

# UTILIZAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS NA CARACTERIZAÇÃO PERFIS TÍPICOS DE CARGA

Samuel Sandmann Cembranel



Departamento de Engenharia Eletrotécnica  
Mestrado em Engenharia Eletrotécnica – Sistemas Eléctricos de Energia

2019



Relatório elaborado para satisfação parcial dos requisitos da Unidade Curricular de DSEE - Dissertação do Mestrado em Engenharia Eletrotécnica - Sistemas Elétricos de Energia do Instituto Superior de Engenharia do Porto (ISEP/IPP) e do Trabalho de Conclusão de Curso do Curso de Graduação em Engenharia Elétrica do Instituto Federal de Educação, Ciência e Tecnologia de Santa Catarina (IFSC – Câmpus Florianópolis). Este trabalho foi elaborado no âmbito do acordo internacional de Dupla Titulação entre o Instituto Superior de Engenharia do Porto (Portugal) e o Instituto Federal de Educação, Ciência e Tecnologia de Santa Catarina (Brasil) como parte dos requisitos para obtenção do título de Mestre em Engenharia Eletrotécnica - Sistemas Elétricos de Energia pelo ISEP/IPP e de Engenheiro Eletricista pelo IFSC.

Candidato: Samuel Sandmann Cembranel, Nº 1171912, [1171912@isep.ipp.pt](mailto:1171912@isep.ipp.pt)

Orientação científica: Sérgio Ramos, [src@isep.ipp.pt](mailto:src@isep.ipp.pt) (ISEP)

Coorientação científica: Fernando Lezama, [flzcl@isep.ipp.pt](mailto:flzcl@isep.ipp.pt) (ISEP)

Orientação científica: Rubiara Cavalcante Fernandes, [piara@ifsc.edu.br](mailto:piara@ifsc.edu.br) (IFSC)



Departamento de Engenharia Eletrotécnica

Mestrado em Engenharia Eletrotécnica – Sistemas Elétricos de Energia

**2019**



Ficha de identificação da obra elaborada pelo autor.

Cembranel, Samuel

**Utilização de Técnicas de Mineração de Dados na Caracterização de Perfis Típicos de Carga / Samuel Cembranel ; orientação de Rubiara Cavalcante Fernandes.** - Florianópolis, SC, 2019.

131 p.

**Trabalho de Conclusão de Curso (TCC) - Instituto Federal de Santa Catarina, Câmpus Florianópolis. Bacharelado em Engenharia Elétrica. Departamento Acadêmico de Eletrotécnica.**

Inclui Referências.

1. **Descoberta do Conhecimento em Banco de Dados.**
2. **Mineração de Dados.** 3. **Agrupamento de Dados.** 4. **Classificação.**
5. **Perfis Típicos de Carga.** I. Cavalcante Fernandes, Rubiara. II. Instituto Federal de Santa Catarina. Departamento Acadêmico de Eletrotécnica. III. Título.

**ATA DE PROVAS DE MESTRADO**Nº: MEESEE / 1 / 2019Data: 2019 - 02 - 01

<b>Hora de Início:</b>	09:30	<b>Elaborado por:</b>	Teresa Alexandra Ferreira Mourão Pinto Nogueira
<b>Hora de Fim:</b>	10:30	<b>Data:</b>	2019-02-01

Presentes
Presidente do Júri: Teresa Alexandra Ferreira Mourão Pinto Nogueira
Orientador: Sergio Filipe Carvalho Ramos
Vogal: Paulo Jorge Machado Oliveira
Vogal: Rubiapiara Cavalcante Fernandes

**Assuntos Tratados:**

Ata da Reunião do Júri para discussão da Dissertação apresentada pelo Licenciado SAMUEL SANDMANN CEMBRANEL, para obtenção do grau de Mestre em Engenharia Eletrotécnica - Sistemas Elétricos de Energia - Mestrado - 2011/2012, intitulada: "Utilização de Técnicas de Mineração de Dados na Caracterização de Perfis Típicos de Carga".

Ao um dia do mês de fevereiro do ano dois mil e dezanove na sala F503 do ISEP, nos termos do Decreto-Lei número setenta e quatro de vinte e quatro de março de dois mil e seis, republicado no anexo ao Decreto-Lei número cento e quinze de sete de agosto de dois mil e treze, reuniu para discussão da dissertação acima referida, o júri nomeado para o efeito com a presença dos elementos acima indicados.

As provas tiveram início às nove horas e trinta minutos, com o Presidente do júri cumprimentando os restantes membros, agradecendo a sua colaboração e desejando felicidades ao candidato.

De seguida, o candidato cumprimentou o júri e passou à apresentação oral da sua dissertação, com duração de vinte minutos.

Foi arguente Paulo Jorge Machado Oliveira, que fez uma apreciação global do trabalho e colocou algumas questões ao candidato, a que este teve oportunidade de responder.

Terminada a discussão, o júri procedeu nos termos da lei (artigos vinte e dois e vinte e quatro do decreto lei citado), deliberando aprovar, por unanimidade, o candidato com a classificação de dezanove (19) valores, tendo o mesmo júri fundamentado esta classificação do seguinte modo:

**ANÁLISE DO RELATÓRIO ESCRITO (40%):** Conteúdo, redação, ilustrações, adequação da metodologia desenvolvida ao problema e resultados - 18,86 valores;

**RELEVÂNCIA DO TRABALHO (30%):** Relevância social e ambiental, criatividade, Qualidade técnica / exigência programática 18,40 valores;

**APRESENTAÇÃO E DEFESA PÚBLICA (30%):** Qualidade dos meios, clareza da apresentação, capacidade de argumentação, domínio das temáticas relacionadas com o trabalho 18,75 valores;

Avaliação global da dissertação de Mestrado:  
 $0,4 \times 18,86 + 0,3 \times 18,4 + 0,3 \times 18,75 = 18,7$  valores

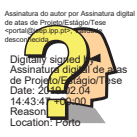
**ATA DE PROVAS DE MESTRADO**

Nº: MEESEE / 1 / 2019

Data: 2019 - 02 - 01

Não havendo outros assuntos a tratar, a reunião foi dada por encerrada às dez horas e trinta minutos.

Da reunião foi exarada a presente ata que, depois de lida em voz alta, vai ser assinada por todos os membros do júri.

Aprovado por:	Data
<p>Ata assinada digitalmente.</p>  <p>Assinatura do autor por Assinatura digital de atas de Provas/Exames/Teses e trabalhos de curso Digitaly signed by Assinatura digital de atas de Provas/Exames/Teses Date: 2019.02.04 14:43:47 -0500 Reason: Location: Porto</p>	2019-02-01

“Faça o que seu coração acha certo, pois de qualquer forma você será criticado. Você será condenado quer faça ou não.” (Eleanor Roosevelt)





## *Agradecimentos*

Gostaria de agradecer, acima de tudo, a meus pais e minha irmã por todo o apoio durante o desenvolvimento desse trabalho.

Esse trabalho foi desenvolvido no âmbito do programa de Dupla Titulação entre o Instituto Federal de Santa Catarina e o Instituto Superior de Engenharia do Porto, então gostaria de agradecer a todos envolvidos na realização e desenvolvimento do programa, pois foram essenciais para o desenvolvimento desse trabalho.

A todos os meus amigos que estiveram presentes durante a realização desse trabalho, gostaria de agradecê-los, pois foram essenciais em momentos difíceis.

Agradeço ao meu orientador, em Portugal, Professor Doutor Sérgio Ramos, e ao meu orientador no Brasil, Professor Doutor Rubiara Cavalcante Fernandes, pela orientação prestada durante o desenvolvimento do trabalho. Gostaria de agradecer também ao meu coorientador, Doutor Fernando Lezama, por todo o auxílio prestado durante o desenvolvimento deste trabalho.

Por fim, muito obrigado a todos que contribuíram nesse trabalho.



## *Resumo*

O conhecimento dos hábitos de consumo de energia elétrica tem-se mostrado uma ferramenta importante nos diversos setores elétricos. A liberalização do setor elétrico, em Portugal e no resto do mundo, culminou no surgimento de novos agentes, que aumentaram a competitividade, sobressaindo aqueles que conseguem fornecer serviços de qualidade a preços baixos.

Nesta dissertação é proposto e avaliado um modelo de caracterização de curvas típicas de carga para consumidores de baixa tensão. A identificação dos padrões de consumo é baseada na aplicação de algoritmos de agrupamento. A base de dados consiste em dados de consumo de energia elétrica de 194 clientes de baixa tensão, localizados nas cidades do Porto, Matosinhos e Vila Real. Com o conhecimento obtido na etapa de agrupamento é elaborado um modelo de classificação, capaz de classificar novos consumidores de acordo com seus dados de consumo. A metodologia de agrupamento é baseada em sete algoritmos particionais e hierárquicos, juntamente com seis índices de validação de agrupamento, capazes de identificar a melhor partição dos dados. Para finalizar o ciclo do reconhecimento de padrões é utilizado um modelo de classificação baseado em árvores de decisão para classificar novos consumidores. Para tornar o modelo simples cada curva de carga é representada por cinco índices capazes de representar o formato das curvas de carga. A metodologia proposta nesse trabalho demonstra ser uma ferramenta eficaz que pode ser utilizada nos mais diversos setores, destacando-se a utilização do conhecimento na otimização da contratação de energia para clientes de baixa tensão. Os dados dos consumidores podem ser constantemente atualizados na tentativa de melhorar o modelo obtido nesse trabalho, obtendo estimativas que consigam representar melhor os consumidores e seus hábitos de consumo.

### *Palavras-Chave*

Descoberta do Conhecimento em Banco de Dados, Mineração de Dados, Agrupamento de Dados, Classificação, Perfis Típicos de Carga.



## *Abstract*

The knowledge of electricity consumption's habits has been an important tool in electrical sectors. The constant liberalization of the electricity sectors, in Portugal and the rest of the world, culminated in the emergence of new agents, which increased the competitiveness, standing those that can provide quality services at low prices.

In this dissertation, a characterization model of typical load curves for low voltage consumers is proposed and evaluated. The identification of consumption patterns is based on clustering analysis. The database consists of electricity consumption data of 194 low voltage consumers, located in the cities of Porto, Matosinhos and Vila Real. With the knowledge obtained in clustering analysis, a classification model is used to classify new consumers according to their consumption data. The clustering methodology is based on seven algorithms, partitional and hierarchical, six clustering validity indices are used to identify the best data partition. To complete the cycle of pattern recognition, the classification model based on decision trees is used in the classification of new consumers. To make the model simple, each load curve is represented by five indices, each index is capable to represent the shape of the load curves. The methodology proposed in this work demonstrates to be an effective tool, and can be used in most diverse sectors, highlighting the use of knowledge in the optimization of the energy contracting for low voltage consumers. Consumer data can be constantly updated in attempt to improve the model obtained in this work, finding estimates that can better represent consumers and their consumption habits.

### ***Keywords***

Knowledge Discovery in Databases, Data Mining, Clustering, Classification, Typical Load Profiles.



# Índice

<b>AGRADECIMENTOS</b> .....	<b>I</b>
<b>RESUMO</b> .....	<b>III</b>
<b>ABSTRACT</b> .....	<b>V</b>
<b>ÍNDICE</b> .....	<b>VII</b>
<b>ÍNDICE DE FIGURAS</b> .....	<b>XI</b>
<b>ÍNDICE DE TABELAS</b> .....	<b>XV</b>
<b>ACRÓNIMOS</b> .....	<b>XVII</b>
<b>1. INTRODUÇÃO</b> .....	<b>1</b>
1.1.OBJETIVOS .....	3
1.2.CONTRIBUIÇÕES .....	4
1.3.ORGANIZAÇÃO DO RELATÓRIO.....	4
<b>2. CARACTERIZAÇÃO DE PERFIS DE CONSUMO</b> .....	<b>5</b>
2.1.INTRODUÇÃO .....	5
2.2.DESCOBERTA DO CONHECIMENTO EM BANCO DE DADOS .....	7
2.3.PROCESSOS DA DESCOBERTA DO CONHECIMENTO EM BANCO DE DADOS .....	8
2.3.1. SELEÇÃO DOS DADOS .....	9
2.3.2. PRÉ-PROCESSAMENTO .....	9
2.3.2.1. ESTIMATIVA DE DADOS EM FALTA .....	10
2.3.2.2. ELIMINAÇÃO DE RÚIDOS.....	11
2.3.2.3. INTEGRAÇÃO DOS DADOS .....	13
2.3.2.4. REDUÇÃO DOS DADOS.....	13
2.3.2.5. TRANSFORMAÇÃO DOS DADOS.....	14
2.3.3. MINERAÇÃO DOS DADOS.....	16
2.3.4. AGRUPAMENTO .....	16
2.3.5. CLASSIFICAÇÃO .....	17
2.3.5.1. ÍNDICES DE FORMATO DE CARGA .....	18
2.3.5.2. ÁRVORE DE DECISÃO .....	19
2.3.6. AVALIAÇÃO DOS PADRÕES.....	21
2.3.7. CONHECIMENTO .....	21
2.4.ESTADO DA ARTE DA CARACTERIZAÇÃO DE PERFIS DE CONSUMO.....	21
2.5.CONSIDERAÇÕES FINAIS .....	26



<b>3. AGRUPAMENTO (CLUSTERING)</b> .....	<b>28</b>
3.1. INTRODUÇÃO .....	28
3.2. ALGORITMOS DE AGRUPAMENTO .....	30
3.3. MEDIDAS DE DISTÂNCIA .....	31
3.3.1. DISTÂNCIA EUCLIDIANA .....	32
3.3.2. DISTÂNCIA DE MINKOWSKI .....	32
3.3.3. DISTÂNCIA DE MANHATTAN .....	33
3.3.4. DISTÂNCIA EUCLIDIANA QUADRÁTICA .....	33
3.3.5. DISTÂNCIA DE CHEBYSHEV .....	33
3.3.6. DISTÂNCIA DE CANBERRA .....	33
3.3.7. DISTÂNCIA DE MAHALANOBIS .....	34
3.3.8. <i>DYNAMIC TIME WARPING (DTW)</i> .....	34
3.4. ÍNDICES DE VALIDAÇÃO DE CLUSTER .....	36
3.4.1. <i>MEAN INDEX ADEQUACY</i> .....	38
3.4.2. <i>CLUSTERING DISPERSION INDICATOR</i> .....	38
3.4.3. <i>DAVIES-BOULDIN INDEX</i> .....	39
3.4.4. <i>DUNN INDEX</i> .....	39
3.4.5. <i>SILHOUETTE INDEX</i> .....	39
3.4.6. <i>CALINSKI-HARABASZ INDEX</i> .....	40
3.5. <i>K-MEANS</i> .....	41
3.5.1. ASSIMILAR OS PONTOS AO CENTROIDE MAIS PRÓXIMO .....	42
3.5.2. CRITÉRIO DE CONVERGÊNCIA .....	43
3.5.3. TEMPO E COMPLEXIDADE DO ESPAÇO .....	43
3.5.4. COMENTÁRIOS ADICIONAIS .....	43
3.5.4.1. MANIPULANDO CLUSTERS VAZIOS .....	43
3.5.4.2. <i>OUTLIERS</i> .....	44
3.5.4.3. NÚMERO IDEAL DE CLUSTERS .....	44
3.6. <i>GAUSSIAN-MEANS</i> .....	45
3.6.1. O ALGORITMO GAUSSIAN-MEANS .....	45
3.6.2. TESTAR OS CLUSTERS PARA O AJUSTE GAUSSIANO .....	46
3.7. ALGORITMOS HIERÁRQUICOS .....	48
3.7.1. ALGORITMO BÁSICO DE AGRUPAMENTO HIERÁRQUICO DIVISIVO .....	49
3.7.2. DEFINIR A PROXIMIDADE ENTRE AGRUPAMENTOS .....	50
3.8. <i>FUZZY C-MEANS</i> .....	52
3.9. OUTRAS TÉCNICAS DE AGRUPAMENTO .....	53
3.9.1. MÉTODOS EVOLUCIONÁRIOS .....	53
3.9.2. MAPAS AUTO-ORGANIZÁVEIS .....	54
3.9.3. MÉTODOS BASEADOS NA DENSIDADE .....	54
3.9.4. MÉTODOS BASEADOS EM GRADE .....	55

---

3.10.CONCLUSÕES .....	55
<b>4. ESTUDO DE CASO .....</b>	<b>56</b>
4.1.DESCRICÃO DOS DADOS .....	56
4.2.PRÉ-PROCESSAMENTO .....	57
4.2.1. TRATAMENTO DOS VALORES INCONSISTENTES .....	57
4.2.2. REDUÇÃO DOS DADOS .....	58
4.2.3. NORMALIZAÇÃO.....	59
4.2.4. INTEGRAÇÃO DAS BASES DE DADOS.....	61
4.3.METODOLOGIA .....	62
4.4.AGRUPAMENTO.....	64
4.4.1. AVALIAÇÃO DOS ALGORITMOS DE AGRUPAMENTO.....	65
4.4.2. AVALIAÇÃO DO ALGORITMO <i>K-MEANS</i> .....	66
4.4.3. AVALIAÇÃO DO ALGORITMO <i>K-MEDOIDS</i> .....	68
4.4.4. AVALIAÇÃO DO ALGORITMO <i>G-MEANS</i> .....	70
4.4.5. AVALIAÇÃO DOS ALGORITMOS HIERÁRQUICOS .....	71
4.4.6. ESCOLHA DO MELHOR ALGORITMO DE AGRUPAMENTO .....	77
4.5.CLASSIFICAÇÃO .....	80
4.6.CONCLUSÕES .....	82
<b>5. CONCLUSÕES .....</b>	<b>84</b>
5.1.CONCLUSÕES E CONTRIBUIÇÕES.....	84
5.2.TRABALHOS FUTUROS .....	86
<b>REFERÊNCIAS DOCUMENTAIS.....</b>	<b>88</b>
<b>ANEXO A. BASE DE DADOS COM DOZE CONSUMIDORES.....</b>	<b>96</b>
<b>ANEXO B. BASE DE DADOS COM DEZ CONSUMIDORES .....</b>	<b>97</b>
<b>ANEXO C. BASE DE DADOS COM CENTO E SETENTA E DOIS CONSUMIDORES.....</b>	<b>98</b>
<b>ANEXO D. DADOS TRATADOS .....</b>	<b>99</b>
<b>ANEXO E. EXEMPLO DE RESULTADO PARA OS ÍNDICES DE VALIDAÇÃO.....</b>	<b>100</b>
<b>ANEXO F. RESULTADO PARCIAL DOS ÍNDICES DE VALIDAÇÃO .....</b>	<b>103</b>
<b>ANEXO G. ÁRVORE DE DECISÃO .....</b>	<b>105</b>



## Índice de Figuras

Figura 1 – Etapas do processo de KDD. Adaptado de [6].	8
Figura 2 – Estimação dos dados faltantes. Em vermelho pode-se verificar os dados reais e em azul os dados estimados pela rede neural artificial MLP. Adaptado de [3].	11
Figura 3 – Metodologia DBOR para a limpeza de dados. Adaptado de [9].	12
Figura 4 – Exemplo de árvore de decisão.	20
Figura 5 – Metodologia de caracterização de perfis de carga. Adaptado de [28].	25
Figura 6 – Metodologia de determinação de perfis típicos de carga. Adaptado de [2].	26
Figura 7 – Árvore dos tipos de classificação. Adaptado de [33].	30
Figura 8 – Duas séries temporais que possuem formatos similares. Adaptado de [40].	35
Figura 9 – Processo de validação de <i>cluster</i> .	36
Figura 10 – Resultado do processo de validação de <i>cluster</i> .	37
Figura 11 – Processo iterativo do algoritmo <i>K-means</i>	42
Figura 12 – Dois conjuntos de dados onde o número de clusters é impropriamente atribuído. Adaptado de [54].	44
Figura 13 – Algoritmos aglomerativos e algoritmos divisivos.	49
Figura 14 – Agrupamento hierárquico para quatro pontos. À direita pode ser visto o dendrograma e à esquerda os pontos analisados.	49
Figura 15 – Visualização gráfica do cálculo das distâncias.	50
Figura 16 – Visualização gráfica da metodologia para estimar valores faltantes.	58

---

Figura 17 – Redução dos dados.	59
Figura 18 – Curvas de carga sem normalização.	60
Figura 19 – Curvas de carga após a normalização.	60
Figura 20 – Integração dos dados.	62
Figura 21 – Metodologia de <i>Clustering</i> e Classificação.	63
Figura 22 – <i>K-means</i> com dois agrupamentos.	66
Figura 23 – <i>K-means</i> para três agrupamentos.	67
Figura 24 – <i>K-means</i> para quatro agrupamentos.	68
Figura 25 – <i>K-medoids</i> para dois agrupamentos.	68
Figura 26 – <i>K-medoids</i> para três agrupamentos.	69
Figura 27 – <i>K-medoids</i> para quatro agrupamentos.	69
Figura 28 – Resultado do algoritmo <i>G-means</i> .	71
Figura 29 – Resultado dos algoritmos hierárquicos para dois agrupamentos.	72
Figura 30 – Resultado dos algoritmos hierárquicos para três agrupamentos.	73
Figura 31 – Resultado dos algoritmos hierárquicos para quatro agrupamentos.	73
Figura 32 – Resultado do algoritmo <i>Average Link</i> para dois agrupamentos.	74
Figura 33 – Resultado do algoritmo <i>Average Link</i> para três agrupamentos.	74
Figura 34 – Resultado do algoritmo <i>Average Link</i> para quatro agrupamentos.	75
Figura 35 – Resultado do algoritmo <i>Ward Link</i> para dois agrupamentos.	75
Figura 36 – Resultado do algoritmo <i>Ward Link</i> para três agrupamentos.	76
Figura 37 – Resultado do algoritmo <i>Ward Link</i> para quatro agrupamentos.	76

---

Figura 38 – Perfis Típicos de Carga para os dias da semana.	77
Figura 39 – Perfis Típicos de Carga para sábados.	78
Figura 40 – Perfis Típicos de Carga para o domingo/feriado.	79
Figura 41 – Percentual de clientes em cada <i>cluster</i> .	79



## *Índice de Tabelas*

Tabela 1 – Índices de caracterização. Adaptado de [15].	19
Tabela 2 – Regras do modelo de classificação. Adaptado de [15]	20
Tabela 3 – Passos do algoritmo <i>K-means</i> . Adaptado de [33].	41
Tabela 4 – Passos do algoritmo <i>G-means</i> . Adaptado de [54].	45
Tabela 5 – Algoritmo hierárquico aglomerativo.	50
Tabela 6 – Passos do algoritmo <i>Fuzzy C-means</i> . Adaptado de [8].	53
Tabela 7 – Dados do problema.	56
Tabela 8 – Avaliação dos algoritmos de <i>clustering</i> através dos índices de validação.	66
Tabela 9 – Índices de formato de carga utilizados.	81
Tabela 10 – Regras de decisão geradas na etapa de treino.	82
Tabela 11 – Matriz de classificação dos dados.	82





## *Acrónimos*

AMR	–	<i>Automatic Meter Reading</i>
AT	–	<i>Alta Tensão</i>
BIC	–	<i>Bayesian Information Criterion</i>
BT	–	<i>Baixa Tensão</i>
CDI	–	<i>Clustering Dispersion Indicator</i>
CHI	–	<i>Calinski-Harabasz Index</i>
DBI	–	<i>Davies-Bouldin Index</i>
DBOR	–	<i>Distance Based Outlier Rejection</i>
DBMSCAN	–	<i>Density-based Micro Spatial Clustering of Applications with Noise</i>
DBSCAN	–	<i>Density-based Spatial Clustering of Applications with Noise</i>
DI	–	<i>Dunn Index</i>
DM	–	<i>Data Mining</i>
DR	–	<i>Demand Response</i>
DTW	–	<i>Dynamic Time Warping</i>
ECDF	–	<i>Empirical Cumulative Distribution Function</i>
KDD	–	<i>Knowledge Discovery in Databases</i>
MIA	–	<i>Mean Index Adequacy</i>
MLP	–	<i>Multilayer Perceptron artificial neural network</i>

MT	– Média Tensão
PI	– Produção Independente
SEE	– Sistemas Elétricos de Energia
SG	– <i>Smart Grid</i>
SI	– <i>Silhouette Index</i>
SOM	– <i>Self-Organizing Maps</i>
SMI	– <i>Similarity Matrix Indicator</i>
TLP	– <i>Typical Load Profile</i>
UPGMA	– <i>Unweighted Pair Group Method Average Algorithm</i>
VPP	– <i>Virtual Power Plant</i>
WCBCR	– <i>Within Cluster Sum of Squares to Between Cluster Variation</i>

# 1. INTRODUÇÃO

Segundo a Entidade Reguladora de Serviços Energéticos em Portugal [1], a modernização do setor elétrico nos países europeus ocorreu graças a abertura dos mercados e a inserção de novas tecnologias nos processos de geração, transmissão e distribuição de energia elétrica. A abertura dos mercados de energia ocorreu de forma progressiva durante os anos de 1995 a 2006. Na maioria dos países esse processo deu-se de forma faseada, começando a incluir primeiramente clientes de maiores consumos (consumidores elegíveis) e níveis de tensão mais elevados. Portugal seguiu uma metodologia idêntica, e desde setembro de 2006 todos os consumidores em Portugal continental passaram a poder escolher livremente o seu fornecedor de energia elétrica.

A utilização do sistema de medidores de leitura automática (AMR)<sup>1</sup> permitiu aos operadores do sistema elevado controlo e visualização dos pontos de consumo. Os AMR's facilitaram a leitura, transmissão e armazenamento de dados para os operadores dos sistemas elétricos. O

---

<sup>1</sup> *Automatic Meter Reading* (AMR), na designação anglo-saxónica.

aumento da quantidade de dados emitidos inviabilizou a operação do sistema baseada apenas na estatística clássica. O advento da produção independente (PI) passou a dificultar a operação, por alterar os fluxos de energia no sistema, inserindo dificuldades na operação dos sistemas elétricos de energia (SEE) por métodos convencionais, sem conhecimento dos perfis de carga dos consumidores. Para isso, foram desenvolvidas técnicas de mineração de dados (DM)<sup>2</sup>, que por meios matemáticos definem perfis de carga, baseados em dados históricos de consumo, auxiliando na análise e operação do sistema.

Olhando pelo lado do consumidor, a inserção dos AMR's, tornou os edifícios não só intensivos em consumo energético, mas também intensivos em informação. Com isso, o consumidor passou a ter maior liberdade em controlar o seu processo de cogeração<sup>3</sup> de energia elétrica. Controle esse, que viabilizou a utilização de novos métodos baseados em otimização de recursos energéticos como forma de aproveitar melhor as fontes de energia presentes, e também baseados em DM e descoberta do conhecimento em bancos de dados (KDD)<sup>4</sup>, onde através do conhecimento dos padrões de consumo pode-se estimar e prever o consumo de energia em períodos específicos. Uma grande vantagem é a otimização da contratação de energia, uma vez que conhecendo os hábitos de consumo, podem ser escolhidas tarifas de energia que forneçam maiores benefícios aos consumidores.

Dessa forma, o presente trabalho apresenta o estudo e implementação de toda uma metodologia de caracterização de consumidores de energia elétrica, baseada em agrupamento de dados<sup>5</sup>. O objetivo é encontrar perfis típicos de consumo em uma base de dados históricos de consumo de energia elétrica. Após a etapa de agrupamento é aplicado um modelo de classificação para a correta classificação de novos consumidores de energia, baseando-se em dados de consumo de energia elétrica. Através dos dados obtidos nessas

---

<sup>2</sup> *Data Mining* (DM), na designação anglo-saxónica.

<sup>3</sup> *O processo de cogeração é definido como o processo de geração e consumo de energia.*

<sup>4</sup> *Knowledge Discovery in Databases* (KDD), na designação anglo-saxónica.

<sup>5</sup> *O termo agrupamento vem da designação anglo-saxónica "clustering".*

etapas o conhecimento sobre os perfis típicos de carga, pode ser utilizado em diversas frentes, tais como: modelos de otimização dos recursos energéticos, previsão de carga, manutenção do sistema, contratação de energia, etc.

O algoritmo de agrupamento e o modelo de classificação serão implementados no *software R Studio*. Este *software* é um *open source* de ambiente de desenvolvimento integrado para R<sup>6</sup>. A linguagem R é rica em funções estatísticas, as quais são desejáveis e utilizadas em todo processo de DM.

## 1.1. OBJETIVOS

O aumento do número de dados em banco de dados introduziu novas dificuldades na estimação e análise das curvas de carga. Para extrair informações úteis e lidar com o grande número de dados disponíveis faz-se necessária a utilização da mineração de dados na descoberta de perfis típicos de carga. Esse trabalho tem como objetivo principal a procura de perfis típicos de carga, com base na análise de dados históricos de consumo armazenados em um banco de dados que representa um conjunto de consumidores. Para isso, serão aplicados algoritmos de agrupamento, bem como a implementação de um modelo de classificação para classificar novos consumidores.

Os objetivos principais do trabalho proposto envolvem os seguintes tópicos:

- Análise do estado da arte;
- Implementação dos algoritmos de agrupamento;
- Implementação do modelo de classificação;
- Desenvolvimento de um estudo de caso;
- Análise dos resultados e conclusões.

---

<sup>6</sup> R é a linguagem de programação utilizada no R studio.

## 1.2. CONTRIBUIÇÕES

O trabalho desenvolvido nesta dissertação resultou em uma metodologia para à caracterização de perfis típicos de carga, assentado no processo de descoberta do conhecimento em banco de dados. Foram identificados algoritmos de *clustering* e classificação que se adequam à caracterização de curvas de carga. Foram identificados e caracterizados os índices que auxiliam na decisão da aferição da melhor partição dos dados, e o melhor algoritmo de *clustering* tendo em conta os dados do problema.

O trabalho desenvolvido nesta dissertação também resultou em um artigo científico, sendo referido da seguinte forma:

- Samuel S. Cembranel, Fernando Lezama, João Soares, Sérgio Ramos, António Gomes, Zita Vale.

*“A short review on data mining techniques for electricity customers characterization”*

IEEE PES GTD Grand International Conference and Exposition Asia 2019

## 1.3. ORGANIZAÇÃO DO RELATÓRIO

O presente trabalho está dividido em cinco capítulos. O primeiro capítulo visa introduzir a informação referente ao trabalho e expor seus objetivos.

O segundo capítulo aborda o processo de mineração de dados utilizado na caracterização de perfis típicos de carga, bem como o estado da arte da caracterização de perfis típicos de consumo.

No terceiro capítulo é estudado o processo de agrupamento. Mostrando a metodologia dos principais algoritmos e as diversas métricas utilizadas no processo.

No quarto capítulo é abordado o estudo de caso que permite analisar e avaliar o desempenho dos algoritmos desenvolvidos, nesse capítulo também são mostrados os resultados obtidos na mineração de dados.

O último capítulo visa apresentar as conclusões, contribuições do trabalho e realizar recomendações para futuros trabalhos.

## 2. CARACTERIZAÇÃO DE PERFIS DE CONSUMO

Este capítulo apresenta as principais técnicas e características do processo de descoberta do conhecimento em banco de dados, apresenta também o estado da arte da caracterização de perfis típicos de consumo.

### 2.1. INTRODUÇÃO

Cada consumidor de eletricidade pode ser identificado do através do seu comportamento de consumo energia elétrica ao longo de um intervalo de tempo. O comportamento de consumo pode ser influenciado por diversos parâmetros, sendo que o principal deles é o tipo de atividade exercida, podendo ser dividida em residencial, comercial e industrial. Cada atividade possui padrões distintos de consumo, por exemplo, uma indústria tende a manter seu consumo constante durante quase todo o dia, enquanto que o consumo de energia de um mercado tende a ser maior no horário comercial e menor no período da noite. Fatores climáticos influenciam diretamente no consumo de energia, pois o mesmo tende a ser maior em períodos de inverno e verão e, ser menor durante a primavera e outono. Fatores técnicos, como o nível de tensão, podem influenciar diretamente no consumo de energia, esses fatores



geralmente caracterizados por baixa tensão (BT), média tensão (MT) ou alta tensão (AT). O consumo de energia elétrica costuma variar durante os dias úteis (segunda à sexta-feira), sábados, domingos e feriados [2].

O estudo dos perfis de consumo está se tornando uma ferramenta poderosa na gestão dos sistemas elétricos, na comercialização e gestão de consumo de energia elétrica. Dentro da gestão dos sistemas elétricos os operadores podem fornecer incentivos aos clientes para alterar seu perfil de consumo, contribuindo para a diminuição do consumo nos horários de pico, otimizando a utilização da rede de energia elétrica. Perfis de consumo também podem ser utilizados na elaboração de tarifas de energia, incentivando o consumidor a otimizar o processo de consumo de energia elétrica. O conhecimento pode ser utilizado nos mercados de energia, fornecendo vantagens aos comercializadores, em contratações mais eficientes. A gestão de energia pelo lado da demanda pode ser feita com maior consciência e efetividade, a partir do momento em que se conhece o perfil de carga, auxiliando a reduzir o consumo de energia nas horas em que o preço da energia elétrica é maior.

A partir de setembro de 2006 todos os consumidores em Portugal continental passaram a poder escolher seu fornecedor de energia elétrica, isso possibilitou o aumento do poder de gestão da energia pelo lado da demanda. A partir do momento em que o mercado foi aberto aos consumidores, os mesmos passaram a poder escolher comercializadores que fornecessem maiores vantagens econômicas para as suas características de consumo [1].

No estudo desenvolvido por [3] é visto que mesmo com o mercado liberalizado, a contratação de energia está fracamente ligada com os padrões de consumo (muitas vezes por falta de conhecimento dos padrões de consumo de energia durante os dias da semana), acarretando em desperdício de dinheiro por contratação excessiva de energia. Com a utilização de técnicas de DM esse problema pode ser reduzido, e pode ser feita a contratação de energia com as tarifas que proporcionam maiores benefícios econômicos aos consumidores.

Visto isso, essa seção apresenta partes fundamentais do processo KDD, visando a caracterização de perfis típicos de consumo, com a utilização de técnicas de DM.

## **2.2. DESCOBERTA DO CONHECIMENTO EM BANCO DE DADOS**

Nas últimas décadas houve um salto na capacidade de processamento e armazenamento de dados, contribuindo para a informatização e automação dos processos nos mais diversos setores. Os processos de reconhecimento de padrões ficaram estagnados frente ao avanço dos sistemas informáticos, para isso, foram desenvolvidos processos fundamentados em KDD, auxiliando o reconhecimento de padrões e gestão dos sistemas. Processos esses, que fizeram a conexão entre os sistemas informáticos e o reconhecimento de padrões [4].

O principal objetivo da utilização do processo KDD é formalizar um conjunto de técnicas que resultam em uma mineração de dados correta, de forma que não sejam encontrados padrões errôneos ou inconsistentes, resultando na má interpretação dos dados. Técnicas de DM são utilizadas, nesse trabalho, para obter perfis típicos de consumo, com base na análise de dados históricos de consumo de energia elétrica. Nos sistemas elétricos o conhecimento do padrão de consumo de energia pode ser utilizado em diversas frentes de estudo, destacando-se: diminuição de perdas na rede, contratação de energia, resposta da demanda, otimização da geração distribuída e predição de cargas.

DM nada mais é do que a formalização matemática de técnicas que auxiliam na identificação de padrões em bancos de dados, utilizada em processos que o trabalho humano de identificação de padrões é inviável em tempo hábil e errôneo (devido ao elevado número de dados analisados). Sua utilização no mundo real está bem consolidada apresentando diversas aplicações em áreas distintas, tais como: marketing, investimentos, detecção de fraudes, manufatura, telecomunicações, sistemas elétricos, entre outros [5].

Os avanços no campo da eletrônica fomentaram a diminuição do custo de tecnologias como: sensores, medidores e processadores. A utilização de novas tecnologias, como os medidores inteligentes possibilitou mensurar, emitir e armazenar os dados do consumo de energia de modo fácil e seguro. A facilidade na aquisição de dados de consumo de energia elétrica aumentou de modo considerável o volume dos bancos de dados. Com intuito de utilizar a informação existente, algoritmos de DM são utilizados para reconhecer os padrões de consumo e fornecer as informações necessárias para melhorar a gestão de energia.

### 2.3. PROCESSOS DA DESCOBERTA DO CONHECIMENTO EM BANCO DE DADOS

Devido às inconsistências presentes nos bancos de dados reais, os algoritmos de reconhecimento de padrão não podem ser aplicados diretamente aos dados. Por esse motivo foram desenvolvidos processos de KDD. As principais inconsistências presentes nos bancos de dados são: ruídos, falta de valores e dados duplicados. Os erros nos dados acabam por distorcer os resultados e os padrões encontrados podem não retratar corretamente os perfis dos consumidores. Devido a isso, é feita a divisão do processo de KDD, diminuindo a chance de erros, e má interpretação dos dados ao longo do processo. A divisão do processo pode ser feita em cinco etapas, Figura 1, que visam alcançar os objetivos mencionados [5] [6] [7].

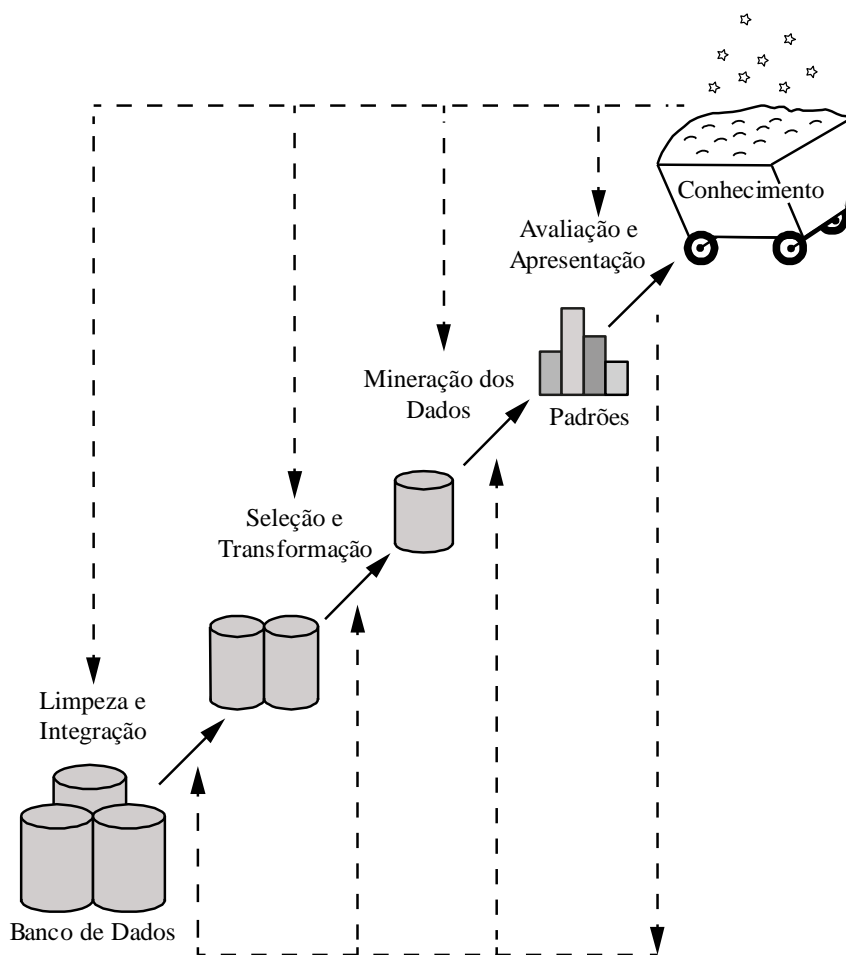


Figura 1 – Etapas do processo de KDD. Adaptado de [6].

A referência [6] define cada etapa da seguinte forma:

- Seleção dos dados – Seleção dos dados desejados na análise;
- Pré-processamento:
  - Limpeza de dados – Remoção de ruídos, valores duplicados e valores inexistentes;
  - Integração de dados – Integração de vários bancos de dados para manter a consistência dos dados do problema;
  - Redução dos dados – Redução do volume de dados para auxiliar no processo de mineração;
  - Transformação dos dados – Condicionamento dos dados para o processo de mineração.
- Mineração dos dados – Aplicação de algoritmos na extração de padrões (agrupamento e classificação);
- Avaliação dos padrões – Avaliação dos dados de interesse para a solução do problema;
- Conhecimento – Utilização do conhecimento obtido no processo para a sugestão de novas soluções.

Devido à importância das etapas do processo de KDD será feita uma abordagem mais detalhada de cada etapa nas seções 2.3.1 a 2.3.7

### **2.3.1. SELEÇÃO DOS DADOS**

Nessa etapa os dados são selecionados de acordo com a análise a ser desenvolvida. Consumidores de energia elétrica estão ligados a parâmetros contratuais de energia, que acabam por influenciar no seu consumo, os parâmetros geralmente são, nível de tensão e potência contratada. Na referência [2] é salientado que para o caso de consumidores de energia elétrica os mesmos podem ser separados de acordo com o consumo de energia, nível de tensão, atividade econômica e outros.

### **2.3.2. PRÉ-PROCESSAMENTO**

Essa etapa é justificada pela existência de inconsistências nos bancos de dados, por exemplo ruídos, valores duplicados e valores inexistentes. Fundamental no processo de KDD, se o pré-processamento não for realizado corretamente pode acarretar na descoberta de padrões inexistentes ou inconsistentes, resultando em erros na interpretação do problema.

Componentes eletrônicos, por exemplo os AMR's, possuem limitações físicas e de processamento. Durante processo de leitura do consumo de energia podem ocorrer erros que acarretam na má qualidade dos dados. A rede elétrica é suscetível a perturbações que podem influenciar as leituras dos medidores, tornando errôneos os dados captados em determinado período. Podem ocorrer falhas de comunicação entre o medidor e o banco de dados, resultando em falta de valores. Medidores eletrônicos estão expostos ao clima (temperatura e umidade) e muitas vezes acabam por avariar e apresentar inconsistências nas leituras ou imprecisão das mesmas.

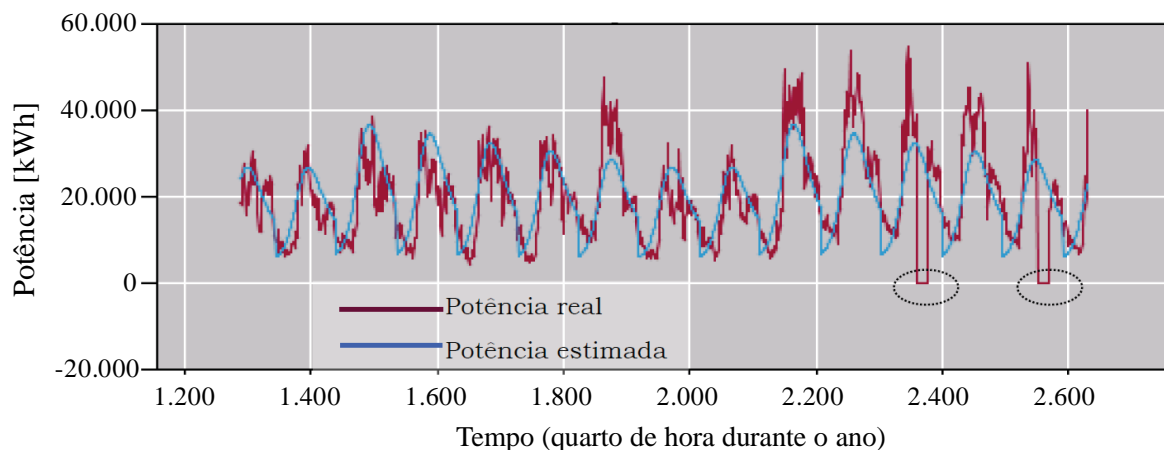
#### **2.3.2.1. ESTIMATIVA DE DADOS EM FALTA**

Nesta etapa é feita a identificação e estimativa dos dados em falta, para corrigir os dados faltantes podem ser utilizadas diversas técnicas. Na referência [6] são apresentadas algumas técnicas utilizadas na correção desse erro, tais como:

- Completar o valor em falta manualmente;
- Estimar uma constante global para substituir os valores faltantes;
- Usar a média dos pontos vizinhos para estimar o valor faltante;
- Calcular a média de todas as amostras e atribuir ao valor faltante;
- Usar técnicas de regressão para encontrar o valor mais provável e substituir o valor faltante.

No trabalho realizado por [8], na etapa de limpeza dos dados as curvas de carga diárias que possuem valores de 0 MW são eliminadas, com intuito de evitar interpretações errôneas ou equivocadas dos dados. É enfatizado também, que dados faltantes ocorrem por falhas de comunicação ou erros de leitura do medidor de energia.

Na referência [3] é feita a utilização de uma rede neural artificial perceptron multicamada (MLP)<sup>7</sup> para estimar os valores faltantes de potência (Figura 2). Como pode ser visto o algoritmo desenvolvido consegue estimar com certa precisão os dados faltantes (curva em azul) e evitar erros nas etapas posteriores do processo de mineração.



**Figura 2 – Estimação dos dados faltantes. Em vermelho pode-se verificar os dados reais e em azul os dados estimados pela rede neural artificial MLP. Adaptado de [3].**

### 2.3.2.2. ELIMINAÇÃO DE RUÍDOS

Ruídos são variáveis aleatórias ou valores discrepantes que estão presentes nos bancos de dados reais. Os mesmos podem comprometer a mineração, pois esses dados aleatórios podem causar má interpretação dos dados reais. Nos sistemas elétricos os ruídos podem ocorrer por fatores como: perturbações na rede e defeitos nos equipamentos de medição.

Apresentado por [6], a correção de ruídos pode ser feita por diversas técnicas, tais como:

- Categorização: esse método suaviza um valor de dados consultando os valores em torno dele;
- Regressão: a suavização do dado pode ser feita através de um método de regressão, por exemplo a regressão linear;

---

<sup>7</sup> Multi Layer Perceptron artificial neural network (MLP), na designação anglo-saxónica.

- Análise de ruídos: essa análise é usada para a detecção de pontos fora da curva, uma técnica usada é o agrupamento de dados.

A referência [9] apresenta uma metodologia de rejeição de ruídos baseadas na distância (DBOR)<sup>8</sup> no âmbito de redes elétricas inteligentes, onde os dados de consumo de energia elétrica são coletados por meio de AMR's. A metodologia DBOR pode ser vista na Figura 3, onde é dividida em cinco passos. No primeiro passo são inseridos os dados a passarem pela limpeza e os mesmos são divididos em classes. O segundo passo consiste em encontrar o centro de cada classe. No terceiro passo é determinada a associação de cada elemento com a classe correspondente, associação é calculada pela distância do elemento ao centro da classe, caso a distância for maior do que a maior distância estabelecida ( $\alpha$  ou  $\beta$ ) o dado é considerado como ruído. No quarto passo os ruídos são eliminados do conjunto de dados, e no quinto passo são identificados os dados corretos de cada grupo.

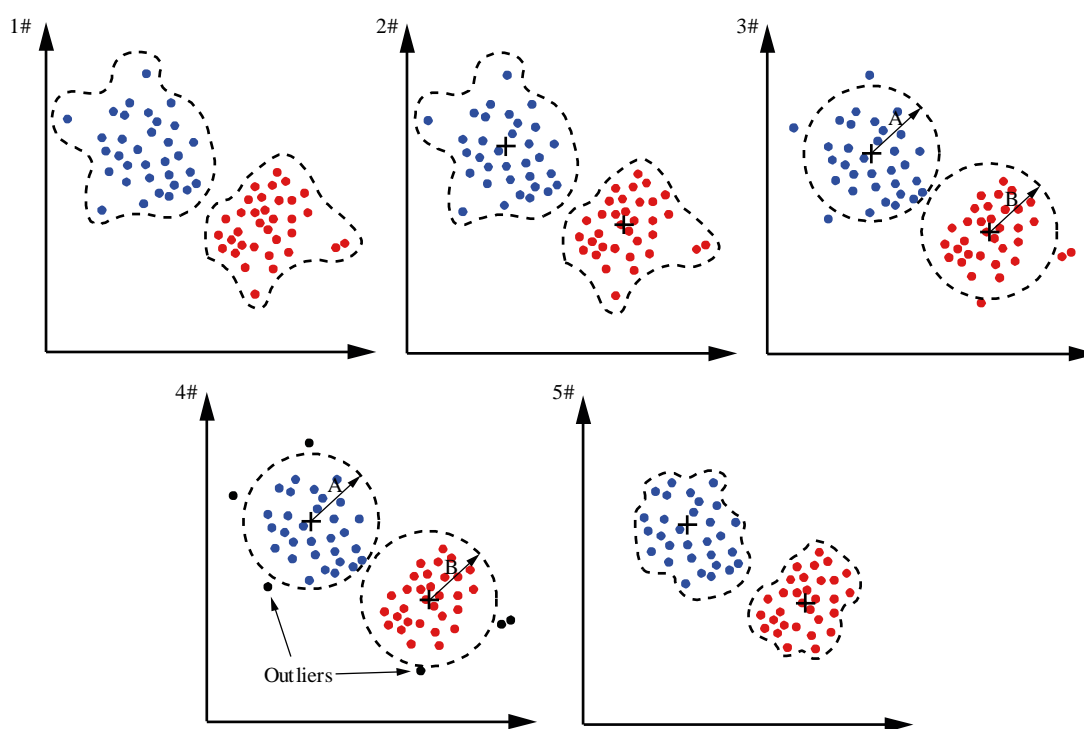


Figura 3 – Metodologia DBOR para a limpeza de dados. Adaptado de [9].

<sup>8</sup> Distance Based Outlier Rejection (DBOR), na designação anglo-saxónica.

### 2.3.2.3. INTEGRAÇÃO DOS DADOS

A integração de dados consiste em integrar dados com outros bancos de dados, pois apenas um sistema de descoberta independente pode ser suscetível a erros. A etapa de integração de dados inclui problemas devido à incompatibilidade dos bancos de dados [5]. Devido aos problemas de integração, a referência [10] estima que 70% do tempo gasto no processo de KDD é devido à integração e preparação dos bancos de dados.

### 2.3.2.4. REDUÇÃO DOS DADOS

Bancos de dados possuem grande volume de dados, tornando o processo de KDD complexo e demorado. Técnicas de redução de dados devem ser utilizadas para reduzir o volume de dados para um volume muito menor, facilitando a análise [5]. O consumo de energia pode ser medido durante um determinado intervalo de tempo (1 dia, 1 semana, 1 mês, 1 ano, ...), dependendo do tempo da etapa de medição dos dados o banco de dados pode adquirir grande volume e tornar o processo de mineração inviável e demorado. No entanto, o consumo de energia de um determinado consumidor é cíclico, por exemplo, o consumo de uma residência tende a se repetir ou ser parecido durante os dias da semana, então se a coleta de dados é feita no período de uma semana pode-se encontrar uma curva de carga média dessa residência, curva de carga para um dia, tornando a análise dos dados mais simples e eficiente.

Para realizar a redução dos dados são aplicadas algumas estratégias, incluindo redução da dimensionalidade, redução da numerosidade e compactação de dados. As técnicas de redução dos dados descritas por [6] são:

- Redução da dimensionalidade: é o processo de reduzir variáveis aleatórias ou atributos em consideração. Os métodos aplicados para a redução da dimensionalidade são transformada de *wavelet* e *principal components analysis*. Esses métodos transformam os dados originais para dados que ocupam um espaço menor;
- Redução da numerosidade: técnicas que substituem o volume original dos dados para alternativas ou formas de representação menores. Métodos paramétricos utilizam um modelo para estimar os dados, com isso apenas os parâmetros de dados precisam ser armazenados, ao invés dos dados reais. Um exemplo de modelo é a regressão log-linear.



Métodos não paramétricos armazenam representações reduzidas dos dados. Exemplos desses métodos são histogramas, agrupamentos, amostragem e agregação de cubo de dados;

- Compressão de dados: transformações são aplicadas para obter uma representação comprimida dos dados originais. As transformações podem ser sem perdas, se os dados puderem ser reconstruídos sem perda de informação, como também podem ser com perdas, quando os dados não puderem ser reconstruídos sem perda de informação. Métodos de redução da dimensionalidade e redução da numerosidade podem ser consideradas técnicas de compressão dos dados.

#### 2.3.2.5. TRANSFORMAÇÃO DOS DADOS

A transformação dos dados é efetuada para tornar o processo de mineração mais eficiente e facilitar a compreensão dos dados. Conforme apresentado na referência [11] a transformação pode ser feita por diversas técnicas, como:

- Suavização (*smoothing*): visa remover ruídos e utiliza métodos como *binning*, regressões e agrupamento;
- Construção de atributos: novos atributos são desenvolvidos e adicionados para auxiliar o processo de mineração;
- Agregação: utiliza a agregação dos dados, por exemplo, dados diários generalizados para dados semanais;
- Normalização: troca de escala para os dados, os mesmos são normalizados, por exemplo, para valores entre 0 e 1;
- Discretização: os valores numéricos são trocados por rótulos de intervalo, como exemplo 0-5, 6-10, etc.;
- Geração de hierarquia de conceito para dados nominais: os atributos são generalizados para que o processo possa ser mais eficiente.

Devido à característica dos dados que serão estudados nesse trabalho (dados numéricos), a transformação de dados mais adequada na análise do trabalho em questão é a normalização. A normalização visa dar aos dados o mesmo peso, um dos benefícios da normalização é a transformação da escala dos dados. Por exemplo, o consumo de energia pode assumir valores

elevados, chegando a escalas de quilo e mega, o que aumenta a escala de análise dos dados, com isso pode ser difícil ter a noção da importância de um dado. Outro fato é que estão sendo comparados padrões de consumo, e não montantes de energia consumida em cada horário.

Uma normalização muito utilizada é a *min-max* (1).

$$z'_i = \frac{z_i - \min_A}{\max_A - \min_A} \quad (1)$$

em que:

$z'_i$  – Representa o dado normalizado;

$z_i$  – Representa o dado a ser normalizado;

$\min_A$  – Representa o menor valor encontrado no intervalo de dados a serem normalizados;

$\max_A$  – Representa o maior valor encontrado no intervalo de dados a serem normalizados.

A grande vantagem dessa normalização é que o maior valor analisado possui valor igual a 1 e o menor valor analisado possui valor 0. Com essa normalização fica fácil ter a noção da dimensionalidade de um dado dentro de um conjunto de dados.

Outra normalização utilizada no processo é a normalização *z-score* (2).

$$z'_i = \frac{z_i - \bar{A}}{\sigma_A} \quad (2)$$

em que:

$z'_i$  – Representa o dado normalizado;

$z_i$  – Representa o dado a ser normalizado;

$\bar{A}$  – É a média dos dados;

$\sigma_A$  – É o desvio padrão dos dados.

A normalização *z-score* é útil quando os atributos máximos e mínimos do conjunto de dados não são conhecidos, ou quando existem ruídos na base de dados a ser analisada [8].

### **2.3.3. MINERAÇÃO DOS DADOS**

A etapa de mineração de dados consiste na análise do banco de dados, e na construção de modelos explicativos. Nessa fase é feita a descoberta do conhecimento através da identificação dos padrões existentes nos dados. O modelo é avaliado, medindo a capacidade que o mesmo tem para explicar os dados conhecidos ou dados ainda desconhecidos [12].

Nesse trabalho serão utilizados dois modelos de mineração de dados, agrupamento e classificação. Nessa seção é apresentado o conceito de agrupamento, pois devido a dimensão e importância dessa técnica é feita uma abordagem mais profunda no Capítulo 3. Na seção 2.3.5 são apresentadas as principais técnicas de classificação e o algoritmo utilizado nesse trabalho.

### **2.3.4. AGRUPAMENTO**

Um agrupamento é um conjunto de dados que possuem semelhanças entre si. Um agrupamento pode ser tratado como uma classe implícita de um conjunto de dados [6]. Métodos de agrupamento são baseados na metodologia de “dividir para conquistar”, nesse caso os dados são agrupados ou separados conforme as semelhanças e diferenças entre os mesmos. Através disso são identificadas regiões dos dados com diferentes características, facilitando a definição de funções de distribuição dos dados [12] [13].

Durante os anos vários métodos de agrupamento foram desenvolvidos, alguns deles são:

- Particionais;
- Hierárquicos;
- Difusos;
- Baseados na densidade;
- Baseados em grade;
- Baseados em redes neurais;
- Evolutivos;

### 2.3.5. CLASSIFICAÇÃO

Classificação é uma forma de análise de dados que extrai modelos que descrevem classes de dados importantes. Esses modelos são denominados classificadores e determinam as classes desejadas. A referência [12] define um classificador como uma função  $F$  de um conjunto de instâncias  $D$  para um conjunto de rótulos das classes  $C$  ( $F: D \rightarrow C$ ).

A classificação de dados é um processo dividido em duas etapas. A primeira etapa é a etapa de aprendizado, onde um modelo de classificação é construído com base em dados já conhecidos. Na segunda etapa é feita a classificação, em que o modelo obtido na primeira fase é usado para prever os rótulos de classe para os dados. Na primeira etapa, um classificador é construído, descrevendo um conjunto predeterminado de classes. Essa etapa é denominada etapa de aprendizado ou fase de treinamento, em que um algoritmo de classificação aprende com um conjunto de treinamento. O conjunto de treinamento é formado por sequências de dados do banco de dados. Uma sequência de dados é representada por um vetor  $X = (x_1, x_2, \dots, x_n)$ , representado por  $n$  medições feitas na sequência de dados, a partir de  $n$  atributos do banco de dados,  $(A_1, A_2, \dots, A_n)$ . Cada sequência  $X$  é associada a uma classe predefinida, conforme determinado por outro atributo do banco de dados chamado de atributo de rótulo de classe. O atributo do rótulo de classe é um valor discreto e não ordenado, é categórico em que cada valor serve como categoria ou classe. As sequências do conjunto de treinamento são amostradas aleatoriamente no banco de dados sob análise. No processo de aprendizagem cada rótulo de classe é fornecido durante o treinamento, logo essa etapa é conhecida como aprendizagem supervisionada, o que difere dos modelos de agrupamento, onde a aprendizagem é não supervisionada [6].

Segundo [11] os métodos de classificação incluem, mas não se limitam a:

- Árvores de decisão;
- Classificação bayseana;
- Redes Neurais Artificiais;
- Classificação baseada em regras de associação;
- Redes Neurais;
- Classificação por análise de vizinhança;
- Classificação baseada em raciocínio;

- Algoritmos genéticos;
- Técnicas *Fuzzy*.

Apresentado na referência [2] existem alguns atributos que classificam diagramas de cargas, em que os principais são:

- Regime de funcionamento da carga – Esses atributos fornecem informações sobre a aquisição dos dados, por exemplo, o dia da semana, mês, estação do ano, hora de aquisição dos dados (durante o período de um dia);
- Características técnicas – Essas características apresentam as informações técnicas do consumidor, como demanda contratada, nível de tensão, tipo de tarifa, entre outros;
- Curvas de carga – Esses atributos estão ligados com o padrão de consumo dos consumidores. Atributos como índices de formato de carga (apresentados na seção 2.3.5.1) são usados para representar a curva de carga de um consumidor, os mais utilizados são fator de carga, impacto do almoço e impacto noturno;
- Condições atmosféricas – Esses atributos apresentam as relações do consumo de energia elétrica com os fatores climáticos;

#### **2.3.5.1. ÍNDICES DE FORMATO DE CARGA**

Consumidores de energia elétrica estão ligados a classes pré-definidas de consumo, normalmente os hábitos de consumo estão relacionados com os contratos de eletricidade. Um modelo de caracterização de novos consumidores pode ser feito através de duas técnicas, primeiramente, aprendizagem não supervisionada, seguida de um modelo de classificação. Um modelo de caracterização de perfis de carga pode ser feito com a utilização de índices de caracterização [14].

Na tentativa de criar um modelo de classificação capaz de vincular consumidores a classes pré-definidas, uma nova representação dos perfis de carga deve ser utilizada. A representação necessita ser simples e inteligível, facilitando o processo de classificação. Para isso, índices de formato de carga são usados para representar os diagramas de carga. Esses índices são derivados dos diagramas de carga e foram propostos em diversos trabalhos como forma de expressar os diagramas de carga [14] [15] [16] [17] [18] [19] [20] [21].

Na referência [14] são utilizados quinze índices de caracterização para fornecer informações sobre os perfis de carga. Primeiramente são definidos os intervalos de tempo para cada intervalo são definidos, o valor máximo, mínimo e médio. A referência [20] apresenta um algoritmo de caracterização de perfis de carga automático, onde são utilizados quatro índices de caracterização para classificar corretamente os consumidores de energia elétrica.

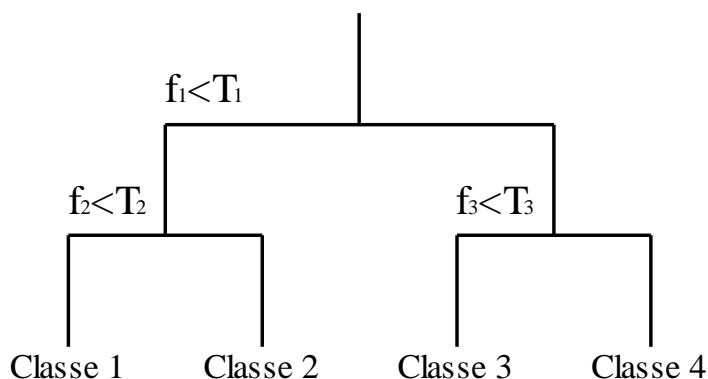
Na referência [15] são definidos seis índices de caracterização (Tabela 1), através de perfis típicos de carga obtidos através de um modelo de *clustering*. Índices esses que fornecem informações sobre a forma da curva de carga diária e sobre o padrão de consumo de cada consumidor. Os índices são diários e incluem impactos do horário de almoço e da noite.

**Tabela 1 – Índices de caracterização. Adaptado de [15].**

Parâmetro	Expressão de Cálculo	Período de Aquisição
Diário $P_{média}/P_{máx}$	$f_1 = \frac{P_{média,dia}}{P_{max,dia}}$	1 dia
Diário $P_{min}/P_{máx}$	$f_2 = \frac{P_{min,dia}}{P_{max,dia}}$	1 dia
Diário $P_{min}/P_{média}$	$f_3 = \frac{P_{min,dia}}{P_{média,dia}}$	1 dia
Impacto noturno	$f_4 = \frac{1}{3} \frac{P_{média,noite}}{P_{média,dia}}$	1 dia (8 horas durante a noite, das 11 p.m. até 6 a.m.)
Impacto do Almoço	$f_5 = \frac{1}{8} \frac{P_{média,almoço}}{P_{média,dia}}$	1 dia (3 horas durante o almoço, das 12 a.m. até 3 p.m.)
Diário $P_{média}/P_{inst}$	$f_6 = \frac{P_{média,dia}}{P_{inst}}$	1 dia

### 2.3.5.2. ÁRVORE DE DECISÃO

Uma árvore de decisão é uma estrutura similar a um fluxograma. Em cada nó da árvore é feito um teste, após cada teste os dados vão sendo separados, cada ramo da árvore representa o resultado de um teste. No final de todos os testes chega-se ao nó final que contém um rótulo de classe. Um exemplo de uma árvore de decisão pode ser visto na Figura 4, onde são feitos três testes para classificar os dados de acordo com quatro rótulos.



**Figura 4 – Exemplo de árvore de decisão.**

Os testes são feitos de acordo com os valores dos índices de formato de carga extraídos dos perfis típicos de carga obtidos na etapa de *clustering*. Na referência [15] é utilizada uma árvore de decisão para fazer classificar 229 consumidores de MT em oito rótulos de classe. Na Tabela 2 estão presentes o conjunto de regras para o modelo de classificação desenvolvido no trabalho.

**Tabela 2 – Regras do modelo de classificação. Adaptado de [15]**

Se $f_3 \leq 0,48$ e $f_2 \leq 0,13$ e $f_5 \leq 0,55$ e $f_1 \leq 0,35$ e $f_4 \leq 0,31$	<b>Cluster 8</b>
Se $f_3 \leq 0,48$ e $f_2 \leq 0,13$ e $f_5 \leq 0,55$ e $f_1 \leq 0,35$ e $f_4 > 0,31$	<b>Cluster 9</b>
Se $f_3 \leq 0,48$ e $f_2 \leq 0,13$ e $f_5 \leq 0,55$ e $f_1 > 0,35$	<b>Cluster 5</b>
Se $f_3 \leq 0,48$ e $f_2 \leq 0,13$ e $f_5 > 0,55$ e $f_5 \leq 0,6$	<b>Cluster 7</b>
Se $f_3 \leq 0,48$ e $f_2 \leq 0,13$ e $f_5 > 0,55$ e $f_5 > 0,076$ e $f_2 \leq 0,06$	<b>Cluster 6</b>
Se $f_3 \leq 0,48$ e $f_2 \leq 0,13$ e $f_5 > 0,55$ e $f_5 > 0,076$ e $f_2 > 0,06$	<b>Cluster 7</b>
Se $f_3 \leq 0,48$ e $f_2 > 0,13$ e $f_4 \leq 0,24$	<b>Cluster 4</b>
Se $f_3 \leq 0,48$ e $f_2 > 0,13$ e $f_4 > 0,24$	<b>Cluster 5</b>
Se $f_3 > 0,48$ e $f_3 \leq 0,78$ e $f_2 \leq 0,44$	<b>Cluster 3</b>
Se $f_3 > 0,48$ e $f_3 \leq 0,78$ e $f_2 > 0,44$	<b>Cluster 2</b>
Se $f_3 > 0,48$ e $f_3 > 0,78$	<b>Cluster 1</b>

O modelo foi testado e provou ter precisão de 94,83% para os dias da semana e 95,45% para os dias do final de semana, mostrando um resultado satisfatório. Os fatores de carga ( $f_1, f_2, f_3, f_4, f_5$  e  $f_6$ ) são calculados de acordo com o apresentado na Tabela 1.

### **2.3.6. AVALIAÇÃO DOS PADRÕES**

Interpretando os dados obtidos na etapa anterior chega-se ao conhecimento, pois tem-se um grande grupo de dados representado por poucos dados. Nessa etapa os dados obtidos são avaliados para verificar a consistência dos resultados. Se os padrões obtidos forem inconsistentes as etapas anteriores são repetidas até chegar a um resultado coerente.

### **2.3.7. CONHECIMENTO**

No final de todo processo são obtidos os padrões típicos de consumo. Nessa etapa final deve ser pensado como são apresentados os dados, determinar como as informações obtidas serão utilizadas. Esse conhecimento pode ser utilizado de diversas formas, pode ser utilizado para prever a demanda no dia seguinte, otimizar a rede e reduzir perdas de energia.

## **2.4. ESTADO DA ARTE DA CARACTERIZAÇÃO DE PERFIS DE CONSUMO**

O conhecimento dos hábitos de consumo vem se tornando uma ferramenta essencial na operação dos sistemas elétricos para os mais diversos fins, tais como: previsão de carga, manutenção das redes, gestão de energia e outros. Nos últimos anos, diversos trabalhos foram desenvolvidos, visando a utilização de técnicas de DM na caracterização de perfis de carga, para as mais diversas aplicações.

No estudo apresentado na referência [22] são utilizadas técnicas de mineração de dados para personalizar tarifas de eletricidade. No problema proposto é utilizado um algoritmo de agrupamento espacial baseados na densidade para aplicações com ruídos (DBSCAN)<sup>9</sup> para análise do perfil de carga a fim de encontrar os padrões de consumo dos usuários através de seus dados históricos. Após essa etapa é realizada a análise estatística do consumo para capturar a regularidade das curvas de carga. As curvas de carga são entradas para um modelo de programação não linear inteira mista para otimizar a estrutura dos preços da tarifa por tempo de uso na busca por preços de varejo de eletricidade.

---

<sup>9</sup> *Density-based Spatial Clustering of Applications with Noise* (DBSCAN), na designação anglo-saxónica.



Uma proposta interessante é feita no trabalho [17], em que são identificados perfis de cargas de consumidores de média tensão, e através desses dados desenvolver um modelo de classificação automático para novos consumidores. O modelo é implementado utilizando dados de 1.022 consumidores de média tensão, contendo dados de consumo do período de um ano. A utilização do modelo é feita para fornecer alternativas para as companhias de distribuição, onde os perfis de carga podem ser utilizados nos contratos de energia entre os distribuidores e os consumidores nos mercados liberalizados.

A maioria das empresas de energia elétrica tem implementado sistemas de AMR's. Sistemas que são instalados em clientes de alta tensão (indústrias, hospitais e grandes centros comerciais) e residências. A referência [23] tenta encontrar o melhor algoritmo de *clustering* para a geração de Perfil de Carga Típico (TLP)<sup>10</sup> em redes inteligentes (SG)<sup>11</sup>. Os algoritmos abordados no estudo incluem algoritmos hierárquicos, *K-means* e *Fuzzy C-means*. De acordo com o trabalho apresentado o algoritmo hierárquico foi considerado a melhor abordagem para garantir um tempo de processamento consistente. A técnica de agrupamento *K-means* apresentou a opção mais eficiente para se minimizar o erro absoluto entre o perfil de carga típico e o perfil de carga real. É enfatizado também que o conhecimento do perfil de consumo auxilia na operação do sistema, os perfis podem ser utilizados para aumentar a eficiência energética da rede e até mesmo para resposta da demanda (DR)<sup>12</sup>.

Na referência [24] é proposto a utilização de técnicas de agrupamento para encontrar perfis típicos de consumo em edifícios, com objetivo de auxiliar a gestão de energia no contexto de uma cidade inteligente. É destacado que conhecer o perfil de carga dos consumidores auxilia no desenvolvimento de estratégias de gestão pelo lado da demanda. Esse conhecimento pode ser utilizado na redução de perdas de energia e também pode auxiliar os operadores do sistema de transmissão e distribuição na gestão das redes e dos mercados de

---

<sup>10</sup> *Typical Load Profile* (TLP), na designação anglo-saxónica.

<sup>11</sup> *Smart Grids* (SG), na designação anglo-saxónica.

<sup>12</sup> *Demand Response* (DR), na designação anglo-saxónica.

energia através do gerenciamento de uma usina virtual (VPP)<sup>13</sup>. Conhecendo a curva de carga o consumidor pode adquirir energia em modelos tarifários mais adequados ao seu consumo (agendamento de cargas), reduzindo custos com a contratação de energia.

Os trabalhos [25] e [26] apresentam a utilização de DM para a redução do consumo de energia em edifícios. Inicialmente é feita a medição do consumo de energia elétrica em cada consumidor. Posteriormente os dados são armazenados em um banco de dados, a partir desse ponto é iniciado o processo de mineração. Após essa etapa tem-se um determinado número de curvas de carga, que representam os grupos de consumidores. Os dados obtidos tornam-se entradas para modelos de predição e otimização que minimizam o desperdício de energia no edifício.

Na referência [9] também é feita a utilização de DM para a previsão de cargas em uma SG. Primeiramente são lidos os dados dos consumidores e armazenados em um banco de dados, os dados de consumo servem de entrada para um modelo de predição que realiza a gestão da rede. Com a utilização dessas ferramentas é possível efetuar a comunicação entre consumidores, operador da rede de transporte e geradores de energia, para a realização de DR e melhor utilização dos recursos energéticos disponíveis.

Na referência [27] é feita uma revisão da utilização de algoritmos de agrupamento para traçar perfis de consumidores residenciais. São levantados quatro pontos principais da utilização desses algoritmos. Os mesmos podem ser utilizados no design de tarifas de energia elétrica visando a redução da curva de carga. Utilização na previsão de carga, em que o operador do sistema traça os perfis de carga e juntamente com algoritmos de previsão faz a análise da demanda para o próximo período. Os algoritmos podem ser utilizados em DR para a melhor utilização dos recursos energéticos. Os algoritmos também ajudam a prever cargas dos consumidores que não possuem medidores inteligentes, agregando-os em conjuntos de consumidores que possuam consumo semelhante, para obter melhor representação do sistema.

---

<sup>13</sup> *Virtual Power Plant* (VPP), na designação anglo-saxónica. Uma VPP pode ser interpretada como um sistema que integra diversas fontes de geração, capaz de interagir com a rede elétrica na forma de um agregador.

Uma ferramenta automática para a classificação de curvas de carga de consumidores é apresentada na referência [28]. Os dados passam por um processo similar de KDD apresentado anteriormente (Figura 5). Na primeira fase os dados são selecionados e passam por uma redução. Na segunda etapa é aplicado um algoritmo de agrupamento, e através de índices de validação de *cluster* é determinado o melhor número de partições para representar os dados. Após a etapa de agrupamento, novos consumidores podem ser classificados através de um modelo de classificação. A grande vantagem do algoritmo desenvolvido no trabalho é que com a informação obtida através de agrupamento pode ser feita a classificação de novos consumidores, apenas conhecendo seus dados históricos de consumo.

Uma metodologia similar à anterior é proposta na referência [29], porém nesse trabalho o objetivo é obter a caracterização de perfis de carga residenciais usando dados de medidores inteligentes. O algoritmo desenvolvido pode ser dividido em três partes principais: etapa de agrupamento, etapa de caracterização e etapa de classificação. Na etapa de agrupamento são testados três algoritmos, sendo eles: *K-means*, *K-medoids* e *Self Organizing Maps (SOM)*. Os algoritmos são comparados através da utilização do índice de validação de *clustering* Davies-Bouldin. Na comparação são avaliados quais algoritmos apresentam a melhor partição e qual o número ideal de agrupamentos para o caso de estudo. Na etapa de caracterização os agrupamentos que apresentam menor tamanho são combinados para reduzir o número de perfis similares. E na etapa de classificação são determinados quais perfis de consumo são utilizados na maior parte do tempo no período analisado.

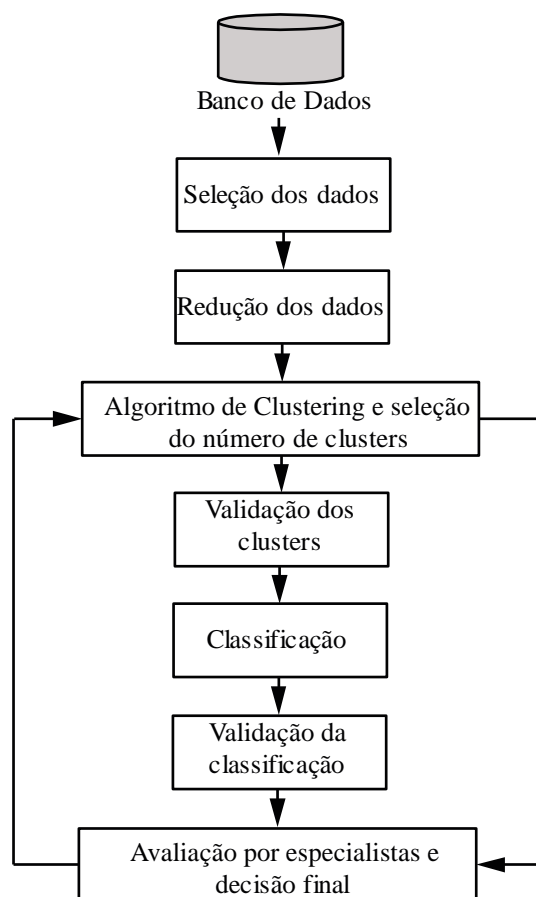


Figura 5 – Metodologia de caracterização de perfis de carga. Adaptado de [28].

Em [30] é feita a comparação da eficiência de métodos de agrupamento como *K-means*, *fuzzy K-means*, e sete algoritmos hierárquicos. Para avaliar os algoritmos são utilizados seis índices de validação. A comparação dos algoritmos mostrou que o método *K-means* desenvolvido obteve melhor desempenho para as medidas *Mean Index Adequacy* (MIA), *Clustering Dispersion Index* (CDI) e *Within Cluster sum of squares to Between Cluster variation* (WCBCR), o modelo *adaptive vector quantization* obteve melhores resultados para a função do erro quadrático médio, e *unweighted pair group method average algorithm* (UPGMA) para SMI.

Uma metodologia de caracterização de perfis de carga para consumidores de MT e AT é proposta por [2]. No trabalho é feita a comparação entre onze algoritmos de *clustering*, utilizando doze índices de validação de cluster. Um esquemático do trabalho pode ser visualizado na Figura 6. O algoritmo que obteve melhores resultados no processo de agrupamento foi o algoritmo *K-means*, utilizando quatro partições. Salienta-se ainda a

necessidade da divisão das curvas de carga para os dias da semana, sábado e domingo, pois os perfis nesses obedecem comportamentos totalmente diferentes uns dos outros.

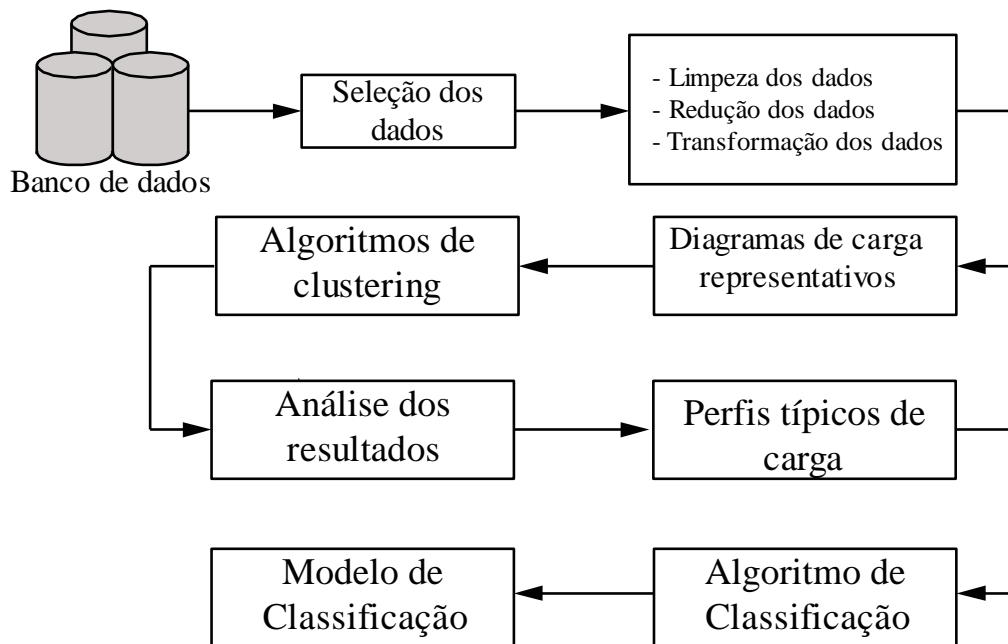


Figura 6 – Metodologia de determinação de perfis típicos de carga. Adaptado de [2].

Como alternativa aos métodos tradicionais de agrupamento, é apresentado na referência [31] um algoritmo evolucionário para traçar perfis típicos de carga, o algoritmo utiliza evolução diferencial e fronteira de Pareto para otimizar duas funções. A primeira função otimizada calcula as distâncias totais somadas entre os padrões e o seu centro de cluster correspondente e a segunda avalia o grau em que os pontos de dados vizinhos foram colocados no mesmo cluster. Na referência [32] é proposto uma técnica para encontrar de perfis típicos de carga elétrica baseado no modelo de algoritmo evolucionário com agrupamento de colônia de formigas.

## 2.5. CONSIDERAÇÕES FINAIS

Neste capítulo, foi feita uma breve revisão do processo de KDD com objetivo na utilização de técnicas de mineração de dados na caracterização perfis típicos de carga. Foi enfatizada a importância das etapas de pré-processamento de dados e sua influência nos resultados na etapa de mineração de dados. Foram abordadas as principais técnicas de estimativa de valores omissos e também foram apresentadas duas técnicas utilizadas no pré-processamento de curvas de carga. Abordaram-se algumas das técnicas utilizadas na eliminação de ruídos

nos bancos de dados, bem como a importância e influência dessa análise na etapa de mineração dos dados. A redução de dados é uma etapa crucial para a análise de curvas de carga, devido ao grande número de dados contidos para cada cliente (uma série temporal com 24, 48 ou 96 dados para cada dia do período de análise). Foram apresentadas duas formas de normalizar os dados, a normalização é muito importante nessa análise, pois é desejado fazer a comparação dos formatos de carga e não a comparação entre os montantes de consumo, por isso essa etapa torna-se tão importante.

Nessa seção também foi abordado o conceito de classificação de dados, foram citados alguns métodos de classificação, porém foi mantido o foco em árvores de decisão, pois será utilizado um modelo que cria uma árvore de decisão para a classificação de novos consumidores.

Abordou-se os principais atributos que caracterizam diagramas de carga, como regime de funcionamento, dados comerciais, atributos do diagrama de carga, condições atmosféricas e periodicidade de consumo entre os dias da semana. Para o modelo de classificação serão utilizados atributos relacionados com os diagramas de carga, e foi mostrado como calcular dos principais atributos encontrados na literatura analisada.

Por último e não menos importante, foram analisados alguns dos trabalhos que possuíssem semelhanças com o estudo que será desenvolvido nesse trabalho. Justificando a utilização das técnicas que serão estudadas e implementadas nesse trabalho nas seções seguintes.

Com isso, esse trabalho visa desenvolver um modelo baseado em descoberta do conhecimento em banco de dados, utilizando técnicas de mineração de dados como agrupamento e classificação, para encontrar perfis típicos de carga e classificar corretamente novos consumidores de energia elétrica. A próxima seção visa apresentar os conceitos e as técnicas de agrupamento que serão utilizadas no modelo proposto.

# 3. AGRUPAMENTO (*CLUSTERING*)

Este capítulo visa apresentar os principais conceitos das técnicas de agrupamento, e também os algoritmos utilizados nesse trabalho na caracterização de perfis típicos de consumo de energia elétrica.

## 3.1. INTRODUÇÃO

A prática de classificar objetos de acordo com as similaridades é base para muitas áreas da ciência. Organizar dados em grupos é um dos métodos fundamentais para entender e aprender. Análise de agrupamento ou *clustering* é a formalização de algoritmos e métodos matemáticos para realizar agrupamento, ou classificar objetos seguindo uma metodologia. Um objeto é descrito como um conjunto de medidas, ou por relações entre ele e outros objetos. Análise de agrupamento não usa rótulos para classificar objetos, através da utilização de prioridades. A não utilização de rótulos de classe distingue a análise de agrupamento de análises discriminantes, como reconhecimento de padrões e análise de decisão. O objetivo da análise de agrupamento é encontrar uma organização conveniente e

válida dos dados, e não estabelecer regras para separar dados em categorias. Esses algoritmos são voltados a encontrar a estrutura implícita dos dados [33].

As técnicas de análise de agrupamento estão preocupadas com a exploração de conjuntos de dados para avaliar se podem ou não ser resumidos, de forma significativa, em termos de um número relativamente pequeno de grupos. Os indivíduos de um grupo devem possuir características semelhantes e serem diferentes, em alguns aspectos, de indivíduos em outros grupos [34].

Conforme apresentado por [35] um agrupamento pode ser definido de três maneiras:

- Um agrupamento pode ser definido como um conjunto de dados ou partições que são mais similares entre si do que com qualquer outro dado de outra partição;
- Pode ser definido como um conjunto de dados de forma que a distância entre quaisquer dois pontos no agrupamento é menor que a distância entre qualquer ponto no agrupamento e qualquer ponto que não esteja nele;
- Um agrupamento também pode ser descrito como uma região em um espaço multidimensional que contém uma densidade relativamente alta de pontos, separada por regiões que contém uma densidade relativamente baixa de pontos.

Como um agrupamento é uma coleção de objetos de dados que são semelhantes entre si dentro do agrupamento e, diferentes de objetos em outros agrupamentos, dessa forma o mesmo pode ser tratado como uma classe implícita de um banco de dados. Nesse sentido, a análise de agrupamento é às vezes chamada de classificação automática. Isso se caracteriza por ser uma vantagem distinta de outros métodos, como por exemplo, a classificação [6].

Pelo fato de ser uma forma de classificação automática, não supervisionada, as técnicas de agrupamento vem sendo usadas em diversas áreas, tais como: buscas web, biologia, marketing, astronomia, psiquiatria, arqueologia, bioinformática e genética [6] [34]. Nesse âmbito engenheiros e pesquisadores viram nas técnicas de agrupamento soluções eficientes para a caracterização de consumidores de energia, através da análise de dados históricos de consumo. Com isso, nos trabalhos [22-32] são implementadas diversas técnicas de



agrupamento com objetivo de caracterizar perfis típicos de carga, para solucionar os mais diversos problemas encontrados no dia-a-dia.

### 3.2. ALGORITMOS DE AGRUPAMENTO

Durante os anos várias técnicas de agrupamento foram desenvolvidas, porém existem duas técnicas básicas foram muito difundidas e utilizadas na literatura, as técnicas hierárquicas e particionais (Figura 7) [33].

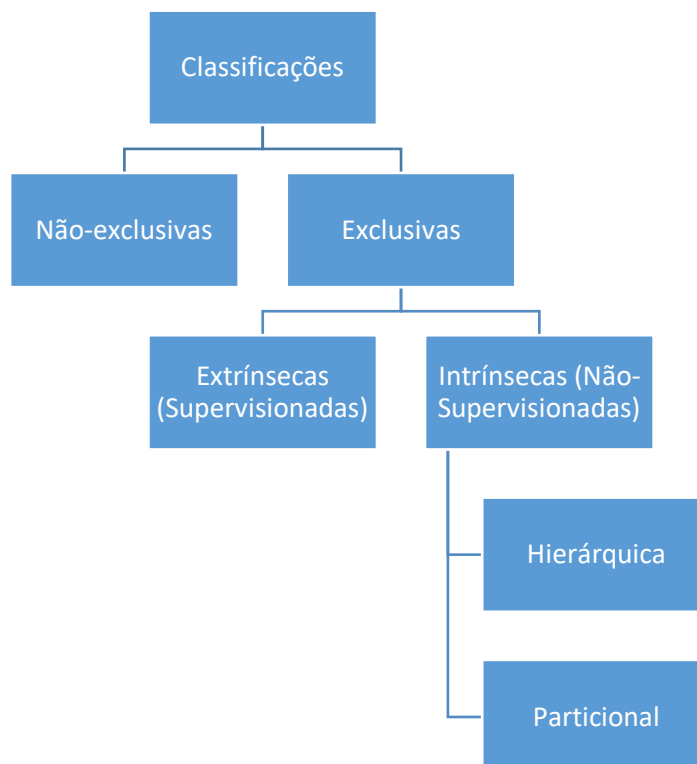


Figura 7 – Árvore dos tipos de classificação. Adaptado de [33].

Conforme apresentado por [33], a definição dos tipos de classificação é feita da seguinte maneira:

- Exclusiva *versus* não-exclusiva – A classificação exclusiva é a partição de um conjunto de dados, onde cada objeto pertence exatamente a um subconjunto ou *cluster*. Na classificação não-exclusiva, um objeto pode ser assimilado ou pertencer a vários subconjuntos ou classes. Um exemplo de classificação não-exclusiva são os métodos de agrupamento *Fuzzy*;

- Intrínseca *versus* extrínseca – Classificação intrínseca usa somente a matriz de proximidade para realizar a classificação. Classificação intrínseca muitas vezes é chamada de “aprendizagem não supervisionada”, pois não são utilizados rótulos para estabelecer as prioridades das partições. Classificação extrínseca utiliza rótulos de classe, bem como, a matriz de proximidade para efetuar a classificação. A diferença é que um classificador extrínseco necessita de um “professor”, enquanto que a classificação intrínseca usa somente a matriz de proximidade;
- Hierárquico *versus* particional – Classificação intrínseca é dividida em hierárquica e particional. A classificação hierárquica é considerada como um conjunto de sequências de partições, enquanto que a classificação particional é uma partição única.

Além dos métodos hierárquicos e particionais existem vários outros, tais como: métodos baseados em grade, baseados na densidade, métodos evolucionários, métodos difusos, métodos baseados em redes neurais artificiais, entre outros.

### 3.3. MEDIDAS DE DISTÂNCIA

Na tentativa de identificar agrupamentos que podem estar presentes nos dados é necessário ter o conhecimento de quão próximos os elementos estão uns com os outros, ou quão distantes estão. Muitos métodos de *clustering* tem como ponto de partida uma matriz  $n \times n$ , cujos elementos refletem, em certo sentido, uma medida quantitativa de proximidade. Dois indivíduos são “próximos” quando sua diferença ou distância é pequena, ou sua semelhança é grande [34].

As medidas de distância devem satisfazer três propriedades fundamentais, sendo elas: propriedade da não-negatividade, identidade de indiscernível e desigualdade triangular. Definindo um espaço  $n$  dimensional contendo os objetos  $i, j$  e  $k$ , em que  $d$  é a distância entre dois objetos, então as três propriedades segundo [6], são:

- Não-negatividade:  $d(i, j) \geq 0$  : a distância entre dois objetos no espaço deve ser um número não negativo.
- Identidade de indiscerníveis:  $d(i, i) = 0$  : a distância entre um objeto e ele mesmo deve ser igual a zero.

- Desigualdade triangular:  $d(i, j) \leq d(i, k) + d(k, j)$  : a distância de ir diretamente de um objeto  $i$  a um objeto  $j$  no espaço não é superior do que a distância entre esses dois objetos com um desvio sobre um outro objeto  $k$ , pois a menor distância entre dois pontos é uma reta.

Fundamentados os três princípios das métricas de distância, passamos a analisar algumas métricas de distância apresentadas na literatura e utilizadas por algoritmos de agrupamento. As métricas abordadas aqui são: distância Euclidiana, distância de Minkowski, distância de Manhattan, distância Euclidiana quadrática, distância de Chebyshev, distância de Canberra, distância de Mahalanobis e *Dynamic Time Warping*.

### 3.3.1. DISTÂNCIA EUCLIDIANA

Uma das métricas mais utilizadas na análise de agrupamento é a distância Euclidiana (3). A distância Euclidiana é um caso particular da distância de Minkowski, quando  $p$  possui valor igual a 2 [36]. Em que  $x_{ik}$  e  $x_{jk}$  representam os valores do  $k$ -ésimo atributo das séries temporais  $i$  e  $j$ , a distância Euclidiana é calculada pela raiz quadrada do somatório dos quadrados das diferenças entre os valores  $x_{ik}$  e  $x_{jk}$ .

$$d_{Eucl}(i, j) = \left( \sum_{k=1}^n (x_{ik} - x_{jk})^2 \right)^{1/2} \quad (3)$$

Uma observação referente às particularidades da distância de Minkowski, deve-se estar atento, pois quanto maior for o valor de  $p$  maior será a sensibilidade da métrica a distâncias maiores.

### 3.3.2. DISTÂNCIA DE MINKOWSKI

Definindo dois vetores no espaço com o mesmo comprimento  $n$ , a distância de Minkowski pode ser calculada por (4) [36].

$$d(i, j) = \left( \sum_{k=1}^n |x_{ik} - x_{jk}|^p \right)^{1/p} \quad p \geq 1 \quad (4)$$

Nota-se que se  $p$  assume valor igual a 1 tem-se a distância de Manhattan e quando  $p$  assume valor igual a 2, essa métrica passa a ser a distância Euclidiana.

### 3.3.3. DISTÂNCIA DE MANHATTAN

A distância de Manhattan é um caso particular da distância de Minkowski, quando  $p$  assume valor igual a um, e também é conhecida como *City Cab distance* ou *City Block distance* (5) [36].

$$d_{Man}(i, j) = \sum_{k=1}^n |x_{ik} - x_{jk}| \quad (5)$$

### 3.3.4. DISTÂNCIA EUCLIDIANA QUADRÁTICA

A distância Euclidiana Quadrática é considerada como uma medida para aumentar a influência de valores mais distantes, destacando a diferença entre os grupos [2]. Porém, como mencionado anteriormente (seção 3.3.1) elevando a distância Euclidiana ao quadrado, aumenta-se a sensibilidade da métrica para valores discrepantes.

$$d_{Eucl.quad}(i, j) = d_{Eucl}(i, j)^2 = \sum_{k=1}^n (x_{ik} - x_{jk})^2 \quad (6)$$

### 3.3.5. DISTÂNCIA DE CHEBYSHEV

A distância de Chebyshev ou distância suprema (também referida como  $L_{max}$  ou norma  $L_{\infty}$ ) é um caso particular da distância Minkowski para  $p \rightarrow \infty$ . Para calculá-lo, encontramos o atributo  $k$  que dá a máxima diferença nos valores entre os dois objetos [6]. Essa diferença é a distância suprema, definida mais formalmente como:

$$d_{Max}(i, j) = \lim_{p \rightarrow \infty} \left( \sum_{k=1}^n |x_{ik} - x_{jk}|^p \right)^{1/p} = \max_k |x_{ik} - x_{jk}| \quad (7)$$

### 3.3.6. DISTÂNCIA DE CANBERRA

A distância de Canberra é calculada pela equação (8).

$$d_{Can}(i, j) = \begin{cases} 0 & \text{para } x_{ik} = x_{jk} = 0 \\ \sum_{k=1}^n |x_{ik} - x_{jk}| / (|x_{ik}| + |x_{jk}|) & \text{para } x_{ik} \neq 0 \text{ ou } x_{jk} \neq 0 \end{cases} \quad (8)$$

Se  $x_{ik}$  ou  $x_{jk}$  for igual a zero o coeficiente se torna igual a 0. Nos outros casos o valor do coeficiente fica dentro do intervalo  $[0,1]$ . O somatório fornecerá um número que pertence ao

intervalo  $[0,n]$ , em que  $n$  é a dimensão dos objetos. Um problema dessa métrica é que a distância é muito sensível a pequenas variações quando as duas coordenadas estiverem próximas de zero. Uma adversidade dessa métrica ocorre em certos casos, quando as coordenadas possuem a mesma distância, porém os módulos são diferentes, nesses casos as coordenadas apresentarão contribuições diferentes [2].

### 3.3.7. DISTÂNCIA DE MAHALANOBIS

A distância de Mahalanobis (9) é uma métrica que ao contrário da distância Euclidiana leva em consideração a correlação entre os conjuntos de dados [2].

$$d_{Mah}(i,j) = \left( (x_{ik} - x_{jk})C^{-1}(x_{ik} - x_{jk})^T \right)^{1/2} \quad (9)$$

Essa distância tem como objetivo, a correção de algumas das limitações da distância Euclidiana, pois leva em consideração a dimensão dos dados (caso for feita a normalização dos dados na etapa de pré-processamento, essa limitação da distância Euclidiana já é corrigida). Porém o cálculo das matrizes de covariância  $C$  pode exigir certo esforço computacional, aumentando o tempo de cálculo e processamento dos algoritmos [37].

### 3.3.8. DYNAMIC TIME WARPING (DTW)

A distância de Minkowski e seus casos particulares apresentam algumas limitações que tornam seu uso inapropriado em alguns casos, sendo elas [38]:

- 1) A distância Euclidiana funciona somente quando os vetores a serem comparados apresentam o mesmo tamanho;
- 2) A distância Euclidiana pode ser fortemente influenciada pela escala (amplitude), a similaridade em um intervalo mais baixo pode ser superada pela dissimilaridade subsequente em um intervalo maior;
- 3) A distância Euclidiana é sensível ao deslocamento temporal entre as séries de dados.

A DTW foi introduzida na comunidade de DM no trabalho [39]. Esse método permite o alinhamento não-linear entre duas séries temporais para acomodar sequências que apresentam similaridades entre si [40].

Dadas duas séries temporais,  $X = \{x_1, x_2, \dots, x_i, \dots, x_n\}$  e  $Y = \{y_1, y_2, \dots, y_j, \dots, y_m\}$ , de tamanho  $n$  e  $m$ , respetivamente, a DTW explora a informação contida em uma matriz de distância,  $n \times m$ , [41]. Um caminho de “deformação”,  $W$ , mapeia ou alinha elementos dos vetores  $X$  e  $Y$ , de forma que a distância entre os vetores é minimizada ( $W = \{w_1, w_2, \dots, w_k\}$ ). O objetivo do algoritmo é minimizar a função (10) [39].

$$DTW(X, Y) = \min \left( \sqrt{\sum_{k=1}^K w_k} \right) \quad (10)$$

O caminho  $W$  pode ser obtido usando programação dinâmica, a metodologia para determinar o caminho pode ser vista com maior detalhe em [40].

Na Figura 8 pode ser verificado que a distância Euclidiana compara o  $k$ -ésimo ponto da primeira série temporal com o  $k$ -ésimo ponto da segunda série temporal produzindo uma medida de dissimilaridade pessimista. Por outro lado, a DTW permite uma medida de distância mais intuitiva [40].

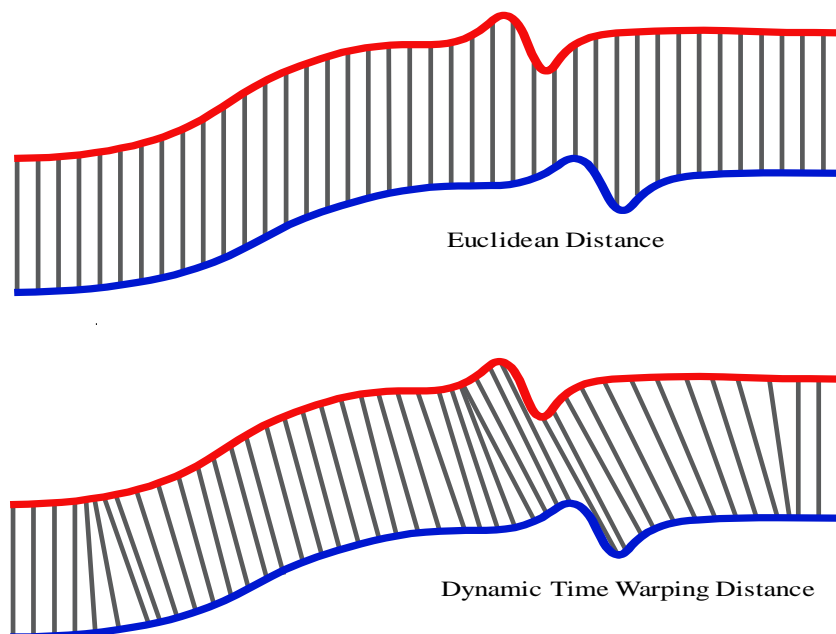


Figura 8 – Duas séries temporais que possuem formatos similares. Adaptado de [40].

### 3.4. ÍNDICES DE VALIDAÇÃO DE CLUSTER

A escolha do número ideal de agrupamentos não é um processo fácil de ser feito sem mecanismos matemáticos, principalmente quando o conjunto de dados a ser analisados possui certa dimensão. Para isso são utilizados mecanismos que possibilitam informar qual o melhor número de partições para um determinado conjunto de dados. Um modelo de processo de validação de *cluster* pode ser visto na Figura 9.

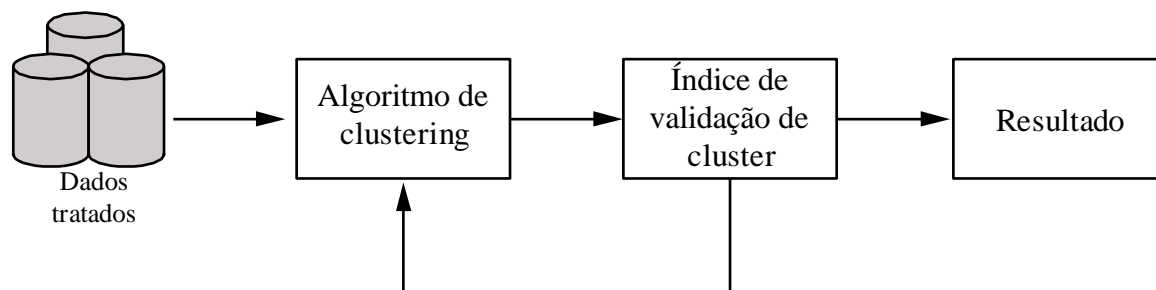


Figura 9 – Processo de validação de *cluster*.

O processo de validação de *cluster* é um processo iterativo. Os dados tratados na fase de pré-processamento passam pelo algoritmo de agrupamento. Primeiramente os dados são divididos em duas partições, o agrupamento passa pelo índice de validação e o resultado é guardado. Na segunda iteração o número de agrupamentos aumenta para três agrupamentos, os dados do agrupamento passam pelo índice de validação e o resultado é guardado. O processo é repetido tantas vezes quanto for desejado. No final desse processo existe um índice de validação para cada número de partições, e com base nisso é possível identificar o melhor número de partições. A Figura 10 apresenta o resultado do processo de validação mencionado.

Para realizar o processo de validação foi utilizado o índice de Davies-Bouldin, que será tratado com detalhe a seguir. Cinco agrupamentos já conseguiriam representar o conjunto de dados analisados, porém o resultado ótimo é com vinte agrupamentos. A inserção de mais do que cinco agrupamentos poderia acarretar em agrupamentos com *outliers* (será tratado nas próximas seções).

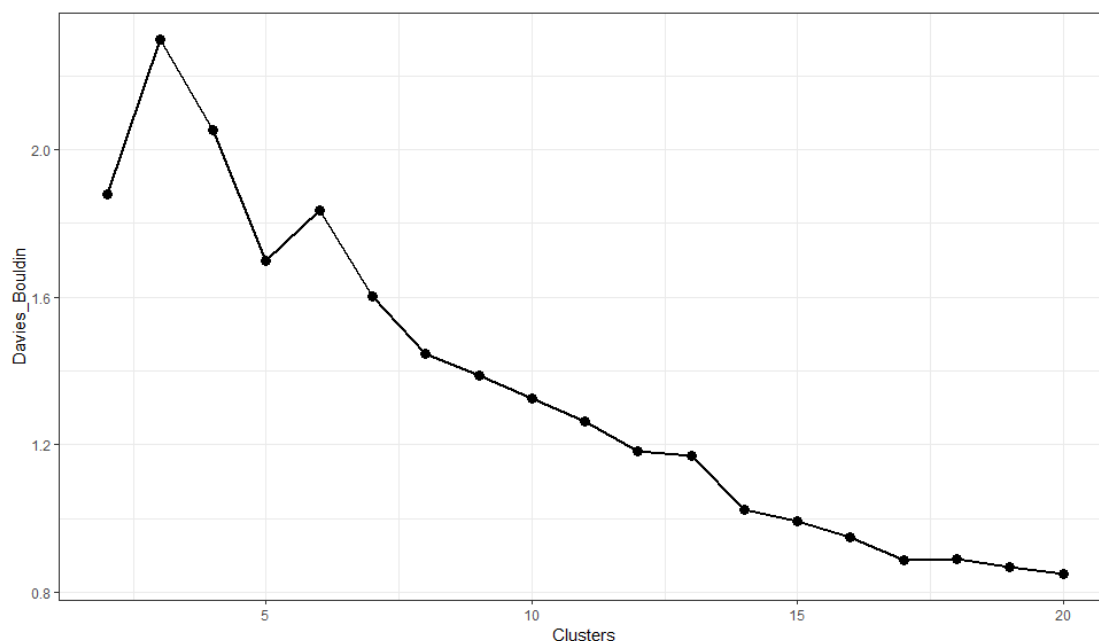


Figura 10 – Resultado do processo de validação de *cluster*.

Os índices de validação de agrupamento podem ser divididos em três tipos básicos [33]:

- Critério de validação externo: esses métodos comparam uma partição de dados com uma outra partição conhecida. Por exemplo, um critério externo mede o grau de correspondência entre os números de uma partição, obtidos de um algoritmo de agrupamento, e uma partição que se acredita ser a certa;
- Critério de validação interno: esses métodos avaliam o resultado de uma partição, somente com base nos dados da própria partição. Por exemplo, o método pode medir o resultado de um agrupamento apenas com base na matriz de proximidade dos dados;
- Critério de validação relativo: critérios relativos decidem qual o melhor resultado de agrupamento de um conjunto de dados, ou qual é mais apropriada. Por exemplo, um critério relativo medirá quantitativamente se um algoritmo hierárquico de link único ou de link completo apresenta melhores agrupamentos.

Nas seções seguintes, 3.4.1 a 3.4.6, serão abordados os índices de validação de cluster que serão utilizados nesse trabalho, sendo eles: *Clustering Dispersion indicator* (CDI), *Mean Index Adequacy* (MIA), *Davies-Bouldin index* (DBI), *Dunn index* (DI), *Silhouette index* (SI) e *Calinski-Harabasz Index* (CHI).



### 3.4.1. MEAN INDEX ADEQUACY

Dado que  $d^2(\mathbf{c}^k, \mathbf{L}^k)$  representa a distância entre o diagrama de carga representativo  $\mathbf{c}^k$  e o conjunto de dados associado a ele  $\mathbf{L}^k$ , definido como o centro geométrico das distâncias entre  $\mathbf{c}^k$  e cada membro de  $\mathbf{L}^k$ , temos que:

$$d^2(\mathbf{c}^k, \mathbf{L}^k) = \sqrt{\frac{1}{n^k} \sum_{m=1}^{n^k} d^2(\mathbf{c}^k, \mathbf{l}^m)} \quad (11)$$

O índice MIA (12) é calculado como o somatório da distância entre cada dado pertencente a um *cluster* e seu centroide [20].

$$MIA = \sqrt{\frac{1}{K} \sum_{k=1}^K d^2(\mathbf{c}^k, \mathbf{L}^k)} \quad (12)$$

Quanto menor for o valor desse índice melhor será o resultado da partição obtida no processo de *clustering*. O uso desse índice pode ser verificado nas referências [20], [42], [43], [44], [45] e [46].

### 3.4.2. CLUSTERING DISPERSION INDICATOR

O CDI depende da distância entre os diagramas de carga no mesmo *cluster* e a distância entre os diagramas de cargas representativos da classe [20].

$$CDI = \frac{1}{\hat{d}(\mathbf{C})} \sqrt{\frac{1}{K} \sum_{k=1}^K \hat{d}^2(\mathbf{L}^{(k)})} \quad (13)$$

Quanto menor for o valor desse índice melhor será o resultado da partição obtida no processo de *clustering*. O uso desse índice pode ser verificado nas referências [20], [42], [43], [44], [45] e [46].

### 3.4.3. DAVIES-BOULDIN INDEX

O índice de Davies-Bouldin é calculado por (14). Para esse índice quanto menor for o valor obtido, melhor será o resultado da partição. Minimizando esse índice os agrupamentos serão mais distintos uns dos outros, ou seja, as partições serão melhores [47].

Para  $i, j = 1, \dots, K$

$$DBI = \frac{1}{K} \sum_{k=1}^K \max_{i \neq j} \left( \frac{\hat{d}(\mathbf{L}^{(i)}) + \hat{d}(\mathbf{L}^{(j)})}{d(\mathbf{c}^{(i)}, \mathbf{c}^{(j)})} \right) \quad (14)$$

O uso desse índice pode ser verificado nas referências [17], [29], [42], [43], [44], [45] e [48].

### 3.4.4. DUNN INDEX

O índice de Dunn [49] utiliza a distância mínima entre os objetos de diferentes agrupamentos, como separação entre os agrupamentos, e usa o diâmetro máximo entre todos os agrupamentos como compactação *intracluster* [17] [28] [48].

$$DI = \min_i \left\{ \min_j \left( \frac{d_{\mathbf{c}^{(i)}, \mathbf{c}^{(j)}}}{\max_{l=1, \dots, k} \text{diam}(\mathbf{C}^{(l)})} \right) \right\} \quad (15)$$

Em que:

$$d_{\mathbf{c}^{(i)}, \mathbf{c}^{(j)}} = \min_{x \in \mathbf{c}^{(i)}, y \in \mathbf{c}^{(j)}} d(x, y) \quad (16)$$

$$\text{diam}(\mathbf{C}^{(l)}) = \max_{x, y \in \mathbf{c}^{(l)}} d(x, y) \quad (17)$$

Quanto maior for o valor desse índice melhor será o resultado do agrupamento obtido.

### 3.4.5. SILHOUETTE INDEX

O *Silhouette Index* [50] valida o desempenho do agrupamento com base na diferença das distâncias entre os agrupamentos e as distâncias dentro do agrupamento. Quanto maior for o valor desse índice, melhor será o resultado do agrupamento obtido [48]. Esse índice é calculado por (18).

$$SI = \frac{1}{K} \sum_i \frac{1}{N_i} \sum_{x \in C_i} \frac{b(x) - a(x)}{\max(a(x), b(x))} \quad (18)$$

Onde os valores para  $a(x)$  e  $b(x)$  são calculados por (19) e (20), respetivamente.

$$a(x) = \frac{1}{N_i} \sum_{y \in C_i, y \neq x} d(x, y) \quad (19)$$

$$b(x) = \min_{j, j \neq i} \left[ \frac{1}{N_j} \sum_{x \in C_j, y \neq x} d(x, y) \right] \quad (20)$$

em que:

- $K$  – é o número de clusters;
- $N_i$  – é o número de objetos no cluster  $i$ ;
- $d(x, y)$  – é a distância entre os elementos  $x$  e  $y$ .

O uso desse índice pode ser verificado nas referências [17] e [28].

### 3.4.6. CALINSKI-HARABASZ INDEX

Considerando um conjunto de dados  $D$ , e definindo  $N$  como o número de objetos em  $D$ , temos que  $c$  é o centro do conjunto de dados  $D$ , o índice de Calinski-Harabasz [51] é definido como:

$$CHI = \frac{\sum_i N_i d^2(\mathbf{c}^{(i)}, c) / (K - 1)}{\sum_i \sum_{N_i} d^2(\mathbf{L}^{(n)}, \mathbf{c}^{(i)}) / (N - K)} = \frac{N - K}{K - 1} \frac{\sum_i N_i d^2(\mathbf{c}^{(i)}, c)}{\sum_i \sum_{N_i} d^2(\mathbf{L}^{(n)}, \mathbf{c}^{(i)})} \quad (21)$$

em que:

- $K$  – número de *clusters*;
- $\mathbf{L}^{(n)}$  – perfil de carga  $n$  dentro do *cluster*  $i$ ;
- $N_i$  – número de objetos no *cluster*  $i$ ;
- $\mathbf{c}^{(i)}$  – centro do *cluster*  $i$ ;

- $d(x, y)$  – distância entre os elementos  $x$  e  $y$ .

Do mesmo modo que o DI para esse índice quanto maior for o valor do índice melhor será o resultado da partição. O uso desse índice pode ser verificado nas referências [45] e [48].

### 3.5. *K-MEANS*

O problema do algoritmo *K-means* e de outros métodos particionais pode ser tratado como: “dado um conjunto de  $n$  objetos em um espaço  $d$ -dimensional, determinar a partição dos padrões em um  $K$  grupos ou *clusters*, de modo que cada padrão em um *cluster* sejam mais semelhante entre si do que em padrões em *clusters* diferentes” [33]. Sendo assim, cada *cluster* conterá pelo menos um dado. Muitos métodos particionais são baseados na distância entre os pontos, dado o número  $K$  de partições a serem encontradas, são criadas partições iniciais e é calculada a distância entre os pontos. Posteriormente é usada a técnica de realocação iterativa, para reduzir o erro da distância entre os pontos [6].

O algoritmo de agrupamento *K-means* é um algoritmo simples, primeiramente escolhem-se os centroides iniciais, onde  $K$  é um parâmetro especificado pelo usuário. O parâmetro  $K$  é o número de partições desejadas pelo usuário. Cada ponto é assimilado ao centroide mais próximo e o conjunto de pontos assimilados a um centroide é denominado de *cluster*. O centroide de cada *cluster* é atualizado baseado nos pontos assimilados ao mesmo. O processo é repetido até os pontos não mudarem de *cluster*, ou até o centroide permanecer o mesmo [52].

**Tabela 3 – Passos do algoritmo *K-means*. Adaptado de [33].**

---

#### **Algoritmo Básico *K-means***

---

- 1: Selecionar as partições iniciais em  $K$  *clusters*.
  - 2: Repetir 3 a 6 até que o critério de parada seja atingido
  - 3: Gerar uma nova partição, assimilando cada padrão ao centro de *cluster* mais próximo.
  - 4: Calcular os novos centros dos *clusters* como centroides dos *clusters*.
  - 5: Repetir os 3 e 4 até que um valor ótimo da função de critério seja encontrado.
  - 6: Ajustar o número de *clusters* mesclando e dividindo os clusters existentes ou removendo clusters pequenos ou discrepantes.
-

O processo iterativo do algoritmo pode ser visto na Figura 11. Inicialmente são dados 4 pontos iniciais, ou pontos sementes, para o algoritmo criar uma partição inicial. Os pontos sementes são gerados aleatoriamente ou podem ser estrategicamente escolhidos. Nesse método, diferentes pontos iniciais podem gerar diferentes resultados ou partições.

O algoritmo *K-means* assimila todos os padrões ao centro do *cluster* mais próximo, fornecendo uma função de similaridade. Na primeira iteração pode ser visto que os quatro *clusters* estão muito próximos, porém a partir da segunda iteração o algoritmo encontra uma melhor partição dos dados, atingindo o critério de parada na terceira iteração.

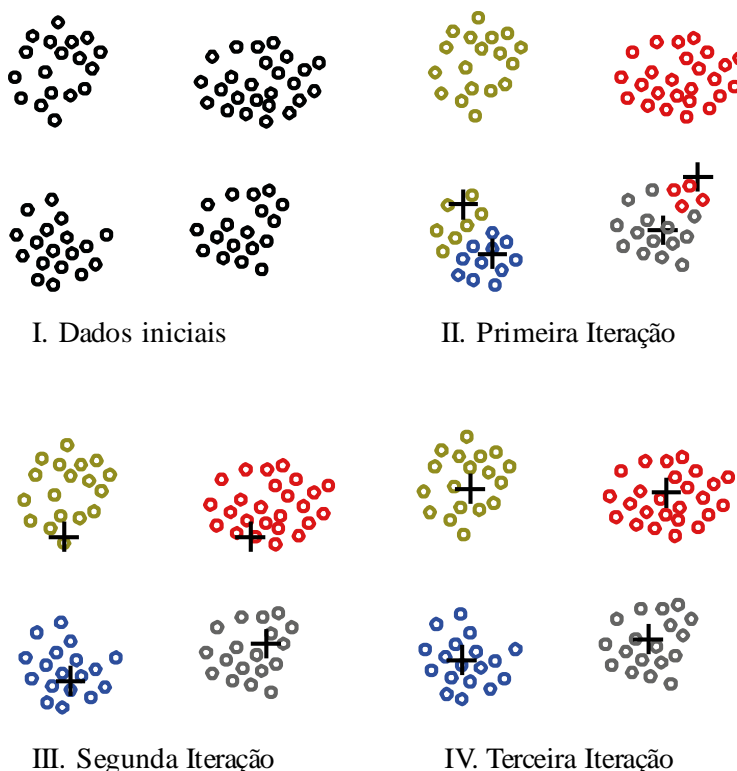


Figura 11 – Processo iterativo do algoritmo *K-means*

### 3.5.1. ASSIMILAR OS PONTOS AO CENTROIDE MAIS PRÓXIMO

Para assimilar cada ponto ao centroide mais próximo, é necessário quantificar a distância entre os pontos. Para medir a distância entre os pontos podem ser utilizadas as métricas apresentadas na seção 3.3. A métrica mais utilizada na literatura é a distância Euclidiana, por sua fácil implementação e baixo tempo de processamento, porém em problemas que o uso da distância Euclidiana se torna inviável, pode ser utilizada a DTW.

O objetivo do método de agrupamento é encontrar uma partição que contenha  $K$  clusters, de modo que minimizem a soma dos erros quadráticos de todos os clusters, ou minimizar a distância de todos os pontos aos seus centros de cluster [33].

### 3.5.2. CRITÉRIO DE CONVERGÊNCIA

O algoritmo converge quando as somas quadráticas dos erros de todos os *clusters* chegam em um ponto de mínimo ou o erro para de mudar. Nesse algoritmo não existe garantia de convergência para a melhor solução, mas é possível mudar os pontos iniciais e comparar os resultados para encontrar partições que melhor representem os dados.

### 3.5.3. TEMPO E COMPLEXIDADE DO ESPAÇO

O algoritmo *K-means* não necessita de grande armazenamento de espaço, uma vez que só são armazenados os dados do problema e os centros calculados. O armazenamento para esse algoritmo é dado pela expressão  $E((n + K)z)$ , em que  $n$  representa o número de dados e  $z$  o número de atributos. O tempo de cálculo do algoritmo também não é demasiadamente grande. A expressão do tempo do algoritmo é dada por  $E(I * K * n * z)$ , onde  $I$  é o número de iterações necessárias para a convergência. Para a maioria dos casos o algoritmo converge rapidamente e o número de iterações é baixo [52].

### 3.5.4. COMENTÁRIOS ADICIONAIS

#### 3.5.4.1. MANIPULANDO CLUSTERS VAZIOS

Segundo a referência [52] um dos problemas do algoritmo *K-means* é que podem ser obtidos *clusters* vazios, se nenhum ponto for alocado à partição durante o processo iterativo. Caso isso ocorra, uma estratégia precisa ser adotada para substituir o centro, caso contrário o erro quadrático será elevado. Para substituir o centro pode ser escolhido o ponto mais distante de qualquer centroide para ser o novo centro, ou pode ser dividido o *cluster* com maior erro quadrático, obtendo dois novos *clusters* para substituir o *cluster* vazio. Na ocorrência de vários *clusters* vazios esse processo deve ser repetido tanto quanto for necessário.

### 3.5.4.2. OUTLIERS

Quando o critério do erro quadrático é usado, ruídos podem influenciar nas partições encontradas. Em particular, quando os mesmos estão presentes, o resultado dos centroides dos *clusters* podem não ser tão representativos quanto deveriam ser, logo o erro quadrático será maior. Outro fator que pode afetar o desempenho do algoritmo é que na maioria dos casos os algoritmos *K-means* utilizam a distância Euclidiana para calcular a similaridade entre os dados, distância essa que apresenta sensibilidade a ruídos. Por esses motivos, é interessante encontrar os ruídos e eliminá-los de antemão (etapa de pré-processamento dos dados) [52].

### 3.5.4.3. NÚMERO IDEAL DE CLUSTERS

Uma desvantagem do algoritmo *K-means* é atribuir um correto número de partições. Um baixo nível de generalização, pode resultar em uma análise pobre ou inconclusiva e uma elevada generalização pode resultar em partições erradas. Na Figura 12 é possível notar, no gráfico da esquerda, que existem mais clusters do que o necessário, onde apenas um cluster já apresentaria uma boa representação dos dados. No gráfico da direita existem menos clusters do que o necessário para representar os dados.

Para solucionar esse problema foram desenvolvidos índices que auxiliam na escolha do número de *clusters*, seção 3.4. Foram desenvolvidos também, algoritmos que encontram automaticamente o número de partições, como os algoritmos *X-means* proposto por [53] e *G-means* proposto por [54] (seção 3.6).

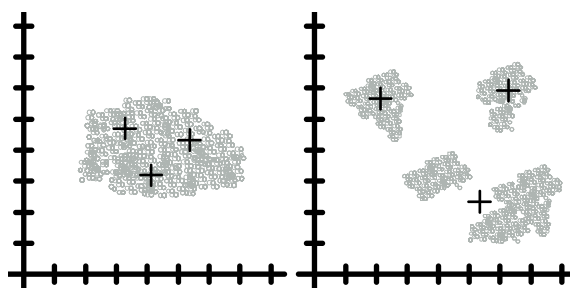


Figura 12 – Dois conjuntos de dados onde o número de clusters é impropriamente atribuído. Adaptado de [54].

Uma das alternativas para encontrar automaticamente o número de *clusters* é o algoritmo *X-means*. Esse algoritmo encontrará em um intervalo o número ideal de *clusters* com base em um índice de validação de *cluster*, como *Bayesian Information Criterion* (BIC) [53].

### 3.6. GAUSSIAN-MEANS

A escolha do número correto de *clusters* a serem utilizados geralmente não é um processo óbvio, e caracteriza uma das principais desvantagens dos problemas de *clustering*. A escolha do número de  $K$  *clusters* automaticamente não é um processo fácil. O algoritmo se baseia em um teste estatístico para a hipótese de que um subconjunto de dados segue uma distribuição gaussiana. *G-means* executa o *K-means* com o aumento progressivo do número de *clusters*,  $K$ , até que o teste aceite a hipótese de que os dados atribuídos a cada centro do *cluster* seguem a distribuição gaussiana [54].

#### 3.6.1. O ALGORITMO GAUSSIAN-MEANS

O algoritmo *G-means* começa com um número baixo de centros (através do algoritmo *K-means*) e aumenta o número de centros a cada iteração. A cada iteração do algoritmo se divide em dois centros cujos dados parecem não seguir uma distribuição gaussiana. Entre cada iteração, executa-se o *K-means* em todo o conjunto de dados e todos os centros para refinar a solução atual. Pode-se inicializar o algoritmo com um *cluster*, ou escolher um valor maior para  $K$ , se já existir algum conhecimento sobre o alcance de  $K$  [54].

Tabela 4 – Passos do algoritmo *G-means*. Adaptado de [54].

---

Algoritmo <i>G-means</i>
1: Seja $C$ o conjunto inicial de centros.
2: $C \leftarrow K\text{-means}(C, X)$ .
3: Seja $\{x_i \mid \text{classe}(x_i) = j\}$ o conjunto de dados atribuídos ao centro $c_j$ .
4: Usar o teste estatístico para detetar se cada $\{x_i \mid \text{classe}(x_i) = j\}$ segue uma distribuição Gaussiana (ao nível de confiança $\alpha$ ).
5: Se parecer gaussiana, manter $c_j$ . Caso contrário trocar $c_j$ por dois centros.
6: Repetir do passo 2 até que não seja adicionado mais nenhum centro

---



O algoritmo toma decisões baseado em um teste estatístico, que verifica se os dados atribuídos a um centro seguem uma distribuição gaussiana. Se os dados seguem a distribuição gaussiana o cluster não é dividido. Caso os dados não sigam a distribuição gaussiana o algoritmo divide o centro em dois novos centros, na tentativa de representar melhor os dados. O algoritmo *G-means* irá executar o algoritmo *K-means* até os dados atribuídos aos centros sigam distribuições gaussianas [54].

O algoritmo *K-means* pressupõe que os dados em cada *cluster* seguem uma distribuição esférica em torno de cada centro. Menos restritivamente, o algoritmo de maximização de expectativas gaussianas assume que os pontos de dados em cada *cluster* têm uma distribuição gaussiana multidimensional com uma matriz de covariância que pode ou não ser fixa ou compartilhada. O teste de distribuição gaussiana é válido para suposição de matriz de covariância. O teste considera também o número de pontos de dados,  $n$ , incorporando o valor no cálculo do valor crítico de teste. Isso evita que o algoritmo *G-means* tome decisões erradas sobre *clusters* com poucos dados [54].

### 3.6.2. TESTAR OS CLUSTERS PARA O AJUSTE GAUSSIANO

Para especificar o funcionamento completo do algoritmo *G-means*, é necessário um teste para saber se os dados atribuídos a um centro são amostrados a partir de um Gaussiano. As alternativas são:

- $H_0$ : Os dados ao redor de um centro seguem a distribuição Gaussiana;
- $H_1$ : Os dados ao redor de um centro não seguem a distribuição Gaussiana.

Se a hipótese  $H_0$  for aceita, então um único centro é suficiente para representar os dados, e não se deve partir o *cluster* em dois. Se a hipótese  $H_1$  for aceita, então é necessário dividir o *cluster*.

O teste usado é baseado na estatística de Anderson-Darling. Esse teste unidimensional foi mostrado empiricamente como o mais poderoso teste de normalidade baseado na função de

distribuição cumulativa empírica (ECDF)<sup>14</sup>. Dada uma lista de valores  $x_i$  que foram convertidos para significar média 0 e variância 1, seja  $x_{(i)}$  o  $i$ -ésimo valor ordenado. Seja  $z_i = F(x_{(i)})$ , onde  $F$  é a função de distribuição cumulativa  $N(0, 1)$  [54]. Então a estatística é:

$$A^2(Z) = -\frac{1}{n} \sum_{i=1}^n (2i - 1) [\log(z_i) + \log(1 - z_{n+1-i})] - n \quad (22)$$

Na referência [55] é mostrado que para o caso onde  $\mu$  e  $\sigma$  são estimados a partir dos dados, como é feito em *clustering*, a estatística deve ser corrigida para:

$$A_*^2(Z) = A^2(Z) \left( 1 + \frac{4}{n} - \frac{25}{n^2} \right) \quad (23)$$

Como apresentado por [54], dado um conjunto de dados  $X$  em  $d$  dimensões que pertence ao centro  $c$ , a hipótese do teste procede da seguinte forma:

1. Escolha um nível significativo de  $\alpha$  para o teste;
2. Inicializar dois centros, chamados “filhos” de  $c$ . O passo 3 é um bom modo de obter os dois centros;
3. Rodar o *K-means* nesses dois centros em  $X$ . Isso pode ser executado até a conclusão, ou para algum ponto de parada antecipada, se desejado. Seja  $c_1, c_2$  os centros filhos escolhidos pelo algoritmo *K-means*;
4. Fazer  $v = c_1 - c_2$  ser um vetor  $d$ -dimensional contendo os dois centros. Essa é a direção que o *K-means* acredita ser importante para o *clustering*. Então projete  $X$  em  $v$ :  $x'_i = \langle x_i, v \rangle / \|v\|^2$ .  $X'$  é a representação unidimensional dos dados projetados em  $v$ . Transforme  $X'$ , de modo que tenha média 0 e variância 1;

---

<sup>14</sup> *Empirical Cumulative Distribution Function* (ECDF), na designação anglo-saxónica.

5. Fazer  $z_i = F(x'_{(i)})$ . Se  $A_*^2(Z)$  está no intervalo de valores não críticos no nível de confiança, aceite  $H_0$ , mantenha o centro original e descarte  $\{c_1, c_2\}$ . Caso contrário, rejeitar  $H_0$  e manter  $\{c_1, c_2\}$  no lugar do centro original.

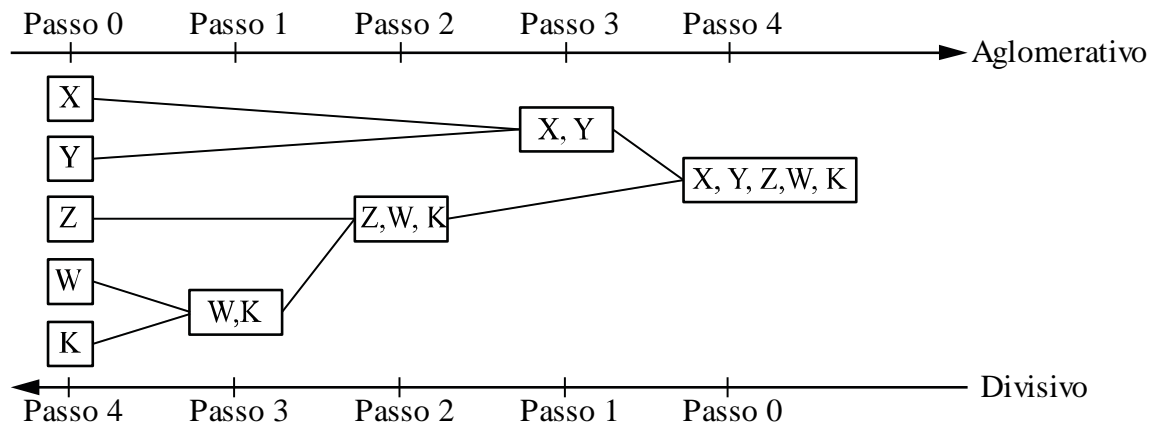
A maneira de inicializar os centros filhos é através de  $c \pm m$ , onde  $c$  é o centro e  $m$  é calculado. O método de cálculo do valor  $m$  coloca os dois centros em seus locais esperados em  $H_0$ , encontrando os principais componentes  $s$  dos dados (com o autovalor  $\lambda$ ) e escolhe  $m = s\sqrt{2\lambda/\pi}$ .

### 3.7. ALGORITMOS HIERÁRQUICOS

Técnicas de *clustering* hierárquicas são a segunda categoria mais importante de métodos de *clustering*. Tal como acontece com o *K-means*, essas abordagens são relativamente antigas em comparação com muitos algoritmos, mas ainda são muito utilizadas em diversas áreas. Existem basicamente duas aproximações para gerar algoritmos de *clustering* hierárquico, tais como [52]:

- Aglomerativo: a análise começa com os pontos em agrupamentos individuais, e a cada passo, os agrupamentos vão sendo juntados aos pares. Essa técnica requer a definição de proximidade entre os agrupamentos.
- Divisivo: a análise começa com todos os pontos em um só agrupamento, e a cada passo, agrupamento é dividido até que restem apenas agrupamentos com um só ponto. Nesse caso é necessário decidir qual agrupamento será dividido em cada passo e como será feita a divisão.

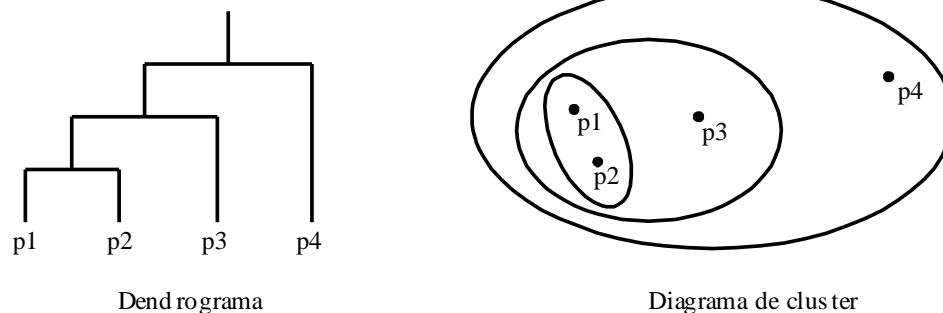
A análise mencionada anteriormente sobre os algoritmos aglomerativos e divisivos pode ser vista na Figura 13.



**Figura 13 – Algoritmos aglomerativos e algoritmos divisivos.**

Técnicas de agrupamento hierárquico vem sendo muito usadas na representação de perfis típicos de carga e sua utilização pode ser vista nos trabalhos [16], [30], [56] e [57]. Essas técnicas de agrupamento geralmente são mostradas como um diagrama de árvore, conhecido por dendrograma.

Um dendrograma ilustrativo pode ser visto na Figura 14, onde são analisados quatro pontos realizando o agrupamento dos mesmos.



**Figura 14 – Agrupamento hierárquico para quatro pontos. À direita pode ser visto o dendrograma e à esquerda os pontos analisados.**

### 3.7.1. ALGORITMO BÁSICO DE AGRUPAMENTO HIERÁRQUICO DIVISIVO

As técnicas de agrupamento aglomerativo hierárquico e suas variações geralmente seguem a seguinte abordagem: começar com todos os pontos individuais como *clusters*, em seguida são juntados os *clusters* mais próximos, esse processo se repete até que reste apenas um único cluster com todos os pontos. Um algoritmo básico proposto por [52] pode ser visto na Tabela 5.

Tabela 5 – Algoritmo hierárquico aglomerativo.

<b>Algoritmo de Agrupamento Aglomerativo Hierárquico</b>
1: Computar a matriz de proximidade, se necessário:
2: <b>Repetir</b>
3: Mesclar dois agrupamentos mais próximos
4: Atualizar a matriz de proximidade para refletir a proximidade entre os novos agrupamentos e o cluster original.
5: <b>Até</b> que só reste apenas um único agrupamento

### 3.7.2. DEFINIR A PROXIMIDADE ENTRE AGRUPAMENTOS

A grande diferença entre os métodos hierárquicos, além de serem divisivos ou aglomerativos, se encontra na forma em que distância para mesclar o *cluster* é computada. Muitas técnicas de agrupamento hierárquico e aglomerativo, como *min*, *max* e *média de grupo*, vêm de uma visualização gráfica. *min* define a proximidade do cluster como a proximidade entre dois pontos mais próximos que estão em clusters diferentes, ou usando termos gráficos, a borda mais curta entre dois nós em diferentes subconjuntos de nós. Alternativamente, o *max* leva a proximidade entre os dois pontos mais distantes em *clusters* diferentes ser a proximidade do cluster ou, usando termos gráficos, a borda mais longa entre dois nós em diferentes subconjuntos de nós. Alternativamente aos nomes *min* e *max* são utilizados os nomes elo único (do inglês – *single link*) ou elo completo (do inglês – *complete link*). Outra abordagem baseada em gráfico é a média de grupo, que define a proximidade do *cluster* como as proximidades médias pares (comprimento médio das bordas) de todos os pares de pontos de diferentes clusters. Esses três métodos podem se visualizados na Figura 15 [52].

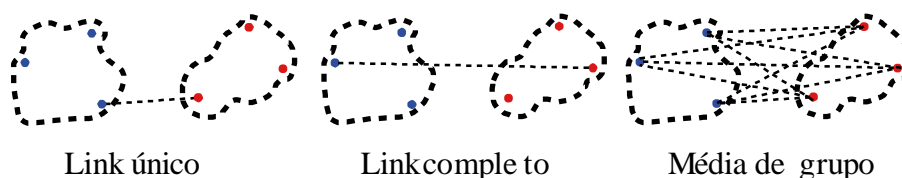


Figura 15 – Visualização gráfica do cálculo das distâncias.

A referência [58] apresenta a formalização matemática do que foi descrito, e outras técnicas para calcular a distância são apresentadas. Conforme a mesma referência dois *clusters*  $C_i$  e

$C_j$  são mesclados para formar um novo cluster  $C_n$ , então a distância desse novo cluster para qualquer outro cluster existente pode ser computada em diversas formas:

- Algoritmo de Elo Único (*Single link algorithm*)

$$d(C_n, C_l) = \min\{d(C_l, C_i), d(C_l, C_j)\} \quad (24)$$

- Algoritmo de Elo Completo (*Complete link algorithm*)

$$d(C_n, C_l) = \max\{d(C_l, C_i), d(C_l, C_j)\} \quad (25)$$

- *Unweighted pair group method average algorithm* (UPGMA)

$$d(C_n, C_l) = \frac{n_i * d(C_l, C_i) + n_j * d(C_l, C_j)}{n_i + n_j} \quad (26)$$

- *Weighted pair group method average algorithm* (WPGMA)

$$d(C_n, C_l) = (d(C_l, C_i) + d(C_l, C_j)) * 0.5 \quad (27)$$

- *Unweighted pair group method centroid algorithm* (UPGMC)

$$d(C_n, C_l) = \frac{|C_i| * d(C_l, C_i) + |C_j| * d(C_l, C_j)}{|C_i| + |C_j|} - |C_i| * |C_j| * \frac{d(c_i, c_j)}{(|C_i| + |C_j|)^2} \quad (28)$$

Onde  $d(c_i, c_j)$  é a distância euclidiana entre os clusters  $i$  e  $j$

- *Weighted pair group method centroid algorithm* (WPGMC)

$$d(C_n, C_l) = \frac{d(C_l, C_i) + d(C_l, C_j)}{2} - \frac{d(c_i, c_j)}{4} \quad (29)$$

- Algoritmo de variância mínima de WARD (*Ward minimum variance algorithm*):

$$d(C_n, C_l) = \left( (|C_l| + |C_i|) * d'(C_l, C_i) + (|C_l| + |C_j|) * d'(C_l, C_j) - |C_l| * d'(C_i, C_j) \right) * (|C_l| + |C_i| + |C_j|)^{-1} \quad (30)$$

Em que  $d'(c_a, c_b)$  é calculado por (31).

$$d'(c_a, c_b) = |c_a| * |c_b| * d(c_a, c_b) * (|C_a| + |C_b|)^{-1} \quad (31)$$

### 3.8. FUZZY C-MEANS

Nos métodos difusos de *clustering* cada curva de carga pode ser alocada em mais de um agrupamento ou em nenhum. Nessa técnica cada curva de carga pertence a um certo grupo com um certo grau, determinado por uma função de pertinência [44] [59] [60]. O algoritmo *Fuzzy C-means* [61] tenta minimizar a função que representa a distância entre cada dado e o centro do agrupamento, pesado pela função de pertinência do dado em questão (32).

$$C^{(r)} = \sum_{k=1}^N \sum_{i=1}^K \mu_{ik}^m \|x_k - s_i\|^2 \quad (32)$$

em que:

- $x_k$  – diagrama de carga do k-ésimo consumidor;
- $s_i$  – centro do agrupamento  $i$ ;
- $\mu_{ik}$  – representa o valor da função de pertinência da curva de carga  $k$  no cluster  $i$ ;
- $N$  – número de dados;
- $K$  – representa o número de agrupamentos;
- $m$  – parâmetro de “fuzzificação”;
- $\|*\|$  – qualquer norma que expressa a similaridade entre o dado e o centro (seção 3.3).

Em cada passo do processo iterativo os valores de pertinência são atualizados.

$$\mu_{ik}^m = \frac{1}{\sum_{i=1}^K \left( \frac{d_{x_i, s_k}^2}{d_{x_j, s_k}^2} \right)^{\frac{1}{m-1}}} \quad (33)$$

$$y_k = \frac{\sum_{i=1}^K \mu_{i,k} \cdot x_k}{\sum_{i=1}^K \mu_{i,k}} \quad (34)$$

A referência [62] mostra que se o valor de  $m$  for igual a 1 o resultado é praticamente igual ao algoritmo *K-means*, aumentando o valor de  $m$ , aumenta-se o tamanho das classes, fazendo com que elas fiquem mais abrangentes.

**Tabela 6 – Passos do algoritmo *Fuzzy C-means*. Adaptado de [8].**

---

**Algoritmo de Agrupamento *Fuzzy C-means***

---

1: Escolher os valores para  $K$  e  $m$ , e inicializar  $\mu^{(0)}$

2: **Repetir**

3: Calcular os centros dos agrupamentos ( $s_i^{(r)}$ )

4: Atualizar a matriz de partição para a próxima iteração  $\mu_{ik}^{(r+1)}$ .

Para todo  $i, k$  notar que:

$$\sum_{k=1}^N \mu_{ik}^{(r+1)} = 1, \mu_{ik}^{(r+1)} \in [0,1]$$

5: **Parar se**  $\|C^{(r+1)} - C^{(r)}\| < \varepsilon$ , caso contrário  $r = r + 1$ .

---

### 3.9. OUTRAS TÉCNICAS DE AGRUPAMENTO

Durante os anos muitos métodos de agrupamentos foram desenvolvidos, e o estudo de todas as técnicas em um só trabalho o tornaria muito extenso e maçante. Vale ressaltar que essas técnicas existem e muitas delas são utilizadas em processos de identificação de perfis de carga.

Os algoritmos de agrupamento particionais e hierárquicos são os principais métodos utilizados, pois são simples e eficientes. Por outro lado, métodos que apresentam bons resultados frente aos métodos tradicionais são os evolucionários. Métodos baseados na densidade são muito utilizados para identificar ruídos e também apresentam bons resultados nas tarefas de identificação de perfis de carga.

Em seguida são abordados, de forma sucinta, quatro outros métodos de agrupamentos encontrados na literatura.

#### 3.9.1. MÉTODOS EVOLUCIONÁRIOS

Evolução é um processo de constante melhoria. Algoritmos evolucionários usam técnicas de otimização baseadas na Teoria da Evolução de Darwin [63]. A ideia principal desses algoritmos é otimizar funções para encontrar a melhor alternativa de *clustering*. O algoritmo pode otimizar uma função ou um grupo de funções. O primeiro conhecido como otimização com objetivo único e o segundo como otimização multiobjetivo [64].



Como alternativa aos métodos tradicionais de agrupamento [31] apresenta um algoritmo de evolucionário para traçar perfis típicos de carga, o algoritmo utiliza evolução diferencial e fronteira de Pareto para otimizar duas funções. [32] propõe encontrar perfis típicos de carga elétrica baseado no modelo de algoritmo evolucionário com agrupamento de colônia de formigas.

### 3.9.2. MAPAS AUTO-ORGANIZÁVEIS

Os mapas auto-organizáveis (SOM) foram introduzidos pelo professor Teuvo Kohonen. Também conhecidos como Kohonen SOM, o algoritmo é uma rede neural artificial que projeta, ou mapeia, um conjunto de dados de alta dimensão,  $d$ -dimensional, em um espaço com dimensão reduzida, geralmente monodimensional ou bidimensional, com uma rede pré-definida  $M_1 \times M_2$  de neurons, que facilita a visualização e interpretação dos resultados. O processo de redução da dimensionalidade dos dados iniciais, é uma técnica de compressão de dados, conhecida como quantização de vetor. Cada unidade da rede é composta por um vetor de peso e sua localização na rede. Na rede os neurós são calculados por aprendizagem competitiva, onde apenas o neurón vencedor é ativado. O algoritmo de aprendizagem atualiza o peso do neurón vencedor, e também, atualiza peso dos neuróns da vizinhança com o inverso da proporção da distância deles ao neurón vencedor. Essa é uma técnica bem difundida e muito utilizada na caracterização de curvas de carga, sua aplicação pode ser vista nos trabalhos [8] [18] [19] [29] [46] [59] [60] [65] [66].

### 3.9.3. MÉTODOS BASEADOS NA DENSIDADE

Métodos baseados em densidade identificam regiões densas de objetos no espaço dados. A ideia é encontrar regiões de alta densidade separadas das regiões de baixa densidade. As regiões com alta densidade são consideradas *clusters*, e regiões com baixa densidade contêm dados de ruído ou *outliers* [67]. [68] propõe a utilização de um algoritmo baseado na densidade (*density-based micro spatial clustering of applications with noise – DBMSCAN*) para detetar irregularidades no consumo de energia elétrica. Outro método baseado na densidade é utilizado por [22] para identificar perfis de carga típicos e juntamente com modelos de otimização otimizar os preços de varejo dos mercados energia.

#### **3.9.4. MÉTODOS BASEADOS EM GRADE**

Métodos baseados em grade quantizam o espaço em um número finito de células formadas por uma grade. As operações de agrupamento são feitas no espaço quantizado. A vantagem dessa abordagem é o baixo tempo de processamento, que é independente do número de objetos e depende apenas do número de células e suas dimensões no espaço quantizado [6].

#### **3.10. CONCLUSÕES**

Nessa seção foram abordadas as técnicas de *clustering* encontradas na literatura, foi feita a caracterização das técnicas de *clustering* e da teoria e características de cada técnica. A apresentação e avaliação das técnicas é abordada no Capítulo 4, onde é feito o estudo de caso.

Primeiramente foram analisadas as principais técnicas e medidas de distância, que refletem de certo modo a similaridade entre os dados. Foi elaborada a abordagem matemática, bem como evidenciada a característica das métricas estudadas.

A escolha do melhor algoritmo de *clustering*, bem como do melhor número de partições não é um processo fácil de ser efetuado sem meios matemáticos que possam quantificar a qualidade das partições obtidas pelos algoritmos. Foi abordado o processo de validação de *clustering*, processo esse adotado para o caso de estudo, em que foram utilizados seis índices capazes de quantificar os agrupamentos obtidos nas análises.

Por fim foram analisadas as principais técnicas de *clustering* de dados utilizadas na literatura com objetivo de identificar perfis típicos de carga. Foram abordadas técnicas tradicionais, como as particionais e hierárquicas, suas variações, e também foram abordados, de forma breve, métodos difusos, evolucionários, baseados em inteligência artificial, baseados em densidade e grade.

Na seção seguinte será efetuado o caso de estudo utilizando as técnicas abordadas até agora, desde o pré-processamento, *clustering* e classificação dos dados. O objetivo é encontrar o algoritmo que possa fornecer a melhor partição dos dados em análise, bem como o melhor número de partições.

## 4. ESTUDO DE CASO

Neste capítulo será realizado o estudo de caso, serão abordadas algumas das técnicas apresentadas nas seções anteriores, com objetivo de identificar os padrões de consumo de energia elétrica de consumidores de baixa tensão.

### 4.1. DESCRIÇÃO DOS DADOS

Os dados utilizados no estudo de caso, são de clientes de baixa tensão, e provém de três bancos de dados analisados. O período da análise foi de um mês e o período de amostragem era diferente para cada banco de dados. No total, com os três bancos de dados o conjunto de análise possui 194 consumidores de baixa tensão das cidades do Porto, Matosinhos e Vila Real.

Tabela 7 – Dados do problema.

Banco de dados	Número de clientes	Período da análise	Período de amostragem	Grandezas Medidas
1	172	1 mês	15 min	Potência Ativa
2	12	1 mês	60 min	Potência Aparente
3	10	1 mês	5 min	Potência Aparente

A grandeza medida em cada aquisição foi a potência ativa, para o primeiro banco de dados, e a potência aparente, para o segundo e terceiro banco de dados, com isso a análise do padrão de consumo será limitada apenas aos diagramas de carga dos consumidores. Se estivessem disponíveis informações contratuais, poderiam ser feitas análises complementares, por exemplo, poderia ser verificado se o contrato de energia de cada consumidor é adequado para seu padrão de consumo. Também poderia ser feita a oferta de tarifas que mais se adequassem aos padrões de consumo de cada consumidor e otimizar a contratação de energia.

As planilhas com os dados de cada consumidor podem ser vistas nos anexos A, B e C, onde pode ser visualizada a característica diferente de cada banco de dados que foi utilizado nessa análise.

## **4.2. PRÉ-PROCESSAMENTO**

Esta etapa, descrita com detalhe na seção 2.3.2 caracteriza-se por ser essencial no processo de descoberta do conhecimento, pois tem como objetivo tratar e condicionar os dados para a análise dos padrões existentes no conjunto de dados.

### **4.2.1. TRATAMENTO DOS VALORES INCONSISTENTES**

A primeira etapa é a identificação dos valores em falta. Valores faltantes nos bancos de dados podem ocorrer por diversas causas, dentre elas a mais comum é a falha de comunicação entre o equipamento que efetua a medida e o banco de dados. Outro fator que pode ocasionar esse problema é a avaria do equipamento de medição. O banco de dados foi verificado com intuito de remover valores faltantes e valores inconsistentes para que os mesmos não interferissem na etapa de mineração dos dados.

Primeiramente foram eliminados consumidores com valores inconsistentes para a análise, foram encontrados sete consumidores com consumo de energia igual a zero para todo o período de aquisição dos dados. Esses consumidores foram removidos para que não influenciassem na análise, pois seus valores de consumo estavam inconsistentes.

Posteriormente à primeira análise, foram tratadas falhas curtas. As falhas curtas caracterizam-se por valores nulos com intervalo menor ou igual a uma hora. Quando a falha

era detetada utilizava-se a média entre os pontos vizinhos para fazer a estimação do valor faltante. Um exemplo, da metodologia utilizada pode ser verificado na Figura 16, onde estão presentes a curva real, linha contínua em vermelho, e a curva estimada, linha pontilhada em azul.

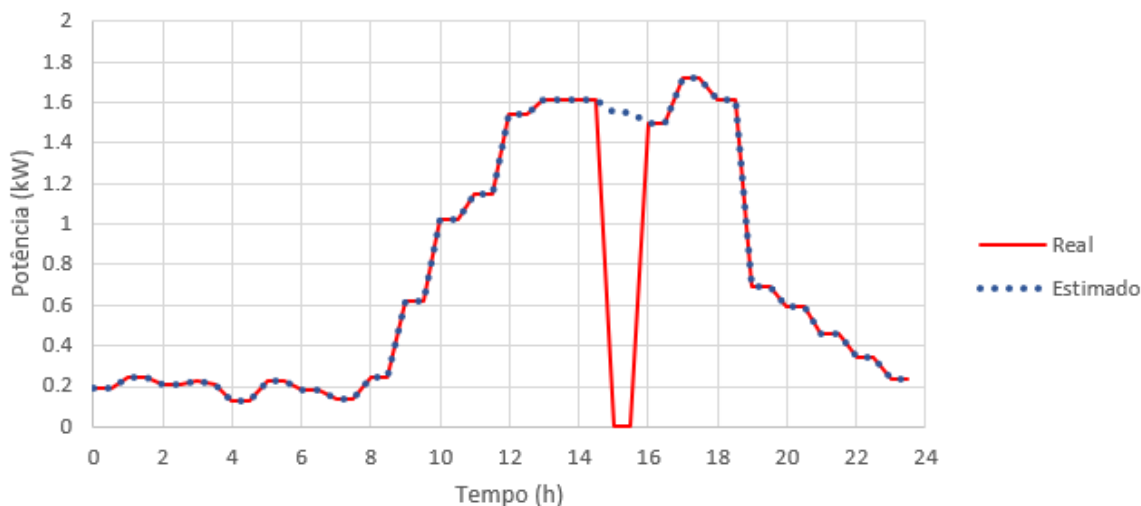


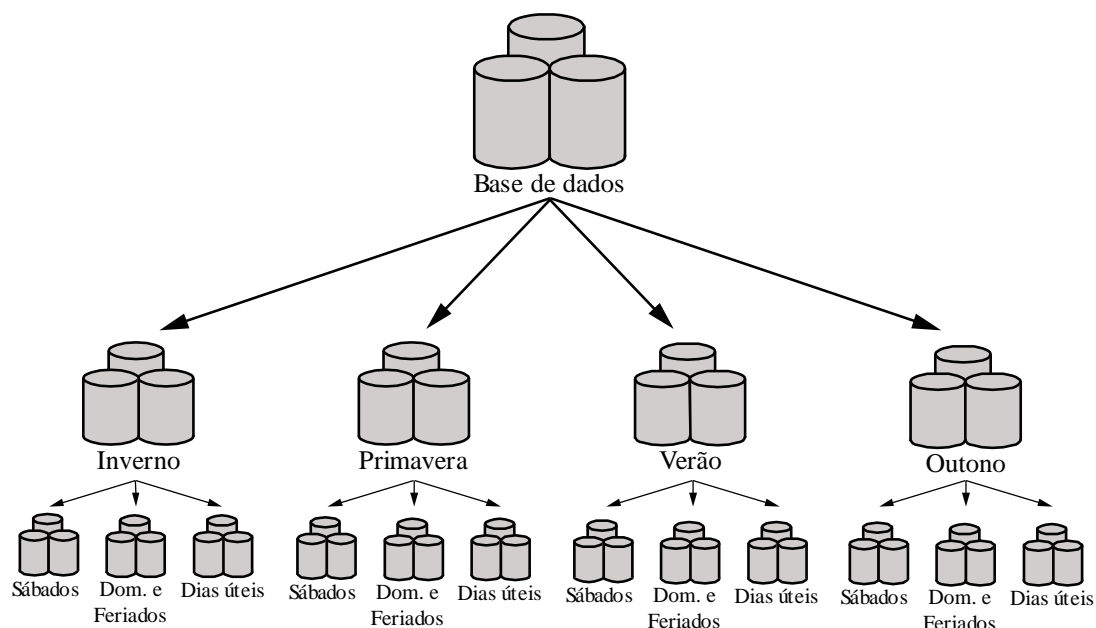
Figura 16 – Visualização gráfica da metodologia para estimar valores faltantes.

#### 4.2.2. REDUÇÃO DOS DADOS

O consumo de energia elétrica pode ser influenciado por diversos fatores, dentre eles as estações do ano. A temperatura média anual afeta a utilização da climatização que reflete no consumo de energia elétrica. Uma divisão inicial pode ser feita separando as curvas de carga conforme as estações do ano (inverno, primavera, verão e outono), pois possuem diferentes padrões climáticos. Outra separação que a ser efetuada é do consumo de energia em formato semestral, agrupando as estações em outono/inverno e primavera/verão.

Os padrões de consumo podem ser diferentes durante os dias da semana, para consumidores baixa tensão, que acabam alterando seus hábitos de consumo durante dias úteis, sábados, domingos e feriados.

Com isso é possível notar a existência de fatores que influenciam o consumo de energia elétrica, um para um horizonte temporal de três meses e outro para um horizonte temporal de uma semana, justifica-se dividir os dados conforme apresentado na Figura 17.



**Figura 17 – Redução dos dados.**

Devido à existência de dados de consumo de energia elétrica de apenas um mês, durante o inverno, fez-se apenas a divisão dos dados apenas entre dias úteis, sábados e domingos/feriados.

Cada consumidor ainda apresenta um diagrama de carga para cada dia analisado, inviabilizando a análise dos dados. Então fez-se necessária a aplicação de outra redução de dados. Nessa redução foi obtido um perfil representativo para os dias úteis, sábados, domingos/feriados para cada consumidor. O perfil representativo foi obtido através da média aritmética dos perfis para cada período analisado.

#### **4.2.3. NORMALIZAÇÃO**

A normalização é uma técnica necessária na caracterização de perfis típicos de carga, onde o objetivo principal é comparar os padrões de consumo de energia, e não o montante de energia consumida.

Um exemplo pode ser visto na Figura 18, onde estão presentes duas curvas de carga. Verifica-se montantes de consumo diferentes para cada consumidor, porém é notório um comportamento de consumo semelhante durante as horas do dia. Se fosse aplicado o algoritmo de agrupamento a esses dois perfis, certamente eles seriam alocados em

agrupamentos diferentes, porém como a intenção não é comparar montante de consumo e sim padrões de consumo.

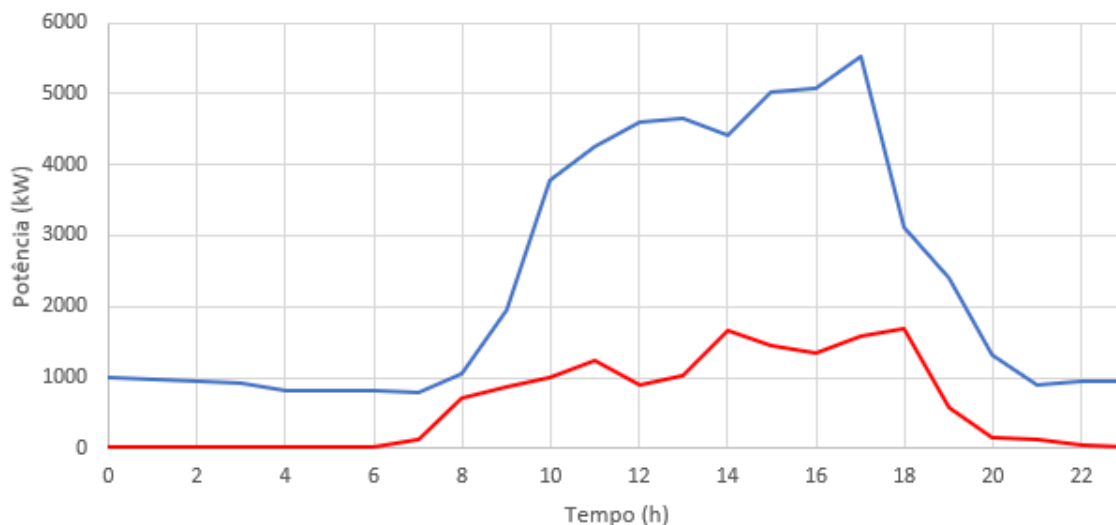


Figura 18 – Curvas de carga sem normalização.

Aplicando a técnica de normalização conhecida como *min-max* pode-se ver claramente, na Figura 19, que os padrões de consumo dos dois consumidores são muito semelhantes. Após essa etapa, caso fosse aplicado o algoritmo de agrupamento, provavelmente as duas curvas de carga seriam agrupadas juntas.

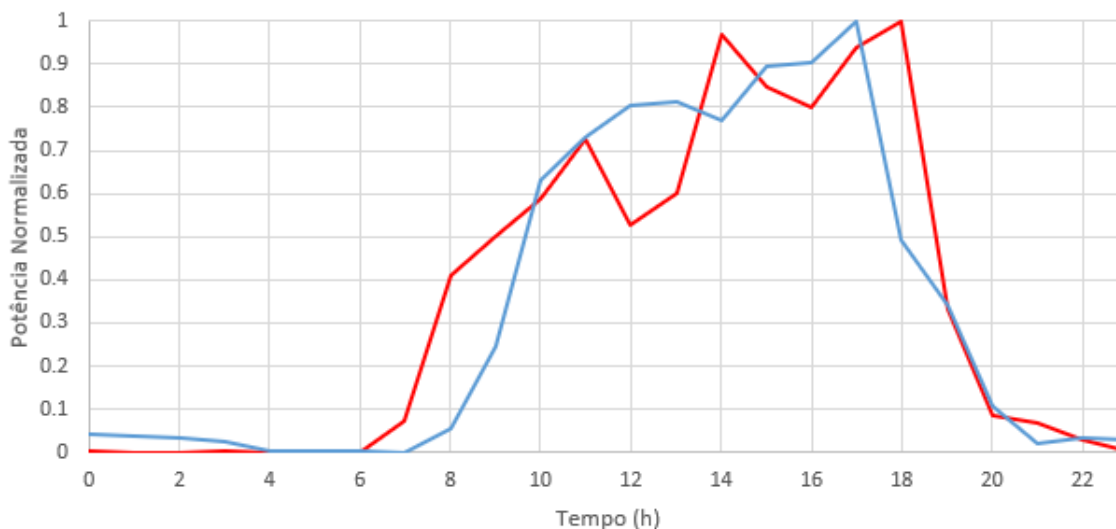


Figura 19 – Curvas de carga após a normalização.

A normalização utilizada nesse trabalho foi a normalização *min-max*, a grande vantagem dessa normalização é que a escala dos dados passa dos valores lidos de potência para valores entre 0 e 1, facilitando a visualização das curvas de carga e a comparação dos padrões de consumo. A formulação dessa normalização pode ser vista na equação (35).

$$z'_i = \frac{z_i - \min_A}{\max_A - \min_A} \quad (35)$$

em que:

$z'_i$  – Representa o dado normalizado;

$z_i$  – Representa o dado a ser normalizado;

$\min_A$  – Representa o menor valor encontrado no intervalo de dados a serem normalizados;

$\max_A$  – Representa o maior valor encontrado no intervalo de dados a serem normalizados.

#### **4.2.4. INTEGRAÇÃO DAS BASES DE DADOS**

Diferentes bases de dados foram utilizadas no problema, com objetivo de diminuir a possibilidade de erros e aumentar a consistência da análise. Porém a integração das bases de dados acaba por não ser um processo fácil, devido ao modo que cada base de dados é formada, geralmente os dados não são armazenados de forma igual, sendo necessário realizar uma fase de pré-processamento para que possa ser feita a integração dos diferentes bancos de dados.

Como mencionado anteriormente, foram utilizadas três bases de dados para efetuar esse estudo. As três bases de dados possuíam características diferentes, tanto no número de consumidores e frequência de amostragem. Devido a isso, primeiramente foi feito o pré-processamento dos dados de forma que fossem completados os valores inexistentes, após isso os dados foram reduzidos e normalizados. Após efetuar o pré-processamento os dados foram passados para um formato que tornou possível a unificação das bases de dados e por fim prosseguir para a etapa de mineração de dados. O modelo de tratamento dos dados pode ser visto na Figura 20.



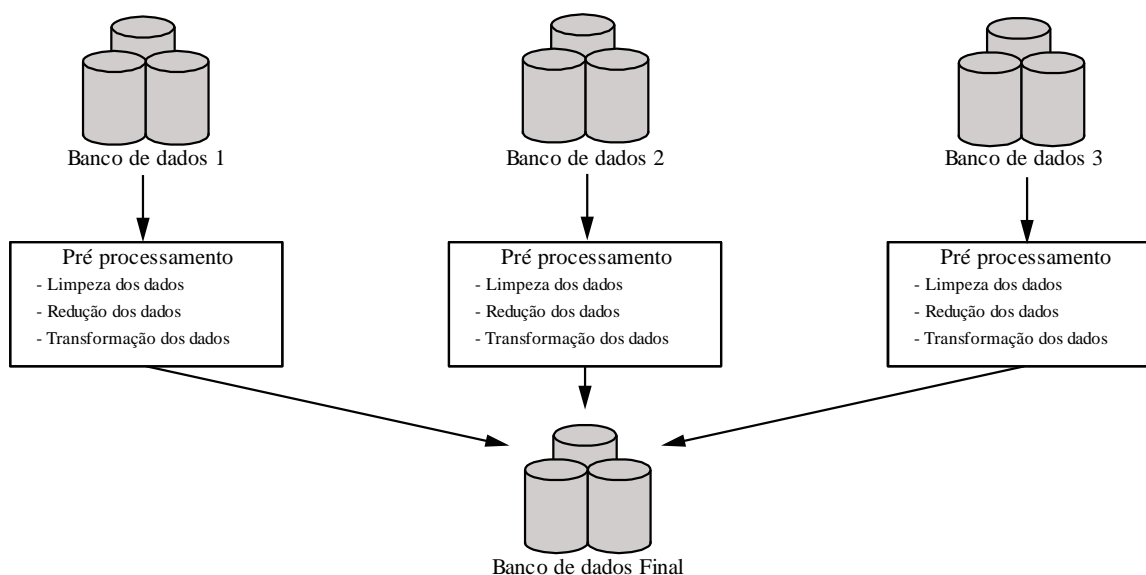


Figura 20 – Integração dos dados.

Os dados pré-processados podem ser vistos no anexo D, onde estão presentes alguns dos 194 consumidores. Os dados foram discretizados em intervalos de uma hora, para facilitar a análise, percepção e visualização dos dados.

### 4.3. METODOLOGIA

Através de um banco de dados, com volume de consumidores considerável, é possível efetuar uma análise de *clustering* e classificação. Para que a análise possa ser efetuada faz-se necessário a utilização de uma metodologia, de forma que cada etapa possa ser feita corretamente. A metodologia utilizada nesse trabalho está apresentada na Figura 21.

A etapa de seleção dos dados consiste em dividir consumidores de acordo com certos parâmetros, nesse estudo foram selecionados apenas consumidores de baixa tensão para a caracterização de perfis típicos de carga. Em seguida esses dados são pré-processados para que sejam removidas as inconsistências que afetam a análise. Na etapa de pré-processamento os dados ainda são reduzidos e transformados, facilitando a compreensão e análise dos resultados.

Saindo da etapa de pré-processamento tem-se os diagramas representativos de cada consumidor para dias úteis, sábados e domingos/feriados, que são os dados de entrada para a etapa de *clustering*. Nessa etapa são utilizados “X” algoritmos de *clustering*, na tentativa de identificar qual obtém melhor resultado para o banco de dados. São definidos o número

mínimo e o número máximo de agrupamentos que serão testados e cada algoritmo separa os dados de “ $k_{min}$ ” até “ $k_{max}$ ” agrupamentos. As partições geradas nos algoritmos de *clustering* são avaliadas através de “Z” índices de validação.

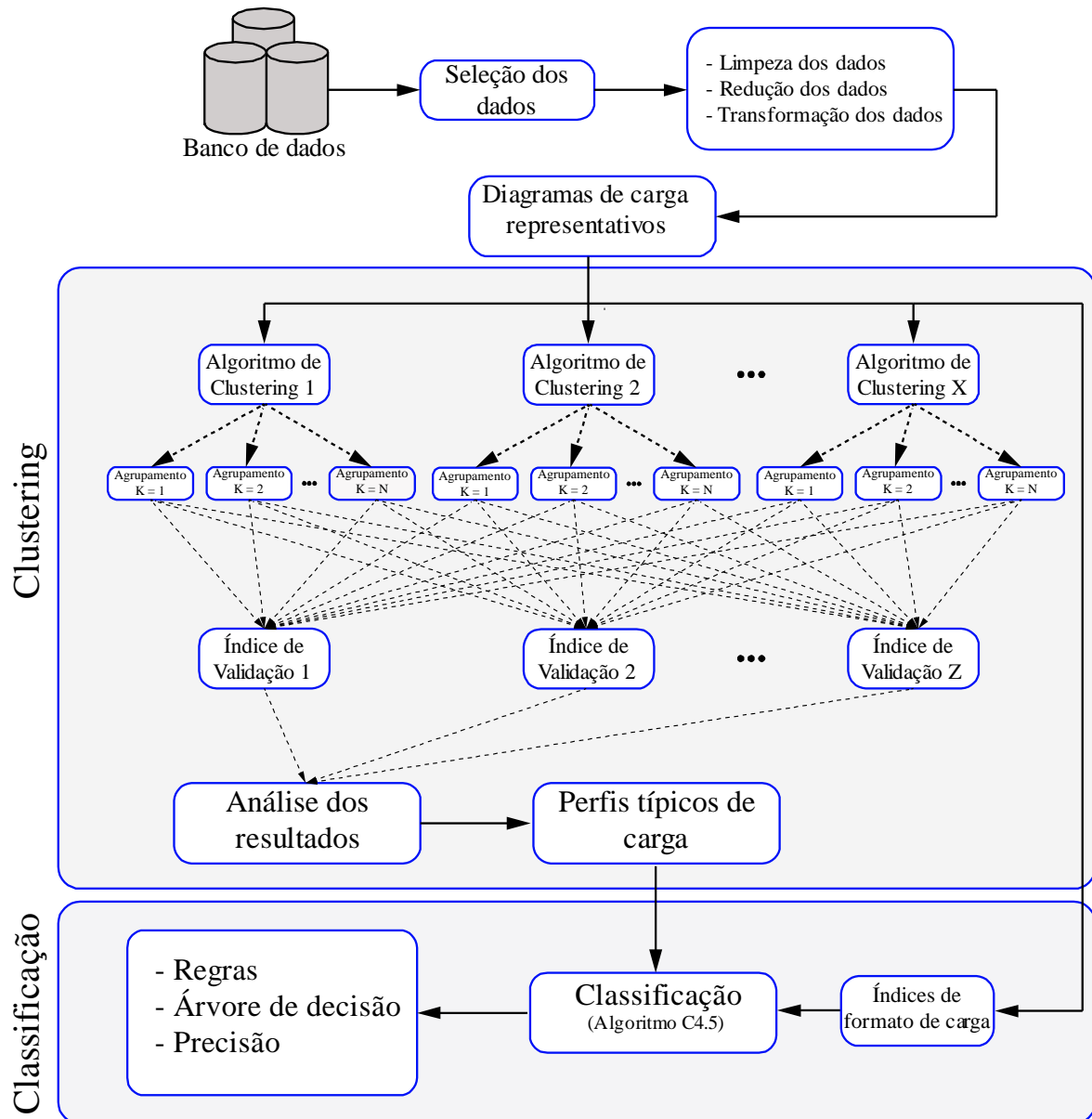


Figura 21 – Metodologia de *Clustering* e *Classificação*.

Obtidos os resultados dos índices de validação para cada algoritmo de *clustering* e para cada número de agrupamentos, os resultados são avaliados por especialistas, com objetivo de identificar as melhores partições obtidas pelos algoritmos. Decidido qual algoritmo obteve o melhor resultado e qual será o número de partições, são seleccionados os diagramas

representativos de carga, resultantes do algoritmo de *clustering* que melhor agrupou os dados.

Terminada a etapa de *clustering*, segue-se para a etapa de classificação. Nessa etapa o algoritmo de classificação é treinado com dados provenientes da etapa de *clustering* e com os diagramas representativos de carga. Os diagramas representativos são discretizados em uma série de índices de formato de carga que auxiliam no treinamento do algoritmo de classificação. Após o treinamento, é gerado um conjunto de regras capaz de classificar novos consumidores com base nos resultados obtidos na etapa de *clustering*.

#### 4.4. AGRUPAMENTO

No Capítulo 3 foram apresentados diferentes algoritmos de agrupamento de dados, entre eles algoritmos particionais, hierárquicos, difusos, baseados em redes neurais artificiais, evolucionários, baseados em densidade e baseados em grade. Para esse caso de estudo foram selecionados alguns algoritmos, sendo eles:

- Algoritmo particional *K-means* (KM);
- Algoritmo particional *K-medoids* (KD);
- Algoritmo particional *G-means* (GM);
- Algoritmo hierárquico *Average link* (AL);
- Algoritmo hierárquico *Single link* (SL);
- Algoritmo hierárquico *Complete link* (CL);
- Algoritmo hierárquico *Ward link* (WL).

Seis desses algoritmos são facilmente encontrados na literatura de agrupamento de curvas de carga, porém não foram encontradas referências da utilização do algoritmo *G-means* na mesma literatura, então será efetuada a análise do seu desempenho na caracterização de perfis típicos de carga.

Para a validação dos agrupamentos foram utilizados seis índices de validação com critérios de avaliação de minimização e maximização, sendo eles:

- *Mean Index Adequacy* (MIA) – critério de minimização;
- *Clustering Dispersion Indicator* (CDI) – critério de minimização;
- *Davies-Bouldin Index* (DBI) – critério de minimização;
- *Silhouette Index* (SI) – critério de maximização;
- *Dunn Index* (DI) – critério de maximização;
- *Calinski-Harabasz Index* (CHI) – critério de maximização.

#### **4.4.1. AVALIAÇÃO DOS ALGORITMOS DE AGRUPAMENTO**

A escolha do número de agrupamentos, bem como da melhor partição, juntamente com o algoritmo que apresenta a melhor partição dos dados foi feita de forma numérica e gráfica, através dos índices de validação. Também foram analisadas as partições obtidas por cada algoritmo de forma visual. A metodologia utilizada na validação está apresentada na Figura 21, em que cada partição dos dados feita por cada algoritmo produz um valor para cada índice.

Foram realizadas duas análises, primeiramente com o número de agrupamentos variando entre 2 e 12 e uma segunda análise com o número de agrupamentos variando entre 3 e 10. As duas análises foram efetuadas para verificar se as novas partições obtidas apresentassem melhor qualidade. Na primeira análise o algoritmo *K-means* obteve melhor desempenho para os índices de Davies-Bouldin e Dunn, com duas partições, e para o índice de Calinski-Harabasz com três partições. Na segunda análise o algoritmo *K-means* obteve melhor resultado para os índices *Silhouette* e Calinski-Harabasz com três partições. Outro algoritmo que apresentou bons resultados para os índices de validação foi o algoritmo *Average Link*, em que na primeira análise obteve melhores resultados para o índice CDI com três agrupamentos e na segunda análise obteve melhores resultados para o CDI e para DI. Com resultados parecidos com o algoritmo *Average Link* tem-se o algoritmo *Single Link*, obtendo melhor resultado para o índice MIA, na primeira análise e para o índice MIA e DBI na

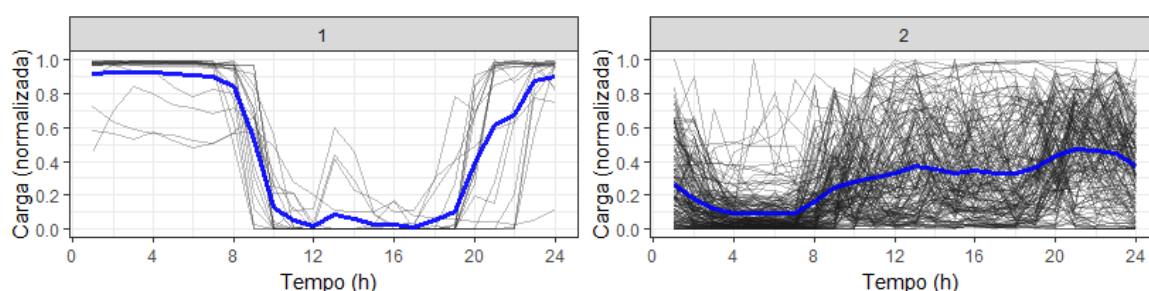
segunda análise. No geral o algoritmo *K-means* obteve melhor resultado frente aos outros algoritmos analisados.

**Tabela 8 – Avaliação dos algoritmos de *clustering* através dos índices de validação.**

Variação do número de agrupamentos	Índice de validação	MIA	CDI	DBI	SI	DI	CHI
2-12 agrupamentos	Algoritmo	SL	AV	KM	KD	KM	KM
	Número de Agrupamentos	4	3	2	2	2	3
3-9 agrupamentos	Algoritmo	SL	AV	SL	KM	AV	KM
	Número de Agrupamentos	4	3	4	3	3	3

#### 4.4.2. AVALIAÇÃO DO ALGORITMO *K-MEANS*

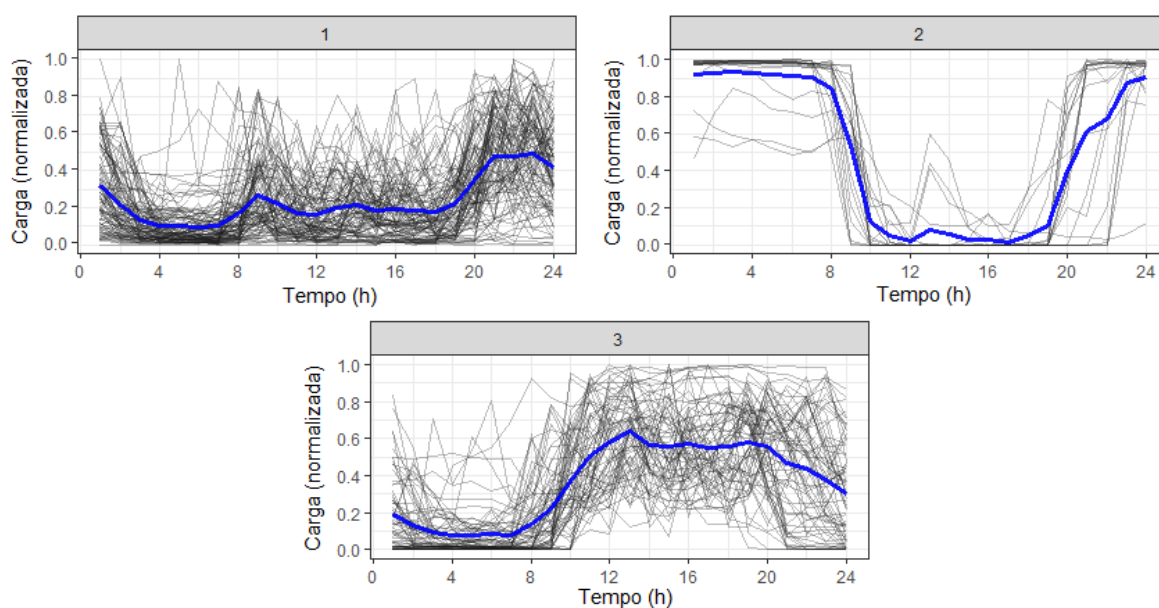
Para comprovar se o algoritmo *K-means* fez a divisão correta dos dados, foram feitas três análises. Na primeira análise o parâmetro *K* possuía valor igual a dois, ou seja, o algoritmo deveria dividir os dados em duas partições. Na Figura 22 está o resultado das duas partições, para o primeiro agrupamento nota-se uma boa divisão e os dados possuem certa semelhança com a média encontrada pelo algoritmo. No segundo agrupamento nota-se que os dados não são parecidos com o centro encontrado pelo algoritmo. Com isso fica perceptível que o algoritmo não consegue representar ou efetuar a correta divisão dos dados em apenas duas partições.



**Figura 22 – *K-means* com dois agrupamentos.**

Com três partições o algoritmo conseguiu encontrar três padrões de curvas de carga distintos (Figura 23). A primeira partição feita pelo algoritmo apresenta características de uma

atividade residencial, apresentando aumento do consumo de energia nas primeiras horas do dia, uma leve diminuição no horário do almoço e um pico de consumo no horário da noite que reduz a partir das 22 horas. A segunda partição encontrada pelo algoritmo apresenta um consumo “atípico”, pois a maior parte do consumo de energia é efetuado no período noturno e existe um consumo residual durante o dia. A terceira partição pode ser caracterizada como uma atividade comercial, pois, no período da madrugada apresenta um consumo residual de energia e o consumo começa a se elevar durante o período da manhã, mantendo-se constante até as 20 horas quando começa a diminuir.



**Figura 23 – *K-means* para três agrupamentos.**

Para quatro divisões, Figura 24, o algoritmo apresentou partições distintas, porém é notório certa similaridade entre as partições 1 e 3, não fornecendo tanto ganho de informação. O algoritmo não apresentou ganho significativo para os índices de validação aumentando o número de partições de três para quatro.

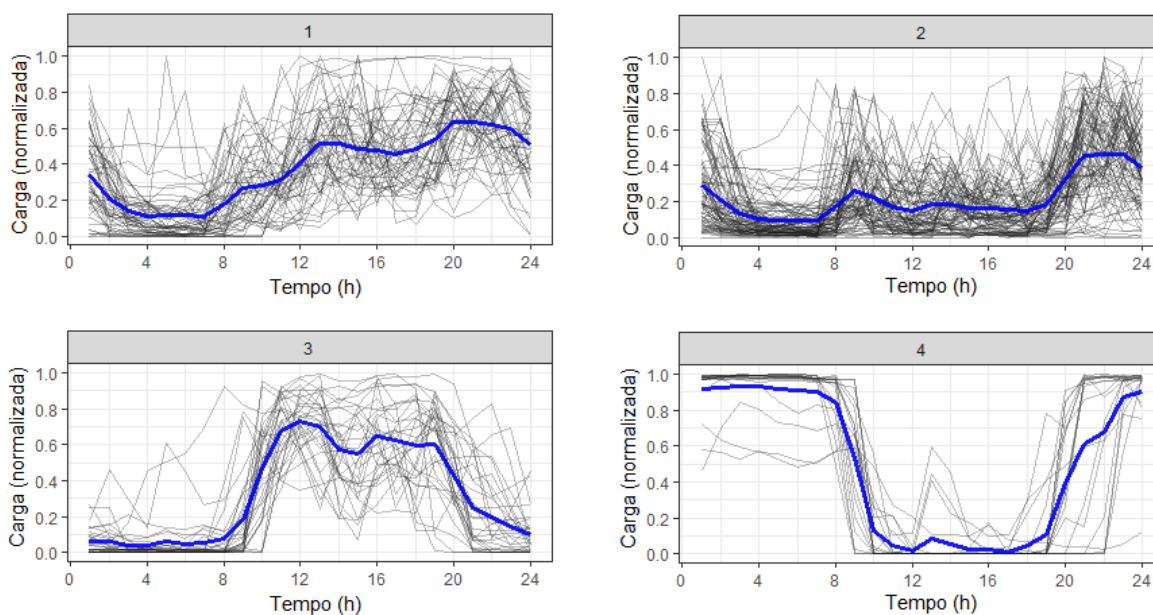


Figura 24 – *K-means* para quatro agrupamentos.

#### 4.4.3. AVALIAÇÃO DO ALGORITMO *K-MEDOIDS*

O algoritmo *K-medoids* apresentou partições semelhantes ao algoritmo *K-means*, o que já era esperado, pois o mesmo é uma variação do *K-means*. Visualmente nota-se a principal diferença entre os dois algoritmos, o *K-means* calcula um centro através da média dos dados, enquanto que o *K-medoids* escolhe um dado como centro. Outra diferença é o cálculo das distâncias entre as séries temporais, o algoritmo *K-means* utiliza a distância Euclidiana, enquanto que o *K-medoids* utiliza a distância de Manhattan.

Na Figura 25 fica perceptível que esse algoritmo, assim como o *K-medoids* não consegue efetuar boa divisão dos dados, considerando apenas duas partições.

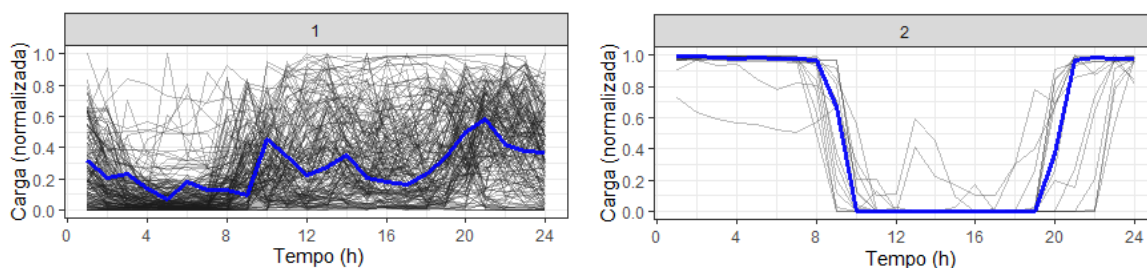


Figura 25 – *K-medoids* para dois agrupamentos.



Para três partições o algoritmo consegue encontrar três padrões distintos de consumo (Figura 26), com resultados semelhantes aos encontrados pelo algoritmo *K-means*.

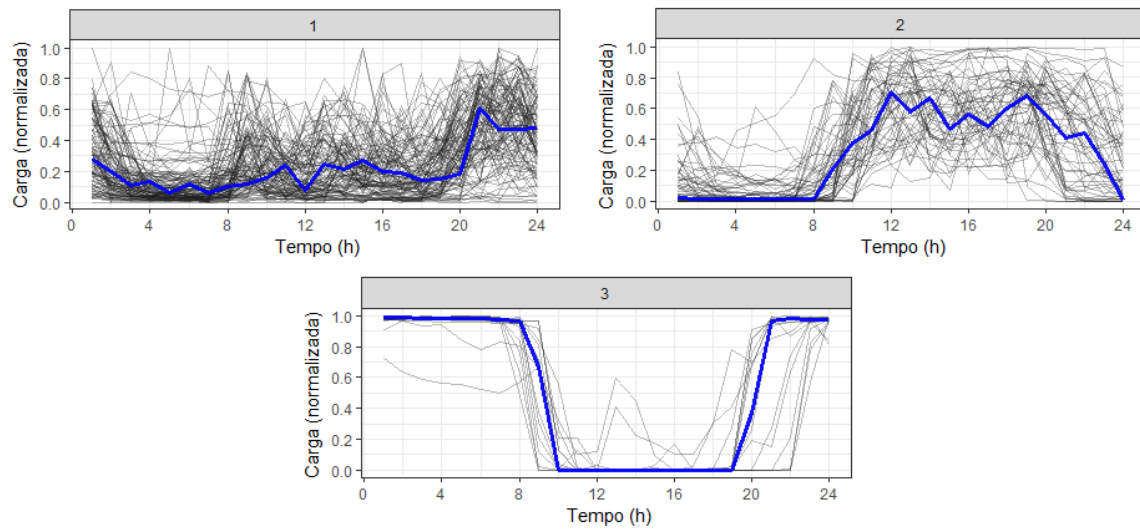


Figura 26 – *K-medoids* para três agrupamentos.

Na Figura 27 estão presentes as quatro partições efetuadas pelo algoritmo, ficando perceptível a diminuição do ganho de informação quando é aumentado o número de agrupamentos. Nota-se que os agrupamentos um e dois possuem grande semelhança o que justifica a utilização de três agrupamentos.

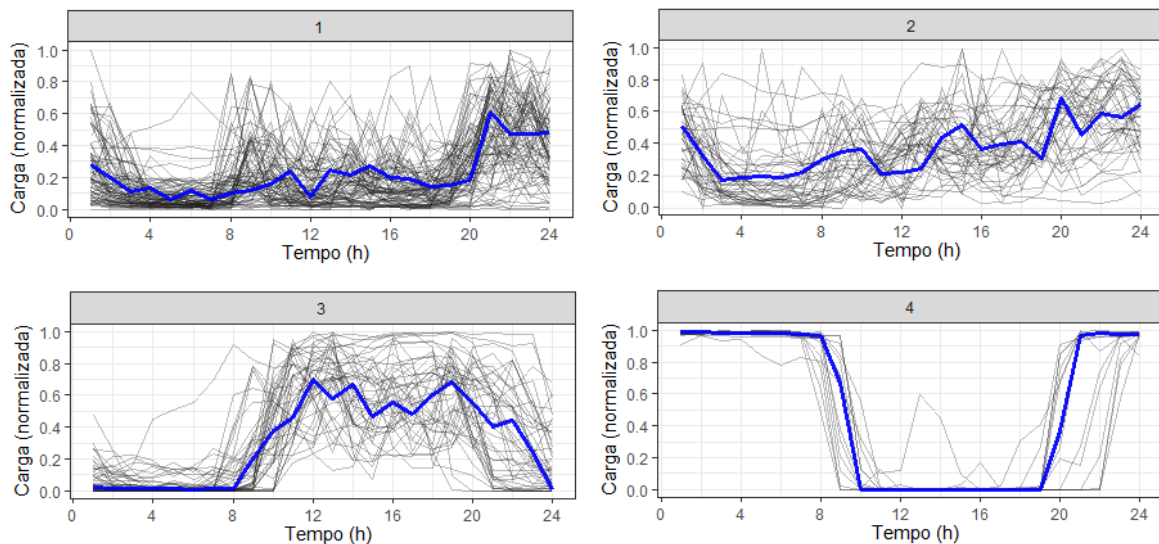


Figura 27 – *K-medoids* para quatro agrupamentos.



#### 4.4.4. AVALIAÇÃO DO ALGORITMO *G-MEANS*

O algoritmo *G-means* apresenta duas particularidades frente aos outros algoritmos analisados nesse estudo de caso. O algoritmo *G-means* nada mais é do que o algoritmo *K-means* com um teste para verificar se os dados agrupados seguem uma distribuição gaussiana. Outro ponto é que não é necessário atribuir o número de partições desejadas, isso pode ser uma vantagem ou uma desvantagem. Vantagem quando apenas deseja-se efetuar a divisão dos dados sem um número máximo ou mínimo de agrupamentos a serem obtidos. Desvantagem, nesse caso para a nossa análise quando é desejado obter um número determinado de partições, por exemplo, não seria interessante uma comercializadora de energia possuir dez perfis de carga, quando só existem três ou quatro modalidades tarifárias tornando os resultados da análise de *clustering* inviável (através do algoritmo *G-means*).

Porém é válida a análise de um algoritmo de clustering que não necessita de um parâmetro pré-determinado, que determina em quantas vezes o conjunto de dados deve ser dividido. Na Figura 28 está disposto o resultado dos agrupamentos do algoritmo. O conjunto de dados foi dividido em 13 agrupamentos e nota-se que os dados foram realmente bem divididos em cada agrupamento, por outro lado existem muitos agrupamentos que poderiam ser considerados apenas como um, por exemplo, os agrupamentos um e três, os agrupamentos nove e onze e os agrupamentos cinco e dez. A divisão ficou bem definida, porém devido à existência de duas ou três modalidades tarifárias a utilização de treze agrupamentos se torna inviável.

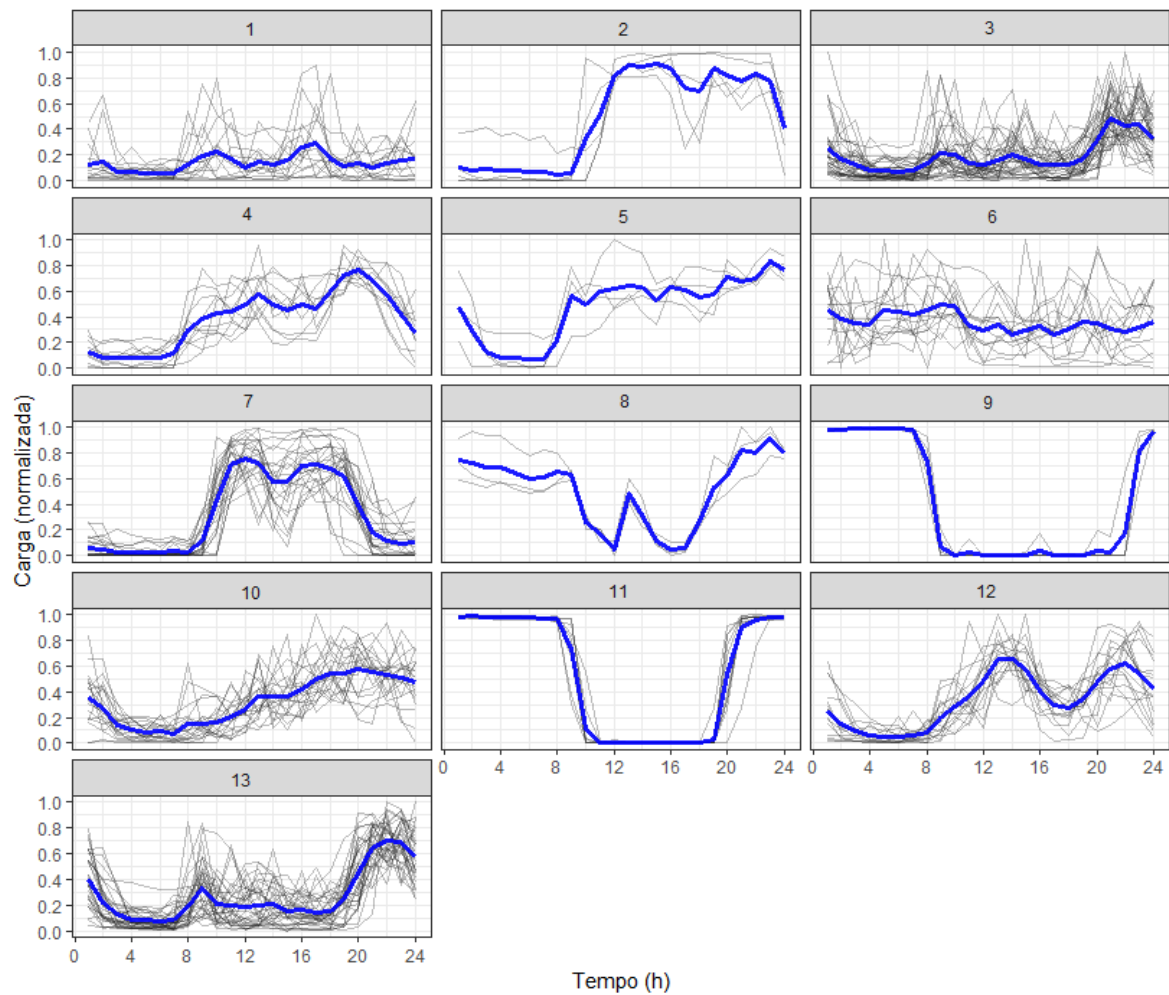


Figura 28 – Resultado do algoritmo *G-means*.

#### 4.4.5. AVALIAÇÃO DOS ALGORITMOS HIERÁRQUICOS

A avaliação dos algoritmos hierárquicos consiste na análise do dendrograma de cada algoritmo e nos agrupamentos formados. Na Figura 29 estão apresentados os dendrogramas de cada algoritmo hierárquico analisado. É notório que assim como os algoritmos particionais os algoritmos hierárquicos também não conseguem efetuar uma boa divisão dos dados, considerando apenas dois agrupamentos. O algoritmo que conseguiu realizar uma divisão “equitativa”, considerando o número de consumidores em cada agrupamento, foi o algoritmo *Complete link*, e o algoritmo que obteve o pior resultado da divisão foi o *Single Link*.

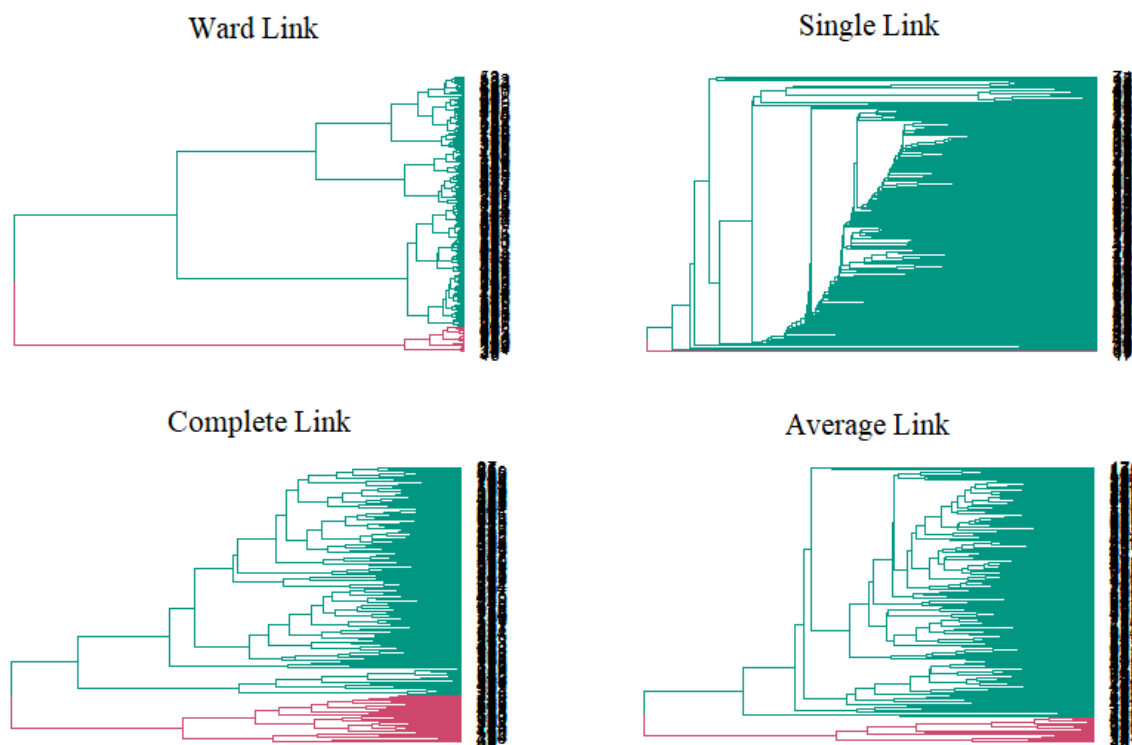


Figura 29 – Resultado dos algoritmos hierárquicos para dois agrupamentos.

Para três agrupamentos, Figura 30, os algoritmos que obtiveram melhores partições foram os algoritmos *Ward* e *Complete Link*. Vê-se claramente uma divisão mais “equitativa” entre os três agrupamentos. Por outro lado, o algoritmo *Single Link* continua apresentando uma fraca divisão dos dados.

Efetuada a análise para quatro agrupamentos, Figura 31, nota-se que somente o algoritmo *Ward Link*, faz uma boa divisão entre os quatro agrupamentos, enquanto que os outros algoritmos apresentam um agrupamento muito grande (em azul) e outros agrupamentos pequenos, possivelmente agrupamentos com *outliers* ou perfis muito diferentes.

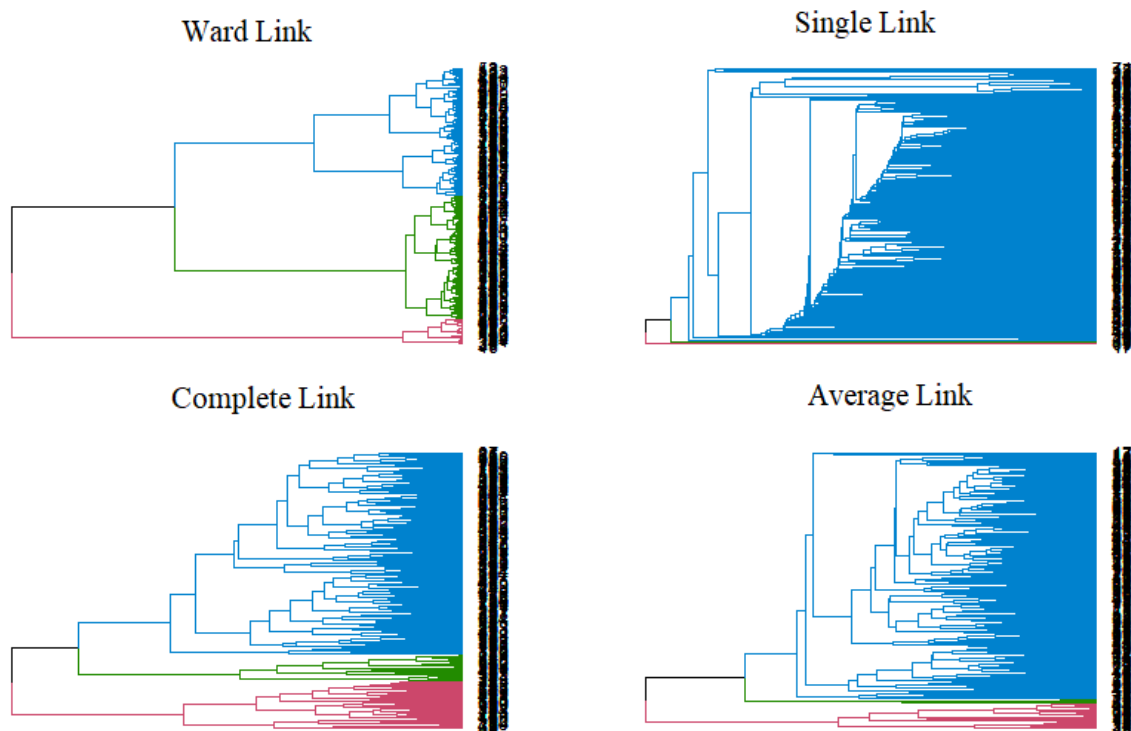


Figura 30 – Resultado dos algoritmos hierárquicos para três agrupamentos.

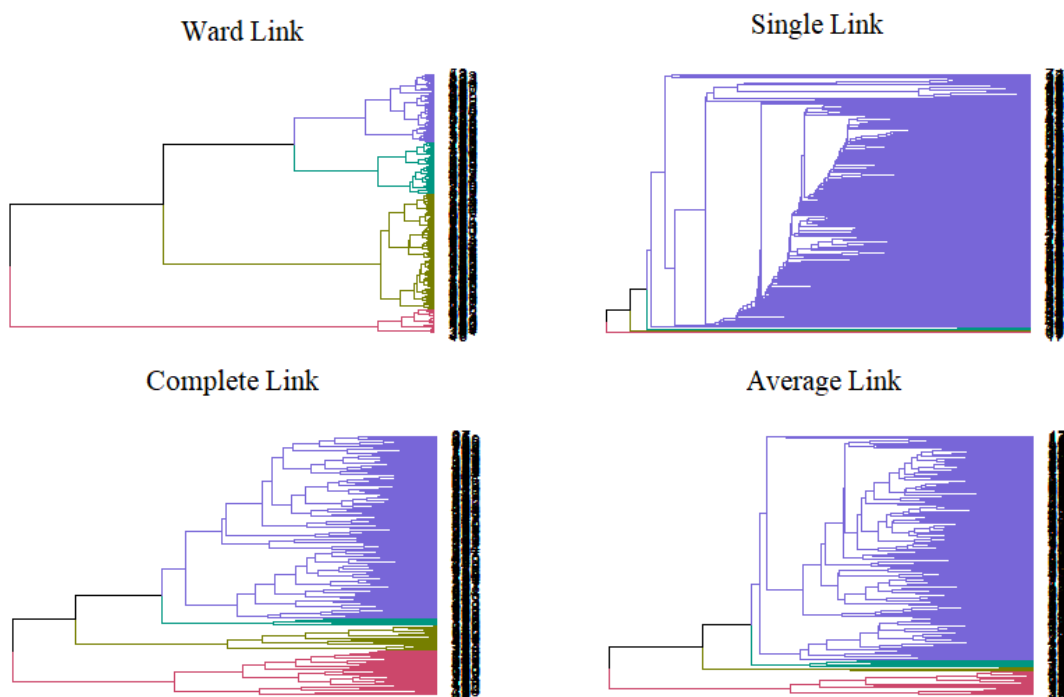
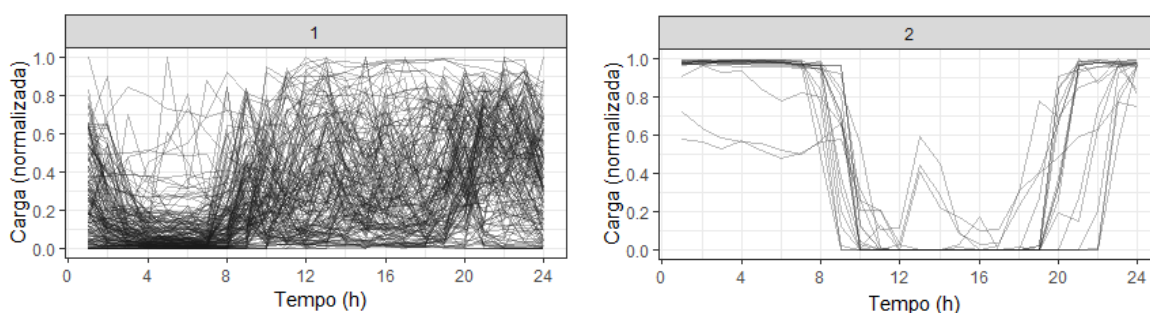


Figura 31 – Resultado dos algoritmos hierárquicos para quatro agrupamentos.

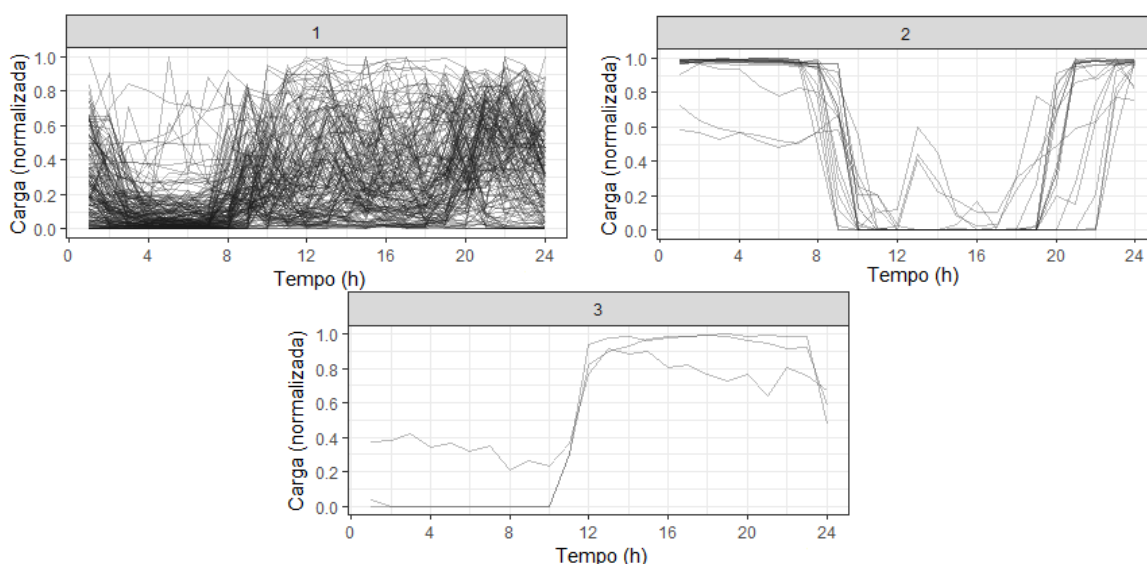
Para efetuar uma análise mais minuciosa do desempenho dos algoritmos, foram plotadas as partições dos algoritmos hierárquicos que obtiveram melhor desempenho nas divisões, algoritmo *Ward* e *Complete Link*. Foram descartados dessa análise os algoritmos *Single* e *Average Link*, pois não apresentaram boas partições na análise dos dendrogramas.

A Figura 32 apresenta os resultados para as duas partições encontradas pelo algoritmo *Average Link*, nota-se que são muito semelhantes as partições encontradas pelos algoritmos hierárquicos.



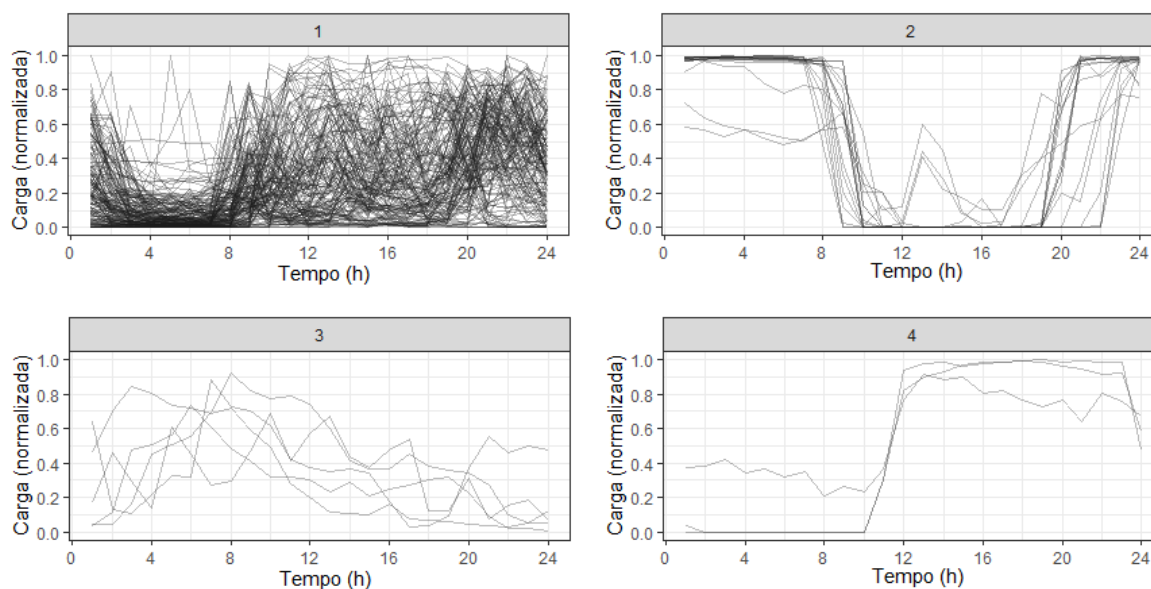
**Figura 32 – Resultado do algoritmo *Average Link* para dois agrupamentos.**

Quando é efetuada a análise para três partições, Figura 33, o algoritmo não consegue efetuar corretamente a divisão, apresentando um terceiro *cluster* com poucos dados, não apresentando ganho de informação.



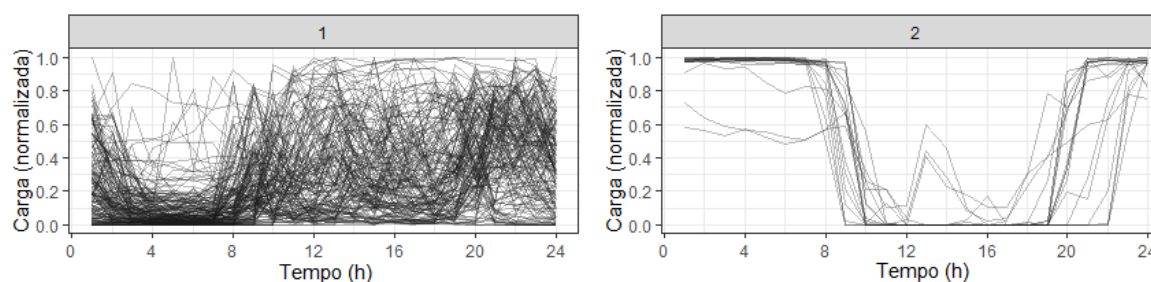
**Figura 33 – Resultado do algoritmo *Average Link* para três agrupamentos.**

Para quatro partições, Figura 34, o resultado é tão ruim quanto para três partições, pois, o primeiro agrupamento não foi dividido, o segundo se mantém muito parecido ao da análise com dois agrupamentos, o terceiro e o quarto agrupamento apresentam poucos dados.



**Figura 34 – Resultado do algoritmo *Average Link* para quatro agrupamentos.**

Quando analisado o algoritmo *Ward Link* para duas partições, Figura 35, nota-se uma divisão semelhante a apresentada pelos algoritmos particionais, porém, o algoritmo não consegue representar os dados apenas com duas partições.



**Figura 35 – Resultado do algoritmo *Ward Link* para dois agrupamentos.**

Analisando para três partições, Figura 36, é possível notar que o algoritmo *Ward Link* apresenta boa divisão dos dados, ao contrário dos outros algoritmos hierárquicos analisados nesse trabalho. As partições encontradas são muito semelhantes às encontradas pelos algoritmos particionais.



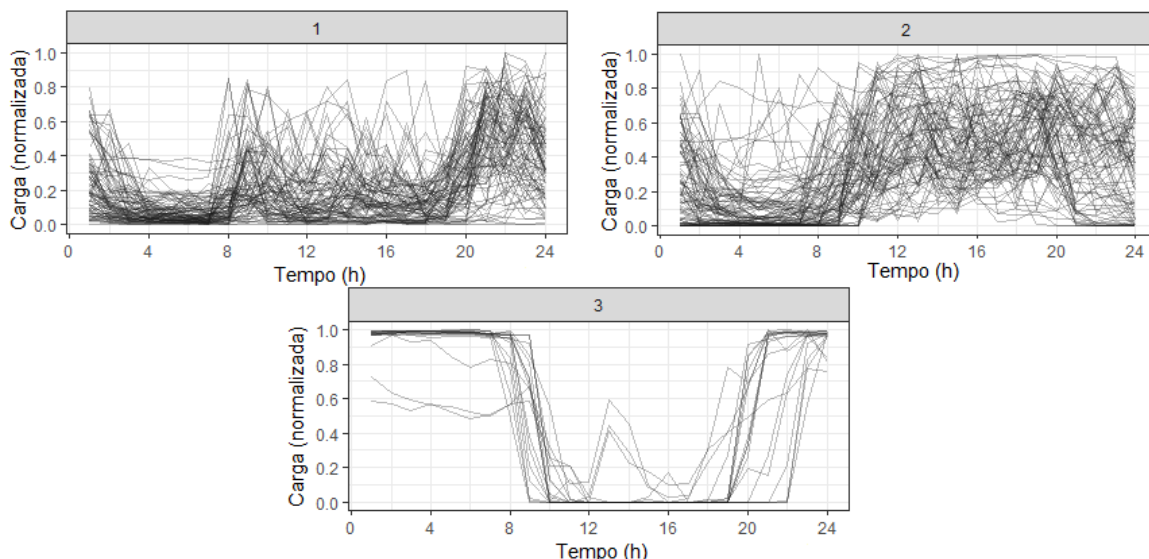


Figura 36 – Resultado do algoritmo *Ward Link* para três agrupamentos.

Na Figura 37 está apresentado o resultado do algoritmo *Ward Link* para quatro partições, e diferentemente do algoritmo *Average Link* esse algoritmo consegue realizar melhor a divisão dos dados. Novamente os resultados apresentados por esse algoritmo se assemelham aos resultados encontrados pelos algoritmos particionais.

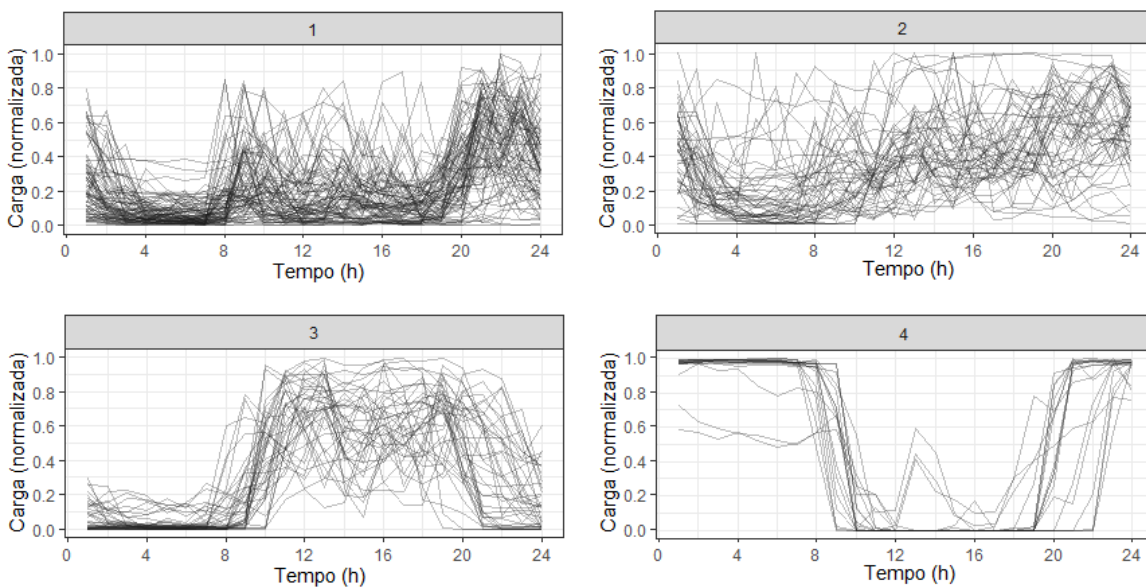


Figura 37 – Resultado do algoritmo *Ward Link* para quatro agrupamentos.

#### 4.4.6. ESCOLHA DO MELHOR ALGORITMO DE AGRUPAMENTO

Efetuada a análise gráfica e numérica dos índices de validação, bem como a análise das partições obtidas por cada algoritmo, chegou-se a conclusão que o algoritmo que obteve melhor desempenho nas duas análises foi o algoritmo *K-means*, com três partições. Foram selecionadas as curvas típicas de carga para os dias da semana, sábados e domingos/feriados para efetuar a análise dos padrões obtidos.

Na Figura 38 estão presentes os perfis típicos de carga para os dias da semana. Percebe-se claramente a existência de três perfis diferentes. O primeiro perfil, caracterizado por um consumo elevado nas horas da madrugada e da noite, e um consumo residual durante o dia, pode ser caracterizado como um perfil “anormal” para consumidores de baixa tensão (estima-se que esse seja o perfil de iluminação pública). O segundo perfil de carga, em azul, apresenta características de consumidores residenciais, pois, durante a madrugada apresenta um consumo residual de energia, e nas primeiras horas do dia apresenta um aumento de carga. Nesse segundo perfil o consumo se mantém praticamente constante durante as nove horas da manhã até as seis horas da tarde, a partir desse ponto o consumo cresce rapidamente, pois é nesse horário que a maioria dos moradores se encontram em suas residências e consomem energia. O terceiro perfil de carga, em amarelo, apresenta traços de uma atividade comercial, onde no início da manhã o consumo aumenta gradativamente e se mantém praticamente constante até o começo da noite, onde decresce gradativamente.

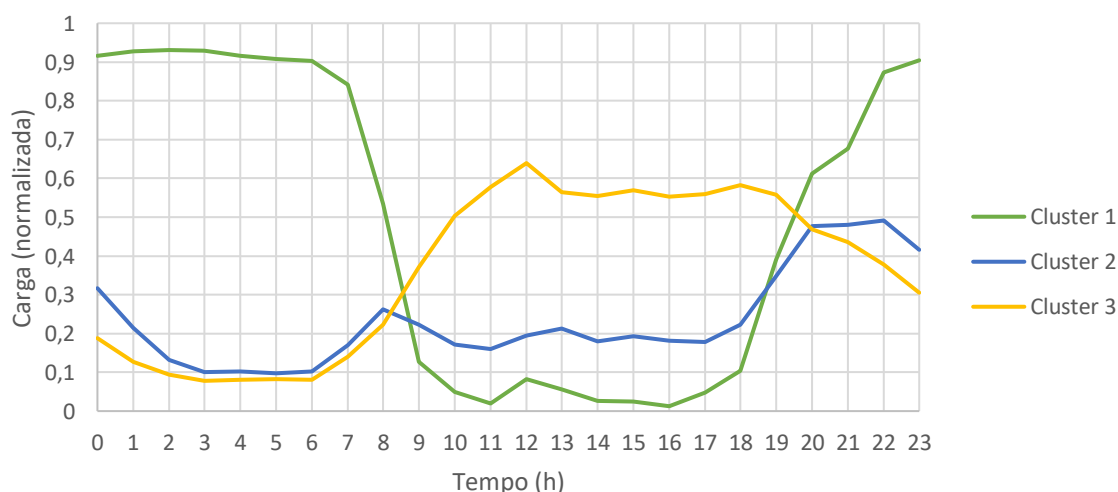
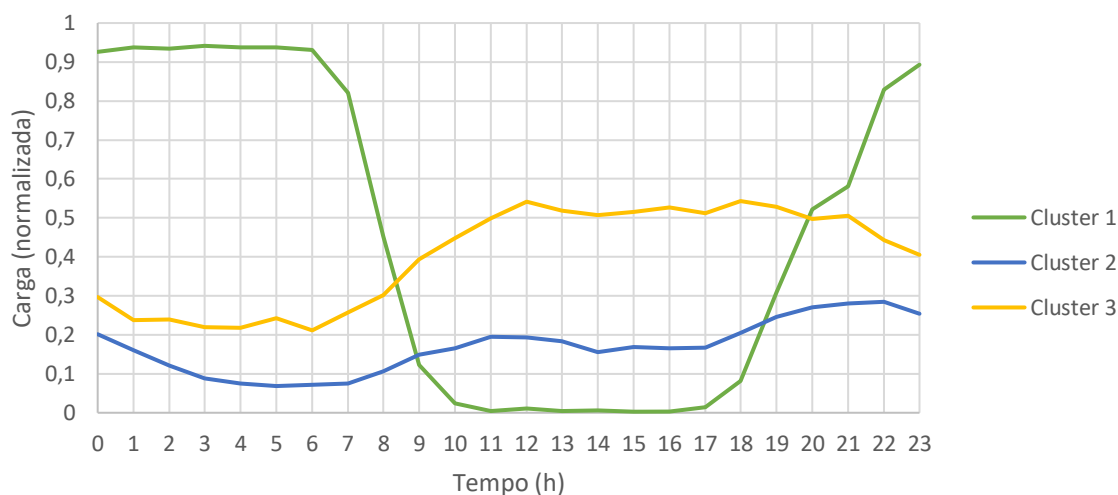


Figura 38 – Perfis Típicos de Carga para os dias da semana.



Na Figura 39 são apresentados os perfis típicos de carga para os sábados. Nota-se que o primeiro perfil, em verde, continuou praticamente igual o perfil encontrado durante para os dias da semana. Para o segundo perfil de carga, em azul, vê-se claramente a diminuição do pico de consumo para os sábados, frente aos dias da semana. Referente ao terceiro perfil de carga, em amarelo, é possível notar uma diminuição do volume de consumo, porém esse perfil se mantém praticamente igual ao perfil de carga encontrado para os dias da semana.



**Figura 39 – Perfis Típicos de Carga para sábados.**

Dos perfis típicos obtidos para os domingos e feriados, Figura 40, nota-se o perfil “atípico”, em verde, manteve-se praticamente igual. No tocante ao segundo perfil de carga, em azul, nota-se certa diminuição do consumo nas horas da tarde, entre as treze e as dezoito horas, porém no restante do dia o perfil se manteve muito similar aos sábados. Para terceiro perfil de carga, em amarelo, nota-se uma certa diminuição de consumo entre as treze e as dezoito horas. Cabe ressaltar que os perfis agrupados no *cluster 2* e *3*, apresentam certa similaridade quanto ao padrão de consumo, porém apresentam montantes diferentes de consumo durante os períodos horários, o que acabou distinguindo um perfil do outro (apenas para sábados e domingos, para os dias da semana os hábitos de consumo são totalmente diferentes).

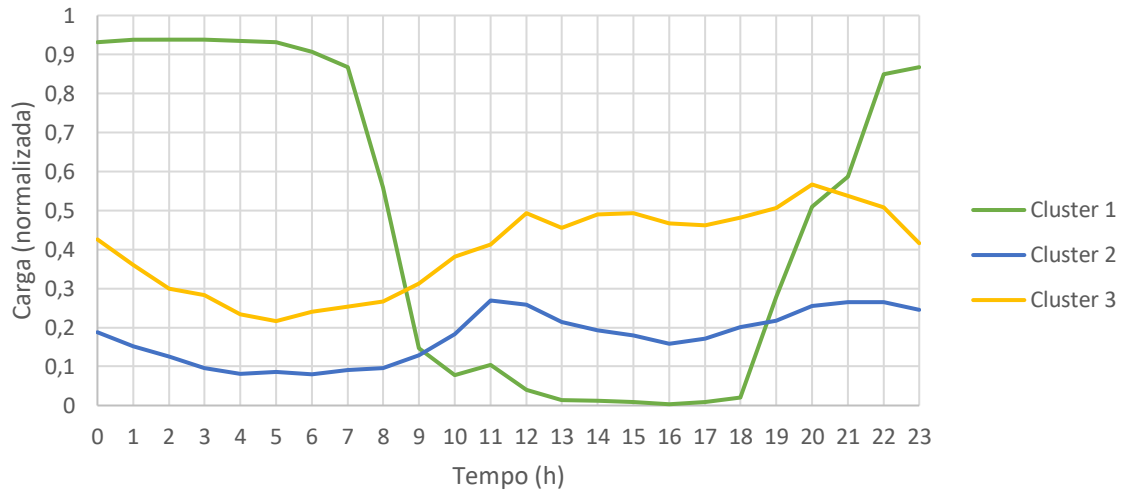


Figura 40 – Perfis Típicos de Carga para o domingo/feriado.

Na Figura 41 estão dispostos os percentuais de clientes em cada *cluster* para os dias da semana, sábados e domingos/feriados, de fora para dentro, respetivamente. Para os dias da semana nota-se que o perfil “atípico” representou 10% dos consumidores, e manteve-se praticamente constante para os sábados e domingos/feriados representando cerca de 9% dos consumidores. Pode-se caracterizar esse perfil como uma atividade industrial, com prevalência de suas atividades nos períodos noturnos. O *cluster 2* representa 36% dos consumidores para os dias da semana, nota-se um aumento dos consumidores desse *cluster* para os sábados e domingos. O *cluster 3* representou 54% dos consumidores para os dias da semana, enquanto que para os sábados e domingos representou 28% e 24% respetivamente.

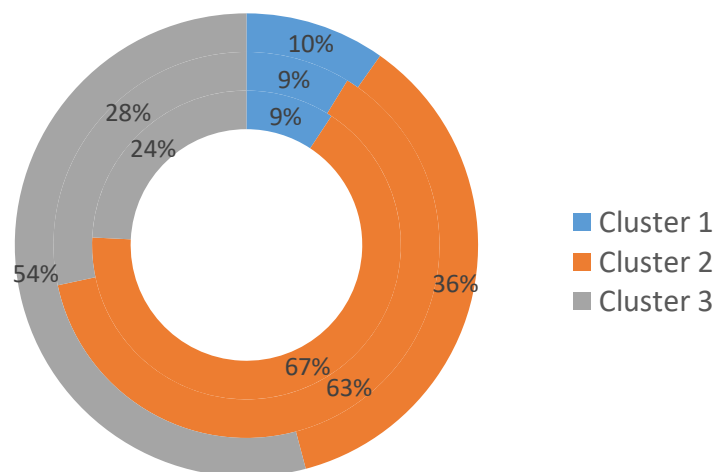


Figura 41 – Percentual de clientes em cada *cluster*.

## 4.5. CLASSIFICAÇÃO

Para fechar o ciclo da mineração de dados é feita a etapa de classificação. A etapa de *clustering* é considerada como aprendizagem não supervisionada, onde os algoritmos “aprendem” somente com base nos dados, e extraem a informação pertinente aos perfis de consumo. O algoritmo de *clustering* separou as curvas de carga com base nas similaridades entre as mesmas, essa informação será utilizada na etapa de treino do algoritmo de classificação.

Nessa etapa será utilizado um modelo de classificação, juntamente com a informação proveniente do *clustering* para que possa ser feita a classificação de novos consumidores com base em apenas os seus dados históricos e em um modelo de classificação.

Para classificar os diagramas de carga foi utilizado o algoritmo C4.5, o mesmo cria uma árvore de decisão capaz de classificar novos consumidores. O algoritmo opera em duas etapas, primeiro o treino onde o é passado um conjunto de dados  $X = [x_1, x_2, \dots, x_n]$  já classificados. Cada  $x_i$  consiste em um vetor n-dimensional onde cada atributo representa as características da curva de carga, assim como a classe em que o  $x_i$  pertence. Em cada nó da árvore de decisão o algoritmo escolhe o atributo dos dados que melhor particiona o grupo de curvas de carga em subgrupos. O algoritmo continua dividindo os dados repetindo o processo anterior [69].

Primeiramente é realizado o *clustering* com objetivo principal de separar os dados em suas classes de acordo com suas similaridades. Posteriormente as curvas de carga são discretizadas em cinco fatores de formato de carga,  $f = [f_1, f_2, f_3, f_4, f_5]$ , com objetivo de reduzir os dados e obter uma forma mais simples de representar cada curva de carga. Os índices de formato de carga foram calculados de acordo com a Tabela 9.

**Tabela 9 – Índices de formato de carga utilizados.**

<b>Parâmetro</b>	<b>Expressão de Cálculo</b>	<b>Período de Aquisição</b>
Diário $P_{média}/P_{máx}$	$f_1 = \frac{P_{média,dia}}{P_{max,dia}}$	1 dia
Diário $P_{min}/P_{máx}$	$f_2 = \frac{P_{min,dia}}{P_{max,dia}}$	1 dia
Impacto noturno	$f_3 = \frac{1}{3} \frac{P_{média,noite}}{P_{média,dia}}$	1 dia (8 horas durante a noite, das 11 p.m. até 6 a.m.)
Diário $P_{min}/P_{média}$	$f_4 = \frac{P_{min,dia}}{P_{média,dia}}$	1 dia
Impacto do Almoço	$f_5 = \frac{1}{8} \frac{P_{média,almoço}}{P_{média,dia}}$	1 dia (3 horas durante o almoço, das 12 a.m. até 3 p.m.)

Para treinar o modelo de classificação foram selecionados aleatoriamente 2/3 dos dados e 1/3 dos dados foram utilizados para efetuar a previsão das classes. Para treinar o modelo é passado o vetor  $V = [f_1, f_2, f_3, f_4, f_5, c]$ , que contém os índices de formato de carga normalizados e a classe de cada curva de carga pertence.

Na etapa de treino o modelo de classificação utilizou apenas três índices de formato de carga para criar a árvore de decisão. Os atributos utilizados foram  $f_1, f_3$  e  $f_5$ , com um percentual de utilização de 63,85%, 100,00% e 81,54% respetivamente. Com isso nota-se que os índices  $f_2$  e  $f_5$  poderiam ser retirados da análise. As regras de decisão geradas na fase de aprendizagem estão dispostas na Tabela 9.

Após a obtenção do modelo de classificação, foram selecionados 1/3 dos dados aleatoriamente para efetuar a classificação dos mesmos. A matriz de classificação dos dados pode ser vista na Tabela 11.

**Tabela 10 – Regras de decisão geradas na etapa de treino.**

Regras de decisão	Rótulo de Classe
Se $f_3 > 0,444$ & $f_1 \leq 0,363$	<b>cluster 2</b>
Se $f_3 > 0,444$ & $f_1 > 0,363$	<b>cluster 1</b>
Se $f_3 \leq 0,444$ & $f_5 \leq 0,084$	<b>cluster 2</b>
Se $f_3 \leq 0,444$ & $f_5 > 0,084$ & $f_3 \leq 0,114$	<b>cluster 3</b>
Se $f_3 \leq 0,444$ & $f_5 > 0,084$ & $f_3 > 0,114$ & $f_1 \leq 0,328$	<b>cluster 2</b>
Se $f_3 \leq 0,444$ & $f_5 > 0,084$ & $f_3 > 0,114$ & $f_1 > 0,328$ & $f_3 > 0,280$ & $f_1 \leq 0,480$	<b>cluster 2</b>
Se $f_3 \leq 0,444$ & $f_5 > 0,084$ & $f_3 > 0,114$ & $f_1 > 0,328$ & $f_3 > 0,280$ & $f_1 > 0,480$	<b>cluster 3</b>
Se $f_3 \leq 0,444$ & $f_5 > 0,084$ & $f_3 > 0,114$ & $f_1 > 0,328$ & $f_3 \leq 0,280$ & $f_1 > 0,404$	<b>cluster 3</b>
Se $f_3 \leq 0,444$ & $f_5 > 0,084$ & $f_3 > 0,114$ & $f_1 > 0,328$ & $f_3 \leq 0,280$ & $f_1 \leq 0,404$ & $f_3 \leq 0,213$	<b>cluster 3</b>
Se $f_3 \leq 0,444$ & $f_5 > 0,084$ & $f_3 > 0,114$ & $f_1 > 0,328$ & $f_3 \leq 0,280$ & $f_1 \leq 0,404$ & $f_3 > 0,213$	<b>cluster 2</b>

**Tabela 11 – Matriz de classificação dos dados.**

Classe Real	Predição		
	1	2	3
1	4	0	0
2	1	32	1
3	0	2	24

O modelo de classificação classificou corretamente 60 curvas de carga e incorretamente 4 curvas de carga atingindo uma precisão de 93,75% para os dias da semana. Uma curva de carga que pertencia a classe 2 foi classificada como classe 1, duas curvas de carga que pertenciam a classe 3 foram classificadas como classe 2 e uma curva de carga pertencente a classe 2 foi classificada como classe 3. No geral o modelo obteve boa precisão para os dias da semana, para os sábados e domingos/feriados o percentual de acerto do modelo foi de 85,71% e 87,77%, respetivamente.

#### 4.6. CONCLUSÕES

Neste capítulo foram apresentadas as técnicas desenvolvidas, neste trabalho, através do estudo de caso. Primeiramente foram descritos os dados apresentados no problema. Em seguida foi descrita a metodologia de pré-processamento dos dados, que se apresentou de essencial importância na correção dos erros contidos no banco de dados, efetuando a

correção de valores inconsistentes e valores faltantes. Os dados foram normalizados e passados para um formato padrão, tornando possível realizar a integração entre os três bancos de dados.

Foram utilizados diversos algoritmos de *clustering* para identificar qual seria a melhor opção para encontrar os perfis típicos de carga, juntamente com seis índices de validação, no intuito de identificar o melhor número de partições dos dados. O algoritmo que obteve melhor resultado entre os algoritmos analisados foi o algoritmo *K-means*, o mesmo obteve melhor resultado para os índices de validação, frente aos outros algoritmos apresentados. A análise das partições não se limitou apenas a análise numérica dos índices de validação, sendo feita a análise visual das partições obtidas para os algoritmos que obtiveram os melhores resultados, e dentre os analisados o algoritmo *K-means* efetuou as melhores divisões dos dados.

Após a etapa de *clustering* foi implementado um modelo de classificação para fechar o ciclo da mineração de dados, e utilizar o conhecimento obtido nesse estudo para a classificação de novos consumidores. Para representar as curvas de carga de cada consumidor, foram utilizados cinco fatores de carga, porém para esse caso, o modelo de classificação fez a utilização de apenas três índices na etapa de aprendizagem e geração das regras de classificação. O modelo obtido é relativamente simples e apresentou bons resultados na classificação de novos consumidores, apresentando acerto 93,75% na classificação.

## 5. CONCLUSÕES

Este capítulo apresenta as conclusões referentes ao trabalho desenvolvido, levando em conta os objetivos alcançados.

### 5.1. CONCLUSÕES

O conhecimento dos perfis de consumo de energia elétrica se faz importante em diversos setores, desde os mercados de energia elétrica onde o conhecimento dos padrões de consumo auxilia na contratação eficiente de energia, operação e manutenção da rede. A previsão da produção de energia também pode ser feita com base na análise dos padrões de consumo.

O processo de descoberta de conhecimento em bancos de dados não é um processo fácil e simples de se realizar. Uma metodologia deve ser adotada, de modo que sejam evitados erros na interpretação do problema, obtenção de padrões inexistentes ou incorretos. Foi apresentada uma metodologia geral de descoberta de conhecimento, foram analisados trabalhos que tinham como foco a caracterização de perfis típicos de consumo e as metodologias adotadas. Foi desenvolvida uma metodologia para corrigir os valores incorretos no banco de dados analisado, removendo ruídos, valores faltantes e dados inconsistentes. A normalização dos dados foi de fundamental importância para a comparação

dos padrões existentes entre as curvas de carga, pois o principal objetivo era comparar os padrões e não montantes de consumo.

Após o tratamento dos dados foram implementados oito algoritmos de agrupamento, três algoritmos particionais, quatro hierárquicos e um algoritmo difuso, porém por não apresentar resultados satisfatórios esse último foi removido da análise. O desempenho dos algoritmos foi avaliado de forma matemática e visual. De forma matemática através de seis índices de validação, que quantificaram a qualidade das partições feitas por cada algoritmo, e de forma visual através da análise das partições obtidas por cada algoritmo. Dentre os sete algoritmos analisados dois particionais e apenas um hierárquico apresentaram resultados satisfatórios na análise dos índices de validação e dos agrupamentos formados, sendo eles o algoritmo *K-means*, *K-medoids* e o algoritmo hierárquico *Ward link*. No geral o algoritmo *K-means* apresentou o melhor resultado, tanto para os índices de validação, quanto na análise visual das partições obtidas. Uma contribuição importante desse trabalho foi a utilização de um algoritmo de *clustering* “automático”, o algoritmo *G-means*, pois não haviam sido encontradas referências de sua utilização na identificação de padrões de consumo de energia elétrica. O algoritmo *G-means* apresentou boas partições na análise visual e valor satisfatório para os índices de validação, porém para a análise desenvolvida nesse trabalho, o algoritmo não forneceu grande informação, pois dividiu os dados em muitos grupos, sendo alguns deles semelhantes uns com os outros, inviabilizando a utilização do mesmo para essa análise. Também ficou evidenciado que os algoritmos hierárquicos, em geral, não conseguiram efetuar boa divisão de séries temporais, nesse caso as curvas de carga.

Através do algoritmo *K-means* obtiveram-se três padrões distintos de consumo. Um padrão atípico, onde a maior parte do consumo era durante o período da noite. O segundo padrão é muito similar a um consumo comercial, apresentando consumo residual de energia no período da madrugada, e no decorrer da manhã ocorre um aumento do consumo que se estabiliza por volta das dez horas da manhã e se mantém praticamente constante até o final da tarde, diminuindo gradativamente nas horas da noite. O terceiro padrão de consumo pode ser caracterizado como residencial, pois possui um consumo residual no período da madrugada, a partir das sete horas da manhã o consumo aumenta e mantém-se constante ao longo do dia e durante a noite apresenta um pico de consumo (horário em que maior parte dos consumidores se encontram em casa).



Por fim, para fechar o ciclo do processo da descoberta do conhecimento em bancos de dados, utilizou-se um modelo de classificação, para utilizar o conhecimento obtido na etapa de *clustering*. As curvas de carga de cada consumidor foram discretizadas em cinco índices que representam o formato de cada curva, esses índices foram passados para que o algoritmo pudesse fazer a classificação dos consumidores. Dois terços dos dados foram selecionados aleatoriamente para treinar o algoritmo, que gerou um conjunto de regras capazes de classificar novos consumidores de energia. Um terço dos dados foi classificado, com objetivo de verificar a precisão das regras geradas. O algoritmo obteve 93,75% de acerto na classificação das curvas de carga para os dias da semana e para os sábados e domingos/feriados obteve um acerto de 85,71% e 87,77%, respetivamente.

## 5.2. CONTRIBUTOS

Esse trabalho contribuiu para a formulação de toda uma metodologia baseada no processo de descoberta do conhecimento em banco de dados e técnicas de mineração de dados, com objetivo na caracterização de perfis típicos de consumo. Foram evidenciados algoritmos que apresentam melhores resultados na análise de *clustering* e também foi implementado um modelo de classificação, que obteve bons resultados na classificação de novos consumidores.

Outra contribuição desse trabalho foi o desenvolvimento dos algoritmos em uma plataforma *Open Source*, o que facilita a disseminação das técnicas abordadas nesse estudo, sendo que grande parte da literatura analisada utilizava plataformas pagas, o que limita, de certo modo, a disseminação do conhecimento.

## 5.3. TRABALHOS FUTUROS

Durante o desenvolvimento desse trabalho foram abertas outras possibilidades de estudo, ficando aberta a janela para:

- Melhorias nas etapas de pré-processamento dos dados, pois foram identificados apenas valores faltantes para o intervalo de uma hora, que poderiam ser desenvolvidas abordagens que conseguissem estimar intervalos maiores, uma vez que a qualidade dos dados reflete diretamente na análise de *clustering*;

- Análise para um período maior de aquisição dos dados, uma vez que foi possível efetuar apenas a análise para o intervalo de um mês. O ideal seria possuir dados de consumo de um ano, onde fosse possível identificar o efeito das variações climáticas nos padrões de consumo de energia elétrica;
- Melhorias no cálculo da distância entre as séries temporais, através da utilização métricas mais robustas na quantificação da distância entre séries temporais, como a *Dynamic Time Warping*.
- Análise e comparação entre algoritmos automáticos de *clustering*, sendo eles o *X-means* e o *G-means*, sendo que o segundo já foi implementado nesse trabalho;
- Análise de algoritmos de *clustering* baseados em redes neurais artificiais e algoritmos evolucionários.

## *Referências Documentais*

- [1] ERSE, “Entidade Reguladora de Serviços Energéticos,” [Online]. Available: <http://www.erse.pt/pt/electricidade/liberalizacaodosector/Paginas/default.aspx>. [Acedido em 24 Sep 2018].
- [2] S. F. C. Ramos, “Agregação e gestão inteligente de consumos de energia elétrica em mercados liberalizados,” Universidade de Lisboa - Instituto Superior Técnico: Tese de Doutoramento, Lisboa, 2015.
- [3] S. Ramos, Z. Vale, J. Santana e J. Duarte, “Data Mining Contributions to Characterize MV Consumers and to Improve the Suppliers-Consumers Settlements,” em *IEEE Power Engineering Society General Meeting*, Tampa, FL, 2007.
- [4] J. A. F. Herrera, “Uso de Data Warehousing e Data Mining na busca de relações e conhecimento em um ambiente de comércio eletrônico,” USP - Dissertação de Mestrado, São Carlos, 2003.
- [5] U. Fayyad, G. Piatetsky-Shapiro e P. Smyth, “From data mining to knowledge discovery in databases,” *AI Magazine*, vol. 17, nº 3, pp. 37-54, 1996.
- [6] J. Han, M. Kamber e J. Pei, *Data mining concepts and techniques*, Waltham, MA: Elsevier Inc, 2011.
- [7] A. Aki, K. M. R. D, K. R. Y, K. C. R. e S. T., “Analyzing the real time electricity data using data mining techniques,” *International Conference On Smart Technologies For Smart Nation (SmartTechCon)*, 17-19 Agosto 2017.
- [8] T. Zhang, G. Zhang, J. Lu, X. Feng e W. Yang, “A new index and classification approach for load pattern analysis of large electricity customers,” *IEEE Transactions on Power Systems*, vol. 27, nº 1, pp. 153-160, 2012.

- [9] A. I. Saleh, A. H. Rabie e K. . M. Abo-Al-Ez, “A data mining based load forecasting strategy for smart electrical grids,” *Advanced Engineering Informatics*, vol. 30, nº 3, pp. 422-448, 2016.
- [10] G. Fernandez, *Data Mining using SAS applications*, USA: Chapman & Hall, 2003.
- [11] S. d. C. Côrtes, R. M. Porcaro e S. Lifschitz, “Mineração de dados - funcionalidades, técnicas e abordagens,” PUC-Rio, Rio de Janeiro, 2002.
- [12] C. Antunes, *Data mining e data warehousing da exploração de dados à descoberta de informação*, Lisboa: Universidade de Lisboa, 2008.
- [13] R. T. Ng e J. Han, “Efficient and effective clustering methods for spatial data mining,” em *Proceedings of the 20th International Conference on Very Large Data Bases*, Santiago, 1994.
- [14] M. Piao e K. H. Ryu, “Local characterization-based load shape factor definition for electricity customer classification,” *IEEE Transactions on Electrical and Electronic Engineering*, pp. S110-S116, 4 May 2016.
- [15] S. Ramos e Z. Vale, “Data mining techniques application in power distribution utilities,” em *IEEE/PES Transmission and Distribution Conference and Exposition*, Chicago, 2008.
- [16] S. Ramos, V. Zita, J. Santana e J. Duarte, “Data mining contributions to characterize MV consumers and to improve the suppliers-consumers settlements,” em *IEEE Power Engineering Society General Meeting*, Tampa, FL, 2007.
- [17] S. Ramos, J. M. Duarte e Z. Vale, “A data-mining-based methodology to support MV electricity customers’ characterization,” *Energy and Buildings*, vol. 91, pp. 16-25, 2015.

- [18] V. Figueiredo, F. Rodrigues, Z. Vale e J. B. Gouveia, “An electric energy consumer characterization framework based on data mining techniques,” *IEEE Transactions on Power Systems*, vol. 20, nº 2, pp. 596 - 602, 2005.
- [19] G. Chicco, R. Napoli, F. Piglione, P. Postolache, M. Scutariu e C. Toader, “Emergent electricity customer classification,” *IEE Proceedings - Generation, Transmission and Distribution*, vol. 152, nº 2, pp. 164 - 172, 2005.
- [20] G. Chicco, R. Napoli, P. Postolache, M. Scutariu e C. Toader, “Customer characterization options for improving the tariff offer,” *IEEE Transactions on Power Systems*, vol. 18, nº 1, pp. 381 - 387, 2003.
- [21] M. Bicego, A. Farinelli, E. Grosso, D. Paolini e S. Ramchurn, “On the distinctiveness of the electricity load profile,” *Pattern Recognition*, vol. 74, pp. 317-325, 2018.
- [22] J. Yang, J. Zhao, F. Wen e Z. Dong, “A model of customizing electricity retail prices based on load profile clustering analysis,” *IEEE Transactions on Smart Grid*, pp. 1-1, 10 Abril 2018.
- [23] Y.-I. Kim, J.-M. Ko e S.-H. Choi, “Methods for generating TLPs (typical load profiles) for smart grid-based energy programs,” em *IEEE Symposium on Computational Intelligence Applications In Smart Grid (CIASG)*, Paris, 2011.
- [24] A. Capozzoli, M. S. Piscitelli e S. Brandi, “Mining typical load profiles in buildings to support energy management in the smart city context,” em *International Conference on Sustainability in Energy and Buildings*, Chania, 2017.
- [25] C. Fan, F. Xiao, Z. Li e J. Wang, “Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review,” *Energy and Buildings*, pp. 296-308, 2018.
- [26] M. Ashouri, F. Haghghat, B. C. Fung, A. Lazrak e H. Yoshino, “Development of building energy saving advisory: A data mining approach,” *Energy and Buildings*, pp. 139-151, 2018.

- [27] A. Rajabi, L. Li, J. Zhang, J. Zhu, S. Ghavidel e M. J. Ghadi, “A Review on clustering of residential electricity customers and its applications,” em *International Conference on Electrical Machines and Systems*, Sydney, 2017.
- [28] F. Biscarri, I. Monedero, A. García, J. I. Guerrero e C. León, “Electricity clustering framework for automatic classification of customer loads,” *Expert Systems With Applications*, vol. 86, pp. 54-63, 2017.
- [29] F. McLoughlin, A. Duffy e M. Conlon, “A clustering approach to domestic electricity load profile characterisation using smart metering data,” *Applied Energy*, vol. 141, pp. 190-199, 2015.
- [30] G. J. Tsekouras, N. D. Hatziargyriou e E. N. Dialynas, “Two-stage pattern recognition of load curves for classification of electricity customers,” *IEEE Transactions on Power Systems*, pp. 1120-1128, Aug 2007.
- [31] F. Lezama, A. Y. Rodríguez-González e E. M. de Cote, “Load pattern clustering using differential evolution with pareto tournament,” em *IEEE Congress on Evolutionary Computation (CEC)*, Vancouver, BC, 2016.
- [32] G. Chicco , O.-M. Ionel e R. Porumb, “Electrical load pattern grouping based on centroid model with ant colony clustering,” *IEEE Transactions on Power Systems*, vol. 28, n° 2, pp. 1706-1715, May 2013.
- [33] A. K. Jain e R. C. Dubes, *Algorithms for clustering data*, New Jersey: Prentice-Hall, Inc, 1988.
- [34] B. S. Everitt, S. Landau, M. Leese e D. Stahl, *Cluster analysis*, London: Wiley, 2011.
- [35] B. S. Everitt, *Cluster analysis*, New York: John Wiley & Sons, Inc., 1974.
- [36] P. Roelofsen, *Time series clustering*, Amsterdam: Vrije Universiteit, Master Thesis, Mar, 2018.

- [37] R. Linden, “Técnicas de agrupamento,” *Revista de Sistemas de Informação da FSMA*, nº 4, pp. 18-36, 2009.
- [38] C.-S. Perng, H. Wang, S. R. Zhang e D. S. Parker, “Landmarks: a new model for similarity-based pattern querying in time series databases,” *Proc. 2000 ICDE*, pp. 33-42, 2000.
- [39] D. J. Berndt e J. Clifford, “Using dynamic time warping to find patterns in time series,” AAAI, New York, 1994.
- [40] C. A. Ratanamahatana e E. Keogh, “Making time-series classification more accurate using learned constraints,” *SIAM*, pp. 11-22, 2004.
- [41] C. Cassisi, P. Montalto, M. Aliotta, A. Cannata e A. Pulvirenti, “Similarity measures and dimensionality reduction techniques for time series data mining,” 2012. [Online]. Available: <https://www.intechopen.com/books/advances-in-data-mining-knowledge-discovery-and-applications/similarity-measures-and-dimensionality-reduction-techniques-for-time-series-data-mining>.
- [42] R. Granell, C. J. Axon e D. C. H. Wallom, “Impacts of raw data temporal resolution using selected clustering methods on residential electricity load profiles,” *IEEE Transactions on Power Systems*, vol. 30, nº 6, pp. 3217-3224, 2015.
- [43] G. Chicco, “Overview and performance assessment of the clustering methods for electrical load pattern grouping,” *Energy*, vol. 42, pp. 68-80, 2012.
- [44] G. J. Tsekouras, N. D. Hatziargyriou e E. N. Dialynas, “Two-stage pattern recognition of load curves for classification of electricity customers,” *IEEE Transactions on Power Systems*, vol. 22, nº 3, 1120-1128 Aug 2007.
- [45] I. P. Panapakidis e G. C. Christoforidis, “Optimal selection of clustering algorithm via multi-criteria decision analysis (MCDA) for load profiling applications,” *Applied Sciences*, vol. 8, nº 2, pp. 1-42, 2018.

- [46] I. Dent, T. Craig, U. Aickelin e T. Rodden, “Variability of behaviour in electricity load profile clustering; who does things at the same time each day?,” *Industrial Conference on Data Mining*, pp. 70-84, 3 Sep 2014.
- [47] D. L. Davies e D. W. Bouldin, “A cluster separation measure,” *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 1, nº 2, 1979.
- [48] Y. Liu, Z. Li, H. Xiong, X. Gao e J. Wu, “Understanding of internal clustering validation measures,” em *IEEE International Conference on Data Mining*, Sydney, NSW, Australia, 2010.
- [49] J. C. Dunn, “Well separated clusters and optimal fuzzy partitions,” *J. Cybern*, vol. 4, pp. 95-104, 1974.
- [50] P. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” *Comput. Appl. Math.*, vol. 20, nº 1, pp. 54-65, 1987.
- [51] T. Calinski e J. Harabasz, “A dendrite method for cluster analysis,” *Comm. in Statistics*, vol. 3, nº 1, pp. 1-27, 1974.
- [52] P. Tan, M. Steinbach e V. Kumar, *Introduction to data mining*, Boston: Pearson, 2006.
- [53] R. Li, F. Li e N. D. Smith, “Multi-resolution load profile clustering for smart metering data,” *IEEE Transactions on Power Systems*, vol. 31, nº 6, pp. 4473-4482, 2016.
- [54] G. Hamerly e C. Elkan, “Learning the K in K-Means,” em *Neural Information Processing Systems (NIPS)*, 2003.
- [55] M. A. Stephnes, “EDF statistics for goodness of fit and some comparisons,” *American Statistical Association*, p. 69(347):730–737, 1974.
- [56] G. Chicco, R. Napoli e P. Federico, “Comparisons among clustering techniques for electricity customer classification,” *IEEE Transactions on Power Systems*, vol. 21, nº 2, pp. 933-940, 2006.



- [57] G. Chicco, “Overview and performance assessment of the clustering methods for electrical load pattern grouping,” *Energy*, vol. 42, pp. 68-80, 2012.
- [58] R. Granell, C. J. Axon e D. C. H. Wallom, “Impacts of raw data temporal resolution using selected clustering methods on residential electricity load profiles,” *IEEE Transactions on Power Systems*, vol. 30, n° 6, pp. 3217-3224, 2015.
- [59] S. Verdu, M. Garcia, C. Senabre, A. Marin e F. Franco, “Classification, filtering, and identification of electrical customer load patterns through the use of Self-Organizing Maps,” *IEEE Transactions on Power Systems*, vol. 21, n° 4, pp. 1672 - 1682, 2006.
- [60] G. Chicco, R. Napoli e F. Piglione, “Comparisons among clustering techniques for electricity customer classification,” *IEEE Transactions on Power Systems*, vol. 21, n° 2, pp. 933-940, 2006.
- [61] J. C. Bezdec, *Recognition with fuzzy objective function algorithms*, New York: Plenum, 1981.
- [62] G. M. d. F. Paula, “Curvas típicas de carga para o planejamento operacional do sistema de distribuição,” Tese de Mestrado, Universidade de São Paulo, São Paulo, 2006.
- [63] T. Bäck, D. B. Fogel e Z. Michalewicz, *Evolutionary computation 1: basic algorithms and operators*, Bristol and Philadelphia: Institute of physics publishing, 2000.
- [64] A. Garcia-Piquer, J. Bacardit, A. Fornells e E. Golobardes, “Scaling-up multiobjective evolutionary clustering algorithms using stratification,” *Pattern Recognition Letters*, vol. 93, pp. 69-77, 2017.
- [65] G. J. Tsekouras, P. B. Kotoulas e C. D. Tsirekis, “A pattern recognition methodology for evaluation of load profiles and typical days of large electricity customers,” *Electric Power Systems Research*, vol. 78, n° 9, p. 1494 –1510, 2009.
- [66] T. Räsänen, D. Voukantsis , H. Niska, K. Karatzas e M. Kolehmainen, “Data-based method for creating electricity use load profiles using large amount of customer-

specific hourly measured electricity use data,” *Applied Energy*, vol. 87, n° 11, pp. 3538-3545, 2010.

[67] P. S. Bradley, U. Fayyad e C. Reina, “Scaling clustering algorithms to large databases,” *AAAI*, 1998.

[68] D. D. Sharma e S. N. Singh, “Aberration detection in electricity consumption using clustering technique,” *International Journal of Energy Sector Management*, vol. 9, n° 4, pp. 451-470, 2015.

[69] J. R. Quinlan, *C4.5: programs for machine learning*, San Mateo, California: Morgan Kaufmann Publishers, 1993.

## Anexo A. Base de dados com doze consumidores

Neste anexo é mostrado um exemplo da base de dados com doze consumidores.

<b>Dia da semana</b>	<b>Horário</b>	<b>Consumidor 1</b>	<b>...</b>	<b>Consumidor 12</b>
Sábado	00:00	0,28002	...	1,03334
Sábado	01:00	0,39000	...	0,78334
Sábado	02:00	0,35801	...	0,63333
Sábado	03:00	0,27399	...	1,56667
Sábado	04:00	0,26601	...	1,81667
Sábado	05:00	0,26199	...	0,71667
Sábado	06:00	0,72149	...	0,63334
Sábado	07:00	0,91002	...	0,68334
Sábado	08:00	0,74288	...	4,00000
Sábado	09:00	0,37025	...	1,41667
Sábado	10:00	0,50147	...	4,01667
Sábado	11:00	0,40916	...	4,66667
Sábado	12:00	0,39246	...	2,96667
Sábado	13:00	0,45366	...	4,03333
Sábado	14:00	0,28524	...	9,16667
Sábado	15:00	0,55766	...	3,95000
Sábado	16:00	0,31400	...	4,63334
Sábado	17:00	0,51876	...	2,06667
Sábado	18:00	0,37509	...	1,91667
Sábado	19:00	1,30599	...	3,30000
Sábado	20:00	5,46002	...	6,30000
Sábado	21:00	2,19402	...	3,88334
Sábado	22:00	1,67400	...	5,61667
Sábado	23:00	0,84600	...	2,56667

## Anexo B. Base de dados com dez consumidores

Neste anexo é mostrado um exemplo da base de dados com dez consumidores.

<b>Dia da semana</b>	<b>Horário</b>	<b>Consumidor 1</b>	<b>...</b>	<b>Consumidor 10</b>
Sábado	00:00	0,35822	...	0,97586
Sábado	00:05	0,35606	...	1,49508
Sábado	00:10	0,36045	...	0,99136
Sábado	00:15	0,35921	...	3,17265
Sábado	00:20	0,35910	...	0,99433
Sábado	00:25	0,34501	...	1,11753
Sábado	00:30	0,35042	...	1,67140
Sábado	00:35	0,41806	...	1,03094
Sábado	00:40	0,47004	...	1,52710
Sábado	00:45	0,46663	...	1,07059
Sábado	00:50	0,45951	...	0,95488
Sábado	00:55	0,44085	...	0,94352
Sábado	01:00	0,33280	...	1,02286
Sábado	01:05	0,32894	...	0,98244
Sábado	01:10	0,12687	...	0,99686
Sábado	01:15	0,08727	...	1,53586
Sábado	01:20	0,08724	...	1,02079
Sábado	01:25	0,08767	...	1,46025
Sábado	01:30	0,08803	...	0,95320
Sábado	01:35	0,15097	...	1,51796
Sábado	01:40	0,18434	...	1,01827
Sábado	01:45	0,09161	...	0,99409
Sábado	01:50	0,17821	...	1,00518
Sábado	01:55	0,21473	...	0,99329

## Anexo C. Base de dados com cento e setenta e dois consumidores

Neste anexo é mostrado um exemplo da base de dados com cento e setenta e dois consumidores.

<b>Dia da semana</b>	<b>Horário</b>	<b>Consumidor 1</b>	<b>...</b>	<b>Consumidor 172</b>
Sábado	00:00	0,50933	...	0,86667
Sábado	00:15	0,34800	...	1,06667
Sábado	00:30	0,32667	...	1,06667
Sábado	00:45	0,40533	...	0,86667
Sábado	01:00	0,24667	...	1,26667
Sábado	01:15	0,34667	...	0,73333
Sábado	01:30	0,24533	...	1,26667
Sábado	01:45	0,23467	...	2,86667
Sábado	02:00	0,29867	...	3,93333
Sábado	02:15	0,20533	...	3,60000
Sábado	02:30	0,30800	...	3,60000
Sábado	02:45	0,24000	...	4,13333
Sábado	03:00	0,22533	...	3,86667
Sábado	03:15	0,29067	...	3,53333
Sábado	03:30	0,23200	...	3,60000
Sábado	03:45	0,25067	...	3,66667
Sábado	04:00	0,25467	...	3,60000
Sábado	04:15	0,24267	...	3,60000
Sábado	04:30	0,26400	...	3,80000
Sábado	04:45	0,23067	...	3,73333
Sábado	05:00	0,27200	...	6,40000
Sábado	05:15	0,23733	...	4,26667
Sábado	05:30	0,25600	...	3,73333
Sábado	05:45	0,24933	...	4,13333

## Anexo D. Dados tratados

Neste anexo é mostrado um exemplo dos dados pré-processados e tratados.

Consumidor	00:00	01:00	02:00	03:00	04:00	...	08:00	23:00
<b>1</b>	0.65796	0.18933	0.19037	0.18829	0.19037	...	0.15379	0.25107
<b>2</b>	0.08141	0.10571	0.08931	0.08991	0.05437	...	0.10692	0.10085
<b>3</b>	0.35284	0.24855	0.12187	0.15799	0.18449	...	0.09392	0.79914
<b>4</b>	0.43207	0.18009	0.14674	0.14427	0.14304	...	0.16477	0.58226
<b>5</b>	0.00846	0.01033	0.01505	0.01410	0.01410	...	0.00658	0.12125
<b>6</b>	0.26618	0.13473	0.09302	0.04803	0.04575	...	0.31067	0.67771
<b>7</b>	0.30499	0.08539	0.24096	0.10673	0.18433	...	0.11867	0.54160
<b>8</b>	0.32911	0.14030	0.13515	0.10342	0.17758	...	0.14349	0.36846
<b>9</b>	0.13745	0.08623	0.07466	0.12211	0.03820	...	0.30295	0.15191
<b>10</b>	0.31553	0.20870	0.23159	0.14123	0.07025	...	0.09205	0.36810
<b>11</b>	0.18085	0.20463	0.08159	0.18674	0.10493	...	0.11365	0.43783
<b>12</b>	0.00510	0.00285	0.00214	0.00332	0.00309	...	0.33310	0.00368
<b>13</b>	0.20877	0.01491	0.01667	0.02018	0.01842	...	0.67544	0.86711
<b>14</b>	0.15330	0.05262	0.01991	0.00455	0.00711	...	0.23663	0.23493
<b>15</b>	0.03903	0.03513	0.03104	0.02398	0.00725	...	0.05037	0.03030
<b>16</b>	0.98447	0.99482	0.99482	0.98965	0.98817	...	0.84911	0.97929
<b>17</b>	0.03312	0.11104	0.47857	0.50974	0.55844	...	0.40974	0.00649
<b>18</b>	0.97522	0.97522	0.98907	0.99198	0.97886	...	0.00000	0.98469
<b>19</b>	0.30119	0.28810	0.29286	0.19524	0.17738	...	0.62738	0.35595
<b>20</b>	0.98093	0.98759	0.99012	0.99426	0.99747	...	0.12224	0.96806
...	...	...	...	...	...	...	...	...
<b>194</b>	0.13752	0.16106	0.07101	0.08457	0.13244	...	0.09873	0.20048

## Anexo E. Exemplo de resultado para os índices de validação

Neste anexo é mostrado um exemplo dos resultados para os índices de validação, o melhor número de partições está circulado na figura, sendo que os índices MIA, CDI e DBI são índices em que o melhor resultado se dá pela minimização do valor e para os índices SI, DI e CHI o melhor resultado se dá pela maximização do valor.

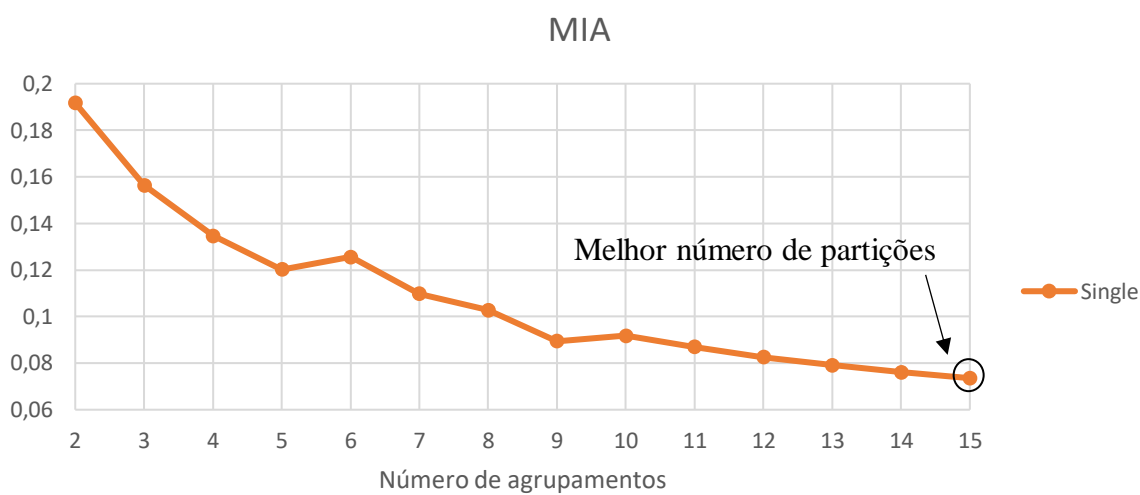


Figura E.1 – Exemplo de resultado para o algoritmo *Single Link* usando o índice MIA.

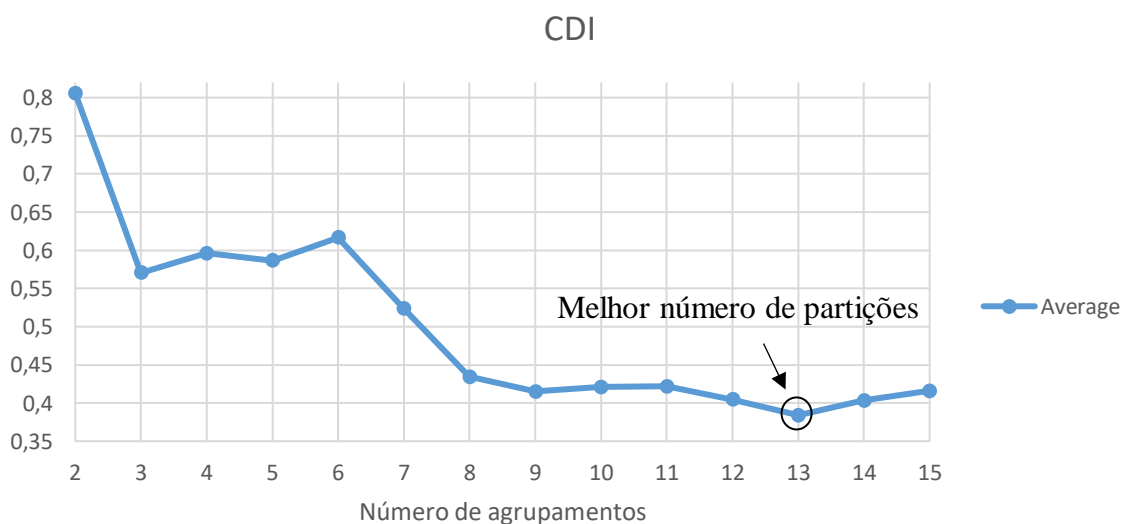


Figura E.2 – Exemplo de resultado para o algoritmo *Average Link* usando o índice CDI.

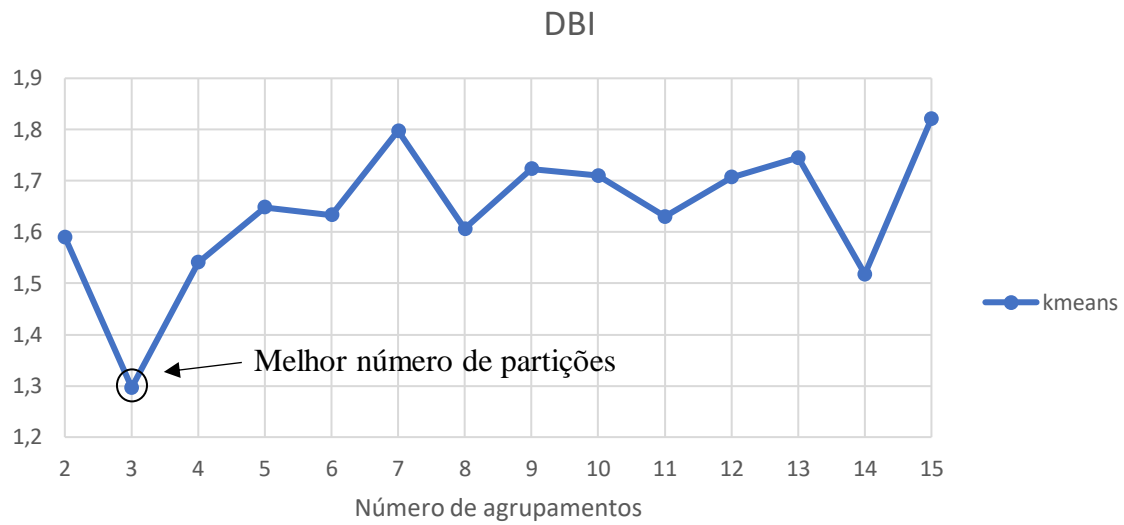


Figura E.3 – Exemplo de resultado para o algoritmo *K-means* usando o índice DBI.

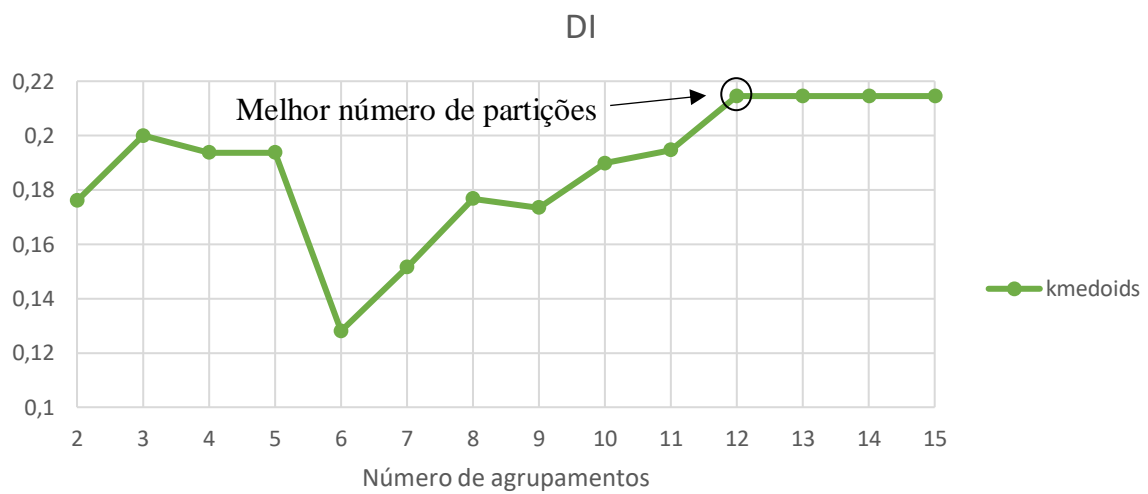


Figura E.4 – Exemplo de resultado para o algoritmo *K-medoids* usando o índice Dunn.



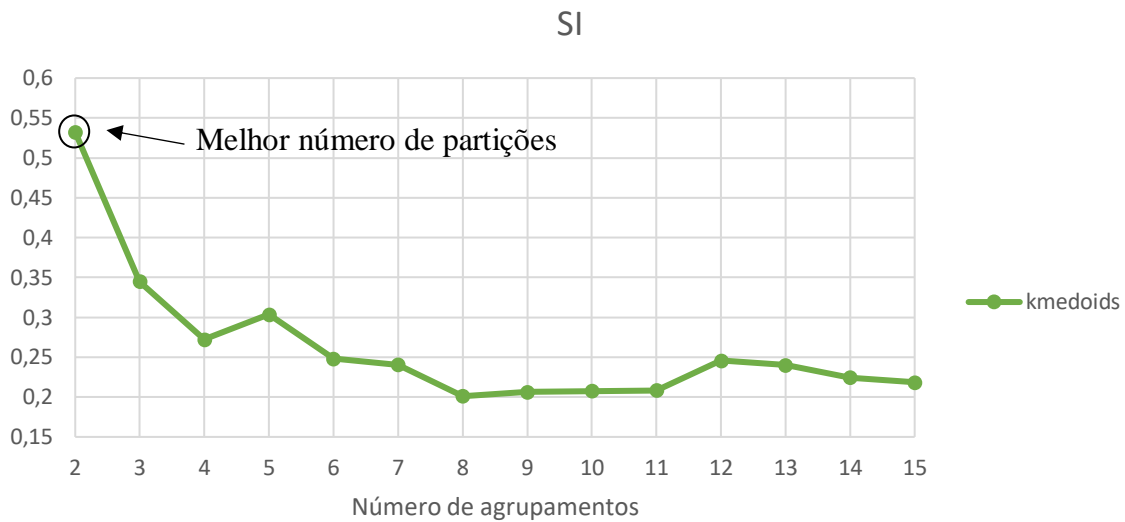


Figura E.5 – Exemplo de resultado para o algoritmo *K-medoids* usando o índice SI.

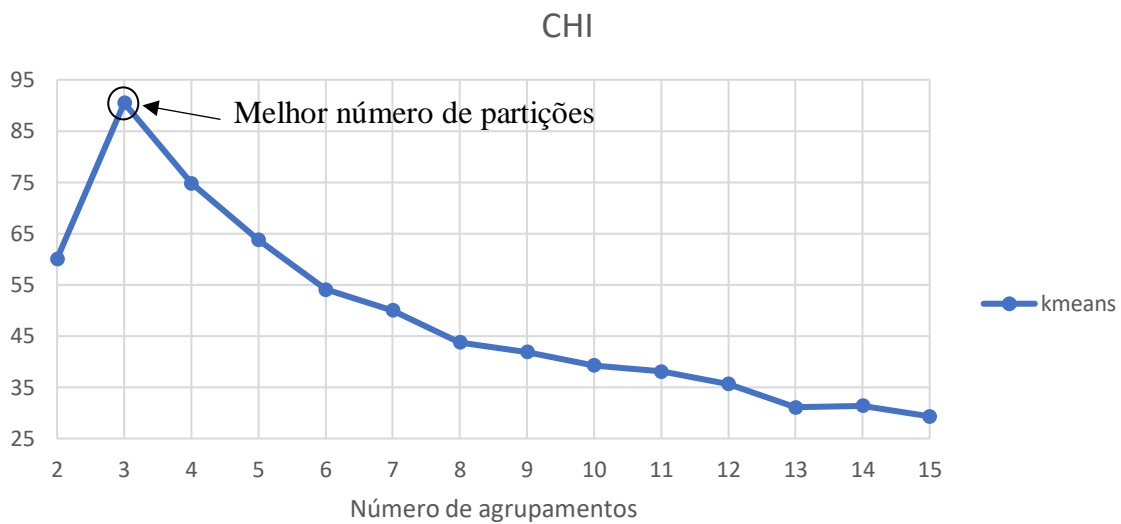


Figura E.6 – Exemplo de resultado para o algoritmo *K-means* usando o índice CHI.

## Anexo F. Resultado parcial dos índices de validação

Neste anexo é mostrado uma parte dos resultados para os índices de validação (de 2 a 9 *clusters*), onde pode ser verificado a sigla de cada algoritmo e a sigla do índice de validação. Alguns resultados não estão presentes, pois não foi possível calcular alguns índices para *clusters* com somente um dado, ou para o caso do algoritmo *G-means* que não conseguia dividir os dados em determinado número de partições.

	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>
<b>AL MIA</b>	0,2091	0,1825	0,1812	0,1756	0,1743	0,1479	0,1349	0,1287
<b>AL CDI</b>	0,8058	0,5707	0,5961	0,5865	0,6168	0,5237	0,4345	0,4151
<b>AL DBI</b>	0,7652	0,8164	1,1307	1,1343	1,1655	0,9728	0,9191	0,9215
<b>AL DI</b>	0,2673	0,2883	0,2883	0,2717	0,2873	0,2873	0,2873	0,2873
<b>AL SI</b>	0,5114	0,4550	0,3696	0,3367	0,2830	-	-	-
<b>AL CHI</b>	85,6475	49,1991	36,4918	48,3506	40,0332	33,7291	31,4540	28,3935
<b>SL MIA</b>	0,1917	0,1564	0,1347	0,1203	0,1256	0,1098	0,1028	0,0894
<b>SL CDI</b>	1,3669	0,8804	0,5389	0,4473	0,4676	0,4112	0,3922	0,3359
<b>SL DBI</b>	0,8858	0,8658	0,7910	0,7718	0,9740	0,8816	0,8905	0,8174
<b>SL DI</b>	0,2983	0,2807	0,2697	0,2668	0,2564	0,2520	0,2502	0,2462
<b>SL SI</b>	-	-	-	-	-	-	-	-
<b>SL CHI</b>	1,0594	1,1305	2,6607	2,4220	2,5753	2,2432	2,0492	1,8305
<b>CL MIA</b>	0,2266	0,1955	0,1905	0,1829	0,1787	0,1789	0,1722	0,1591
<b>CL CDI</b>	1,4409	0,6625	0,6992	0,6408	0,6635	0,6869	0,6639	0,5394
<b>CL DBI</b>	1,3310	1,1580	1,1963	1,2214	1,4574	1,5717	1,5437	1,4255
<b>CL DI</b>	0,1472	0,1938	0,2032	0,2112	0,2476	0,2324	0,2334	0,2453
<b>CL SI</b>	0,2825	0,3652	0,3090	0,2822	0,2363	0,1982	0,1981	0,2461
<b>CL CHI</b>	46,3731	81,4526	60,1324	49,6991	51,9207	45,6021	41,9288	39,8358
<b>WL MIA</b>	0,2091	0,1969	0,1906	0,1880	0,1800	0,1625	0,1606	0,1576
<b>WL CDI</b>	0,8058	0,7584	0,7301	0,7680	0,7125	0,5526	0,5669	0,5746
<b>WL DBI</b>	0,7652	1,4314	1,6538	1,5134	1,6507	1,4828	1,7360	1,6855
<b>WL DI</b>	0,2673	0,1582	0,1582	0,2044	0,2044	0,2044	0,1751	0,1751
<b>WL SI</b>	0,5114	0,3072	0,2593	0,2167	0,1949	0,2599	0,2169	0,1981
<b>WL CHI</b>	85,6475	78,3966	68,9218	59,4148	51,9803	46,9210	43,5912	40,5808

	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>
<b>KM MIA</b>	0,2139	0,1980	0,1743	0,1789	0,1770	0,1631	0,1702	0,1625
<b>KM CDI</b>	0,7631	0,7468	0,5981	0,6871	0,7250	0,6374	0,7114	0,6969
<b>KM DBI</b>	0,7221	1,2968	1,2527	1,6829	1,6719	1,6706	1,8299	1,7850
<b>KM DI</b>	0,3318	0,1996	0,1996	0,2060	0,1873	0,2184	0,1548	0,1922
<b>KM SI</b>	0,4935	0,4623	0,3257	0,2193	0,2014	0,2208	0,1450	0,1585
<b>KM CHI</b>	87,0288	90,4713	66,1742	64,6809	54,3523	47,7525	43,6524	42,3088
<b>KD MIA</b>	0,2068	0,1944	0,1862	0,1661	0,1633	0,1598	0,1582	0,1540
<b>KD CDI</b>	0,7781	0,6888	0,6938	0,5566	0,5702	0,5449	0,5630	0,5564
<b>KD DBI</b>	0,7350	1,2611	1,6101	1,4045	1,5639	1,6108	1,8773	1,8987
<b>KD DI</b>	0,1761	0,2000	0,1937	0,1937	0,1280	0,1516	0,1768	0,1735
<b>KD SI</b>	0,5321	0,3447	0,2722	0,3035	0,2482	0,2404	0,2010	0,2064
<b>KD CHI</b>	85,2121	88,4146	72,4718	60,4632	53,7861	50,4405	45,3623	42,8695
<b>GM MIA</b>	0,2143	-	0,1730	-	-	-	0,1501	-
<b>GM CDI</b>	0,8412	-	0,5922	-	-	-	0,5215	-
<b>GM DBI</b>	0,8026	-	1,2126	-	-	-	1,4149	-
<b>GM DI</b>	0,3226	-	0,2000	-	-	-	0,2069	-
<b>GM SI</b>	0,4928	-	0,3457	-	-	-	0,2728	-
<b>GM CHI</b>	86,8631	-	65,8126	-	-	-	46,6647	-

# Anexo G. Árvore de decisão

Neste anexo é mostrada a árvore de decisão gerada na etapa de classificação.

