

**SUMARIZAÇÃO PARA DIFERENTES HORIZONTES
TEMPORAIS**

APLICAÇÃO A DADOS DA REDE ELÉCTRICA NACIONAL

por

Elsa Manuela Pereira de Aguiar

Dissertação de Mestrado em Análise de Dados e Sistemas de Apoio à
Decisão

Orientada por

Professor Doutor João Manuel Portela da Gama

2010

NOTA BIOGRÁFICA

Elsa Manuela Pereira de Aguiar nasceu em Vila Nova de Gaia em 1987. Concluiu o 1º ciclo de estudos em 2008, licenciando-se em Gestão pela Universidade de Aveiro. Nesse mesmo ano ingressou na Faculdade de Economia da Universidade do Porto, iniciando o 2º ciclo de estudos no Mestrado em Análise de Dados e Sistemas de Apoio à Decisão.

AGRADECIMENTOS

A elaboração desta dissertação não teria sido possível sem a colaboração e apoio de algumas pessoas, a quem aqui gostaria de agradecer.

Em primeiro lugar, ao meu orientador e director de mestrado, Professor Doutor João Gama pela disponibilidade, apoio e motivação sempre demonstrados.

Seguidamente, a todo o corpo docente do mestrado que com os seus ensinamentos me permitiram adquirir conhecimentos que me foram úteis neste trabalho.

Não esquecendo ainda o enquadramento do trabalho no projecto Knowledge Discovery from Ubiquitous Data Streams (PTDC/EIA-EIA/098355/2008).

À Márcia, colega de licenciatura e mestrado, pelos conhecimentos partilhados e pela ajuda na revisão deste documento.

Finalmente, aos meus pais e amigos pela preocupação, incentivo e apoio incondicionais.

RESUMO

Esta dissertação tem por objecto de estudo a sumarização de dados para diferentes horizontes temporais. Para esse efeito foram utilizados dados relativos ao consumo de energia eléctrica em Portugal nos anos de 2004 e 2005, recolhidos em períodos fixos de tempo.

A principal metodologia utilizada que serve como ponto de partida para a análise desenvolvida, denominada de histogramas por níveis, consiste na agregação dos dados originalmente recolhidos em sucessivos níveis com unidades temporalmente crescentes. Desta forma, pretende-se detectar padrões de consumo característicos no que respeita às diferentes granularidades consideradas.

No decorrer do trabalho, recorrendo à aplicação do método de agrupamento das k-Médias aos resultados de cada nível do histograma e à correlação dos mesmos, foi possível caracterizar comparativamente diferentes perfis de consumo de energia eléctrica, acompanhados do erro cometido por utilizar os perfis em vez dos dados originais.

Explorando os perfis diários resultantes do segundo nível de agregação do histograma e, interpretando a aplicação do método das k-Médias, foi estudada a distribuição de dias da semana entre dois a sete grupos. Os diferentes agrupamentos de perfis diários foram ainda analisados em termos de comportamento de consumo verificado em dias representativos.

Com este trabalho pretende-se aliar a utilização de metodologias com construções simples à obtenção de conhecimento relevante para o funcionamento dos mercados de energia eléctrica.

Palavras-Chave: Agrupamento, Histogramas por Níveis, Perfis, Sumarização

ABSTRACT

This thesis aims to study the summarization of data for different time horizons. For this purpose was used portuguese electricity supply data, which is collected at fixed time points and is available for years 2004 and 2005.

The main adopted methodology that serves as a starting point for the developed analysis and is called layered histograms, consists in the aggregation of originally collected data in successive levels with increasing time-units. Thus, we intend to detect defining patterns in each considered data granularity.

By applying k-Means clustering to the results of each level of the histogram and by computing the correlations between them, was comparatively characterized distinct electricity supply profiles, together with the error associated to the use of the profiles instead of the original data.

We also study the distribution of the weekdays between two and seven clusters, through the exploration of the daily profiles that result from the application of k-Means clustering to the second level of aggregation of the histogram. The different clusters of the daily profiles are further analyzed in terms of the consumption behavior verified in representative days.

The main goal of this work is to combine the use of simple methodologies to the discovery of relevant and useful knowledge about the studied phenomena.

Keywords: Clustering, Layered Histograms, Profiles, Summarization

CONTEÚDO

Nota Biográfica.....	ii
Agradecimentos.....	iii
Resumo	iv
Abstract	v
1 Introdução	1
1.1 Motivação	1
1.2 Descrição do Problema.....	1
1.3 Principais Contribuições da Dissertação	2
1.4 Estrutura da Dissertação	3
2 Enquadramento Teórico	4
2.1 Tratamento de Dados Temporais	4
2.1.1 Janelas Temporais	5
2.1.2 Histogramas	7
2.2 Estudo de Perfis	9
2.2.1 Discriminação de Tipos de Consumidores	14
2.2.2 Aplicação para Previsão	15
2.2.3 Integração com a Gestão	16
2.2.4 Integração com a Tomada de Decisão.....	17

3	Metodologias.....	18
3.1	Análise Exploratória.....	18
3.1.1	Análise Preliminar.....	18
3.1.2	Preparação dos Dados	19
3.1.3	Estudo da Auto-Correlação dos Dados	19
3.2	Aplicação de Histograma Por Níveis aos Dados	21
3.2.1	Método k-Médias	24
3.2.2	Coeficiente de Silhueta.....	26
3.2.3	Agrupamento a partir das Correlações.....	28
3.2.4	Apresentação dos Perfis	29
3.2.5	Caracterização dos Perfis	32
3.3	Distribuição dos Dias da Semana.....	33
3.3.1	Dias Representativos	34
3.4	Ferramentas.....	35
4	Estudo do Consumo de Energia Eléctrica	36
4.1	Análise Exploratória de Dados	36
4.1.1	Análise Preliminar.....	36
4.1.2	Preparação dos Dados	38
4.1.3	Estudo da Auto Correlação dos dados.....	39
4.2	Aplicação de Histograma Por Níveis aos Dados	43
4.2.1	Perfis Horários (Nível 1)	43
4.2.2	Perfis 6 Horas (Nível 2).....	48
4.2.3	Perfis Diários (Nível 3)	52
4.2.4	Perfis Dias Semana (Nível 4)	55

4.2.5	Perfis Semana e Fim-de-Semana (Nível 5)	58
4.2.6	Perfis Semanais (Nível 6)	61
4.2.7	Perfis Mensais (Nível 7)	64
4.2.8	Perfis Trimestrais (Nível 8)	66
4.2.9	Perfis Semestrais (Nível 9)	69
4.3	Refinar perfis	69
4.4	Distribuição de Dias da Semana	75
4.4.1	2 Grupos	75
4.4.2	3 Grupos	77
4.4.3	4 Grupos	80
4.4.4	5 Grupos	82
4.4.5	6 Grupos	85
4.4.6	7 Grupos	88
4.4.7	Resumo de Resultados.....	91
4.5	Dias Representativos dos Grupos	93
5	Considerações Finais	99
5.1	Conclusões.....	99
5.2	Trabalho Futuro.....	101
	Apêndices.....	102
	Apêndice A – Calendários de 2004 e 2005.....	102
	Apêndice B – Listagem de Feriados.....	104
	Apêndice C – Distribuição Mensal por Dias da Semana (2 Grupos)	105
	Referências bibliográficas	106

ÍNDICE DE FIGURAS

Figura 2.1 – Representação das <i>Landmark e Sliding Windows</i>	5
Figura 2.2 – Tipos de <i>Tilted Time Windows</i>	6
Figura 2.3 – Múltiplas funções do sistema	11
Figura 2.4 – Exemplo de agrupamento das células.....	12
Figura 2.5 – Esquema representativo do processo de estimação de consumo.....	14
Figura 2.6 – Exemplo da Segmentação de Consumidores	16
Figura 3.1 – Exemplo Auto-Correlação para Perfis Mensais (2004).....	20
Figura 3.2 – Esquema Exemplificativo de um Histograma por Níveis.....	21
Figura 3.3 – Exemplo de visualização dos dados agregados (nível 3).....	24
Figura 3.4 – Coeficiente de Silhueta Médio por Número de Grupos - Perfis 6 Horas (2004)	27
Figura 3.5 – Exemplo de Análise do Coeficiente de Silhueta Médio por Número de Grupos - Perfis 6 Horas (2004).....	27
Figura 3.6 – Exemplo de Agrupamento k-Médias para Perfis Horários	30
Figura 3.7 – Exemplo Perfis Médios Horários (2 Grupos).....	31
Figura 3.8 – Exemplo da Distribuição Diária por Grupo (2005).....	34
Figura 3.9 – Exemplo dos Dias Representativos de 2004 (3 Grupos)	35
Figura 4.1 – Sumário Global do conjunto de dados ren.....	37
Figura 4.2 – Primeiras e últimas linhas do conjunto de dados ren.....	37
Figura 4.3 – Estatísticas Descritivas Básicas da Variável Valor Observado.....	37
Figura 4.4 – Evolução do Consumo de Energia Eléctrica (2004 e 2005)	38
Figura 4.5 – Estatísticas Descritivas Básicas do Valor Observado (2004 e 2005)	39
Figura 4.6 – Exemplo Auto-Correlação de Perfis Horários	39

Figura 4.7 – Exemplo Auto-Correlação de Perfis 6 Horas.....	40
Figura 4.8 – Exemplo Auto-Correlação de Perfis Diários.....	40
Figura 4.9 – Exemplo Auto-Correlação de Perfis Dias da Semana	41
Figura 4.10 – Exemplo Auto-Correlação de Perfis Semana e Fim-de-Semana.....	41
Figura 4.11 – Exemplo Auto-Correlação de Perfis Mensais	42
Figura 4.12 – Exemplo de Agregação para o Nível 1	43
Figura 4.13 – Evolução do Consumo de Energia Eléctrica em Janeiro de 2004 (Horária)	44
Figura 4.14 – Coeficiente de Silhueta Médio por Número de Grupos - Perfis Horários (2004)	45
Figura 4.15 – Análise do Coeficiente de Silhueta Médio por Número de Grupos - Perfis Horários (2004).....	45
Figura 4.16 – Coeficiente de Silhueta Médio Global.....	45
Figura 4.17 – Agrupamento k-Médias para Perfis Horários.....	46
Figura 4.18 – Perfis Médios Horários	47
Figura 4.19 - Evolução do Consumo de Energia Eléctrica em Agosto de 2004 (Períodos de 6 Horas).....	48
Figura 4.20 – Agrupamento k-Médias para Perfis 6 Horas.....	50
Figura 4.21 – Perfis Médios 6 Horas.....	51
Figura 4.22 – Evolução do Consumo de Energia Eléctrica em Agosto de 2005 (Diária)	52
Figura 4.23 – Agrupamento k-Médias para Perfis Diários.....	53
Figura 4.24 – Perfis Médios Diários	54
Figura 4.25 – Evolução do Consumo de Energia Eléctrica em Junho de 2005 (Dias da Semana)	55
Figura 4.26 – Agrupamento k-Médias para Perfis Dias da Semana	56

Figura 4.27 – Perfis Médios Dias da Semana	57
Figura 4.28 – Evolução do Consumo de Energia Eléctrica em 2004 e 2005 (Semana e Fim-de-Semana)	58
Figura 4.29 – Agrupamento k-Médias para Perfis Semana e Fim-de-Semana	59
Figura 4.30 – Perfis Médios Semana e Fim-de-Semana	60
Figura 4.31 – Evolução do Consumo de Energia Eléctrica - 2004 e 2005 (Semanal)	61
Figura 4.32 – Agrupamento k-Médias para Perfis Semanais	62
Figura 4.33 – Perfis Médios Semanais	63
Figura 4.34 – Evolução do Consumo de Energia Eléctrica - 2004 e 2005 (Mesal)	64
Figura 4.35 – Agrupamento k-Médias para Perfis Mensais	65
Figura 4.36 – Perfis Médios Mensais	66
Figura 4.37 – Evolução do Consumo de Energia Eléctrica – 2004 e 2005 (Trimestral).67	
Figura 4.38 – Agrupamento k-Médias para Perfis Trimestrais	67
Figura 4.39 – Perfis Médios Trimestrais	68
Figura 4.40 – Resultados de Agregação do Nível 9 (2004 e 2005)	69
Figura 4.41 – Exemplo do Desdobramento dos Perfis Diários (Distâncias Euclidenas) 73	
Figura 4.42 – Exemplo do Desdobramento dos Perfis Diários (Correlações).....	74
Figura 4.43 – Distribuição Dias da Semana (2 Grupos) – Distâncias Euclideanas (2004)	76
Figura 4.44 – Distribuição Dias da Semana (2 Grupos) – Correlações (2004)	77
Figura 4.45 – Distribuição Dias da Semana (3 Grupos) – Distâncias Euclideanas (2005)	78
Figura 4.46 – Distribuição Dias da Semana (3 Grupos) – Correlações (2005)	79
Figura 4.47 – Distribuição Dias da Semana (4 Grupos) – Distâncias Euclideanas (2004)	81
Figura 4.48 – Distribuição Dias da Semana (4 Grupos) – Correlações (2004)	82

Figura 4.49 – Distribuição Dias da Semana (5 Grupos) – Distâncias Euclidianas (2005)	83
Figura 4.50 – Distribuição Dias da Semana (5 Grupos) – Correlações (2005)	85
Figura 4.51 – Distribuição Dias da Semana (6 Grupos) – Distâncias Euclidianas (2004)	86
Figura 4.52 – Distribuição Dias da Semana (6 Grupos) – Correlações (2004)	87
Figura 4.53 – Distribuição Dias da Semana (7 Grupos) – Distâncias Euclidianas (2005)	89
Figura 4.54 – Distribuição Dias da Semana (7 Grupos) – Distâncias Euclidianas (2005)	90
Figura 4.55 – Dias Representativos de 2004 – Distâncias Euclidianas (2 Grupos)	96
Figura 4.56 – Dias Representativos de 2004 – Correlações (2 Grupos)	97
Figura 4.57 – Dias Representativos de 2005 – Distâncias Euclidianas (3 Grupos)	98
Figura 4.58 – Dias Representativos de 2005 – Correlações (3 Grupos)	98

ÍNDICE DE TABELAS

Tabela 2.1 – Definição dos grupos.....	12
Tabela 3.1 – Exemplo cálculo da média de agregação (nível 1).....	23
Tabela 3.2 – Exemplo cálculo da média de agregação (nível 2).....	23
Tabela 3.3 – Dados originais agregados em Períodos de 6 Horas (2004).....	29
Tabela 3.4 – Tabela de Correlações dos dados agregados em Períodos de 6 Horas (2004)	29
Tabela 3.5 – Exemplo da Distribuição de Dias da Semana (2005).....	34
Tabela 4.1 – Caracterização do Agrupamento em 4 Grupos (Perfis Horários)	47
Tabela 4.2 – Caracterização do Agrupamento em 4 Grupos (Perfis 6 Horas)	50
Tabela 4.3 – Caracterização do Agrupamento em 4 Grupos (Perfis Diários)	54
Tabela 4.4 – Caracterização do Agrupamento em 3 Grupos (Perfis Dias da Semana)...56	
Tabela 4.5 – Caracterização do Agrupamento em 3 Grupos (Perfis Semana e Fim-de-Semana)	60
Tabela 4.6 – Caracterização do Agrupamento em 2 Grupos (Perfis Semanais).....	62
Tabela 4.7 – Caracterização do Agrupamento em 2 Grupos (Perfis Mensais).....	65
Tabela 4.8 – Caracterização do Agrupamento em 2 Grupos (Perfis Trimestrais).....	68
Tabela 4.9 – Erros de Refinamento de Perfis Horários	70
Tabela 4.10 – Erros de Refinamento de Perfis 6 Horas	70
Tabela 4.11 – Erros de Refinamento de Perfis Diários	70
Tabela 4.12 – Erros de Refinamento de Perfis Dias da Semana.....	71
Tabela 4.13 – Erros de Refinamento de Perfis Semana e Fim-de-Semana	71
Tabela 4.14 – Erros de Refinamento de Perfis Semanais.....	71
Tabela 4.15 – Erros de Refinamento de Perfis Mensais.....	71
Tabela 4.16 – Erros de Refinamento de Perfis Trimestrais	72

Tabela 4.17 – Erros dos Perfis (Distâncias Euclidianas e Correlações).....	72
Tabela 4.18 – Distribuição Diária por Grupo em 2004 (2 Grupos) – Distâncias Euclidianas	75
Tabela 4.19 – Distribuição Diária por Grupo em 2004 (2 Grupos) – Correlações	77
Tabela 4.20 – Distribuição Diária por Grupo em 2005 (3 Grupos) – Distâncias Euclidianas	78
Tabela 4.21 – Distribuição Diária por Grupo em 2004 (3 Grupos) – Correlações	79
Tabela 4.22 – Distribuição Diária por Grupo em 2004 (4 Grupos) – Distâncias Euclidianas	80
Tabela 4.23 – Distribuição Diária por Grupo em 2004 (4 Grupos) – Correlações	82
Tabela 4.24 – Distribuição Diária por Grupo em 2005 (5 Grupos) – Distâncias Euclidianas	83
Tabela 4.25 – Distribuição Diária por Grupo em 2005 (5 Grupos) – Correlações	84
Tabela 4.26 – Distribuição Diária por Grupo em 2004 (6 Grupos) – Distâncias Euclidianas	86
Tabela 4.27 – Distribuição Diária por Grupo em 2004 (6 Grupos) – Correlações	87
Tabela 4.28 – Distribuição Diária por Grupo em 2005 (7 Grupos) – Distâncias Euclidianas	88
Tabela 4.29 – Distribuição Diária por Grupo em 2005 (7 Grupos) – Correlações	90
Tabela 4.30 – Dias Representativos dos Grupos (2004)	93
Tabela 4.31 – Dias Representativos dos Grupos (2005)	94
Tabela 4.32 – Dias Representativos Comuns	95

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

A emergência de técnicas de extracção de conhecimento de dados a par com o crescente volume de dados recolhidos e gerados pelas empresas nas mais variadas áreas de negócio reflecte a importância que tem vindo a ser dada à análise de sequências ordenadas de dados recolhidos ao longo do tempo. Estes fluxos contínuos de dados são gerados naturalmente por diversas aplicações. Exemplos disso são os registos de chamadas gerados continuamente na área das telecomunicações ou a monitorização de tráfego de internet entre outras aplicações como a monitorização médica, as bases de dados de sensores, a monitorização financeira entre outras mais.

Estes conjuntos de dados com características próprias impõem novos desafios e têm vindo a ser alvo de exploração de várias metodologias de extracção de conhecimento de dados, no sentido de encontrar padrões subjacentes que permitam acompanhar mais proximamente o comportamento dos dados e gerar conhecimento relevante para o domínio em estudo.

A par da possibilidade de efectuar consultas aos dados, não devemos descurar a avaliação e manutenção de todos os padrões frequentes não só nos dados que temporalmente já foram ultrapassados, mas também, e dinamicamente, no que respeita aos novos dados que vão chegando a cada instante de tempo, para que possa haver um acompanhamento em tempo real dos dados em estudo.

1.2 DESCRIÇÃO DO PROBLEMA

O mercado da energia eléctrica é um domínio que necessita de actualizações constantes das previsões de consumo, para que seja possível prever a procura por parte da rede eléctrica (ou seja, dos consumidores) de forma a adequar a oferta disponível em termos de energia que é produzida pelos operadores, não bastando que exista por parte destes uma monitorização mensal ou até mesmo semanal dos perfis de energia, este acompanhamento deve ser mais minucioso.

O principal problema apresentado para a realização deste trabalho consistirá na definição e estudo de perfis de consumo de energia eléctrica para diferentes horizontes temporais a partir de dados de consumo recolhidos em períodos de 15 minutos, no sentido de avaliar a evolução desses perfis com granularidades diárias, semanais ou mensais, entre outras possíveis. Neste sentido, os pontos fundamentais de análise serão os padrões e perfis de consumo de energia, não esquecendo que existem diferentes horizontes temporais que condicionam os resultados obtidos.

Os dados utilizados para esta análise são disponibilizados pela Rede Eléctrica Nacional (REN), uma empresa dedicada à actividade de transporte de electricidade, efectuada através da Rede Nacional de Transporte (RNT). Possibilitando, desta forma, a existência de um elo de ligação entre os centros de produção e os de consumo, alimentando a rede de distribuição e equilibrando a procura e oferta de energia no nosso país.

1.3 PRINCIPAIS CONTRIBUIÇÕES DA DISSERTAÇÃO

As principais contribuições deste estudo são essencialmente as seguintes:

- Exploração de padrões de consumo de energia para diferentes granularidades temporais, com base em agregações simples dos dados originais em histogramas por níveis;
- Definição de perfis para horizontes temporais distintos assente na aplicação de métodos de agrupamento aos resultados dos dados agregados em cada nível de granularidade, comparativamente com os resultados das correlações desses mesmos dados;
- Distribuição de dias da semana por agrupamentos criados com base em perfis diários e definição dos dias representativos para cada grupo.

1.4 ESTRUTURA DA DISSERTAÇÃO

A dissertação encontra-se organizada em cinco secções. A presente secção é iniciada com a apresentação da motivação subjacente, seguida de uma breve descrição do problema em estudo, assim como das principais contribuições deste trabalho. Posteriormente, é feito um enquadramento teórico que nos dá conta do estado de arte no que respeita à temática abordada, em termos de tratamento de dados temporais bem como de estudo de perfis, sendo ainda descritas as metodologias utilizadas, nas Secções 2 e 3, respectivamente.

O progresso do trabalho prático desenvolvido é reunido na Secção 4, que se desdobra, primeiramente, numa breve análise exploratória dos dados que fazem parte do estudo. De seguida, são apresentados os resultados da agregação das observações originais em cada um dos níveis temporais definidos no histograma, considerados relevantes para a envolvente em análise.

Entretanto são propostos possíveis perfis de consumo resultantes da aplicação do método de agrupamento das k-Médias aos dados agregados em cada nível anteriormente definido e às respectivas correlações, sendo analisados os grupos obtidos, em termos de tamanho e de média, bem como os limites de agrupamento e o erro cometido de utilizar os perfis em vez dos valores reais. Como variante foram construídos perfis tendo em conta o desdobramento em traçados médios acima e abaixo da média, acompanhados igualmente do erro cometido e analisados comparativamente.

A fase posterior corresponde à análise da distribuição dos diferentes dias da semana pelas partições entre dois a sete grupos dos dados, resultante da aplicação do método das k-Médias aos perfis diários encontrados previamente. Para cada partição definida é tido em conta o número de elementos, a média do grupo e analisada a distribuição dos dias da semana pelos grupos. Por último, são apresentados os dias mais representativos de cada um dos grupos definidos a partir dos perfis diários encontrados.

Além da análise crítica aos resultados realizada ao longo deste estudo são apresentadas, na última secção, as principais conclusões retiradas, bem como orientações para trabalho futuro.

2 ENQUADRAMENTO TEÓRICO

Antes de partir numa tentativa de desenvolver algo de novo e com algum valor acrescentado para o tema em estudo torna-se necessário ter uma noção daquilo que tem vindo a ser feito não só em termos abrangentes, no que toca ao tratamento de dados recolhidos ao longo do tempo, mas também concretamente, no que respeita à utilização de uma abordagem que permita a sumarização para diferentes horizontes temporais. Neste sentido, é necessário recolher informação acerca dos trabalhos relacionados desenvolvidos neste âmbito. O estudo que foi realizado aponta para a existência de várias vertentes de investigação que irão ser seguidamente apresentadas.

2.1 TRATAMENTO DE DADOS TEMPORAIS

Apesar dos dados tratados neste estudo não serem objectivamente caracterizados como um fluxo contínuo de dados, podemos basear-nos em algumas abordagens a este tipo de dados para ter uma noção das técnicas utilizadas no tratamento de dados temporais na sua generalidade.

Por definição, fluxos contínuos de dados são sequências infinitas de registos recolhidos em períodos fixos ou não de tempo, sendo, por isso, ilimitados. Não sendo possível garantir o armazenamento do conjunto de dados no seu todo, impõem-se restrições de memória. Para colmatar esta lacuna, foram idealizadas técnicas que permitem manter apenas um sumário dos dados históricos. No entanto, na maioria das aplicações, os dados mais antigos são considerados menos relevantes que os dados mais recentes, situação que deve ser tida em conta na construção dos modelos ou estatísticas, daí existirem duas abordagens mais comuns para lidar com essa problemática. A primeira baseia-se na atribuição de pesos decrescentes ao longo do tempo, através de vários modelos, como o exemplo do modelo de decrescimento exponencial explorado em Gilbert et al. (2001) [29]. A outra, geralmente mais utilizada, assenta no modelo de janelas deslizantes [4, 20].

2.1.1 JANELAS TEMPORAIS

As janelas temporais são frequentemente utilizadas para realizar consultas em fluxos contínuos de dados, sendo a consulta efectuada sobre um subconjunto finito de elementos que constituem a janela. Neste tipo de abordagem, associado a cada elemento existe um instante temporal que define quando cada registo é válido. Nestas condições, as consultas consideradas relevantes são executadas sobre os registos dentro de cada janela. Existem vários modelos que exploram este tipo de janelas, destacando-se como mais relevantes, em Gama & Rodrigues (2007) [25], os modelos de *landmark windows*, *sliding windows* e *tilted windows*.

Nas *landmark windows* são identificados pontos relevantes (*landmarks*) no conjunto de dados e o operador de agregação usa todos os registos verificados após o *landmark*. No entanto, a maior parte das vezes não estamos interessados em obter estatísticas sobre todo o passado, mas apenas do passado recente. As *sliding windows* reflectem essa filosofia, podendo ser utilizadas numa lógica de tamanho fixo ou variável das janelas. Neste modelo os elementos chegam continuamente e apenas os N elementos mais recentes são usados para responder a consultas. A figura que se segue evidencia as diferenças entre estes dois primeiros modelos sugeridos.

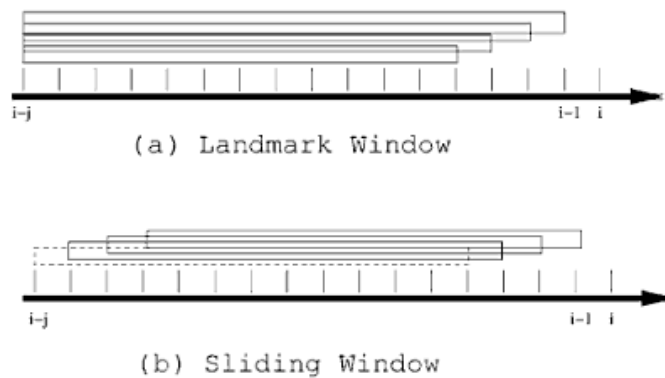


Figura 2.1 – Representação das *Landmark* e *Sliding Windows*

Os modelos anteriores utilizam um “esquecimento” catastrófico. Como alternativa, surgem os modelos de *tilted time windows* [13] onde a escala temporal é comprimida. Desta forma, os dados mais recentes são armazenados dentro da janela ao mais fino pormenor, denominado de granularidade, sendo a informação mais antiga, por sua vez, armazenada menos detalhadamente, de uma forma agregada. Existem duas variantes deste modelo, desdobrando-se em *natural tilted time windows* e *logarithm tilted windows*, apresentadas na figura a seguir:

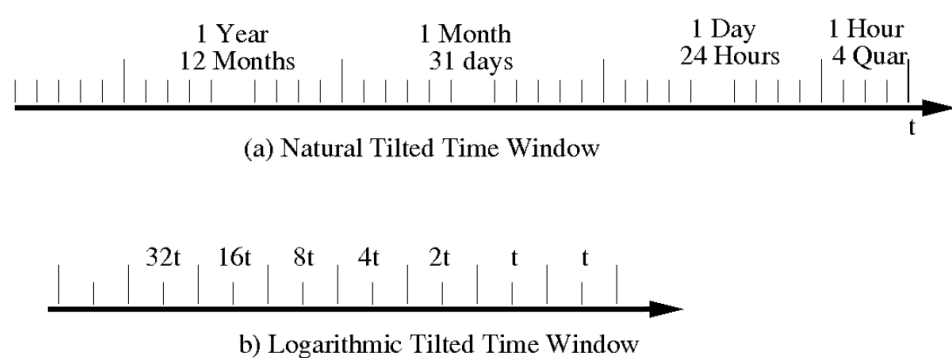


Figura 2.2 – Tipos de *Tilted Time Windows*

No primeiro caso, os dados são armazenados com uma granularidade que obedece à taxonomia natural do tempo. Como ilustrado na Figura 2.2 (a), a título exemplificativo, a última hora é guardada com uma granularidade de 15 minutos (4 células¹), o último dia é armazenado em horas (24 células), o último mês em dias (31 células) e o último ano em meses (12 células). No segundo caso – Figura 2.2 (b) – a compressão da escala temporal é semelhante, havendo, no entanto, uma igualdade em termos de tamanho das células. O tempo é comprimido com base numa função exponencial de base fixa 2 e factor exponencial variável n ($n = 1, \dots, N$) (e.g., $2^0, 2^1, 2^2$, etc). Assim, as células agregam e armazenam dados de limites de tempo de tamanho crescente.

¹ Células correspondem a divisões/compartimentos, neste caso, de uma linha temporal.

Em Giannella et al. (2003) [26] é demonstrado que a utilização de *logarithmic tilted time windows* reduz o número de *tilted time windows* usadas, comparativamente com as *natural tilted time windows*, denotando-se por parte das primeiras uma abordagem mais eficiente em termos de espaço. A metodologia proposta pelos autores consiste na manutenção incremental de *tilted time windows* para cada padrão frequente com diferentes granularidades, no sentido de ser possível responder a um conjunto de consultas aos dados.

2.1.2 HISTOGRAMAS

Explorados os procedimentos, em termos de janelas temporais, que podem ser utilizados no processamento dos dados, é importante ter em conta as técnicas que podem ser utilizadas para dar resposta às consultas aos dados. Estas técnicas podem ser de dois tipos: redução de dados e resumos, sendo que ambas são executadas sobre sumários de dados. São apresentadas como técnicas de redução de dados mais comuns a amostragem, sinopses e histogramas e *wavelets*. Existe muito trabalho relacionado em termos de manutenção de histogramas e outras estruturas de sinopse [28, 29, 30, 32, 41].

De acordo com a metodologia proposta para este trabalho é destacada a utilização de histogramas como técnica de sumarização de dados temporais, especialmente de fluxos contínuos de dados, no sentido de conseguirem captar a distribuição dos dados. Através da contagem da frequência de cada valor do conjunto, a sua representação é aproximada, idealmente, de uma forma eficiente em termos de memória disponível e de erro global decorrente da aproximação.

Os histogramas exponenciais foram sugeridos em Datar et al. (2002) [20] para resolver problemas de contagem, no sentido de manter agregados e estatísticas sobre fluxos contínuos de dados, respeitantes aos últimos N elementos dos dados vistos até ao momento, utilizando um modelo de *sliding window*.

Por outro lado, Bifet (2009) [7] utiliza uma estrutura de dados variante dos histogramas exponenciais, sugerindo um algoritmo de janela deslizante adaptativa em que é utilizado um estimador com memória e detecção de mudança.

O problema de construção de histogramas sobre conjuntos de dados de tamanho conhecido, ou seja, finito, acaba por ser trivial, sendo considerado um problema clássico de construção de histogramas, como será apresentado neste trabalho. No entanto, quando se passa para conjuntos de dados com um tamanho desconhecido, onde os dados chegam continuamente formando um fluxo contínuo de dados, por definição infinito, o problema de construção de histogramas passa a ser mais complexo.

Em Guha et al. (2006) [31] são sugeridas algumas características que algoritmos deste tipo devem possuir, entre elas: serem capazes de operar numa quantidade fixa de memória, serem rápidos de forma a acompanharem a natureza de tempo real do fluxo contínuo de dados; e incrementais no sentido de que com novos dados produzidos os resultados intermediários possam ser reutilizados. A acrescentar, a natureza dos algoritmos de fluxos contínuos de dados pode ser aglomerativa ou de janela fixa [30], sendo que, no primeiro caso, são considerados todos os itens desde o início do fluxo de dados e no segundo apenas são considerados os últimos M elementos, sobretudo devido a restrições de memória.

Outras metodologias foram propostas [42], subdividindo-se a janela deslizante em janelas de pequena dimensão com tamanho fixo (janelas básicas), preconizando a manutenção de estatísticas como a média, desvio padrão ou correlações entre pares de fluxos contínuos de dados não só para as janelas básicas mas também para a janela deslizante, de forma a facilitar a eliminação de dados antigos e facilitar a incorporação de dados recentes.

Em termos abrangentes é apresentada por Babcock et al. (2002) [4] uma pesquisa do trabalho realizado em algoritmos de fluxos contínuos de dados. Datar et al. (2002) [20] consideram o problema da manutenção de estatísticas simples com recurso a janelas deslizantes, através da utilização de histogramas exponenciais, no sentido de criar algoritmos óptimos em termos de memória para estatísticas fundamentais como contagem, soma dos números inteiros positivos, média, entre outras.

Além das estatísticas básicas, vários problemas foram efectivamente resolvidos recorrendo a janelas deslizantes. Em Babcock et al. (2003) [6] foi estendida a aplicabilidade desta técnica a outras estatísticas como variância e a k -mediana. Gibbons

& Tirthapura (2002) [27] propuseram um melhoramento do modelo de janela deslizante [20] providenciando algoritmos de memória e tempo ótimos para os problemas de contagem e soma.

Chi et al. (2004) [14] consideraram o problema de conjuntos de itens frequentes. Algoritmos de contagem de frequência e quantis foram propostos por Arasu & Manku (2004) [2]. Datar & Muthukrishnan (2002) [19] resolveram problemas de raridade e similaridade. Domingos & Hulten (2000) [21] e Hulten et al. (2001) [33] utilizaram também janelas deslizantes, mas no contexto da aprendizagem de árvores de decisão sobre fluxos contínuos de dados. Foram ainda apresentados algoritmos de amostragem aleatória simples para janelas deslizantes [5].

2.2 ESTUDO DE PERFIS

Até aqui foram explanadas algumas das técnicas que têm vindo a ser objecto de investigação no que toca ao tratamento (armazenamento/sumarização) de dados temporais. A partir daqui serão exploradas as variadas metodologias desenvolvidas no que respeita à obtenção de perfis de consumo exclusivamente na área de distribuição de energia eléctrica.

O agrupamento e caracterização do comportamento dos consumidores de energia eléctrica estão ainda a ser objecto de investigação. No entanto, os perfis de consumo foram identificados como um método adequado para lidar com os clientes sem ter que enfrentar o impedimento trazido pelo período de tempo envolvido na medição e registo da energia consumida através de equipamentos instalados ao longo da rede de distribuição de energia. Além disso, esta metodologia também serve como ferramenta para as empresas de distribuição melhorarem as suas estratégias de mercado e oferecerem novos serviços, assim como para desenvolverem novas tarifas [9], [15] num mercado liberalizado. Métodos muito diferentes, desde os mais convencionais aos mais sofisticados [8], [9], [10] têm sido usados para a modelação da criação de perfis.

A metodologia de criação de perfis foi usada durante muitos anos, antes da liberalização de mercados, para a formulação de tarifas. No entanto, a partir da desregulamentação, é acentuada a necessidade de reconhecimento de perfis de consumo de electricidade pelas razões acima referidas. O processo de criação de perfis é intuitivamente dividido em duas etapas que se desdobram na determinação dos grupos de consumidores e na sua alocação em cada grupo específico. O Reino Unido, Taiwan e Portugal são alguns dos países onde têm vindo a ser desenvolvidos variados estudos na área de construção de perfis de consumo de energia eléctrica e, conseqüentemente, aplicadas diversas técnicas de agrupamento.

No mercado de oferta do Reino Unido é importante compreender as variações de consumo no sentido de desenvolver estruturas de preço competitivas, integradas numa estratégia de *marketing* agressivo. Neste sentido, é apresentada em B. Pitt & D. Kitschen (1999) [38] a aplicação de um método centrado no agrupamento de perfis de consumo através de árvores de decisão adaptativas. A criação de ferramentas para apoiar os vendedores e consumidores de energia irá motivar uma participação mais activa por parte da procura nos mercados de electricidade.

Em Chicco et al. (2003) [15] foi conduzido um estudo utilizando perfis de consumo, com um procedimento de agrupamento automático, para caracterizar os padrões de consumo dos clientes e calcular a margem de uma empresa de distribuição para a fixação de tarifas para diferentes grupos de clientes. Além disso, um estudo [16] dedicado ao desenvolvimento de estratégias de mercado também foi realizado através de agrupamento.

Em Taiwan, um sistema de levantamento de necessidades de consumo foi implementado desde 1993 para determinar vários grupos de clientes com diferentes características de consumo [8]. Este sistema serviu várias funções – Figura 2.3 – tais como o planeamento do sistema [11], operação e manutenção do sistema [12], estrutura de tarifas e marketing [9]. Este estudo usou uma análise estatística para obter o padrão de consumo de cada cliente e representar o próprio comportamento de consumo [10].

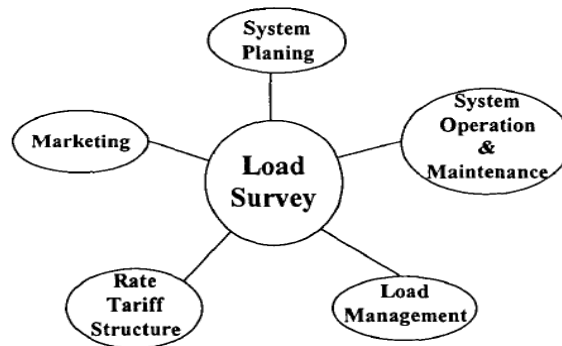


Figura 2.3 – Múltiplas funções do sistema

Em Portugal, entretanto, apesar do mercado ainda não se encontrar num clima de liberalização, foram utilizadas várias técnicas de extracção de conhecimento de dados para desenvolver a caracterização de consumo de electricidade. Diversos algoritmos foram testados em diferentes operações de agrupamento [24]. A selecção do melhor algoritmo foi baseada na análise comparativa de desempenho, utilizando duas medidas de adequação: uma medida de densidade do agrupamento (*Mean Index Adequacy - MIA*) e uma medida de dispersão do grupo (*Clustering Dispersion Indicator - CDI*) apresentadas em [15].

Em M. A. Matos et al. (2005) [36] é descrita a prática portuguesa no que respeita ao estabelecimento e aplicação de perfis de consumo no funcionamento do mercado, sendo concebida uma nova abordagem de agrupamento para construir diagramas de consumo típicos que utiliza os diagramas reais de consumo para criar grupos definidos no espaço de informação de facturação, que maximizem os diagramas de consumo em cada grupo.

São determinadas três fases de um projecto de agrupamento de consumos, sendo primeiramente recolhidos os dados, seguido do processamento do agrupamento e terminando com testes de avaliação de desempenho. A fase de processamento pode ser baseada em grupos pré-definidas (baseadas nos resultados de facturação da empresa) ou através dos diagramas de consumo. A primeira abordagem pode gerar uma grande variedade de diagramas em cada grupo, ao passo que a segunda pode gerar bons resultados.

O conceito chave desta metodologia consiste em estabelecer uma grelha de pequenas células de agrupamento – Figura 2.4 e Tabela 2.1 – no espaço de informação de facturação que são posteriormente agregadas por um procedimento de optimização que minimiza uma medida de compacticidade dos diagramas em cada grupo. Cada célula é caracterizada por uma sub-amostra de diagramas, um diagrama médio e a variância intra-grupo correspondente. Um conceito chave adicional do trabalho é descrever como os perfis de consumo desenvolvidos são usados para a operação no mercado, em especial para fins de decisão.

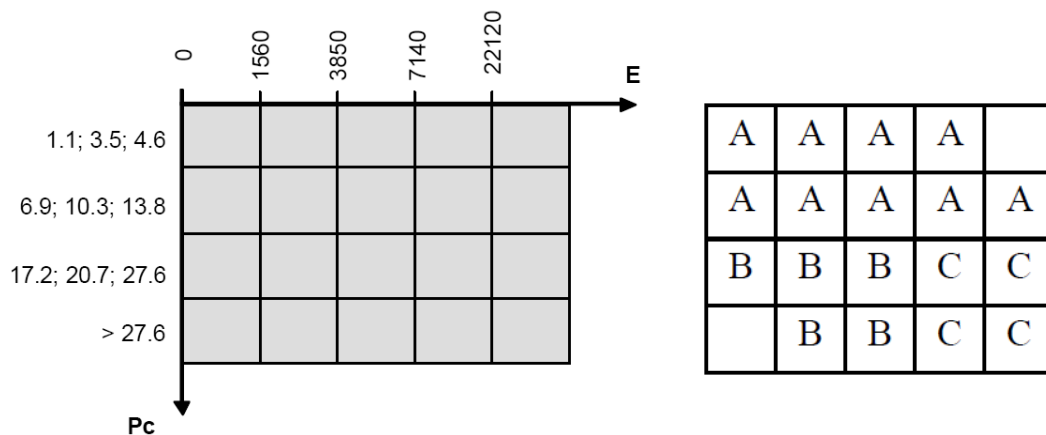


Figura 2.4 – Exemplo de agrupamento das células

	Pc	E
Class A	≤ 13.8	all values
Class B	>13.8	≤ 7140
Class C	>13.8	>7140

Tabela 2.1 – Definição dos grupos

Um outro estudo elaborado por Nizar et al. (2006) [37] apresenta a exploração de perfis de consumo de energia eléctrica para diferentes tipos de consumidores, sendo comparados alguns métodos de criação de perfis, através da análise e avaliação de técnicas de agrupamento. O objectivo deste estudo é determinar os melhores métodos de criação de perfis e técnicas de extracção de conhecimento de dados para agrupar, detectar e prever perdas não técnicas no sector da distribuição de energia, devido a falhas e erros de medição de facturação, bem como reunir conhecimentos sobre o

comportamento do consumidor e suas preferências, de modo a obter uma vantagem competitiva no mercado liberalizado.

São feitas comparações entre os métodos de agrupamento *k-Means*, *Cobweb* e o algoritmo *Expectation-Maximization* (EM) no mesmo conjunto de dados. Os resultados obtidos são diferentes em termos de número de grupos a seleccionar. O tempo de execução para o *Cobweb* é o mais alto, sendo o *k-Means* a técnica que funciona melhor nestes conjuntos. De acordo com os dados disponíveis, os consumidores de electricidade foram agrupados de acordo com a hora do dia e as características de consumo. É também demonstrada a existência de diferenças significativas entre os três métodos. No futuro, este desenvolvimento no estudo servirá para que as irregularidades ou anormalidade neste padrão de consumo disparem um alerta para o sistema.

É apresentado por Falcão & Henriques (2001) [22] um projecto baseado na aplicação de técnicas de Redes Neurais e *Fuzzy Sets*² para gerar curvas de padrões de consumo para grupos de consumidores com base no seu consumo mensal de energia e numa grande base de dados de curvas de consumo das campanhas de medição levadas a cabo na rede de distribuição eléctrica. O estimador de carga utiliza essas curvas para ajustar a carga individual nos pontos de carga especificada.

Primeiramente, é estabelecido um conjunto de curvas de consumo diário que será utilizado mais tarde para dar uma estimativa aproximada da carga imediata num determinado ponto de carga. A técnica de agrupamento utilizada é baseada no *Kohonen's Self Organizing Map* (SOM). O segundo passo na metodologia de estimação de consumo consiste no desenvolvimento de um procedimento de agrupamento que agrupe os consumidores individuais em determinados grupos através do uso de *fuzzy sets* associados aos atributos de consumo de cada um dos grupos.

Os dados de distribuição dos consumidores são o *input* do classificador *fuzzy*, de forma a que, com base num conjunto de curvas de consumo típicas, seja definida a melhor curva para descrever esse comportamento de consumo particular. Esta curva de consumo é ajustada para o consumo de kw/h mensal. Os resultados obtidos para os

² *Fuzzy Sets* ou Conjuntos Difusos.

consumidores conectados num transformador de distribuição particular são adicionados e integrados no Estimador de Programação Linear *Fuzzy*. Na Figura 2.5 é representado todo o processo descrito neste estudo [22].

A metodologia descrita foi implementada e testada com dados reais de uma empresa de distribuição brasileira e os resultados obtidos até ao momento são encorajadores, mas são necessários mais testes para confirmar a adequação da metodologia para aplicações práticas.

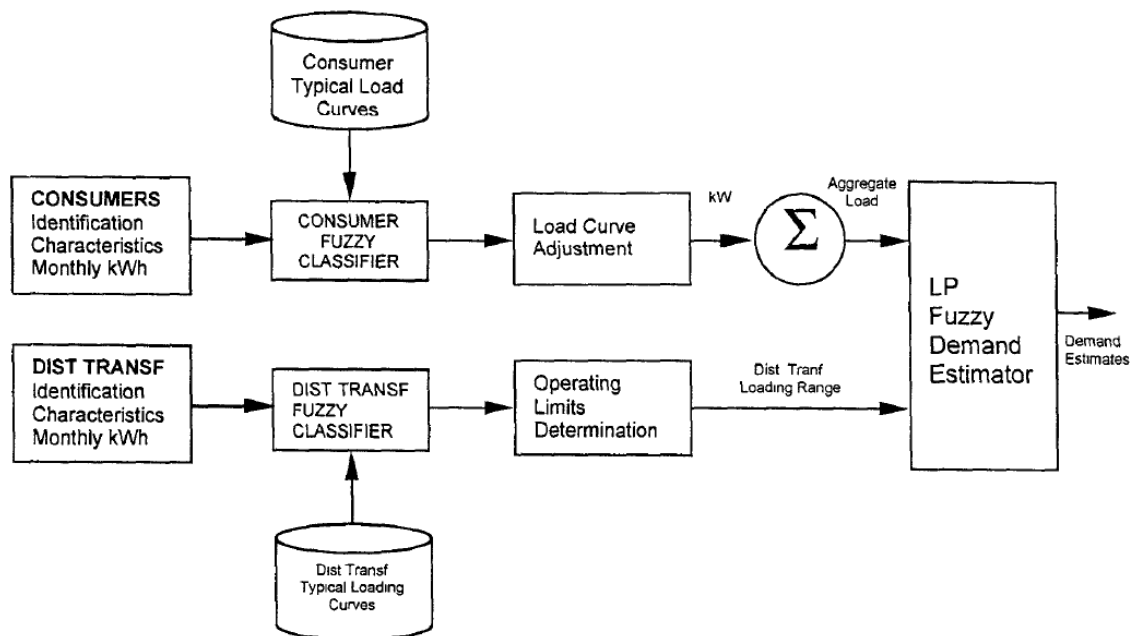


Figura 2.5 – Esquema representativo do processo de estimação de consumo

2.2.1 DISCRIMINAÇÃO DE TIPOS DE CONSUMIDORES

A especificação de diferentes tipos de consumidores, para uma posterior obtenção de perfis pode tornar-se uma metodologia algo vantajosa, sendo à partida discriminados grupos de clientes que desenvolvem o mesmo tipo de actividade e posteriormente aplicadas técnicas de obtenção de perfis dentro de cada agrupamento.

Num estudo desenvolvido neste sentido por Jardini et al. (2000) [34] são apresentados resultados no que toca à determinação de curvas diárias de consumo para consumidores residenciais, comerciais e industriais, tendo por base, medições de campo realizadas pelas concessionárias de Energia Eléctrica do Estado de São Paulo, Brasil. É também apresentada uma metodologia para a agregação destes consumos no sentido de determinar a carga esperada nos equipamentos ou a carga real numa parte da rede de distribuição, sendo usadas para isso, as curvas diárias representativas da actividade de cada consumidor e o consumo de energia mensal dos consumidores conectados.

Esta análise estatística levou à recomendação das curvas representativas de consumidores por intervalo de consumo residencial e por tipo de actividade nos sectores comercial e industrial. Estas curvas podem ser utilizadas para obter curvas de consumo diário em qualquer ponto da rede através da agregação de consumo dos consumidores. Este trabalho propõe que as curvas representativas de consumo diário dos consumidores sejam definidas estatisticamente pelas curvas de média e desvio padrão.

2.2.2 APLICAÇÃO PARA PREVISÃO

Em Álvarez et al. (2007) [1] é analisada uma série temporal de preços de electricidade, onde o principal objectivo é encontrar uma representação que preserve a informação original e que descreva a forma da série temporal o mais aproximado possível, através da utilização dos algoritmos *k-Means* e *Fuzzy C-means*. A escolha do número de grupos a analisar, como uma das decisões mais críticas a tomar nas técnicas de agrupamento é resolvida a partir da medida de silhueta para o método *k-Means*.

Neste estudo torna-se importante encontrar os grupos de preços que mostrem um comportamento semelhante sob determinadas circunstâncias (e.g., a divergência entre dias úteis e fins-de-semana e feriados), podendo esta informação ser utilizada para previsão futura da progressão dos preços para o dia seguinte. Outros autores utilizaram métodos diferentes para previsão, mas apenas este trabalho parte da descoberta de padrões no preço da electricidade para melhorar os modelos de previsão.

Torna-se, assim, claro que a utilização de técnicas de agrupamento para previsão é tão poderosa como útil. Através de ambos os agrupamentos foi obtido o mesmo erro, sendo possível observar que os dias da semana têm um comportamento oposto aos dos fins-de-semana e festividades.

2.2.3 INTEGRAÇÃO COM A GESTÃO

A adaptação destas técnicas à prática de gestão da empresa é bastante relevante no que toca, por exemplo, ao estabelecimento de estratégias de *marketing* para retenção de clientes – Figura 2.6. O método proposto por Kitayama et al. (2002) [35] inserido num ambiente de desregulamentação do mercado de energia eléctrica no Japão implica a utilização de árvores de decisão com base nos dados do perfil do cliente como forma de estudar a mudança de fornecedor de energia por parte dos clientes, encarando os clientes não apenas como consumidores uniformes, mas que apresentam diferentes perfis. O método de extracção de conhecimento de dados adoptado ajuda no agrupamento dos clientes preferenciais nos respectivos segmentos a partir da determinação das medidas que devem ser tidas em conta na definição de segmentação de clientes. Este processo segue uma lógica de *Customer Relationship Management - CRM* (Gestão de Relação com o Cliente), uma vez que pretende alcançar uma compreensão profunda dos clientes que auxilie a delineação e o planeamento de uma estratégia de negócio.

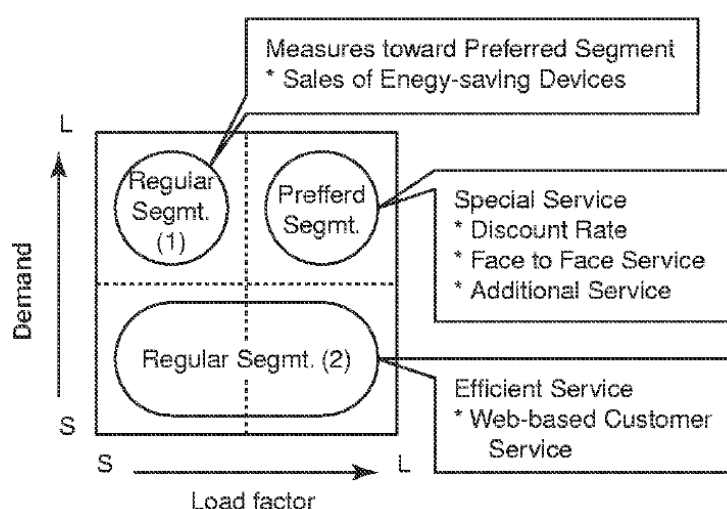


Figura 2.6 – Exemplo da Segmentação de Consumidores

2.2.4 INTEGRAÇÃO COM A TOMADA DE DECISÃO

A integração das técnicas desenvolvidas no processo de decisão das empresas fornecedoras de energia eléctrica torna-se essencial para que as investigações levadas a cabo sejam frutíferas. Neste sentido, V. Figueiredo et al. [23] apresentam um Sistema de Apoio à Decisão (SAD) para ajudar tanto as empresas de electricidade como os clientes na definição dos melhores contratos e tarifas. Este SAD é suportado por um quadro de caracterização do consumidor.

As regras geradas pela utilização de árvores de decisão formam o conhecimento base do SAD. Os perfis de consumo, juntamente com a informação dos contratos e dos preços da electricidade formam a base de dados do SAD. Desta forma, este sistema permitirá o agrupamento de diferentes consumidores, com base nos seus perfis de consumo e o teste de diferentes tarifas e contratos. Os resultados finais do SAD centram-se numa análise económica comparativa entre diferentes contratos e conselhos sobre os contratos mais económicos para cada grupo de consumidores. A metodologia proposta baseia-se em projectos anteriores de perfis de consumo [39, 16].

Esta aplicação tem como intenção futura o fornecimento do SAD como uma aplicação *Web* para ser consultada em qualquer local por qualquer pessoa, usando um simples browser *Web* que permita aos consumidores finais simular diferentes cenários que se adequem ao seu comportamento de consumo.

Com a aprendizagem das práticas levadas a cabo no tratamento de dados temporais e na obtenção de perfis a partir de dados do sector energético foi possível solidificar conhecimentos básicos para o tratamento da temática abordada neste trabalho.

Nesta Secção foram dadas a conhecer as linhas de investigação dentro do problema em estudo. No entanto, neste trabalho, é dado um passo em frente na obtenção de perfis visto ser realizada a partir de diferentes níveis de granularidade.

3 METODOLOGIAS

Neste capítulo são dadas a conhecer as metodologias utilizadas no trabalho de sumarização de dados de consumo de energia eléctrica de acordo com os diferentes horizontes temporais definidos, nomeadamente no que respeita a ferramentas de análise exploratória, procedimentos de preparação dos dados, de aplicação de histogramas por níveis aos mesmos, de obtenção e avaliação de agrupamentos/perfis e de apreciação da distribuição dos dias da semana pelos agrupamentos determinados.

3.1 ANÁLISE EXPLORATÓRIA

3.1.1 ANÁLISE PRELIMINAR

De forma a realizar uma análise preliminar dos dados foram aplicadas aos dados funções de sumarização em termos das variáveis que constituem o conjunto de dados através da função *str()* do *software* R. Esta função permite a visualização global das variáveis constituintes. Funções semelhantes podem ser efectuadas através dos comandos *head()* e *tail()* que permitem visualizar as seis primeiras e seis últimas linhas de observações para todas as variáveis do conjunto. Foram ainda obtidas as seguintes estatísticas descritivas básicas da variável de interesse para o estudo, disponíveis através da função *summary()* do R:

Mínimo: valor mais baixo no conjunto de valores tomados por determinada variável;

Primeiro Quartil, também designado por quartil inferior, é o valor correspondente aos 25% da amostra ordenada de elementos;

Mediana ou segundo quartil, é o valor até ao qual se encontra 50% da amostra ordenada;

Média: valor médio no conjunto de valores tomados por determinada variável;

Terceiro Quartil, também denominado de quartil superior, é o valor a partir do qual se encontram 25% dos valores mais elevados, ou seja, o valor correspondente aos 75% da amostra ordenada;

Máximo é valor mais elevado no conjunto de valores tomados por determinada variável.

3.1.2 PREPARAÇÃO DOS DADOS

Após a análise preliminar e no sentido de preparar os dados para aplicações futuras, foi feita a divisão do conjunto de dados, para os anos em estudo (2004 e 2005). Posteriormente, foi analisada graficamente a evolução do consumo de energia eléctrica ao longo de cada ano através de um gráfico que regista o valor observado para os períodos de 15 minutos contemplados na totalidade de cada ano, sendo também caracterizados através das mesmas estatísticas descritivas básicas anteriores e analisados em termos comparativos.

3.1.3 ESTUDO DA AUTO-CORRELAÇÃO DOS DADOS

Para finalizar o estudo preliminar dos dados, foi feita uma análise à auto-correlação verificada para os diferentes perfis e diferentes períodos temporais. A função de auto-correlação é uma medida de interesse no que respeita à análise de séries temporais, como é o caso dos dados em estudo. Esta função descreve o grau de correlação existente entre os valores de determinada série temporal observados em diferentes instantes de tempo. No instante 0 é atingido o valor máximo de 1, indicativo de correlação perfeita da variável com ela mesma. O valor mínimo, -1, significa uma anti-correlação perfeita. Por outro lado, o valor 0 indica uma total ausência de correlação.

Admitindo que a variável a analisar é uma variável aleatória X_t , discreta estacionária, dependente do tempo t e com média μ , a **função de auto-correlação** $R(k)$ [3] é definida pela fórmula (3.a) que se segue, onde $E[...]$ é o valor esperado, k é a deslocação temporal e $\sigma_{X_t}^2$ é a variância da variável X_t . Resumidamente, esta formulação é definida como a razão entre a auto-covariância e a variância da variável em estudo.

$$(3.a) \quad R(k) = \frac{E[(X_t - \mu)(X_{t+k} - \mu)]}{\sigma_{X_t}^2}$$

A interpretação dos resultados foi feita visualmente, através da análise dos correlogramas – Figura 3.1 – correspondentes a diferentes horizontes temporais, como apresentado de seguida, através da função `acf()`, disponível no R.

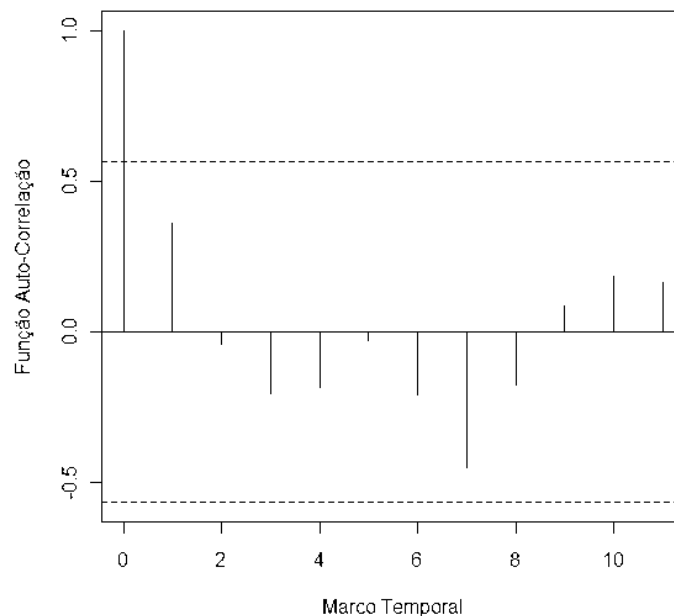


Figura 3.1 – Exemplo Auto-Correlação para Perfis Mensais (2004)

A tracejado encontram-se as chamadas “bandas de significância” que delimitam os valores de auto-correlação considerados significativos.

3.2 APLICAÇÃO DE HISTOGRAMA POR NÍVEIS AOS DADOS

A principal metodologia utilizada neste estudo, que servirá como ponto de partida para o desenvolvimento de todo o trabalho prático, são os **histogramas por níveis**, uma derivação das *natural tilted time windows* referidas na secção anterior, no sentido em que existe um acompanhamento das granularidades temporais de acordo com a taxonomia natural do tempo.

A principal diferença entre estas duas metodologias reside no facto de que na abordagem proposta as agregações dos dados, para horizontes temporais crescentes, são guardadas por níveis, ao invés de serem guardadas numa linha temporal horizontal, onde o passado mais recente é armazenado em unidades temporais mais detalhadas e o passado mais longínquo é armazenado em unidades temporais mais dilatadas. O mesmo se verifica nos histogramas por níveis, mas numa extensão entre o nível inicial inferior e final superior de agregação.

Para cada nível correspondente aos horizontes temporais a considerar, são agrupados os dados originais do conjunto, sendo utilizada a **média** como **estatística de agregação**. A Figura 3.2 que se segue é um esquema ilustrativo do que se pretende fazer.

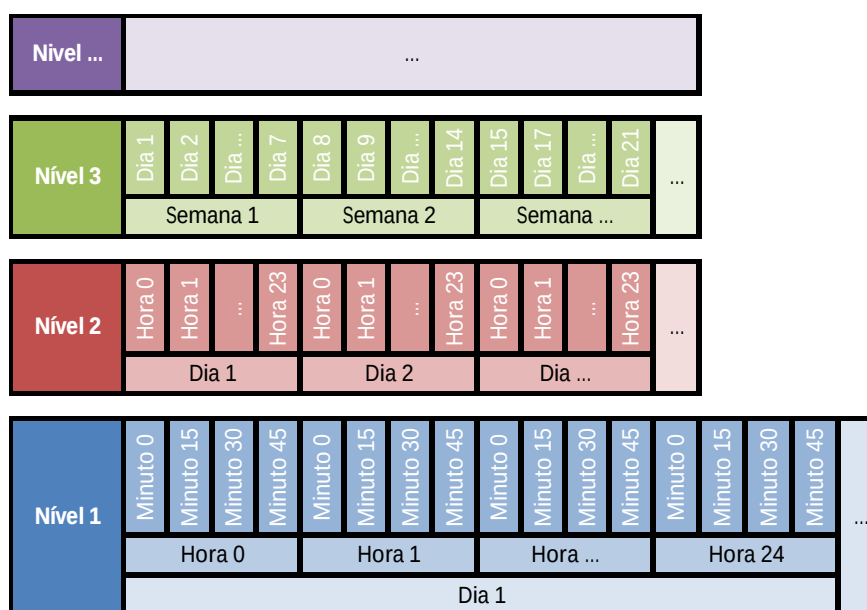


Figura 3.2 – Esquema Exemplificativo de um Histograma por Níveis

A título exemplificativo, considerando que os dados são recolhidos em períodos de 15 minutos, no primeiro nível de agregação, esses mesmos dados são agregados numa unidade temporal superior, por exemplo, por hora. Num nível posterior, os dados anteriormente agregados por hora, são reunidos por dias e assim sucessivamente, até um último nível de agregação justificável para o domínio em estudo.

De acordo com os dados em análise, foram definidos os seguintes **níveis de agregação**: Perfis Horários (Nível 1), Perfis 6 Horas (Nível 2), Perfis Diários (Nível 3), Perfis Dias da Semana (Nível 4), Perfis Semana e Fim-de-Semana (Nível 5), Perfis Semanais (Nível 6), Perfis Mensais (Nível 7), Perfis Trimestrais (Nível 8) e Perfis Semestrais (Nível 9).

Considerando $X_{a,d,t,m}$ a variável representativa do valor observado em termos de consumo de energia eléctrica para o ano a , dia d , valores possíveis da unidade temporal t em causa e o minuto m , temos que a estatística de agregação para o primeiro nível do histograma será dada pela seguinte fórmula:

$$(3.b) \quad \bar{X}_{a,d,t} = \frac{1}{n} \sum X_{a,d,t,m} \text{ em que}$$

n corresponde ao número de períodos agregados na unidade temporal do primeiro nível (hora), ou seja, os quatro períodos de 15 minutos que constituem cada hora;

a refere-se aos anos de 2004 ou 2005;

d assume os valores possíveis para o mês relativo, variando entre um mínimo de 1 e um máximo de 31, consoante o mês em causa;

t adopta os valores entre 0 a 23 horas;

m toma os valores 0, 15, 30 ou 45.

A tabela que se segue – Tabela 3.1 – pretende exemplificar a aplicação da fórmula de cálculo da estatística de agregação para a primeira hora, do primeiro dia do mês de Janeiro de 2004 no primeiro nível do histograma.

	$X_{a,d,t,m}$	$\bar{X}_{a,d,t}$
$a = 2004, d = 1, t = 0, m = 0$	$X_{2004,1,0,0} = 4402.5$	$\begin{aligned} \bar{X}_{2004,1,0} &= \\ \frac{1}{4} \sum (X_{2004,1,0,0}, X_{2004,1,0,15}, \\ &X_{2004,1,0,30}, X_{2004,1,0,45}) = \\ &= \frac{4402.5 + 4381.3 + 4334.2 + 4285.9}{4} = \\ &= \mathbf{4350.975} \end{aligned}$
$a = 2004, d = 1, t = 0, m = 15$	$X_{2004,1,0,15} = 4381.3$	
$a = 2004, d = 1, t = 0, m = 30$	$X_{2004,1,0,30} = 4334.2$	
$a = 2004, d = 1, t = 0, m = 45$	$X_{2004,1,0,45} = 4285.9$	

Tabela 3.1 – Exemplo cálculo da média de agregação (nível 1)

A partir dos resultados deste primeiro nível é possível construir os restantes níveis das agregações pretendidos, como exemplificado na Tabela 3.2 para o segundo nível de agregação.

	$\bar{X}_{a,d,t}$	Perfis 6 Horas
$a = 2004, d = 1, t = 0$	$\bar{X}_{2004,1,0} = 4350.975$	$\begin{aligned} \bar{X}_{2004,1,0-5} &= (4350.975 + \\ &+ 4095.275 + 3785.45 + 3511.25 + \\ &+ 3338.25 + 3240.375) / 6 = \\ &\mathbf{3720.2625} \end{aligned}$
$a = 2004, d = 1, t = 1$	$\bar{X}_{2004,1,1} = 4095.275$	
$a = 2004, d = 1, t = 2$	$\bar{X}_{2004,1,2} = 3785.45$	
$a = 2004, d = 1, t = 3$	$\bar{X}_{2004,1,3} = 3511.25$	
$a = 2004, d = 1, t = 4$	$\bar{X}_{2004,1,4} = 3338.25$	
$a = 2004, d = 1, t = 5$	$\bar{X}_{2004,1,5} = 3240.375$	
$a = 2004, d = 1, t = 6$	$\bar{X}_{2004,1,6} = 3211.55$	$(\dots)$
$a = 2004, d = 1, t = 7$	$\bar{X}_{2004,1,7} = 3188.175$	
$a = 2004, d = 1, t = 8$	$\bar{X}_{2004,1,8} = 3091.1$	
$a = 2004, d = 1, t = 9$	$\bar{X}_{2004,1,9} = 3267.4$	
$a = 2004, d = 1, t = 10$	$\bar{X}_{2004,1,10} = 3617.35$	
$a = 2004, d = 1, t = 11$	$\bar{X}_{2004,1,11} = 4005.15$	
$(\dots)$		

Tabela 3.2 – Exemplo cálculo da média de agregação (nível 2)

As conclusões que podem ser retiradas a partir desta primeira abordagem são essencialmente baseadas na visualização de **gráficos** que mostram, para as unidades temporais dos níveis em análise, a **evolução do consumo de energia eléctrica**, de forma a permite-nos retirar conclusões mais intuitivamente. A figura que se segue é um exemplo disso, sendo possível avaliar e identificar os picos de maior e menor consumo de energia, posteriormente caracterizados.

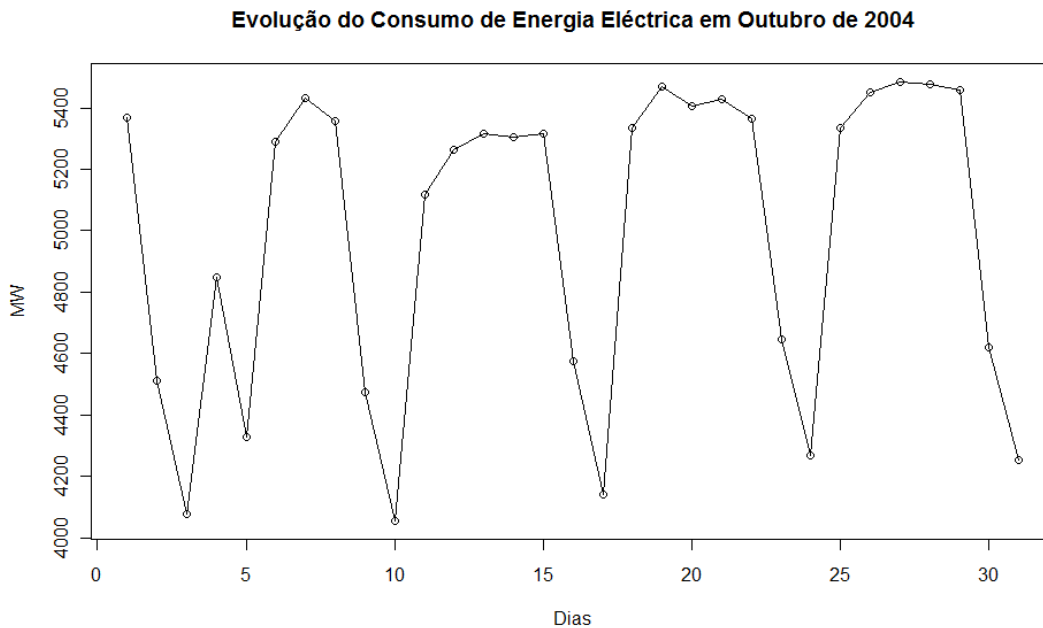


Figura 3.3 – Exemplo de visualização dos dados agregados (nível 3)

3.2.1 MÉTODO K-MÉDIAS

A aplicação deste tipo de metodologia a dados recolhidos sequencialmente ao longo do tempo, tem como finalidade a extracção de padrões que caracterizem a periodicidade dos actos envolvidos. Numa tentativa de explorar a **obtenção de perfis** de consumo de energia eléctrica para as diversas granularidades justificáveis, é sugerida a utilização de técnicas de agrupamento que permitam reunir o conjunto de observações em grupos homogéneos. Neste sentido, aos resultados obtidos através dos histogramas por níveis, será aplicado o **método** iterativo das **k-Médias**. Este baseia-se na criação de partições dos dados, sendo estes particionados em tantos grupos quanto o número de perfis pretendidos.

O método das k-Médias é um algoritmo de agrupamento não hierárquico que consiste em agrupar observações em grupos que sejam, por um lado, homogéneos, no sentido de que os elementos de cada grupo devem ter um grau de associação elevado entre si, sendo muito semelhantes, e por outro lado sejam bem separados, de modo a que elementos de grupos distintos se assemelhem o menos possível. Neste algoritmo parte-se de um conjunto inicial de centros e afecta-se cada objecto ao grupo cujo centro

está mais próximo, sendo o centro do grupo recalculado após cada afectação. Os objectos são inicialmente alocados ao grupo com a média mais próxima. Tendo em conta a alocação inicial, o procedimento consiste em procurar, iterativamente, a k -ésima partição com a soma de quadrados localmente óptima dentro do agrupamento, movendo os pontos de um grupo para outro.

Existem vários algoritmos que podem ser referenciados no uso deste método [18]. O mais comum é o apresentado por *MacQueen* (1967), mas também pode ser utilizado o dado por *Lloyd* (1957) ou *Forgy* (1965). O algoritmo de **Hartigan e Wong (1979)** é utilizado por defeito na função *kmeans()* do R, por se considerar que, usualmente, é o que obtém melhores resultados. Desta forma, dados n objectos com p variáveis medidas em cada objecto por $x_{i,j}$ onde $i = 1, 2, \dots, n$; $j = 1, 2, \dots, p$, o algoritmo k-Médias aloca cada objecto a um dos k grupos de forma a minimizar a soma dos quadrados dentro do agrupamento.

$$(3.c) \quad Sum(k) = \sum_{i=0}^n \sum_{j=0}^p (x_{i,j} - \overline{x_{k,j}})^2 \text{ onde}$$

$\overline{x_{k,j}}$ é a variável média j de todos os elementos do grupo k .

Às agregações obtidas em cada um dos níveis definidos para o histograma é aplicado o método das k-Médias, conforme o número de grupos distintos que se deseja criar. Para o domínio em estudo, foram analisados os resultados dos agrupamentos entre duas a nove partições e seleccionado o melhor número de grupos a analisar através do coeficiente de silhueta³.

³ Coeficiente de Silhueta originalmente denominado de *Silhouette Width*.

3.2.2 COEFICIENTE DE SILHUETA

O Coeficiente de Silhueta (Rousseeuw, 1987) [40] é uma medida de validação interna dos resultados de uma análise de agrupamentos e como tal tem em conta apenas o conjunto de dados e a partição do agrupamento como *inputs*, usando informação intrínseca dos dados para avaliar a qualidade do agrupamento. Esta medida para além de avaliar a validade do agrupamento pode também ser utilizada para decidir qual o melhor número de grupos a seleccionar.

Parte-se do cálculo do coeficiente de silhueta individualmente para cada exemplo. Posteriormente, obtém-se o coeficiente de silhueta médio para cada grupo e, finalmente, o coeficiente de silhueta médio global para todos os dados, dada pela média simples do valor de silhueta calculado para todos os objectos no conjunto de dados. Um **maior coeficiente de silhueta médio global** indica o **melhor agrupamento**. Portanto, o número de grupos com o máximo coeficiente de silhueta médio global é tomado como o número óptimo de grupos a considerar. De seguida, é apresentado um exemplo dos resultados obtidos através da medida em causa – Figura 3.4 e Figura 3.5 – comparativamente, no que respeita ao agrupamento obtido a partir das distâncias euclidianas e das correlações.

O valor de silhueta $S(i)$ mede o grau de confiança na atribuição de determinado grupo a uma observação em particular. Neste sentido, o valor de silhueta, representado pela fórmula que se segue, encontra-se no intervalo $[-1, 1]$ que deve ser maximizado. Um valor de silhueta próximo de 1 indica observações bem agrupadas no conjunto e próximo de -1, observações mal agrupadas. Com um valor de silhueta próximo de 0, as observações podem ser atribuídas da mesma forma a outro agrupamento mais próximo.

$$(3.d) \quad S(i) = \frac{b_i - a_i}{\max(b_i, a_i)} \quad \text{onde}$$

a_i é a distância média entre i e todas as outras observações no mesmo grupo e b_i é a distância média mínima do ponto i a todos os pontos noutro agrupamento, o agrupamento vizinho mais próximo, isto é,

$$(3.e) \quad b_i = \min_{C_k \in \mathcal{C} \setminus C(i)} \sum_{j \in C_k} \frac{dist(i,j)}{n(C_k)} \text{ em que}$$

$C(i)$ é o grupo que contém a observação i , $dist(i,j)$ é a distância entre as observações i e j , e $n(g)$ é a cardinalidade do grupo g .

```
> sw.6h04
```

	2	3	4	5	6	7
Dist.Euclideana	0.5622983	0.5062017	0.4823318	0.4351980	0.4008704	0.4062860
Correlações	0.6389852	0.6832272	0.5751936	0.6000179	0.5789515	0.5145134

	8	9
Dist.Euclideana	0.4179918	0.4375868
Correlações	0.5080519	0.5009007

Figura 3.4 – Coeficiente de Silhueta Médio por Número de Grupos - Perfis 6 Horas (2004)

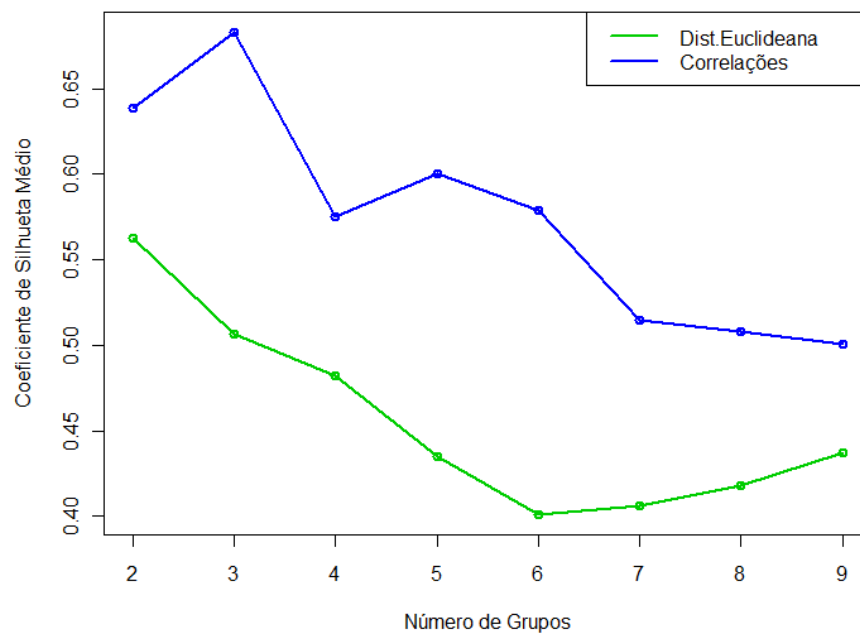


Figura 3.5 – Exemplo de Análise do Coeficiente de Silhueta Médio por Número de Grupos - Perfis 6 Horas (2004)

3.2.3 AGRUPAMENTO A PARTIR DAS CORRELAÇÕES

A par com a análise dos agrupamentos, a partir dos resultados agregados para cada um dos níveis do histograma, foram ainda analisados os resultados de agrupamento a partir das correlações dessas agregações.

A análise do agrupamento a partir das correlações é uma metodologia que reflecte a relação existente entre as tendências de subida ou descida do consumo de energia eléctrica, ao passo que o agrupamento dos dados segundo as distâncias euclidianas, apenas reflecte o distanciamento entre curvas características da evolução do consumo de energia. Neste sentido, com perspectivas de obter melhores resultados para o agrupamento dos perfis obtidos foi feito um estudo comparativo dos agrupamentos obtidos a partir das correlações entre dias, segundo a função de correlação de Pearson ρ [17], denominada no R por $cor()$. Temos que:

$$(3.f) \quad \rho = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{cov(X,Y)}{\sqrt{var(X) \cdot var(Y)}}, \quad \rho \in [-1, 1]$$

em que $x_{i=1,\dots,n}$ e $y_{i=1,\dots,n}$ são os valores para ambas as variáveis e

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^n x_i \quad \text{e} \quad \bar{y} = \frac{1}{n} \cdot \sum_{i=1}^n y_i$$

Esta associação linear entre duas variáveis assume valores entre -1 e 1, correspondentes à variação entre uma correlação negativa ou positiva perfeita entre as variáveis, sendo que 0 indica a não dependência linear entre as mesmas.

Aplicando esta metodologia aos dados em análise, organizados como apresentado exemplificativamente para os perfis de 6 horas na Tabela 3.3, são construídas tabelas de correlação – Tabela 3.4 – entre os dias em estudo para os anos de 2004 e 2005.

Dias	1	2	3	...	366
Períodos 6 Horas					
0-5 Horas	3720.329	3512.704	3847.708	...	4464.075
6-11 Horas	3396.787	4496.096	3890.742	...	5032.496
12-17 Horas	4072.150	5252.783	4392.221	...	5357.771
18-23 Horas	4928.425	5919.279	5578.375	...	5904.496

Tabela 3.3 – Dados originais agregados em Períodos de 6 Horas (2004)

Dias	1	2	3	...	366
1	1	0.7895323	0.9282536	...	0.8158503
2	0.7895323	1	0.9495478	...	0.9935525
3	0.9282536	0.9495478	1	...	0.9701611
...	...	...	...	...	...
366	0.8158503	0.9935525	0.9701611	...	1

Tabela 3.4 – Tabela de Correlações dos dados agregados em Períodos de 6 Horas (2004)

3.2.4 APRESENTAÇÃO DOS PERFIS

A partir dos valores obtidos para ambas as abordagens serão obtidos os agrupamentos entre duas a nove partições segundo o método das k-Médias enunciado anteriormente, escolhido o número de agrupamentos a analisar segundo o coeficiente de silhueta e realizado um estudo comparativo entre os resultados.

Graficamente é obtida uma análise comparativa para o agrupamento de dois perfis horários a partir das distâncias euclidianas ou das correlações – Figura 3.6 – sendo apresentado para cada dia (desdobrado em 24 horas⁴), do mês de Outubro exemplificativamente, o agrupamento do dia no grupo correspondente.

⁴ A escala do gráfico varia de 1 a 24, no entanto, a escala temporal dos dados correspondente é de 0 a 23 horas.

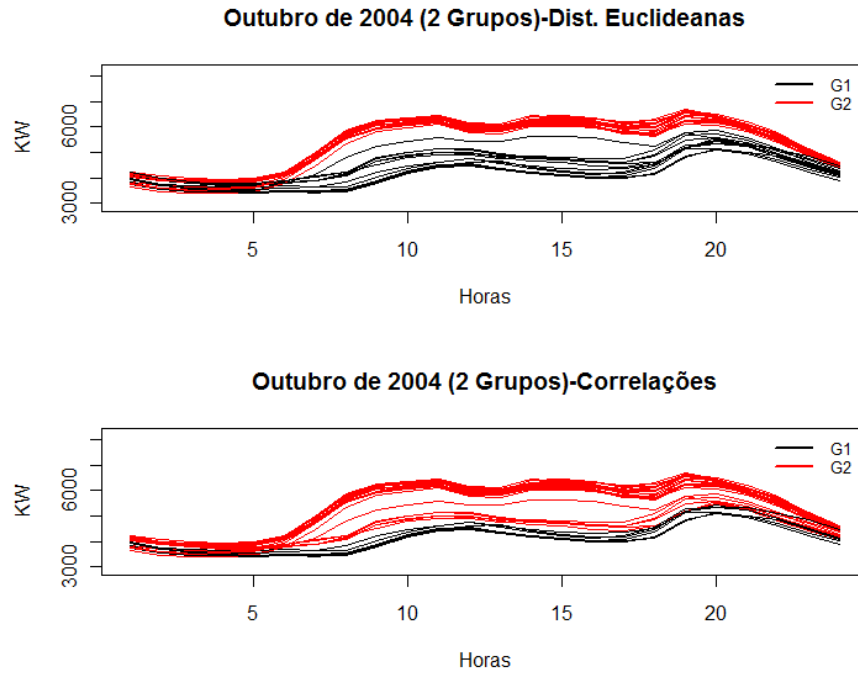


Figura 3.6 – Exemplo de Agrupamento k-Médias para Perfis Horários

Numa abordagem conclusiva podem ser obtidos os perfis médios horários de consumo de energia eléctrica verificados para cada diferente nível de granularidade, bastando aplicar a média aos valores observados em cada hora, dentro de cada conjunto.

Admitindo $X_{na,g}$ como a variável representativa do valor observado em determinado **nível de agregação na** e pertencente ao **grupo g** , temos que o perfil médio de consumo é dado pela variável $\bar{X}_{na,g}$ calculada da seguinte forma:

$$(3.g) \quad \bar{X}_{na,g} = \sum \frac{1}{n} X_{na,g} \text{ em que}$$

n diz respeito ao número de elementos que constituem o grupo, no nível de agregação $na = 1, 2, 3, 4, 5, 6, 7, 8$, correspondentes aos níveis de agregação definidos na Secção 3.2. Exclui-se, no entanto, o nível 9 (semestral) porque não foi possível aplicar o método de agrupamento a este nível de granularidade, visto apenas existirem duas observações para agregar em cada ano. O número de grupos k a analisar, dado pelos resultados do coeficiente de silhueta são os seguintes:

Para

na = 1	→	k = 1, 2, 3, 4
na = 2	→	k = 1, 2, 3, 4
na = 3	→	k = 1, 2, 3, 4
na = 4	→	k = 1, 2, 3
na = 5	→	k = 1, 2, 3
na = 6	→	k = 1, 2
na = 7	→	k = 1, 2
na = 8	→	k = 1, 2

Esta formulação deve ser seguida para o cálculo dos perfis médios segundo o nível de agregação em causa tanto para a abordagem do agrupamento a partir das distâncias euclidianas como a partir das correlações. Na Figura 3.7 são apresentados graficamente os dois perfis médios horários obtidos para cada abordagem.

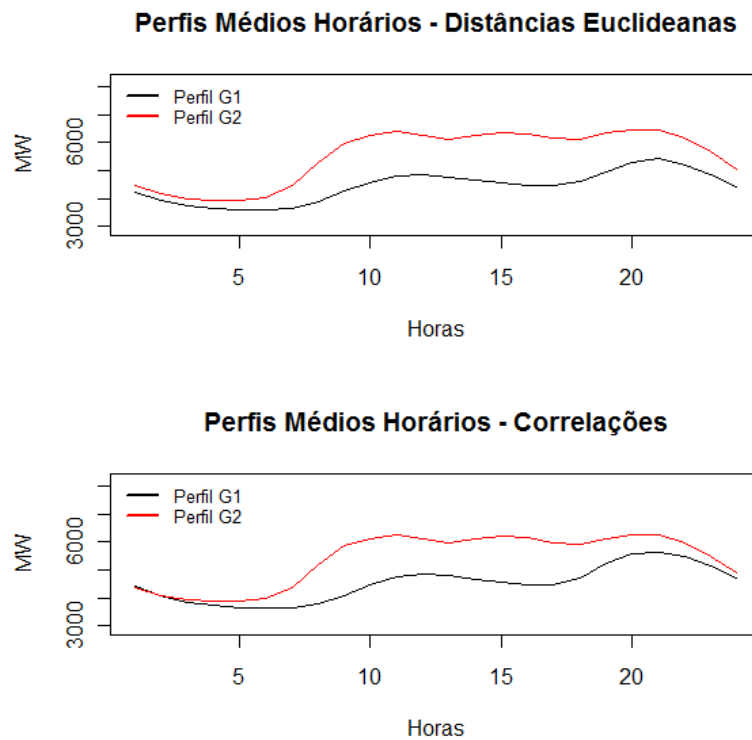


Figura 3.7 – Exemplo Perfis Médios Horários (2 Grupos)

3.2.5 CARACTERIZAÇÃO DOS PERFIS

Juntamente com estas informações gráficas é possível identificar outros elementos importantes para a caracterização dos grupos encontrados. O número de observações e a média dos agrupamentos são obtidos directamente como resultado do método das k-Médias. Por outro lado, os limites mínimo e máximo de agrupamento, assim como o erro cometido, foram determinados.

O limite mínimo ou máximo de agrupamento num determinado conjunto é dado pelo menor ou maior **valor observado**, respectivamente, para os elementos do **grupo g** (X_g).

$$(3.h) \quad \textit{Limite mínimo (ou máximo) de agrupamento no grupo } g \\ = \min (\text{ou max}) \{X_g\}$$

O erro cometido, em termos relativos, de utilizar os perfis em vez dos valores reais será calculado pelo somatório dos desvios entre a **média dos grupos** ($\bar{X}_g$) obtidos e os valores reais dos elementos correspondentes.

$$(3.i) \quad \textit{Erro do perfil} = \text{sum} \{\bar{X}_g - X_g\}$$

A definição de perfis imposta primeiramente parte de uma construção básica que pode sugerir variações diversas. Neste estudo são apresentadas duas variantes constituídas por perfis criados a partir da divisão traçado médio no agrupamento original, em traçados médios acima e abaixo da média desse agrupamento. Neste refinamento de perfis, o erro cometido será igualmente calculado pela soma dos desvios, mas desta vez, entre a média dos traçados acima ($\bar{T}_{acima}$) ou abaixo ($\bar{T}_{abaixo}$) da média e os valores reais acima (T_{acima}) ou abaixo (T_{abaixo}) da média do agrupamento, como apresentado de seguida pelas fórmulas (3.j) e (3.k).

$$(3.j) \quad \textit{Erro do perfil acima da média} = \text{sum} \{\bar{T}_{acima} - T_{acima}\}$$

$$(3.k) \quad \textit{Erro do perfil abaixo da média} = \text{sum} \{\bar{T}_{abaixo} - T_{abaixo}\}$$

Confrontando os erros antes e após o desdobramento é possível avaliar a evolução do erro após este procedimento, sendo importante saber se o erro aumentou, diminuiu ou se manteve. Neste sentido, foram comparados os erros absolutos antes (3.l) e após o desdobramento (3.m), sendo este último constituído pela soma dos erros absolutos após desdobramento acima e abaixo da média, correspondentes aos valores absolutos dos erros dos perfis acima e abaixo da média.

$$(3.l) \quad \textit{Erro Absoluto antes do desdobramento} = |\textit{sum} \{\bar{X}_g - X_g\}|$$

$$\begin{aligned} (3.m) \quad & \textit{Soma dos Erros Absolutos após o desdobramento} \\ &= \textit{Erro Absoluto acima da média} + \textit{Erro Absoluto abaixo da média} = \\ &= |\textit{sum} \{\bar{T}_{acima} - T_{acima}\}| + |\textit{sum} \{\bar{T}_{abaixo} - T_{abaixo}\}| \end{aligned}$$

3.3 DISTRIBUIÇÃO DOS DIAS DA SEMANA

Tirando partido dos dados recolhidos é importante, para o domínio em causa, o estudo da distribuição dos dias da semana no que respeita aos perfis diários encontrados no nível três do histograma por níveis. Para isso foram criados agrupamentos entre dois a sete grupos, representativos do número de dias da semana, utilizando novamente o método das k-Médias sobre os dados em causa, para os dois anos em separado.

Os resultados obtidos com a aplicação desta metodologia permitem-nos concluir acerca do número de elementos que constituem cada agrupamento, bem como das médias de consumo de energia eléctrica correspondentes, mas principalmente, dos dias da semana que caracterizam cada grupo definido.

A Figura 3.8 apresenta a distribuição dos dias da semana do ano de 2005 por quatro grupos, ao passo que a Tabela 3.5 complementa a informação com o número de dias da semana que constituem cada conjunto e a respectiva percentagem dentro de cada agrupamento.

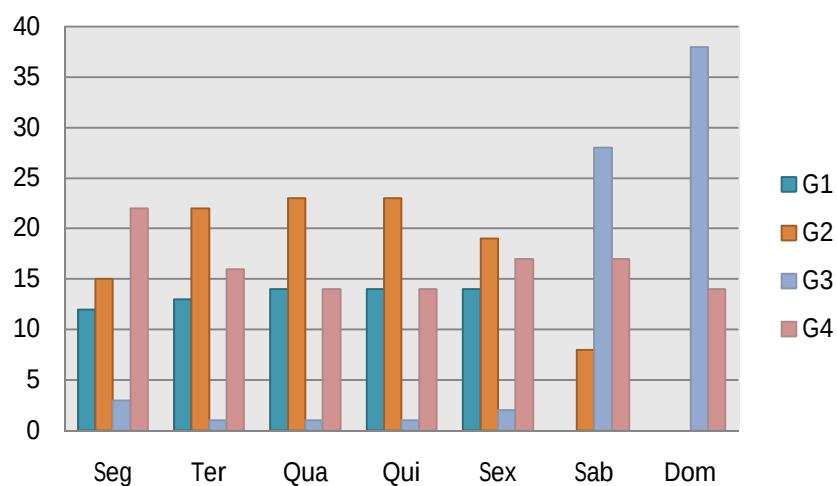


Figura 3.8 – Exemplo da Distribuição Diária por Grupo (2005)

	G1	% G1	G2	% G2	G3	% G3	G4	% G4
Seg	3	6%	28	19%	21	16%	0	0%
Ter	1	2%	31	21%	19	14%	1	3%
Qua	1	2%	30	20%	21	16%	0	0%
Qui	2	4%	30	20%	19	14%	1	3%
Sex	1	2%	30	20%	21	16%	0	0%
Sab	13	25%	0	0%	31	23%	9	29%
Dom	32	60%	0	0%	0	0%	20	65%
Total	53	100%	149	100%	132	100%	31	100%

Tabela 3.5 – Exemplo da Distribuição de Dias da Semana (2005)

3.3.1 DIAS REPRESENTATIVOS

De forma a complementar o estudo apresentado e usufruindo da partição entre dois a sete grupos foram encontrados os dias mais representativos de cada um dos agrupamentos, tendo sido calculados pela diferença mínima absoluta entre a média do agrupamento e o valor observado nos perfis diários que constituem cada agrupamento.

$$(3.n) \quad \text{Dia representativo do grupo } g = \min \{ \bar{X}_g - X_g \}$$

São ainda analisadas representações gráficas conjuntas dos dias representativos obtidos para as diferentes partições – Figura 3.9.

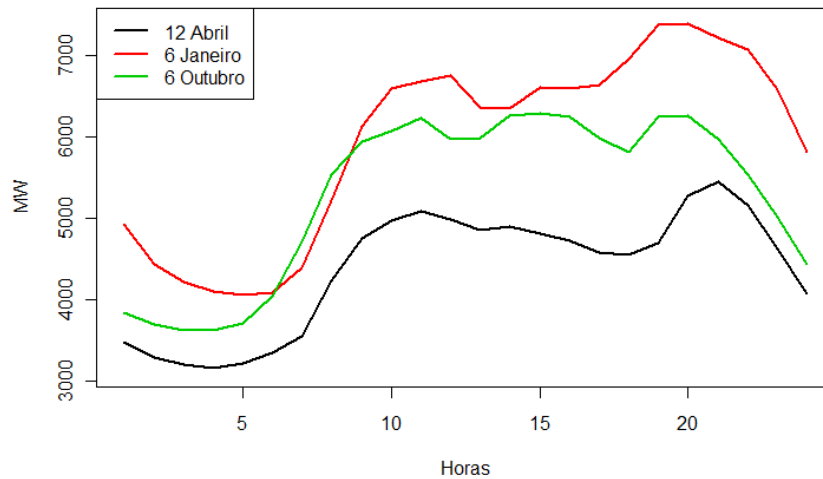


Figura 3.9 – Exemplo dos Dias Representativos de 2004 (3 Grupos)

3.4 FERRAMENTAS

Para a execução das tarefas propostas neste trabalho, a principal ferramenta utilizada para a importação dos dados, a sua preparação, aplicação de metodologias e apresentação de resultados foi o *software* R 2.10.1, tendo sido também utilizado o Microsoft Office Excel para algumas tarefas auxiliares de cariz menos técnico.

A aplicação aos dados em estudo das metodologias acima descritas serviu como ponto de partida para a materialização das conclusões retiradas no estudo do consumo de energia eléctrica divulgado de seguida.

4 ESTUDO DO CONSUMO DE ENERGIA ELÉCTRICA

O estudo do consumo de energia eléctrica levado a cabo encontra-se sintetizado e estruturado em cinco componentes primordiais, iniciando-se pela análise exploratória dos dados, a aplicação de histogramas por níveis aos dados e o refinamento dos perfis encontrados dessa forma, passando pela distribuição dos dias da semana nos agrupamentos determinados, assim como os dias representativos dos conjuntos.

4.1 ANÁLISE EXPLORATÓRIA DE DADOS

Através técnicas quantitativas e gráficas é feita uma aproximação aos dados em análise neste estudo, no sentido de obter de informações estruturais relevantes. Neste sentido, é levada a cabo uma análise preliminar, descrita a preparação dos dados e estudada a auto-correlação dos dados, visto estarmos perante uma série temporal.

4.1.1 ANÁLISE PRELIMINAR

O conjunto de dados, objecto de estudo, foi denominado por *ren* e encontra-se organizado numa estrutura com 70176 observações descritas por **sete variáveis**, correspondentes ao **ano**, **mês**, **dia**, **dia da semana**, **hora**, **minuto** e **valor observado** – Figura 4.1 – de energia eléctrica fornecida para distribuição.

Os dados a analisar dizem respeito ao período de tempo decorrido entre 1 de Janeiro de 2004 e 31 de Dezembro de 2005, sendo as observações originalmente recolhidas nos períodos de 0, 15, 30 e 45 minutos da amplitude temporal referida. Estamos perante uma sequência ordenada de dados recolhidos continuamente ao longo do tempo – Figura 4.2 – apesar de finita. Nos Apêndices A e B são apresentados os calendários e as listagens de feriados dos anos em análise.

```
> str(ren)
'data.frame': 70176 obs. of 7 variables:
 $ Year : int 2004 2004 2004 2004 2004 2004 2004 2004 2004 2004 2004 ...
 $ Month : int 1 1 1 1 1 1 1 1 1 1 1 ...
 $ Day : int 1 1 1 1 1 1 1 1 1 1 1 ...
 $ Weekday: Factor w/ 7 levels "Fri","Mon","Sat",...: 5 5 5 5 5 5 5 5 5 5 5 ...
 $ Hour : int 0 0 0 0 1 1 1 1 2 2 ...
 $ Minute : int 0 15 30 45 0 15 30 45 0 15 ...
 $ Value : num 4402 4381 4334 4286 4200 ...
```

Figura 4.1 – Sumário Global do conjunto de dados ren

```
> head(ren)
  Year Month Day Weekday Hour Minute Value
1 2004     1  1    Thu    0      0 4402.5
2 2004     1  1    Thu    0     15 4381.3
3 2004     1  1    Thu    0     30 4334.2
4 2004     1  1    Thu    0     45 4285.9
5 2004     1  1    Thu    1      0 4199.5
6 2004     1  1    Thu    1     15 4137.6
> tail(ren)
  Year Month Day Weekday Hour Minute Value
70171 2005    12 31    Sat   22     30 5488.9
70172 2005    12 31    Sat   22     45 5424.9
70173 2005    12 31    Sat   23      0 5424.0
70174 2005    12 31    Sat   23     15 5382.9
70175 2005    12 31    Sat   23     30 5290.5
70176 2005    12 31    Sat   23     45 5189.2
```

Figura 4.2 – Primeiras e últimas linhas do conjunto de dados ren

Numa primeira abordagem ao conjunto de valores observados, da análise das estatísticas apresentadas – Figura 4.3 – pode-se concluir essencialmente que na totalidade dos **dois anos** foram consumidos, em **média, 5325 MegaWatts (MW)** de energia eléctrica por cada período de 15 minutos, variando entre um mínimo de 2872 MW e um máximo de 8528 MW.

```
> summary(ren$Value)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 2872   4327   5295   5325   6222   8528
```

Figura 4.3 – Estatísticas Descritivas Básicas da Variável Valor Observado

4.1.2 PREPARAÇÃO DOS DADOS

Com o objectivo de tornar mais fácil o seu estudo e manipulação, os dados foram divididos nos anos correspondentes e denominados de ren04 e ren05, com um total de 35136 e 35040 observações, respectivamente. É de salientar que a diferença de 96 observações entre os dois anos reside no facto de 2004 ter sido um ano bissexto, logo com um dia a mais que o posterior ($24 \text{ horas} \times 4 \text{ períodos de 15 minutos por hora} = 96$).

Ao analisar graficamente a evolução do consumo de energia eléctrica para os dois anos em separado – Figura 4.4 – não é possível retirar grandes conclusões acerca do padrão de variação do consumo, podendo este considerar-se relativamente semelhante.

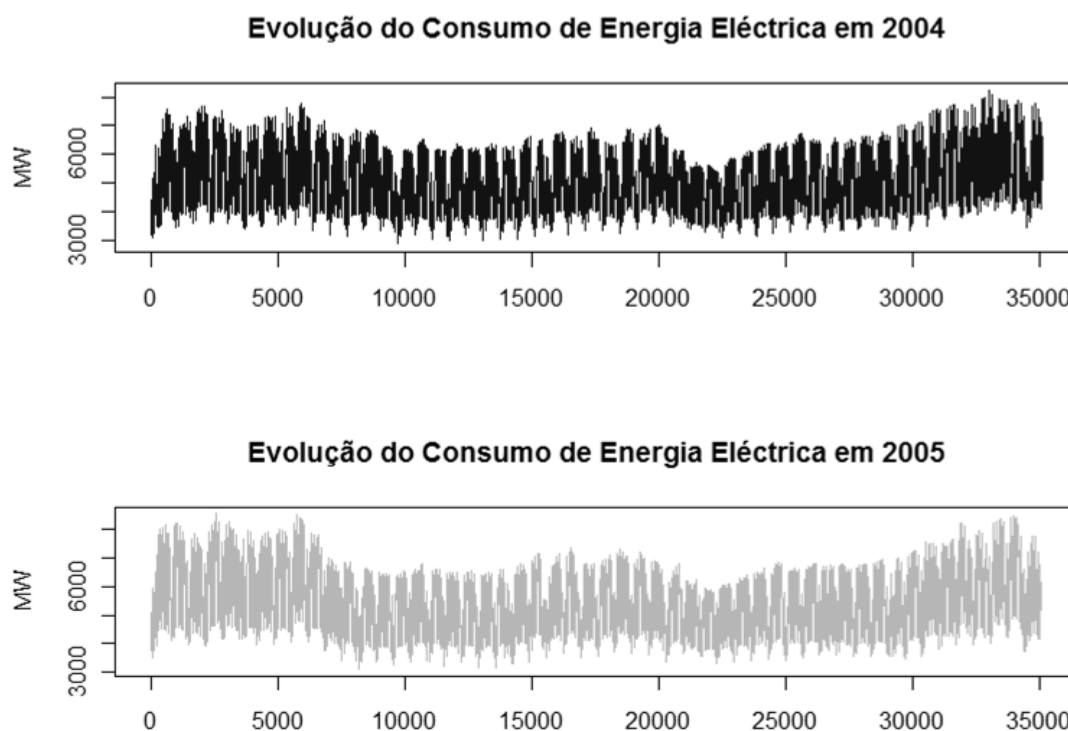


Figura 4.4 – Evolução do Consumo de Energia Eléctrica (2004 e 2005)

No entanto, a sumarização em termos de estatísticas descritivas básicas dos dados para cada ano – Figura 4.5 – indica-nos que de **2004 para 2005** houve, na generalidade, um **aumento da energia eléctrica consumida**, passando de um mínimo de 2872 MW em 2004 para 3131 MW em 2005, assim como de um máximo de 8250 MW para 8258 MW, respectivamente. O mesmo se reflecte no aumento da média de energia consumida em cerca de 300 MW. As amplitudes de variação de 5378 MW para 2004 e de 5127 MW para 2005, apesar de idênticas, registam um valor inferior no último ano.

```
> summary(ren04$Value)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 2872   4180   5143   5176   6100   8250
> summary(ren05$Value)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 3131   4476   5444   5473   6353   8528
```

Figura 4.5 – Estatísticas Descritivas Básicas do Valor Observado (2004 e 2005)

4.1.3 ESTUDO DA AUTO CORRELAÇÃO DOS DADOS

Em termos de perfis horários verificados no mês de Janeiro de 2004, para os marcos temporais correspondentes ao período das 0 às 5 horas, a correlação da variável com ela mesma é significativa e positiva, apesar de decrescente. Entre as 8 e as 15 horas o consumo de energia eléctrica encontra-se negativamente correlacionado. A partir das 18 e até às 23 horas, a variável em estudo correlaciona-se positiva e crescentemente com ela própria. Para as restantes horas (6, 7, 16 e 17), não são assumidos valores de auto-correlação considerados significativos.

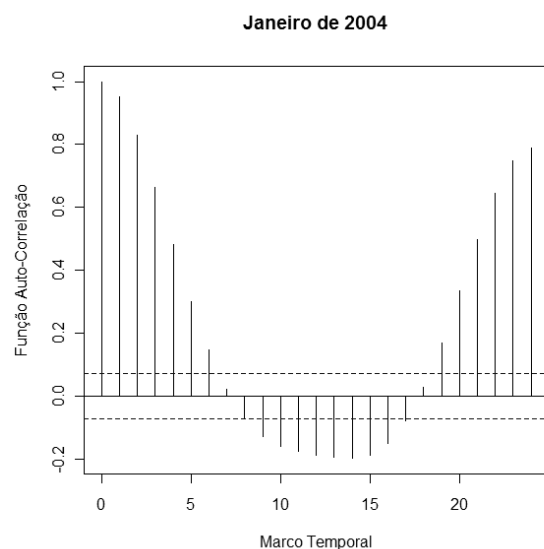


Figura 4.6 – Exemplo Auto-Correlação de Perfis Horários

O consumo de energia eléctrica no mês de Abril de 2004, para os perfis correspondentes aos períodos entre as 0 e as 5 horas e as 18 e as 23 horas é auto-correlacionado positivamente, apesar deste último período apresentar uma correlação mais forte. Os restantes dois períodos de 6 horas não apresentam uma auto-correlação significativa.

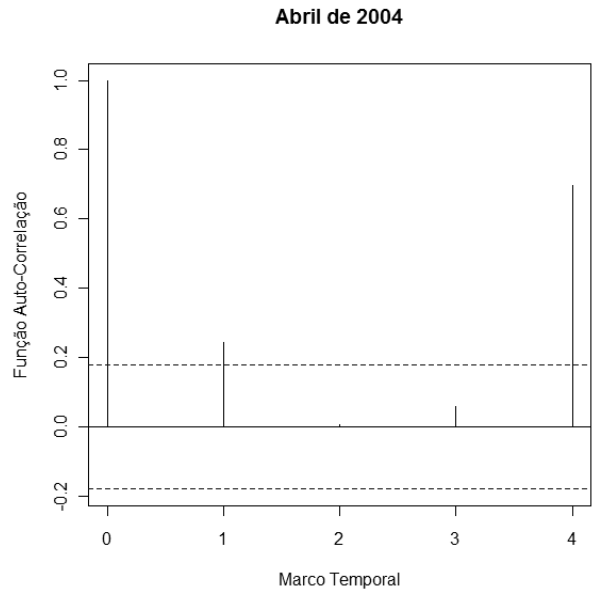


Figura 4.7 – Exemplo Auto-Correlação de Perfis 6 Horas

Relativamente ao mês de Junho de 2005, apenas se verifica para os dias 1, 3, 4, 7, 14, uma correlação relativamente significativa da variável consumo de energia eléctrica com ela mesma, sendo os restantes dias não significativos em termos de auto-correlação.

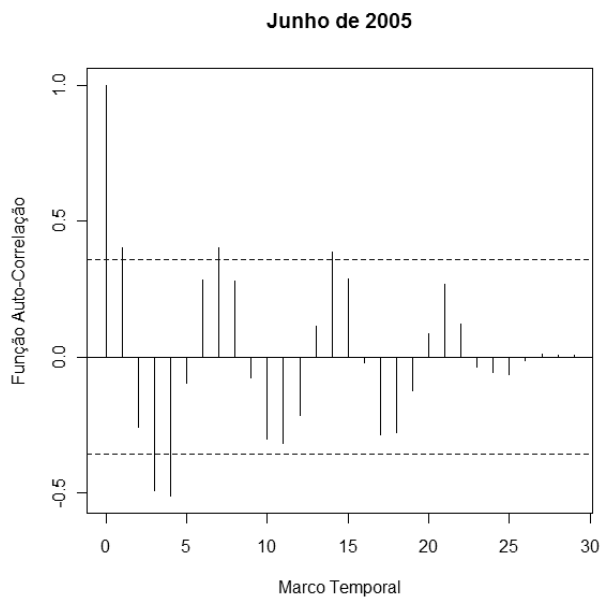


Figura 4.8 – Exemplo Auto-Correlação de Perfis Diários

Em termos de perfis de dias da semana, para o ano de 2005, os dias que apresentam uma maior auto-correlação para o consumo de energia eléctrica são as segundas-feiras, os sábados e os domingos, não sendo os restantes dias da semana significativos.

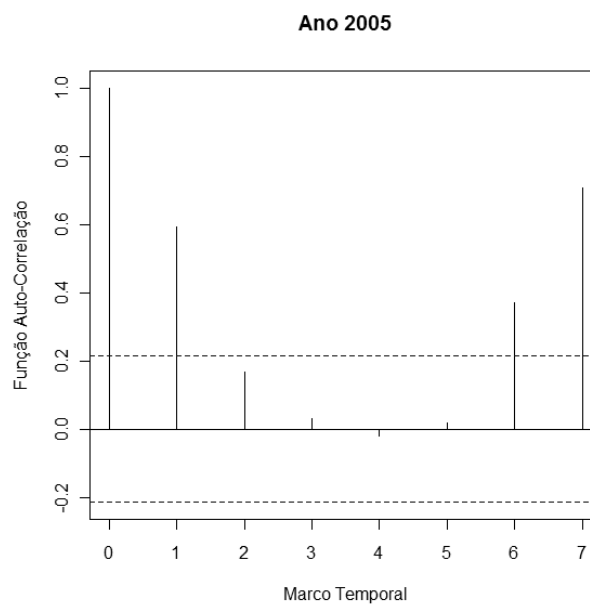


Figura 4.9 – Exemplo Auto-Correlação de Perfis Dias da Semana

Analisando os perfis Semana e Fim-de-Semana para o ano de 2004 verifica-se que o consumo de energia eléctrica dos dias de semana se encontram correlacionados negativamente, ao passo que os fins-de-semana se auto-correlacionam positiva e mais significativamente.

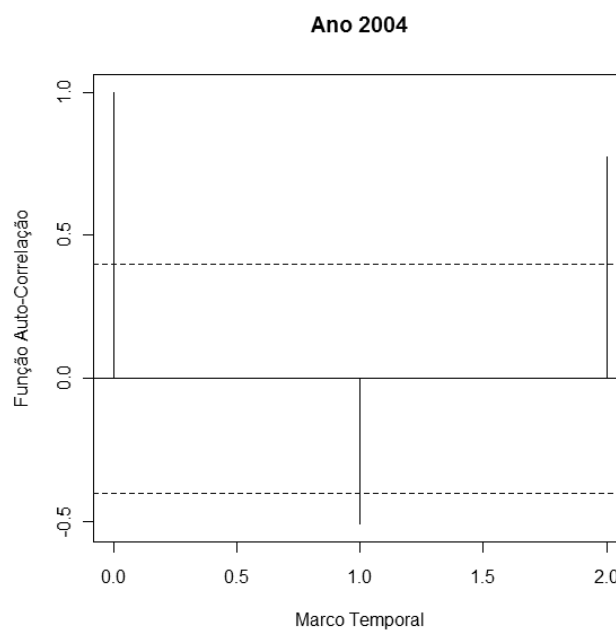


Figura 4.10 – Exemplo Auto-Correlação de Perfis Semana e Fim-de-Semana

No que respeita aos meses do ano de 2005 não se assinala nenhum valor de auto-correlação significativa, o mesmo se verificando para os perfis trimestrais e semestrais dos anos de 2004 e 2005.

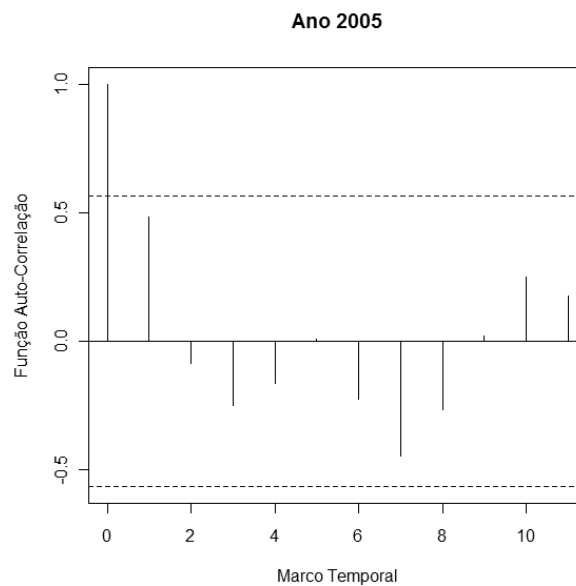


Figura 4.11 – Exemplo Auto-Correlação de Perfis Mensais

Realizado um estudo exploratório aos dados, foram identificados os traços gerais característicos. O próximo passo consiste basicamente na estruturação dos dados para a obtenção, análise e caracterização de perfis de consumo, sem esquecer que, de acordo com as granularidades temporais em causa.

4.2 APLICAÇÃO DE HISTOGRAMA POR NÍVEIS AOS DADOS

Com o intuito de estudar os perfis de utilização de energia eléctrica para diferentes horizontes temporais será criado um histograma por níveis, de forma a agregar as observações, utilizando-se a média como estatística de agregação, tal como descrito na Secção 3. Esses resultados de evolução do consumo são interpretados e complementados com o estudo dos agrupamentos correspondentes aos perfis de consumo.

4.2.1 Perfis Horários (Nível 1)

Os dados originais são apresentados, como referido anteriormente, em períodos de 15 minutos. Neste sentido, no nível 1 são agrupados, por média, os valores observados em cada hora do período temporal considerado, obtendo-se assim perfis horários de consumo de energia eléctrica. Dado o elevado volume de dados envolvido e para facilitar a sua interpretação, a elaboração deste primeiro nível foi repartida em meses. A Figura 4.12 que se segue exemplifica os resultados obtidos para as 24 horas do primeiro dia do mês de Janeiro de 2004 num primeiro nível de agregação. No entanto, para que seja possível retirar, intuitivamente, algumas conclusões foram analisadas as representações gráficas correspondentes, como é o exemplo da Figura 4.13 apresentada adiante, onde estão assinalados os períodos de menor consumo de energia.

```
> jan04
, , 1

      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]      [,8]
Hour    0.000     1.000     2.00     3.00     4.00     5.000     6.00     7.000
Value 4350.975 4095.275 3785.45 3511.25 3338.65 3240.375 3211.55 3188.175
      [,9]     [,10]     [,11]     [,12]     [,13]     [,14]     [,15]     [,16]     [,17]
Hour    8.0      9.0      10.00     11.00     12.0      13.000     14.000     15.000     16.000
Value 3091.1 3267.4 3617.35 4005.15 4213.2 4062.125 3914.475 3901.825 3935.125
      [,18]     [,19]     [,20]     [,21]     [,22]     [,23]     [,24]
Hour   17.00    18.000    19.00     20.000    21.000    22.000    23.00
Value 4406.15 4938.825 5073.65 5136.225 5059.425 4857.675 4504.75
      (...)

```

Figura 4.12 – Exemplo de Agregação para o Nível 1

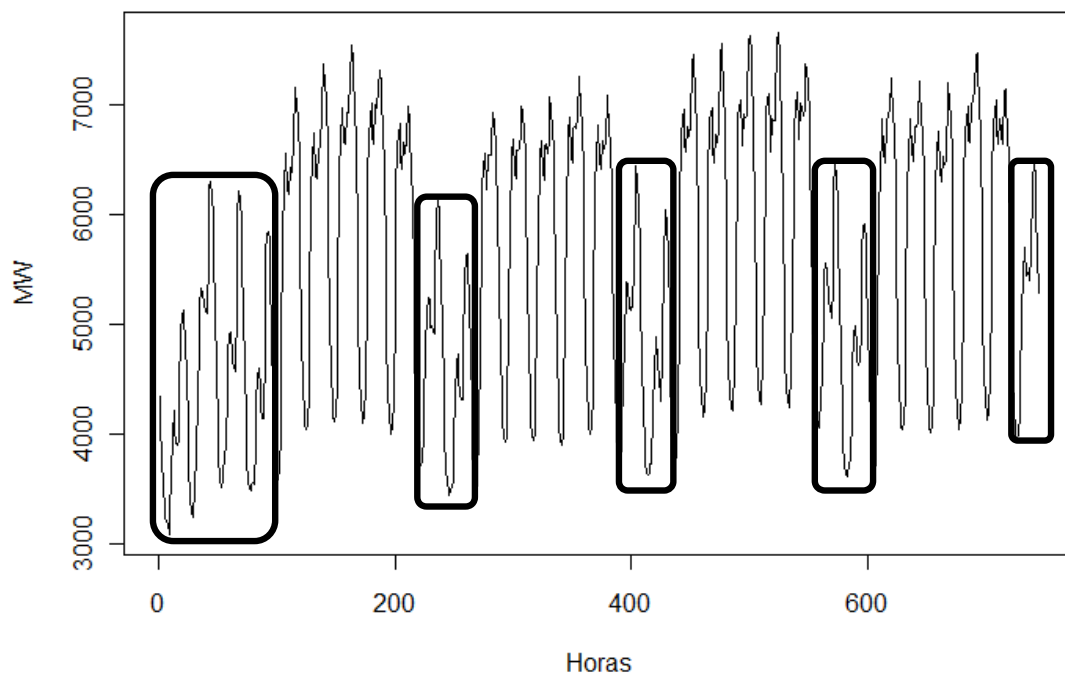


Figura 4.13 – Evolução do Consumo de Energia Eléctrica em Janeiro de 2004 (Horária)

Analisando a série de observações agregadas por hora pode identificar-se um padrão evolutivo que se repete ao longo dos vários meses de ambos os anos. O **fim** do mês de **Dezembro** e o **início** do mês de **Janeiro** correspondentes à época festiva de **passagem de ano**, juntamente com as épocas do **Carnaval**, **Páscoa** e **Natal**, bem como **feriados nacionais** e **fins-de-semana** são caracterizados pela **diminuição** no consumo de energia eléctrica. Estas evidências devem-se ao facto de existir um decréscimo do consumo nos períodos assinalados e uma consequente diminuição das necessidades de distribuição de energia. No entanto, existem **picos de variação semelhantes entre períodos diários**, algo difíceis de caracterizar para este nível mas que serão explorados mais adiante.

De acordo com a metodologia definida na Secção anterior, com a aplicação do método de agrupamento aos dados obtidos no primeiro nível do histograma (ou seja, através das distâncias euclidianas) e aos dados resultantes da aplicação de correlações é analisado comparativamente o coeficiente de silhueta para o número de grupos entre dois e nove, exemplificativamente para o ano de 2004 – Figura 4.14 e Figura 4.15.

```

> sw.hora04
      2      3      4      5      6      7
Dist.Euclidean 0.5396435 0.4909321 0.4658177 0.4650795 0.4056387 0.3759690
Correlações    0.6241594 0.6712717 0.6748658 0.5149831 0.4668380 0.4626796
      8      9
Dist.Euclidean 0.3484198 0.3921393
Correlações    0.4629331 0.4240992

```

Figura 4.14 – Coeficiente de Silhueta Médio por Número de Grupos - Perfis Horários (2004)

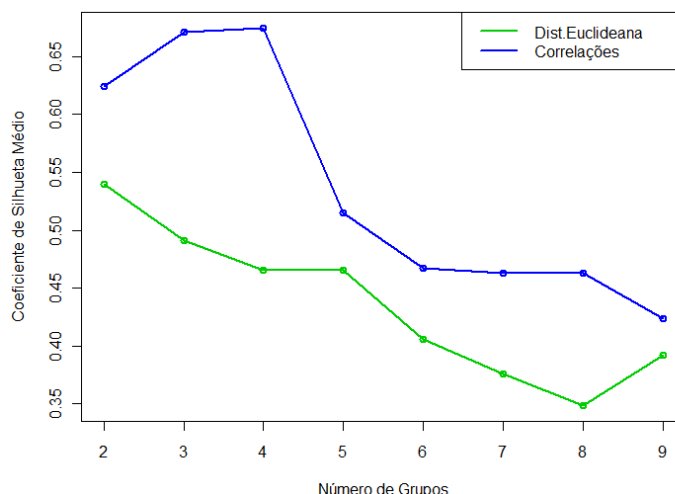


Figura 4.15 – Análise do Coeficiente de Silhueta Médio por Número de Grupos - Perfis Horários (2004)

No conjunto das abordagens em estudo, o valor ótimo do coeficiente de silhueta é verificado para o agrupamento em 4 partições obtido a partir das correlações. Em termos médios globais, como podemos comprovar na Figura 4.16, verifica-se a tendência de que são obtidos melhores resultados a partir do agrupamento das correlações.

```

> mean(sw.hora04['Dist.Euclidean',])
[1] 0.4354550
> mean(sw.hora04['Correlações',])
[1] 0.5377287

```

Figura 4.16 – Coeficiente de Silhueta Médio Global

É apresentada, a título elucidativo, uma representação gráfica para o mês de Agosto de 2004 – Figura 4.17 – onde são confrontados os agrupamentos obtidos a partir da partição em quatro grupos de acordo com as distâncias euclidianas e as correlações.

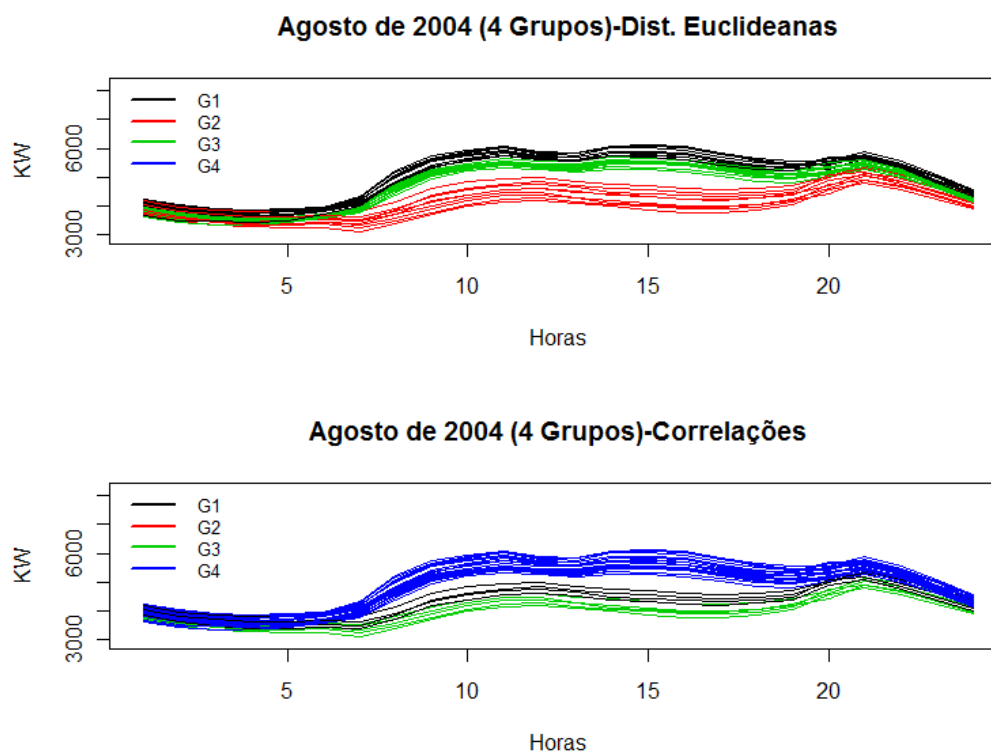


Figura 4.17 – Agrupamento k-Médias para Perfis Horários

As duas abordagens envolvem agrupamentos em apenas três partições, no que diz respeito ao mês de Agosto de 2004, visualmente distintas, sendo que o agrupamento a partir das distâncias euclidianas pressupõe a existência de um grupo em que os consumos são inferiores (G2), outros dois com consumos intermédios (G3) e superiores (G1). Por outro lado, o agrupamento a partir das correlações subdivide o grupo de consumos inferiores em dois agrupamentos (G3 e G1), reunindo os outros dois de consumos intermédios e superiores num só (G4). A análise visual deste tipo de conclusões é intuitiva mas torna-se exaustiva se a efectuarmos para os restantes meses do ano. É neste sentido que se torna útil recorrer a medidas que caracterizem os resultados de agrupamento obtidos – Tabela 4.1.

Ano 2004		4 Perfis - Dist. Euclideanas				4 Perfis - Correlações			
Perfis Horários	Grupos	G1	G2	G3	G4	G1	G2	G3	G4
Limites de classificação	Mínimo	3317,825	2908,925	3242,275	3529,375	3152,025	2908,925	3053,875	3317,825
	Máximo	6952,275	5855,075	6988,6	8214,125	8214,125	6988,6	7167,975	6952,275
Tamanho Grupos		149	69	59	89	130	30	55	151
Média Grupos		5319,585	4208,275	4846,423	5904,946	5521,258	4597,897	4434,479	5264,049
Erro Relativo por Grupo		2,67E-10	2,09E-10	4,23E-10	-1,41E-10	-1,97E-10	-3,05E-11	-4,11E-11	-3,76E-10
Erro Absoluto Total		1,04E-09				6,44E-10			

Tabela 4.1 – Caracterização do Agrupamento em 4 Grupos (Perfis Horários)

O agrupamento em quatro partições pressupõe perfis de consumo algo distintos, nomeadamente no número de dias que constituem os perfis bem como na média dos grupos, sendo que, globalmente e em termos absolutos, o **erro** cometido pelos perfis encontrados a partir do agrupamento das **correlações é inferior** ao cometido a partir das distâncias euclideanas. De acordo com os perfis definidos são obtidos os perfis horários médios de consumo de energia eléctrica – Figura 4.18.

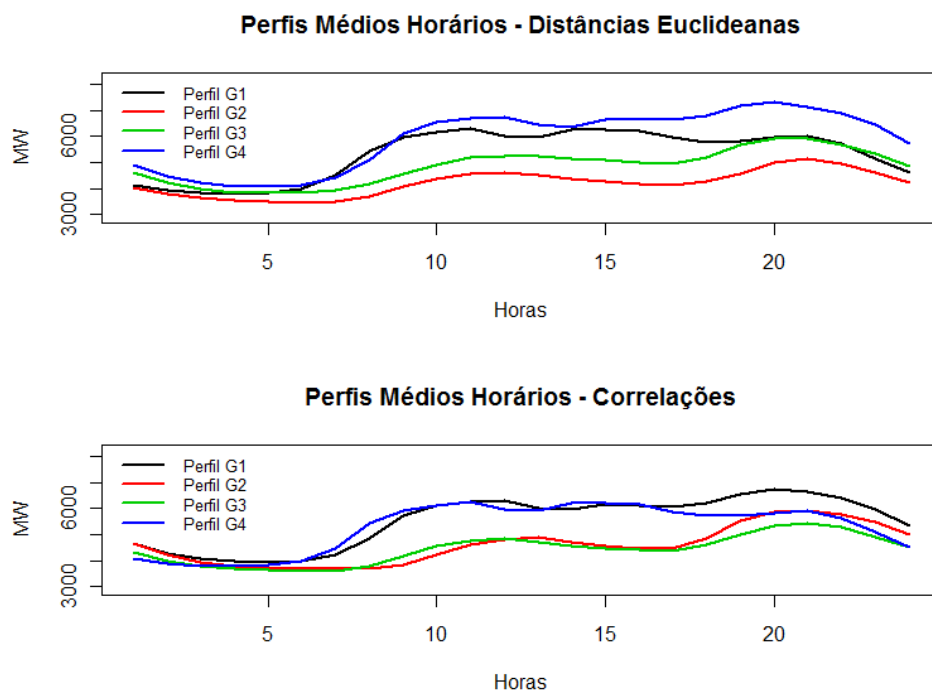


Figura 4.18 – Perfis Médios Horários

Os perfis médios horários obtidos seguem um comportamento semelhante, existindo, no entanto, para a abordagem das correlações, uma aproximação entre os perfis G1 e G4 com níveis de consumo superiores e os perfis G2 e G3 com níveis inferiores.

4.2.2 Perfis 6 Horas (Nível 2)

No sentido de compreender as oscilações diárias do consumo de energia eléctrica, foram consideradas agregações por períodos de 6 horas, continuando a ter em conta a divisão em meses. Os períodos definidos são entre as 0 e 5 horas, 6 e 11 horas, 12 e 17 horas, 18 e 23 horas. É apresentada como exemplo uma representação gráfica resultante de agregações para este nível de granularidade – Figura 4.19.

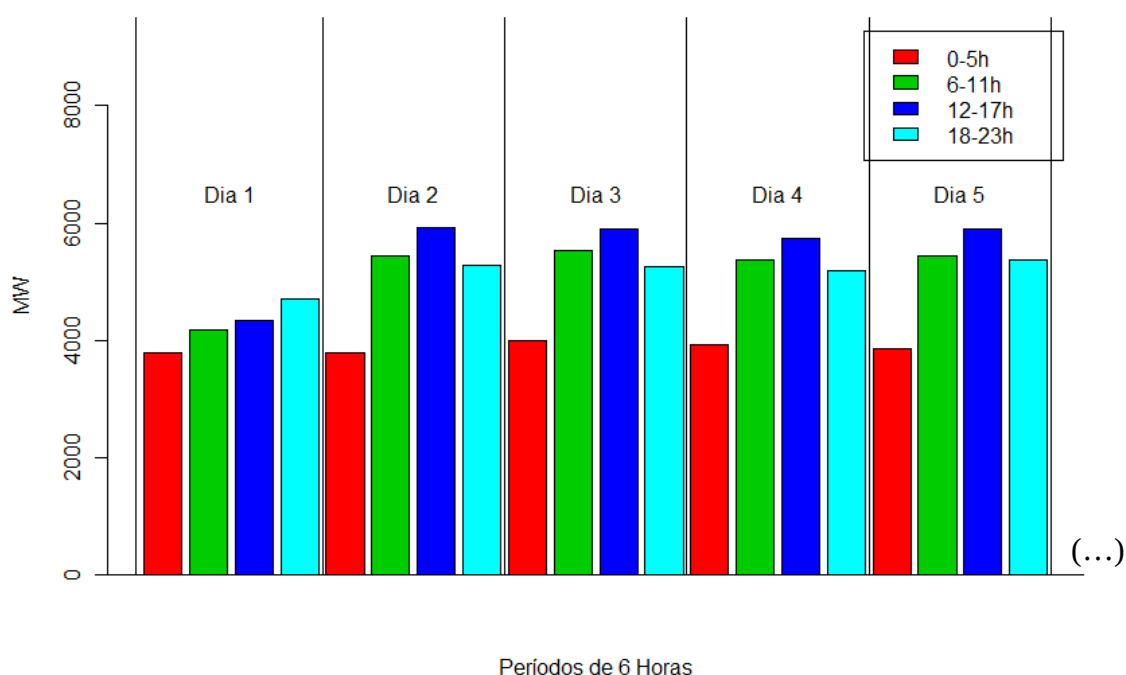


Figura 4.19 - Evolução do Consumo de Energia Eléctrica em Agosto de 2004 (Períodos de 6 Horas)

De um modo geral, é reconhecido que durante todo o ano, o período entre as **0 e as 5 horas** é caracterizado pelo **menor consumo** de energia eléctrica por ser a altura do dia mais calma, em que a maioria das pessoas está a descansar. **Segue-se** o período entre as **6 e as 11 horas** em que existem mais pessoas acordadas e em actividade laboral. Por outro lado, no início e fim do ano, nomeadamente nos meses de **Janeiro, Fevereiro, Dezembro, inícios de Março e fins de Novembro**, a **tendência** é para que o período das **18 às 23 horas supere** maioritariamente o período das **12 às 17 horas** em termos de consumo de energia eléctrica. Esta tendência é justificada pelo facto de serem meses mais frios e este período horário corresponder a uma altura do dia em que as pessoas estão em casa e recorrem a aquecimento. No entanto, o período das **12 às 17 horas supera** o das **18 às 23 horas**, em épocas mais amenas, correspondente aos **finais de Março, meses de Abril a Outubro e inícios de Novembro**. É de acrescentar, que o padrão de evolução do consumo de energia eléctrica é semelhante para ambos os anos em estudo.

Como exemplo da aplicação do método das k-Médias aos dados agregados no segundo nível do histograma para a obtenção de perfis, são apresentadas as representações gráficas resultantes para quatro agrupamentos – Figura 4.20 – definido como o número de grupos óptimo para o agrupamento a partir das correlações, de acordo com o coeficiente de silhueta. Em termos médios globais, os resultados da mesma medida para o agrupamento a partir das correlações são mais favoráveis do que para o agrupamento a partir das distâncias euclidianas.

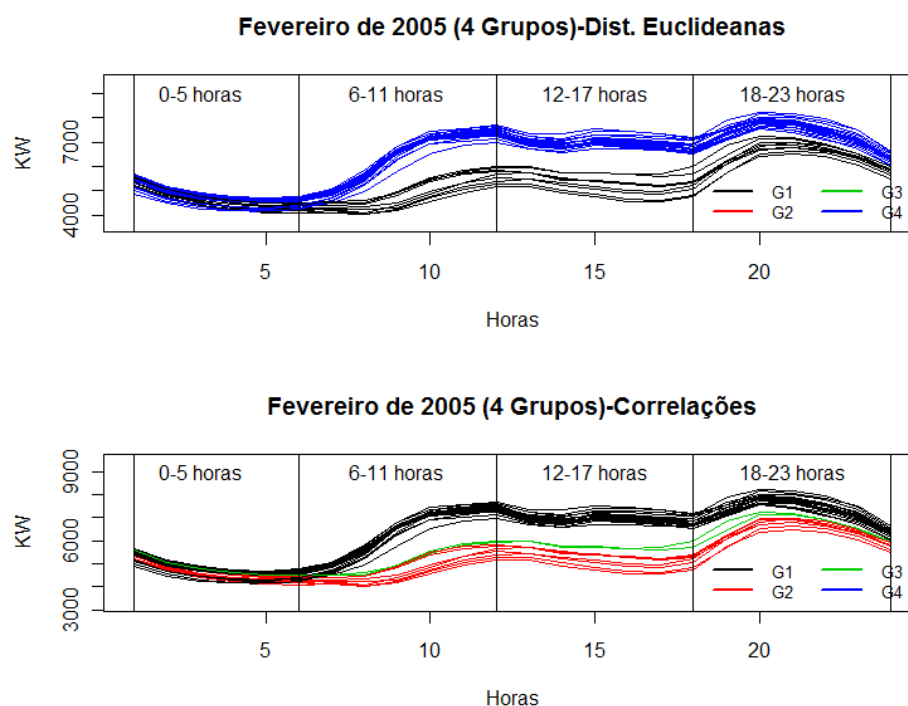


Figura 4.20 – Agrupamento k-Médias para Perfis 6 Horas

Da análise exemplificativa da Figura 4.20, no que respeita ao mês de Fevereiro de 2005, verifica-se para o agrupamento obtido a partir das correlações a divisão em dois grupos (G2 e G3) de consumos inferiores que correspondem à união num grupo só (G1) no agrupamento a partir das distâncias euclideanas, mantendo-se a tendência de existência de um grupo de observações com consumos superiores para ambas as abordagens.

Ano 2005		4 Perfis - Dist. Euclideanas				4 Perfis - Correlações			
Perfis 6 Horas	Grupos	G1	G2	G3	G4	G1	G2	G3	G4
Limites de classificação	Mínimo	4157,817	3484,233	3667,85	4112,663	3519,537	3484,233	3521,387	3667,85
	Máximo	7038,083	5843,971	6962,087	7946,346	7946,346	7038,083	6811,529	6962,087
Tamanho Grupos		39	79	166	81	132	32	52	149
Média Grupos		5325,828	4518,276	5517,926	6384,289	5872,349	5127,173	4593,907	5501,017
Erro Relativo por Grupo		-7,19E-11	1,06E-10	-1,32E-10	-1,25E-10	-6,82E-12	-7,73E-12	-5,91E-12	-1,09E-11
Erro Absoluto Total		4,35E-10				3,14E-11			

Tabela 4.2 – Caracterização do Agrupamento em 4 Grupos (Perfis 6 Horas)

Com diferentes números de dias que os constituem, bem como respectivas médias de consumo dos agrupamentos, verifica-se um **comportamento mais próximo da realidade** no que respeita aos perfis de consumo obtidos a partir da aplicação do método das k-Médias às **correlações** dos dados originais. São ainda apresentados os perfis médios de consumo de energia eléctrica para uma granularidade de 6 horas – Figura 4.21.

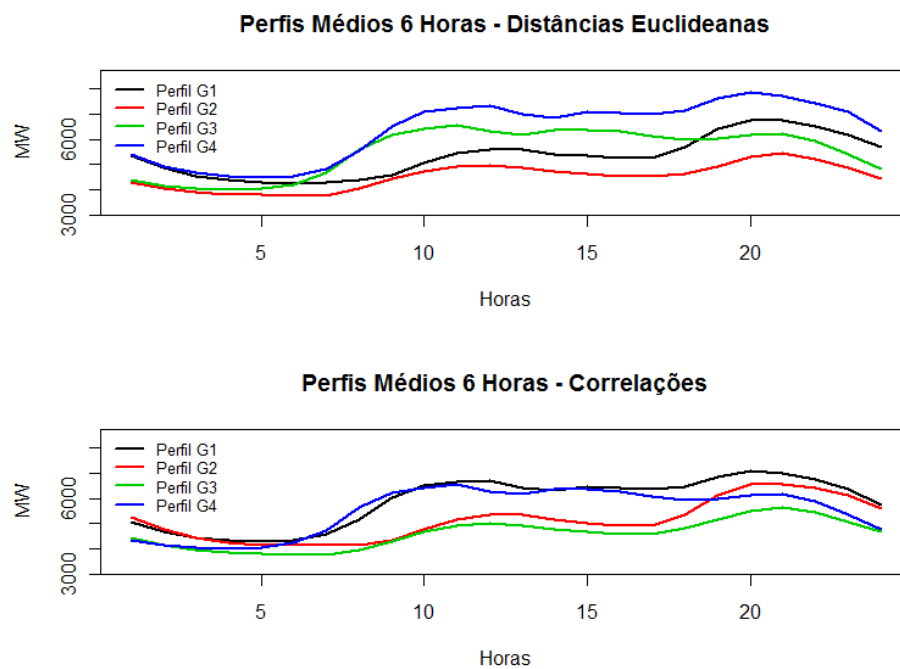


Figura 4.21 – Perfis Médios 6 Horas

Os perfis médios de 6 horas encontrados revelam uma aproximação entre os dois perfis de nível superior (G1 e G4) e os de nível inferior (G2 e G3), para a abordagem das correlações relativamente à anterior.

4.2.3 Perfis Diários (Nível 3)

Para este nível, os valores reunidos anteriormente são reagrupados por dia. Com o aumento da granularidade das observações pode reconhecer-se mais facilmente as variações de consumo de energia eléctrica ao longo de cada mês, nomeadamente os **aumentos em dias de semana e as diminuições aos fins-de-semana, épocas festivas e feriados nacionais** (assinalados na Figura 4.22) que já tínhamos identificado anteriormente.

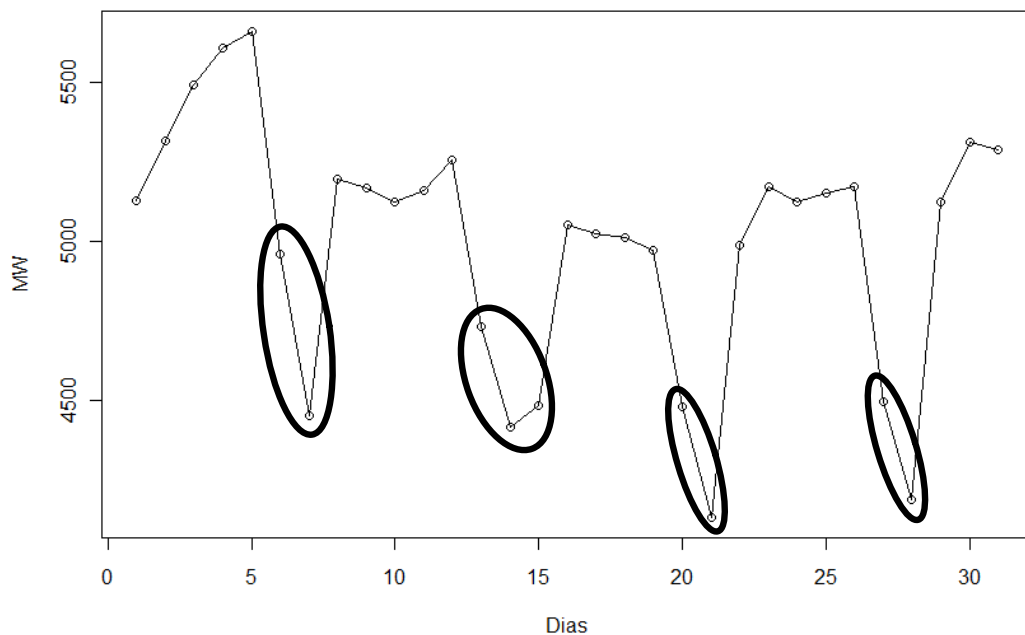


Figura 4.22 – Evolução do Consumo de Energia Eléctrica em Agosto de 2005 (Diária)

Aplicando o método de agrupamento das k-Médias às agregações dos dados originais obtidas a este nível e às respectivas correlações é possível observar a sua distribuição perante os grupos definidos, como podemos verificar na Figura 4.23, exemplificativa do agrupamento em quatro grupos, sendo este o número de grupos óptimo definido a partir do coeficiente de silhueta, no que respeita ao agrupamento a partir das correlações. É de acrescentar que apesar de ser obtido um valor máximo desta medida para o agrupamento em quatro grupos a partir das correlações, em termos médios globais, os resultados são mais favoráveis para o agrupamento a partir das distâncias euclidianas.

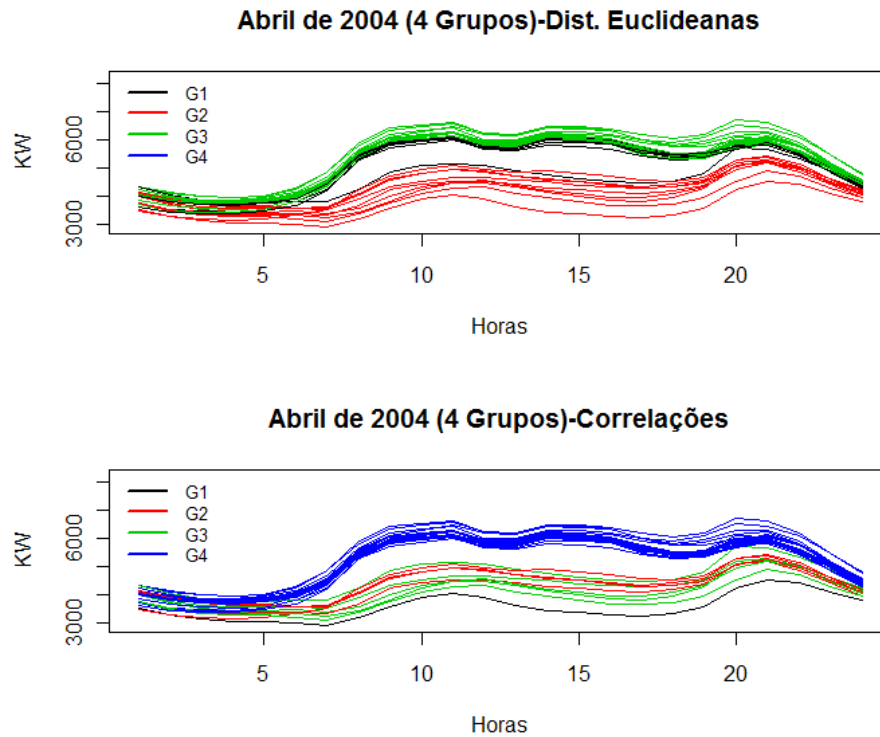


Figura 4.23 – Agrupamento k-Médias para Perfis Diários

O agrupamento em quatro partições para o mês de Abril de 2004 no que toca aos perfis diários é algo confusa, sendo que para a abordagem das distâncias euclidianas existe um grupo predominante de dias com consumos mais baixos de energia eléctrica (G2) e outro de consumos mais elevados (G3), que se confundem, ambos, com dias agrupados noutro grupo com consumos intermédios (G1). No que respeita ao agrupamento a partir das correlações existe um agrupamento distinto de dias com consumos de energia superiores (G4), sendo que os dias com consumos inferiores se confundem entre os restantes três grupos. A caracterização dos perfis encontrados – Tabela 4.3 – pressupõe a existência de grupos com números de elementos e médias distintos. Com erros absolutos aproximados, a abordagem baseada no agrupamento a partir das **correlações** é a que apresenta um **erro inferior**.

Ano 2004		4 Perfis - Dist. Euclidianas				4 Perfis - Correlações			
Perfis Diários	Grupos	G1	G2	G3	G4	G1	G2	G3	G4
Limites de classificação	Mínimo	4576,377	3567,501	5133,735	5728,691	2908,925	3152,025	3053,875	3317,825
	Máximo	5125,906	4517,188	5693,096	6465,375	6988,6	8214,125	7167,975	6952,275
Tamanho Grupos		72	70	161	63	30	129	56	151
Média Grupos		4860,264	4212,54	5401,462	6032,002	4597,897	5530,74	4432,043	5264,049
Erro Relativo por Grupo		-8,28E-11	-7,96E-11	-1,69E-10	3,82E-11	-3,18E-12	-9,82E-11	-1,18E-11	-2,05E-11
Erro Absoluto Total		3,70E-10				1,34E-10			

Tabela 4.3 – Caracterização do Agrupamento em 4 Grupos (Perfis Diários)

De acordo com os agrupamentos obtidos neste nível são definidos os perfis médios diários de consumo de energia eléctrica – Figura 4.24.

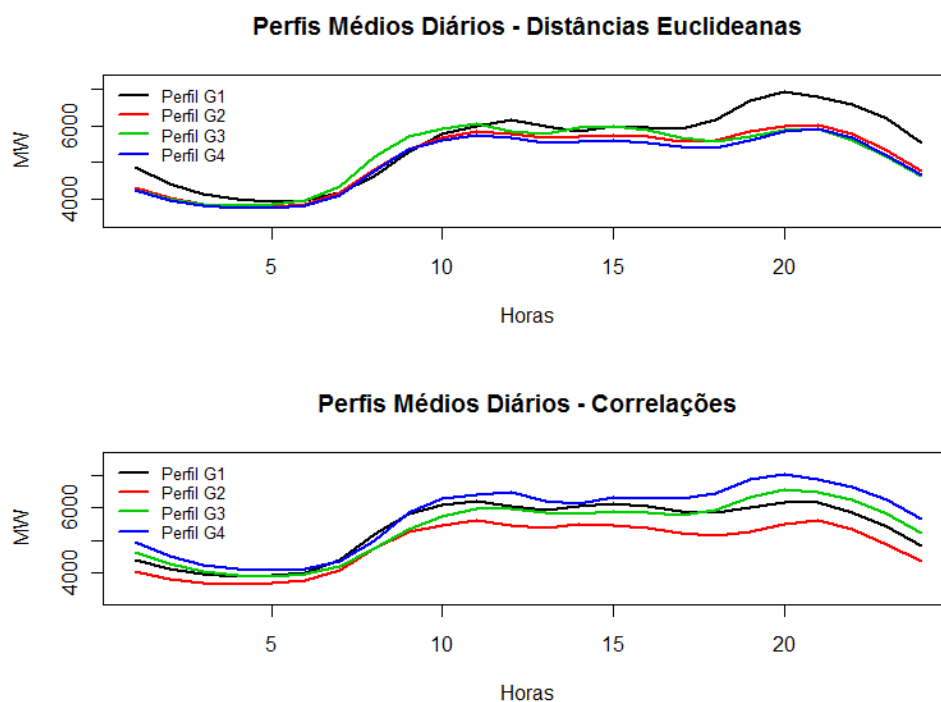


Figura 4.24 – Perfis Médios Diários

Os perfis médios obtidos a partir das distâncias euclidianas apresentam-se bastante sobrepostos segundo uma evolução horária semelhante, destacando-se a partir das 16 horas, aproximadamente, o perfil G1 com um nível de consumo superior. No entanto, a abordagem a partir das correlações, com o mesmo padrão evolutivo, destaca de uma forma mais clara, diferentes níveis de consumo, embora semelhantes.

4.2.4 Perfis Dias Semana (Nível 4)

Num nível superior, em termos de evolução do consumo de energia eléctrica por dia de semana, verifica-se uma semelhança de variação em todos os meses, denotando-se um **nível de consumo superior de segunda a sexta-feira** e uma **diminuição** de consumo durante **sábados e domingos**, que afecta proporcionalmente a distribuição da energia eléctrica. Na Figura 4.25, é apresentada uma representação gráfica do mês de Junho de 2005 que exemplifica as conclusões retiradas.

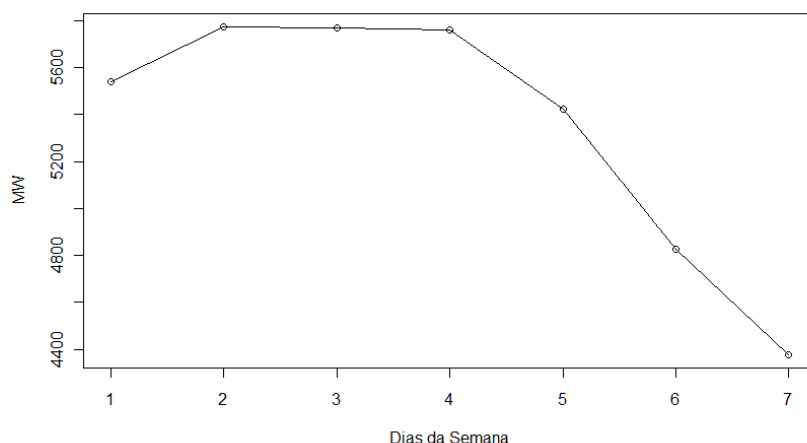


Figura 4.25 – Evolução do Consumo de Energia Eléctrica em Junho de 2005 (Dias da Semana)

Como consequência da aplicação do algoritmo de agrupamento das k-Médias aos resultados obtidos no quarto nível do histograma por níveis e às correlações dos dados, é apresentada uma representação gráfica exemplificativa para três grupos – Figura 4.26 – bem como a caracterização – Tabela 4.4 – para todos os grupos em termos de limites de agrupamento, tamanho dos grupos, média e erro cometidos. O valor máximo do coeficiente de silhueta é atingido para o agrupamento em três grupos a partir das correlações para o ano de 2004. Seguindo a mesma tendência, verifica-se que o valor médio global para esta medida é superior neste ano para a mesma abordagem, revertendo-se a situação para o ano de 2005, onde o coeficiente de silhueta médio global é superior para o agrupamento a partir das distâncias euclidianas.

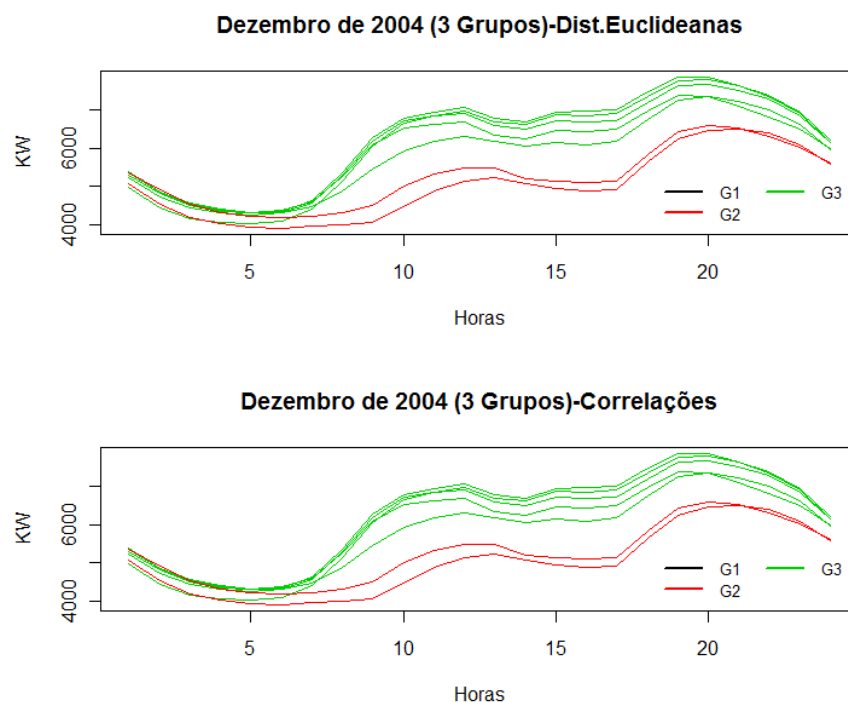


Figura 4.26 – Agrupamento k-Médias para Perfis Dias da Semana

Para o exemplo apresentado não há reparos a fazer comparativamente aos resultados obtidos no que respeita ao agrupamento a partir das duas abordagens definidas, para o mês de Dezembro de 2004, sendo coincidentes. Ainda assim, no que respeita ao número de dias incluídos em cada perfil de consumo obtido, os resultados variam, assim como as médias associadas. Apesar de bastante próximos, o **erro** absoluto total cometido é **superior** para os perfis obtidos a partir das **correlações** dos dados.

Ano 2004		3 Perfis - Dist. Euclideanas			3 Perfis - Correlações		
Perfis Dias da Semana	Grupos	G1	G2	G3	G1	G2	G3
Limites de classificação	Mínimo	3886,723	4834,355	5478,855	3372,5	3132,28	3429,49
	Máximo	4727,477	5459,705	6260,005	6601,288	6582,181	7829,375
Tamanho Grupos		19	37	28	35	17	32
Média Grupos		4322,004	5180,665	5748,495	5230,316	4476,498	5505,565
Erro Relativo por Grupo		-8,19E-12	-3,37E-11	3,64E-12	3,18E-11	-9,09E-13	-1,68E-11
Erro Absoluto Total		4,55E-11			4,96E-11		

Tabela 4.4 – Caracterização do Agrupamento em 3 Grupos (Perfis Dias da Semana)

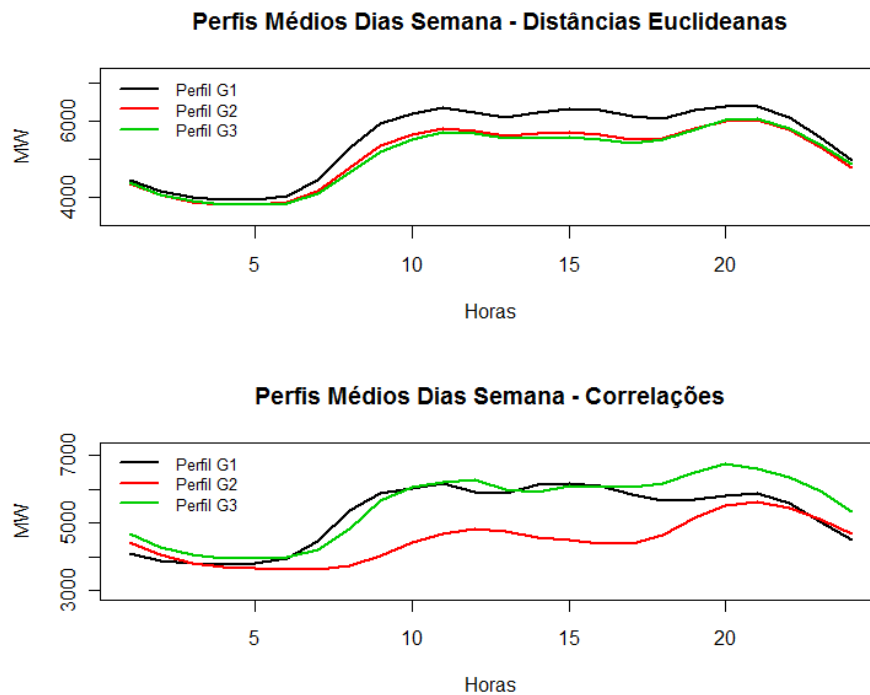


Figura 4.27 – Perfis Médios Dias da Semana

A abordagem a partir das distâncias euclidianas em termos de perfis médios de dias da semana apresenta uma evolução bastante semelhante, nomeadamente no que respeita aos perfis G2 e G3 que se sobrepõem, destacando-se apenas o perfil G1 com um nível de consumo superior a partir das 5 horas. Os perfis médios de dias da semana encontrados a partir das correlações demonstram a existência do perfil G2 com níveis de consumo bastante inferiores entre as 4 e as 20 horas. Os perfis G1 e G3 com um comportamento mais semelhante, apesar de que com oscilações, destacam-se a aproximadamente a partir das 16 horas, com um consumo superior para o perfil G3.

4.2.5 Perfis Semana e Fim-de-Semana (Nível 5)

O comportamento verificado para este nível de agregação vem comprovar a existência de uma **diminuição** significativa do consumo de energia eléctrica durante os **fins de semana**, não havendo nenhum mês em que o consumo semanal seja inferior ao do fim-de-semana. É de realçar que a distribuição é semelhante ao longo de ambos os anos, apesar do nível de energia utilizada ser superior para o ano de 2005. Em termos mensais, repercute-se o que foi concluído anteriormente, verificando-se um maior consumo nos meses mais frios e um menor consumo nos mais amenos.

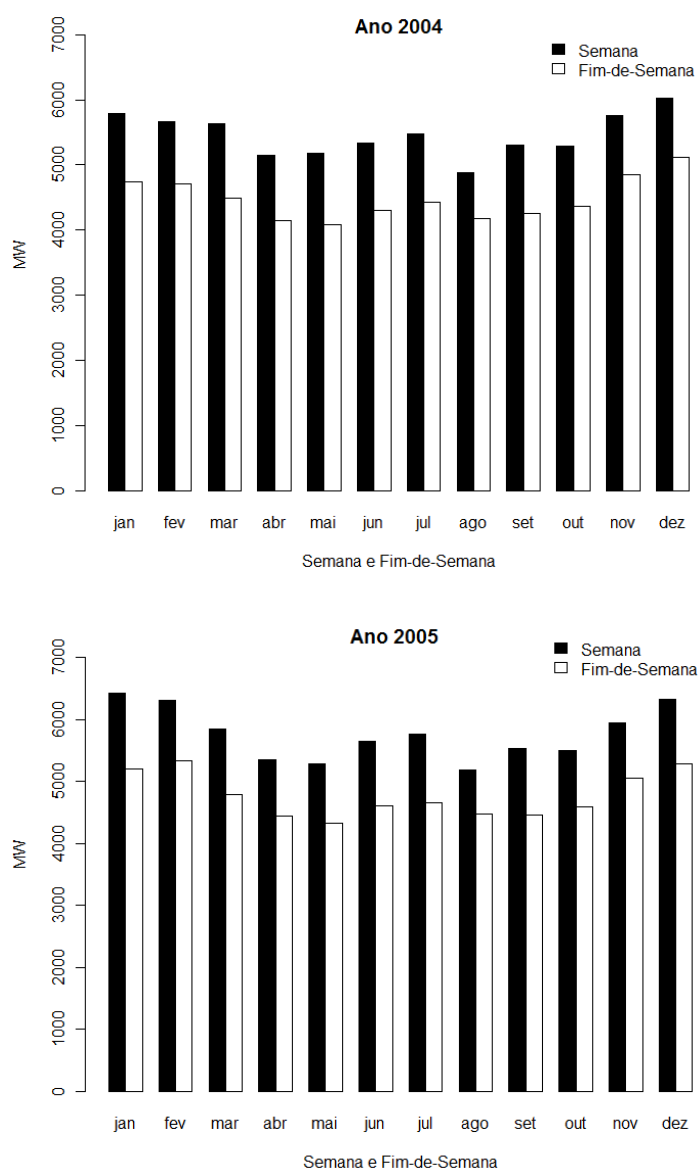


Figura 4.28 – Evolução do Consumo de Energia Eléctrica em 2004 e 2005 (Semana e Fim-de-Semana)

Com o agrupamento efectuado aos resultados ao nível de semanas e fins-de-semana foram alcançadas características que figuram na Tabela 4.5 e é apresentada uma representação gráfica exemplificativa para três grupos – Figura 4.29. Três foi o número óptimo de grupos indicado pela medida de silhueta para o agrupamento a partir das correlações, apesar dos resultados médios globais serem mais favoráveis para as distâncias euclidianas.

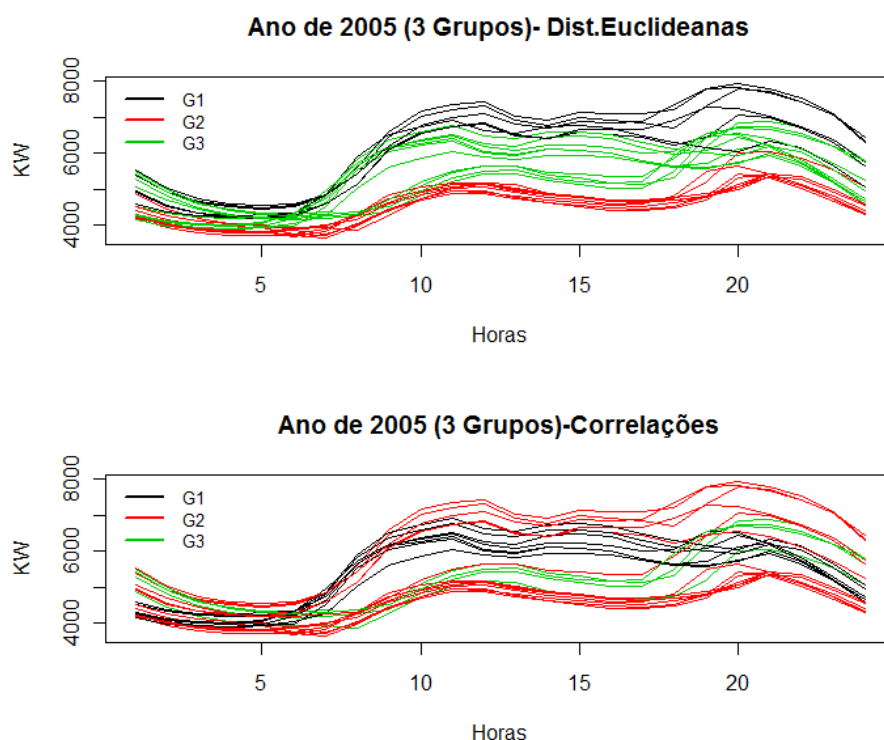


Figura 4.29 – Agrupamento k-Médias para Perfis Semana e Fim-de-Semana

Da análise ao agrupamento obtida para as duas abordagens desenvolvidas durante o ano de 2005 verifica-se uma divisão visualmente mais clara no que respeita aos resultados de agrupamento a partir das distâncias euclidianas, onde são observados três grupos com consumos inferiores (G2), intermédios (G3) e superiores (G1). Relativamente ao agrupamento a partir das correlações, o Grupo 2 possui perfis com médias de consumo baixas e elevadas, havendo intermedicamente perfis agrupados nos Grupos 1 e 3. Com características diferentes para ambas as abordagens, o agrupamento que reproduz um **menor erro** absoluto total é a obtida com base nas **distâncias euclidianas**.

Ano 2005		3 Perfis - Dist. Euclidianas			3 Perfis - Correlações		
Perfis Semana e Fim-de-Semana	Grupos	G1	G2	G3	G1	G2	G3
Limites de classificação	Mínimo	5770,911	4332,006	5047,366	3861,42	3640,576	3871,756
	Máximo	6431,393	4789,096	5652,444	6908,667	7939,934	6882,837
Tamanho Grupos		6	8	10	7	13	4
Média Grupos		6105,72	4545,632	5338,817	5468,74	5225,439	5164,042
Erro Relativo por Grupo		-3,64E-12	-9,09E-13	0	3,18E-12	4,82E-11	-2,73E-12
Erro Absoluto Total		4,55E-12			5,41E-11		

Tabela 4.5 – Caracterização do Agrupamento em 3 Grupos (Perfis Semana e Fim-de-Semana)

Como resultado dos perfis obtidos a partir dos agrupamentos são apresentados os Perfis Médios de Semana e Fim-de-Semana – Figura 4.30.

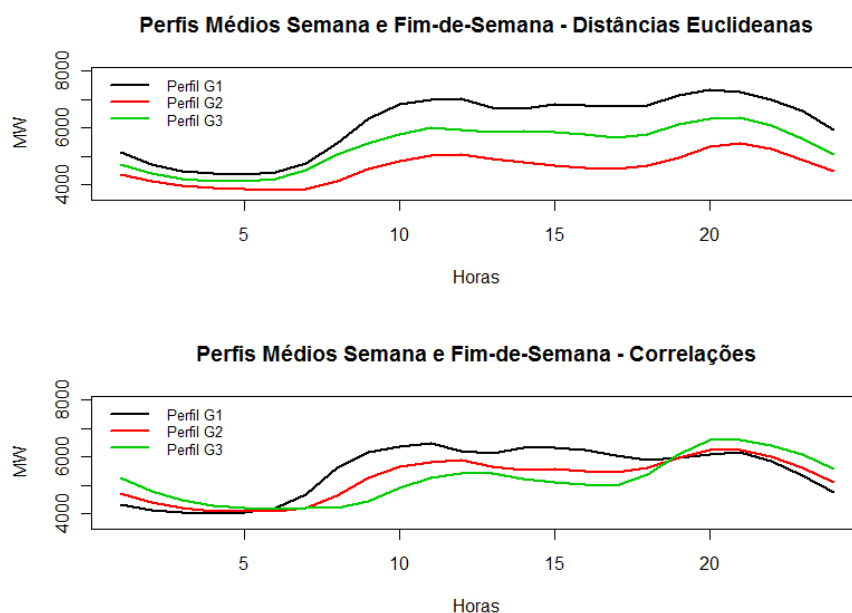


Figura 4.30 – Perfis Médios Semana e Fim-de-Semana

Os perfis médios de semana e fim-de-semana, de acordo com as distâncias euclidianas, demonstram uma compartimentação bastante clara em termos de níveis de consumo inferior (G2), intermédio (G3) e superior (G1), mas com um comportamento bastante semelhante. Por outro lado, os perfis médios encontrados a partir das correlações cruzam-se aproximadamente nas 5 e 18 horas, momento entre os quais o perfil G3 passa a ter um nível de consumo inferior, o perfil G1 passa para um nível superior de consumo, mantendo-se o perfil G2 com um consumo intermédio, aproximando-se todos às 12 horas.

4.2.6 Perfis Semanais (Nível 6)

A agregação semanal sugere que os **meses** do ano **iniciais e finais** são caracterizados por um **maior nível de consumo de energia**, ao passo que, nos **meses intermédios**, os **níveis de consumo** são **mais baixos**, existindo, no entanto, quebras identificáveis para as semanas de cada mês.

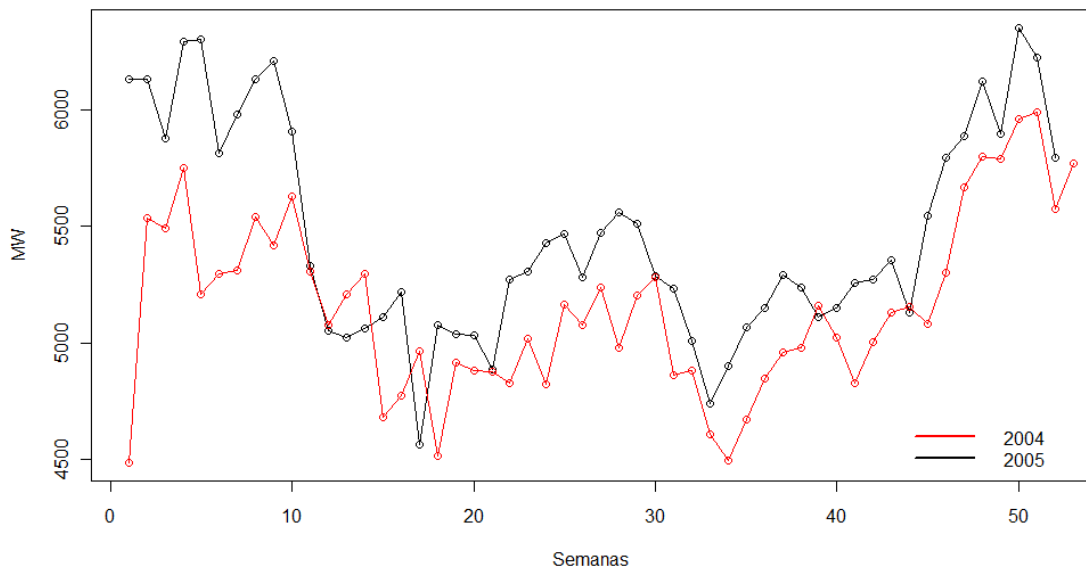


Figura 4.31 – Evolução do Consumo de Energia Eléctrica - 2004 e 2005 (Semanal)

Para a caracterização dos perfis semanais é apresentada uma representação gráfica exemplificativa do agrupamento em dois conjuntos – Figura 4.32 – bem como os limites de agrupamento, número de elementos que os constituem, média e erros associados - Tabela 4.6. O número óptimo de agrupamentos a analisar foi definido pelo máximo do coeficiente de silhueta obtido para o agrupamento a partir das distâncias euclidianas. A mesma tendência se verifica em termos médios globais.

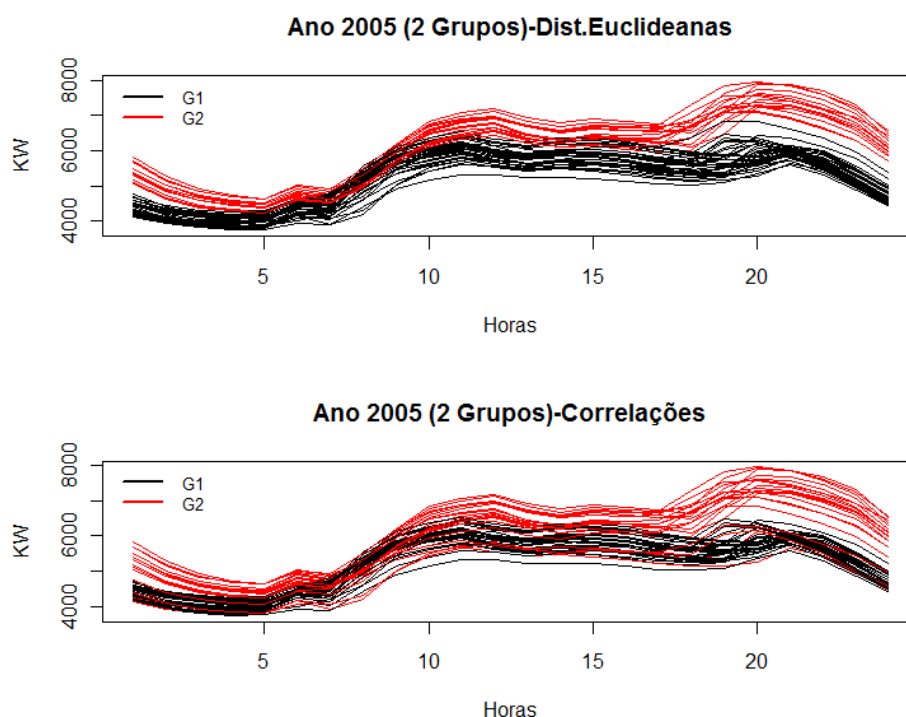


Figura 4.32 – Agrupamento k-Médias para Perfis Semanais

Da análise do agrupamento para ambas as representações em estudo é possível denotar algumas diferenças nos agrupamentos, mantendo-se, no entanto, a prevalência de um grupo de perfis com médias de consumo inferiores (G1) e outro com médias superiores (G2). O agrupamento a partir das **distâncias euclidianas** é a que proporciona um **menor erro** absoluto.

Ano 2005		2 Perfis - Dist. Euclideanas		2 Perfis - Correlações	
Perfis Semanais	Grupos	G1	G2	G1	G2
Limites de classificação	Mínimo	4560,25	5797,192	4560,25	5030,534
	Máximo	5560,613	6356,381	5910,833	6356,381
Tamanho Grupos		35	15	30	20
Média Grupos		5183,25	6063,561	5194,096	5839,733
Erro Relativo por Grupo		-3,64E-12	8,19E-12	-1,50E-11	4,27E-11
Erro Absoluto Total		1,18E-11		5,78E-11	

Tabela 4.6 – Caracterização do Agrupamento em 2 Grupos (Perfis Semanais)

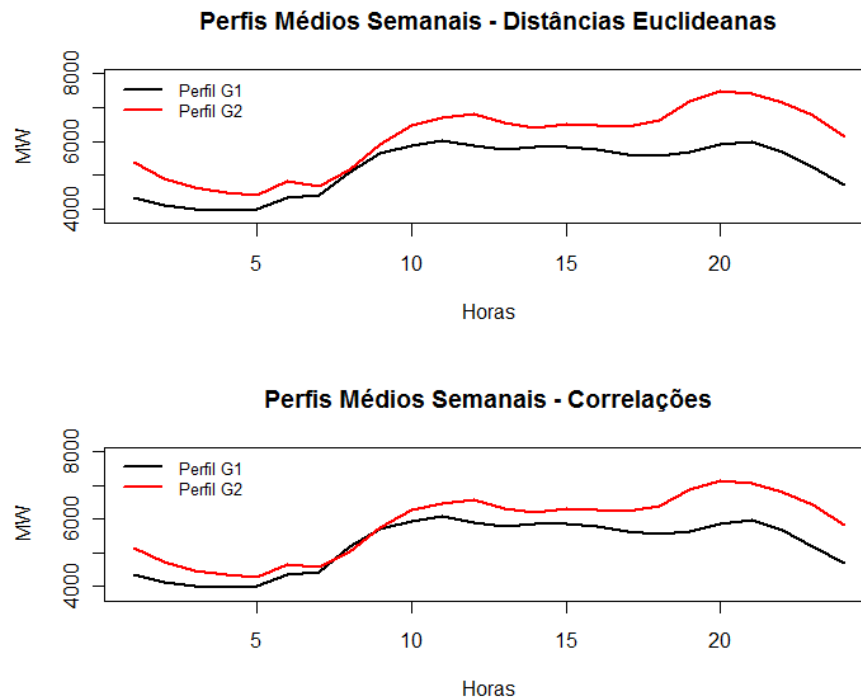


Figura 4.33 – Perfis Médios Semanais

Os perfis médios encontrados a nível semanal são bastante semelhantes para ambas as abordagens, aproximando-se bastante às 7 horas, passando mesmo o perfil G2 abaixo do perfil G1.

4.2.7 Perfis Mensais (Nível 7)

Com a agregação dos valores por meses do ano, torna-se viável a comparação em termos de níveis mensais médios de transporte de energia – Figura 4.34.

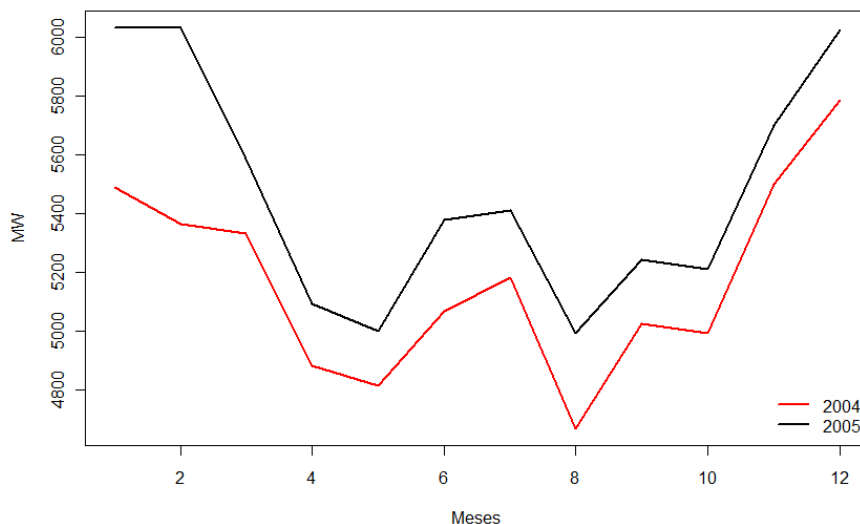


Figura 4.34 – Evolução do Consumo de Energia Eléctrica - 2004 e 2005 (Mesal)

Para o ano de 2004, assiste-se a uma **descida gradual** do consumo de energia eléctrica de aproximadamente 5491 MW em **Janeiro** para cerca de 4817 MW em **Maio**, havendo uma **pequena subida** nos meses de **Junho e Julho** que culmina numa **descida** para o mês de **Agosto**, sendo atingido um **mínimo** de 4668,628 MW. No mês seguinte (**Setembro**) o consumo de energia **volta a aumentar**, verificando-se uma **última descida** em **Outubro**, seguida de dois **aumentos substanciais** para os meses de **Novembro e Dezembro**, onde se atingem **máximos** acima dos 5500 MW e 5700 MW, respectivamente. Por sua vez, o ano de 2005, começa com **valores idênticos** para os primeiros dois meses (**Janeiro e Fevereiro**), fixando-se à volta dos 6000 MW (valores superiores ao ano anterior). Os meses seguintes (de **Março a Dezembro**) seguem a mesma tendência do ano anterior, com uma **descida** do consumo de electricidade entre **Março e Maio**, **subidas** nos meses de **Junho, Julho e Setembro**, novamente **descidas** para os meses de **Agosto e Outubro**, terminando o ano com a **subida** em **Novembro e Dezembro** para um valor semelhante ao inicial. Neste sentido, pode concluir-se que em épocas climatericamente mais frias o consumo de energia tende a aumentar, ocorrendo o contrário em épocas mais amenas.

Numa tentativa de caracterização de perfis mensais através da utilização de métodos de agrupamento, é apresentada uma representação gráfica exemplificativa para dois grupos – Figura 4.35 – bem como elementos que caracterizam os agrupamentos – Tabela 4.7. Em termos médios globais, de acordo com a medida de silhueta, o agrupamento através das distâncias euclidianas é o que produz resultados mais favoráveis, apesar do valor óptimo de grupos a analisar ser atingido para o agrupamento a partir das correlações.

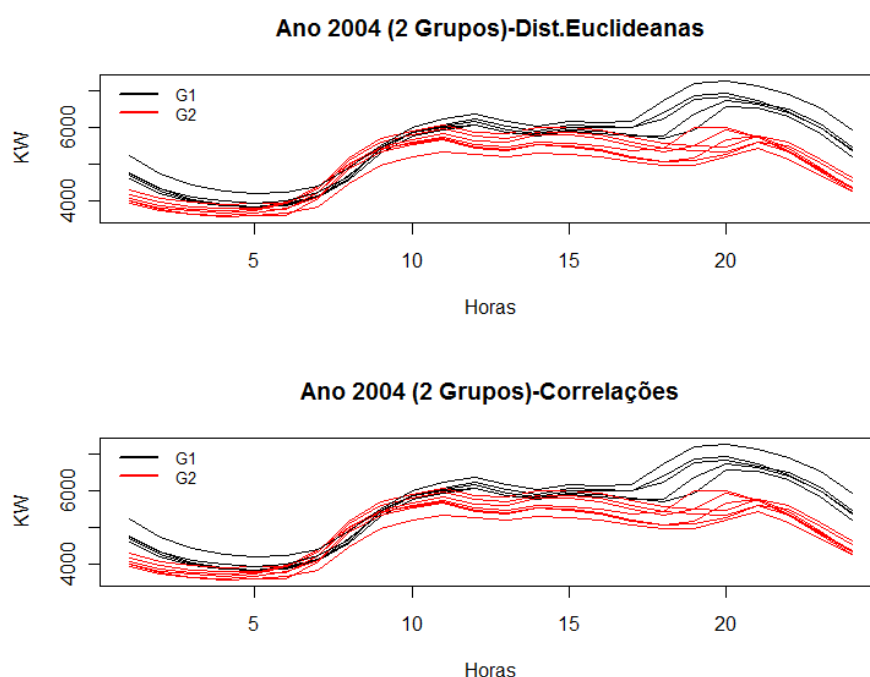


Figura 4.35 – Agrupamento k-Médias para Perfis Mensais

Em termos de perfis mensais não são verificadas diferenças nos agrupamentos obtidos a partir das diferentes abordagens.

Perfis Mensais	Grupos	G1	G2	G1	G2
Limites de classificação	Mínimo	5333,966	4668,628	3853,72	3851,977
	Máximo	5785,558	5184,348	7263,485	6060,444
Tamanho Grupos		5	7	5	7
Média Grupos		5495,557	4948,694	5495,557	4948,694
Erro Relativo por Grupo		0	-9,09E-13	8,41E-11	-1,27E-11
Erro Absoluto Total		9,09E-13		9,69E-11	

Tabela 4.7 – Caracterização do Agrupamento em 2 Grupos (Perfis Mensais)

Com grupos com o mesmo número de elementos e médias iguais para ambas as abordagens, o **erro** absoluto total acaba por ser **inferior** para o agrupamento a partir das **distâncias euclidianas**.

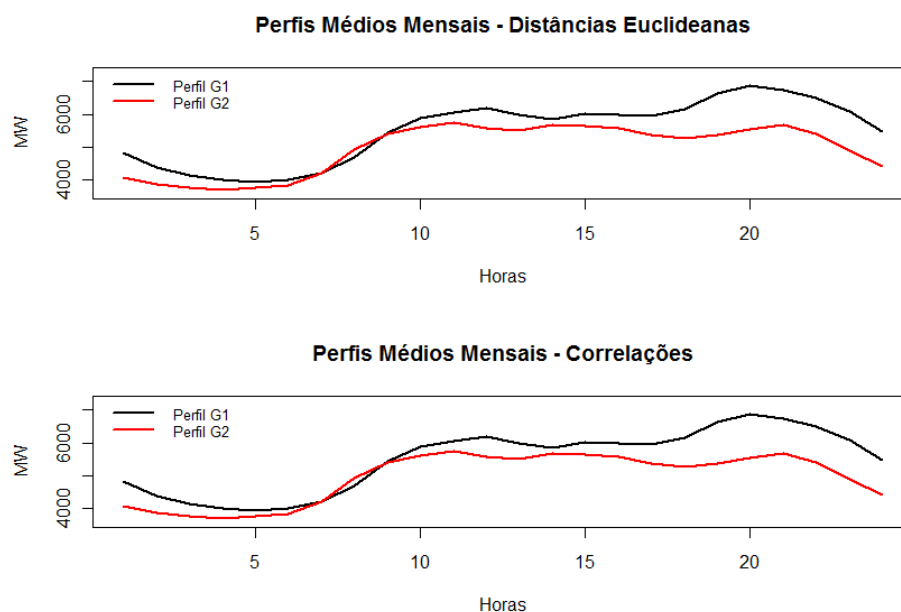


Figura 4.36 – Perfis Médios Mensais

Ao nível mensal, os perfis médios encontrados são semelhantes, não sendo detectadas, visualmente, divergências.

4.2.8 PERFIS TRIMESTRAIS (NÍVEL 8)

A agregação em trimestres não sugere uma grande diferenciação no consumo de energia eléctrica em ambos os anos. O padrão de evolução para esta granularidade sugere uma **diminuição** do consumo de energia eléctrica do **1º para o 2º trimestre**, havendo uma **subida** entre os dois trimestres subsequentes (**3º e 4º trimestres**), como seria de esperar, após a evolução em termos mensal analisada anteriormente.

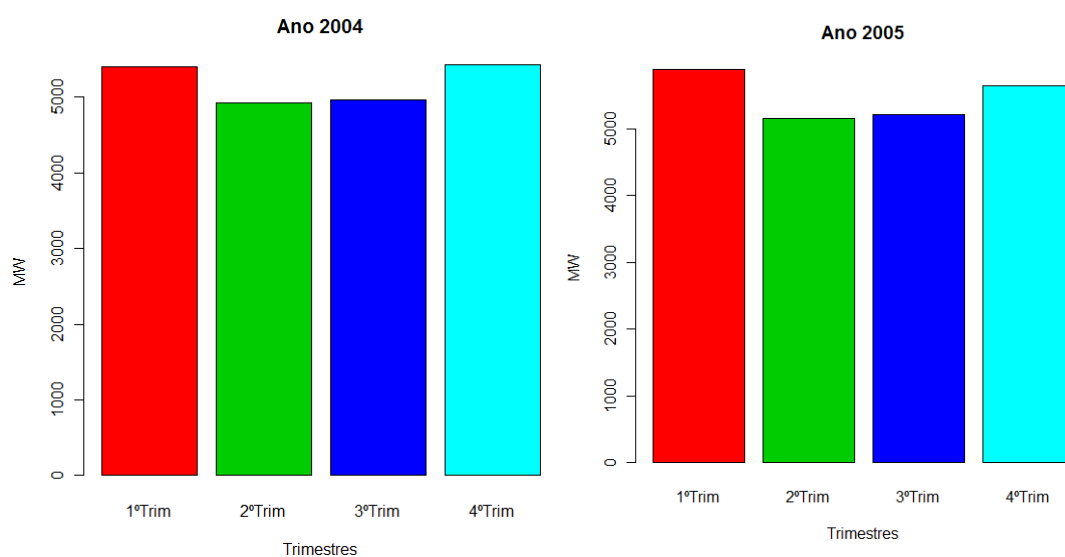


Figura 4.37 – Evolução do Consumo de Energia Eléctrica – 2004 e 2005 (Trimestral)

Como resultado do agrupamento aplicado aos dados agregados no nível trimestral é apresentada uma representação gráfica para dois grupos – Figura 4.38 – sendo este o valor óptimo obtido para os agrupamentos abordados. Em termos médios globais, o coeficiente de silhueta obtém melhores resultados para o ano de 2004 no que respeita às distâncias euclidianas, verificando-se o contrário para o ano de 2005.

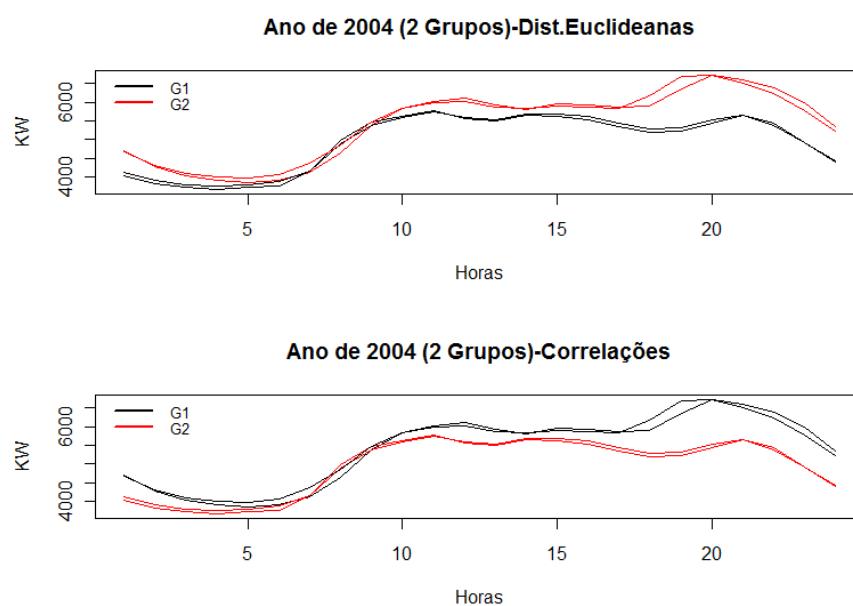


Figura 4.38 – Agrupamento k-Médias para Perfis Trimestrais

Analisando a Figura 4.38 constata-se que são obtidas os mesmos agrupamentos no que respeita aos perfis trimestrais, ainda que os grupos possuam denominações diferentes.

Ano 2004		2 Perfis - Dist. Euclídeanas		2 Perfis - Correlações	
Perfis Trimestrais	Grupos	G1	G2	G1	G2
Limites de classificação	Mínimo	4922,563	5396,664	3857,729	3672,398
	Máximo	4960,102	5426,886	6718,154	5779,292
Tamanho Grupos		2	2	2	2
Média Grupos		4941,332	5411,775	5411,775	4941,332
Erro Relativo por Grupo		0	-9,09E-13	1,09E-11	9,09E-13
Erro Absoluto Total		9,09E-13		1,18E-11	

Tabela 4.8 – Caracterização do Agrupamento em 2 Grupos (Perfis Trimestrais)

Com o mesmo número de elementos que confluem na mesma média, são obtidos diferentes erros absolutos totais, destacando-se os agrupamentos através das **distâncias euclídeanas** com um **erro inferior**.

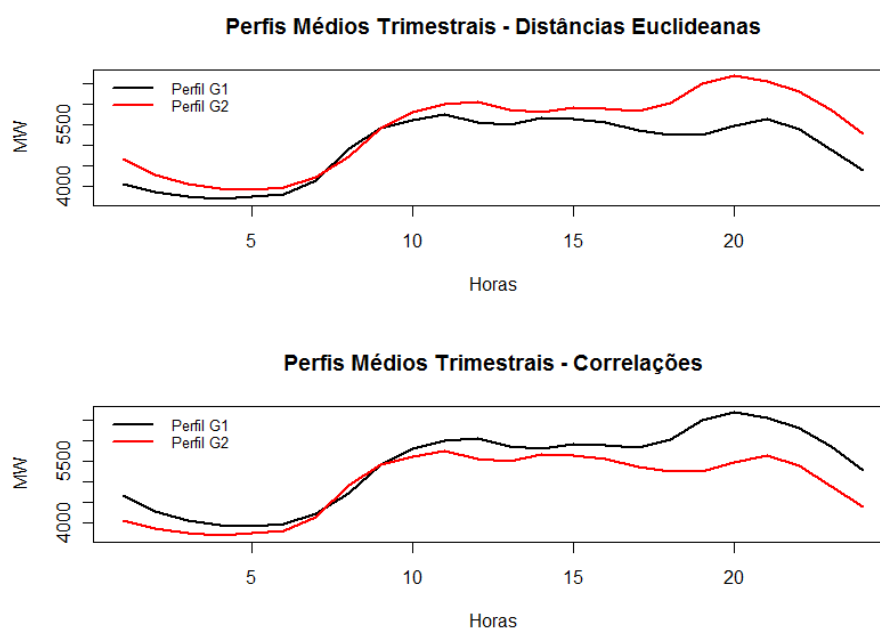


Figura 4.39 – Perfis Médios Trimestrais

Como seria de esperar, os perfis médios trimestrais não aparentam diferenças, apesar dos perfis terem designações diferentes para ambas as abordagens.

4.2.9 Perfis Semestrais (Nível 9)

Para a agregação das observações por semestre verifica-se, como era expectável, uma média de consumo de energia eléctrica nos dois **semestres** do ano de **2005 superior** à de **2004**, apesar de se verificar um pequeno **aumento do 1º para o 2º semestre** no primeiro ano em análise, ao contrário do ano posterior, em que existe uma **diminuição** de valor **entre semestres**.

```
> sem04          > sem05
      1ºSem      2ºSem      1ºSem      2ºSem
Value 5159.613 5193.494 Value 5521.133 5431.95
```

Figura 4.40 – Resultados de Agregação do Nível 9 (2004 e 2005)

Dado o reduzido número de observações na agregação para o nível 9 (duas para cada ano) não é possível aplicar o método de agrupamento aos dados de forma a agrupá-los separadamente para cada ano.

Até este ponto, o estudo permitiu analisar a evolução dos dados agrupados e comparar os perfis encontrados consoante a granularidade temporal e a abordagem de agrupamento pretendidas. De seguida é proposto e avaliado um refinamento desses perfis.

4.3 REFINAR PERFIS

Numa tentativa de encontrar variantes para os perfis encontrados, cada um deles foi desdobrado em outros dois, compostos pela média dos traçados acima e abaixo da média. Os erros cometidos por escolher estes perfis em detrimento dos valores reais são apresentados nas tabelas que se seguem.

		Erro Perfis Horários		Erro Absoluto	Erro Absoluto Total Após Desdobramento	Erro Absoluto Total Antes Desdobramento
		Acima	Abaixo			
Distâncias Euclidianas	1º Perfil	-1,23E-10	2,57E-10	3,81E-10	5,78E-10	1,04E-09
	2º Perfil	-7,28E-12	-1,86E-11	2,59E-11		
	3º Perfil	2,86E-11	4,00E-11	6,87E-11		
	4º Perfil	6,37E-12	9,69E-11	1,03E-10		
Correlações	1º Perfil	2,09E-10	-2,04E-10	4,13E-10	9,17E-10	6,44E-10
	2º Perfil	-2,36E-11	-6,59E-11	8,96E-11		
	3º Perfil	-1,47E-10	-5,55E-11	2,03E-10		
	4º Perfil	-1,18E-11	2,00E-10	2,12E-10		

Tabela 4.9 – Erros de Refinamento de Perfis Horários

		Erro Perfis 6 Horas		Erro Absoluto	Erro Absoluto Total Após Desdobramento	Erro Absoluto Total Antes Desdobramento
		Acima	Abaixo			
Distâncias Euclidianas	1º Perfil	3,64E-12	2,73E-12	6,37E-12	1,01E-10	4,35E-10
	2º Perfil	-2,14E-11	3,46E-11	5,59E-11		
	3º Perfil	-9,09E-12	5,46E-12	1,46E-11		
	4º Perfil	-2,00E-11	4,55E-12	2,46E-11		
Correlações	1º Perfil	-6,00E-11	4,55E-13	6,05E-11	1,71E-10	3,14E-11
	2º Perfil	-3,64E-12	-5,91E-12	9,55E-12		
	3º Perfil	1,91E-11	-1,32E-11	3,23E-11		
	4º Perfil	-4,46E-11	-2,46E-11	6,91E-11		

Tabela 4.10 – Erros de Refinamento de Perfis 6 Horas

		Erro Perfis Diários		Erro Absoluto	Erro Absoluto Total Após Desdobramento	Erro Absoluto Total Antes Desdobramento
		Acima	Abaixo			
Distâncias Euclidianas	1º Perfil	9,09E-12	-8,19E-12	1,73E-11	6,59E-11	3,70E-10
	2º Perfil	0,00E+00	-5,91E-12	5,91E-12		
	3º Perfil	-9,09E-12	1,09E-11	2,00E-11		
	4º Perfil	-8,19E-12	-1,46E-11	2,27E-11		
Correlações	1º Perfil	-2,73E-12	3,64E-12	6,37E-12	3,49E-11	1,34E-10
	2º Perfil	-9,09E-13	-9,91E-12	1,08E-11		
	3º Perfil	-7,28E-12	4,55E-13	7,73E-12		
	4º Perfil	-8,19E-12	-1,82E-12	1,00E-11		

Tabela 4.11 – Erros de Refinamento de Perfis Diários

		Erro Perfis Dias da Semana		Erro Absoluto	Erro Absoluto Total Após Desdobramento	Erro Absoluto Total Antes Desdobramento
		Acima	Abaixo			
Distâncias Euclidianas	1º Perfil	2,73E-12	3,64E-12	6,37E-12	1,91E-11	4,55E-11
	2º Perfil	-6,37E-12	-9,09E-13	7,28E-12		
	3º Perfil	2,73E-12	-2,73E-12	5,46E-12		
Correlações	1º Perfil	-2,00E-11	-4,55E-13	2,05E-11	1,08E-10	4,96E-11
	2º Perfil	6,37E-12	-1,68E-11	2,32E-11		
	3º Perfil	-3,09E-11	3,32E-11	6,41E-11		

Tabela 4.12 – Erros de Refinamento de Perfis Dias da Semana

		Erro Perfis Semana e Fim-de-Semana		Erro Absoluto	Erro Absoluto Total Após Desdobramento	Erro Absoluto Total Antes Desdobramento
		Acima	Abaixo			
Distâncias Euclidianas	1º Perfil	9,09E-13	0	9,09E-13	1,82E-12	4,55E-12
	2º Perfil	9,09E-13	0	9,09E-13		
	3º Perfil	0	0	0		
Correlações	1º Perfil	2,73E-12	9,09E-13	3,64E-12	2,00E-11	5,41E-11
	2º Perfil	-9,09E-13	-1,09E-11	1,18E-11		
	3º Perfil	9,09E-13	-3,64E-12	4,55E-12		

Tabela 4.13 – Erros de Refinamento de Perfis Semana e Fim-de-Semana

		Erro Perfis Semanais		Erro Absoluto	Erro Absoluto Total Após Desdobramento	Erro Absoluto Total Antes Desdobramento
		Acima	Abaixo			
Distâncias Euclidianas	1º Perfil	-9,09E-13	6,37E-12	7,28E-12	8,19E-12	1,18E-11
	2º Perfil	0	9,09E-13	9,09E-13		
Correlações	1º Perfil	4,18E-11	8,64E-12	5,05E-11	6,59E-11	5,78E-11
	2º Perfil	-1,00E-11	5,46E-12	1,55E-11		

Tabela 4.14 – Erros de Refinamento de Perfis Semanais

		Erro Perfis Mensais		Erro Absoluto	Erro Absoluto Total Após Desdobramento	Erro Absoluto Total Antes Desdobramento
		Acima	Abaixo			
Distâncias Euclidianas	1º Perfil	-9,09E-13	-9,09E-13	1,82E-12	3,64E-12	9,09E-13
	2º Perfil	-1,82E-12	0	1,82E-12		
Correlações	1º Perfil	4,55E-12	7,28E-12	1,18E-11	2,36E-11	9,69E-11
	2º Perfil	1,00E-11	-1,82E-12	1,18E-11		

Tabela 4.15 – Erros de Refinamento de Perfis Mensais

		Erro Perfis Trimestrais		Erro Absoluto	Erro Absoluto Total Após desdobramento	Erro Absoluto Total Antes Desdobramento
		Acima	Abaixo			
Distâncias Euclidianas	1º Perfil	0	0	0	0	9,09E-13
	2º Perfil	0	0	0		
Correlações	1º Perfil	-9,09E-13	-1,82E-12	2,73E-12	4,09E-12	1,18E-11
	2º Perfil	-9,09E-13	4,55E-13	1,36E-12		

Tabela 4.16 – Erros de Refinamento de Perfis Trimestrais

Analisando o **erro** absoluto total **após o desdobramento** e comparando-o com o erro total antes do desdobramento verifica-se maioritariamente que, o primeiro é **inferior** para o agrupamento a partir das **distâncias euclidianas**, sendo superior para o agrupamento através das correlações. Isto verifica-se para os **perfis horários, 6 horas, dias da semana e semanais**. Contrariamente, verifica-se para os **perfis mensais** que o erro absoluto após desdobramento é **superior** ao erro antes do desdobramento para o agrupamento a partir das **distâncias euclidianas** e inferior para o agrupamento a partir das correlações. Por outro lado, para os **perfis diários, semana e fim-de-semana e trimestrais** o erro absoluto total após desdobramento é **inferior** para as **duas abordagens** em estudo.

Fazendo uma análise comparativa entre os erros dos perfis – Tabela 4.17 – podemos concluir que os perfis horários relativamente aos restantes são os que expressam um erro superior tanto para a abordagem de agrupamento a partir das distâncias euclidianas como a partir das correlações. Com um erro algo superior encontram-se os perfis de 6 horas apenas para as distâncias euclidianas e os perfis diários para ambas. Os restantes erros cometidos pela escolha dos perfis em vez dos valores reais demonstram-se inferiores a todos os outros.

	Distâncias Euclidianas	Correlações
Perfis Horários	1,04E-09	6,44E-10
Perfis 6 Horas	4,35E-10	3,14E-11
Perfis Diários	3,70E-10	1,34E-10
Perfis Dias da Semana	4,55E-11	4,96E-11
Perfis Semana e Fim-de-Semana	4,55E-12	5,41E-11
Perfis Semanais	1,18E-11	5,78E-11
Perfis Mensais	9,09E-13	9,69E-11
Perfis Trimestrais	9,09E-13	1,18E-11

Tabela 4.17 – Erros dos Perfis (Distâncias Euclidianas e Correlações)

A título exemplificativo são apresentados os perfis resultantes do desdobramento acima e abaixo da média para os perfis diários do ano de 2004 de acordo com o agrupamento a partir das distâncias euclidianas – Figura 4.41 – e das correlações – Figura 4.42.

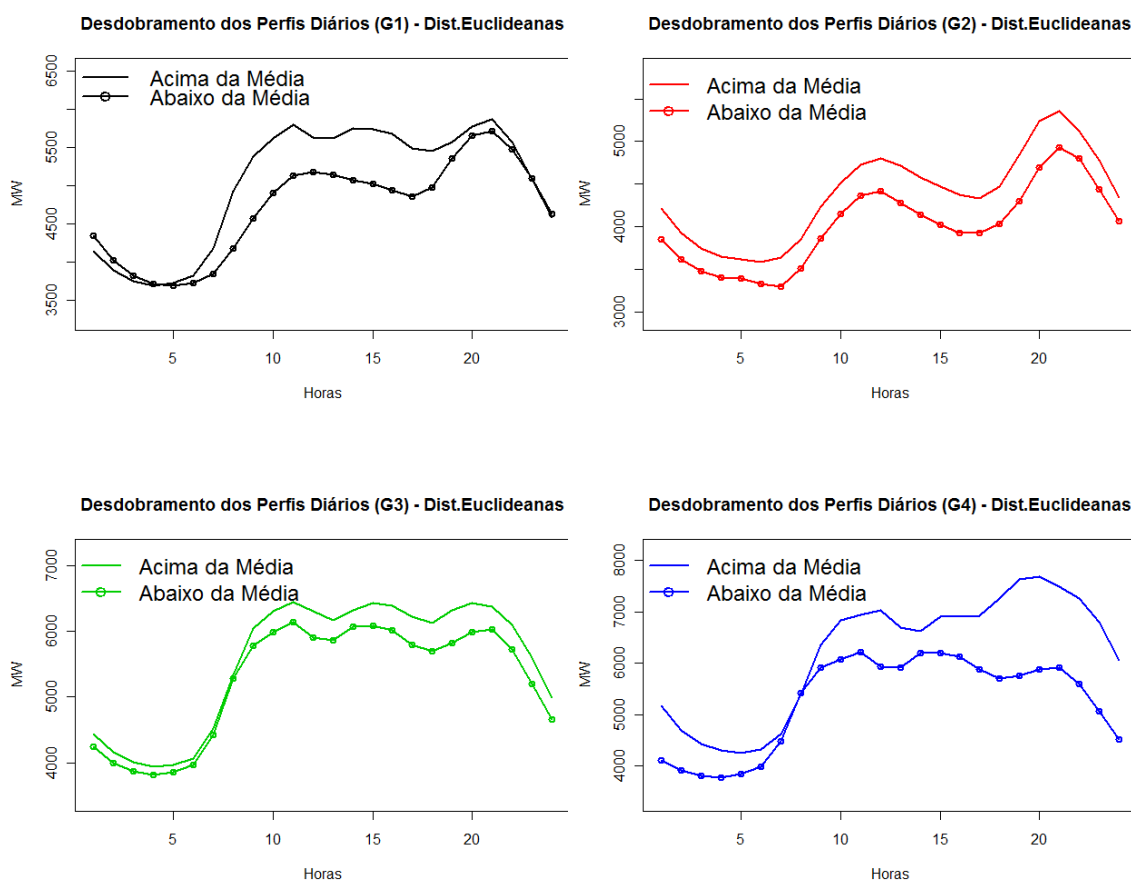


Figura 4.41 – Exemplo do Desdobramento dos Perfis Diários (Distâncias Euclidenas)

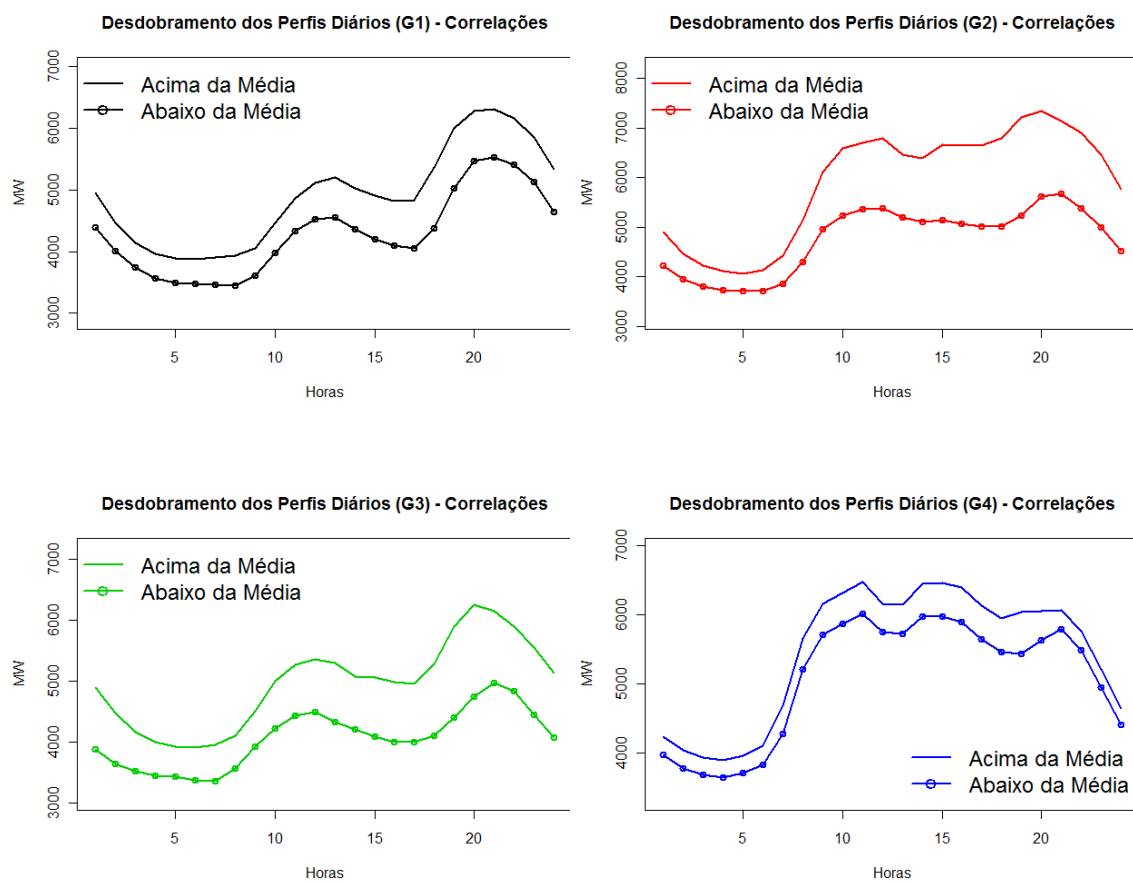


Figura 4.42 – Exemplo do Desdobramento dos Perfis Diários (Correlações)

Após a análise comparativa entre os perfis obtidos antes e após desdobramento acima e abaixo da média torna-se relevante concentrarmo-nos na distribuição dos dias da semana pelos perfis/agrupamentos criados.

4.4 DISTRIBUIÇÃO DE DIAS DA SEMANA

Tirando partido do agrupamento dos dados originais num nível de granularidade diária é sugerida a análise de distribuição dos dias da semana⁵ entre dois e sete agrupamentos (o número de dias da semana existentes).

4.4.1 2 GRUPOS

A agregação dos perfis diários do ano de 2004 em dois conjuntos, segundo a abordagem das distâncias euclidianas, permite-nos concluir que o Grupo 1, com 249 observações e o nível médio de consumo de energia superior – 5527 MW – é constituído maioritariamente por dias de semana, havendo ainda alguns sábados e domingos representados neste conjunto. Em termos mensais, entre Março e Outubro existem apenas dias de semana, apesar de com menor ocorrência para o mês de Agosto. Com dias de semana e sábados, encontram-se os meses de Janeiro, Fevereiro e Novembro, sendo o mês de Dezembro o que apresenta ocorrências para todos os dias (Apêndice C).

Distâncias Euclidianas				
	G1	% G1	G2	% G2
Seg	45	18%	7	6%
Ter	47	19%	5	4%
Qua	49	20%	3	3%
Qui	48	19%	5	4%
Sex	47	19%	6	5%
Sab	10	4%	42	36%
Dom	3	1%	49	42%
Total	249	100%	117	100%

Tabela 4.18 – Distribuição Diária por Grupo em 2004 (2 Grupos) – Distâncias Euclidianas

⁵ Dias da semana é uma expressão utilizada para todos os dias de Segunda a Domingo. Por outro lado, dias de semana é uma denominação para os dias de Segunda a Sexta-Feira. Não é utilizada a expressão dias úteis porque os dias de semana agrupados podem englobar feriados.

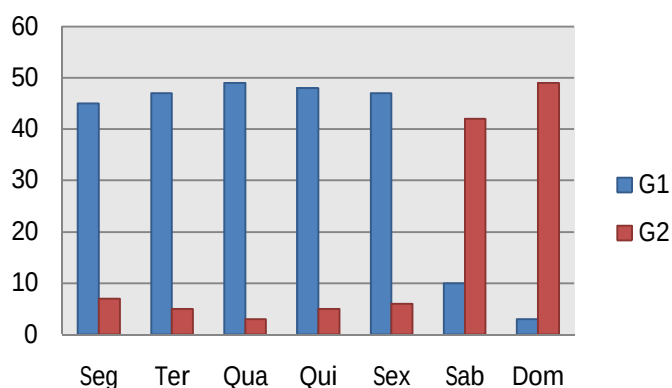


Figura 4.43 – Distribuição Dias da Semana (2 Grupos) – Distâncias Euclidianas (2004)

O segundo agrupamento com um total de 117 observações e o nível médio de consumo inferior – 4429 MW – é caracterizado essencialmente por dias de fim-de-semana e também grande parte dos feriados (excepto os dos dias 1 e 8 de Dezembro), sendo uma parte minoritária representada por dias de semana. Desta forma, todos os meses incluem na sua maioria fins-de-semana, sendo que em Agosto existe o mesmo nível de ocorrências para todos os dias. Os meses de Março, Maio, Julho, Setembro e Dezembro para este grupo incluem apenas fins-de-semana (Apêndice C).

Por outro lado, o agrupamento a partir das correlações pressupõe a existência do Grupo 1 com 89 observações e o nível de consumo médio inferior – 4517 MW – constituído na sua maioria por dias de final de semana (principalmente domingos) e a totalidade dos feriados. A divisão mensal é clara, destacando-se maioritariamente por fins-de-semana, sendo que Julho, Agosto e Setembro apenas incluem domingos. O Grupo 2 com 277 observações e o nível médio superior – 5388 MW – é caracterizado essencialmente por dias de semana e um número bastante significativo de sábados. Neste sentido, os meses de Janeiro a Março, Novembro e Dezembro apresentam ocorrência de apenas dias de semana, sendo que os restantes meses incluem dias de semana e sábados.

Correlações				
	G1	% G1	G2	% G2
Seg	1	1%	51	18%
Ter	2	2%	50	18%
Qua	3	3%	49	18%
Qui	3	3%	50	18%
Sex	4	4%	49	18%
Sab	24	27%	28	10%
Dom	52	58%	0	0%
Total	89	100%	277	100%

Tabela 4.19 – Distribuição Diária por Grupo em 2004 (2 Grupos) – Correlações

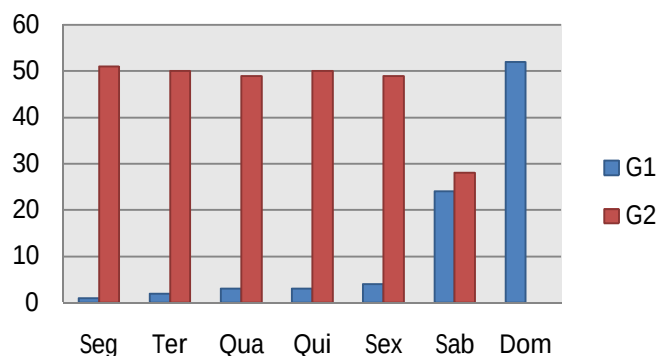


Figura 4.44 – Distribuição Dias da Semana (2 Grupos) – Correlações (2004)

4.4.2 3 GRUPOS

Como consequência da agregação dos perfis diários para o ano de 2005, de acordo com a abordagem a partir das distâncias euclidianas, é possível identificar que o Grupo 1 com 86 observações e o menor nível médio de consumo – 4553 MW – é constituído, na sua maioria, por fins-de-semana e alguns dos restantes dias, incluindo a maioria dos feriados do ano, não havendo meses que se destaquem em termos de distribuição mensal. O segundo conjunto com 194 observações e o maior nível de consumo – 6369 MW – por sua vez, possui um número maior de dias de semana (incluindo os feriados de 8 de Fevereiro, 1 e 8 de Dezembro), alguns sábados e menos domingos. O mês de Janeiro destaca-se pela presença apenas de fins-de-semana, ao passo que Abril, Maio, Agosto, Setembro e Outubro são caracterizados apenas por dias de semana. O agrupamento 3, por último, com 85 observações e um nível intermédio de consumo de

energia relativamente aos restantes – 5489 MW – é representado apenas por dias de semana para os meses de Janeiro a Março, Junho, Julho, Novembro e Dezembro.

Distâncias Euclidianas						
	G1	% G1	G2	% G2	G3	% G3
Seg	4	5%	33	17%	15	18%
Ter	1	1%	34	18%	17	20%
Qua	1	1%	30	15%	21	25%
Qui	2	2%	35	18%	15	18%
Sex	3	3%	32	16%	17	20%
Sab	33	38%	20	10%	0	0%
Dom	42	49%	10	5%	0	0%
Total	86	100%	194	100%	85	100%

Tabela 4.20 – Distribuição Diária por Grupo em 2005 (3 Grupos) – Distâncias Euclidianas

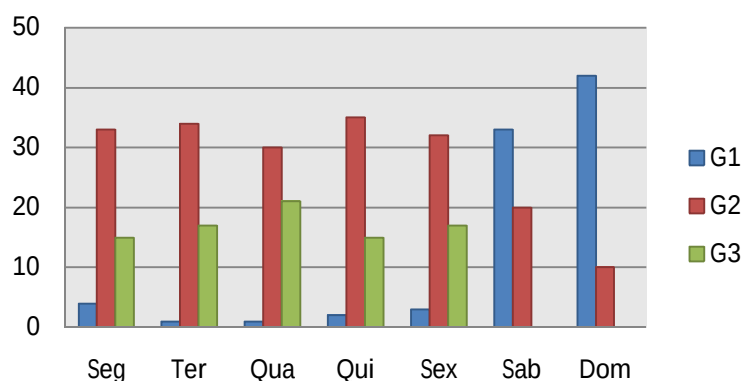


Figura 4.45 – Distribuição Dias da Semana (3 Grupos) – Distâncias Euclidianas (2005)

O agrupamento a partir das correlações origina a existência do Grupo 1 com um nível intermédio de consumo – 5501 MW – alcançado pelas 149 observações divididas apenas em dias de semana originários dos meses de Março a Outubro. O segundo grupo com 137 observações e o nível de consumo médio superior – 5827 MW – é constituído maioritariamente por dias de semana (incluindo quatro feriados – 25 de Abril, 10 de Junho, 15 de Agosto e 5 de Outubro), bastantes sábados e apenas um domingo. Os meses de Janeiro a Março, assim como Novembro e Dezembro incluem apenas dias de semana, tendo os restantes uma maior ocorrência de sábados e alguns dias mais. O

Grupo 3 com 79 observações e o nível de consumo inferior – 4808 MW – possui poucos dias de semana, alguns sábados e bastantes domingos, sendo incluída a maioria dos feriados. Neste sentido, os meses de Janeiro a Março, Novembro e Dezembro são constituídos na sua maioria por fins-de-semana e alguns dias de semana, e os meses de Abril a Outubro maioritariamente por domingos.

Correlações						
	G1	% G1	G2	% G2	G3	% G3
Seg	28	19%	24	18%	0	0%
Ter	31	21%	19	14%	2	3%
Qua	30	20%	22	16%	0	0%
Qui	30	20%	19	14%	3	4%
Sex	30	20%	21	15%	1	1%
Sab	0	0%	31	23%	22	28%
Dom	0	0%	1	1%	51	65%
Total	149	100%	137	100%	79	100%

Tabela 4.21 – Distribuição Diária por Grupo em 2004 (3 Grupos) – Correlações

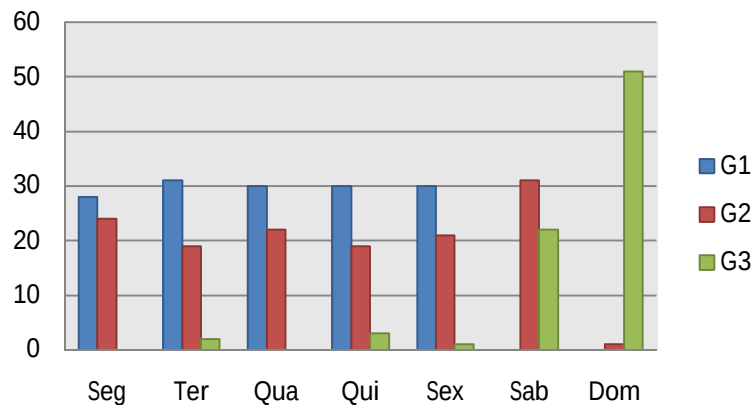


Figura 4.46 – Distribuição Dias da Semana (3 Grupos) – Correlações (2005)

4.4.3 4 GRUPOS

A distribuição em quatro agrupamentos dos perfis diários do ano de 2004 de acordo com as distâncias euclidianas determina que o Grupo 1 com 72 observações e com o segundo nível de consumo de energia mais baixo – 4860 MW – é constituído na sua maioria por fins-de-semana e segundas, bem como alguns dias de semana (incluindo o feriado de 24 de Fevereiro) para todos os meses, excepto Abril Maio e Agosto. Por sua vez, o segundo grupo na sua totalidade com 70 observações que agrupam o nível mais baixo de consumo – 4213 MW – é representado maioritariamente por domingos, bastantes sábados e alguns dias de semana (incluindo a maioria dos feriados) para todos os meses. O Grupo 3, por sua vez, com 161 observações que agregam o segundo nível médio de consumo superior – 5401 MW – contém essencialmente dias de semana (incluindo feriados de 1 e 8 de Dezembro), alguns sábados e um domingo. Os meses de Março a Outubro são constituídos apenas por dias de semana, sendo os restantes caracterizados por alguns dias de semana e fins-de-semana. O quarto grupo com 63 observações e com o nível de consumo superior – 6032 MW – inclui apenas dias de semana, com contribuição unicamente dos meses de Janeiro a Março, Novembro e Dezembro, não havendo ocorrência de feriados.

Distâncias Euclidianas								
	G1	% G1	G2	% G2	G3	% G3	G4	% G4
Seg	15	21%	2	3%	25	16%	10	16%
Ter	6	8%	1	1%	31	19%	14	22%
Qua	6	8%	0	0%	34	21%	12	19%
Qui	6	8%	2	3%	30	19%	15	24%
Sex	8	11%	1	1%	32	20%	12	19%
Sab	21	29%	23	33%	8	5%	0	0%
Dom	10	14%	41	59%	1	1%	0	0%
Total	72	100%	70	100%	161	100%	63	100%

Tabela 4.22 – Distribuição Diária por Grupo em 2004 (4 Grupos) – Distâncias Euclidianas

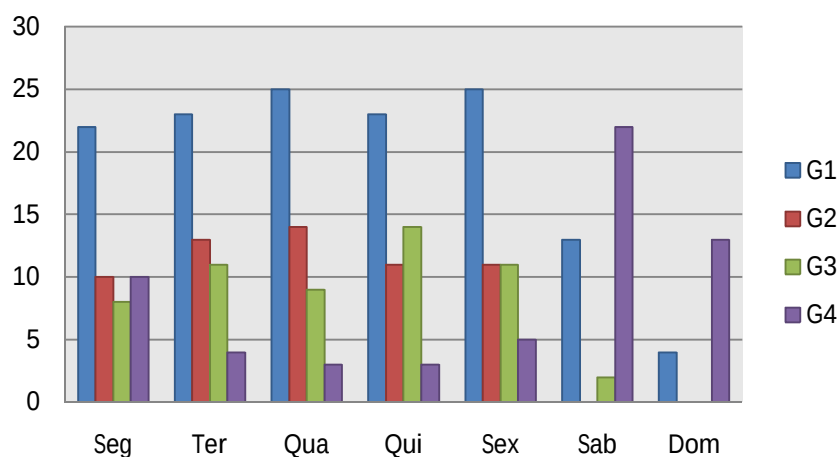


Figura 4.47 – Distribuição Dias da Semana (4 Grupos) – Distâncias Euclidianas (2004)

Com o agrupamento em quatro grupos a partir das correlações é reconhecido o Grupo 1 com o segundo nível médio de consumo inferior – 4598 MW – descrito por 30 observações, constituído na grande maioria por domingos derivado aos meses de Janeiro a Março, Novembro e Dezembro, incluindo os feriados desses períodos e o de 11 de Abril. O segundo grupo com 129 observações e o nível médio de consumo superior – 5531 MW – é representado apenas por dias de semana e sábados, não incluindo feriados. Os dias de semana têm origem nos meses de Janeiro a Março, Novembro e Dezembro e os sábados nos restantes. No Grupo 3 as 56 observações que agregam o nível médio de consumo de energia inferior – 4432 MW – são repartidas essencialmente entre sábados (Janeiro a Março, Novembro e Dezembro), domingos (Abril a Outubro) e poucos dias de semana, abrangendo seis feriados do ano. O quarto grupo com o maior número de observações – 151 – e o segundo nível médio de consumo superior – 5264 MW – é apenas constituído por dias de semana, de acordo com os meses de Março a Outubro, sem ocorrência de feriados.

Correlações								
	G1	% G1	G2	% G2	G3	% G3	G4	% G4
Seg	1	3%	21	16%	0	0%	30	20%
Ter	1	3%	20	16%	1	2%	30	20%
Qua	2	7%	19	15%	0	0%	31	21%
Qui	1	3%	21	16%	1	2%	30	20%
Sex	0	0%	19	15%	4	7%	30	20%
Sab	3	10%	29	22%	20	36%	0	0%
Dom	22	73%	0	0%	30	54%	0	0%
Total	30	100%	129	100%	56	100%	151	100%

Tabela 4.23 – Distribuição Diária por Grupo em 2004 (4 Grupos) – Correlações

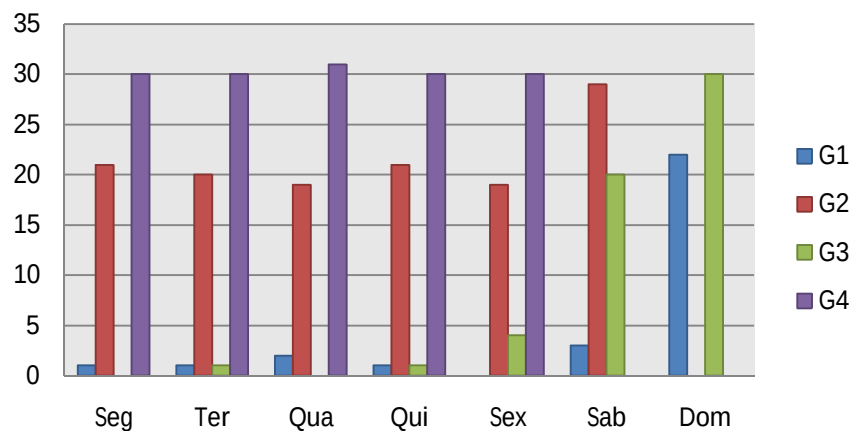


Figura 4.48 – Distribuição Dias da Semana (4 Grupos) – Correlações (2004)

4.4.4 5 GRUPOS

A divisão dos perfis diários do ano de 2005 em cinco grupos, segundo a abordagem das distâncias euclidianas, pressupõe a existência do primeiro conjunto com 135 observações que convergem num nível intermédio de consumo – 5491 MW – que se destaca por possuir um número elevado de dias de semana (abrangendo apenas o feriado de 8 de Dezembro) e alguns sábados. Desta forma, destacam-se Janeiro e Fevereiro com apenas por fins-de-semana, os meses de Abril a Outubro com apenas dias de semana. O Grupo 2 de 59 observações e o maior nível de consumo de energia – 6505 MW – é representado apenas por dias de semana correspondente aos meses de Janeiro a Março, Novembro e Dezembro, não incluindo nenhum feriado. O terceiro grupo com 55 observações que agregam o segundo maior nível de consumo – 5937 MW

– é constituído por alguns dias de semana (incluindo o feriado de 1 de Dezembro) e poucos sábados repartidos pelos vários meses, excepto Abril e Agosto sem ocorrências. O Grupo 4 numa totalidade de 60 observações que convergem no segundo nível inferior de consumo – 5003 MW – contém maioritariamente fins-de-semana, algumas segundas-feiras e poucos dias de semana repartidos pelos vários meses, entre os quais quatro feriados – 8 de Fevereiro, 25 de Março, 10 de Junho e 25 de Dezembro. O quinto grupo com o nível médio inferior de consumo – 4392 MW – num conjunto de 56 observações destaca-se pelo elevado número de domingos e alguns sábados (Janeiro, Março, Junho, Julho e Setembro), assim como poucos dias de semana repartidos, incluindo-se ainda a totalidade dos restantes feriados.

Distâncias Euclidianas										
	G1	% G1	G2	% G2	G3	% G3	G4	% G4	G5	% G5
Seg	22	16%	10	17%	8	15%	10	17%	2	4%
Ter	23	17%	13	22%	11	20%	4	7%	1	2%
Qua	25	19%	14	24%	9	16%	3	5%	1	2%
Qui	23	17%	11	19%	14	25%	3	5%	1	2%
Sex	25	19%	11	19%	11	20%	5	8%	0	0%
Sab	13	10%	0	0%	2	4%	22	37%	16	29%
Dom	4	3%	0	0%	0	0%	13	22%	35	63%
Total	135	100%	59	100%	55	100%	60	100%	56	100%

Tabela 4.24 – Distribuição Diária por Grupo em 2005 (5 Grupos) – Distâncias Euclidianas

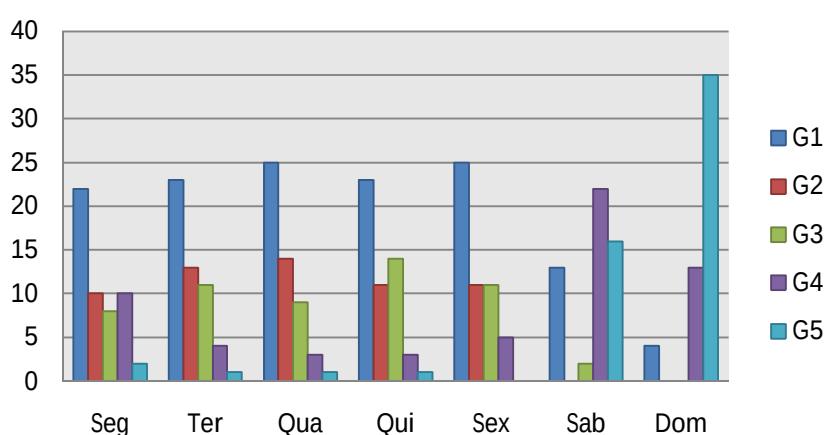


Figura 4.49 – Distribuição Dias da Semana (5 Grupos) – Distâncias Euclidianas (2005)

Da distribuição dos dias a partir das correlações conclui-se que o Grupo 1 com 31 observações e um nível de consumo de 5107 MW, contém maioritariamente domingos, alguns sábados e poucos dias de semana repartidos por todos os meses, excepto os de Abril a Outubro, sendo abrangidos a maior parte dos feriados. No segundo grupo, a totalidade das 53 observações agregadas no nível inferior de consumo – 4616 MW – é distribuída maioritariamente por domingos, alguns sábados (Janeiro a Março e Novembro) e poucos dias de semana, havendo ocorrência de oito feriados. O Grupo 3 com 96 observações e o segundo nível médio de consumo superior – 5601 MW – é constituído apenas por dias de semana (Abril a Julho, Setembro e Outubro), sem a presença de feriados. Da mesma forma, o quarto grupo, de 53 observações com o nível de consumo médio intermédio – 5320 MW – é também representado apenas por dias de semana (Março a Junho e Agosto a Outubro) e não inclui feriados. O Grupo 5 com o maior número de observações – 132 – e o maior nível médio de consumo – 5872 MW – destaca-se pela existência de dias de semana (Janeiro a Março, Novembro e Dezembro) e sábados, acrescentando ainda o feriado de 10 de Junho.

Correlações										
	G1	% G1	G2	% G2	G3	% G3	G4	% G4	G5	% G5
Seg	0	0%	3	6%	11	11%	17	32%	21	16%
Ter	1	3%	1	2%	20	21%	11	21%	19	14%
Qua	0	0%	1	2%	22	23%	8	15%	21	16%
Qui	1	3%	2	4%	19	20%	11	21%	19	14%
Sex	0	0%	1	2%	24	25%	6	11%	21	16%
Sab	9	29%	13	25%	0	0%	0	0%	31	23%
Dom	20	65%	32	60%	0	0%	0	0%	0	0%
Total	31	100%	53	100%	96	100%	53	100%	132	100%

Tabela 4.25 – Distribuição Diária por Grupo em 2005 (5 Grupos) – Correlações

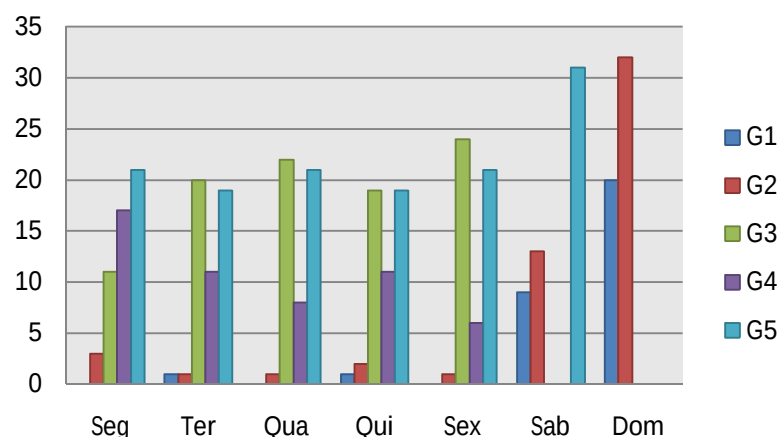


Figura 4.50 – Distribuição Dias da Semana (5 Grupos) – Correlações (2005)

4.4.5 6 GRUPOS

Os resultados de distribuição dos dias da semana do ano de 2004 pelos seis agrupamentos, a partir das distâncias euclidianas, indicam que o Grupo 1 de 92 observações com um nível médio de consumo de 5169 MW é representado essencialmente por dias de semana (abrangendo o feriado de 1 de Dezembro) dos quais são responsáveis os meses de Março a Outubro. Os segundo e quinto grupos, com 35 e 28 observações que agregam os dois níveis superiores de consumo – 5901 e 6195 MW, respectivamente – são constituídos apenas por dias de semana (não incluindo nenhum feriado) pertencentes a todos os meses, excepto entre Abril e Junho, Agosto e Outubro no segundo grupo e Fevereiro, Abril a Outubro no quinto grupo. Os Grupos 3 e 4 com os dois níveis inferiores de consumo – 4658 e 4134 MW, respectivamente – numa totalidade de 58 e 54 observações são caracterizados maioritariamente por fins-de-semana, principalmente sábados no terceiro (incluindo também o feriado de 24 de Fevereiro) e domingos no quarto (abrangendo nove feriados do ano). O sexto grupo descrito por 99 observações que agrupam um nível de consumo de 5510 MW é constituído essencialmente por dias de semana (bem como o feriado de 8 de Dezembro) e alguns sábados repartidos por vários meses.

Distâncias Euclidianas												
	G1	% G1	G2	% G2	G3	% G3	G4	% G4	G5	% G5	G6	% G6
Seg	21	23%	7	20%	5	9%	2	4%	3	11%	14	14%
Ter	14	15%	6	17%	4	7%	1	2%	8	29%	19	19%
Qua	15	16%	6	17%	2	3%	0	0%	6	21%	23	23%
Qui	15	16%	9	26%	2	3%	2	4%	6	21%	19	19%
Sex	17	18%	7	20%	3	5%	1	2%	5	18%	20	20%
Sab	7	8%	0	0%	30	52%	11	20%	0	0%	4	4%
Dom	3	3%	0	0%	12	21%	37	69%	0	0%	0	0%
Total	92	100%	35	100%	58	100%	54	100%	28	100%	99	100%

Tabela 4.26 – Distribuição Diária por Grupo em 2004 (6 Grupos) – Distâncias Euclidianas

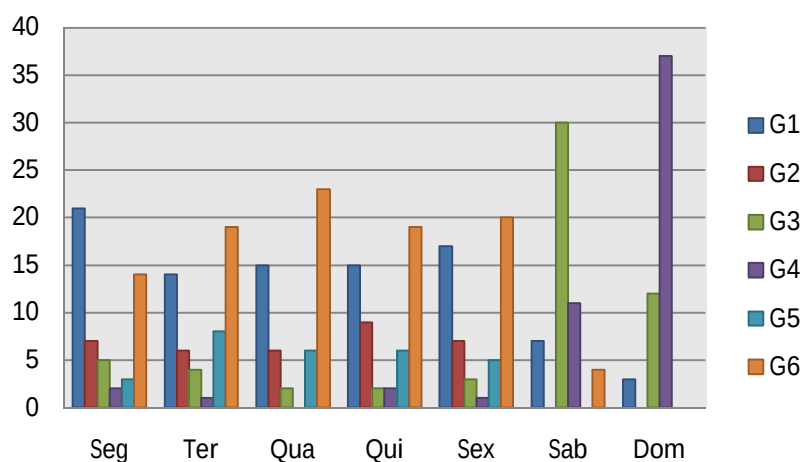


Figura 4.51 – Distribuição Dias da Semana (6 Grupos) – Distâncias Euclidianas (2004)

Segundo a abordagem a partir das correlações, os primeiro e terceiro grupos com 30 e 28 observações de 5411 e 5009 MW, respectivamente, são descritos apenas por alguns dias de semana (sem ocorrência de feriados) de acordo com os meses de Abril a Julho e Outubro e os meses de Março a Maio, Agosto a Outubro, respectivamente. O Grupo 2, com o segundo maior número de observações – 93 – e um nível de consumo de 5293 MW é também constituído somente por dias de semana (Abril a Outubro), não incluindo feriados. O quarto grupo com o segundo nível médio de consumo inferior – 4598 MW – agrupado em 30 observações, das quais maioritariamente domingos correspondentes aos meses de Janeiro a Abril, Novembro e Dezembro, períodos que incluem a ocorrência de sete dos feriados para este grupo. O Grupo 5 com o maior

número de observações – 129 – e o maior nível médio de consumo – 5531 MW – é descrito apenas por dias de semana e sábados (Janeiro a Março, Novembro e Dezembro), não abrangendo feriados. No sexto grupo, a totalidade das 56 observações agrupam o menor nível médio de consumo – 4432 MW – e são repartidas maioritariamente por fins-de-semana de todos os meses, onde são incluídos os restantes seis feriados do ano.

Correlações												
	G1	% G1	G2	% G2	G3	% G3	G4	% G4	G5	% G5	G6	% G6
Seg	0	0%	21	23%	9	32%	1	3%	21	16%	0	0%
Ter	6	20%	18	19%	6	21%	1	3%	20	16%	1	2%
Qua	8	27%	19	20%	4	14%	2	7%	19	15%	0	0%
Qui	6	20%	19	20%	5	18%	1	3%	21	16%	1	2%
Sex	10	33%	16	17%	4	14%	0	0%	19	15%	4	7%
Sab	0	0%	0	0%	0	0%	3	10%	29	22%	20	36%
Dom	0	0%	0	0%	0	0%	22	73%	0	0%	30	54%
Total	30	100%	93	100%	28	100%	30	100%	129	100%	56	100%

Tabela 4.27 – Distribuição Diária por Grupo em 2004 (6 Grupos) – Correlações

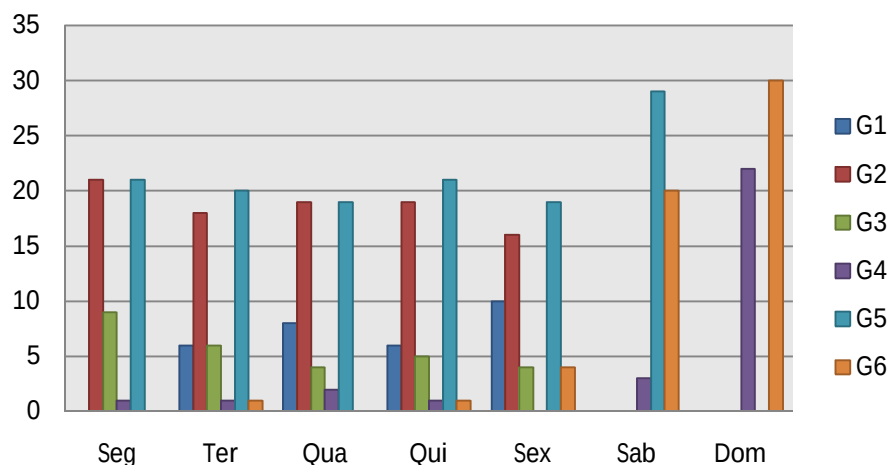


Figura 4.52 – Distribuição Dias da Semana (6 Grupos) – Correlações (2004)

4.4.6 7 GRUPOS

Finalmente, como resultado da divisão dos perfis diários do ano 2005 em 7 grupos, a partir das distâncias euclidianas, identifica-se o Grupo 1 com 58 observações e o maior nível médio de consumo de energia – 6510 MW – que inclui apenas dias de semana (Janeiro a Março, Novembro e Dezembro) e nenhum feriado. O segundo grupo com o nível intermédio de consumo – 5392 MW – é descrito por 72 observações repartidas maioritariamente por dias de semana (Março a Outubro, Novembro e Dezembro) e alguns fins-de-semana (Janeiro e Fevereiro), abrangendo o feriado de 8 de Dezembro. Os Grupos 3 e 5, com os menores níveis de consumo – 4286 e 4699 MW – numa totalidade de 38 e 36 observações são descritos por poucos dias de semana, e maioritariamente domingos e sábados, respectivamente. No terceiro grupo são incluídos seis dos feriados do ano e outros quatro no quinto. Os quarto e sétimo grupos de 72 e 47 observações com níveis médios de consumo de 5624 e 5978 MW são representados maioritariamente por dias de semana e alguns sábados, sendo que o quarto não inclui feriados e sétimo abrange os de 25 de Março e 1 de Dezembro. O Grupo 6, por sua vez, com o terceiro nível inferior de consumo – 5097 MW – obtido pelas 42 observações, inclui maioritariamente fins-de-semana e alguns dias de semana (entre os quais o feriado de 8 de Fevereiro) repartidos pelos diferentes meses.

Distâncias Euclidianas														
	G1	% G1	G2	% G2	G3	% G3	G4	% G4	G5	% G5	G6	% G6	G7	% G7
Seg	9	16%	13	18%	2	5%	10	14%	1	3%	9	21%	8	17%
Ter	13	22%	12	17%	1	3%	14	19%	0	0%	4	10%	8	17%
Qua	14	24%	11	15%	0	0%	14	19%	1	3%	3	7%	9	19%
Qui	11	19%	11	15%	1	3%	15	21%	0	0%	3	7%	11	23%
Sex	11	19%	14	19%	0	0%	12	17%	2	6%	3	7%	10	21%
Sab	0	0%	7	10%	2	5%	7	10%	26	72%	10	24%	1	2%
Dom	0	0%	4	6%	32	84%	0	0%	6	17%	10	24%	0	0%
Total	58	100%	72	100%	38	100%	72	100%	36	100%	42	100%	47	100%

Tabela 4.28 – Distribuição Diária por Grupo em 2005 (7 Grupos) – Distâncias Euclidianas

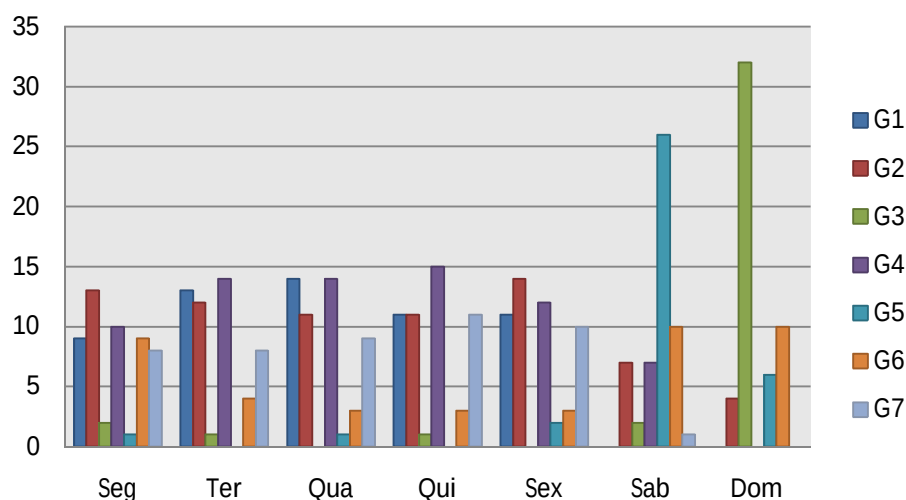


Figura 4.53 – Distribuição Dias da Semana (7 Grupos) – Distâncias Euclidianas (2005)

De acordo com os resultados de distribuição diária a partir das correlações o primeiro e quinto grupos descritos por 56 e 77 observações com níveis médios de consumo de 5635 e 5480 MW, respectivamente, são constituídos apenas por dias de semana, não havendo ocorrência de feriados. O Grupo 2 de apenas 16 observações que agregam o terceiro menor nível de consumo – 5132 MW – é também representado apenas por dias de semana do mês de Agosto e uma terça-feira de Março, não abrangendo feriados. Para os terceiro e quarto grupos com os níveis médios de consumo inferiores – 5107 e 4618 MW – obtidos pelas 31 e 52 observações, respectivamente, é identificada uma maior frequência para os fins-de-semana repartidos pelos diferentes meses do ano, incluindo em conjunto 12 dos 14 feriados do ano. Os Grupos 6 e 7 com médias de consumo de 5591 e 6041 MW, respectivamente, são representados por 53 e 80 observações que incluem maioritariamente dias de semana e sábados. Os restantes feriados de 10 de Junho e 15 de Agosto integram o sétimo grupo.

Correlações														
	G1	% G1	G2	% G2	G3	% G3	G4	% G4	G5	% G5	G6	% G6	G7	% G7
Seg	4	7%	4	25%	0	0%	2	4%	20	26%	9	17%	13	16%
Ter	10	18%	3	19%	1	3%	1	2%	18	23%	6	11%	13	16%
Qua	13	23%	3	19%	0	0%	1	2%	14	18%	7	13%	14	18%
Qui	12	21%	3	19%	1	3%	2	4%	15	19%	6	11%	13	16%
Sex	17	30%	3	19%	0	0%	1	2%	10	13%	10	19%	11	14%
Sab	0	0%	0	0%	9	29%	13	25%	0	0%	15	28%	16	20%
Dom	0	0%	0	0%	20	65%	32	62%	0	0%	0	0%	0	0%
Total	56	100%	16	100%	31	100%	52	100%	77	100%	53	100%	80	100%

Tabela 4.29 – Distribuição Diária por Grupo em 2005 (7 Grupos) – Correlações

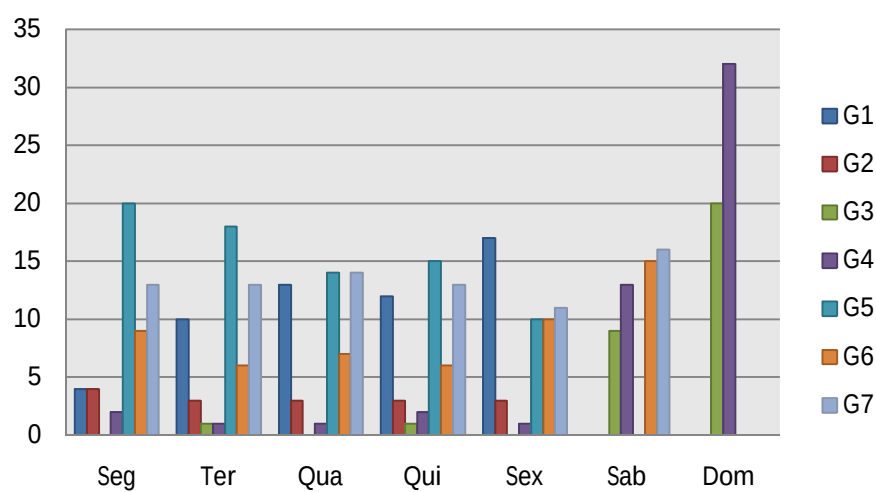


Figura 4.54 – Distribuição Dias da Semana (7 Grupos) – Distâncias Euclidianas (2005)

4.4.7 RESUMO DE RESULTADOS

Resumidamente, podem ser retiradas as seguintes conclusões relativamente à distribuição dos dias da semana pelos agrupamentos entre dois e sete grupos:

2 Grupos
Um grupo com o menor número de elementos e um nível médio de consumo de energia mais baixo é constituído essencialmente por fins-de-semana e feriados. Por outro lado, um grupo composto por o maior número de elementos e caracterizado por um nível médio de consumo mais elevado engloba essencialmente dias de semana.
3 Grupos
Um conjunto com o maior nível médio de consumo de energia abrange essencialmente dias de semana e possivelmente alguns sábados. Outro tipo de grupo, com níveis de consumo de energia inferiores é caracterizado maioritariamente por fins-de-semana e feriados. Por outro lado, um grupo com o nível de consumo médio intermédio é constituído só por dias de semana.
4 Grupos
Dois grupos com os níveis médios de consumo de energia mais baixos são compostos essencialmente por fins-de-semana e a maior parte dos feriados do ano. Por outro lado, os dois grupos com os maiores níveis de consumo de energia são representados, maioritariamente, por dias de semana e alguns sábados.

5 Grupos

Dois agrupamentos com os dois níveis de consumo médio mais baixos são caracterizados essencialmente pela presença de fins-de-semana (principalmente domingos) e feriados. Um grupo com níveis de consumo intermédios é constituído por dias de semana, podendo incluir alguns sábados. Outros dois grupos que apresentam níveis médios de consumo de energia mais elevados são compostos maioritariamente por dias de semana e alguns sábados.

6 Grupos

Dois grupos com os dois níveis médios de consumo mais baixos são representados por fins-de-semana e feriados. Outros dois grupos com níveis de consumo de energia intermédios englobam na sua maioria dias de semana. Dois grupos-tipo com elevados níveis de consumo são constituídos maioritariamente por dias de semana e alguns sábados.

7 Grupos

Dois grupos com os dois níveis de consumo de energia mais baixos englobam, na sua maioria, fins-de-semana e feriados. Três grupos com os maiores níveis de consumo de energia são representados essencialmente por dias de semana e eventualmente alguns sábados. Um grupo-tipo com consumos intermédios é caracterizado por dias de semana. Outro grupo com o terceiro nível médio inferior de consumo é constituído por dias de semana.

Nesta fase foi possível avaliar a distribuição dos dias da semana pelos diferentes conjuntos definidos, tornando-se interessante verificar concretamente qual a evolução diária dos dias representativos nos agrupamentos.

4.5 DIAS REPRESENTATIVOS DOS GRUPOS

Para ambos os anos em estudo, e para os agrupamentos definidos foram encontrados os dias mais representativos, tendo em conta a média dos grupos e os valores observados em termos de consumo de energia eléctrica nos perfis diários. Como medida de representatividade foi utilizado o mínimo absoluto da diferença entre esses dois valores, como definido na Secção 3. Deste modo, são apresentados de seguida os resultados obtidos:

Distâncias Euclidianas - 2004				Correlações - 2004		
	Dia	Dia Semana	Média (MW)	Dia	Dia Semana	Média (MW)
2 Grupos	30-Jun	Quarta-Feira	5523,433	25-Dez	Sábado (Feriado)	4517,188
	04-Jan	Domingo	4427,261	20-Abr	Terça-Feira	5394,92
3 Grupos	12-Abr	Segunda-Feira	4369,758	12-Out	Terça-Feira	5264,714
	06-Jan	Terça-Feira	5933,806	02-Nov	Terça-Feira	5476,556
	06-Out	Quarta-Feira	5289,123	14-Nov	Domingo	4506,82
4 Grupos	06-Mar	Sábado	4858,951	18-Jan	Domingo	4596,224
	14-Mar	Domingo	4238,399	25-Mar	Quinta-Feira	5532,176
	18-Dez	Terça-Feira	5401,198	25-Jul	Domingo	4452,938
	18-Nov	Quinta-Feira	6030,364	12-Out	Terça-Feira	5264,714
5 Grupos	17-Out	Domingo	4142,445	17-Abr	Sábado	4400,31
	23-Nov	Terça-Feira	6058,191	26-Jan	Segunda-Feira	5804,664
	22-Set	Quarta-Feira	5535,804	08-Nov	Segunda-Feira	5360,214
	03-Jan	Sábado	4679,179	13-Nov	Sábado	4919,049
	14-Abr	Quarta-Feira	5183,223	05-Mai	Quarta-Feira	5356,594
6 Grupos	28-Abr	Quarta-Feira	5170,407	04-Jun	Sexta-Feira	5414,883
	15-Jan	Quinta-Feira	5888,959	08-Set	Quarta-Feira	5297,469
	19-Ago	Quinta-Feira	4654,233	31-Ago	Terça-Feira	5046,878
	17-Out	Domingo	4142,445	18-Jan	Domingo	4596,224
	22-Jan	Quinta-Feira	6198,549	25-Mar	Quinta-Feira	5532,176
	26-Mar	Sexta-Feira	5507,075	25-Jul	Domingo	4452,938
7 Grupos	11-Jan	Domingo	4389,873	16-Nov	Terça-Feira	5923,41
	13-Ago	Sexta-Feira	4749,181	12-Out	Terça-Feira	5264,714
	09-Mai	Domingo	4015,932	13-Mar	Sábado	4722,312
	15-Jan	Quarta-Feira	5888,959	30-Out	Sábado	4618,672
	22-Jan	Quinta-Feira	6198,549	24-Out	Domingo	4266,773
	23-Fev	Segunda-Feira	5519,009	29-Fev	Domingo	4607,516
	10-Nov	Quarta-Feira	5589,925	26-Mar	Sexta-Feira	5507,075

Tabela 4.30 – Dias Representativos dos Grupos (2004)

Distâncias Euclidianas - 2005				Correlações - 2005		
	Dia	Dia Semana	Média (MW)	Dia	Dia Semana	Média (MW)
2 Grupos	07-Jun	Terça-Feira	5879,404	29-Set	Quinta-Feira	5673,446
	10-Jun	Sexta-Feira (Feriado)	4790,157	13-Nov	Domingo	4807,603
3 Grupos	10-Jul	Domingo	4552,055	28-Abr	Quinta-Feira	5502,714
	03-Jun	Sexta-Feira	5488,240	08-Nov	Terça-Feira	5823,756
	25-Fev	Sexta-Feira	6372,232	13-Nov	Domingo	4807,603
4 Grupos	31-Jan	Segunda-Feira	6453,81	05-Out	Quarta-Feira (Feriado)	4629,486
	05-Jul	Terça-Feira	5730,223	28-Abr	Quinta-Feira	5502,714
	15-Ago	Segunda-Feira (Feriado)	4486,733	10-Nov	Quinta-Feira	5869,279
	30-Mai	Segunda-Feira	5282,694	04-Dez	Domingo	5096,108
5 Grupos	18-Mar	Sexta-Feira	5490,695	04-Dez	Domingo	5096,108
	22-Fev	Terça-Feira	6507,027	05-Out	Quarta-Feira (Feriado)	4629,486
	14-Nov	Segunda-Feira	5923,183	28-Jul	Quinta-Feira	5598,403
	18-Ago	Quinta-Feira	5013,328	05-Mai	Quinta-Feira	5321,979
	09-Nov	Quarta-Feira	4385,465	10-Nov	Quinta-Feira	5869,279
6 Grupos	18-Ago	Quinta-Feira	5013,328	05-Mai	Quinta-Feira	5321,979
	18-Mar	Sexta-Feira	5490,695	05-Out	Quarta-Feira (Feriado)	4629,486
	08-Mar	Terça-Feira	6386,02	16-Nov	Quarta-Feira	6042,962
	14-Nov	Segunda-Feira	5923,183	17-Mar	Quinta-Feira	5581,277
	09-Out	Domingo	4385,465	04-Dez	Domingo	5096,108
	15-Dez	Quinta-Feira	6703,677	28-Jul	Quinta-Feira	5598,403
7 Grupos	10-Jan	Segunda-Feira	6507,029	30-Jun	Quinta-Feira	5636,745
	10-Mai	Terça-Feira	5391,32	01-Ago	Segunda-Feira	5127,554
	23-Out	Domingo	4296,796	04-Dez	Domingo	5096,108
	20-Out	Quinta-Feira	5626,043	05-Out	Quarta-Feira (Feriado)	4629,486
	30-Jun	Quinta-Feira	5636,745	19-Abr	Terça-Feira	5479,599
	04-Dez	Domingo	5096,108	17-Mar	Quinta-Feira	5581,277
	20-Jul	Quarta-Feira	5975,891	16-Nov	Quarta-Feira	6042,962

Tabela 4.31 – Dias Representativos dos Grupos (2005)

Ao analisar a Tabela 4.30 e a Tabela 4.31 podemos verificar quais os dias mais representativos em cada um dos conjuntos. É importante referir que os dias que se revelam como mais representativos em cada um dos agrupamentos não correspondem necessariamente aos dias da semana mais significativos assinalados anteriormente para cada um dos grupos mas, na maioria dos casos, são correspondentes.

Analisando comparativamente esses resultados de entre os agrupamentos entre dois e sete grupos, dentro da mesma abordagem e entre abordagens diferentes são apresentados de seguida os dias mais representativos que se repetem.

		Dias	Grupos
2004	Distâncias Euclidianas	15-Jan	6 e 7
		22-Jan	
		17-Out	5 e 6
	Correlações	18-Jan	4 e 6
		25-Mar	
		25-Jul	
		12-Out	3 e 7
	Distâncias Euclidianas e Correlações	26-Mar	6 (Dist. Euclidianas)
			7 (Correlações)
2005	Distâncias Euclidianas	18-Mar	5 e 6
		18-Ago	
		14-Nov	
	Correlações	13-Nov	2 e 3
		28-Abr	3 e 4
		10-Nov	4 e 5
		05-Mai	5 e 6
		28-Jul	
		17-Mar	6 e 7
		16-Nov	
		05-Out	4, 5, 6 e 7
		04-Dez	
	Distâncias Euclidianas e Correlações	30-Jun	7 (Dist. Euclidianas)
			7 (Correlações)
		04-Dez	7 (Dist. Euclidianas)
			4,5,6 e 7 (Correlações)

Tabela 4.32 – Dias Representativos Comuns

Após identificar os dias representativos no geral e os dias representativos mais comuns entre grupos e abordagens é importante analisar o comportamento diário dos dias representativos dentro de alguns agrupamentos exemplificativos.

Analizando as representações gráficas dos dias representativos para dois agrupamentos no ano de 2004, verifica-se para a abordagem a partir das distâncias euclidianas – Figura 4.55 – que o dia 30 de Junho (quarta-feira), representativo do Grupo 1, é responsável por um nível de consumo substancialmente superior ao dia 4 de Janeiro (domingo), respectivo ao Grupo 2, entre aproximadamente a uma e as 17 horas. O dia 4 de Janeiro segue um comportamento evolutivo ao longo das 24 horas que se distingue por um prolongamento dos fenómenos de aumento e diminuição de níveis de comportamento.

Por outro lado, de acordo com o agrupamento a partir das correlações – Figura 4.56 – os dois dias definidos como representativos indiciam da mesma forma uma diferenciação clara entre níveis de consumo, havendo, no entanto, cruzamento de níveis aproximadamente entre as três e as 21 horas, período no qual o dia 20 de Abril (terça-feira) supera o nível de consumo do feriado do dia 25 de Dezembro (sábado). Para esta abordagem, tal como na anterior, o dia com um nível de consumo inferior apresenta um atraso na reprodução do comportamento evolutivo horário relativamente ao dia com um nível de consumo superior.

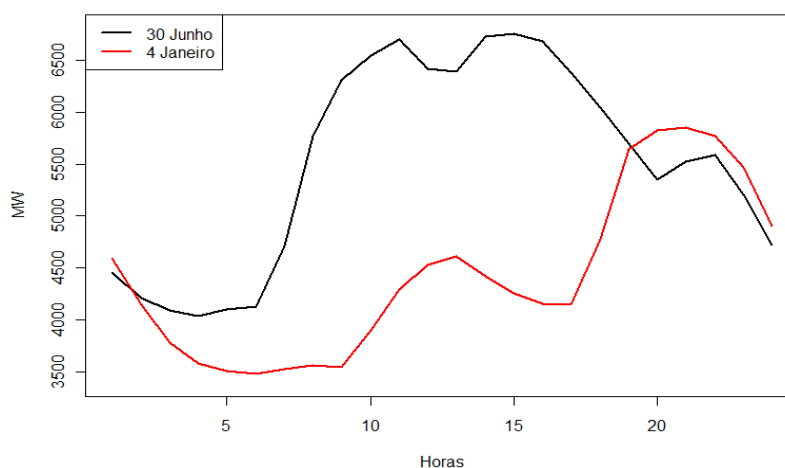


Figura 4.55 – Dias Representativos de 2004 – Distâncias Euclidianas (2 Grupos)

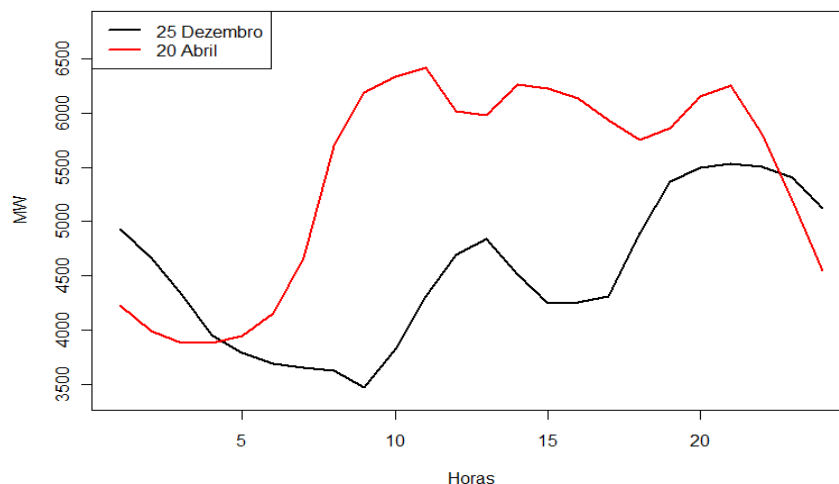


Figura 4.56 – Dias Representativos de 2004 – Correlações (2 Grupos)

Os dias mais representativos para a divisão em três grupos dos perfis diários de 2005, de acordo com as distâncias euclidianas – Figura 4.57 – apresentam um comportamento semelhante. O dia 10 de Julho (domingo) é o que mais se distingue dos restantes, não só pelo nível de consumo inferior aos restantes, mas também por um atraso, bem como alisamento da evolução horária verificada para os outros dois dias. Os dias 25 de Fevereiro e 3 de Junho, ambos sextas-feiras, são os que apresentam picos de consumo mais semelhantes, diferenciando-se não só por níveis de consumo superiores e intermédios, respectivamente, mas também pelo cruzamento de curvas de consumo por volta das 7 horas.

A abordagem a partir das correlações – Figura 4.58 – para o agrupamento em três grupos dos perfis diários do ano de 2005 destaca dias representativos com comportamentos de consumo significativamente distintos, com alguns cruzamentos ao longo das 24 horas do dia. Os dias 28 de Abril (quinta-feira) e 8 de Novembro (terça-feira), com os dois níveis de consumo de energia eléctrica maioritariamente superiores sobrepõem-se por volta das cinco, nove e 13 horas. O dia 13 de Novembro (domingo), com um consumo quase sempre inferior aos restantes, cruza o dia 28 de Abril por volta da uma hora e ao final do dia, às 18, 19 e 20 horas.

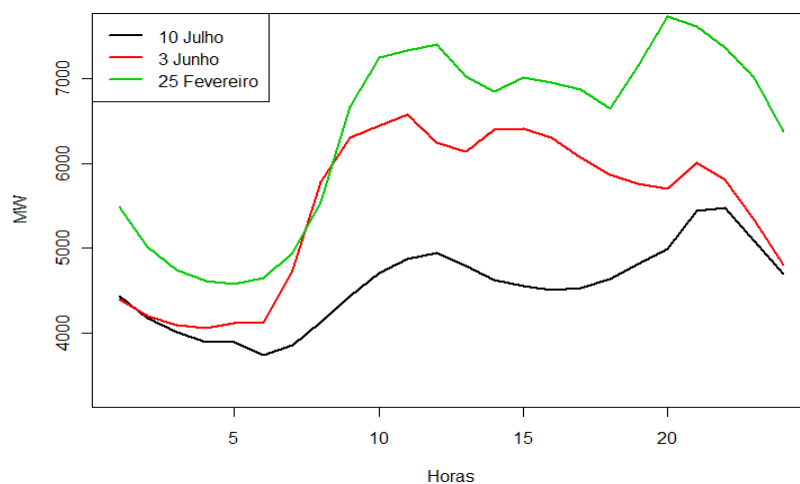


Figura 4.57 – Dias Representativos de 2005 – Distâncias Euclideanas (3 Grupos)

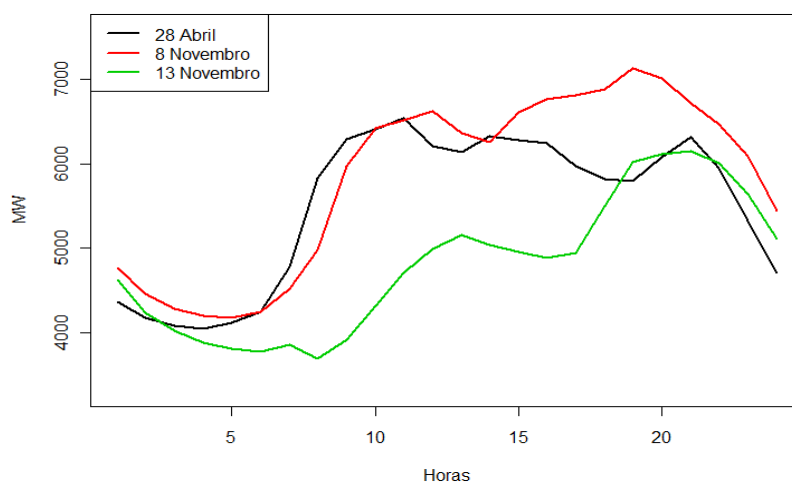


Figura 4.58 – Dias Representativos de 2005 – Correlações (3 Grupos)

Nesta Secção, foi desenvolvido todo o trabalho prático deste estudo que se iniciou com uma visão geral sobre os dados em causa, seguida pela aplicação do histograma por níveis aos dados e da análise comparativa dos perfis de agrupamento a partir das distâncias euclideanas e das correlações. Foram ainda comentados os resultados de refinamento dos perfis encontrados, analisada a distribuição dos dias de semana pelos perfis diários e comentados os resultados em termos de dias representativos.

5 CONSIDERAÇÕES FINAIS

5.1 CONCLUSÕES

Com este trabalho foi possível demonstrar a importância da utilização de histogramas por níveis no tratamento de dados recolhidos em períodos fixos de tempo, no sentido de extrair não só padrões de consumo de energia eléctrica, mas também perfis com granularidades temporais adequadas. As principais conclusões que podem ser retiradas em relação aos padrões de consumo de energia eléctrica são as seguintes:

- Em termos anuais, o ano de 2005 superou os níveis de 2004;
- Existem valores mais baixos de consumo de energia aos fins-de-semana, épocas festivas (Passagem de Ano, Carnaval, Páscoa e Natal) e feriados nacionais;
- Os meses iniciais e finais, correspondentes aos meses mais frios, registam níveis superiores aos meses intermédios, mais amenos;
- Decréscimo do 1º para o 2º trimestre e aumento nos dois trimestres posteriores;
- Os níveis de consumo são inferiores para o período das 0 às 5 horas, seguido do das 6 às 11 horas, variando os dois restantes períodos consoante a época do ano.
- O período das 12 às 17 horas supera os níveis de consumo das 18 às 23 horas nas épocas mais amenas (Final de Março, Abril, Outubro e início de Novembro), acontecendo o contrário nas épocas mais frias (Janeiro, Fevereiro, inícios de Março, fins de Novembro e Dezembro).

Foram ainda caracterizados perfis de consumo de energia eléctrica correspondentes aos diferentes horizontes temporais, bem como variantes dos mesmos, acima e abaixo da média, a partir das distâncias euclidianas e das correlações dos dados. Neste sentido, em termos de erro cometido pela escolha dos perfis em vez dos valores reais, o agrupamento a partir das correlações demonstrou-se mais adequado nos perfis com granularidades horárias, de 6 horas e diárias, obtendo-se melhores resultados no agrupamento a partir das distâncias euclidianas para os restantes perfis de dias da semana, semana e fim-de-semana, semanais, mensais e trimestrais. Desta forma, pode-se concluir que o agrupamento a partir das correlações dos dados é mais adequado para níveis do histograma com granularidades mais refinadas, verificando-se o contrário no agrupamento a partir das distâncias euclidianas.

Com o refinamento dos perfis encontrados, o erro absoluto após desdobramento comparativamente com o erro absoluto antes do desdobramento, indica algumas alterações nos resultados obtidos, sendo o agrupamento a partir das distâncias euclidianas o que apresenta melhor adequação para os perfis horários, 6 horas, dias da semana e semanais. Contrariamente, com o refinamento dos perfis o erro cometido nos perfis mensais demonstra menor eficácia para o agrupamento a partir das distâncias euclidianas e maior a partir das correlações. Por último, são sempre obtidas melhores performances após desdobramento para ambas as abordagens em termos de agrupamento no que respeita aos perfis diários, semana e fim-de-semana e trimestrais.

Neste sentido, pode-se resumidamente afirmar que o desdobramento de perfis é manifestamente mais favorável para os perfis diários, semana e fim-de-semana e trimestrais. Os restantes perfis horários, 6 horas, dias da semana e semanais, dependentes do agrupamento a partir das distâncias euclidianas demonstram uma melhor performance, contrariamente aos perfis mensais que apresentam melhores resultados para o agrupamento a partir das correlações.

No que respeita à distribuição dos dias da semana em agrupamentos entre dois a sete grupos conclui-se, na generalidade, que:

- Grupos constituídos por dias de semana têm nível médio de consumo superior;
- Grupos constituídos por fins-de-semana e feriados têm um menor nível médio de consumo igualmente inferior.

Por último, foram ainda apresentados os dias considerados mais representativos em cada um dos agrupamentos diários, assim como os dias comuns entre agrupamentos e abordagens, sendo que dentro dos mesmos agrupamentos os dias representativos apresentam algumas semelhanças de comportamento, embora que com algum desfasamento em termos de evolução horária ao longo do dia e com níveis de consumo distintos.

5.2 TRABALHO FUTURO

Existe ainda muito trabalho que pode ser realizado no âmbito da sumarização para diferentes horizontes temporais e, nomeadamente, no que respeita a dados de consumo de energia eléctrica. Algumas linhas de investigação interessantes são seguidamente enunciadas:

- Aplicação de outros métodos de agrupamento, além das k-Médias, para encontrar perfis de consumo, e comparar os resultados;
- Definir outros refinamentos de perfis para além do desdobramento na média dos traçados acima e abaixo da média e verificar se são obtidos erros menores, e consequentemente, perfis mais fiáveis;
- Transformação das metodologias utilizadas numa aplicação que permita a sumarização de dados em “tempo real”.

APÊNDICES

APÊNDICE A – CALENDÁRIOS DE 2004 E 2005

Ano 2004

Janeiro						
Seg	Ter	Qua	Qui	Sex	Sab	Dom
1			1	2	3	4
2	5	6	7	8	9	10
3	12	13	14	15	16	17
4	19	20	21	22	23	24
5	26	27	28	29	30	31

Fevereiro						
Seg	Ter	Qua	Qui	Sex	Sab	Dom
5						1
6	2	3	4	5	6	7
7	9	10	11	12	13	14
8	16	17	18	19	20	21
9	23	24	25	26	27	28

Março						
Seg	Ter	Qua	Qui	Sex	Sab	Dom
10	1	2	3	4	5	6
11	8	9	10	11	12	13
12	15	16	17	18	19	20
13	22	23	24	25	26	27
14	29	30	31			

Abril						
Seg	Ter	Qua	Qui	Sex	Sab	Dom
14				1	2	3
15	5	6	7	8	9	10
16	12	13	14	15	16	17
17	19	20	21	22	23	24
18	26	27	28	29	30	

Maio						
Seg	Ter	Qua	Qui	Sex	Sab	Dom
18					1	2
19	3	4	5	6	7	8
20	10	11	12	13	14	15
21	17	18	19	20	21	22
22	24	25	26	27	28	29
23	31					

Junho						
Seg	Ter	Qua	Qui	Sex	Sab	Dom
23		1	2	3	4	5
24	7	8	9	10	11	12
25	14	15	16	17	18	19
26	21	22	23	24	25	26
27	28	29	30			

Julho						
Seg	Ter	Qua	Qui	Sex	Sab	Dom
27				1	2	3
28	5	6	7	8	9	10
29	12	13	14	15	16	17
30	19	20	21	22	23	24
31	26	27	28	29	30	31

Agosto						
Seg	Ter	Qua	Qui	Sex	Sab	Dom
31						1
32	2	3	4	5	6	7
33	9	10	11	12	13	14
34	16	17	18	19	20	21
35	23	24	25	26	27	28
36	30	31				

Setembro						
Seg	Ter	Qua	Qui	Sex	Sab	Dom
36			1	2	3	4
37	6	7	8	9	10	11
38	13	14	15	16	17	18
39	20	21	22	23	24	25
40	27	28	29	30		

Outubro						
Seg	Ter	Qua	Qui	Sex	Sab	Dom
40				1	2	3
41	4	5	6	7	8	9
42	11	12	13	14	15	16
43	18	19	20	21	22	23
44	25	26	27	28	29	30

Novembro						
Seg	Ter	Qua	Qui	Sex	Sab	Dom
45	1	2	3	4	5	6
46	8	9	10	11	12	13
47	15	16	17	18	19	20
48	22	23	24	25	26	27
49	29	30				

Dezembro						
Seg	Ter	Qua	Qui	Sex	Sab	Dom
49			1	2	3	4
50	6	7	8	9	10	11
51	13	14	15	16	17	18
52	20	21	22	23	24	25
53	27	28	29	30	31	

Ano 2005

Janeiro							
	Seg	Ter	Qua	Qui	Sex	Sab	Dom
53						1	2
1	3	4	5	6	7	8	9
2	10	11	12	13	14	15	16
3	17	18	19	20	21	22	23
4	24	25	26	27	28	29	30
5	31						

Fevereiro							
	Seg	Ter	Qua	Qui	Sex	Sab	Dom
5		1	2	3	4	5	6
6	7	8	9	10	11	12	13
7	14	15	16	17	18	19	20
8	21	22	23	24	25	26	27
9	28						

Março							
	Seg	Ter	Qua	Qui	Sex	Sab	Dom
9		1	2	3	4	5	6
10	7	8	9	10	11	12	13
11	14	15	16	17	18	19	20
12	21	22	23	24	25	26	27
13	28	29	30	31			

Abril							
	Seg	Ter	Qua	Qui	Sex	Sab	Dom
13					1	2	3
14	4	5	6	7	8	9	10
15	11	12	13	14	15	16	17
16	18	19	20	21	22	23	24
17	25	26	27	28	29	30	

Maio							
	Seg	Ter	Qua	Qui	Sex	Sab	Dom
17							1
18	2	3	4	5	6	7	8
19	9	10	11	12	13	14	15
20	16	17	18	19	20	21	22
21	23	24	25	26	27	28	29
22	30	31					

Junho							
	Seg	Ter	Qua	Qui	Sex	Sab	Dom
22			1	2	3	4	5
23	6	7	8	9	10	11	12
24	13	14	15	16	17	18	19
25	20	21	22	23	24	25	26
26	27	28	29	30			

Julho							
	Seg	Ter	Qua	Qui	Sex	Sab	Dom
26					1	2	3
27	4	5	6	7	8	9	10
28	11	12	13	14	15	16	17
29	18	19	20	21	22	23	24
30	25	26	27	28	29	30	31

Agosto							
	Seg	Ter	Qua	Qui	Sex	Sab	Dom
31	1	2	3	4	5	6	7
32	8	9	10	11	12	13	14
33	15	16	17	18	19	20	21
34	22	23	24	25	26	27	28
35	29	30	31				

Setembro							
	Seg	Ter	Qua	Qui	Sex	Sab	Dom
35				1	2	3	4
36	5	6	7	8	9	10	11
37	12	13	14	15	16	17	18
38	19	20	21	22	23	24	25
39	26	27	28	29	30		

Outubro							
	Seg	Ter	Qua	Qui	Sex	Sab	Dom
39						1	2
40	3	4	5	6	7	8	9
41	10	11	12	13	14	15	16
42	17	18	19	20	21	22	23
43	24	25	26	27	28	29	30
44	31						

Novembro							
	Seg	Ter	Qua	Qui	Sex	Sab	Dom
44		1	2	3	4	5	6
45	7	8	9	10	11	12	13
46	14	15	16	17	18	19	20
47	21	22	23	24	25	26	27
48	28	29	30				

Dezembro							
	Seg	Ter	Qua	Qui	Sex	Sab	Dom
48				1	2	3	4
49	5	6	7	8	9	10	11
50	12	13	14	15	16	17	18
51	19	20	21	22	23	24	25
52	26	27	28	29	30	31	

APÊNDICE B – LISTAGEM DE FERIADOS

Listagem de feriados de 2004:

1 de Janeiro, Quinta-Feira
24 de Fevereiro, Terça-Feira
9 de Abril, Sexta-Feira
11 de Abril, Domingo
25 de Abril, Domingo
1 de Maio, Sábado
10 de Junho, Quinta-Feira
15 de Agosto, Domingo
5 de Outubro, Terça-Feira
1 de Novembro, Segunda-Feira
1 de Dezembro, Quarta-Feira
8 de Dezembro, Quarta-Feira
25 de Dezembro, Sábado

Listagem de feriados de 2005:

1 de Janeiro, Sábado
8 de Fevereiro, Terça-Feira
25 de Março, Sexta-Feira
27 de Março, Domingo
25 de Abril, Segunda-Feira
1 de Maio, Domingo
26 de Maio, Quinta-Feira
10 de Junho, Sexta-Feira
15 de Agosto, Segunda-Feira
5 de Outubro, Quarta-Feira
1 de Novembro, Terça-Feira
1 de Dezembro, Quinta-Feira
8 de Dezembro, Quinta-Feira
25 de Dezembro, Domingo

APÊNDICE C – DISTRIBUIÇÃO MENSAL POR DIAS DA SEMANA (2 GRUPOS)

Distâncias Euclidianas - 2004 (2 Grupos)

G1	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez	Total
Seg	4	4	5	3	5	4	4	1	4	3	4	4	45
Ter	4	3	5	4	4	5	4	2	4	3	5	4	47
Qua	4	4	5	4	4	5	4	1	5	4	4	5	49
Qui	4	4	4	5	4	3	5	1	5	4	4	5	48
Sex	4	4	4	4	4	3	5	1	4	5	4	5	47
Sáb	3	2	0	0	0	0	0	0	0	0	2	3	10
Dom	0	0	0	0	0	0	0	0	0	0	0	3	3

G2	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez	Total
Seg	0	0	0	1	0	0	0	4	0	1	1	0	7
Ter	0	1	0	0	0	0	0	3	0	1	0	0	5
Qua	0	0	0	0	0	0	0	3	0	0	0	0	3
Qui	1	0	0	0	0	1	0	3	0	0	0	0	5
Sex	1	0	0	1	0	1	0	3	0	0	0	0	6
Sáb	2	2	4	4	5	4	5	4	4	5	2	1	42
Dom	4	5	4	4	5	4	4	5	4	5	4	1	49

Correlações - 2004 (2 Grupos)

G1	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez	Total
Seg	0	0	0	0	0	0	0	0	0	0	1	0	1
Ter	0	1	0	0	0	0	0	0	0	1	0	0	2
Qua	0	0	0	0	0	0	0	0	0	0	0	3	3
Qui	1	0	0	0	0	1	0	0	0	0	0	1	3
Sex	1	0	0	1	0	0	0	0	0	0	0	2	4
Sáb	5	4	4	2	1	0	0	0	0	0	4	4	24
Dom	4	5	4	4	5	4	4	5	4	5	4	4	52

G2	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez	Total
Seg	4	4	5	4	5	4	4	5	4	4	4	4	51
Ter	4	3	5	4	4	5	4	5	4	3	5	4	50
Qua	4	4	5	4	4	5	4	4	5	4	4	2	49
Qui	4	4	4	5	4	3	5	4	5	4	4	4	50
Sex	4	4	4	4	4	4	5	4	4	5	4	3	49
Sáb	0	0	0	2	4	4	5	4	4	5	0	0	28
Dom	0	0	0	0	0	0	0	0	0	0	0	0	0

REFERÊNCIAS BIBLIOGRÁFICAS

- [1] Álvarez, F. M., Troncoso, A., Riquelme, J. C., & Riquelme, J. M. (2007). Discovering Patterns in Electricity Price Using Clustering Techniques. *Proceedings of the International Conference on Renewable Energies and Power Quality*, (pp. 245-252).
- [2] Arasu, A., & Manku, G. S. (2004). Approximate counts and quantiles over sliding windows. *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of Database Systems (PODS)*, (pp. 286-296).
- [3] *Autocorrelation*. (24 de Agosto de 2010). Obtido em 30 de Agosto de 2010, de Wikipedia: <http://en.wikipedia.org/wiki/Autocorrelation>
- [4] Babcock, B., Babu, S., Datar, M., Motwani, R., & Widom, J. (2002). Models and Issues in Data Stream Systems. *Proceedings of the Twenty-first ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)* (pp. 1-16). ACM.
- [5] Babcock, B., Datar, M., & Motwani, R. (2002). Sampling from a moving window over streaming data. *Proceedings of the thirteenth annual ACM-SIAM Symposium on Discrete Algorithms (SODA)* (pp. 633 - 634). Society for Industrial and Applied Mathematics (SIAM).
- [6] Babcock, B., Datar, M., Motwani, R., & O'Callaghan, L. (2003). Maintaining Variance and k-Medians over Data Stream Windows. *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of Database Systems (PODS)* (pp. 234 - 243). ACM.
- [7] Bifet, A. (2009). Adaptive Learning and Mining for Data Streams and Frequent Patterns. *PhD Thesis*. Universitat Politècnica de Catalunya.
- [8] Chen, C. S., Kang, M., Hwang, J., & Huang, C. (1999). Implementation of the Load Survey System in Taipower. *IEEE Transmission and Distribution Conference vol.1*, (pp. 300-304).

- [9] Chen, C., Hwang, J., & Huang, C. (1997). Application of load survey systems to proper tariff design. *IEEE Transactions on Power Systems*, Vol. 12, No. 4, (pp. 1746-1751).
- [10] Chen, C., Hwang, J., Tzeng, Y., Huang, C., & Cho, M. (1996). Determination of Customer Load Characteristics by Load Survey System at Taipower. *IEEE Transactions on Power Delivery*, Vol.11, No. 3, (pp. 1430-1436).
- [11] Chen, C., Kang, M. S., Hwang, J. C., & Huang, C. W. (2001). Temperature effect to distribution system load profiles and feeder losses. *IEEE Transactions on Power Systems* , (pp. 916-921).
- [12] Chen, C., Kang, M., Hwang, J., & Huang, C. (2001). Temperature adaptive switching operation for distribution systems. *IEEE Transactions on Power Delivery volume: 16, issue: 4*, (pp. 694-699).
- [13] Chen, Y., Dong, G., Han, J., W. Wah, B., & Wang, J. (2002). Multi-Dimensional Regression Analysis of Time-Series Data Streams. *Proceedings of the 28th VLDB Conference*, (pp. 323-334).
- [14] Chi, Y., Wang, H., Yu, P. S., & Muntz, R. R. (2004). Moment: Maintaining Closed Frequent Itemsets over a Stream Sliding Window. *Proceedings of Fourth IEEE International Conference on Data Mining (ICDM'04)*, (pp. 59-66).
- [15] Chicco, G., Napoli, R., Postolache, P., Scutariu, M., & Toader, C. (2003). Customer characterization options for improving the tariff offer. *IEEE Transactions on Power Systems*, Vol. 18, No. 1 , 381-387.
- [16] Chicco, G., Napoli, R., Postolache, P., Scutariu, M., & Toader, C. (2001). Electric energy customer characterisation for developing dedicated market strategies. *IEEE Porto Power Tech Proceedings*.
- [17] Coeficiente de correlação de Pearson. (15 de Julho de 2010). Obtido em 30 de Agosto de 2010, de Wikipédia: http://pt.wikipedia.org/wiki/Coeficiente_de_correla%C3%A7%C3%A3o_de_Pearson

[18] *Data Mining Algorithms In R/Clustering/K-Means*. (10 de Janeiro de 2010). Obtido em 30 de Agosto de 2010, de Wikibooks: http://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Clustering/K-Means

[19] Datar, M., & Muthukrishnan, S. (2002). Estimating Rarity and Similarity over Data Stream Windows. *Proceedings of the 10th Annual European Symposium on Algorithms (ESA)*, (pp. 323–334).

[20] Datar, M., Gionis, A., Indyk, P., & Motwani, R. (2002). Maintaining stream statistics over sliding window. *Proceedings of the 13th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2002)* (pp. 635-644). Society for Industrial and Applied Mathematics (SIAM).

[21] Domingos, P., & Hulten, G. (2000). Mining high-speed data streams. *Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD)* (pp. 71 - 80). ACM.

[22] Falcao, D., & Henriques, H. (2001). Load estimation in radial distribution systems using neural networks and fuzzy set techniques. *Proceedings of IEEE Power Engineering Society Summer Meeting*, (pp. 1002-1006).

[23] Figueiredo, V., Rodrigues, F., Pinto, R., & Vale, Z. A Data Mining Decision Support Tool to Study Electricity Retail Contracts.

[24] Figueiredo, V., Rodrigues, F., Vale, Z., & Gouveia, J. (2005). An electric energy consumer characterization framework based on data mining techniques. *IEEE Transactions on Power Systems volume: 20, issue: 2*, (pp. 596-602).

[25] Gama, J., & Rodrigues, P. P. (2007). Data Stream Processing. In J. Gama, & M. M. Gaber (Eds.), *Learning from Data Streams: Processing Techniques in Sensor Networks* (pp. 25-38). Springer.

[26] Giannella, C., Han, J., Pei, J., Yan, X., & Yu, P. S. (2003). Mining Frequent Patterns in Data Streams at Multiple Time Granularities. *Workshop on Next Generation Data Mining* (pp. 191-212). MIT Press.

- [27] Gibbons, P. B., & Tirthapura, S. (2002). Distributed streams algorithms for sliding windows. *Proceedings of the fourteenth annual ACM Symposium on Parallel Algorithms and Architectures (SPAA)* (pp. 63-72). ACM.
- [28] Gilbert, A. C., Guha, S., Indyk, P., Kotidis, Y., Muthukrishnan, S., & Strauss, M. (2002). Fast, small-space algorithms for approximate histogram maintenance. *Proceedings of the thirty-fourth annual ACM symposium on Theory of Computing (STOC)* (pp. 389 - 398). ACM.
- [29] Gilbert, A. C., Kotidis, Y., Muthukrishnan, S., & Strauss, M. (2001). Surfing Wavelets on Streams: One-Pass Summaries for Approximate Aggregate Queries. *Proceedings of the 27th International Conference on Very Large Data Bases (VLDB)*, (pp. 79-88).
- [30] Guha, S., & Koudas, N. (2002). Approximating a Data Stream for Querying and Estimation: Algorithms and Performance Evaluation. *Proceedings of the 18th International Conference on Data Engineering (ICDE)*, (pp. 567-576).
- [31] Guha, S., Koudas, N., & Shim, K. (2006). Approximation and streaming algorithms for histogram construction problems. In *ACM Transactions on Database Systems (TODS)* (Vol. 31, pp. 396-438). ACM Press.
- [32] Guha, S., Koudas, N., & Shim, K. (2001). Data Streams and Histograms. *Proceedings of the thirty-third annual ACM Symposium on Theory of Computing (STOC)* (pp. 471 - 475). ACM.
- [33] Hulten, G., Spencer, L., & Domingos, P. (2001). Mining time-changing data streams. *Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD)*, (pp. 97-106).
- [34] Jardini, J., Tahan, C., Gouvea, M., Ahn, S., & Figueiredo, F. (2000). Daily load profiles for residential, commercial and industrial low voltage consumers. *IEEE Transactions on Power Delivery* volume: 15, issue: 1, (pp. 375-380).

- [35] Kitayama, M., Matsubara, R., & Izui, Y. (2002). Application of Data Mining to Customer Profile Analysis in the Power Electric Industry. *IEEE Power Engineering Society Winter Meeting vol.1*, (pp. 632-634).
- [36] Matos, M., Fidalgo, J., & Ribeiro, L. (2005). Deriving LV Load Diagrams for Market Purposes using Commercial Information. *Proceedings of the 13th International Conference on Intelligent Systems Application to Power Systems*, (pp. 105-110).
- [37] Nizar, A. H., Dong, Z. Y., & Zhao, J. H. (2006). Load Profiling and Data Mining Techniques in Electricity Deregulated Market. *IEEE Power Engineering Society General Meeting*, (pp. 287-293).
- [38] Pitt, B., & Kitschen, D. (1999). Application of Data Mining Techniques to Load Profiling. *Proceedings of the 21st IEEE International Conference on Power Industry Computer Applications (PICA)*, (pp. 131-136).
- [39] Rodrigues, F., Duarte, J., Figueiredo, V., Vale, Z., & Cordeiro, M. (2003). A Comparative Analysis of Clustering Algorithms Applied to Load Profiling. *Proceedings of the 3rd International Conference on Machine Learning and Data Mining in Pattern Recognition*, (pp. 73-85).
- [40] Rousseeuw, J. P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20 , 53-65.
- [41] Thaper, N., Indyk, P., Guha, S., & Koudas, N. (2002). Dynamic multidimensional histograms. *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data* (pp. 428-439). Association for Computing Machinery (ACM).
- [42] Zhu, Y., & Shasha, D. (2002). StatStream: Statistical Monitoring of Thousands of Data Streams in Real Time. *Proceedings of 28th International Conference on Very Large Data Bases (VLDB)* (pp. 358-369). Morgan Kaufmann.