



Universidade do Porto

Faculdade de Engenharia

FEUP

Caracterização da Contribuição dos Consumidores BT para as Perdas na Rede

Marco Alexandre Martins Marques

*Dissertação submetida para satisfação parcial dos requisitos
do programa do curso de
Mestrado em Gestão de Informação*

Dissertação realizada sob a supervisão de
José Nuno Moura Marques Fidalgo
Professor Auxiliar

Departamento de Engenharia Electrotécnica e de Computadores
Faculdade de Engenharia da Universidade do Porto

Porto, Agosto 2005

Abstract

The new electricity market rises a new set of concerns regarding Information Management related with electricity consumers namely the need to acquire and analyse data allowing an objective base regarding economic tariffs and network use.

With that idea real load diagrams were gathered in several campaigns allowing the implementation of a classification process in order to differentiate automatically the consumers according to their own load profile.

In a first step the logical outliers are removed followed by clustering technics in order to gather consumers with the same pattern behaviour giving a set of generic profiles.

Later on several models are built and used to define the consumer participation regarding network losses in order to find reference values to the possibility of a direct taxation of power and or energy. Several factors are computed based on the real data gathered such loss factor, load factor, day energy allowing the definition of the consumer classes.

The correspondence study between those factors and effective network losses may allow to generalize several conclusions. That study was made both using classical Linear regression and Neural Networks.

Keywords: Load Profiles, Neural Network, Loss factors, Load factors

Resumo

A abertura do mercado de electricidade suscita um conjunto de novas preocupações relacionadas com a gestão da informação sobre os consumidores, nomeadamente a necessidade de adquirir e tratar dados que permita construir uma base de conhecimento consistente, que possibilite a definição de tarifas relativas à utilização da rede.

Para esse efeito, são analisados os diagramas reais obtidos em campanhas de medição para implementar um processo de classificação dos diversos tipos de consumidores, segundo a evolução temporal do seu diagrama de carga.

Numa primeira etapa são filtrados os *outliers* após o que se segue a aplicação de técnicas de *clustering* de modo a visualizar mais facilmente grupos de consumidores de comportamento semelhante.

Posteriormente é feita a construção de modelos capazes de caracterizar a contribuição dos consumidores para as perdas nas redes, com o intuito de encontrar valores de referência para a valorização tarifária de energia e potência. Assim, são calculados, a partir de dados reais, diversos factores tais como: factor de perdas, factor de carga, entre outros que permitem caracterizar as diversas classes de consumidores. Pretende-se estimar o valor das perdas a atribuir a cada consumidor com base nas medidas de energia e em factores caracterizadores de diagramas. A ideia base consiste em utilizar estes factores, assumindo-se que se encontram determinados para cada tipo de consumidor, e a partir daí, analisar a viabilidade da repartição das perdas. Assim tenta-se analisar as diversas relações entre as perdas atribuídas a cada consumidor e tais factores. Esta análise será efectuada por regressão linear e com redes neuronais.

O estudo da correspondência entre estas variáveis e as perdas efectivas na rede onde se inserem, permitirá generalizar as inter-relações podendo concluir que possível estimar parcialmente as contribuições para as perdas com base nessas mesmas variáveis.

Palavras chave: Perfis de Consumo, Rede Neuronal, Factores de Perda, Factores de Carga

Agradecimentos

Expresso deste modo o meu agradecimento a todos os que tornaram possível o desenvolvimento e conclusão desta tese de mestrado.

Agradeço ao meu orientador, professor José Nuno Fidalgo da Faculdade de Engenharia da Universidade do Porto, pelo seu apoio e conselhos no decurso dos trabalhos.

Aos meus colegas da Unidade de Sistemas de Energia do INESC Porto expresso a minha gratidão pelo apoio, disponibilidade e compreensão manifestada.

A todos os meus amigos pelo apoio prestado na elaboração deste documento.

Conteúdo

Índice de Tabelas	9
Índice de Figuras	10
1 Introdução	12
1.1 Objectivo	12
1.2 Âmbito	12
1.3 Estrutura da Tese	13
2 Caracterização de Consumidores	14
2.1 Introdução	14
2.2 Estratégias e Ferramentas dos Mercados	15
2.2.1 Elasticidade dos Consumidores	17
2.2.2 Aumento da Elasticidade pelos Comercializadores	18
2.2.3 Novas formas de relacionamento comercial	20
2.3 Perfis de Consumo	21
2.4 Campanha de Medição	24
2.5 Tratamento dos Dados	25
2.6 Redução de Dimensionalidade	26
2.7 Normalização	26
2.8 Selecção de Atributos	28
2.8.1 Atributos relacionados com o Diagrama de Carga	28
2.8.2 Atributos relacionados com os Dados Comerciais	29
2.8.3 Outros Atributos	31
2.9 Factores Caracterizadores de Diagramas	31
2.10 Factor de Carga e Factor de Perdas	32

2.10.1	Factor de Carga	32
2.10.2	Factor de Perdas	33
2.10.3	Determinação do Factor de Perdas	35
2.11	Modelos para a caracterização de perfis	36
2.11.1	Abordagem Estática	37
2.11.2	Abordagem Dinâmica	37
2.11.3	Modelo por análise de <i>clusters</i> , <i>Load Profiling</i>	38
2.12	Índices de Avaliação dos Resultados	40
3	Data Mining	42
3.1	DM e Estatística	43
3.2	DM e Técnicas de Apoio à Decisão	44
3.3	Principais Operações de DM	45
3.3.1	Associação e Análise Sequencial	45
3.3.2	Classificação	45
3.3.3	Clustering	46
3.3.4	Visualização	47
3.4	Componentes Primários dos Algoritmos de Data Mining	47
3.5	Técnicas de Data Mining	47
3.5.1	Regressão	48
3.5.1.1	Formalização dos Problemas de Regressão	49
3.5.1.2	Métodos Paramétricos Globais	50
3.5.1.3	Modelos Não Paramétricos	51
3.5.1.4	Modelos Aditivos	52
3.5.2	Algoritmos de Vizinhança mais Próxima	52
3.5.3	Redes Neurais	53
3.5.3.1	Single Layer Perceptron	54
3.5.3.2	Multi Layer Perceptron	55
3.5.3.3	Aplicação de Redes Neurais	55
3.5.4	Partições Hierárquicas e Árvores de Decisão	56
3.5.5	Algoritmos Genéticos	58
3.5.6	Fuzzy Sets	59
3.5.7	Rough Sets	61
3.6	Qual a Técnica a Aplicar ?	61

4	Clustering	63
4.1	Fases do Processo	64
4.1.1	Extracção de Atributos	64
4.1.2	Avaliação de Similaridade entre registos	64
4.1.3	Avaliação dos resultados	65
4.2	Classificação de Clustering	66
4.2.1	Algoritmos Hierárquicos	66
4.2.1.1	AGNES, <i>Agglomerative NESting</i>	69
4.2.1.2	DIANA	70
4.2.1.3	MONA	71
4.2.2	Algoritmos de Partição	72
4.2.2.1	K-Means	73
4.2.2.2	PAM <i>Partitioning Around Medoids</i>	74
4.2.2.3	CLARA	75
4.2.2.4	FANNY, <i>Fuzzy Analysis Clustering</i>	77
4.2.3	Outros Algoritmos	78
4.3	Aplicação de Clustering	79
5	Perdas	81
5.1	Perdas na Rede	81
5.1.1	Tipos de Perdas	82
5.1.2	Localização	82
5.1.3	Componentes	82
5.2	Cálculo das Perdas Técnicas	82
5.2.1	A complexidade da estimativa de perdas técnicas	84
5.2.2	Determinação de perdas por período horário	84
5.2.3	Métodos de Alocação de Perdas	85
5.2.3.1	Alocação por Nível de Tensão	86
5.2.3.2	Postage Stamp	86
5.2.3.3	Modified Postage Stamp	88
5.2.3.4	MW.Km	89
5.2.3.5	MW^2 .km	91
5.2.3.6	Participações Médias	92
5.2.3.7	Participações Médias Modificadas	93
5.2.3.8	Outros Métodos	94

6	Repartição Automática de Perdas	95
6.1	Descrição dos Dados	95
6.1.1	Limpeza dos Dados	97
6.1.2	Pré-Processamento dos Dados	98
6.2	Construção do conjunto de padrões	99
6.3	Modelos Testados	99
6.3.1	Modelo de Regressão Linear	100
6.3.2	Rede Neuronal SLP	100
6.3.3	Rede Neuronal MLP	102
6.4	Conclusões	103
7	Conclusão	105
7.1	Futuros Desenvolvimentos	106
	Referências	110

Lista de Tabelas

2.1	Factores de ajustamento em horas para perdas em Portugal	21
2.2	Opções de Tarifa [BT]	31
6.1	Amostra da organização dos dados de entrada	97
6.2	Erros Médios dos métodos abordados	103

Lista de Figuras

2.1	Mercado de Energia	18
2.2	Perfil diário calculado a partir de DM	40
3.1	Single Layer Perceptron	54
3.2	Multi Layer Perceptron MLP	55
3.3	Representação de uma Árvore de Decisão	57
3.4	Representação Fuzzy: com pontos em vários clusters	60
3.5	Representação Fuzzy de um ponto	60
4.1	Diferença dos processos hierárquicos	67
4.2	Dendrograma exemplo de dados [MT] do projecto CCR	68
4.3	AGNES Banner	70
4.4	Representação de consumidores através de <i>PAM</i>	75
4.5	Exemplo do agrupamento de alguns consumidores através do algoritmo <i>CLARA</i>	76
4.6	Amostra de clusters do tipo de consumidores [MT]	77
4.7	Algoritmo FANNY	78
4.8	Exemplo de classes de consumidores [MT] obtidos com algoritmo de clus- tering <i>CLARA</i>	79
6.1	Esquema da Rede seleccionada	96
6.2	Sequência de Operações	98
6.3	Processo com Modelo de Regressão	101
6.4	Topologia de uma Rede Neuronal 4–13–1	102
6.5	Aproximação com rede Neuronal SLP	102
6.6	Esquema da topologia da rede MLP utilizada	103

Lista de abreviaturas

AT -Alta Tensão
BPN - BackPropagation Network
BT -Baixa Tensão
BTE -Baixa Tensão Especial
CSV - Comma Separated Values
DM - Data Mining
DSM - Demand Side Management
Fc -Factor de Carga
Fp - Factor de Perdas
GA - Genetic Algorithm
GUI - Graphical User Interface
GIS - Geographical Information System
KNN - K Nearest Neighbour
LL - Log Likelihood
LM - Linear Model
LSM - Load Shape Models
LVQ - Learning Vector Quantization
MAT - Muito Alta Tensão
MARS - Multivariate Adaptive Regression Splines
MLE - Maximum Likelihood Estimates
MLP - Multi Layer Perceptron
MSE - Mean Square Error
MT -Média Tensão
NNC - Nearest Neighbor Classifier
OLAP - On-Line Analytical Processing
PCT - Perfil Consumo Típico
RN - Redes Neurais
SDM - Spatial Data Mining
SOM - Self Organizing Map
SQL -Structured Query Language
STLF - Short Term Load Forecasting
SLP - Single Layer Perceptron
TLP - Typical Load Profiles

Capítulo 1

Introdução

1.1 Objectivo

O objectivo principal deste trabalho é propor uma metodologia para a caracterização e contribuição dos consumidores para as perdas de energia eléctrica em redes de Baixa Tensão (BT).

Esta metodologia será baseada na obtenção de estimativas das perdas atribuíveis aos consumidores, a partir da informação sobre a energia consumida e de outros parâmetros complementares.

A motivação para o desenvolvimento deste trabalho é a estimativa e repartição das perdas de energia utilizando a informação relacionada com os diagramas de carga sabendo-se que a forma destes diagramas tem influência directa nas perdas. No entanto, estes nem sempre estão disponíveis imediatamente. Contudo, os factores caracterizadores dos diagramas (factor de carga, factor de perdas e outros) podem ser determinados e adoptados para as diversas classes de consumidores e, desse modo, obter a informação necessária sem recorrer directamente aos diagramas reais.

Um outro objectivo deste trabalho consiste na síntese de técnicas de *Data Mining*, e em particular de *clustering* dado a sua importância no processos de tratamentos de dados necessários á caracterização de consumidores.

1.2 Âmbito

Este trabalho foi desenvolvido no âmbito do Projecto CCR [1] - Caracterização de Consumidores e Redes no âmbito de um contrato com a EDP (Energias de Portugal) motivado pela abertura do Mercado Ibérico de electricidade. O conhecimento sobre o comportamento dos consumidores pode contribuir para uma melhor estruturação das tarifas, uma gestão optimizada das cargas e uma maior integração com a distribuição. Face à emergente abertura

do Mercado Ibérico a *EDP* decidiu avançar com o projecto visando estes objectivos:

1. Caracterização dos consumidores incluindo estimativa de factores de carga;
2. Estimativa de perdas;
3. Caracterização de procura.

A caracterização dos consumidores pressupôs duas tarefas:

- Obter diagramas de carga médios dos diversos tipos de consumidores e respectivos factores caracterizadores;
- Perfis Típicos para operação em mercado.

A estimativa de perdas procura obter perdas típicas para as redes e alocação das perdas para cada tipo de consumidor. Assim, no seguimento do segundo ponto, neste trabalho, procura-se responder ao problema da estimativa de repartição de perdas, pelos diversos consumidores ligados a uma determinada rede. O terceiro ponto pretende obter diagramas de consumidores a partir de inquéritos e relacioná-los com os diagramas obtidos por medição directa (diagramadores).

1.3 Estrutura da Tese

O capítulo inicial apresenta uma pequena introdução à matéria a estudar, mostrando os objectivos principais. No capítulo 2 é descrito o processo de caracterização dos consumidores e explicada a utilização dos perfis de consumo, os quais permitem determinar os factores de carga e os factores de perda para os diversos tipos de consumidores [BT], posteriormente, utilizados para o cálculo das perdas. No capítulo 3 é descrita a utilização de algumas técnicas e operações de *Data Mining*. No capítulo seguinte são referidas as técnicas de *clustering* que foram utilizadas de modo a agrupar e validar os dados aos perfis de consumo existentes. O capítulo 5, relativo às perdas, evidencia os cálculos e os métodos de alocação de perdas. No capítulo 6 é descrita a aplicação prática, no caso em estudo. Por último, é apresentada a Bibliografia consultada.

Capítulo 2

Caracterização de Consumidores

Neste capítulo introduz-se o conceito de perfil de consumo sendo justificada a sua importância nos sistemas de energia actuais. Este conceito provém da necessidade de resposta prática dos mercados de energia para a simplificação dos processos de tarifação face à volatilidade dos preços temporais da energia em mercado aberto e pelo facto dos consumidores [BT] não disporem de contadores horários. Em cada uma das fases, são apresentadas diferentes técnicas e modelos. Ao longo deste capítulo serão abordados os diferentes elementos necessários para a caracterização de consumidores, desde a necessidade da campanha de medição até ao cálculo dos factores de carga e de perdas.

2.1 Introdução

Os consumidores de energia eléctrica, de tipos e características heterogéneas, apresentam diagramas de carga diários tipicamente diferentes, os quais possuem múltiplas formas que dependem dos hábitos de consumo, dos dias da semana, das condições climáticas e da época do ano. O conhecimento detalhado do comportamento dos consumidores permite às empresas de distribuição uma maior eficiência no planeamento e operação das suas redes, constituindo além disso uma vantagem competitiva em termos de mercado, para as empresas de comercialização de energia eléctrica.

Habitualmente, a classificação dos consumidores é feita *à priori*, pelas empresas de distribuição, baseada nas características comerciais destes, tais como, o nível de tensão a que se encontra ligado, a potência contratada, entre outras.

Existe uma grande diversidade de metodologias para a classificação de consumidores e inferência de curvas de consumo. Os critérios de classificação são diversos, dependendo da informação comercial e estrutura tarifária disponível e, posteriormente, da estrutura e desenvolvimento do mercado local de energia eléctrica. Consequentemente, o número de conjuntos ou classes varia muito conforme o local, a região ou o País. Esta classificação não permite, por si, obter um conhecimento profundo do comportamento dos consumidores,

pelo que é necessário conhecer o diagrama de cargas diário para os diferentes dias da semana e épocas do ano, e para os diversos tipos de consumidores.

2.2 Estratégias e Ferramentas dos Mercados

Com o Mercado Ibérico emergente e quase a terminar o período de transição torna-se crucial definir regras e seria conveniente que estas se baseassem em estudos relativos aos consumidores.

Durante vários anos o sector eléctrico foi caracterizado pela existência de um regime monopolista, verticalmente integrado, no qual a produção, o transporte e a distribuição de energia eram actividades desenvolvidas apenas por uma única empresa.

A utilização dessa estrutura estava associada com a existência de economias de escala, a facilidade de despacho dos sistemas de produção, a segurança e a garantia de fornecimento de energia a todos os consumidores. Os custos da energia, suportados pelos consumidores, são definidos nas tarifas que descrevem o custo médio da sua produção, transporte e distribuição. Com a separação entre a distribuição e a comercialização de energia eléctrica as várias empresas comercializadoras passam a competir mais intensamente nos mercados de retalho. Para que as empresas sejam bem sucedidas necessitam de diversas estratégias e ferramentas que permitam, por um lado, reduzir o risco inerente ao funcionamento de mercado, e por outro, conseguir obter e aumentar a quota de mercado.

De referir ainda que a compra da energia eléctrica baseada num preço fixo não existe forma de modificar o comportamento dos consumidores.

A alteração da dinâmica do funcionamento dos mercados retalhistas (para o consumidor final), torna aqueles mais competitivos e permite aos consumidores uma maior participação no mercado. Para tal, é necessário criar um conjunto de estruturas e de ferramentas de apoio ao funcionamento do mercado, quer para os consumidores, quer para as empresas comercializadoras.

O primeiro desafio encontrado nos mercados de retalho é o aumento da elasticidade dos consumidores face à volatilidade do preço de energia eléctrica. O aumento da elasticidade dá lugar, normalmente, a um conjunto de benefícios ao funcionamento dos mercados competitivos, em geral, contribuindo para um maior controlo e adaptação de preços. Relativamente ao consumidor, esse aumento contribui para diminuir os gastos totais com a energia eléctrica e para uma utilização mais eficiente dos recursos.

O segundo desafio é a competição das empresas retalhistas pelo conjunto de consumidores pertencentes ao próprio mercado. Devido ao referido anteriormente é importante conhecer essas estratégias de diferenciação que visam o conhecimento do comportamento e agrupamento dos diferentes tipos de consumidores. Posteriormente, conduzirá à realização de diversos estudos que procuram criar e caracterizar os diferentes perfis de consumidores.

As actividades de transporte *REN* e de distribuição *EDP* ainda permanecem em

monopólio, devidamente regulado, uma vez que as economias de escala existentes tornam estas actividades mais eficientes, quando desenvolvidas em bloco.

Como resultado da reestruturação, a responsabilidade global de servir as necessidades dos consumidores individuais deixará de pertencer a uma única empresa.

A concorrência nas actividades de produção e de comercialização de energia eléctrica faculta o aparecimento de vários produtores e comercializadores, permitindo aos consumidores escolher o seu fornecedor. Os comercializadores, utilizando os serviços de rede dos distribuidores (e dos transportadores), competem entre si pela captação de mais clientes, procurando aumentar as respectivas quotas de mercado. A segmentação do mercado de fornecimento de electricidade tenderá a ocorrer, com os comercializadores a definirem tarifas diferenciadas e adaptadas às diferentes necessidades dos seus potenciais consumidores. Este facto representa uma inversão na lógica tradicional do negócio da energia eléctrica, deixando este de ser desenvolvido de acordo com as regras ditadas pela empresa monopolista, para passar a ser gerido de acordo com os requisitos dos consumidores (por exemplo, dependentes das suas especificidades de consumo).

Para diminuir o risco, os retalhistas procuram obter contratos de tarifa única a longo prazo com as empresas produtoras. Por outro lado, tentam criar contratos de venda com preço único de modo a poderem partilhar o risco com os compradores. Idealmente, o retalhista tenta acompanhar a variação do preço e a quantidade de electricidade vendida, com a variação do preço e a quantidade de electricidade comprada. No entanto, a natureza aleatória da carga dificulta esta coordenação, pelo que o retalhista é forçado a comprar ou a vender o excedente no mercado. Devido à volatilidade do preço no mercado *SPOT*, a gestão dos desequilíbrios entre a oferta e a procura representa um factor de risco para o retalhista. Assim sendo, para reduzir este factor de risco, o retalhista deve possuir ferramentas que lhe permitam fazer uma previsão da carga, com o maior grau de precisão possível. Contudo, o problema de previsão da carga, em ambiente de mercado é mais complexo do que o problema de previsão convencional.

Num mercado retalhista competitivo, poderá não ser a mesma empresa a fornecer todos os consumidores numa determinada região nem mesmo um determinado posto de transformação. Cada empresa fornece energia eléctrica a um determinado número de consumidores ligados de forma aleatória à rede de distribuição. Neste caso, uma alternativa seria a previsão poder ser realizada de *baixo para cima*, depois da agregação das previsões individuais de cada consumidor ou grupo de consumidores pertencentes à quota de mercado de uma empresa. Esta previsão seria dispensável no caso da existência de um esquema de contagem em tempo real, o qual implicaria o abandono da existência de perfis.

No entanto, a instalação destes tipos de contagem não é economicamente viável em pequenos consumidores como os consumidores domésticos, os quais representam em Portugal cerca de 40 % de consumo total. Deste modo, é impossível aos retalhistas prever com rigor e de um modo directo a procura de cada um dos seus clientes. Para conseguir esta previsão, os consumidores devem ser divididos num conjunto de classes representadas

pelos respectivos perfis de consumo, os quais serão previamente determinados pela entidade reguladora do mercado.

O retalhista deve identificar esses grupos da forma mais precisa possível, determinar os respectivos perfis de consumo e a sua variância nas diferentes situações de carga. Para além do factor de risco, os retalhistas têm ainda que lidar com a concorrência. O retalhista competitivo vai permitir aos consumidores a escolha e a troca de fornecedor de energia eléctrica, sempre que este apresente uma vantagem. Os retalhistas têm que recorrer a diversas alternativas ou estratégias baseadas em apenas dados temporais para conseguir a diferenciação.

Como objectivo principal surge a maximização do lucro, pelo que a relação com os consumidores, a resposta dos mercados, a criação de novos produtos e serviços são os novos desafios que estas empresas têm que enfrentar.

As empresas que escolham uma estratégia de diferenciação para conquistar o mercado devem ter uma relação próxima com os seus consumidores, ao procurar conhecer as suas necessidades. Estas desenvolvem processos de inovação que permitem criar produtos direccionados a cada um dos seus segmentos. Neste contexto de transição de mercados entra o estudo dos perfis de consumidor, de modo a conseguir prever as diversas alterações na rede e no mercado.

2.2.1 Elasticidade dos Consumidores

No período anterior à reestruturação do sector eléctrico, os preços da energia eléctrica eram definidos através de um conjunto de tarifas globais, as quais incluíam todos os custos associados à produção, transporte e distribuição. Geralmente, as tarifas eram estabelecidas no ano $t-1$, para o ano t e demarcavam uma discriminação apenas temporal nos preços da energia (horas de vazio, cheias e ponta).

A reestruturação do sector eléctrico e o emergir dos comercializadores de energia eléctrica ainda não alterou significativamente esta realidade, pelo menos para os consumidores de menor dimensão (em especial os BT (domésticos e comerciais)). O facto dos consumidores não sentirem, em cada momento, os custos reais da electricidade faz com que apresentem um comportamento não elástico.

Assim, para que os consumidores utilizem a electricidade de acordo com as variações no seu custo é necessário que tenham acesso à informação sobre os custos reais da electricidade em intervalos de tempo curtos (por exemplo uma hora ou mesmo meia hora). Contudo, os consumidores apenas obtêm vantagem se os contadores forem horários. Além do fornecimento da informação sobre os custos da energia em intervalos relativamente curtos, os consumidores devem fazer parte de uma estrutura que lhes permita responder de forma rápida e simples a tais alterações. (ver [2])

2.2.2 Aumento da Elasticidade pelos Comercializadores

O papel a desempenhar pelos comercializadores de electricidade é muito importante na definição das formas de transmissão para os seus clientes dos custos reais (ou quase) da electricidade em cada momento. Os comercializadores de energia devem celebrar contratos com os consumidores que lhes permitam controlar a carga destes, quando tal for necessário. Este mecanismo fornece ao retalhista alguma flexibilidade para gerir o risco associado aos preços. Assim, uma carga agregada mais elástica pode resultar em preços de mercado menores para a electricidade, beneficiando todos os participantes. Cada retalhista agrega várias cargas individuais, as quais são constituídas por consumidores domésticos, comerciais e industriais: $q(x_{in}, p_i)$, (no qual $q(x_i)$ é a quantidade de energia e p_i é o preço da energia) e uma curva da carga agregada, $QD(p)$ é obtida para cada retalhista com base nos valores das cargas individuais dos seus consumidores e nos termos dos contratos estabelecidos com estes. Estas curvas agregadas são utilizadas pelos comercializadores para efectuarem as suas ofertas de compra de energia no mercado grossista.

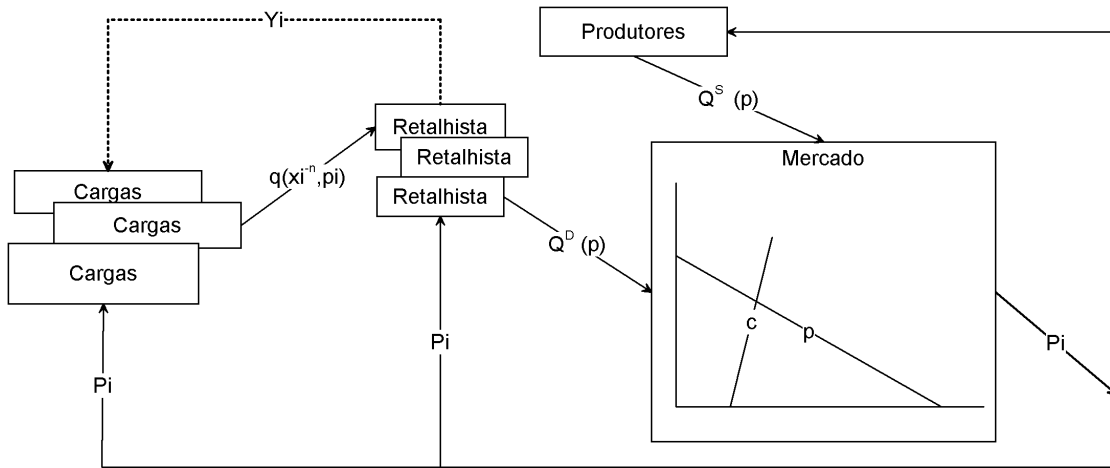


Figura 2.1: Mercado de Energia

No mercado *OMEL*, a determinação do preço e das quantidades a transaccionar resulta da intersecção da curva da Oferta $Q^S(p)$ com a curva da carga total C . Se o preço for elevado, os comercializadores e os consumidores finais têm incentivos para reduzir a carga e consequentemente os custos. A redução de carga pode ser conseguida, de forma directa, pelos clientes da parte retalhista, reduzindo o pedido de energia $q(x_i)$, ou de forma indirecta, através de *sinais de controlo* enviados pelo retalhista às cargas dos seus clientes (y_i), controlando os equipamentos ligados.

O grau de controlo da carga e a redução ou deslocação do consumo de energia depende dos tipos de contrato estabelecidos entre os comercializadores e os seus clientes. As tarifas podem incorporar esquemas de tarifas verdes, de tarifas diferenciadas de acordo com o

período do dia, dia da semana, mês, estação do ano ou mesmo tarifas em tempo real. No caso das tarifas verdes, o consumidor opta pela compra de energia proveniente de fontes renováveis (na totalidade ou em parte) e, eventualmente, suporta um custo adicional [3]. Nas tarifas diferenciadas, os comercializadores efectuam uma segmentação do mercado de fornecimento de energia eléctrica e desenvolvem tarifas, adaptadas aos perfis de consumo de alguns dos seus clientes.

A segmentação de mercado leva a uma mudança na filosofia das tarifas, mudando o centro das atenções do vendedor para o consumidor, deixando as tarifas de ser fixas. Em Portugal, a separação das tarifas é efectuada com recurso à definição de horas de vazio e de horas fora do vazio, para os consumidores domésticos com tarifa bi-horária e, no caso de grandes consumidores, em horas de super vazio, vazio, cheias e de ponta.

Estas tarifas apresentam uma estrutura estática que não permite ajustar, em tempo real, todas as variações nos preços da electricidade pelo que é que possível proceder ao cálculo das variações sazonais, mensais e diárias nos custos de produção, transmissão e nos preços do mercado da electricidade. A divergência entre os preços em tempo real e as tarifas globalmente estáticas é significativa, dando lugar a um funcionamento não óptimo do mercado. Este facto torna importante a definição de tarifas com características parcialmente dinâmicas que permitam aos consumidores dispor de uma maior percepção sobre os custos reais suportados em cada momento, o que aumenta a sua elasticidade da procura de energia.

O estabelecimento de sistemas de tarifas em tempo real é o culminar do processo de segmentação do mercado de comercialização de electricidade e, principalmente, do processo do aumento da elasticidade dos consumidores. Em Ernoult [4] é apresentado um trabalho de determinação de índices para representação de diagramas de carga. Este conjunto de índices permite caracterizar e tenta explicar diversos aspectos do comportamento em termos de elasticidade dos consumidores numa base diária e semanal.

Neste caso, cada consumidor é representado por um vector de atributos, constituído por um grupo de índices, com vista a facilitar a proposta de novos produtos como:

- *Produtos Simples*: onde os consumidores podem comprar electricidade, produtos, serviços e informações, ao proporcionar um contacto mais simples com o fornecedor de electricidade.
- *Produtos Especializados*: no caso em que existam segmentos de mercado que podem ter preferência por produtos especializados estando dispostos a pagar mais por eles, por exemplo: garantia de que a energia eléctrica que se consome é produzida através de energias renováveis, apesar de ser mais cara, pode atrair consumidores pertencentes a segmentos de mercado e melhorar a imagem da empresa junto do publico em geral.

As estratégias e as ferramentas para o bom funcionamento e o sucesso das empresas de retalho devem ser pensadas e construídas com base num conhecimento mais profundo

dos consumidores. Assim, o estudo da carga, *à posteriori*, procura fornecer conhecimento baseado nos dados obtidos, assumindo um papel relevante no presente e no futuro. O conhecimento de como e quando os consumidores utilizam a electricidade é importante para as empresas comercializadoras nos mercados retalhistas. No entanto, este conhecimento evolui de forma dinâmica, pelo que as ferramentas utilizadas pelas empresas devem ser susceptíveis de actualização.

2.2.3 Novas formas de relacionamento comercial

A implementação de tarifas diferenciadas e, em especial das tarifas em tempo real, implica a necessidade de instalação de estruturas de automação e de comunicação que permitam informar os consumidores sobre os preços da electricidade, em cada momento, e receber o feedback respectivo.

A transmissão aos consumidores de informação sobre os preços da electricidade em cada momento pode ser conseguida de diversas formas, mediante o recurso à emissão de sinais de rádio, utilização de redes de telecomunicações já existentes ou mesmo tirando partido dos próprios condutores da rede de distribuição. Os consumidores devem possuir equipamento capaz de receber os sinais referentes a essa informação e em função disso adaptar o seu perfil de consumo. Aqueles agentes económicos dispõem de sistemas de produção e ou armazenamento de electricidade podem, em função do preço da electricidade em cada momento (fornecido pelos comercializadores), decidir se pretendem comprar, produzir ou utilizar a electricidade armazenada para suprirem as suas necessidades.

O facto dos custos de acesso às redes (de transporte e de distribuição), dos custos de compra de energia no mercado grossista e dos custos das perdas poderem variar temporalmente (ao longo do dia, dia da semana, mês e estação do ano, entre outras), permite aos comercializadores de electricidade criar tarifas diferenciadas no tempo. Os distribuidores e os transportadores de electricidade podem, se os reguladores do sector o permitirem, aplicar taxas de acesso às redes, diferenciadas temporalmente, dado que os custos de rede variam em função dos níveis de carga. A definição de tarifas de utilização de rede diferenciadas temporalmente (tarifas nas horas fora de vazio superiores às aplicadas nas horas de vazio) permite que os operadores das redes forneçam informação económica capaz de incentivar uma melhor utilização da capacidade das redes e, eventualmente, diferir investimentos para reforço desta capacidade (determinados pela ponta de carga que o sistema tem de satisfazer num determinado momento).

No caso dos custos da energia no mercado grossista, a variação resulta essencialmente da aplicação da lei da oferta e da procura, uma vez que, nas horas de menor consumo alguns produtores estão dispostos a vender a sua energia a preços inferiores aos praticados nas horas de ponta. Por outro lado, nas horas de pouca procura, as centrais com maiores custos marginais de produção de electricidade não são despachadas, resultando daqui um preço de mercado mais reduzido que nas horas de ponta. O facto da carga das redes eléctricas variar de acordo com a hora do dia, dias da semana, meses e estações do ano faz

Tabela 2.1: Factores de ajustamento em horas para perdas em Portugal

	Ponta	Cheia	Vazio Normal	Super Vazio
γ^h MAT	1, 8	1, 7	2, 3	2, 3
γ^h AT/RNT	2, 1	2, 0	2, 6	2, 6
γ^h AT	2, 2	1, 5	1, 0	1, 0
γ^h MT	5, 4	2, 5	2, 2	2, 2
γ^h BT	9, 3	6, 87	4, 3	4, 3

com que as perdas tenham também variações.

O facto dos custos de acesso às redes, das perdas e da própria energia no mercado grossista poderem variar temporalmente permite, aos comercializadores, a definição de tarifas de energia com preços que variam de acordo com a hora do dia, dia da semana, mês e estação do ano, entre outras.

Por exemplo, em Portugal, a regulamentação do sector eléctrico evidencia as variações nas perdas através dos designados *factores de ajustamento para perdas*.

A Tabela 2.1, publicada por Despacho da *ERSE* N.º 14 / 2003 [5], apresenta os valores dos factores de ajustamento referidos.

Este facto, associado à concorrência entre comercializadores de energia (que procuram aumentar a sua quota de mercado), define potencialmente a criação de tarifas diferenciadas para diferentes tipos de consumidores. Qualquer tipo de consumidor, incluindo os consumidores domésticos pode usufruir das tarifas especiais. Como foi anteriormente referido, o culminar da definição de tarifas especiais é a concepção de tarifas em tempo real, às quais estarão associadas maiores elasticidades.

O estabelecimento de tarifas diferenciadas, tarifas em tempo real, contratos de controlo de carga com os clientes permite aos comercializadores obter um conjunto de vantagens, surgindo novas formas de relacionamento comercial entre consumidores, comercializadores e distribuidores.

Os distribuidores também podem definir um conjunto de tarifas diferenciadas, adaptadas a diferentes segmentos de clientes, incluindo as tarifas em tempo real e os contratos de controlo de carga. Para tal os consumidores têm de dispor de uma estrutura capaz de responder aos preços ou às alterações de carga, limitando os impactos na rede. Entre o consumidor e o distribuidor podem ser estabelecidos contratos de ligação à rede especiais, nos quais o consumidor se disponibiliza a reduzir a potência de ligação requerida em determinados períodos de tempo, em contrapartida de menores custos de ligação.

2.3 Perfis de Consumo

Esta parte do estudo apresenta o conceito de *Perfil de Consumo Típico* e descreve sinteticamente os estudos feitos por alguns autores.

Com o objectivo de representar os diferentes padrões de consumo de electricidade através de um conjunto de classes, Allera [6] define um Perfil de Consumo Típico PCT como *o padrão de consumo de electricidade, de um consumidor ou grupo de consumidores, durante um determinado período de tempo*. Esses perfis são habitualmente determinados para representar o diagrama de cargas diário, correspondente a um determinado dia, ou como representação de um dia tipo, como por exemplo, o diagrama de carga típico de um dia da semana. Deste modo, o conceito tem sido aplicado há muito pelas empresas de electricidade na formulação e aplicação de tarifas e possui especial importância em ambientes de mercado.

A determinação de perfis de consumo é importante em diversas áreas dos Sistemas de Energia, tais como:

- Previsão de carga de curto prazo *Short Term Load Forecasting* [STLF]: onde a informação sobre as características de evolução da carga dos diferentes consumidores pode contribuir para o aumento do desempenho de previsão optimizando a eficiência da produção e a própria exploração do sistema eléctrico de energia;
- Planeamento da produção de electricidade: onde o conhecimento do comportamento dos perfis de consumo é importante na previsão da evolução da carga e do seu valor máximo (ponta);
- Implementação de medidas de gestão da procura *Demand Side Management DSM* e de gestão de energia mais adequadas: onde o conhecimento do perfil de consumo permite a utilização das estratégias e medidas de DSM mais apropriadas e o aconselhamento do consumidor. Este aspecto é tanto mais importante se considerarmos a possibilidade do estabelecimento de tarifas de electricidade em tempo real, as quais indicam aos consumidores o custo real da electricidade, em cada instante.
- Estabelecimento de mercados competitivos: onde a utilização de perfis de consumo pode permitir a participação de pequenos consumidores no mercado, sem a necessidade de instalação de contagem em tempo real, o que permite às empresas distribuidoras e comercializadoras que actuam nesses mercados, partilhar o risco inerente às variações de preço com o consumidor;
- Definição de estratégias contratuais e serviços ao consumidor o conhecimento de diferentes perfis de consumidores, pode permitir a criação de estruturas tarifárias adaptadas às características de cada grupo de consumidores, constituindo uma vantagem competitiva para as empresas comercializadoras de electricidade, com benefícios económicos para os próprios consumidores.

A relação entre os diagramas de carga, factores atmosféricos, factores temporais e as características de cada consumidor são *Não lineares* e possuem diversas inter-relações, o que justifica os inúmeros trabalhos de investigação realizados. A maioria destes surge no

âmbito de estudos para previsão de cargas ou em *Load Profiling* onde se tenta inferir os efeitos dos factores atmosféricos na forma e na amplitude dos perfis base.

A partir destes dados, é construído um modelo que permite gerar um conjunto de PCT, com o respectivo nível de confiança, para diferentes dias da semana, meses e estações.

O número de classes típicas não deve ser demasiado grande, mas deve ser suficiente para representar os diferentes padrões de consumo existentes no conjunto inicial. Por outro lado cada PCT deve representar um grupo relativamente homogéneo de consumidores.

De seguida, apresenta-se uma pequena lista de outras características de *PCT*:

- cada PCT deve ser único;
- a identificação das características de um consumidor, para ser associado a um PCT deve ser fácil e simples de efectuar, de modo a ser viável a sua aplicação na população de consumidores;
- o conjunto dos PCT deve permitir a representação da totalidade dos consumidores integrados;
- a precisão necessária para a estimativa dos PCT deve ser estudada, dependendo do tipo de aplicação adoptada.

Actualmente, são utilizadas diversas abordagens para a determinação de um PCT. De acordo com a abordagem utilizada, é possível dividir este tipo de estudos numa abordagem estática e dinâmica.

A abordagem estática baseia-se em dados históricos de estudo da carga, habitualmente diferenciados pela hora, pelo dia da semana, pelo mês em que foram recolhidos. O tipo de PCT assim obtido é utilizado para representar o diagrama de carga dos consumidores e possui como principal limitação a impossibilidade de ser ajustado de acordo com as condições meteorológicas.

A abordagem dinâmica também requer a recolha, validação dos dados e a determinação dos PCT numa base temporal diária, assim utiliza dados históricos, mas permite incorporar a influência das condições meteorológicas, através de uma função de resposta às mesmas associada a um modelo. Esta função é muitas vezes gerada com base em técnicas estatísticas e considera as alterações meteorológicas numa base diária ou mesmo horária. Esta abordagem, normalmente mais rigorosa, deve-se ao conhecimento de que as alterações nas condições meteorológicas influenciam fortemente o consumo e a forma de utilização da electricidade alterando, deste modo, o perfil de consumo.

A selecção da abordagem mais adequada depende de diversos factores como o custo, os dados e o equipamento disponível, a precisão desejada e o domínio de aplicação, não sendo possível definir qual a melhor abordagem de uma forma genérica. Um projecto de caracterização de consumidores deve envolver a participação activa das empresas distribuidoras e comercializadoras, para a obtenção dos dados necessários, e o planeamento de cada uma

das fases do projecto, assim como uma estreita colaboração entre estas empresas e a equipa que desenvolve o processo de classificação.

2.4 Campanha de Medição

Para se obter os dados para o estudo da carga é necessário um processo de definição da amostra da população a estudar, a instalação de aparelhagem de medição adequada e a recolha e o tratamento dos dados. Este é o primeiro passo para a realização de um projecto de caracterização dos consumidores. Este tipo de estudos deve ser mantido durante o período necessário para a consolidação e assegurar a significância estatística da amostra. Usualmente, pretende-se que o período referido seja o maior possível (no mínimo de um ano) de modo a ser possível inferir alterações sazonais, através da instalação de diagramadores em vários consumidores pertencentes à amostra em estudo. O volume de dados obtido é normalmente bastante elevado dificultando a gestão, manutenção e tratamento da informação.

A primeira fase do estudo da carga é a definição do processo de recolha de dados, na qual devem ser estabelecidos os seguintes parâmetros:

- período total de recolha de dados (dias, meses, anos...);
- grandezas a monitorizar;
- cadência da recolha de informação (5 minutos, quarto de hora, vinte minutos, meia hora, hora completa);
- características técnicas dos aparelhos de medição (diagramadores), as quais afectam a precisão das medidas.

A escolha do período e da cadência das medições deve estar relacionada, na medida em que a dimensão máxima *tratável* do conjunto pode requerer a diminuição da cadência de medição com vista a aumentar o período de tempo total de recolha de dados. Este tipo de parâmetro é definido de acordo com os objectivos e modelo de classificação a adoptar.

Não independente da primeira, a segunda fase do estudo da carga inclui a definição da amostra de consumidores a monitorizar, de dimensão M , que é feita com base em técnicas estatísticas de amostra, de forma a garantir que a amostra seleccionada seja representativa da totalidade da população, que se pretende estudar.

A definição da amostra é seguida da instalação dos diagramadores e da recolha dos dados durante o período de tempo definido. Devem ainda ser definidos nesta fase todos agentes externos que possam influenciar as grandezas a monitorizar, tais como as condições atmosféricas, dias especiais e eventos que possam influenciar o consumo de electricidade.

O resultado das campanhas de medição é um conjunto que inclui toda a informação recolhida para cada um dos consumidores que compõem a amostra. Deste modo, existe

para cada consumidor um conjunto de leituras recolhidas de acordo com a cadência definida, ao longo do período de tempo do estudo. Estes dados correspondem aos diagramas de cargas diários para o número de dias em que se processou a recolha de informação.

2.5 Tratamento dos Dados

Quando é necessário proceder à utilização de dados reais, adquiridos e armazenados de forma automática, é necessário criar um procedimento de detecção e filtragem de dados anómalos. Este procedimento é frequentemente designado como *Limpeza de Dados* e corresponde a uma das fases do processo de *Data Mining*. Os dados anómalos podem ser originados nos sistemas de medição, no processo de transmissão, no comportamento fortuito e irregular de um consumidor, a existência de dias heterogéneos ou mesmo de falhas de serviço da rede de alimentação.

As técnicas de filtragem de dados e detecção de dados anómalos podem incluir regras heurísticas, *Fuzzy Filtering* ou mesmo *Redes Neurais* [7].

Um exemplo de regra heurística seria a utilização do valor da potência que permite definir um critério de aceitabilidade de uma medida. Esse critério é baseado na comparação média da amplitude de potência com a amplitude média das potências das medidas recolhidas ao longo do dia (ou de um intervalo de tempo menor). Se o valor da diferença exceder um valor limite, a leitura é classificada como anómala e definida como inválida.

Após a detecção dos registos com dados anómalos, deve ser decidida a eliminação ou manutenção dos correspondentes diagramas de cargas. Esta decisão é baseada no volume de dados disponível, se existir um número suficientemente elevado de diagramas de carga para cada consumidor, o que estiver afectado por dados anómalos será removido. Caso contrário, é necessário adoptar uma estratégia de substituição destes dados, baseada num modelo de regressão, construído a partir dos restantes dados ou noutro tipo de artefacto.

O tratamento feito aos dados omissos vai depender do seu volume, do tipo de dados e dos algoritmos que vão ser utilizados no modelo de caracterização dos consumidores. Se o volume de dados em falta for significativo, é importante criar uma técnica suficientemente robusta para que os novos dados façam sentido. Muitas vezes utilizam-se técnicas semelhantes às utilizadas em previsão de cargas para obter estimativas dos dados em falta.

As técnicas de substituição mais utilizadas são baseadas em abordagens estatísticas ou em *Redes Neurais*. A técnica a aplicar vai depender do tipo de dados em falta. Por exemplo, no caso de se utilizarem técnicas estatísticas, se se tratar de grandezas contínuas pode-se aplicar uma regressão linear, mas se os dados em falta forem discretos, poderá ser utilizada uma regressão logística.

Em alternativa, podem ser utilizadas *Redes Neurais*, que a partir dos dados existentes, criam modelos que permitem estimar dados em falta.

A fase de *Limpeza de Dados* é essencial para que os dados anómalos, o ruído e os dados

omissos não afectem as fases seguintes do processo de caracterização de consumidores.

2.6 Redução de Dimensionalidade

O conjunto de dados obtido é baseado nas leituras realizadas a partir da amostra de consumidores ao longo de um período de tempo. A dimensão da amostra é função do número de consumidores que a constituem, da cadência de recolha de dados e do número total de dias que constitui o intervalo total de tempo das campanhas de medição.

Assim sendo, existe um elevado volume de dados constituído por um número de diagramas de carga diário total, igual ao número de consumidores da amostra multiplicado pelo número de total de dias em que se realizaram as campanhas.

Deste modo, os resultados de um estudo com 10 consumidores, durante um ano (365 dias), com uma cadência de 15 minutos, é igual a cerca de 349 440 valores no total.

Tendo em conta a abordagem a adoptar para estudar o problema, é possível, nesta fase, fazer um pré-processamento dos dados reduzindo-se o volume inicial, sem perda essencial de informação, e mesmo simplificar a sua aplicação nas fases seguintes, tendo em conta que o detalhe da informação obtida depende da facilidade e da capacidade no tratamento dos dados.

Nesta fase, os dados podem ainda ser separados *temporalmente* e por outras condições, tais como, as épocas do ano ou mesmo as semanas. Após esta separação, poderão ser realizados estudos individuais para as diferentes condições de carga que permitam obter perfis de consumo diferentes para cada situação. Se os dados não forem separados de acordo com as condições de carga, estas podem ser introduzidas como atributos (ver 2.5) e separadas *à posteriori* pelo modelo.

A redução de dimensionalidade também pode ser conseguida através da utilização de diagramas representativos de cada consumidor num período pré-estabelecido (em cada semana por exemplo).

Cada um destes diagramas representativos pode ser determinado através da média dos diagramas de carga diários, para cada um dos períodos em estudo, por consumidor.

2.7 Normalização

Nesta fase, procede-se à normalização, ou seja à conversão em escalas comparáveis, sobretudo em estudos onde a forma do diagrama transporta a informação mais relevante.

Assim sendo, diversas tentativas de normalização foram efectuadas com vários métodos, tendo-se verificado que este procedimento pode ocasionar perdas de informação importantes, pelo que no estudo em apreço não foi efectuada nenhuma normalização, mantendo-se o conjunto de dados original.

No entanto, a escolha do factor de normalização a utilizar nem sempre é simples e deve ser feita tendo em conta os dados disponíveis, a análise que se pretende fazer, os algoritmos de *DM* que são utilizados e o tipo de resultados finais pretendidos.

Apesar de existirem diversos factores de normalização utilizados no tratamento de dados em trabalhos, os quais descrevem modelos para a obtenção de perfis de consumidores, não existe consenso quanto aos factores a utilizar.

Estes podem ser descritos de diversas maneiras como :

- Divisão do diagrama de carga diário do consumidor pela energia total que este consumiu no período em estudo. No entanto existem problemas devido à existência de perfis de consumo de consumidores distintos, com a mesma forma mas com consumos totais diários de energia muito diferentes. Assim, este tipo de normalização perde a noção de *amplitude* do diagrama.
- Divisão do diagrama de cargas de cada dia pela energia total consumida nesse dia. Neste caso existe o risco de existirem dois consumidores com diagramas de carga diferentes em termos de volume mas após a normalização aparentam ser semelhantes.
- *Potência Média*: obtida no período em estudo pela qual são divididos os diagramas de cargas representativos. Este factor permite manter a informação relativa à energia total, para qualquer consumidor, em todas as condições de carga. No entanto, apresenta o risco de perder a informação relativa às diferentes formas de diagrama de carga entre consumidores com valores de potência mais afastados do valor da potência média utilizada na normalização.
- *Potência de Ponta*: Neste caso, o valor da potência pela qual se dividem todos os diagramas de carga representativos não deve ser o valor mais elevado de leitura ou máximo, o qual pode corresponder a uma situação anómala, mas sim ao máximo valor de potência dos diagramas representativos. Este factor permite obter e manter a informação relativa à energia total, para qualquer consumidor, em todas as condições de carga (dia da semana, semana do ano, estação do ano, épocas especiais do calendário). Como desvantagem apresenta a perda parcial da informação relativamente a consumidores com valores de potência de ponta baixos, a informação relativa às diferenças entre a forma dos diagramas de carga.
- *Potência de Ponta do diagrama representativo de cada consumidor*: Neste caso todos os diagramas representativos normalizados assumem valores dentro do intervalo $[0, 1]$. Com este factor de normalização é possível reter toda a informação relativa à forma, mas não é possível obter os resultados dos perfis de consumo em função dos valores originais das potências dos consumidores.

A selecção do factor de normalização a utilizar deve ser feita de acordo com o modelo de determinação do perfil do consumidor que se pretende implementar, procurando sempre

minimizar a perda de informação. Em resultado da multiplicação de todos os consumidores pelo mesmo factor, apenas se verifica uma redução de escala, mantendo-se as amplitudes relativas dos diagramas de carga de cada um.

Para dissipar a dúvida da necessidade da normalização e para que se possam comparar os diagramas de carga representativos de cada consumidor é necessário proceder a sua conversão para escalas comparáveis.

Se o objectivo for a eliminação do efeito do volume do consumo, o tipo de normalização a utilizar deve permitir comparar os diagramas representativos de acordo com a sua forma. Por outro lado, se a opção for a de utilizar como factor de normalização a potência máxima obtida no período de monitorização, corre-se o risco de seleccionar um valor obtido numa situação anómala, que pode conduzir a um processo de normalização pouco preciso.

Neste contexto, optou-se por recorrer à potência máxima encontrada no conjunto dos diagramas representativos.

Após a realização de diversas experiências de normalização observou-se que os resultados obtidos não revelaram qualquer vantagem relativamente aos valores reais. Assim, optou-se pela utilização dos estudos sem normalização de dados.

2.8 Selecção de Atributos

Nesta fase, procede-se a uma tentativa de redução do volume dos dados e à procura de relações úteis para a representação dos mesmos, dependendo dos objectivos propostos no modelo neste sentido, esta operação complementar do processo de redução de dimensionalidade descrito em 2.6.

A informação disponível para caracterizar os consumidores é dividida em:

- informação relacionada com o padrão de consumo de cada consumidor;
- com as características comerciais definidas pela empresa de electricidade;

Deste modo é necessário representar cada consumidor por um conjunto de atributos relacionados com a informação indicada. O processo de selecção e representação dos atributos depende da informação disponível e do modelo a implementar.

2.8.1 Atributos relacionados com o Diagrama de Carga

Os atributos são utilizados para introduzir no modelo a informação de como varia o consumo ao longo de um determinado período de tempo, tal como, um dia, uma semana ou um mês.

Num diagrama de carga é representada a evolução do consumo ao longo de uma semana. Esta informação permite distinguir o padrão de consumo de cada consumidor

para diferentes condições de carga. Nos diferentes trabalhos descritos na bibliografia, este tipo de informação é normalmente introduzida por vectores de atributos do tipo contínuo.

Estes atributos podem ser utilizados pelo modelo de diferentes formas, onde:

- M - número de clientes da amostra
- H - número de valores lidos, o que varia com a cadência de recolha das leituras para diagramas de 24 horas
 - (se esta for de 30 minutos vem $H=48$)
 - (se esta for de 15 minutos vem $H=96$)

O valor de H , resultante da cadência de recolha de informação, pode ser reduzido caso se aceite uma taxa de amostragem dos diagramas mais baixa, diminuir o esforço computacional. Em vez de se visualizar o diagrama de cargas por leituras de 15 em 15 minutos, pode-se representar através de valores de hora a hora, o que reduz a quantidade de registos de dados para 25%, ou mesmo de duas em duas horas o que diminui para $\frac{1}{8}$.

Em alternativa à utilização dos valores das leituras efectuadas, como atributos para a representação dos diagramas de cargas, podem ser utilizados factores caracterizadores de diagramas 2.9 determinados a partir destes.

Estes factores permitem representar a forma do diagrama de carga através de um número muito menor de atributos, quando comparada com o inicial. A definição deste tipo de factores difere entre países e até entre empresas de electricidade. No estudo de atribuição de perdas descrito no Cap. 6, estes factores são utilizados pelo meio de representação da forma do diagrama.

2.8.2 Atributos relacionados com os Dados Comerciais

Os atributos relacionados com os dados comerciais são obtidos a partir dos dados de facturação de cada consumidor, armazenados na base de dados da empresa de electricidade. Este tipo de informação varia de País para País e de empresa para empresa, pois está relacionada com o tipo de classificação utilizado por cada empresa para diferenciar os seus clientes.

No caso português, existem os seguintes atributos relacionados com os dados comerciais:

- Código de identificação do cliente;
- Nível de tensão;
- Código da actividade (CAE);
- Potência contratada;

- Tipo de Tarifa;
- Energia mensal consumida;
- Data de ligação;
- Rede onde se encontra ligado;
- Número de fases;
- Escalão;

Da lista anterior convém referir que:

Nível de tensão a que os consumidores são alimentados: os consumidores podem ser alimentados em muito alta tensão [MAT], alta tensão [AT], média tensão [MT], baixa tensão especial [BTE] e baixa tensão [BT]. Esta informação é normalmente utilizada para separar os dados e elaborar estudos individualizados, de acordo com o nível de tensão.

Código de actividade: designa o tipo de actividade a que os consumidores estão associados. No caso de [BT] o tipo de actividade é designado através de um conjunto de letras e é tratado como um atributo discreto, sob um domínio de regras, que permite obter um sistema de classificação, que pode atribuir novos consumidores às classes obtidas. [1]

Potência contratada: indica o valor da potência disponibilizada pela rede a cada consumidor. A forma de atribuição da potência contratada difere consoante o nível de tensão. Os consumidores de [MT] fixam a potência contratada através do valor da sua potência de ponta. Os consumidores de [BT] devem seleccionar o valor da potência aquando da celebração do contrato, não podendo ultrapassar esse valor por limitação do contador.

Neste caso, a potencia contratada pode assumir um dos valores discretos de uma escala estabelecida pela EDP. (ver tabela 2.2)

Em [BT] existem várias tarifas diferentes:

- Tarifa Simples, na qual o preço da energia é o mesmo para todas as horas;
- Tarifa Bi-horária, em que o período de tarifação é dividido por horas (fora de vazio e vazio), com preços diferentes.
- Tarifa Tri-horária, em que o período de tarifação é também dividido por horas (pontas, vazio e cheias), com preços diferentes.

Energia anual ou mensal consumida: indica o valor obtido pelos contadores que registam o valor da energia consumida e facturada no período de um mês. A energia mensal consumida é tratada como um atributo contínuo, de domínio real não negativo. No entanto, nem todos os factores referidos anteriormente são relevantes para a análise em curso.

Tabela 2.2: Opções de Tarifa [BT]

Opção Tarifária	Lim. Potência
Social	1,15 a 2,3 kVA
Simples	1,15 a 20,7 kVA
Bi-horária	3,45 a 20,7 kVA
Simples	27,6 a 41,4 kVA
Médias Utilizações	27,6 a 41,4 kVA
Longas Utilizações	27,6 a 41,4 kVA
Médias Utilizações	<41,4 kW
Longas Utilizações	>41,4 kW
Sazonal Simples	3,45 a 20,7 kVA
Sazonal Bi-horária	3,45 a 20,7 kVA
Sazonal Tri-horária	3,45 a 41,4 kVA
Iluminação Pública	

2.8.3 Outros Atributos

Dado que a estação do ano e o dia da semana têm influência no comportamento dos consumidores, se os dados não forem tratados de forma separada, devem ser introduzidos atributos que permitam identificar os factores anteriormente referidos. Pode ser criado um atributo categórico referente ao dia da semana: (segunda-feira, ..., sábado, domingo) e outro referente ao mês: (Janeiro, Fevereiro, ..., Dezembro). De acordo com o modelo que se pretende desenvolver, podem surgir ainda outros atributos relacionados com os tipos de cargas existentes nas instalações.

2.9 Factores Caracterizadores de Diagramas

Os factores caracterizadores de diagramas permitem qualificar sucintamente os diagramas segundo determinadas vertentes, evidenciando cada um dos factores específicos. Estes servem para observar certo tipo de atributos, tais como a carga ou as perdas. As fórmulas de cálculo utilizadas, no CCR [8], na determinação dos diversos factores de caracterização de diagramas foram:

$$F_c = \frac{P_{med}}{P_{max}} \quad (2.1)$$

Factor de Carga de um Diagrama

$$F_p = \sum_{h=1}^n \frac{(P_h/P_{max})^2}{n} \quad (2.2)$$

Factor de Perdas de um Diagrama Diário

$$F_s = \frac{Max_h \sum_i P_{ih}}{\sum_i Max_h P_{ih}} \quad (2.3)$$

Factor de simultaneidade de um conjunto de diagramas i

$$CP_i = \frac{P_i(h = \text{ponta da rede})}{P_{max} Rede} \quad (2.4)$$

Contribuição do consumidor i para a Ponta($\sum_i cp_i = 1$)

$$FOP_i = \frac{P_i(h = \text{ponta da rede})}{P_i Max} \quad (2.5)$$

Factor de ocupação da ponta do consumidor i ($fop_i = 1$) quando a ponta do consumidor coincide temporalmente com a da rede.

onde: h designa um posto horário genérico; n intervalo de tempo a analisar.

Neste trabalho foram apenas utilizados os factores de carga e de perdas, os quais são referidos ao longo do trabalho.

2.10 Factor de Carga e Factor de Perdas

2.10.1 Factor de Carga

Define-se Factor de Carga - Fc , como sendo a relação entre a Potência Média (P_{med}) e a Potência Máxima (P_{max}), ao longo de um período de tempo T :

$$Fc = \frac{P_{med}}{P_{max}} = \frac{\frac{1}{T} \times \int_0^T P(t) dt}{P_{max}} \quad (2.6)$$

onde:

F_c - Factor de carga

P_{med} - Consumo médio [kW] ou potência média

P_{max} - Consumo máximo [kW] ou potência máxima

$P(t)$ - Potência no instante t [kW]

T - Período de tempo [h]

No caso em estudo utiliza-se um diagrama discreto definido de 15 em 15 minutos, podendo ser facilmente convertido em diagrama de hora em hora ou noutra unidade temporal.

Na medida em que a potência média é inferior à potência máxima (ou igual, no caso de cargas constantes), o factor de carga varia nos seguintes limites:

$$0 < F_c \leq 1 \quad (2.7)$$

O integral de $P(t)$ - (potência instantânea) representa a energia fornecida ao sistema durante o período T (kWh), o mesmo também é válido em [pu] :

$$\int_0^T P(t)dt = E[kWh] \quad (2.8)$$

A expressão 2.6 é simplificada em:

$$F_c = \frac{E}{T \times P_{max}} \quad (2.9)$$

Esta é a forma mais comum de determinação do F_c , visto que E e P_{max} são grandezas geralmente utilizadas nas subestações e também nos pontos de fornecimento dos consumidores mais importantes e no caso em estudo para os consumidores da amostra.

2.10.2 Factor de Perdas

Em cada instante t o sistema de distribuição recebe das subestações alimentadoras uma potência $P1(t)$ e fornece aos consumidores uma potência $P2(t)$, inferior a $P1(t)$.

A diferença entre estas duas grandezas, também variável ao longo do tempo, representa a $L(t)$ - perda técnica de consumo no instante t :

$$L(t) = P1(t) - P2(t)[kW] \quad (2.10)$$

onde:

$L(t)$ - Perda técnica de consumo no instante t [kW]

$P1(t)$ - Potência recebida pelo sistema de distribuição no instante t [kW]

$P_2(t)$ - Potência fornecida aos consumidores ligados ao sistema de distribuição no instante t [kW]

NOTA: A definição só é correcta se for considerado o consumo efectivamente fornecido, independentemente de ser ou não facturado.

As definições relativas a $P(t)$ - Potência de carga (diagrama de carga, curva de permanência de carga e factor de carga) podem ser aplicadas a função $L(t)$ - perda técnica de consumo, sendo formalmente válidas as diversas expressões:

$$Fp = \frac{L_{med}}{L_{max}} = \frac{\frac{1}{T} \times \int_0^T P(t) dt}{L_{max}} \quad (2.11)$$

$$0 < Fp \leq 1 \quad (2.12)$$

$$\frac{1}{T} \times y \int_0^T P(t) dt = e [kWh] \quad (2.13)$$

$$Fp = \frac{e}{T \times L_{max}} \quad (2.14)$$

onde:

Fp - Factor de perdas

L_{med} - Perda média de consumo [kW]

L_{max} - Perda máxima de consumo [kW]

$L(t)$ - Perda instantânea de consumo [kW]

e - Perda de energia no período T [kWh]

T - período de tempo [h]

De salientar que o factor de perdas é simplesmente o factor de carga das perdas.

Quando as perdas são independentes da carga como acontece, por exemplo, com as *perdas no ferro*, nos núcleos dos transformadores de distribuição e reguladores de tensão, as perdas nas baterias de condensadores, entre outras, são independentes (ou quase) da carga.

Admitindo que estas perdas são constantes, verifica-se:

$$Fp = 1 \quad (2.15)$$

No caso das perdas dependentes da carga (caso mais geral das perdas *ohmicas* nos condutores das redes e nos equipamentos existentes), verifica-se:

$$0 < Fp \leq 1 \quad (2.16)$$

Caso se aceite que as cargas apresentam factor de perdas praticamente constante, ou seja os diagramas de cada consumidores mantêm uma determinada forma (tipicidade) considera-se a seguinte relação entre a perda e o consumo:

$$L(t) \cong K[P(t)]^2 \quad (2.17)$$

na qual:

$L(t)$ - Perdas de consumo no instante t

$P(t)$ - Potência no instante t

K - Constante de proporcionalidade

2.10.3 Determinação do Factor de Perdas

O processo de cálculo do factor de perdas pela definição da equação 2.11 é, na prática inviável, uma vez que obrigaria ao cálculo do diagrama de perdas para cada caso.

A solução mais prática, consiste em realizar o cálculo anteriormente descrito em diagramas de cargas típicos, ao obter como resultado custos unitários das perdas técnicas, os quais incorporam o custo das perdas associadas da energia.

O factor de perdas (das perdas dependentes da carga) é calculado a partir das expressões 2.6 e 2.17:

$$\frac{\frac{1}{T} \times \int_0^T P(t) dt}{L_{max}} = \frac{\frac{1}{T} \int_0^T [P(t)]^2 dt}{L_{max}} \quad (2.18)$$

As expressões 2.6 e 2.18 podem ser simplificadas substituindo o integral por um somatório, caso exista uma sequência de medições horária regular (como exemplo: de 15 em 15 minutos).

Assim:

$$Fc = \frac{\sum_{n=1}^7 P(n)}{N \times P_{max}} \quad (2.19)$$

$$Fp = \frac{\sum_{n=1}^7 [P(n)]^2}{N \times [P_{max}]^2} \quad (2.20)$$

Na ausência de um conhecimento completo das curvas de carga procura-se quantificar o factor de perdas- Fp das perdas variáveis, relacionando-o somente com o factor de carga- Fc .

O trabalhar das expressões Fc e Fp permite deduzir a expressão:

$$Fc^2 \leq Fp \leq Fc \quad (2.21)$$

Com a impossibilidade matemática de definir o factor de perdas a partir unicamente do factor de carga, este problema tem sido resolvido de forma empírica, ou seja, através da utilização dos parâmetros k ou α :

$$Fp = k \times Fc + (1 - k) \times Fc^2; \quad 0 < k < 1 \quad (2.22)$$

ou

$$Fp = Fc^\alpha; \quad 1 < \alpha < 2 \quad (2.23)$$

Os novos parâmetros k introduzidos nas expressões 2.22 e α em 2.23 representam o grau de indeterminação de 2.21 e devem ser estimados periodicamente a partir de diagramas de carga reais.

Uma vez que os custos das perdas são normalmente referidos a períodos anuais é necessário conhecer os dados relativos aos dias úteis e aos sábados, domingos e feriados.

Neste documento propõe-se o valor típico de $k = 0,3$ para [BT], podendo haver alterações se os dados reais assim o requererem.

Os factores de carga e de perdas estão sujeitos a variações, o que obriga a uma actualização periódica destes, devido nomeadamente a:

- Sazonalidade da Carga: fenómeno de ocorrência limitada, que obriga a desdobrar o ano tantas vezes quantas as necessárias, além de se considerar dias úteis, sábados, domingos e feriados. A existência de cargas sazonais provoca uma redução no factor de carga e no factor de perdas, considerados em bases anuais.
- Crescimento da Carga: aumento regular da carga (e consequente perdas técnicas variáveis), ao longo do tempo, provoca pequenas reduções nos factores de carga e de perdas, previamente calculados em bases anuais.

2.11 Modelos para a caracterização de perfis

O tipo de modelo a desenvolver depende do objectivo de aplicação dos resultados e dos dados disponíveis. Na bibliografia da especialidade podem-se encontrar trabalhos que descrevem a aplicação de diversas técnicas para a criação de modelos, que permitem estimar o diagrama de cargas típico, a partir de um conjunto de variáveis explicativas e dependendo do tipo do modelo a utilização das condições meteorológicas, dos factores temporais, entre outras. No entanto é possível obter perfis através dos diagramadores utilizados em campanhas de medição com dados reais de consumo.

Tendo em consideração a importância que os modelos desempenham na definição dos perfis de consumo, como foi referido em 2.3, de seguida apresenta-se uma descrição sucinta das abordagens anteriormente mencionadas.

2.11.1 Abordagem Estática

Os modelos estáticos baseiam-se apenas em dados históricos do estudo da carga, habitualmente diferenciados pela hora, pelo dia da semana, pelo mês em que foram recolhidos.

Este tipo de abordagem precisa de informação temporal, a qual pode ser fornecida com a instalação de diagramadores específicos para o efeito. No entanto existem dados que apenas podem ser recolhidos através de inquérito directo aos consumidores, caso se pretenda considerar também a informação sobre a utilização da energia.

Nos trabalhos apresentados por Schick em [9] e [10], a estimativa dos diagramas diários de utilização das cargas é obtida a partir dos dados relativos a cada aparelho eléctrico de utilização doméstica, medidos numa amostra de consumidores residenciais.

Este modelo combina um estudo longitudinal, ao representar a evolução da carga ao longo do tempo, com um estudo sectorial, que representa a evolução da carga com as características de utilização dos consumidores domésticos. Os diagramas de carga, para os diferentes aparelhos eléctricos, são obtidos ao agregar os perfis de consumo de todas as instalações eléctricas monitorizadas, que possuam um determinado tipo de aparelho, subtraindo-lhe o perfil de consumo agregado das instalações eléctricas, que não possuam esse tipo de aparelho.

A recolha de dados é feita através de um sistema de inquérito aos consumidores, o qual procura identificar as características do consumo. O inquérito é estruturado com base em técnicas de amostragem, para definir quais os consumidores que são submetidos à recolha de dados. O padrão de consumo, para cada consumidor em estudo, é construído com base em análise estatística sobre os dados recolhidos, durante o inquérito. O padrão de consumo sazonal de cada classe de consumidores é determinado por integração dos padrões de consumo do mesmo tipo. A potência consumida por cada classe de consumidores é então estimada com base no padrão de consumo dessa classe e na energia consumida por todos os consumidores que a integram. O modelo permite, ainda, obter o perfil de consumo típico para todo o sistema de energia, ao agregar os perfis de consumo de todas as classes encontradas. Esta abordagem pode ser aplicada directamente aos perfis baseados nas campanhas de medição.

2.11.2 Abordagem Dinâmica

Como foi referido anteriormente uma abordagem dinâmica utiliza não só dados históricos, mas também requer outras variáveis.

Para a criação de modelos dinâmicos do diagrama de carga existem diversas metodolo-

gias que permitem estudar a contribuição de diferentes tipos de aparelhos ligados à rede, como por exemplo, os termoacumuladores e os recuperadores de calor, ar condicionado.

No modelo baseado na caracterização da procura, o aquecimento da água e do ambiente e o ar condicionado são representados através de modelos complexos.

Esta abordagem assume particular importância nos consumidores domésticos, os quais dispõem de aparelhos, como o *ar condicionado*, que podem influenciar bastante a forma como a energia eléctrica é consumida.

O conhecimento deste tipo de informação é essencial para estudos de gestão da procura, para a implementação de medidas de conservação de energia, para a criação de estruturas tarifárias e mesmo de novos serviços para o consumidor.

Deste modo existe um perfil base baseado em dados históricos ao qual se pode adicionar informação extra utilizando como variáveis de previsão as condições meteorológicas, as características térmicas dos edifícios, os equipamentos existentes ao dispor de cada consumidor e as informações relativas à estrutura tarifária em vigor.

No modelo baseado nos diagramadores a informação não se encontra separada em componentes como no anterior, no entanto é recebida a informação global sobre o consumo real. Com a adição desta nova informação é possível alterar a forma do diagrama ajustando-a mais correctamente à realidade. Quando existem tarifas que variam com o período horário, como no caso português, em que existe a tarifa bi-horária para consumidores domésticos, podem-se encontrar *picos de carga* nas horas em que se inicia a tarifa mais económica.

Em Calloway [11] e Chong [12] podem-se encontrar descrições deste tipo de modelos em que são utilizadas variáveis explicativas, como as características térmicas dos edifícios, que não são habitualmente registadas pelas empresas distribuidoras, pelo que não é simples proceder à sua implementação, sendo necessário recorrer a inquéritos para colmatar essa falha de informação.

2.11.3 Modelo por análise de *clusters*, *Load Profiling*

Independente da abordagem seleccionada a construção de um *PCT* pode ser feita por *clustering* as diferenças entre os vários modelos baseados na análise de clusters consistem, essencialmente, na técnica de *clustering* aplicada e nas variáveis de previsão seleccionadas.

Muller [13] propõe uma técnica de *clustering* conceptual, ao utilizar diversos conjuntos de variáveis de previsão. O conjunto de padrões a segmentar é constituído por diagramas de carga diários, recolhidos ao longo do tempo, para uma amostra de consumidores. A normalização do diagrama de cargas pode ser feita através da sua divisão pelo consumo médio diário. No *clustering* baseado em factores climatéricos são utilizadas diversas variáveis climatéricas, registadas diariamente, relacionadas com a temperatura e a luminosidade, sendo obtida uma primeira segmentação dos dados.

Em Pitt [14] é proposto um modelo de classificação, baseado na criação de perfil de consumidor, representados por uma grande diversidade de atributos. Cada consumidor

é representado por um vector de atributos de diferentes tipos. O modelo de previsão do diagrama de cargas diário é baseado num algoritmo adaptativo, não paramétrico de regressão, *Multivariate Adaptive Regression Splines* MARS que cria um modelo do efeito das condições atmosféricas no consumo, de forma independente dos restantes atributos.

A determinação dos diferentes perfis de consumo é feita através de uma técnica de *clustering* hierárquico, baseada em *Árvores de Decisão Adaptativas*, dividindo de forma recursiva o conjunto inicial, ao utilizar diversos algoritmos de *clustering* aplicados localmente. Este modelo permite estimar para determinados subconjuntos de consumidores a sua curva de carga (e respectivo intervalo de confiança), para diferentes condições atmosféricas, épocas do ano e dias da semana. O conjunto de consumidores encontrados nos *Nós Extremidade* da árvore de decisão, permite definir um conjunto de perfis de consumidores, ao representar as diferentes classes de consumo.

Em Gasperic [15] é apresentada uma abordagem de determinação de perfis de consumo diários, ao utilizar um processo de *clustering* hierárquico, baseado num procedimento em três passos, para agrupar diagramas de carga normalizados numa estrutura hierárquica de *clusters* em forma de árvore.

Deste modo a partir dos dados dos diagramadores o objectivo é classificar os perfis de consumo em grupos coerentes denominados *Typical Load Profiles* TLP, ou *PCT*, representados por 96 valores de carga, distribuídos ao longo de 24 horas.

O algoritmo é aplicado a um conjunto de dados, resultante de uma campanha de medição, ao serem recolhidos diagramas de carga diários para a amostra de consumidores. Estes diagramas são separados de acordo com o dia da semana e a estação do ano, segundo os seguintes grupos: dias úteis (de segunda a sexta-feira), sábados e domingos. O TLP é então determinado através da combinação do tipo de actividade com o grupo a que pertence.

Chicco em [16] apresenta um modelo de caracterização de consumidores para aplicação ao estudo de opções tarifárias. Este modelo é baseado na caracterização de consumidores, ao criar classes representadas pelo seu perfil de consumo diário típico. A criação dos diferentes perfis de consumo é feita através de um algoritmo de *clustering*.

Os dados relativos aos consumidores são recolhidos através de uma campanha de medições, tendo por base uma amostra de consumidores definida através de um processo de amostragem baseado em estratificação dos consumos [17].

Estes dados são utilizados para criar os diagramas representativos relativos a cada tipo de consumidor, estes diagramas atribuídos ao mesmo *cluster* e pertencem à mesma classe de consumo. O processo de *clustering* proposto é baseado em aprendizagem não-supervisionada em [18] e possui uma estrutura de funcionamento partitiva.

Na figura 2.2 pode-se observar um exemplo de um diagrama diário dos dados da campanha de medição do projecto CCR [1] relativo a consumidores [MT].

Este processo não requer a inicialização do número de *clusters* e consegue determinar automaticamente os centros dos *clusters*. A medida de Distância Euclidiana, utilizada no

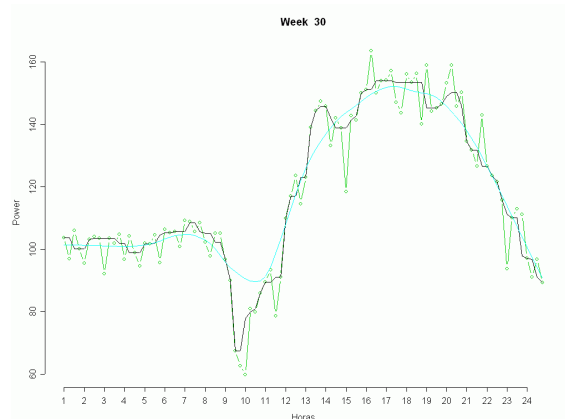


Figura 2.2: Perfil diário calculado a partir de DM

processo original, foi alterada através da introdução de um índice, que diferencia os pesos de cada atributo, para que o impacto dos atributos com maior variância seja aumentado no cálculo da Distância Euclidiana pesada. Para cada *cluster* obtido, é definido um diagrama de carga representativo, determinando a média dos diagramas dos consumidores incluídos. O conjunto de diagramas representativos obtido define as diferentes classes. O estudo das estruturas tarifárias é feito com base nos diagramas representativos de cada classe.

2.12 Índices de Avaliação dos Resultados

O processo de avaliação dos resultados obtidos varia de acordo com o modelo utilizado, podendo basear-se em valores finais de erro, na avaliação do desempenho das técnicas utilizadas ou no conhecimento de peritos. Os perfis de consumo não pretendem fornecer valores com elevada precisão, para um dado ponto horário, mas sim uma estimativa da forma como se distribui o consumo ao longo do dia.

Deste modo, a medição da sua precisão depende de vários factores, divididos em: factores relacionados com a sua determinação e factores relacionados com a sua utilização.

As principais fontes de erro que podem ocorrer na construção dos modelos são :

- Erro introduzido pelo sistema de recolha de dados: este advém do erro introduzido pela aparelhagem de medição, pelo erros nas leituras e pelos erros na transmissão. Pode ser atenuado na fase de limpeza de dados, mas não há garantias de ser completamente eliminado;
- Erro de amostra; apesar dos cuidados prestados na selecção dos elementos da população que compõem a amostra, não é possível conseguir que esta represente com exactidão o comportamento da população como um todo, na medida em que a

amostra é seleccionada tendo por base o consumo (variável unidimensional) e é avaliada com base nos diagramas (variável multidimensional);

- Imprecisão do modelo: o modelo utilizado na determinação dos perfis de consumo não consegue determinar os diferentes perfis sem um certo erro;
- Desvio no perfil: o comportamento da população que se pretende caracterizar pode sofrer alterações ao longo do tempo, aumentando o desvio entre o perfil de consumo obtido e a população que este pretende representar.

Estas fontes de erro influenciam negativamente todos os modelos e consequentemente os índices de desempenho finais de avaliação.

Nos modelos de operação não supervisionada a avaliação dos resultados é mais difícil, pois não existe um *Valor-Objectivo* que possa ser utilizado como base de comparação para determinar o erro cometido. A utilização de um conjunto de dados de teste é validada para testar a generalização dos modelos. Nestes modelos, os peritos podem desempenhar um papel importante numa primeira abordagem para a validação dos resultados.

Adicionalmente, é possível definir alguns índices baseados nas medidas de distância, permitindo avaliar a proximidade entre elementos do mesmo *cluster* e mesmo a distância entre *clusters* diferentes. Este tipo de índices permite essencialmente comparar resultados obtidos, ao utilizar atributos ou algoritmos diferentes.

Assim, nesta metodologia é definida uma forma de agrupar *à priori* os consumidores, baseados no histórico com dados reais.

Capítulo 3

Data Mining

Nesta secção existe uma breve descrição do conceito de *DM* e da inter-relação com outras disciplinas, tais como, a Estatística e os Sistemas de Apoio à Decisão. Finalmente, são descritas, de forma sucinta, as técnicas mais utilizadas em *DM*.

Recentemente, com a evolução dos sistemas informáticos, assistiu-se a uma crescente capacidade de geração e armazenamento de dados muito além do que estava previsto. Com os avanços científicos nos sistemas de recolha de dados, o volume de dados armazenado aumentou de uma forma quase incontrolável. Este aumento de capacidade de recolha e armazenamento de dados não é devidamente aproveitado se não for acompanhado por uma evolução dos métodos de análise.

Os estudos baseados apenas em simples bases de dados, as quais são suficientes para a criação de relatórios, não permitem analisar o conteúdo de forma a dar lugar a novo conhecimento. Deste modo, surgiu uma nova geração de técnicas e ferramentas que auxiliam, de forma automática e parcialmente inteligente, os utilizadores na análise de grandes volumes de dados.

Com estas ferramentas, a exploração toma um carácter semi-autónomo através dos algoritmos de *DM*. Estes algoritmos aplicados aos dados revelam informação *não visível* e relações indirectas e não triviais, implícitas e previamente desconhecidas nos conjuntos a analisar. Esta aplicação tem um grande sucesso na área do Apoio à Decisão, ao fornecer *novo conhecimento*, após a análise de dados históricos. No entanto, *DM* não é considerado como uma nova tecnologia. O conceito de extracção de informação tem sido utilizado desde há muito em estudos científicos (de carácter biológico, médico, de engenharia, entre outros). Apenas se pode considerar como nova a convergência de diversos factores que criaram novas oportunidades para a utilização de *DM* no mundo empresarial.

As empresas aperceberam-se que *Dados* e *Informação* são *coisas* diferentes, ao tentar transformar os dados recolhidos nos últimos 30 anos em informação, tornando-a num factor de competitividade. As facilidades introduzidas pelas novas tecnologias no armazenamento de dados permitiram que, para um utilizador final, seja fácil proceder à análise dos dados

e extrair a informação valiosa para a gestão. Neste momento, *DM* é utilizado em variadas e diferentes aplicações, desde proceder à segmentação de mercados, ao planeamento do lançamento de novos produtos e serviços, até à detecção de fraudes.

No caso específico da liberalização dos mercados de electricidade, esta tecnologia terá um papel importante na manutenção da competitividade das empresas e poderá ser aplicada directamente, às áreas de controlo de custos, à qualidade de serviço e à criação de novos mercados, à descoberta de nichos de mercado, à identificação de novos produtos e serviços. Através do estudo e análise de conjuntos de dados dos clientes com *DM* é mais simples a identificação das necessidades dos consumidores, o que permite fornecer um serviço de electricidade adequado.

3.1 DM e Estatística

Segundo Glymour [19] a estatística tem *pouco a oferecer* aos algoritmos de *DM*, na medida em que procura as relações nos dados, mas, em contrapartida, tem muito para oferecer no que se refere aos testes de hipótese, desde a avaliação dos resultados obtidos até à sua aplicação. Para Rodrigues [20] as fronteiras entre *Data Mining* e a Estatística não são claras; ambas as disciplinas têm por objectivo descobrir as relações nos dados e muitas das técnicas de análise de dados utilizadas nas ferramentas de *DM* são especializadas em aproximações estatísticas. Por outro lado, *DM* utiliza ideias, ferramentas e métodos de outras áreas tais como a computação, as tecnologias de Bases de Dados e a Inteligência Artificial, não incluindo, contudo, muitas das áreas que fazem parte da estatística original.

A Estatística tem um papel muito importante na selecção dos dados que compõem a amostra, na preparação dos dados e na análise do conhecimento directo. Ao longo do tempo, muito do trabalho dos estatísticos foi centrado apenas na avaliação e adequação dos modelos aos dados e no teste de hipóteses.

No que diz respeito ao presente estudo, a importância da Estatística reflecte-se nos seguintes pontos:

- Selecção de dados: criação de esquemas de amostra;
- Limpeza de dados: técnicas de detecção de *Outliers*;
- Estimativa de parâmetros de ruído e suavização dos dados;

Estes são pontos claramente relevantes, para a avaliação dos resultados de *DM*, para a filtragem dos dados, para o processo de *DM* na selecção de parâmetros e finalmente, para o ajuste dos modelos aos dados.

Nas análises estatísticas, existem condições a ser verificadas *à priori*, tais como, a distribuição dos dados, a qual deve ser conhecida à partida, e o desenvolvimento e teste de hipóteses são feitos através de um processo de análise pré-feito. Em *DM*, pelo contrário,

assume-se uma distribuição de dados fonte e gera-se uma hipótese a partir dos dados históricos.

3.2 DM e Técnicas de Apoio à Decisão

Actualmente, as bases de dados são acompanhadas por várias ferramentas de pesquisa capazes de revelar a informação existente através de diversos critérios definidos pelos utilizadores. Para simplificar a operação e torná-la mais *amigável*, algumas destas ferramentas fornecem um *Graphical User Interface* - *GUI* para ajudar os utilizadores.

No caso das bases de dados relacionais estas ferramentas são conhecidas por ferramentas de *Structured Query Language* SQL. Nas bases de dados multi dimensionais, as ferramentas são denominadas por *Online Analytical Processing* [OLAP]. Uma vez que o objectivo do *DM* é extrair informação não trivial pode surgir a seguinte questão:

Qual a diferença entre *DM* e *SQL*? A utilização de *SQL* é importante para questões em que são impostas restrições na pesquisa dos dados para obter a resposta correcta. Em contrapartida, *DM* é indicado para questões de natureza exploratória.

Por exemplo, se uma empresa comercializadora de energia eléctrica pretende saber quais os clientes que consomem mais do que um determinado valor de energia por mês, pode obter esta informação ao utilizar *SQL*. Por outro lado, se a empresa pretende saber quais as interrelações entre os consumidores que pagam habitualmente as facturas, dentro do prazo, e os que pagam habitualmente fora do prazo terá de utilizar *DM*.

Enquanto que o *SQL* é muito útil para extrair informação directa mas simples, o *DM* permite examinar informação *menor* e indirecta.

As ferramentas de *OLAP* tornaram-se muito populares nos últimos anos, uma vez que permitem aos utilizadores manipular os dados de um *Data Warehouse*, de forma a obter várias perspectivas dos mesmos. Nesta perspectiva, diferentes dimensões correspondem a diferentes características, como exemplo: vendas, tipos de produtos, localização geográfica, entre outras.

Para permitir análises com relação a este tipo de informações, as ferramentas *OLAP* armazenam os dados num formato que corresponde a uma estrutura hiper-dimensional. Apesar das ferramentas *OLAP*, tal como *DM*, fornecerem respostas a questões a partir dos dados, estas não geram um processo de aprendizagem, nem criam novo conhecimento.

O *DM* distingue-se das ferramentas de *OLAP*, uma vez que o ênfase é centrado na descoberta de relações não directas nos dados. As questões são respondidas ao procurar novas relações entre os diferentes atributos dos registos do conjunto. Assim, como o propósito das ferramentas de *OLAP* não é o de substituir as bases de dados relacionais, o mesmo se passa com *Data Mining*. *DM* não é uma alternativa às ferramentas interrogativas, mas sim um complemento.

3.3 Principais Operações de DM

Apesar da complexidade envolvida, a maioria dos problemas que ocorrem a nível científico, podem ser incluídos nas seguintes operações: associação, classificação, análise sequencial, segmentação ou *clustering* e visualização. Cada uma destas operações é implementada através de várias metodologias, como análise estatística de dados, aprendizagem automática, *Redes Neurais* e reconhecimento de padrões. Por recorrer a metodologias de diferentes áreas científicas, *DM* é visto como uma área multidisciplinar.

As operações de *DM* podem ser divididas em dois tipos fundamentais:

- Operação Supervisionada - este tipo de operação necessita da identificação de um objectivo específico, antes da execução da técnica que implementa a operação de *DM* em questão.

Como exemplo de operação supervisionada existe a *Classificação Clássica*.

- Operação não Supervisionada - não precisa de um objectivo pré-definido nem informação *á priori*.

As técnicas que implementam estas operações analisam um conjunto de dados e descobrem relações sem quaisquer restrições relativas ao que essas relações possam inferir. Como exemplos desta operação existe a segmentação ou *clustering* e a análise de associações. Estas operações permitem descobrir novas relações, embora lhes ofereçam pouco significado, o que só é conseguido com análises suplementares.

3.3.1 Associação e Análise Sequencial

A operação de *Associação* é uma operação não supervisionada, que tem por objectivo encontrar relações entre os diferentes atributos de um conjunto. Esta operação é muito utilizada nas áreas de Marketing e Vendas, na procura de associações padrão, como por exemplo: 90 % dos clientes que comprem o produto *X* também comprem o produto *Y*. Este tipo de associações tem grande utilidade para organizar produtos nas prateleiras de um supermercado. Em [21] esta operação foi introduzida para gerar regras de associação em dados recolhidos nas caixas dos supermercados e em [22] é descrita uma aplicação para a previsão de produtos vendidos em conjunto.

Esta aplicação é sobretudo um exercício de processamento massivo de dados, pelo que pode recorrer a técnicas estatísticas, quer na optimização do processo de análise, quer na determinação de níveis de confiança e suporte de regras geradas.

3.3.2 Classificação

A classificação é um dos problemas de *DM* mais explorado, dado o seu vasto campo de aplicação e à fácil compreensão dos seus resultados [23]. Muitos dos problemas de

Data Mining podem ser denominados como problemas de Classificação. Como exemplo: a classificação de clientes bancários relativamente ao crédito com baixo, alto ou médio risco. Esta é uma função de aprendizagem que divide (ou distingue) os dados de acordo com um número pré-definido de características. Tem como objectivo construir modelos de classificação (classificadores), que aplicados a novos registos *não classificados* estabelecem automaticamente a classe a que pertencem.

Como exemplo, no caso da classificação de clientes mencionado anteriormente, o modelo de classificação é criado com base nos dados existentes sobre os consumidores relativos ao histórico do seu comportamento com o crédito. Este modelo de classificação é utilizado para atribuir uma classe a potenciais clientes de forma a poderem ser aceites ou rejeitados.

A classificação é uma aprendizagem direccionada, ou seja, os algoritmos são previamente testados com um conjunto de casos conhecidos e pré-classificados, denominado como *conjunto de treino* permitindo extrair *regras de classificação*. Com estas regras é construído um *modelo de Classificação*. *Árvores de Decisão*, *Redes Neurais* e raciocínio baseado em casos são as técnicas mais utilizadas para fazer classificação.

3.3.3 Clustering

Este tipo de problemas de *DM* tem por objectivo separar um conjunto de dados em subgrupos ou classes com interesse e significado. Todos os membros de cada subgrupo devem partilhar características comuns.

Clustering é uma operação de *DM* não supervisionada. Nesta operação, os algoritmos criam as classes (subconjuntos de registos que apresentam valores mais próximos em certos atributos), produzindo um esquema de classificação que divide o conjunto de dados [24] segundo as classes. A maior diferença relativamente à classificação reside no facto de esta ser supervisionada.

Esta operação é indicada para situações em que não é possível obter um conjunto de treino com registos pré-classificados. Os resultados do *clustering* podem ser utilizados para sumariar o conteúdo de cada segmento do conjunto ao considerar as características mais relevantes de cada *cluster*. Por outro lado, esta operação pode servir para preceder uma operação de classificação, ao produzir regras de classificação para cada um dos *clusters* descobertos. O *clustering* é normalmente realizado através de *Redes Neurais* ou *Fuzzy Logic*.

As técnicas de *clustering* podem ser combinadas com técnicas de classificação. Por exemplo, [25] é apresentado um exemplo de aplicação de uma operação de *clustering* para determinação de um conjunto de perfis de consumidores, seguida de uma operação de classificação que caracteriza cada um dos perfis obtidos com base nas características dos diagramas de carga dos consumidores.

O *Clustering* deve ser considerado como um exercício de exploração, sendo esta razão

uma das operações de *DM* com maior peso, complexidade e exigência computacional, actualmente.

3.3.4 Visualização

A aplicação de técnicas de visualização a problemas de *DM* inspirou-se no facto de os seres humanos terem maior facilidade em compreender as estruturas na forma visual. Este tipo de operações baseia-se na criação de ferramentas que permitem ao utilizador a visualização dos dados, sob várias perspectivas.

A descoberta de padrões nos dados é elaborada pelo utilizador enquanto analisa os dados sob a forma visual. No extremo, a aplicação desta operação em *DM* consiste na criação de um imenso ambiente de realidade virtual onde o utilizador se move na procura de novas relações entre os dados. A sua aplicação é comum em bases de dados espaciais, nomeadamente, nas imagens obtidas via satélite, em equipamento médico e mesmo em sistemas de informação geográfica [*GIS*].

Os principais desafios deste tipo de questões, quando comparados com os problemas clássicos de *DM*, são: transformar a sua complexidade e volume numa projecção (quase sempre a duas dimensões) onde determinadas relações entre os dados aparecem de forma evidente.

3.4 Componentes Primários dos Algoritmos de Data Mining

Nesta secção são apresentadas as descrições dos algoritmos que permitem efectuar as operações de *DM*. Segundo [20] pode-se indicar como componentes primários de qualquer algoritmo de *DM* os seguintes:

- Representação do Modelo
- Avaliação do Modelo

Representação do Modelo: a linguagem L é utilizada para descrever as relações encontradas. Assim, quanto maior for a capacidade de L representar o modelo, mais difícil e complexa é a sua utilização, logo a procura de novas relações e a interpretação do próprio modelo torna-se mais complexa.

Avaliação do modelo: esta componente estima o rigor com que um determinado modelo e os seus parâmetros consegue alcançar os objectivos do processo.

3.5 Técnicas de Data Mining

Várias técnicas que suportam as operações de *DM* estão constantemente a aparecer. Estas podem ser oriundas de diferentes disciplinas como a estatística, o reconhecimento de

padrões e a aprendizagem automática.

Algumas das técnicas mais conhecidas são :

- Regressão;
- Algoritmos de vizinhança mais próxima;
- *Clustering*;
- Redes Neurais;
- Árvores de Decisão;
- Algoritmos Genéticos;
- *Fuzzy Logic*;
- *Rough Sets*.

Cada técnica é implementada de diferentes formas; uma *Árvore de Decisão* ou uma *Rede Neuronal* é implementada com algoritmos diferentes. Por outro lado, uma aplicação de *DM* pode recorrer a diferentes técnicas para realizar diferentes operações. Uma aplicação pode recorrer a um algoritmo de *clustering* para fazer a segmentação inicial de clientes, classificar os *clusters* para identificar os factores mais importantes que influenciam as escolhas dentro de cada um deles e em seguida, construir um modelo optimizado para prever comportamentos futuros.

3.5.1 Regressão

A regressão é definida como a aplicação de métodos que procuram relações entre uma variável dependente (ou resposta) e um conjunto de variáveis independentes (ou de previsão). Estes estudos são habitualmente baseados num conjunto de medidas, relativas a um conjunto de objectos.

Através de alguns valores de entrada, a regressão é utilizada para estimar valores contínuos de variáveis. A regressão também pode ser utilizada para prever valores futuros de variáveis, tais como aumentos de preço de determinados produtos, níveis de stock, entre outros. Quando a estes modelos são aplicados valores de entrada correntes obtêm-se como resultado previsões futuras de comportamento.

Para a *previsão* e outros fins existem vários métodos de regressão, derivados da estatística, tais como a regressão linear e regressão polinomial. No caso de operações de *DM* podem surgir dificuldades adicionais devido a elevada dimensão do domínio (ou espaço de previsão) e às grandes variações possíveis no contradomínio (espaço de resposta), pelo que os métodos de regressão clássicos são de difícil aplicação. Nestes casos são mais

indicadas técnicas de regressão não clássicas que possuam maior flexibilidade [14], tais como a regressão passo-a-passo, a regressão hierárquica, a regressão não paramétrica, entre outras.

3.5.1.1 Formalização dos Problemas de Regressão

O objectivo de um processo de regressão é a obtenção de um modelo baseado numa amostra de objectos, pertencentes a uma função de regressão desconhecida. Esta amostra (conjunto de treino) é representada por um conjunto de pares com a forma x_i, y_i em que x_i é o vector com os valores dos atributos (variáveis de previsão) e y_i é o respectivo valor de resposta. Em problemas de regressão, a formulação matricial é utilizada com frequência para simplificar as expressões. Pode-se então representar X como a matriz de entrada, cuja i -ésima linha representa o vector de entrada x_i . Se tivermos n vectores, X terá dimensão $n \times a$, onde a é o número de atributos. Os valores objectivo das variáveis de entrada são guardados numa matriz Y de dimensão $n \times 1$. Pode-se ainda representar o conjunto de dados D como uma matriz D de dimensão $n \times (a + 1)$. Considerando um sistema de aprendizagem por regressão, como uma função que faz o mapeamento de um vector de entrada $x_i \in X$ num número real $y \in Y$.

A análise de regressão pretende estimar ou prever o valor médio da variável dependente y , com base nos valores das variáveis dependentes x_i tal que:

$$E(y/x_1, \dots, x_n) \quad (3.1)$$

onde E representa a estimativa estatística.

Como a verdadeira função de regressão é desconhecida, as estimativas são baseadas numa amostra denominada *Conjunto de Treino* ou *Training Set*.

A relação de regressão entre os atributos e a variável objectivo é usualmente representada por:

$$y_i = r(\beta, x_i) + \varepsilon_i \quad (3.2)$$

onde, $r(\beta, x_i)$ é o modelo de regressão para as variáveis de entrada com $x_i, i = (1 \dots a)$; β representa os parâmetros do modelo; ε é o erro.

O objectivo principal deste processo de regressão é procurar os melhores parâmetros β para o modelo, de acordo com um critério de preferência definido. A maioria dos processos de regressão, utiliza o erro de previsão como critério de busca dos parâmetros do modelo. Para a avaliação da precisão dos modelos de regressão existem diversos critérios.

O critério mais utilizado é o *Mean Square Error* [MSE] (Método dos Mínimos Quadrados), dado por:

$$MSE(r) = \frac{\sum_{i=1}^n (Y_i - \hat{y}_i)^2}{n - 2} \quad (3.3)$$

onde Y_i , com $i=1 \dots n$, representa o conjunto de dados; $\hat{y}_i$ - a previsão do modelo de regressão na avaliação do caso y .

3.5.1.2 Métodos Paramétricos Globais

Estes métodos procuram adaptar uma função única para todo o conjunto de treino. Isso implica que sejam assumidas considerações significativas, relativamente à forma da função de regressão desconhecida, podendo condicionar a precisão se estas não forem as mais adequadas. No entanto, estes métodos são adequados para modelos de funções simples, são fáceis de interpretar e permitem obter soluções computacionais rápidas. Um exemplo clássico deste tipo de métodos é a regressão linear, que é habitualmente baseada no critério designado por Método dos Mínimos Quadrados MSE para encontrar o vector de parâmetros β que minimize a soma dos quadrados dos erros,

$$\sum_{i=1}^n (y_i - (\beta_0 + \beta_1 X_1 + \dots + \beta_a X_a))^2 \quad (3.4)$$

A minimização desta equação apresenta a seguinte solução:

$$\beta = (X'X)^{-1} \times X'Y'$$

onde: X' é a matriz transposta de X e X^{-1} é a matriz inversa de X .

Assim, existem inúmeras variantes desta formulação geral, que diferem apenas na forma da adaptação dos modelos aos dados. No entanto, este tipo de regressão permite apenas prever variáveis contínuas. Outro tipo de regressão paramétrica é a regressão logística ou nominal. Esta é uma forma de regressão que permite prever qualquer tipo de variáveis, incluindo variáveis categóricas (ou nominais). Este tipo de regressão é dividida em *regressão logística binária*, se a variável de resposta assumir valores 0 ou 1, o que permite prever se um determinado acontecimento poderá ou não suceder; *regressão logística multinomial* para casos em que a variável de saída possa assumir mais do que duas classes e *regressão logística ordinal* se as classes foram ordenadas. Este último pode ser utilizado para prever o valor de uma variável dependente, com base nas variáveis independentes, permitindo, ainda, determinar a percentagem da variância, na variável dependente, gerada pelas variáveis independentes.

O processo faz uma ordenação das variáveis independentes, de acordo com a sua importância relativa, prevê efeitos de interacção entre as variáveis e permite compreender o impacto da co-variância nas variáveis de controlo.

Neste tipo de modelos, a variável dependente é habitualmente transformada numa variável *logit* (logaritmo natural da probabilidade da variável dependente ocorrer ou não), através das expressões :

$$Y = a + BX + e \quad (3.5)$$

onde Y é uma variável dependente binária com o valor = 1 se o evento acontecer e = 0 se o evento não acontecer; a é o coeficiente do termo constante; B é o vector de coeficientes das variáveis independentes; X é o vector das variáveis independentes; e é o termo de erro.

$$\ln\left(\frac{P}{1-p}\right) = a + BX + e \quad (3.6)$$

p é a probabilidade do evento Y ocorrer em p com ($Y=1$); $\ln\left(\frac{P}{1-p}\right)$ é o *logit* de Y .

A variável *log* é utilizada no cálculo da estimativa da máxima verosimilhança logarítmica *Maximum Likelihood Estimates* [MLE], que permite estimar a probabilidade de ocorrência de um determinado evento.

Neste ponto pode-se encontrar a primeira diferença em relação a utilização de *MSE*, uma vez que as alterações calculadas dizem respeito ao logaritmo da probabilidade da variável dependente e não à variável dependente.

No cálculo do *MSE* pretende-se minimizar a soma dos quadrados das distâncias entre os pontos e a função de regressão enquanto que no cálculo da *MLE* pretende-se maximizar a verosimilhança logarítmica *Log Likelihood* [LL], que reflecte a semelhança entre as estimativas das probabilidades observadas e os valores previstos ao utilizar as variáveis independentes.

A *MLE* é determinada através de um processo iterativo (como *Newton-Raphson*), que pretende maximizar *LL*, partindo de uma solução inicial, que consiste numa estimativa dos coeficientes *logit*. O processo determina as alterações na direcção e no tamanho dos *clusters*, com o aumento da LL até ser obtida a convergência.

3.5.1.3 Modelos Não Paramétricos

A regressão não paramétrica é parte de uma metodologia de análise de dados conhecida como *Modelização Local* [26].

O princípio deste tipo de regressão baseia-se na obtenção de uma previsão para uma variável x , ao ajustar uma função paramétrica na vizinhança de X . Deste modo, a parametrização é feita localmente e não na totalidade do conjunto de dados. Pode-se dizer que o objectivo geral da regressão local é ajustar uma função de grau p em torno de um caso de teste X_s , ao utilizar o conjunto de treino na sua vizinhança. Por exemplo, se $p = 0$ equivale a uma regressão *kernel* e se $p = 1$ existe uma regressão linear local. Estes métodos apresentam dificuldades em domínios de previsão de elevada dimensão (elevado número de atributos),

devido a esparsidade que apresentam, o que dificulta a definição da vizinhança. Têm ainda a desvantagem de não permitirem obter um modelo compreensível do conjunto de treino.

3.5.1.4 Modelos Aditivos

A ideia base deste tipo de modelos é tirar partido da possibilidade de decompor funções de regressão complexas de uma forma aditiva, em diversas funções simples. Os modelos aditivos consistem na soma de um conjunto de funções de pequena dimensão.

Um modelo aditivo é definido por:

$$r(x) = \alpha + \sum_{j=1}^{\alpha} f_j(X_j), \text{ onde } f_j \text{ é uma função univariável para cada atributo.}$$

Estes modelos podem ser generalizados para funções de vários atributos ou para atributos discretos. Como exemplos de modelos aditivos frequentemente utilizados têm-se as *árvores de Decisão* [27] e a *Multivariate Adaptive Regression Splines* [MARS] [28]. Estes surgem para responder às dificuldades dos modelos locais com espaço de previsão de dimensão elevada. A sua maior limitação é a elevada complexidade computacional.

3.5.2 Algoritmos de Vizinhança mais Próxima

Os algoritmos de vizinhança mais próxima *K-nearest neighbour* [KNN] são técnicas não paramétricas utilizadas em operações de classificação.

Tal como outras técnicas de classificação, estes algoritmos necessitam de uma variável objectivo, na qual se baseia a sua classificação. Cada registo é classificado de acordo com o valor esperado dos registos mais similares ao mesmo. O valor k representa o número de registos de vizinhança utilizados para classificar o novo registo.

Quando $k = 1$ o método é denominado apenas por classificador de vizinhança mais próxima *Nearest Neighbour Classifier* [NNC].

Este método de classificação é muito utilizado em *DM*, sendo definido para o tratamento de dados, não implica suposições em relação aos mesmos e permite criar fronteiras de decisão complexas. A vizinhança mais próxima é definida ao utilizar uma medida de distância K , tal como nos algoritmos de *clustering*. Quanto maior for o valor de k mais exacta é a classificação mas maior é o tempo de computação.

A principal vantagem destes algoritmos digna de referência é a elevada eficiência na classificação de conjuntos de dados de tamanho médio, para valores pequenos de k . Como principais desvantagens, refere-se o elevado esforço computacional, principalmente se k for elevado; a susceptibilidade de criação de ruído nos dados e consequentemente, a incapacidade para escalar devidamente o espaço dos dados.

Outra limitação, ao contrário das *árvores de Decisão* e *Redes Neurais*, é a necessidade de utilizar todos os dados históricos para classificar *novos dados* ao estabelecer quase sempre comparações directas entre os registos. No entanto, desenvolvimentos científicos

recentes, propõem soluções para a resolução das limitações desta técnica, relativas à dimensão da base de dados.

3.5.3 Redes Neurais

As Redes Neurais estão entre as técnicas mais utilizadas para implementar operações de *DM*. Apesar de não proporcionarem a compreensibilidade das *árvores de Decisão* ou das regras de indução, são mais eficientes no tratamento de problemas complexos. Originárias há cerca de 50 anos, num trabalho apresentado por McCulloch e Pitts, em que propunha um modelo abstracto e simplificado de células nervosas e neurónios.

Em *DM*, as *Redes Neurais* são essencialmente utilizadas em operações de classificação, *clustering* e aproximação de funções. Estas são normalmente adequadas para problemas que envolvam a análise de interações complexas entre variáveis de entrada.

As *Redes Neurais*, tal como o cérebro humano, são constituídas por uma série de nós interligados, em que as interligações entre os mesmos vão determinar a capacidade da rede.

Uma *Rede Neuronal* é constituída por um número de unidades de processamento elementares, chamadas *nós*, neurónios ou simplesmente unidades. As ligações entre eles designam-se por sinapses, estando associados pesos às sinapses. Assim, uma *Rede Neuronal* é um conjunto de nós interligados, formando redes multi-camadas ou redes de nível simples, os quais podem ser bidireccionais ou unidireccionais. Para cada tipo de topologia é desenvolvido um processo de aprendizagem para determinar o valor dos pesos (*weights*) das ligações. Um Nó recebe um conjunto de entradas X , cada entrada está ligada por uma sinapse, à qual é atribuído um certo valor denominado peso (*weight*). Existe ainda uma entrada de *bias* WO (polarização) que corresponde a uma entrada permanentemente activada. Se o peso da ligação tiver um valor positivo, diz-se que existe uma ligação *excitadora*, se o peso for negativo existe uma ligação *inibidora*. Num nó são realizadas duas operações, um somatório e o cálculo com uma função normalmente não linear do tipo sigmóide dando lugar aos valores de saída.

A operação da equação 3.7 formaliza a operação *somatório*, calculando a soma pesada de todas as entradas e da *bias*, para um determinado neurónio.

$$nnet = \sum_{i=0}^k w_i x_i \quad (3.7)$$

O cálculo dos valores de saída é realizado através de um mapeamento da função linear de soma pesada com uma função de activação: $y = f_{(net)}$.

Quando a aprendizagem é calculada ao aplicar as entradas a um conjunto de treino, para o qual o valor das saídas é conhecido, existe um processo de aprendizagem supervisionada. Se pelo contrário, o valor das saídas para o conjunto de treino não for especificado, existe um processo de aprendizagem automático, mas não supervisionado. A maioria dos modelos

utiliza processos de aprendizagem supervisionada ou não supervisionada, existindo poucos modelos que utilizam processos combinados.

Neste trabalho são descritos apenas os mais comuns em aplicações de *DM*:

- *Single Layer Perceptron* (SLP),
- *Multi Layer Perceptron* (MLP),
- *Self Organizing Maps* (SOM),
- Modelos com Aprendizagem Combinada,
- *Recurrent Networks*,
- Funções de base radial,
- *Modular Networks*,
- *Learning Vector Quantization* (LVQ),

As secções seguintes descrevem com mais detalhe apenas os casos SLP e MLP, dado terem sido apenas estes os utilizados no trabalho em curso.

3.5.3.1 Single Layer Perceptron

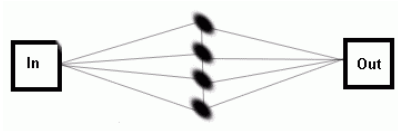


Figura 3.1: Single Layer Perceptron

Um *Single Layer Perceptron* SLP é constituído por um ou mais nós de entrada apenas com uma só camada. Esta é denominada camada escondida ou *hidden layer*, na qual cada nó pode estar ligado a todos entre si ou apenas a alguns nós. No entanto, apenas possui uma única camada entre a entrada e a saída.

O processo de treino é baseado num procedimento iterativo do tipo *correção-erro*, em que cada nó ajusta os seus pesos, cada vez que analisa o erro de classificação e aproximação no conjunto de treino. Este processo de ajustamento dos pesos prossegue recursivamente, até que cada nó seja capaz de fornecer um valor suficiente próximo do valor esperado.

3.5.3.2 Multi Layer Perceptron

O modelo MLP é vulgarmente o mais utilizado. Este modelo consiste na existência de 2 ou mais níveis (camadas) de nós interligados, como se apresenta na 3.2. Geralmente, os nós de um nível estão ligados a todos os nós dos níveis adjacentes por ligações unidireccionais. Existe o nível de entrada e o nível de saída e os níveis entre estes são designados por intermédios ou escondidos *hidden levels*. Uma rede *MLP* pode ter múltiplas camadas na zona escondida ou *hidden*.

O processo da regra Delta, consiste num procedimento de minimização de erro, que procura determinar os pesos de cada nó ao utilizar um método de conjugação de gradiente. Consequentemente, a *regra Delta* não apresenta problemas de convergência, mas é limitada, com o risco de convergir para um mínimo local, e não encontrar o mínimo global. O processo de aprendizagem *Delta* é muito aplicado em problemas de regressão e de classificação.

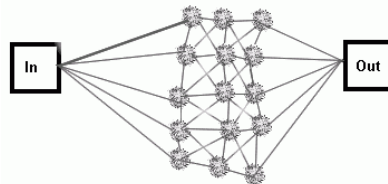


Figura 3.2: Multi Layer Perceptron MLP

Com a topologia de rede MLP, de múltiplos níveis, tem a capacidade de produzir uma relação complexa não linear entre a entrada e a saída. O *treino* neste tipo de redes é feito através de um processo de minimização de erro, constituindo uma generalização da regra *Delta*.

No entanto, a flexibilidade da rede implica uma afinação desta para propriedades muito próximas das propriedades do conjunto de dados treino. Por este motivo não são adequadas para analisar conjuntos de dados simples com poucos registos e é necessário tomar algumas medidas para evitar *Over Fitting* - *sobre ajustamento*.

Pode-se recorrer a um segundo conjunto de dados para proceder à validação e realização de testes, evitando que a rede se ajuste a apenas um conjunto de dados. O treino pára quando o desempenho do modelo com dados de treino e de teste atingir um nível aceitável de desempenho. Este nível, corresponde a uma taxa mínima de erro face ao conjunto de dados de teste.

3.5.3.3 Aplicação de Redes Neurais

A aplicação de uma *Rede Neuronal* deve ser acompanhada dos seguintes pontos:

- selecção do modelo,
- codificação (normalização) das entradas e das saídas,
- definição da taxa de aprendizagem,
- treino,
- avaliação de desempenho.

Em função da implementação pode ser necessário proceder à codificação (normalização) das entradas e das saídas, pois estas podem ser limitadas aos valores pertencentes aos intervalos $[0, 1]$ e ou $[-1, 1]$.

Outro factor importante é a selecção da taxa de aprendizagem, porque esta determina a amplitude das variações dos pesos em cada passo do treino. Uma taxa de aprendizagem desajustada, muito baixa ou muito elevada, pode levar a um tempo de treino muito elevado, ou a uma performance de previsão inferior ao óptimo esperado. Relativamente à definição da taxa de desempenho existem algoritmos que a ajustam automaticamente.

3.5.4 Partições Hierárquicas e Árvores de Decisão

Uma partição hierárquica pode ser facilmente representada por uma árvore, em que a raiz representa o espaço de previsão (domínio) total, cada ramo representa a partição exaustiva e as folhas representam as sub-partições. Os ramos marcam-se com as condições de partição dos atributos. Se na árvore cada ramo tiver apenas um atributo é chamada *Árvore de Decisão*. Se cada partição for dividida apenas em 2 denomina-se como árvore binária. Uma árvore de decisão disponibiliza uma representação gráfica de como os dados foram classificados, apresentando o conjunto de regras de classificação desses mesmos dados. Assim os nós terminais da árvore são denominados nós folha. Existe um processo onde é também possível gerar regras, sem construir a própria árvore, sendo designado genericamente por *Indução de Regras*.

Uma metodologia adaptativa altera a sua estratégia de acordo com a natureza local dos dados, permitindo ao modelo ajustar-se aos mesmos de acordo com as suas características locais. Assim, o modelo pode apresentar maior especificidade, quando os dados são mais densos ou mais heterogéneos e ser mais suave quando os dados apresentam ruído.

Os métodos adaptativos utilizam habitualmente estratégias de partição hierárquica recursiva, também conhecidas como estratégias *divide and conquer*. O seu princípio baseia-se na subdivisão recursiva da *população* em subconjuntos (partições) exclusivos, de modo a que os registos incluídos numa partição, segundo um determinado critério, sejam similares e, os registos em partições diferentes, sejam dissimilares. O processo é repetido para cada partição obtida, mas o tipo de algoritmo utilizado pode variar de acordo com o critério local, baseado nas características dos dados da mesma.

A decisão de quando parar a subdivisão de cada partição é também baseada num critério local. Por exemplo, *parar* quando os registos contidos numa partição são suficientemente similares. Quando os dados são mais densos ou heterogéneos, ocorrem mais partições e se os dados forem esparsos ou homogéneos ocorrem menos partições, o que leva a que as partições finais tenham o mesmo grau de homogeneidade.

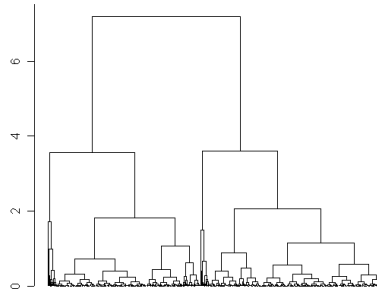


Figura 3.3: Representação de uma Árvore de Decisão

Um algoritmo de construção de uma *Árvore de Decisão* funciona dividindo o conjunto de dados de treino, de modo a construir o modelo que classifica cada registo, de acordo com o valor apresentado no atributo objectivo. Os algoritmos de construção mais utilizados são os que constroem de *cima para baixo*, também conhecidos por *partição geral para específica*. A estratégia é cortar o conjunto original, dividindo-o recursivamente em subconjuntos, até satisfazer o critério de paragem.

Pode-se dividir o processo de construção da árvore nas 4 fases chave:

- Definição do critério a utilizar na construção;
- Qual o critério para parar o crescimento;
- Como cortar ou podar de modo a obter melhor resultados;
- Gerar um método para criar uma regra de decisão em cada partição final.

Ao definir o critério de construção é necessário determinar as variáveis a utilizar e que partições fazer em cada sub conjunto, como é seleccionado um atributo, a partir do qual é elaborada a divisão do conjunto de dados e que condição deve ser utilizada para definir essa partição.

Uma *Árvore de Decisão* tem como objectivo encontrar partições no espaço de previsão, de modo a que os registos incluídos numa partição segundo um determinado critério, sejam similares e os registos em partições diferentes sejam dissimilares.

No caso da árvore de regressão, o objectivo é procurar partições, tais que, em cada nó se encontrem registos cujas respostas se aproximem de um modelo de regressão.

Nas *árvores de classificação*, os nós folha devem conter registos que pertençam à mesma classe, com o mínimo de excepções possível. No caso de se utilizar árvores para *clustering*, cada nó deve conter registos, cujas respostas multi-variadas sejam suficientemente próximas, de acordo com uma medida de distância definida. As árvores apresentam algumas limitações: são pouco flexíveis comparativamente com outras técnicas, no que se refere à repetição para treinos. Se os dados históricos, a partir dos quais a árvore foi construída se alterarem, esta terá de ser novamente construída.

Comparativamente, nas *Redes Neurais*, o reajuste pode ser baseado apenas nos registos mais recentes. O agrupamento dos dados pode ainda inadvertidamente esconder relações existentes entre eles, pelo que as *Redes Neurais* são mais eficientes na criação de modelos efectivos e práticos.

3.5.5 Algoritmos Genéticos

Os *Genetic Algorithms* [GA] pertencem a uma área mais vasta que é a programação evolucionária ao aplicar conceitos da selecção natural e evolução (baseada na evolução das espécies).

Um algoritmo genético manipula uma população de potenciais soluções de optimização para resolver problemas de procura. Cada solução possível é codificada numa *string* de bits binária e é denominada *cromossoma* ou indivíduo, à qual é associada uma função de avaliação que mede a sua qualidade como solução candidata. O processo de procura inicia-se com uma selecção aleatória de soluções possíveis, que é a *Gene Pool* ou População Inicial.

Em cada iteração da procura, o algoritmo analisa uma nova geração (população), criada a partir de operadores genéticos. Tal como no processo de evolução natural das espécies, os membros das sucessivas populações apresentam valores de ajustamento tendencialmente melhores. Após várias gerações de alterações genéticas é encontrada a solução óptima entre as soluções sobreviventes. A aplicação de *GA* a um problema requer a consideração de diversos factores tais como: o esquema de codificação, a escolha da função de ajustamento, a selecção dos indivíduos a cruzar e os operadores genéticos.

Os operadores genéticos utilizados para fazer evoluir as populações são:

Cruzamento consiste na troca de 2 secções entre 2 strings de cromossomas, ao imitar a possibilidade de combinações que podem ser introduzidas por reprodução;

Mutação consiste na selecção aleatória de um bit numa string de cromossomas, alterando-o com uma probabilidade próxima ou igual a uma determinada taxa de mutação.

Este último operador é necessário, porque ao utilizar apenas a operação de *Cruzamento*, pode não ser encontrada uma solução satisfatória. Sem a utilização da mutação existe o risco de paragem num mínimo local. A taxa de mutação deve ser mantida relativamente baixa, de modo a evitar um efeito aleatório não controlável.

Em adição a estes operadores, pode ainda ser utilizado o *Best of the Best*, no qual se

utilizam apenas os melhores membros da população para o cruzamento. Em aplicações de *DM* um cromossoma é codificado numa string de k bits. O valor de k depende dos valores possíveis de cada atributo. No caso das *árvores de Decisão* [29] nas quais é utilizado para determinar as partições dos dados.

Os algoritmos genéticos são adequados para problemas em que se pretenda encontrar a solução óptima num espaço de procura delimitado, com dados relativamente uniformes, pois estes têm de ser codificados em vectores com a mesma dimensão. Têm como pontos fortes a aplicabilidade a problemas de optimização contínuos ou discretos; são menos susceptíveis a paragens em mínimos locais e podem ser utilizados em aplicações de processamento paralelo. Como pontos fracos, a sua aplicação é ainda possível apenas a problemas de reduzida dimensão e devido ao elevado esforço computacional que é exigido comparativamente a outros métodos, tornam-se parcialmente impraticáveis.

3.5.6 Fuzzy Sets

Estes diferem dos conjuntos tradicionais pela forma como são definidos os elementos que os compõem. Num conjunto convencional, um objecto é classificado como pertencente ou não a esse conjunto. Num conjunto difuso (*fuzzy set*), são permitidos vários graus de pertença de um objecto a um conjunto.

Como exemplo no caso do conceito de *dia frio* não é definido dessa forma, pois se considerarmos que 10 graus de temperatura máxima implicam um dia frio, se esta temperatura atingir os 11 graus, não existe um dia quente, mas ainda há a interpretação pessoal de dia frio que não é a mesma para todos.

Para ultrapassar esta limitação, existem os *Fuzzy Sets* que permitem transformar a expressão *Hoje está um dia frio...*, numa noção mais alargada que inclui a imprecisão deste tipo de informação.

Operacionalmente, esta noção é representada por um grau de pertença entre $[0...1]$ e o conjunto inicial perde o valor original [30].

Como exemplo, na figura 3.4 pode observar-se o resultado da intersecção de diversos grupos, nos quais existem pontos individuais que têm graus de pertença diferentes para cada grupo.

Outra aplicação cada vez mais utilizada é a dos sistemas *Neuro Fuzzy*, em que são criadas aplicações híbridas com *Redes Neurais* [31].

Nestes sistemas, as regras *fuzzy* podem ser utilizadas para representar conhecimento inicial sobre o domínio, que é convertido numa *Rede Neuronal*, permitindo passar para a rede neuronal sem ser necessário treiná-la. Para *DM* a principal limitação é a definição das condições *Fuzzy*, sendo necessário a ajuda de um perito na matéria.

A *Lógica Difusa Fuzzy Logic* foi introduzida por Lotfi Zadeh em 1965 e, desde então, tem sido aplicada em numerosos problemas, permitindo ultrapassar algumas limitações dos sistemas lógicos tradicionais. Como objectivo principal visa a introdução de conceitos

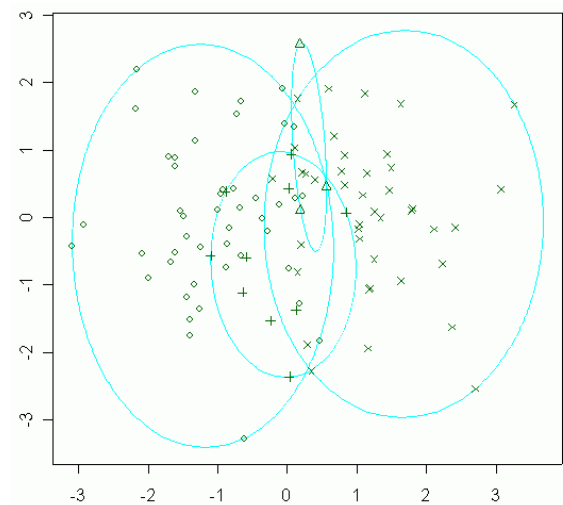


Figura 3.4: Representação Fuzzy: com pontos em vários clusters

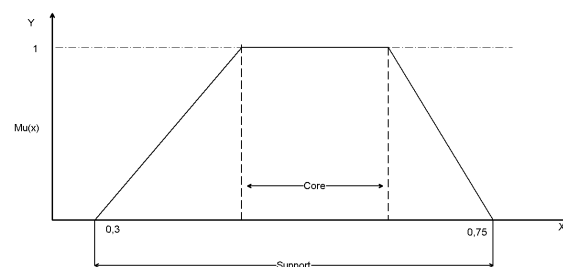


Figura 3.5: Representação Fuzzy de um ponto

frequentemente utilizados que são vagos ou incertos, tais como o peso, a altura ou a temperatura num sistema formal. Ao considerar a noção de *pessoa alta*, essa é considerada uma pessoa com altura superior a 1,90m. Numa abordagem *Fuzzy*, o conceito de *pessoa alta* inclui a totalidade das pessoas, com diferentes graus de pertença, entre as alturas pré-definidas. A maior vantagem da *Fuzzy Logic* é permitir a criação de regras de inferência, para classificação, expressas de uma forma mais próxima da realidade, permitindo trabalhar com dados que contenham informação imprecisa.

Uma das suas aplicações mais interessantes é a problemas de *clustering*, pois a noção de *Fuzzy Clustering* permite atribuir a cada objecto um grau de pertença a cada *cluster* [32].

3.5.7 Rough Sets

A teoria dos *Rough Sets* - Conjuntos Irregulares foi introduzida por Pawlak em meados dos anos 80 [33] e resulta de um longo programa de pesquisa das propriedades lógicas fundamentais dos Sistemas de Informação. Baseia-se na ideia de que ao baixar o grau de precisão dos dados, torna as regularidades nos dados mais fáceis de encontrar e, consequentemente, de caracterizar em termos de regras. No entanto, esta diminuição de precisão tem riscos que podem levar à perda de informação e à diferenciação entre os diferentes objectos.

A teoria dos *Rough Sets* fornece ferramentas para lidar com este tipo de problema, ao deixar a irregularidade variar de uma forma controlada, para encontrar a precisão nos dados que assegura a manutenção de um grau de diferenciação suficiente para a distinção de padrões. Esta última é diversas vezes confundida com a teoria de *Fuzzy Sets* pois ambas lidam com imprecisão nos dados. A diferença entre elas é que nos *Fuzzy Sets* a imprecisão é relativa a objectos da mesma classe e nos *Rough Sets* esta imprecisão existe entre objectos de classes diferentes. Deste modo, existem diversos trabalhos que descrevem aplicações híbridas que combinam as 2 teorias, em que a teoria dos conjuntos irregulares *Rough Sets* é utilizada no pré-processamento para aplicação aos conjuntos imprecisos [34].

Uma das vantagens relativamente a outras teorias é que esta não requer qualquer conhecimento prévio acerca dos dados ou da independência de atributos, como a probabilidade estatística ou o grau de pertença. Outra vantagem é a simplicidade da construção dos seus algoritmos. Como principal aplicação desta teoria temos as operações de redução de atributos [35], a geração de regras de classificação [36] e os problemas de previsão [37].

3.6 Qual a Técnica a Aplicar ?

Após a apresentação das técnicas de *DM* mais comuns, surge naturalmente a questão de qual a melhor técnica para realizar uma determinada tarefa de *Data Mining*. Não existe uma resposta simples e directa a esta questão, pois não existe uma única técnica que se possa apresentar como melhor do que as outras para fazer face a todos os problemas.

Ao ser possível analisar cada técnica de forma comparativa, de acordo com o seu desempenho, com diferentes tipos de conjuntos de dados de entrada, é possível classificá-las relativamente a um conjunto de índices. Estes índices representam características a ter em atenção durante a selecção de uma técnica de *DM*.

Os factores utilizados são:

- *Scalability* representa a capacidade da técnica para lidar com um elevado número de registos;
- *Dimensionality*, indica qual a capacidade que cada técnica possui para lidar com registos de grande dimensão;

- Capacidade para tratar diferentes tipos de atributos (numéricos, discretos);
- *Comprehension Factor* indica a compreensibilidade para o utilizador final do conhecimento descoberto pela técnica;
- *CPU Time* - factor relacionado com o *trabalho* computacional na aprendizagem e na execução;
- *RAM* - Capacidade de memória necessária, factor relacionado com o esforço computacional na aprendizagem da execução.

Cada técnica possui características específicas, que devem ser consideradas e adaptadas às diferentes aplicações, para poder encontrar a que melhor se aplica ao problema em questão. A escolha de um determinado algoritmo depende tanto do tipo de dados disponível como do objectivo final. Por vezes, vários algoritmos são aplicáveis e mesmo argumentos *à priori* podem não ser suficientes para diminuir o número de algoritmos a testar e para obter a informação importante. Assim, deste modo não é possível responder imediatamente à pergunta inicial de uma forma *cega* e categórica.

Capítulo 4

Clustering

Este capítulo apresenta algumas das técnicas mais utilizadas em algoritmos de *clustering* e faz uma breve descrição das diferentes fases do processo. O processo de *clustering* é uma das técnicas de *DM* referidas no capítulo anterior, a qual foi necessária para agrupar os dados com características semelhantes.

No caso do projecto CCR [1] o clustering foi utilizado para obter perfis típicos que caracterizassem os diversos tipos de consumidores [BT].

No entanto nos dados dos consumidores [MT] a técnica de *clustering* foi utilizada na remoção de *outliers*.

Segue-se uma classificação dos diferentes algoritmos abordados no *State of the Art* e por fim é apresentada uma descrição mais detalhada dos algoritmos utilizados neste trabalho.

O *clustering* consiste na divisão de um conjunto de dados em grupos. Cada grupo é denominado *cluster* e inclui objectos similares entre si e dissimilares dos objectos incluídos noutros grupos. Estes algoritmos permitem representar um elevado conjunto de objectos, através de um conjunto restrito de *clusters*, encontrados no conjunto dos dados. Ao contrário da classificação, o *clustering* não necessita da identificação de uma variável objectivo, pelo que também pode ser designado por *aprendizagem não supervisionada*.

Alguns dos métodos de *clustering* baseiam-se na medição da similaridade (*Distância* entre registos) e é da similaridade entre *clusters* para que a cada registo seja atribuído a um *cluster* de forma a minimizar a distância entre registos pertencentes ao mesmo *cluster*.

Um algoritmo de *clustering* é caracterizado pelo tipo de medida de similaridade e pelo método utilizado para proceder ao agrupamento dos dados. A diversidade de técnicas de representação dos dados, de tipos de medidas de similaridade entre os elementos dos dados e de métodos de agrupamento dos elementos dos dados deu origem a um vastíssimo conjunto de algoritmos de *clustering*.

Num problema de *clustering* convencional todas as variáveis podem ser vistas como objectos de previsão e como objectos de resposta e os registos (ou instâncias) podem ser

designados como padrões. Num problema de *clustering* conceptual, existem objectos de previsão diferentes dos objectos de resposta e os registos são utilizados como variáveis de resposta de cada caso. Neste caso, o objectivo é também encontrar *clusters* de padrões similares, mas as regras que indicam em que *cluster* deve ser incluído cada caso dependem apenas das variáveis de previsão, ou seja, são extrínsecos relativamente aos registos. Deste modo, no processo de agrupamento dos registos, são utilizadas apenas as distâncias entre eles, em vez dos valores próprios dos atributos. Um bom método de clustering deve ser capaz de produzir clusters de alta qualidade com elevada similaridade de entre classes. A qualidade do resultado do clustering depende tanto da medida de similaridade utilizada pelo método como da sua implementação. A qualidade também pode ser medida pela habilidade de se descobrir alguns padrões não directamente visíveis ou escondidos.

4.1 Fases do Processo

4.1.1 Extracção de Atributos

Esta fase está relacionada com o número de classes, o número de registos disponíveis e o número, tipo e escala dos atributos disponíveis para o algoritmo de *clustering*. Por vezes, esta informação não é possível ser analisada de uma forma directa e torna-se necessário implementar processos de selecção e extracção de atributos. Um processo de selecção de atributos pretende identificar a partir do conjunto total de atributos, qual o subconjunto mais indicado para utilizar no processo de *clustering*, de forma a torna-lo mais eficiente e económico. Um processo de extracção de atributos consiste na utilização de um conjunto seleccionado dos atributos possíveis ou de uma ou mais transformações dos atributos iniciais, para obter novos atributos que sejam mais apropriados para a implementação do processo de *clustering*.

4.1.2 Avaliação de Similaridade entre registos

Similaridade mede a distância entre duas observações ou seja mede a proximidade entre casos. Casos que são semelhantes têm similaridade elevada implicando que têm uma distância entre si baixa. A correlação de itens pode ser utilizada como medida de similaridade.

A sua medição é usualmente realizada através de uma função de medida de similaridade *Distância*, definida entre pares de registos. O primeiro passo em *clustering* é a definição da similaridade ou matriz de distâncias.

Esta matriz é uma tabela em que os valores de cada célula são a medida da similaridade ou distância entre cada par de casos.

A distancia euclidiana é a medida de distância mais comum. Um par de casos é definido por 2 variáveis que formam os eixos XX e YY . A distância euclidiana é a raiz quadrada

da soma do quadrado da diferença de x mais a diferença do quadrado da distância de y .

Quando existem várias variáveis, a maior em valor absoluto pode *dominar* as restantes pelo que para evitar esse problema, é necessário normalizar-las. Como aplicação no cap. 2.7 existe uma grande variedade de funções de medidas de similaridade e que devem ser seleccionadas de acordo com o problema a tratar.

Binary matching é outro tipo de medida de similaridade, onde 1 indica *correlação* e 0 indica *sem correlação* entre cada par de casos. Podem existir diversas combinações de atributos onde o *similarity score* é o número de correlações dividido pelo número de atributos a relacionar. No *clustering* se for utilizado um algoritmo hierárquico, são realizadas diversas séries de partições, baseadas num critério de agregação ou separação, para juntar ou dividir os *clusters* conforme a medida de similaridade.

No caso dos algoritmos particionais é identificada uma partição que optimiza o critério de *clustering*. O resultado desta fase é a partição dos dados em grupos, ou *clusters* que podem ser rígidos ou difusos. No primeiro caso cada registo é atribuído a um *cluster*, a que passa a pertencer, enquanto que no segundo caso, esse registo tem um determinado grau de pertença a cada um dos *clusters* obtidos.

4.1.3 Avaliação dos resultados

Qualquer algoritmo de *clustering*, quando aplicado a um conjunto de dados, cria *clusters* com esses dados quer o conjunto solução apresentado contenha *clusters* interessantes, ou não. Mesmo quando o conjunto de dados possui *clusters* interessantes, alguns algoritmos podem obter melhores resultados do que outros. Deste modo é importante criar procedimentos que permitam avaliar os resultados do processo de *clustering*.

Como exemplo desse processo de avaliação de resultados poderemos utilizar diversas abordagens tais como:

- Distância média entre padrões do mesmo cluster;
- Distância entre os *centroids* dos diferentes clusters;
- Distância mínima entre todos os elementos do mesmo cluster;
- Distância máxima entre elementos dos *edges* de clusters diferentes;
- entre outras.

Destas abordagens a mais utilizada é a verificação da homogeneidade dos *clusters* através da análise da média das distâncias dos elementos do *cluster* ao *centroid*.

4.2 Classificação de Clustering

Uma definição largamente adoptada para um *clustering óptimo* é uma partição que minimize as distâncias entre os elementos de um *cluster* e maximize a distância entre *clusters* diferentes. No entanto, esta definição permite muitas variações, uma vez que estas distâncias podem ser definidas de diversas formas.

Em Berhkin [38] é possível encontrar uma proposta de classificação muito completa, ao considerar os algoritmos existentes divididos nos seguintes subgrupos:

- Algoritmos de Partição;
- Algoritmos Hierárquicos;
- Algoritmos Aglomerativos;
- Algoritmos Divisivos;
- *Density based algorithms*;
- *Grid based algorithms*;
- Algoritmos baseados em Modelos;
- *Clustering* baseado em restrições;

Estes algoritmos podem ser aplicados a diversos tipos de problemas, apresentando alguns deles características que os tornam mais indicados para determinado tipo de problemas de acordo com as suas particularidades. Os métodos vulgarmente mais utilizados para realizar *clustering* são os métodos Hierárquicos e os métodos de Partição.

Os primeiros conseguem a segmentação através de uma sequência interligada de partições, que podem ser representadas numa estrutura em árvore conhecida como *dendrograma*. Por vezes, para resolver um determinado problema é necessário recorrer a mais do que um algoritmo, e comparar os resultados obtidos com cada um deles. O clustering divide um conjunto em grupos (*clusters*) de observações que são similares entre si.

Em algoritmos Hierárquicos como *AGNES*, *DIANA*, *MONA* é construída uma hierarquia de conjuntos, com o número de clusters a ir de 1 até ao número total de observações.

Nos algoritmos de partição como *PAM*, *CLARA*, é necessário especificar o número de *clusters à priori* como objectivo fornecido pelo utilizador.

4.2.1 Algoritmos Hierárquicos

Estes conseguem a segmentação através de uma sequência interligada de partições, que podem ser representadas numa estrutura em árvore.

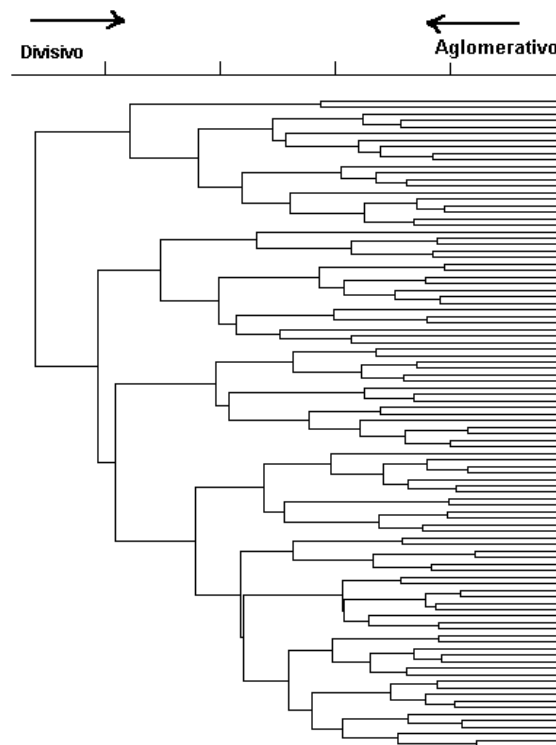


Figura 4.1: Diferença dos processos hierárquicos

Os algoritmos hierárquicos podem ainda ser divididos em 2 tipos: Aglomerativos e Divisivos.

Um processo aglomerativo começa tipicamente com vários *clusters*, que contém apenas um registo, congregando de uma forma recursiva, 2 ou mais *clusters*, segundo uma medida de similaridade. Um processo divisivo, começa apenas com um *cluster* que contém todos os registos, é divide-o recursivamente em novos *clusters*, segundo uma medida de similaridade. Em ambos os processos, a sequência repete-se até que um critério de *paragem* seja atingido. Este critério de *paragem* é frequentemente um número predefinido de *clusters*. Um algoritmo hierárquico fornece um dendrograma, ao representar as divisões sucessivas e os níveis de similaridade em cada partição.

Podendo criar cortes no dendrograma, para um determinado grau de similaridade, é possível obter um número de *clusters* diferente, para o mesmo conjunto de dados. Dependendo da forma como é actualizada a matriz de proximidade no segundo passo, é possível encontrar uma grande variedade de algoritmos dentro dos métodos aglomerativos.

A maioria dos algoritmos de *clustering* hierárquico é baseada em variações dos algoritmos de distância mínima (ver Sneath [39]), variância mínima (ver Ward [40]) e distância

máxima. Destes os mais populares são os dois primeiros, que diferem entre si na forma como é caracterizada a similaridade entre dois *clusters*. Nos algoritmos de ligação mínima, a distância entre 2 *clusters* é a menor distância existente entre 2 registos de 2 *clusters* diferentes. No caso dos algoritmos de ligação máxima a distância entre dois *clusters* é a máxima distância encontrada, entre dois registos de *clusters* diferentes. Em ambos os casos, o objectivo é aglomerar os *clusters*, com base num critério de distância mínima.

Na figura 4.2 é visualizada a aplicação deste tipo de *clustering* ao projecto CCR [1] nos consumidores [MT], apesar de se visualizar a árvore completa na prática esta foi *cortada* na altura 4 de modo a obter um conjunto *tratável* de *clusters*.

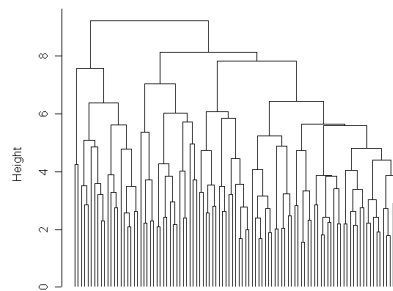


Figura 4.2: Dendrograma exemplo de dados [MT] do projecto CCR

Este procedimento é representado através da seguinte estrutura:

1. Calcular a matriz de similaridade ($n \times n$), que contém as distâncias entre cada par de registos;
2. Tratar cada registo como um *cluster*;
3. Procurar o par de registos mais similar ao utilizar a matriz de similaridade mais actual;
4. Aglomerar os 2 *clusters* mais próximos num só;
5. Recalcular a matriz de proximidade para considerar a operação de aglomeração. Se todos os registos estiverem num só *cluster* ou tiver sido atingido o critério de paragem parar;
6. Se não regressar ao segundo passo.

Este tipo de algoritmos apresenta algumas vantagens: não necessita de definição prévia do número de *clusters*; tem facilidade em utilizar diferentes tipos de medidas de distância de similaridade; aplica-se a qualquer tipo de atributos.

As principais desvantagens deste tipo de algoritmo são:

- o critério de paragem é difuso, se o número de cluster objectivo for mal definido;
- a maioria dos algoritmos não permite voltar aos níveis intermédios para introduzir melhoramentos nos *clusters* obtidos;
- é de difícil aplicação a grandes conjuntos de dados, nos quais poderá ser complicado construir um dendrograma.

4.2.1.1 AGNES, Agglomerative *NESting*

Este algoritmo calcula através de um processo aglomerativo os *clusters* executando uma sequência de fusões.

Num primeiro passo os objectos mais próximos (com a dissimilaridade entre objectos mais baixa) são agregados, ficando $n-1$ clusters, um dos quais possui dois objectos onde os outros possuem apenas um só elemento. Nos passos sucessivos os dois clusters mais próximos voltam a ser unidos.

Comparando com outros métodos aglomerativos este tem as seguintes propriedades:

- utiliza um coeficiente aglomerativo que mede o montante da estrutura de clustering encontrada;
- e além da árvore usual mostra o *Banner*, um formato visual, onde se pode verificar visualmente o número de casos que pertence a cada cluster.

As folhas da árvore de *clustering* são as observações originais. Os dois ramos são unidos na distância entre a junção de dos *clusters* a serem unidos.

O *Banner* mostra a hierarquia dos *clusters* e é equivalente a uma árvore, este formato mostra as distâncias nas quais os registos e os *clusters* são agrupados. Os registos são listados na ordem encontrada pelo algoritmo *AGNES* e os números no vector de alturas são representados pelas barras entre as observações. As folhas da árvore de *clustering* são as observações originais, os dois ramos são juntos na distância entre a junção dos dois *clusters* reunidos. (ver 4.3)

Na figura 4.3 é possível ver a ordem dos clusters encontrada pelo algoritmo *AGNES* num pequeno conjunto de teste dos consumidores [MT] do projecto CCR [1].

Este algoritmo constrói uma hierarquia de *clusters*, em que inicialmente cada observação é um pequeno *cluster* por si só. Os *clusters* são agregados até apenas restar apenas um só grande *cluster* que contém todas as observações. Em cada fase os *clusters* mais próximos são combinados e passam a formar um só grande *cluster*.

Deste modo existem várias opções para a construção dos *clusters* diferindo no cálculo da distância entre elementos: *Average*; *Single*; *Complete*; *Weighted*; *Flexible*.

Com a opção *Average* a distância entre 2 clusters é a *média* das dissimilaridades entre os pontos no primeiro cluster e os pontos no segundo. Com esta opção encontram-se

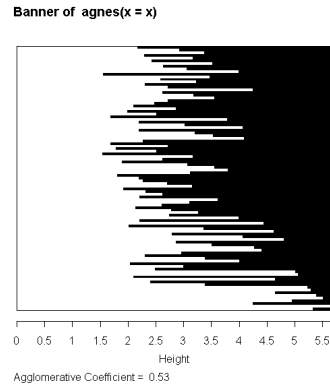


Figura 4.3: AGNES Banner

clusters com forma de bola (*ball shaped*). Na opção *Single* é utilizada a dissimilaridade mais pequena entre um ponto no primeiro *cluster* e um ponto no segundo equivalente ao método do vizinho mais próximo (*nearest neighbour method*). Esta opção é capaz de lidar com *clusters* com forma de batata (*potato form*).

Com a opção *Complete* é utilizada a maior dissimilaridade entre um ponto no primeiro cluster e um ponto no segundo *cluster*, equivalente ao método do vizinho mais afastado (*furthest neighbour method*). Esta abordagem tende a produzir clusters muito compactos com pequeno diâmetro. Assim, cada membro desse tipo de cluster deve estar próximo de qualquer outro membro do mesmo cluster e pontos externos não são incorporados. Os clusters resultantes não são necessariamente bem separados porque estes não são unidos quanto contêm um par de pontos demasiados distantes.

Se os clusters C_1 e C_2 são aglomerados num novo *cluster*, a dissimilaridade entre a sua junção e outro *cluster* Q é dada por (ver em Kaufman [41]) :

$$D(C_1 \cup C_2, Q) = \alpha_1 \times D(C_1, Q) + \alpha_2 \times D(C_2, Q) + \beta \times D(C_1, C_2) + \gamma \times |D(C_1, Q) - D(C_2, Q)| \quad (4.1)$$

onde: os 4 coeficientes $\alpha_1; \alpha_2; \beta; \gamma$ são definidos pelo utilizador.

4.2.1.2 DIANA

Este algoritmo pode ser comparado directamente o *AGNES* (ver 4.2.1.1) pois tem o mesmo tipo de entradas e saídas. No entanto difere na maneira como as hierarquias são construídas sendo neste caso em direcções opostas.

A abordagem divisiva oferece uma vantagem distinta na medida em que a maior parte dos utilizadores estão interessados na estrutura global dos dados, através do cálculo de alguns *grandes clusters* em vez de uma descrição detalhada dos pontos individuais.

DIANA fornece os seguinte elementos:

- coeficiente divisivo que mede a *quantidade de clustering* encontrada;
- a forma de visualização em *banner* 4.3.

Este algoritmo constrói uma hierarquia de clusters, começando com 1 grande cluster que contém todas as N Observações. Desta forma os clusters são divididos até que cada cluster contenha apenas uma so observação.

Em cada fase o cluster com maior diâmetro (sendo este a maior dissimilaridade entre 2 das suas observações) é seleccionado. Para dividir o cluster seleccionado o algoritmo, começar por procurar a observação mais *díspar* (isto é a observação que possui a maior média de dissimilaridade relativamente às outras observações do cluster seleccionado). Deste modo esta observação inicia o grupo de separação (*splinter group*). Nos passos seguintes o algoritmo re-avalia as observações que estão mais próximas do *splinter group* do que do *cluster* inicial.

O resultado é a divisão do *cluster* seleccionado em dois novos clusters. Dependendo da implementação é possível interromper o processo de divisão a meio se não for do interesse do utilizador continuar a dividir o conjunto de dados além de uma determinada fase.

4.2.1.3 MONA

MONothetic Analysis Clustering of Binary Variables

Este algoritmo devolve uma lista representando o clustering hierárquico divisivo de um conjunto apenas de variáveis binárias.

Todas as variáveis têm de ser binárias; quando existem valores em falta este são considerados como (NA s).

[NA s] neste caso é apenas um valor fictício utilizado apenas quando não se sabe o valor real e deve substituir-se aquando da existência de valores omissos. Assim cada observação deve ter pelo menos um valor diferente de NA , nenhuma variável pode ter metade dos seus valores em falta; deve existir pelo menos uma variável com todos os valores completos; uma variável com todos os valores iguais não é permitida.

Este algoritmo é *monothetic* no sentido em que cada divisão é baseada numa só única (bem escolhida) variável, onde maior parte dos algoritmos hierárquicos (incluindo *AGNES* e *DIANA*) são *polythetic*, isto é utilizam todas as variáveis em conjunto.

Este algoritmo começa com 1 grande cluster que contém todas as N Observações, os clusters são então divididos até que todas as observações no mesmo *cluster* tenham valores idênticos para todas as variáveis. Em cada passo todos os *clusters* são divididos de acordo com os valores de uma variável. Um *cluster* é dividido noutro com todas as observações com valor 1 para essa mesma variável e outro *cluster* com todas as observações para ter o valor 0 para essa mesma variável.

A variável utilizada para separar um cluster é variável com a *associação total máxima* relativamente a outras variáveis de acordo com as observações no *cluster* a ser separado.

Este algoritmo não funciona com registos incompletos, deste modo toda a informação tem de estar completa e revista, isto é, todos os valores em falta devem estar preenchidos através da inclusão de *NAS*. Para isso a mesma medida de associação entre variáveis é utilizada como no algoritmo inicial.

Quando a variável f tem valores em falta, a variável g com a maior associação absoluta para f é revista. Quando a associação entre f e g é *positiva* qualquer valor em falta em f é substituído pelo valor de g para a mesma observação.

Quando a associação entre f e g é *negativa* o valor em falta em f é substituído pelo valor de $1-g$ para a mesma observação.

Este algoritmo é totalmente descrito no capítulo 7 de Kaufman [41].

4.2.2 Algoritmos de Partição

Um algoritmo de partição obtém uma partição simples do conjunto de dados e não uma estrutura de *clustering* como os algoritmos hierárquicos. Este tipo de algoritmos, baseia-se na atribuição ao conjunto inicial de dados, de um número de *clusters*, ao tentar minimizar um determinado critério. O número de *clusters* é normalmente preestabelecido, mas a sua determinação pode fazer parte da função de optimização [42]. Esta função de optimização é provavelmente definida localmente para um subconjunto de registos ou globalmente para todo o conjunto. O número de *clusters* assume tipicamente valores entre 2 a $N-1$, em que N é o número total de registos do conjunto inicial.

Para um conjunto inicial de dimensão n e um número predefinido de *clusters* igual a k , é determinadas de forma exaustiva todas as possíveis partições do mesmo e para cada uma seria calculado o índice de dispersão do *cluster* ao seleccionar a partição que apresentasse o menor índice de dispersão.

Este tipo de procura exaustiva só é viável na prática para valores de n e k relativamente pequenos. Deste modo, são utilizados critérios de optimização locais, que permitem estimar qual a *melhor* partição. O algoritmo de partição mais conhecido é o *K-Means*.

Uma possível estrutura deste tipo de algoritmos pode ser a seguinte:

1. seleccionar um número inicial k de *clusters* que formem uma partição;
2. calcular o centro de cada *cluster*;
3. gerar uma nova partição atribuindo cada registo ao *cluster* cujo centro esteja mais próximo;
4. recalcular os centros;
5. repetir os passos anteriores até que todos os registos estejam recolocados;

6. ajustar k dividindo os *clusters* de maior dimensão por agregação dos de menor dimensão e ou remover os *clusters* que apresentem grande disparidade;
7. Repetir desde o segundo passo até ser satisfeito o critério de paragem.

Para criar os *clusters* iniciais no primeiro passo existem diversos métodos. Tendo múltiplos critérios de paragem através da variação de certos parâmetros.

Estes algoritmos apresentam, relativamente aos algoritmos hierárquicos, a vantagem de poderem ser aplicados a conjuntos de dados de maiores dimensões, em que a construção de uma estrutura de *clustering* como o dendrograma não é viável. No entanto apresentam ainda menor sensibilidade a imperfeições nos dados, como o ruído, outliers e *clusters* com sobreposição.

4.2.2.1 K-Means

A designação *K-Means* deve-se ao facto do algoritmo, representar cada um dos K *clusters* C'_j , pela média (ou média pesada) c'_j dos seus pontos, em que c'_j é designado centro do *cluster*. O número de *clusters* é definido e fixado inicialmente, mantendo-se o mesmo ao longo do processo. O algoritmo faz uma partição inicial aleatória e atribui cada registo a um *cluster*, com base na distância de similaridade entre o mesmo e o centro do *cluster*. Este processo repete-se até que ser atingido o critério de convergência.

Este critério é habitualmente o ponto em que não existe nenhuma troca de registos entre *clusters*, ou um valor mínimo para a diminuição da função objectivo de uma iteração para a seguinte. Este algoritmo é muito sensível à partição inicial escolhida. Se esta não for bem escolhida, o algoritmo pode convergir para um mínimo local da função objectivo. Neste sentido existem trabalhos de Brown [43] e Babu [44] que propõem variantes do método, no sentido de conseguir uma boa partição inicial para que o algoritmo encontre quase imediatamente um mínimo global.

O primeiro trabalho propõe a utilização de técnicas de *Simulated Annealing* e o 2º recorre à utilização de algoritmos genéticos para resolver o problema de optimização da partição inicial, ao evitar que a solução obtida seja um mínimo local. A função objectivo utilizada por este algoritmo é erro quadrático total entre o centro do cluster e os respectivos pontos. Este tipo de ajuste *tune-up* permite obter bons resultados em *clusters* isolados e compactos.

Uma abordagem para um número predefinido de *clusters* k segue a seguinte estrutura:

1. Determinar os k centros para a partição inicial, ao utilizar registos seleccionados aleatoriamente.
2. Atribuir cada registo, ao *cluster* cujo centro esta mais próximo do mesmo.
3. Recalcular os centros dos *clusters*, ao utilizar os registos que lhes foram atribuídos.

4. Se o critério de convergência não tiver sido atingido, voltar ao ponto 2.

O algoritmo de *K-Means* exige um esforço computacional proporcional ao produto do número de registos n existentes no conjunto de registos C , pelo número de *clusters* K para cada iteração, o que o torna muito dispendioso computacionalmente para grandes conjuntos de dados. Esta limitação torna difícil a sua aplicação a problemas de *DM*, pelo que existem trabalhos que apresentam abordagens do método que pretendem reduzir o esforço computacional necessário. A popularidade deste algoritmo justifica-se pela sua simplicidade, compreensibilidade e pelas suas bases sólidas, oriundas da análise de variância.

No entanto o algoritmo apresenta algumas limitações, nomeadamente:

- forte dependência da partição inicial;
- possibilidade significativa de convergir para um mínimo local;
- dificuldade em determinar o número de *clusters* ;
- sensibilidade a valores isolados;
- baixa *scalability*;
- funciona apenas com atributos numéricos.

Para atenuar as limitações definidas existem diversas extensões e modificações do algoritmo inicial as quais resolvem parcialmente esse problema.

4.2.2.2 PAM *Partitioning Around Medoids*

Este algoritmo agrupa objectos que são medidos em variáveis intervalo p . De modo a obter k clusters o algoritmo selecciona k objectos no conjunto. Os *clusters* correspondentes são então encontrados associando cada objecto restante ao objecto representativo mais próximo. No entanto nem toda a selecção de clusters k dá lugar a bom *clustering*. A ideia é escolher os objectos representativos de modo a estarem no centro do cluster a definir. A distância média do objecto representativo e os outros objectos do mesmo cluster deve estar sempre a ser minimizada. A descrição detalhada de todo o algoritmo está presente em Kaufman [41] (Cap. 2).

Comparável à abordagem *K-Means* este algoritmo tem as seguintes particularidades:

- aceita uma matriz de dissimilaridades;
- é mais robusto pois minimiza a soma das dissimilaridades em vez da soma do quadrado das distâncias euclidianas;

- mostra graficamente a *silhouette* ;
- permite seleccionar o número de clusters utilizando a mediana da *silhouette* a partir do resultado inicial ou directamente sobre o componentes finais.

Este algoritmo é baseado na procura dos k (*Medoids*) *representative objects* no conjunto de observações. Essas observações em ultima instância devem representar a estrutura da informação contida no conjunto. (ver 4.4)

Depois de encontrar um conjunto de k *medoids*, k clusters são construídos ao associar cada observação ao medoid mais proximo. O objectivo é encontrar k objectos que minimizem a soma das dissimilaridades das observações do objecto mais proximo que os representa. Por omissão, quando os *medoids* não são especificados, o algoritmo pesquisa por um *bom* conjunto inicial de medoids (chamada *fase de construção* - (*build phase*)). De seguida procura um mínimo local para a função objectivo, o que é a solução na qual não há mais que uma simples troca de uma observação com o medoid que diminuirá o objectivo (chamada *fase de troca* - (*swap phase*)). Quando a opção *cluster. only* é verdadeira o resultado é simplesmente um vector especificando o clustering sem mais dados adicionais.

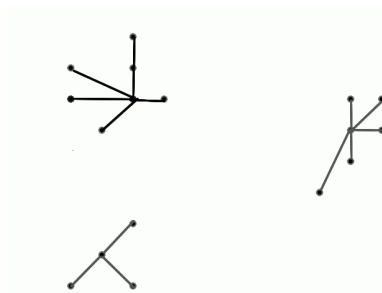


Figura 4.4: Representação de consumidores através de *PAM*

O algoritmo aceita como parâmetro duas medidas métricas para calcular as dissimilaridades entre observações. As opções existentes actuais são *euclidian* e *manhattan*. A distância euclidiana é a raiz da soma dos quadrados das diferenças, a distância de *manhattan* é a soma das diferenças em modo absoluto. Há ainda a possibilidade manter os dados da matriz inicial e se são mostrados no resultado final. Se se especificar esse modo como falso é possível obter resultados mais pequenos e assim salvar tempo de alocação de memória.

4.2.2.3 CLARA

Comparando com outros algoritmos de partição tais com o *PAM*, este consegue lidar com conjuntos muito maiores que os anteriores. Internamente isto é alcançado ao considerar

sub-conjuntos de tamanho fixo de modo que a complexidade e tamanho requeridos fiquem lineares em N em vez de quadráticos $O(N^2)$.

Cada sub-conjunto é partido em k clusters utilizando a mesma abordagem que em *PAM*.

Uma vez que k objectos representativos foram seleccionados do sub-conjunto, cada observação do conjunto inteiro é associada ao *medoid* mais proximo. Este procedimento é feito diversas vezes sendo a soma das dissimilaridades das observações para o *medoid* mais próximo e existe como medida de qualidade de *clustering*. O sub-conjunto para o qual a soma é mínima é mantido para a iteração seguinte.

Uma análise extra é calculada na partição final em que cada sub-conjunto é forçado a conter os medoids do melhor sub-conjunto até ao momento. As diversas observações são escolhidas aleatoriamente e adicionadas a esse conjunto até que o valor *sampsize* tenha sido atingido.

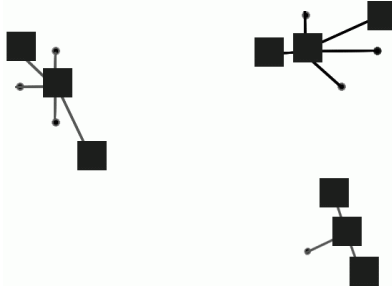


Figura 4.5: Exemplo do agrupamento de alguns consumidores através do algoritmo *CLARA*

O requisito de capacidade do algoritmo é baixo, (para um k pequeno) é cerca de $O(n \times p) + O(j^2)$ onde $j = \text{sampsize}$ e $(n, p) = \text{Dim}(x)$.

A complexidade de computação (ignorando o valor de k) é :

$$O(n \times p \times j^2 \times N) \quad (4.2)$$

onde: N = número de observações.

Para pequenos conjuntos o algoritmo 4.2.2.2 fornece os mesmos tempos e resultados e pode ser utilizado directamente nos dados. Mais detalhes sobre o algoritmo presentes em Kaufman [41] (Cap. 3). Foi este algoritmo, que devido à sua complexidade de computação e à capacidade de tratar elevado número de registos se utilizou numa análise inicial da construção do perfil de consumo típico *PCT*.

Como exemplo (ver figura 4.6) no conjunto de consumidores [MT] foi possível obter automaticamente perfis diários apenas fornecendo ao algoritmo *CLARA* o número de *clusters* a calcular. Nesta visualização os perfis estão separados para os diversos valores de energia.

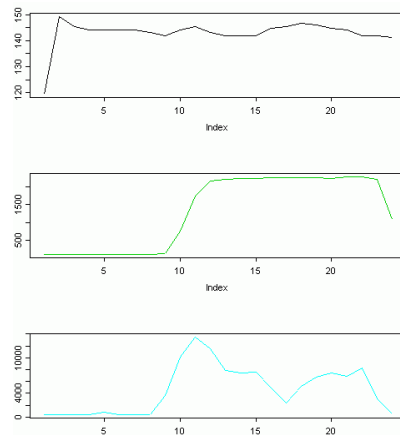


Figura 4.6: Amostra de clusters do tipo de consumidores [MT]

4.2.2.4 FANNY, *Fuzzy Analysis Clustering*

Este algoritmo pode ser aplicado ao mesmo tipo de dados que o *PAM* mas o seu princípio é diferente. Ao utilizar o princípio *fuzzy*; (*fuzziness principle*) significa que consegue evitar decisões *duas*. Em vez de dizer que o objecto α pertence ao cluster 1, este algoritmo indica que α pertence 90 % ao cluster 1, 5% ao cluster 2, 2 % ao cluster 3 mostrando que α é provavelmente marcado para o cluster 1 mas pode existir alguma dúvida nos restantes. No caso do conjunto ser em bruto, cada linha corresponde a uma observação e cada coluna corresponde a uma variável, todas as variáveis devem ser numéricas, valores em falta são permitidos. Neste algoritmo um vector de comprimento $n \times (\frac{n-1}{2})$ é permitido onde n é o número de observações. Sendo interpretado da mesma maneira que uma matriz de dissimilaridades porém neste ultimo caso valores em falta não são permitidos. Neste algoritmo o k (número de clusters) deve estar compreendido entre $0 < k < \frac{n}{2}$ onde n é o número de observações.

Normalmente aceita as medidas normalizadas para cada variável (neste caso por coluna), ao subtrair o valor da mediana e dividindo pelo desvio absoluto da mediana, neste caso com a opção *stand*. No *Fuzzy Clustering* cada observação está *espalhada* em vários clusters. Identificando $u(i,v)$ como nível de pertença da observação i para o cluster *versus* o nível de pertença (não-negativo e para uma observação fixa i a sua soma é 1). Este algoritmo é descrito em Kaufman [41] (Cap. 4). Comparando com outros algoritmos de clustering Fuzzy, *FANNY* tem os seguintes elementos:

- aceita tanto dados em bruto como matrizes de dissimilaridade;
- é mais robusto para *clusters* tipo esférico;

FANNY tenta minimizar a função objectivo :

$$\sum_{v=1 \dots k} \frac{\left(\sum_{i,j} u(i,v)^2 u(j,v)^2 d(i,j) \right)}{\left(2 \times \sum_j u(j,v)^2 \right)} \quad (4.3)$$

onde n é o número de observações, k é o número de clusters e $d(i,j)$ é a dissimilaridade entre observações i e j .

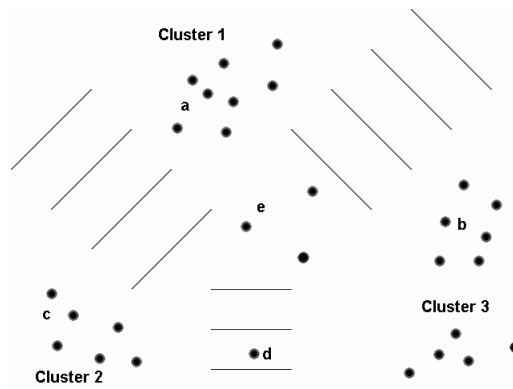


Figura 4.7: Algoritmo FANNY

Como se pode ver na figura 4.7 os elementos (a, b, c, d, e) têm todos diferentes níveis de pertença relativamente às partições definidas. Os elementos (a, b, c) são simples de se identificar com determinado cluster, no entanto os outros 2 têm níveis de pertença nos quais não é possível definir imediatamente a qual cluster pertencem.

4.2.3 Outros Algoritmos

Deste modo, a selecção do algoritmo de *clustering* a utilizar deve ter em consideração a seguintes características:

- tipo de atributos que o algoritmo pode tratar;
- escalabilidade para grandes conjuntos de dados;
- capacidade para trabalhar dados com elevada dimensão;
- capacidade para encontrar *clusters* de forma irregular;
- influência de valores isolados;
- dependência entre os dados;

- tipo de níveis de alocação (rígidos ou *fuzzy*);
- dependência de conhecimento prévio e parâmetros definidos pelo utilizador;
- interpretabilidade dos resultados.

Estas características permitem a comparação entre algoritmos e, desse modo, facilitar a selecção da técnica a aplicar.

4.3 Aplicação de Clustering

Foram testados e aplicados diferentes algoritmos, de modo a proceder a um estudo comparativo, que permitiu concluir qual o algoritmo que permite obter melhores resultados, para o conjunto de dados em estudo.

Neste caso (figura 4.8) a representação do conjunto agregado das diversas classes encontrados pelo algoritmo *CLARA* no conjunto de dados de consumidores [MT] do projecto CCR [1].

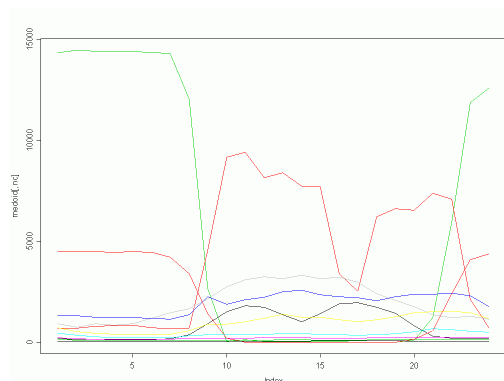


Figura 4.8: Exemplo de classes de consumidores [MT] obtidos com algoritmo de clustering CLARA

Nos testes efectuados, o algoritmo *CLARA* foi o que se revelou com melhor capacidade de resposta devido à capacidade de tratar grandes volumes de dados e de fornecer *clusters* bem caracterizados.

Devido à necessidade de agrupar e classificar *automaticamente* consumidores com características semelhantes aplicaram-se técnicas de *clustering* de modo a validar os PCT s pré-definidos com os existentes no conjunto.

Assim, procedeu-se á comparação entre os perfis encontrados no conjunto de dados reais e os pré-definidos anteriormente.

Deste modo, limitou-se a aplicação apenas dos PCT s que realmente existiam no conjunto reduzindo assim o tempo de testes de comparação.

Capítulo 5

Perdas

5.1 Perdas na Rede

O termo *Perdas na rede* significa a diferença normalmente em termos da grandeza potência, entre a entrada e saída de uma rede eléctrica, a diferença entre o que é produzido e o que é entregue ao consumidor. Assim, define-se *Perda* como sendo a diferença existente entre a grandeza de entrada (energia ou potência injectada na rede) e a grandeza de saída (energia ou potência cedida ao conjunto dos consumidores):

$$\text{Perdas}_i = \text{Entrada}_i - \text{Saída}_i$$

onde i é o índice da grandeza estudada podendo esta ser classificada quanto à localização e componentes.

De uma forma geral, as perdas aumentaram ao longo da ultima década e provavelmente grande parte do sistema de distribuição apresenta, hoje, perdas excessivas, sendo uma das principais razões a existência de cabos a funcionar em regimes próximos da capacidade máxima.

A justificação principal da situação actual deve-se ao facto do nível dos investimentos, no melhoramento do sistema de distribuição com a introdução de novas linhas, serem insuficientes para compensar o aumento da carga.

Assim, no mercado, as empresas ao investir mais na rede ficam com menor lucro bruto posteriormente sofrendo a pressão dos accionistas para aumentar esse mesmo lucro. No entanto, as perdas técnicas continuam a ser avaliadas pelos custos marginais, com larga aceitação internacional e o seu nível económico corresponde ao equilíbrio entre o seu custo e o custo dos investimentos necessários para diminuir os mesmos.

5.1.1 Tipos de Perdas

As perdas podem ser divididas em dois grandes tipos:

Perda Comercial - é a energia ou consumo efectivamente fornecido ao consumidor, mas não considerada nas vendas (como exemplo: ligação directa ou mesmo roubo da energia).

Perda Técnica - é a energia ou consumo perdido no transporte e na transformação, logo inerente ao processo e caracterizada por ocorrer antes do ponto de entrega.

5.1.2 Localização

As perdas localizam-se nos diversos segmentos de um sistema eléctrico, dividindo-se habitualmente em Perdas de Geração, de Transmissão e da Distribuição.

Face ao tipo usual de estrutura das empresas do Sector Eléctrico, no qual o sistema de geração e o sistema de transmissão são tratados geralmente como um único segmento, a classificação das perdas técnicas é vulgarmente considerada em 2 níveis distintos: Perdas na Transmissão e Perdas na Distribuição.

5.1.3 Componentes

As perdas técnicas na distribuição podem ser classificadas de acordo com o componente eléctrico onde as mesmas ocorrem, nomeadamente : condutores da rede primária, transformadores de distribuição, condutores da rede secundária, ramais de ligação, bobina de potencial dos medidores dos consumidores, equipamentos (condensadores, reguladores de tensão, pára-raios, entre outros) e diversas (fuga em árvores e isoladores, efeito corona, conexões).

Destes, os mais significativos são:

- Efeito de Joule - nos condutores;
- Perdas nos transformadores.

As perdas técnicas, em termos de potência, estão divididas em Perdas Fixas L_f (associadas aos transformadores e quase independentes do regime de carga) e Perdas Variáveis L_v função aproximadamente quadrática do regime de carga P .

5.2 Cálculo das Perdas Técnicas

Na realidade apenas existe uma ideia aproximada das perdas técnicas, normalmente utilizam-se os dados de facturação e o balanço energético contendo os valores da energia fornecida ao sistema e a energia facturada, relativos habitualmente a um ano. Para calcular e tentar encontrar valores mais próximos são realizadas simulações em redes modelo ou

mesmo em secções da rede real, no entanto não se consegue ter controle de toda a rede em todos os momentos. Nos sistemas de transmissão, as perdas são calculadas por estudos de fluxo de carga ou estimadas através do balanço energético das linhas se estas estiverem monitorizadas (diferença entre a energia que *entra* incluindo a geração própria de determinado consumidor e a energia que *sai* consumida como vendas mais a energia fornecida à distribuição). Os processos baseados em cálculo (fluxo de carga) apresentam resultados que se devem aproximar à realidade, podendo ser utilizados para análises individuais e localizadas. No entanto, exigem uma extensa base de dados e registos permanentemente actualizados. As metodologias simplificadas (processos estatísticos, modelo geométricos, entre outras) apenas necessitam de um volume reduzido de dados e permitem a estimativa das perdas de forma expedita. No entanto, as mesmas tendem a apresentar resultados satisfatórios apenas quando aplicadas a grandes sistemas e de forma global. De forma a obter estimativas mais robustas podem realizar-se comparações ao utilizar vários processos de estimativa de perdas através de metodologias alternativas, desenvolvidas para serem aplicadas num sistema de distribuição. Por esta razão tendem a apresentar resultados próximos da realidade e que apenas dependem da qualidade dos dados de entrada. Noutros casos, utilizam-se valores médios ou típicos para representar as características do sistema (consumo, comprimento, resistência dos condutores, carga, factor de carga, entre outras).

Diversos elementos podem contribuir para a dificuldade no cálculo de perdas:

- As perdas nos transformadores de distribuição (ferro e cobre) possuem características variáveis, são por vezes de difícil acesso dependendo do fabricante do equipamento.
- As perdas nos condutores da rede de distribuição, apresentam grande dispersão na tipicidade (comprimento, geometria da rede, número de pontos de carga, diâmetro e tipo de condutores, entre outras). Deve estar associado a acções e medições quase permanentes para garantir a qualidade dos dados de entrada no entanto a manutenção dessa base de dados seria bastante onerosa.
- Considera-se que outros tipos de perdas que ocorrem nos sistemas de distribuição (ramais de ligação, medidores, conexões, correntes de fuga, entre outras) apenas contribuem com uma pequena participação no valor total das perdas.

Em síntese o cálculo das perdas é um processo complexo devido aos diversos factores referidos anteriormente. Desta forma, as empresas devem fazer um esforço no sentido de implementar ou ampliar os instrumentos que permitam a gestão de todo o seu sistema eléctrico de distribuição.

Como foi descrito as perdas dependem de muitos parâmetros e características. Muitos destes não são nem conhecidos nem observáveis, o que implica a necessidade de estimativas desses valores e admitindo-se, por vezes intervalos de possibilidade para outros. Daí, as perdas resultantes são recebidas e calculadas em intervalos de valores.

5.2.1 A complexidade da estimativa de perdas técnicas

O estabelecimento de valores típicos de perdas é um processo complexo dependente de diversos parâmetros que podem variar bastante de rede para rede.

Normalmente, pretende-se determinar a energia adicional que é necessária acrescentar nos pontos de injeção nos diferentes níveis de tensão. A estimativa das perdas globais da rede pode ser feita a partir do balanço energético, mas torna-se progressivamente mais difícil quanto mais se avança na desagregação dos elementos da rede. Na verdade, mesmo para um único regime de carga, as perdas dependem como foi referido anteriormente dos parâmetros eléctricos de rede, do comprimento dos cabos, da simultaneidade dos consumos.

Estudos mostram que para a mesma carga total onde se mostra a grande variabilidade das perdas apenas alterando a distribuição da carga na rede. [8] De notar que entre os casos mais e menos desfavoráveis (carga crescente e carga decrescente) as perdas podem variar mais do dobro...

No que respeita à atribuição de perdas aos consumidores finais, também se encontra uma grande variação resultante, sobretudo, da característica quadrática das perdas, da influência da posição do consumidor na rede e da presença ou não de outros consumidores.

Assim, a definição de um valor fixo da potência consumida (ou do quadrado da potência consumida) como contribuição para as perdas do consumidor terá que ter sempre um carácter médio.

5.2.2 Determinação de perdas por período horário

Em diversos países (incluindo Portugal) a atribuição de perdas é feita com base em factores de ajustamento estabelecidos por período horário.

De facto o regulamento prevê isto no ponto 5 do artigo 59º do RARI (Regulamento do Acesso às Redes e às Interligações): *Os factores de ajustamento para perdas são diferenciados por rede, de transporte ou de distribuição, por nível de tensão e por período tarifário, de acordo com o ciclo semanal publicado pela ERSE no despacho anual que estabelece as tarifas e preços da energia eléctrica para o ano seguinte.*

Esta secção apresenta uma das metodologias para o cálculo da percentagem de perdas a considerar por nível de tensão e por período horário, a partir dos resultados dos diagramas representativos, a qual foi utilizada no projecto CCR [8].

Podem-se considerar basicamente duas alternativas para o cálculo das perdas por período horário: com base nos valores típicos dos cenários de *vazio*, *cheia*, *super vazio* ou com base nos diagramas completos, como aconteceu com o projecto CCR [8].

Neste ultimo caso, a metodologia adoptada foi a seguinte:

- Cálculo mensal do consumo por período horário, tipo de consumidor e tipo de dia;
- Cálculo mensal das perdas por período horário, tipo de consumidor e tipo de dia;

- Agregação dos consumos referidos inicialmente;
- Agregação das perdas;
- Cálculo dos factores de ajustamento para as perdas (Fap) por período e nível de tensão, a partir dos valores obtidos nos pontos anteriores.

Em qualquer dos casos é possível estimar as perdas globais com base nos factores de ajustamento ou outros e verificar se concordam ou não com os valores do balanço global.

5.2.3 Métodos de Alocação de Perdas

Ao depender dos diversos grupos de consumidores a alocação de perdas pode ser efectuada com base na seguinte forma:

- Determinação do nível de perdas típico em redes BT, MT e AT.
- Estabelecimento e aplicação dos critérios de distribuição de perdas com base nos valores típicos considerados.

A estimativa do nível de carga típico da diversas redes (BT, MT e AT) é muito importante para efectuar a estimativa das perdas e suas variações, uma vez que ao nível médio de carga não corresponde à média das perdas.

Neste capítulo apenas são referidos métodos relativos à rede de distribuição não considerando métodos marginais. Assim, são descritos os seguintes :

- Alocação por Nível de Tensão;
- Postage Stamp;
- Modified Postage Stamp;
- MW.Km ;
- MW^2 .Km;
- Participações Médias;
- Participações Médias Modificadas.

5.2.3.1 Alocação por Nível de Tensão

A alocação de perdas aos diversos tipos de consumidores pode ser baseada, no Nível de Tensão [8].

No caso do Projecto CCR utilizou-se a seguinte metodologia:

- Os consumidores são agrupados em 3 conjuntos distintos, consoante o nível de tensão de alimentação, desde os consumidores BT até consumidores AT.
- Para cada conjunto considerado, as perdas distribuem-se, pelo diagrama global anual respectivo, proporcionalmente ao quadrado do consumo horário.

A distribuição de perdas é definida pela seguinte expressão:

$$p(g, m, d, h) = \frac{E^2(g, m, d, h)}{\sum E^2(g, m, d, h)_{cons.g}} * P_b(g) * e(g, m, d, h) \quad (5.1)$$

em que:

$p(g, m, d, h)$ - perdas (kWh) do grupo g de consumidores considerado mensalmente m , tipo de dia d e hora h ;

$p_b(g)$ - percentagem base de perdas considerada para o grupo g ;

$E(g, m, d, h)$ - consumo médio dos consumidores do grupo por mês, tipo de dia, hora;

- As perdas dos transformadores são atribuídas aos consumidores dos níveis de tensão iguais ou inferiores à tensão do secundário;

Deste modo, foram obtidos diagramas de perdas (valores médios e valores percentuais) para os diversos tipos de consumidores considerados, em função do mês e do tipo de dia (dia útil, Sábado, Domingo).

Neste caso foi possível implementar esta metodologia porque se dispõe dos diagramas representativos dos diversos tipos de consumidores, o que vulgarmente não acontece.

5.2.3.2 Postage Stamp

Este método parte do princípio de que todos os consumidores alimentados pela rede a utilizam da mesma forma, no que se refere ao percurso, distância, direcção, características físicas e eléctricas da rede.

A diferenciação entre consumidores é efectuada, apenas, pela quantidade de energia ou potência que cada um solicita à rede, o que significa que os custos das perdas eléctricas são alocados aos diferentes consumidores, de forma proporcional à quantidade de energia ou potência solicitada por cada um.

A alocação dos custos associados às perdas aos diferentes consumidores é, portanto, baseada num custo médio, cujo cálculo se resume ao quociente entre o total dos referidos custos e o total da grandeza adoptada para medida da utilização que cada consumidor faz da rede. Como foi referido por J.Saraiva et al [45] este é um método de fácil de aplicação.

Para medir esta utilização, pode usar-se a potência ou a energia eléctrica consumida por cada um dos consumidores num determinado instante de tempo ou período de tempo respectivamente.

Perante estes critérios para a medição da utilização da rede, é importante definir em que período se deve efectuar tal medição. Quando a potência for o critério adoptado, a medida da utilização da rede pode ser efectuada com base no período de ponta de todo o sistema ou com base no período de ponta do utilizador. Se o critério adoptado, para quantificar a utilização da rede por cada utilizador, for a energia torna-se necessário definir de forma clara o período de medição.

A expressão capaz de determinar o custo unitário das perdas, a qual servirá de base para a alocação dos custos de cada um dos consumidores é:

$$p = \frac{P_T}{G} \quad (5.2)$$

no qual :

p é o valor das perdas por unidade de potência ou de energia associada a um determinado consumidor, expresso em W/W ou Wh/Wh .

P_T é o valor total das perdas eléctricas associadas ao sistema de distribuição em W .

G é o valor total da grandeza adoptada para servir de base à medição da utilização da rede pelos consumidores, medida num determinado momento ou período.

A previsibilidade do método é uma das suas vantagens. De facto, o valor de perdas a suportar por cada consumidor é uma função da quantidade de energia ou potência que é por este retirada da rede, e não das perdas efectivas que a sua alimentação vai provocar. Diminui-se assim a volatilidade dos custos a suportar por cada consumidor, o que, do ponto de vista económico, pode ser uma vantagem muito considerável para estes. No entanto pode parecer injusto para os pequenos consumidores (BT) .

Uma outra vantagem deste método é que este permite recuperar a totalidade dos custos considerados, resultantes de uma previsão.

À simplicidade deste método contrapõem-se algumas desvantagens: uma delas define-se com o facto da utilização deste método, para alocação das perdas, não permitir o fornecimento de informação económica adequada aos consumidores das redes de distribuição por forma a incentivar o aumento da eficiência do sistema eléctrico.

De facto, este método acarreta alguns problemas, na medida em que para o caso de dois consumidores que consumam igualmente o mesmo valor de energia, independentemente da energia destinada a cada um percorrer diferentes distâncias e caminhos na rede eléctrica,

suportam igual valor de perdas eléctricas.

A forma de operação deste método leva a que cada consumidor suporte uma parte dos custos de cada um dos elementos da rede de distribuição, mesmo que nunca utilize alguns desses elementos.

Tal facto deve-se à forma de operação do método que efectua uma alocação das perdas globais da rede e não uma alocação por elemento de rede.

Este método garante a existência de uma relação linear entre as perdas suportadas pelo consumidor i e o valor da sua potência no que refere a potência total do sistema.

Por outro lado, para igual valor relativo da potência do consumidor i face ao valor da potência total do sistema, são alocadas a este consumidor iguais perdas, independentemente da relação de grandeza entre as potências dos restantes consumidores alimentados pela rede ou do seu número.

5.2.3.3 Modified Postage Stamp

Como foi referido, anteriormente, o método *Postage Stamp* 5.2.3.2 aloca as perdas a cada um dos consumidores das redes de distribuição em função do valor solicitado de potência ou da energia eléctrica à rede de distribuição.

No entanto, as perdas eléctricas nos componentes da rede, ignorando os efeitos das tensões, são função do quadrado das potências que fluem por eles. Esta é uma alteração ao método *Postage Stamp* de forma a tentar uma maior aproximação à realidade dos sistemas eléctricos de energia.

A modificação do método centra-se na consideração dos quadrados das potências em vez da consideração das potências simples, como grandeza de medição da utilização da rede por parte de cada consumidor.

$$G_j = P^2 \quad (5.3)$$

A determinação do custo unitário das perdas a associar aos vários consumidores é agora efectuada através da expressão :

$$p = \frac{P_t}{\sum_j G_j} \quad (5.4)$$

Esta alteração ao método *Postage Stamp* 5.2.3.2, referente ao método convencional, conduz a uma penalização dos consumidores que absorvem maior potência ou energia na rede, beneficiando, assim, os que absorvem menores quantidades ao contrário do que sucede com o método anterior, para o qual os que absorvem maiores quantidades são os beneficiados.

A alocação de perdas é mais coerente do que a que é calculada pelo método convencional, uma vez que o acréscimo de perdas, resultante do facto de se alimentarem os 2 consumidores em conjunto deixa de ser repartido do mesmo modo entre estes, passando a maior parte a ser suportado pelo consumidor de maior potência.

5.2.3.4 MW.Km

O método *MW.km* pode ser representado em diversas variantes, as quais se distinguem principalmente pela implementação prática.

O seu princípio básico consiste na medição dos MW.km imputáveis a cada consumidor da rede, como forma de comparar a utilização que cada um faz desta. Contrariamente aos anteriores, este método tenta reflectir, não só as potências que são solicitadas pelos consumidores, mas também a *quantidade de rede* que é utilizada por cada um [45]. Na realidade existem duas variantes fundamentais deste método, uma baseada na distância física entre o ponto de injeção e o ponto de recolha da energia e outra baseada nos resultados de estudos de trânsito de potências, referidos por J.Saraiva [45].

A primeira variante, apesar de mais simples de compreender e de aplicar, é pouco realista uma vez que não expressa as alterações reais do sistema originadas pela alimentação de cada consumidor.

Nesta variante do método, as perdas a associar a um consumidor são:

$$P_{pj} = PT_t * \frac{P_j * L_j}{\sum_t P_t * L_t} \quad (5.5)$$

no qual:

P_{pj} é o valor de perdas alocado ao consumidor j .

P_T representa o valor total regulado das perdas.

$P_j * L_j$ representa o produto da potência pela distância associada ao consumidor j , j este que corresponde à soma dos produtos da potência associada ao consumidor pelo comprimento de cada linha incluída num caminho eléctrico definido.

O denominador da expressão representa o somatório dos valores de MW.km associados a cada consumidor fornecido pela rede de distribuição.

A determinação do valor de MW.Km associado a cada consumidor, conduz a que esta variante do método não expresse as alterações reais do sistema, devidas à alimentação de cada consumidor.

De facto, o caminho da energia entre o ponto de injeção e o ponto de ligação é fictício, sem recurso a nenhum tipo de estudos, nomeadamente o cálculo com trânsito de potência, conduzindo à identificação de um caminho que pode não ser o real.

A segunda variante da aplicação do método MW.Km, através do estudo com trânsito

de potência, baseia o seu cálculo nos valores de MW.Km associados a cada consumidor. Esta situação leva a uma realidade mais aproximada, embora não se obtenha a *realidade absoluta*, devido às propriedades não-lineares dos sistemas eléctricos.

A aplicação do método MW.km, segundo esta variante, pode ser feita de diversas formas e depende das condições do sistema em estudo com trânsito de potência e da forma como são tratadas as diferentes variações ocorridas.

Assim sendo existem diversas formas para se determinarem as contribuições das perdas sendo uma delas a realização de um estudo sobre o trânsito de potência com o funcionamento da rede numa situação *base* representativa da operação normal do sistema. No entanto o trânsito de potência por si só não resolve o problema da distribuição dos consumidores.

Com base neste estudo determina-se o valor global de MW.Km associado à rede, no caso base mediante a expressão seguinte.

$$T_t = \sum_{k=1}^n F_k * L_k \quad (5.6)$$

onde:

T_t é o valor de MW.km atribuídos à rede no caso base; F_k é o valor do trânsito de potência na linha k do sistema; L_k é o comprimento da linha k ; n é o número de linhas da rede.

De seguida retira-se sucessivamente cada um dos consumidores e elabora-se um novo estudo de trânsito de potência para o sistema com os restantes, determinando-se o novo valor dos MW.km associados à rede nessas condições.

De salientar, que após calculado o valor de MW.km para um determinado consumidor, a resolução do valor para o seguinte, implica a necessidade de remover este da rede e também a de incluir aquele anteriormente estudado.

Uma segunda forma de obter a contribuição de cada consumidor j para o trânsito de potência global em cada linha do sistema passa pela observação da rede em vazio e a inserção de cada um dos consumidores nesta.

Por outras palavras, coloca-se na rede cada um dos consumidores isoladamente e efectua-se um estudo de trânsito de potência, o qual revela as potências que este consumidor implica em cada linha.

Note-se que, devido à não-linearidade dos sistemas eléctricos, nenhuma destas duas formas utilizadas para determinar o valor de MW.km associados a cada consumidor conduz a resultados próximos dos reais, uma vez que ambas podem conduzir a valores distintos de MW.km a atribuir a cada consumidor.

O quociente entre o valor de MW.Km imputado ao consumidor em estudo e o valor global de MW.Km corresponde a uma percentagem de utilização da rede pelo consumidor.

Saliente-se que, o valor global de MW.Km resulta do somatório dos valores de MW.Km

imputados a cada consumidor, não sendo portanto, o valor determinado através de um estudo de trânsito de potências para um caso base.

O custo de perdas a suportar pelo consumidor j é :

$$P_{pj} = PT_t * \frac{T_j}{\sum_i T_i} \quad (5.7)$$

onde:

P_{pj} é o valor das perdas a suportar pelo consumidor j ; P_T é o valor total das perdas do sistema.

O método MW.Km apresenta algumas vantagens significativas face ao método *Postage Stamp* 5.2.3.2. Uma das vantagens é a possibilidade de fornecer informação económica, de forma a minorar o impacto e as perdas do sistema.

De facto, a contabilização das distâncias para determinar a utilização da rede que cada consumidor apresenta, tende a aproximar os consumidores dos pontos de injeção de energia na rede de distribuição.

Neste caso, uma expressão da rede pode ter a forma:

$$T_j = \sum_{k=1}^n F_{kj} * L_k * h_k \quad (5.8)$$

onde: h_k é um coeficiente que procura reflectir as condições físicas e de estado de cada elemento da rede. A informação económica fornecida é deste modo maior do que a que é fornecida pelo método convencional MW.km 5.2.3.5.

5.2.3.5 $MW^2.km$

Tal como acontece com o método 5.2.3.2, a aplicação do método *MW.km*, na sua versão original, não permite expressar a dependência quadrática das perdas em cada linha com a potência que nestes transita. Esta situação tende a beneficiar os utilizadores da rede que originam trânsito de potência mais elevados, sendo parte das perdas originadas por estes suportadas pelo tipo de utilizadores com menor trânsito de potência, a qual pode ser, parcialmente resolvida ao aplicar o método em $MW^2.Km$ e não em termos de MW.km. A alteração ao método original, permite fazer sentir aos utilizadores da rede, com maior aproximação, a relação quadrática existente entre as perdas de cada linha e o respectivo trânsito de potência. A aplicação do método modificado é em tudo semelhante à do método original, excepto na forma como se determinam os valores de $MW^2.km$ a atribuir.

Neste caso, a expressão da rede é substituída por:

$$T_j = \sum_{K=1}^n F_{kj}^2 * L_k \quad (5.9)$$

Contrariamente ao que sucedia no método Modified Postage Stamp 5.2.3.3, o risco de se penalizar os consumidores que se situam próximos dos pontos de injeção de potência na rede e que solicitam à rede grandes potências, é neste caso mais reduzido. Assim, neste método, não são só as potências solicitadas à rede por um determinado consumidor que influenciam a medição da sua utilização da rede, mas também o trânsito de potência provocado nos ramos e as características físicas. No entanto diversas entidades reguladoras são apologistas do princípio da igualdade no que se refere á distancia entre consumidores e o respectivo posto de transformação.

A introdução desta relação quadrática na determinação do valor de $MW^2.Km$ a atribuir a cada utilizador da rede pode, no entanto, conduzir a situações de inconsistência.

Assim, apesar do valor do trânsito da potência total numa linha permanecer constante, o valor das perdas entre dois consumidores que repartem a essa utilização pode ser diferente.

5.2.3.6 Participações Médias

Bialek [46], D. Kirschen et al [47], Macqueen [48], referem a aplicação do método de *Participações Médias* para a alocação do uso das redes entre os seus diferentes utilizadores. O método baseia-se no fluir da energia eléctrica desde as fontes de alimentação até às cargas, permitindo efectuar uma alocação de custos aos consumidores sustentados por uma determinada rede eléctrica.

A partir de um determinado cenário de injeção e de consumo de potência nos nós da rede, determina-se o trânsito de potências em cada um dos seus elementos através de um estudo de trânsito de potências. De seguida aplica-se um algoritmo de linearização que permite determinar os valores característicos do trânsito de potências nos ramos e das potências das cargas e geradores. Após serem conhecidas as potências características injectadas e consumidas nos diferentes nós e o trânsito de potências característicos nos diferentes ramos, determina-se a fracção com que cada consumidor contribui para o trânsito de potência em cada elemento da rede, utilizando uma proporcionalidade simples.

O princípio de cálculo da contribuição de cada consumidor para o trânsito de potência em cada elemento da rede baseia-se na aceitação de que, uma vez conhecidas as potências que incidem num nó, os diferentes consumidores da rede utilizam os elementos que veiculam essa potência ao manter a proporcionalidade verificada nas potências que incidem no nó, ou seja, de acordo com o princípio das participações médias. Este princípio pode ser generalizado a todos os nós da rede e de forma simples seguir a energia eléctrica ao longo desta. Foram propostos vários algoritmos para efectuar tal seguimento, uns mais formais do que outros, mas todos baseados no princípio das participações médias.

Um algoritmo formal foi proposto por Bialek [46], baseado numa análise matricial da

rede, o qual pode ser dividido em duas variantes, o algoritmo para jusante e o algoritmo para montante.

Quando, para o trânsito de potência numa determinada linha, contribuem mais que dois consumidores, verifica-se que a alocação dos custos, com base nas potências simples, conduz a uma situação em que cada consumidor suporta um total de perdas proporcional ao somatório do quadrado da potência, que faz circular nessa linha acrescida de metade do dobro do produto entre a contribuição desse consumidor e cada uma das contribuições dos restantes.

A aplicação do método das participações médias impede que qualquer consumidor suporte perdas ocorridas em elementos da rede que não utiliza. Este facto não acontecia nos métodos Postage Stamp 5.2.3.2 ou MW.Km 5.2.3.4. A evolução da alocação de perdas, em função do número de consumidores e das potências relativas destes, é análoga à verificada na aplicação do método Postage Stamp, com uma substancial diferença: aqui essa evolução refere-se a um elemento de rede, enquanto no método Postage Stamp se refere a toda a rede. Isto significa que, o valor de perdas alocadas a cada consumidor, em cada linha da rede, não varia com o número de consumidores que utilizam essa linha nem com a relação entre os contributos de cada um para o trânsito de potência na linha, desde que esta se mantenha constante.

5.2.3.7 Participações Médias Modificadas

O método, anterior, das participações médias pode ser modificado para expressar a relação quadrática existente entre as perdas de um elemento. Esta alteração não repercute, realmente, a referida relação, uma vez que as perdas no elemento podem ser proporcionais ao quadrado do seu trânsito de potência e não ao somatório dos quadrados de cada uma das contribuições dos consumidores para o trânsito de potência desse elemento.

A evolução da alocação de perdas, em função do número de consumidores e das potências relativas destes, é análoga à verificada na aplicação do método Postage Stamp modificada 5.2.3.3, nesta evolução refere-se a um elemento de rede, enquanto que no método Postage Stamp 5.2.3.2 se refere a toda a rede.

Neste caso, numa determinada linha admitamos que o estudo de trânsito de potência tem contribuições de n consumidores, consequentemente, o valor para as perdas é calculado por:

$$P_{linha} = \left(\sum_{i=1}^n P_i \right)^2 * k \quad (5.10)$$

Isto significa que o valor de perdas alocadas, a cada consumidor, em cada linha da rede varia com o número de consumidores que utilizam essa linha e com a relação entre os contributos de cada um para o trânsito de potência na linha, apesar deste ser constante.

5.2.3.8 Outros Métodos

Além dos métodos referidos anteriormente existem diferentes famílias de métodos não-marginais tais como métodos de *tracing*; *circuit based*; *grid based*; *Zbus*. Estes seguem abordagens tão variadas desde a inserção de barramentos de referência (*slack bus*) como inteligência artificial. Outras abordagens são baseadas em probabilidade e mesmo em *Teoria de Jogos*, tanto cooperativos como não cooperativos, os quais permitem um nível de abstracção maior dos dados em estudo. Todos tentam fornecer informação económica fiável para o mercado.

Capítulo 6

Repartição Automática de Perdas

Neste capítulo é apresentada uma proposta para a estimativa da contribuição dos consumidores BT para as perdas na rede. Genericamente pretende-se obter valores de perdas aproximados aos que se obteriam utilizando informação dos diagramas completos, a qual que não está habitualmente disponível para dados [BT]. Pretende-se estimar a contribuição de cada consumidor para as perdas em função da energia global da rede e a energia consumida e dos factores caracterizadores de diagramas. Esta metodologia aplica técnicas de DM na criação e caracterização de um conjunto de consumidores que represente a amostra de dados existente, de modo a ser possível caracterizar qual a sua contribuição para as perdas na rede onde estão inseridos. Este capítulo encontra-se organizado da seguinte forma: procedeu-se à introdução e à descrição dos dados a analisar, seguiu-se o processo de preparação dos dados e de selecção de atributos, cálculo dos factores de perda e testados os modelos em estudo.

Nas secções anteriores foram utilizadas diversas técnicas de *DM* para o cálculo de perfis de consumidores, os quais deram resultados aceitáveis do ponto de vista de funcionalidade. Neste trabalho foram testados e utilizados algoritmos de regressão, *clustering* e Redes Neurais.

6.1 Descrição dos Dados

Os dados utilizados neste estudo foram disponibilizados pela empresa responsável pela distribuição de electricidade em Portugal (*EDP Distribuição*).

Os dados relativos aos diagramas são parte de uma campanha de medição, neste caso localizada numa rede semi-urbana, para a qual se possui todos os parâmetros físicos e eléctricos necessários para executar estudos de trânsito de potência. A rede escolhida (fig 6.1) foi aquela de entre as disponíveis que se encontra melhor caracterizada em termos de registos (diagramação) de consumos e dados físicos respeitantes à cablagem. Ao possuir todos os dados integrantes da rede é possível calcular as perdas utilizando programas de

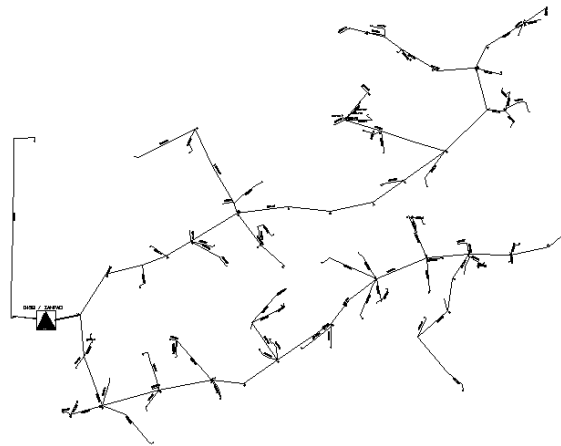


Figura 6.1: Esquema da Rede seleccionada

trânsito de potência.

Os dados foram fornecidos em modo *Raw* em formato TXT, os quais foram descarregados no formato original do diagramador em que cada ficheiro corresponde a um consumidor diferente. Dependendo do diagramador o formato dos dados são diferentes. Deste modo, os dados foram todos transformados, filtrados e ajustados em *Comma Separated Values* [CSV], com intervalos de 15 em 15 minutos, referentes ao período de contagem em que se apresentavam.

Assim, para cada consumidor [BT], a informação disponibilizada consiste em:

- número de identificação do consumidor;
- número do diagramador instalado;
- data de início e de fim do período de monitorização;
- leituras da potência activa instantânea consumida, com uma cadência de 15 minutos, durante a totalidade do período em que o consumidor foi monitorizado;
- energia mensal consumida, em kWh, no período em estudo;
- alguma informação comercial: tipo de utilização [BT], tipo de tarifa e potência contratada, também denominados *dados de facturação*.

Neste estudo foram utilizados os consumos diários de um conjunto de consumidores na rede seleccionada para monitorização. Através de *software* de cálculo de trânsito de potências, implementado com uma abordagem do algoritmo *Newton-Raphson* [49] foram calculadas as perdas da rede com todos os consumidores com as cargas sincronizadas e correspondentes ao momento intervalo temporal.

Tabela 6.1: Amostra da organização dos dados de entrada

Energia Cons.(kWh)	Factor de Perdas	Factor de Carga	Energia Rede (kWh)	N.Cons.
7.689	0.06133067	0.1878390	327.7098	25
6.713	0.09492579	0.2621446	309.4533	26
6.091	0.04835500	0.1778862	316.6887	27
5.613	0.06526763	0.2143184	358.4259	28
6.311	0.12439337	0.3040274	332.2797	29
6.000	0.18756263	0.3890294	369.1122	30
6.452	0.13523845	0.3260562	348.0003	31
14.258	0.04537836	0.1284875	323.1654	32
10.700	0.06545573	0.1764948	359.3274	33
8.033	0.02864843	0.1136000	374.9271	35
7.136	0.08673318	0.2563310	306.8592	36
6.313	0.16091011	0.3556419	362.5440	37
6.076	0.16463925	0.3641813	332.5107	38
7.754	0.19577887	0.3752965	338.1792	39
8.231	0.11960675	0.3085661	317.7390	40
7.535	0.05859856	0.1981643	278.7225	41

Assim, de cada resultado e para cada unidade de tempo ¹ foram calculados os valores correspondentes às perdas.

A partir dos diagramas foram calculados o factor de carga e factor de perdas para cada dia em relação a cada consumidor seleccionado na amostra no momento temporal; para cada dia foi calculado o valor de energia de cada consumidor e a energia total do conjunto da rede no próprio dia.

Estas quatro variáveis determinadas a partir dos diagramas originais são utilizadas como variáveis de entrada no cálculo da contribuição das perdas.

Através dos diagramas foi alimentada a rede seleccionada (ver figura 6.1) e calculadas as perdas repartidas por *repartição quadrática* relativas a cada período de tempo, ao fazer incidir a cada consumidor com o respectivo valor de potência nesse período.

Em resumo, a metodologia proposta engloba os seguintes passos:

1. selecção da rede e dos consumidores;
2. análise do fluxo de carga e cálculo das perdas globais;
3. repartição quadrática das perdas;
4. cálculo do factor de carga e factor de perdas.

6.1.1 Limpeza dos Dados

A fase de limpeza de dados é essencial, para que dados omissos e anómalos não afectem as fases seguintes do processo de caracterização de consumidores. Neste conjunto de

¹Os diagramadores registam o valor de potência de 15 em 15 minutos

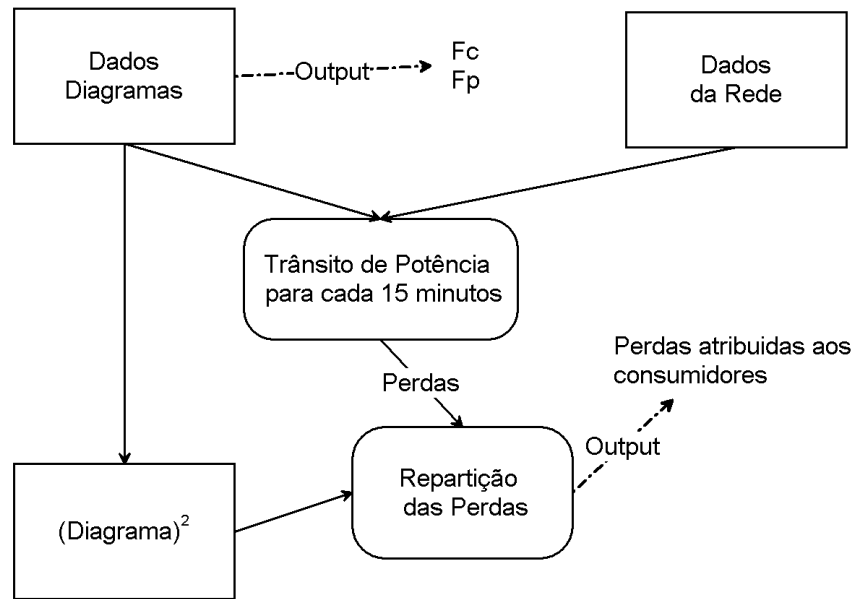


Figura 6.2: Sequência de Operações

dados foram detectadas situações em que, para determinados consumidores, existiam vários dias com valores de leitura nulos, sendo mantido no conjunto inicial devido a serem situações raras representando menos de 3 % dos dados totais. No entanto tiveram de ser recalculados alguns valores por interpolação de modo a não haver falhas temporais no conjunto, permitindo dessa forma manter a integridade dos dados em termos temporais.

6.1.2 Pré-Processamento dos Dados

O conjunto de dados obtido é função do número de consumidores que constituem a amostra, da cadência de recolha de informação e do número total de dias que constitui o intervalo total de tempo das campanhas de medição.

O volume de dados relativo às cargas, (sem contar com os dados da rede, os dados de facturação, entre outros), constituído pelo número de diagramas de carga diário, é igual ao número de consumidores da amostra multiplicado pelo número total de intervalos temporais.

Neste estudo, a cadência de recolha de informação utilizada foi um intervalo de 15 minutos, o que faz com que cada diagrama de carga diário seja representado por 96 registos do valor da potência consumida, num intervalo de tempo com início às 00:00 horas e final às 23:45 do mesmo dia. Para cada consumidor, existe um número de diagramas de carga diários correspondente ao número de dias que constitui o período de monitorização. Para esta análise foi utilizada uma *janela temporal* de três meses para cada consumidor.

6.2 Construção do conjunto de padrões

O modelo criado vai permitir calcular a contribuição dos consumidores para as perdas na rede em que estão inseridos. Diversos modelos foram testados incluindo modelos de regressão linear LM e redes neuronais. Cada um dos modelos *responde* de maneira diferente e com valores de erro diferentes. Assim, inicializa-se a rede com os consumidores no período a analisar (neste caso 30 dias). Para cada intervalo (15 minutos) a partir dos diagramas diários, calcula-se o fluxo de carga e é elaborado o cálculo das perdas.

A cada consumidor k atribuem-se as perdas P_k :

$$P_k = \frac{E_k^2}{\sum_i E_i^2} \times Perdas\ Totais \quad (6.1)$$

Onde: E_k^2 é a energia consumida pelo consumidor k no período temporal; P_k perdas atribuídas ao consumidor k no mesmo período temporal.

Desse valor reparte-se e calcula-se de forma quadrática o valor das perdas, no caso da rede em estudo através da aplicação de um trânsito de potência baseado no algoritmo de *Newton-Raphson* [49].

Com estes valores foi construído um ficheiro de padrões o qual serve para *alimentar* os diversos modelos de classificação. Nesse ficheiro intermédio constam os valores dos factores de carga, factores de perda, a energia total da rede no dia, a energia do diagrama do dia vindos do cálculo do trânsito de potência. Cada conjunto desses 4 valores corresponde aos dados *agregados* de um dia. O ficheiro de saída é constituído pela lista das perdas atribuídas segundo a equação 6.1. Como foram considerados dados de N consumidores durante 30 dias, resulta $N \times 30$ padrões. Destes, metade foi definido como conjunto de treino e a outra parte para testes. Esta divisão foi feita de modo a validar de modo mais credível a capacidade de generalização.

6.3 Modelos Testados

Dentro desta secção são descritos os modelos para a atribuição das perdas com base nas características do diagrama e medidas de energia. A partir dos dados de entrada da tabela 6.1 obtêm-se estimativas das perdas atribuídas a cada consumidor segundo o processo descrito anteriormente 6.2.

De acordo com os resultados dos modelos anteriores em conjuntos de dados semelhantes, os quais deram resultados francamente positivos, foram utilizados apenas os modelos a seguir descritos.

A aplicação do modelo de classificação aos dados foi realizada em duas fases. Numa primeira fase, o modelo foi *treinado* com uma parte do conjunto de dados, designada por

conjunto de treino. Após o treino, obteve-se o modelo que é representado por uma função de regressão linear ou uma *Rede Neuronal*. Na segunda fase, foi realizado o teste do modelo criado com outra parte do conjunto de dados, designada por conjunto de teste, de modo a obter um erro de generalização mais realista.

6.3.1 Modelo de Regressão Linear

A regressão linear é um método normalmente utilizado para análise da variância e co-variância de conjuntos.

Este modelo pode ser visto como uma extensão de uma regressão linear múltipla para uma só variável dependente e perceber o modelo de regressão múltipla é essencial para o entender. O objectivo da regressão múltipla é quantificar a relação entre várias variáveis independentes e uma variável dependente.

Os dados de entrada encontram-se organizados tal como na tabela 6.1. Estes dados são combinados com o valor das perdas atribuídas as cada consumidor em cada dia.

O objecto construído no R [50] por este processo contém os seguintes índices :

- coeficientes: vector com os coeficientes;
- resíduos: resposta menos valores ajustados;
- *rank*: posição de ordem no modelo linear;
- pesos: pesos iniciais especificados;
- *residual degrees of freedom*: graus de liberdade residuais;
- termos: termos utilizados;
- model: modelo utilizado.

Estes índices caracterizam a função encontrada em conjunto com os erros quadráticos e com os pesos iniciais. A saída obtida é uma função linear das entradas. Este modelo foi utilizado para testar a linearidade dos factores a relacionar. Posteriormente, para validar o modelo aplicou-se aos dados reservado para teste. Na figura 6.3 podemos ver uma comparação dos valores originais (pontos) com os valores obtidos por [LM] na linha contínua para o conjunto de teste.

6.3.2 Rede Neuronal SLP

Nesta secção descreve-se a aplicação de uma rede neuronal tipo [SLP] (ver 3.5.3.1) aos dados do caso em estudo. Após testar diversas configurações optou-se por uma rede com topologia 4-13-1 baseada em *feed forward* com função de transferência de tangente

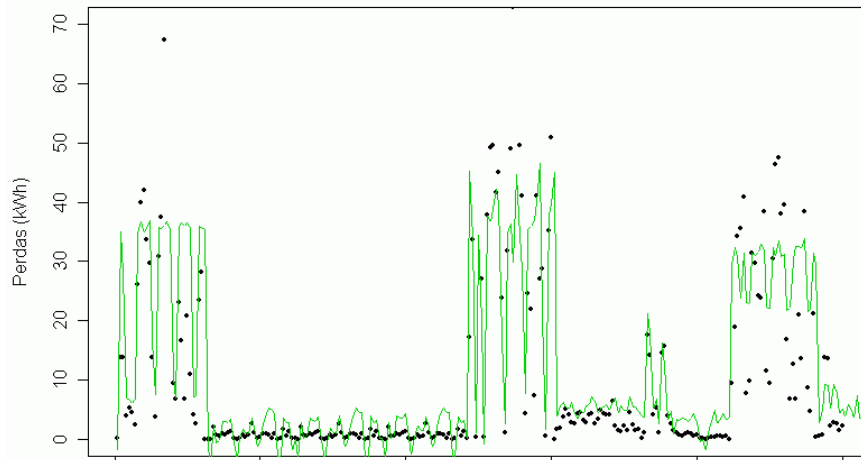


Figura 6.3: Processo com Modelo de Regressão

hiperbólica (*Tan sig*) e otimizada através do método *[BFGS]*, denominado por *algoritmo métrico variável (quasi-Newton)*, publicado em 1970 por Broyden, Fletcher, Goldfarb and Shanno [51]. Este modelo utiliza valores e gradientes da função inicial para construir uma forma da *superfície* a ser otimizada, ajustando-se dinamicamente em cada passagem. Assim, *alimentando* a rede *caso a caso* em que cada linha é um registo com o conjunto dos 4 valores de entrada correspondentes a cada dia com o factor de carga, factor de perdas, energia consumida pelo consumidor em causa e energia total da rede. Como função *target* está o valor correspondente às perdas atribuídas a cada consumidor naquele momento calculadas segundo o método descrito em 6.2. Em cada passagem os pesos são ajustados de modo a ser minimizado o valor do erro.

As redes neuronais utilizadas devolvem um conjunto de valores correspondentes ao número de casos que analisou, o conjunto dos melhores pesos calculados, e o conjunto de erros para cada caso (*residuals*).

Após o treino da rede é possível averiguar qual a sua capacidade de aproximação da rede aos valores iniciais como se pode ver na figura 6.5.

É possível verificar que, de um modo geral, a rede neuronal apresenta uma saída bastante próxima do valor desejado. No entanto, como se pode observar no esquema 6.5 existem valores muito elevados os quais foram obtidos na situação em que um único consumidor absorve uma percentagem muito elevada (cerca de 25%) da carga total da rede. Esse caso constitui uma situação rara, cuja probabilidade de ocorrência é extremamente

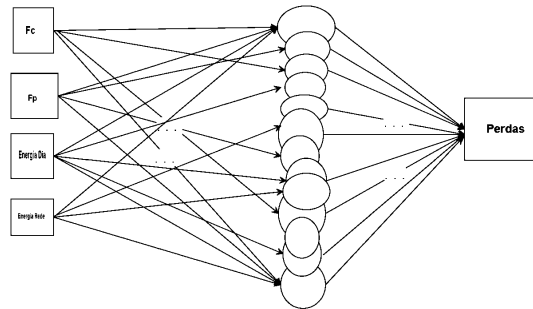


Figura 6.4: Topologia de uma Rede Neuronal 4-13-1

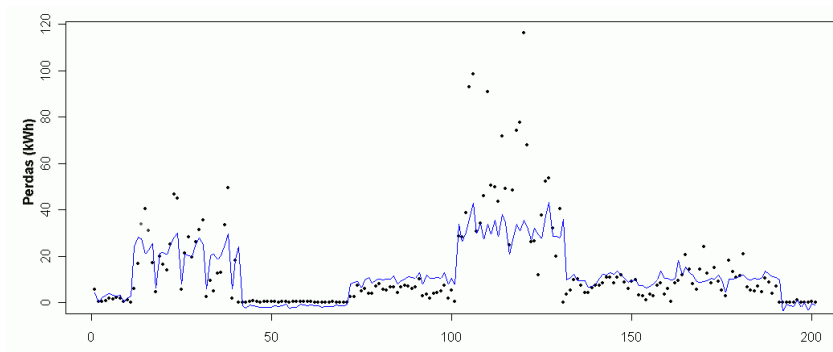


Figura 6.5: Aproximação com rede Neuronal SLP

baixa numa rede *BT*. A topologia (4-13-1) foi escolhida por *Trial and Error*, na qual ao alterar o número de unidades na camada escondida se tenta balançar complexidade da rede neuronal com a aproximação ao valor resultado. Assim, depois dos testes com diferentes topologias, chegou-se a esta, a qual consegue manter um nível de aproximação relativamente elevado, sem causar problemas de *over-fitting*, o que pode ser comprovado através da análise do erro médio.

No entanto, outra topologia pode obter, para o mesmo conjunto de treino e de testes, valores mais ou menos aproximados mas com tempos de execução diferentes. Deste modo, uma topologia mais simples implicaria à partida a necessidade de um maior número de épocas de treino para obter valores próximos dos obtidos.

6.3.3 Rede Neuronal MLP

Neste caso foi utilizada uma rede MLP (ver 3.5.3.2) com duas camadas escondidas com a topologia 4-10-5-1 treinada com base no algoritmo *Adaptive Back Propagation Network*. (ver [52]). Foi escolhida uma rede com uma topologia (4-10-5-1) (fig 6.6) pois esta foi a que respondeu melhor ao conjunto de dados após diversas tentativas com diferentes topologias.

Tabela 6.2: Erros Médios dos métodos abordados

Neural Network MLP	Neural Network SLP	Linear Regression
0.19889	0.27880	2.53076

Do mesmo modo que no método anterior (6.3.2) a rede é alimentada registo a registo com o conjunto dos quatro valores de entrada correspondentes a cada dia com factor de carga, factor de perdas, energia consumida pelo consumidor e energia da rede nesse dia. Como função *target* está o valor correspondente às perdas atribuídas a cada consumidor naquele momento. Em cada passagem os *pesos* são ajustados de modo a ser minimizado o valor total do erro.

Esta abordagem, a rede neuronal *MLP*, tem como inconveniente o aumento do peso do cálculo computacional dado apresentar uma topologia mais elaborada.

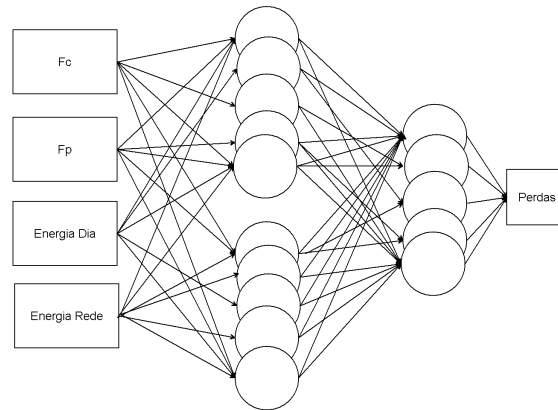


Figura 6.6: Esquema da topologia da rede MLP utilizada

Como se pode verificar o conjunto de métodos consegue responder positivamente e é capaz de prever qual o valor da perdas baseado em dados históricos.

6.4 Conclusões

Os estudos realizados permitem concluir que existe uma forte relação entre o factor de perdas F_p , energia, factor de carga F_c e as perdas atribuídas. Os resultados obtidos sugerem que é possível estimar a repartição de perdas baseada nas características dos diagramas reais ou dos perfis atribuídos e a partir destes obter estimativas sobre as contribuições dos consumidores para as perdas na rede.

Deste modo é possível estimar as perdas atribuídas a cada consumidor com base nos factores caracterizadores dos diagramas e nas medidas de energia. Este facto permite estabelecer diversas metodologias de alocação relativamente simples de implementar. Através

da comparação dos métodos testados chegou-se á conclusão que a rede neuronal é a que apresenta melhores índices, como se pode ver na tabela 6.2, com os valores de erro absoluto médio. Verifica-se que de facto em termos gerais a rede neuronal MLP permite obter uma estimativa das perdas referente a cada consumidor bastante próxima do valor desejado.

Capítulo 7

Conclusão

Este trabalho teve como objectivo a análise dos consumidores de Baixa Tensão e a caracterização das suas contribuições para as perdas na rede onde se inserem.

Utilizando técnicas de *Data Mining*, tentou-se observar a existência de inter-relações entre os diversos factores de caracterização dos consumidores e a sua contribuição para as perdas.

Numa primeira fase e devido á não-homogeneidade dos dados foi necessário utilizar *clustering* de modo a eliminar *outliers* e agrupar em conjunto os diversos registos com características semelhantes de modo a obter perfis típicos. Após diversas experiências, o algoritmo *CLARA* foi o que respondeu com melhores resultados na construção do perfil típico. Contudo o processo de atribuição de perfis permitirá a atribuição de um perfil a cada consumidor, e consequentemente, a determinação dos respectivos factores caracterizadores. Deste modo e considerando que estes factores estão facilmente disponíveis, a metodologia proposta poderá ser aplicada de um modo genérico .

Numa segunda fase a partir dos dados reais fez-se a caracterização dos consumidores através dos factores de perda e carga juntamente com as medidas de energia. Normalmente os diagramas reais não estão disponíveis no nível de baixa tensão.

Nos testes efectuados foi possível estimar as perdas atribuídas a cada consumidor com base nos factores caracterizadores dos diagramas e nas medidas de energia. Como ferramentas de regressão foram utilizadas a regressão linear e redes neuronais, tendo estas ultimas conduzido a resultados de melhor qualidade. Deste modo podem-se analisar quais as relações dos diversos factores do consumidor final e como este contribui na rede relativamente às perdas. Assim, ficou demonstrado que é possível caracterizar qual o papel dos consumidores de baixa tensão no que respeita às suas contribuições para as perdas na rede, através de métodos relativamente simples e generalizável.

7.1 Futuros Desenvolvimentos

Em trabalhos subsequentes seria interessante experimentar outros factores além dos utilizados e obter desta forma resultados mais conclusivos. Seria conveniente analisar a rede em termos dinâmicos, incluindo, por exemplo, novos perfis, de forma a verificar os impactos ao longo do tempo da sua utilização exclusiva. Os esforços de pesquisa deste trabalho mostram que alguns tópicos relevantes ainda necessitam de mais investigações, destacando-se:

- Elaboração de novos testes com outros algoritmos de repartição de perdas.
- Inclusão, no modelo de cálculo de perdas a influência da temperatura, a qual como se sabe actua directamente nas cargas e dado que a passagem de corrente num condutor provoca o seu aquecimento e consequentemente, a alteração no valor da sua resistência, influenciando o cálculo das perdas.
- Extensão dos testes realizados noutras redes.
- Exploração dos mesmos dados através de outras técnicas não-lineares como *Neural Fuzzy*, Árvores de Regressão entre outras.
- Agregação de outro tipo de informação relativa a cada consumidor dando lugar a novos factores e a novas vias de exploração.

Referências

- [1] Manuel Matos; A. Gomes Martins; J. Nuno Fidalgo; Humberto Jorge. Caracterização de consumidores e redes, parte 1-caracterização de consumidores e redes, projecto ccr, relatório técnico. Technical report, INESC Porto, INESC Coimbra, 2005.
- [2] Marija Llic; Jason W. Black; Jill L. Wat. *IEEE Transactions on Power Systems*, 2002.
- [3] I. Silvas G. Tupu P. Stephenson;; I. Lungu, M. Paun;. *CIREN*, pages 18–21, 2001.
- [4] Ernoult M. ; Meslier F. *Analysis and Forecast of Electrical Energy Demand*. Number 4. RGE, 1982.
- [5] ERSE. Despacho erse n.º 14 / 2003, tarifas e preços para a energia eléctrica e outros serviços em 2004. Technical report, ERSE, Novembro 2003.
- [6] Allera S. V. ; Horsburgh A. G. Load profiling for the energy trading and settlements in the uk electricity markets. London, 1998. Proc. Distributed Europe DA/DSM Conference.
- [7] Monticelli A. Electric power systems state estimation. volume 88 of 2, pages 262–282. IEEE, February 2000.
- [8] Manuel Matos; A. Gomes Martins; J. Nuno Fidalgo; Humberto Jorge. Caracterização de consumidores e rede, parte 2-alocação de perdas, projecto ccr, relatório técnico. Technical report, INESC Porto, INESC Coimbra, 2005.
- [9] Schick I. C.; Usoro P. B.; Ruane M. F.; Schweppe F. C. Modelling and weather-normalisation of whole-house metered data for residential end-use load shape estimation. *IEEE Transactions on Power Systems*, 3(1):213–219, 1988.
- [10] Schick I. C.; Usoro P. B.; Ruane M. F.; Hausman J. A. Residential end-use load shape estimation from whole-house metered data. *IEEE Transactions on Power Systems*, 3(3):986–991, 1988.

- [11] Calloway T. M. ; Brice C. W. Physically based model of demand with applications to load management assessment and load forecasting. *IEEE Transactions on Power Apparatus and Systems*, PAS-101(12):4625–4631, 1982.
- [12] Chong C. ; Malhami R. P. Statistical synthesis of physically based load models with applications to cold load pickup. In IEEE, editor, *IEEE Transactions on Power Apparatus and Systems*, volume PAS-103 of 7, pages 1621–1627. IEEE, 1984.
- [13] Muller H. Classification of daily load curves by cluster analysis. pages 381–388. 8's Power Systems Computation Conference, 1990.
- [14] Pitt; B.D. *Applications of Data Mining Techniques to Electric Load Profiling*. Electric and electronic engineering, University of Manchester Institute of Science and Technology, Manchester, 2000.
- [15] Gasperic S. ; Gerbec D. ; Gubina F. Determination of the consumers load profiles. *Telmark discussion forum*, September 2002. www.telmark.org/index.htm.
- [16] Chicco G. ; Napoli R. ; Postulache P. ; Scutariu M. ; Toader C. Customer characterization options for improving the tariff offer. *IEEE Transactions on Power Systems*, 18(1):381–387, February 2003.
- [17] Proc. RIMAPS. *Stratified Sampling of the Electricity Customers for Setting up a Load Profile Survey*, Madeira, Portugal, Sept. 25-28 2000. Bompard E; Carpaneto E. ; Chicco G. ; Napoli R. ; Piglion F. ; Postulache P. ; Scutariu M.
- [18] Pao Y-H ;Sobajic; D. J. Combined use of unsupervised and supervised learning for dynamic security assessment. volume 7, pages 878–884. *IEEE Transactions on Power Systems*, IEEE, May 1992.
- [19] Glymour C. ; Madigan D. ; Pregibon D. ; Smyth P. Statistical inference and data mining. *Communications of the ACM*, 39(11), November 1996.
- [20] Rodrigues F. *Arquitetura Heterogenea para Extração de Conhecimento a partir de Dados*. PhD thesis, Universidade do Minho, December 2000.
- [21] Agrawal R.; Gupta A.; Sarawagi S. Modelling multidimensional databases. pages 232–243, Santiago, Chile, 1997. IEEE.
- [22] Rodrigues F. ; Ramos C. ; Henriques P. R. Knowledge discovery with retail basket analysis. Brasil, 1998. The First International Conference on Data Mining.
- [23] Michie D. ; Spiegelhalter D. J. ; Taylor C. C. *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, New York, 1994.
- [24] Fisher D. H. *Knowledge Acquisition via Incremental Conceptual Clustering*. Stanford, Stanford University, 1987.

- [25] Gouveia; J. et al. Figueiredo V. ; Duarte F J. ; Rodrigues F. ; Vale Z. Electric energy customer characterization by clustering. accepted for publication in ISAP Competitive Environment, Renewable Energy, Distributed Generation, Lemnos, Greece, September 2003.
- [26] Fan J. Local modelling. *Encyclopaedia of Statistical Science*, 1995.
- [27] Morgan J. ; Sonquist J. Problems in the analysis survey data and a proposal. *Journal of the American Statistical Association*, 58:415–434, 1963.
- [28] Friedman J. Multivariate adaptative regression splines. *Annals of Statistics*, 1(19):141, 1991.
- [29] Carvalho D. R. ; Freitas A. A hybrid decision tree / genetic algorithm for coping with the problem of small disjuncts in data mining. pages 1061–1068, Las Vegas, USA, July 2000. Genetic and Evolutionary Computation Conference.
- [30] Zimmermann H. J. *Fuzzy Sets Theory and its Applications*. Kluwer-Nijhoff, Boston, EUA, 1985.
- [31] Nauck D. ; Klawonn F. ; Kruse R. *Foundations of Neuro-Fuzzy Systems*. Wiley, Chichester, 1997.
- [32] Fidalgo J. N. ; Leão M. T. ; Matos M. Assessing error bars in distribution load curve estimation. *Lecture Notes in Computer Science*, pages 1017–1022, 1997.
- [33] Pawlak Z. Rough sets. *International Journal Computer Information Science*, 11:341–356, 1982.
- [34] Pal S. ; Skowron A. *Rough Fuzzy Hybridization*. Springer-Verlag, 1999.
- [35] Bazan; J. G. Dynamic reduce as a tool for extracting laws from decision tables. In Charlotte, editor, *Proc. of the Symp. On Methodologies for Intelligent Systems*, pages 346–355, Berlin, 1998. Springer-Verlag.
- [36] Polkowski L. ; Skowron A. *Rough Sets and Current Trends in Computing*, volume 1 of *Lecture Notes in Artificial Intelligence*. Springer, 1998.
- [37] Browe C. ; Duntsch I. ; Gediga G. *Rough Sets in Knowledge Discovery*, volume 2, chapter IRIS revisited: A comparison of discriminante and enhanced rough data analysis, pages 345–368. Heidelberg:Physica-Verlag, 1998.
- [38] Berkhin P. *Survey of Clustering Data Mining Techniques*. Accrue Software, 2002.
- [39] Sneath P. H. ; Sokal R. R. *Numerical Taxonomy*. Freeman, London, UK, 1973.
- [40] Ward J. H. Jr. Hierarchical grouping to optimise an objective function. *Joint American Statistics Association*, 1(58):236–244, 1963.

- [41] Leonard Kaufman; Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience, 1990.
- [42] Buhmann J. Complexity optimized data clustering by competitive neural networks. *Neural Computation*, 5:75–88, May 1993.
- [43] Brown D. ; Huntley C. A practical application of simulated annealing to clustering. Technical Report IPC-TR-91-003, University of Virginia, Virginia, 1991.
- [44] Babu G. P. ; Murty M. N. A near-optimal initial seed value selection in k-means algorithm using a genetic algorithm. In *Pattern Recogn Letters*, volume 14, pages 163–169, 10 1993.
- [45] João Tomé Saraiva; José Pereira Silva; Teresa Ponce de Leão. *Mercados de Electricidade - Regulação e Tarificação de Uso das Redes*. FEUP Edições, 2002.
- [46] J.w. Bialek. Tracing the flow of electricity. *IEE Procedures Generation Transm. Distrib*, 143(4):313–320, July 1996.
- [47] D.; R. Allan Kirschen and G. Strbac. Contribution of individual generations to loads and flow. *IEEE Trans. on Power Systems*, 12(1):52–60, February 1997.
- [48] C. N. Macqueen and M. R. Irvin. An algorithm for the allocation of distribution system demand and energy losses. *IEEE Trans. Power Syst.*, 11:338–343, February 1996.
- [49] Jonh J. Grainger; William D. Stevenson Jr. *Power Systems Analysis*. McGraw-Hill, 1994.
- [50] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2004. ISBN 3-900051-07-0.
- [51] Broyden CG. The convergence of a class of double-rank minimization algorithm. the new algorithm. *Inst. Math Applications*, (6):222–231, 1970.
- [52] S. Russel and P. Norvig. *Artificial Intelligence: A Modern Approach*. Morgan Kaufmann Publishers, 1994.