressa também se formam 3 classes, que se caracterizam pelo *Grupo Etário*, estando uma classe associada a um não sentimento de pressa que são os mais idosos, outra formada pelos grupos com idades entre os 25 e os 54 anos, aos quais se junta a faixa etária mais de Lisboa e Vale do Tejo, que se sentem na maioria apressados, e por último uma classe com a faixa etária mais jovem (excepto Lisboa e Vale do Tejo) e as faixas dos 55 aos 64 anos, que tanto se sentem apressados como não.

Podemos concluir que o método DIV é mais útil quando se pretende encontrar variáveis que discriminem as classes, enquanto que os métodos em Pirâmide e Hierarquia são mais adequados para generalizar as classes formadas, havendo contudo mais “rigidez” na separação das classes com a Hierarquia do que com a Pirâmide, sendo esta última útil para detectar elementos que se encontrem na “fronteira”. O método *SCLUST* é semelhante à classificação não-hierárquica clássica, mas generalizado aos dados simbólicos, e permite identificar grupos que se caracterizem por certas variáveis.

Capítulo 7

Considerações finais

Neste trabalho, foi efectuado um estudo da teoria da Análise de Dados Simbólicos, com o objectivo de pôr em evidência a sua pertinência na análise de estatísticas oficiais. Em particular, foi estudado como caso prático o Inquérito à Ocupação do Tempo do Instituto Nacional e Estatística (INE). Do ponto de vista teórico, foi efectuado um grande esforço no sentido de harmonizar notação e vocabulário dispersos, de forma a obter uma formalização única e coerente. Posteriormente o trabalho incidiu na aplicação de estatísticas descritivas e de alguns métodos de classificação hierárquica e não-hierárquica aos dados simbólicos obtidos pela agregação dos dados clássicos em dois níveis. As estatísticas descritivas permitiram uma análise prévia dos dados e a sua visualização em representações sob a forma de estrela. Com os métodos de classificação identificaram-se grupos com características próprias na forma como ocupam o seu tempo diário, no sentimento de pressa e disponibilidade de tempo. Como seria de esperar, os métodos de classificação não produzem os mesmos resultados nem identificam os mesmos grupos, pois não têm os mesmos critérios de comparação e de formação de classes.

A grande vantagem na utilização das técnicas de Análise de Dados Simbólicos é a possibilidade de se ter em consideração a variabilidade interna dos grupos que é utilizada para análises posteriores, já que a forma habitual de comparar dados agregados passa pela comparação das médias, medianas ou modas, calculadas em cada grupo. A agregação dos dados permite ainda a junção de questionários diferentes, desde que se utilizem as mesmas variáveis na construção dos grupos. Os níveis de agregação dependem dos objectos em estudo e daquilo que o utilizador pretende analisar.

Como complemento deste trabalho, para além das estatísticas descritivas e métodos

de classificação, poder-se-iam aplicar outras técnicas de Análise de Dados, como a análise factorial ou a análise em componentes principais. Seria ainda interessante, aplicar os mesmos níveis de agregação a outros inquéritos com indivíduos distintos (por exemplo o inquérito ao Emprego) e agregar a informação, sendo assim possível reunir informação proveniente de fontes diversas.

Bibliografia

- [Baker e Hubert (1976)] Baker, F.B., Hubert, L.J. 1976. Measuring the Power of Hierarchical Cluster Analysis. *Journal of the American Statistical Association*, 70, 31-38.
- [Bertrand e Diday (1985)] Bertrand, P. e Diday, E. 1985. A Visual Representation of Compatibiliy Between an Order and a Dissimilarity Index: The Pyramids. *Computational Statistics Quarterly*, 2(1), 31-42.
- [Bertrand e Goupil (2000)] Bertrand, P. e Goupil, F. 2000. Descriptive Statistics for Symbolic Data. In: *Analysis of Symbolic Data*, Bock, H.-H. e Diday, E. (Eds.). Springer, Heidelberg, 6, 106-124.
- [Billard (2007)] Billard, L. 2007. Dependencies and Variation Components of Symbolic Interval-Valued Data. In: *Selected Contributions in Classification and Data Analysis*, Brito, P., Bertrand, P., Cucumel, G., De Carvalho, F.A.T. (Eds.). Series “Studies in Classification and Data Analysis, and Knowledge Organization”, Springer, Heidelberg, 3-12.
- [Billard e Diday (2003)] Billard, L. e Diday, E. 2003. From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis. *Journal of the American Statistical Association*, 98, 470-487.
- [Bock e Diday (2000)] Bock, H.-H. e Diday, E. 2000. *Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer, Heidelberg.
- [Bock e Diday (2000a)] Bock, H.-H. e Diday, E. (Eds.) 2000. Symbolic Objects. In: *Analysis of Symbolic Data*, Bock, H.-H. e Diday, E. (Eds.). Springer, Heidelberg, 4, 54-77.
- [Breiman *et al.* (1984)] Breiman, L., Friedman, J.H., Olshen, R.A e Stone, C.J. 1984. *Classification and Regression Trees*. Wadsworth, Belmont, Calif.

- [Brito (1991)] Brito, P. 1991. Analyse des Données Symboliques. Pyramides d'Héritage. Tese de Doutorado. Universidade de Paris - IX Dauphine.
- [Brito (1994)] Brito, P. 1994. Use of the Pyramids in Symbolic Data Analysis. In: New Approaches in Classification and Data Analysis, Diday, E. *et al.* (Eds.). Springer, Heidelberg, 378-386.
- [Brito (1998)] Brito, P. 1998. Symbolic Clustering of Probabilistic Data. In: Advances in Data Science and Classification, Rizzi, A., Vichi, M. e Bock, H.-H. (Eds.). Springer, Berlin.
- [Brito e De Carvalho (1999)] Brito, P. e De Carvalho, F.A.T. 1999. Symbolic Clustering in the Presence of Hierarchical Rules. In: Studies and Research, Proceedings of the Conference on Knowledge Extraction and Symbolic Data Analysis (KESDA'98). Eurostat, Luxemburgo, 385-390.
- [Brito e De Carvalho (2002)] Brito, P., De Carvalho, F.A.T. 2002. Symbolic Clustering of Constrained Probabilistic Data. In: Exploratory Data Analysis in Empirical Research, Optiz, O., Schwaiger, M. (Eds.). Series "Studies in Classification, Data Analysis and Knowledge Organization". Springer, Heidelberg, 12-21.
- [Brito e De Carvalho (2008)] Brito, P., De Carvalho, F.A.T. 2008. Hierarchical and Pyramidal Clustering. In: Symbolic Data Analysis and the SODAS Software, Diday, E. e Noirhomme-Fraiture, M. (Eds.). Wiley.
- [Brito e Diday (1989)] Brito, P., Diday, E. 1989. Pyramidal Representation of Symbolic Objects. In: Knowledge, Data and Computer-Assisted Decisions, NATO ASI series, Schader, M. e Gaul, W. (Eds.). Springer, Heidelberg, 3-16.
- [Calínski e Harabasz (1974)] Calínski, T., Harabasz, J. 1974. A Dendrite Method for Cluster Analysis. Communications in Statistics 3, 1-27.
- [Calvo (2000)] Calvo, P. 2000. Applications of Symbolic Objects in Official Statistics. EUSTAT. País Basco.
- [Chavent (1997)] Chavent, M. 1997. Analyse des Données Symboliques. Une Méthode Divisive de Classification. Tese de Doutorado. Universidade de Paris - IX Dauphine.
- [Chavent (1998)] Chavent, M. 1998. A Monothetic Clustering Algorithm. Pattern Recognition Letters, 19, 989-996.

- [Chavent (2000)] Chavent, M. 2000. Criterion-Based Divisive Clustering for Symbolic Data. In: Analysis of Symbolic Data, Bock, H.-H. e Diday, E. (Eds.). Springer, Heidelberg, 11, 299-311.
- [Chavent e Lechevallier (2002)] Chavent, M. e Lechevallier, Y. 2002. Dynamical Clustering Algorithm of Interval Data: Optimization of an Adequacy Criterion Based on Hausdorff Distance. In: Classification, Clustering and Data Analysis, Sokolowsky and Bock, H.-H. (Eds.). K. Jaguja, Springer, Heidelberg, 53-59.
- [Chavent *et al.* (2003)] Chavent, M., De Carvalho, F.A.T., Lechevallier, Y., Verde, R. 2003. Trois Nouvelles Méthodes de Classification Automatique de Données Simboliques de Type Intervalle. *Revue de Statistique Appliquée*, 4, 5-9.
- [Cazes *et al.* (1990)] Cazes, P., Chouakria, A., Diday, E. e Schecktmann, Y. 1990. Extension de la Analyse en Composantes Principales à Données Intervalles. *Revue de Statistique Appliquée*, 45(3).
- [Celeux *et al.* (1988)] Celeux, G., Diday, E., Govaert, G., Lechevallier, Y., Ralambondainy, H. 1988. Classification Automatique des Données, Environnement Statistique et Informatique, Dunod, Gauthier-Villards, Paris.
- [De Carvalho (1998)] De Carvalho, F.A.T. 1998. New Metrics for Constrained Boolean Symbolic Objects. Proceedings of the Conference on Knowledge Extraction and Symbolic Data Analysis (KESDA'98). Eurostat, Luxemburgo.
- [De Carvalho *et al.* (2008)] De Carvalho, F.A.T., Lechevallier, Y., Verde, R. 2008. Clustering Methods in Symbolic Data Analysis. In: Symbolic Data Analysis and the SODAS Software, Diday, E. e Noirhomme-Fraiture, M. (Eds.). Wiley.
- [Diday (1971)] Diday, E. 1971. La méthode des Nuées dynamiques. *Revue de Statistique Appliquée*, 19(2), 19(34).
- [Diday (1984)] Diday, E. 1984. Une Représentation Visuelle des Classes Empiétantes: Les Pyramides. RR-291, INRIA, Rocquencourt, França.
- [Diday (1986)] Diday, E. 1986. Orders and Overlapping Clusters by Pyramids. In: Multidimensional Data Analysis, J. De Leeuw *et al.* (Eds.). Leiden:DSWO Press, 201-234.
- [Diday (1993)] Diday, E. 1993. An Introduction to Symbolic Data Analysis. RR-1936, INRIA, Rocquencourt, França.

- [Diday e Noirhomme-Fraiture (2008)] Diday, E. e Noirhomme-Fraiture, M. (Eds.) 2008. Symbolic Data Analysis and the SODAS Software. Wiley.
- [Gowda e Diday (1992)] Gowda, C. e Diday, E. 1992. Symbolic Clustering Using a New Dissimilarity Measure. IEEE Trans. Syst. Man and Cybernetics, 22(2), 368-378.
- [Goupil (2000)] Goupil, F. 2000. Elementary Statistics of Symbolic Data. École SODAS, Ceremade, Universidade de Dauphine -IX, Paris.
- [Hubert e Levin (1976)] Hubert, L., Levin, J.R. 1976. A General Statistical Framework for Assessing Categorical Clustering in Free Recall. Psychological Bulletin, 83, 1072-1080.
- [Lima e De Carvalho (2003)] Lima, N. e De Carvalho, F.A.T. 2003. Analyzing Administrative Managements Through the Symbolic Approach. Electronic Journal of Symbolic Data Analysis, 1(1).
- [Macqueen (1967)] Macqueen, J. 1967. Some Methods for Classification Analysis of Multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1, 281-297. Berkeley: University of California Press.
- [Marcelo (2001)] Marcelo, C. 2001. Especificidades do Emprego em Portugal: A Contribuição da Análise Simbólica de Dados. Revista de Estudos Regionais de Lisboa e Vale do Tejo, 2.
- [Marcelo (2002)] Marcelo, C. 2002. Application of Symbolic Data Analysis to Official Statistical Data. IFCS'2002, Workshop on Symbolic Data Analysis, Cracóvia.
- [Mas e Olaeta (2002)] Mas, M. e Olaeta, H. 2002. Symbolic Analysis of the Time Use in the Basque Country. EUSTAT, País Basco.
- [Matusita (1951)] Matusita, K. 1951. Decision Rules Based on Distance for Problems of Fit, Two Samples and Estimation. Ann. Math. Stat. 3, 1-30.
- [Rodriguez (2000)] Rodriguez, O. 2000. Classification et Modèles Linéaires en Analyse des Données Symboliques. Tese de Doutorado. Universidade de Paris - IX Dauphine.
- [Soussi (2003)] Soussi, S. 2003. Analyse d'une Base de Données Pays par le Logiciel SODAS. Ceremade, Universidade de Paris - IX Dauphine.

- [Stéphan *et al.* (2000)] Stéphan, V., Hébrail, G. e Lechevallier, Y. 2000. Generation of Symbolic Objects from Relational Databases. In: Analysis of Symbolic Data, Bock, H.-H. e Diday, E. (Eds.). Springer, Heidelberg, 5, 78-105.

Apêndice A

Notação

Ω	conjunto de todos os indivíduos ω
n	cardinal de Ω
E	conjunto dos elementos u (classes ou objectos)
k	cardinal de E
$\tilde{Y}$	variável clássica
$\mathcal{Y}$	domínio da variável $\tilde{Y}$
$P(\mathcal{Y})$	todos os subconjuntos $A \subseteq \mathcal{Y}$
$M(\mathcal{Y})$	conjunto de todas as distribuições no conjunto $\mathcal{Y}$
$X = (x_{ij})_{n \times p}$	matriz clássica com n linhas $i = 1, \dots, n$ e p colunas $j = 1, \dots, p$
Y	variável simbólica
B_j	conjunto onde a variável simbólica Y_j toma valores
$\Xi = (\xi_{ij})_{k \times p}$	matriz simbólica com k linhas $i = 1, \dots, k$ e p colunas $j = 1, \dots, p$
$B = B_1 \times \dots \times B_p$	espaço das descrições
$d = (d_1, \dots, d_p)$	descrição ou vector descrição com $d_j \in B_j$