UNIVERSIDADE FEDERAL DO PIAUÍ
CENTRO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

JUAN DE AGUIAR GONÇALVES

ALGORITMO HÍBRIDO PARA FILTRAGEM E AGRUPAMENTO DE CURVAS DE CARGA

TERESINA
2018

JUAN DE AGUIAR GONÇALVES

ALGORITMO HÍBRIDO PARA FILTRAGEM E AGRUPAMENTO DE CURVAS DE CARGA

Trabalho de Dissertação submetido à banca examinadora designada pela Universidade Federal do Piauí, apresentado como parte dos requisitos para obtenção de grau de Mestre em Engenharia Elétrica.

Orientador: Prof. Dsc. Hermes Manoel Galvão Castelo Branco

Coorientador: Prof. Dsc. Aldir Silva Sousa

TERESINA

2018

FICHA CATALOGRÁFICA
Universidade Federal do Piauí
Biblioteca Comunitária Jornalista Carlos Castello Branco
Serviço de Processamento Técnico

G635a Gonçalves, Juan de Aguiar.
Algoritmo híbrido para filtragem e agrupamento de curvas de carga. /Juan de Aguiar Gonçalves. - 2018.
112 f.: il.

Dissertação (Mestrado) – Universidade Federal do Piauí, Mestrado em Engenharia Elétrica, Teresina, 2018.
"Orientação: Prof. Dsc. Hermes Manoel Galvão Castelo Branco".
"Coorientador: Prof. Dsc. Aldir Silva Sousa".

1. Engenharia Elétrica - Curva de carga.
2. Transformada Wavelet. I. Título.

CDD: 621.3

JUAN DE AGUIAR GONÇALVES

ALGORITMO HÍBRIDO PARA PRÉ-PROCESSAMENTO E AGRUPAMENTO DE CURVAS DE CARGA

Esta Dissertação foi julgada adequada para a obtenção do título de Mestre em Engenharia Elétrica, Área de Concentração Sistemas de Energia Elétrica, e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Piauí.

Juan de Aguiar Gonçalves
Pós-Graduando (PPGEE-UFPI)

Banca Examinadora:

Prof. Dr. Hermes Manoel Galvão Castelo Branco
Orientador (PPGEE-UFPI).

Prof. Dr. Aldir Silva Sousa
Co-Orientador (PPGEE-UFPI).

Prof. Dr. Luís Gustavo Mota Souza
Avaliador Interno (PPGEE-UFPI)

Prof. Dr. Guilherme de Alencar Barreto
Avaliador Interno (PPGEE-UFPI)

Prof. Dr. Mário Oleskovicz
Avaliador Externo (EESC-USP)

Teresina, 12 de dezembro de 2018.

A Deus,
Minha esposa Lígia e minha filha Talita,
Aos meus pais, João e Cristina,
Aos irmãos, Hugo, Daniel e Rafael,
Aos meus avós, paternos e maternos,
A todos os familiares e amigos.

AGRADECIMENTOS

Agradeço a Deus por sua inspiração, perseverança e infinita bondade em todos os momentos.

Minha amada esposa Lígia e minha filha Talita por compreenderem minha ausência, me ajudarem e apoiarem nessa jornada.

Meus pais João e Maria Cristina por sempre estarem ao meu lado.

Aos meus irmãos Hugo, Daniel e Rafael, pela motivação.

Ao professor Dsc. Hermes Manoel Galvão Castelo Branco, pela inestimável e competente orientação, sem a qual não seria possível galgar minha vida acadêmica, desde a graduação até hoje.

Ao professor Dsc. Aldir Silva Sousa, que com sua ajuda e intelecto, proporcionou as soluções de difíceis problemas encontrados nessa pesquisa.

Meus docentes, por sempre estarem disponíveis para me ensinar e ajudar.

Aos colegas do programa de pós-graduação em engenharia elétrica da UFPI, pela valiosa amizade.

À UFPI, pela oportunidade ímpar de aprendizado e cidadania promovidos por esse programa de pós-graduação.

À CHESF, que autorizou a coleta dos dados necessários para desenvolvimento desse trabalho.

A todos que direta ou indiretamente contribuíram para a realização dessa pesquisa.

“*O cientista não se destina a resultados imediatos. Ele não espera que suas ideias sejam prontamente aceitas. Seu dever é preparar o terreno para aqueles que virão, e apontar o caminho”*.

(Nikola Tesla)

RESUMO

Esta pesquisa propõe um algoritmo híbrido para filtragem e agrupamento de curvas de carga. Utilizou-se dados reais de um transformador de extra-alta tensão instalado em uma subestação, suscetível ao clima e às peculiaridades do sistema elétrico de potência. O primeiro passo do algoritmo é filtrar os dados através de um processo iterativo utilizando filtro *hampel* para corrigir as curvas de carga e procurar defeitos remanescentes usando o sinal do banco de filtros com Transformada *Wavelet* Discreta (TWD). No segundo estágio, associado ao agrupamento, um processo iterativo realizou a redução da dimensionalidade das curvas de carga, através de um sinal aproximado do banco de filtros TWD, nos seus diversos níveis, seguido do agrupamento dessas curvas, aplicando o algoritmo *k-means*. O resultado da filtragem apresentou a correção otimizada das curvas de carga, além do número de curvas excluídas indicarem anormalidades no sistema de medição. A estratégia de redução de dimensionalidade das curvas de carga, utilizada na metodologia proposta, resultou em melhor eficácia de agrupamento, quando comparada ao agrupamento realizado sem redução de dimensionalidade, assim como, em relação a redução das curvas de carga, através da Análise de Componentes Principais. Os grupos resultantes representaram curvas de carga com tipologias bem definidas, associadas aos dias da semana, classe de carga, meses do ano e estações climáticas, mesmo que o agrupamento tenha sido realizado sem supervisão, ou qualquer informação prévia fornecida ao algoritmo.

Palavras-Chave: Curva de Carga. Transformada *Wavelet*. Filtro de *Hampel*. Redução de Dimensionalidade. Agrupamento. *K-Means*. Sistema Elétrico de Potência.

ABSTRACT

This research proposes a hybrid algorithm for filtering and clustering of load curves. We used real data from an extra-high voltage transformer installed in a substation, susceptible to the weather and peculiarities of the electrical system. The first step of the algorithm is to filter data through an iterative process using hampel filter to correct the load curves and search for remaining defects using the signal from the filter bank with Discrete Wavelet Transform (DWT). In the second stage, associated to the clustering, an iterative process performed the reduction of dimesionality of the load curves, through approach signal of the TWD filter bank, at its various levels, followed by the grouping of these curves, applying the k-means algorithm. The result of the filtration presented the optimized correction of the load curves, in addition to that the number of curves excluded indicated abnormalities in the measurement system. The strategy of dimensionality reduction of the load curves, used in the proposed methodology, resulted in better clustering efficiency, when compared to the clustering realized without reduction of dimensionality, as well as, in relation to the reduction of the load curves, through the Principal Component Analysis. The resulting clusters represent load curves with well-defined typologies, associated with the days of the week, load class, months of the year and climatic seasons, even though the clustering was performed without supervision, or any previous information provided to the algorithm.

Keywords: Charge Curve. Wavelet Transform. Hampel Filter. Dimensionality Reduction. Clustering. K-Means. Power System.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

ANEEL	Agência Nacional de Energia Elétrica
BEA	Subestação Boa Esperança
CAST	*Cluster Affinity Searching Technique*
CDI	*Clustering Dispersion Indicator*
CHESF	Companhia Hidrelétrica do São Francisco
CNO	Subestação Coelho Neto
CPTEC	Centro de Previsões do Tempo e Estudos Climáticos
DTW	*Dynamic Time Warping*
ED	*Euclidean Distance*
EDA	Análises Exploratórias de Dados
EEG	Eletroencefalograma
FCM	*Fuzzy C-Means*
HSCM	*Hyperbolic Smoothing Clustering Method*
INPE	Instituto Nacional de Pesquisas Espaciais
LOESS	*Locally Weighted Regression and Smoothing Scatterplots*
MFL	*Modified Follow the Leader*
MIA	*Mean Index Adequacy*
MS	*MultiShapes*
MSE	*Multi Shapes Extraction*
ONS	Operador Nacional do Sistema
PCA	*Principal Component Analysis*
PRI	Subestação Piripiri
SAVL	Sistema de Alarme Violação de Limites
SAX	*Symbolic Aggregate AproxXimation*
SBD	*Shape-Based Distance*
SE	*Shape Extraction*
SEP	Sistema Elétrico de Potência
SIN	Sistema Interligado Nacional
SSE	Soma do Erros Quadrados
SVD	*Singular Value Decomposition*

SWC	*Silhouette Width Coefficient*
TFD	Transformada de Fourier Discreta
TSA	Subestação Teresina
TSD	Subestação Teresina Dois
TW	Transformada *Wavelet*
TWC	Transformada *Wavelet* Contínua
TWD	Transformada *Wavelet* Discreta
WFA	*Weighted Fuzzy Average*

SUMÁRIO

1. INTRODUÇÃO

O agrupamento de dados é um dos principais problemas na mineração de dados, o qual objetiva determinar um conjunto finito de categorias que particionem dados de acordo com as similaridades entre os seus objetos. A solução de um problema de agrupamento também pode ajudar a resolver outros problemas relacionados, como classificação de padrões e extração de regras (HORTA, 2013).

Agrupar séries temporais é uma solução importante para vários problemas em vários campos de pesquisa, incluindo negócios, ciência médica, finanças e engenharia (HORTA, 2013). Uma série temporal, $F_i = \{f_1 \dots, f_t \dots, f_n\}$, é definida como um conjunto ordenado de números que indicam as características temporais dos objetos a qualquer momento t de uma observação (MORRIS, TRIVEDI, 2009).

As curvas de carga são séries temporais que representam a demanda de energia elétrica ao longo do dia no contexto dos sistemas elétricos de potência. Infelizmente, percebem-se falhas nesses dados que podem ocorrer: no processo de medição, aquisição de dados ou anormalidades oriundas de sinistros nos Sistemas Elétricos de Potência (SEP). Essas falhas se manifestam na forma de anormalidades instantâneas, anormalidades temporárias, lacunas ou dados repetidos (DE OLIVEIRA, 2013).

Contudo, dados com falhas não são apropriados no agrupamento de curvas de carga, já que esse se baseia em padrões, sendo necessário a filtragem das falhas presentes no banco de dados utilizado (DE OLIVEIRA, 2013; LIN et al., 2004).

Dadas curvas de carga livre de observações deficientes, com a análise de agrupamentos, procura-se identificar perfis típicos ou tipologias da carga. Por outro lado, o comportamento da carga sofre influência de diversos fatores. Dentre outros, citam-se os principais:

Fatores temporais: periodicidade diária-semanal, variações sazonais, estações do ano e ocorrência de feriados.

Fatores meteorológicos: temperatura, luminosidade, precipitação, velocidade e direção do vento.

Fatores aleatórios: greve de grande repercussão, transmissão de um programa de televisão de interesse geral, entrada e ou saída de grandes consumidores e etc.

Fatores determinísticos: redução deliberada da tensão, implementação de tarifas horosazonais, apelo ao público e etc.

Fatores econômicos: horário de verão, racionamentos, planos econômicos e etc. (ALMEIDA, 2013).

Consequentemente, surgem na base de dados diversas tipologias oriundas do comportamento da carga. Portanto, é fundamental que o conjunto de dados seja separado em grupos, definindo os diversos padrões de carga.

Essas tipologias, apresentam grande aplicabilidade como, por exemplo, subsidiar o planejamento do sistema elétrico, extrair características para análise preditiva de equipamentos elétricos de potência, realizar o pré-processamento de dados em métodos de previsão de carga, dentre outros.

Por outro lado, séries temporais são padrões de alta dimensionalidade (RANI e SIKKA, 2012; KEOGH e KASETTY, 2003). Sendo assim, é importante aplicar estratégias no intuito de diminuir tais dimensões e com isso garantir melhores resultados para a utilização de algoritmos clássicos de agrupamento (AGHABOZORGI et al., 2014). Uma das soluções consiste na conversão dos dados da série temporal num vetor de características, com dimensão inferior ao vetor original. Esse recurso é denominado abordagem baseada em redução de dimensionalidade (WANG, SMITH e HYNDMAN, 2006; LIAO, 2005).

No entanto, a redução de dimensionalidade pode subtrair informações importantes para a representação das tipologias existentes no banco de dados estudado, originando agrupamentos de séries temporais inadequados (LAI, CHUNG, TSENG, 2010).

Por fim, a análise de agrupamentos deve ser combinada com uma estratégia de validação adequada, supervisionada ou não supervisionada, para verificar a qualidade do agrupamento, através de índices que mensuram cada metodologia de validação (HAN, PEI e KAMBER, 2012).

Esta pesquisa propõe uma metodologia de pré-processamento de curvas de carga dividida em duas etapas, que são a filtragem e o agrupamento das curvas de carga.

A primeira etapa, diz respeito ao processo de filtragem, constituindo um processo iterativo, utilizando filtro de hampel, para correção das curvas de carga e o sinal do detalhe oriundo do banco de filtros da transformada *wavelet* discreta (TWD), aplicado na pesquisa de defeitos remanescentes.

A segunda etapa, refere-se ao processo de agrupamento. O agrupamento foi realizado através de um processo iterativo, utilizando a redução de dimensionalidade de curvas de carga. A redução de dimensionalidade decorreu do sinal de aproximação do banco de filtros TWD, nos seus diversos níveis, seguido do agrupamento das curvas reduzidas utilizando o algoritmo

k-means. Incluso ao processo iterativo de agrupamento, um procedimento de validação não supervisionada, indicou o melhor agrupamento.

O banco de dados utilizado na pesquisa foi aquisitado de um transformador de potência de extra alta tensão instalado na cidade de Teresina no estado do Piauí, Brasil. A coleta dos dados de carregamento fora realizada para dados de corrente elétrica RMS (*Root Mean Square*), das curvas de carga diárias entre 00:00h e 23:55h, no intervalo de cinco em cinco minutos, correspondendo a 288 amostras para cada curva de carga, durante o período de julho de 2010 a julho de 2017. O intervalo de coleta totalizou um banco de dados de 2588 curvas de carga.

Esse transformador supre cargas de concessionárias de distribuição dos estados do Piauí e Maranhão e está localizado em um ponto do SEP susceptível a diversos intempéries e peculiaridades, adequado para a verificação da robustez da metodologia proposta.

O processo iterativo de filtragem apresentou resultados satisfatórios para a correção e limpeza de falhas das curvas de carga, assim como, seus resultados também serviram como indicativo para verificação de problemas associados aos equipamentos envolvidos, desde a medição até a aquisição dos dados.

Um dos aspectos determinantes do agrupamento da metodologia proposta, está na redução de dimensionalidade das curvas de carga. Sendo assim, verificou-se o resultado dessa metodologia quanto ao agrupamento sem redução de dimensionalidade, assim como, em relação à redução da dimensão das curvas de carga, através da Análise de Componentes Principais (*Principal Component Analysis* – PCA). Verificou-se que a metodologia proposta resulta na melhor eficácia de agrupamento, em relação à esses procedimentos verificados.

Apesar da técnica de validação da metodologia proposta não dispor de referências de agrupamento, já que se trata de validação não supervisionada, essa realizou agrupamentos distintos e bem definidos, com características que representam desde o tipo de carga, dia e meses de consumo predominantes, fornecendo valiosas informações a respeito do banco de dados.

1.1 OBJETIVOS

O objetivo geral desse trabalho compreende o desenvolvimento de uma metodologia para análise de curvas de carga, englobando as etapas de filtragem e agrupamento de dados.

Todavia, os objetivos específicos consistem em:

- Propor uma estratégia de tratamento de falhas (lacunas e continuidades[1]) ou *outliers*[2] em curvas de carga utilizando a Transformada *Wavelet* (TW) e filtro de *hampel*;
- Investigar o desempenho da TW para extrair características e reduzir a dimensionalidade das curvas de carga utilizadas na etapa de agrupamento;
- Agrupar as diversas tipologias das curvas de carga do objeto de estudo, em função dos cenários energéticos do sistema elétrico onde esse se encontra instalado, buscando similaridades entre os comportamentos dinâmicos da mesma carga nos diversos dias da semana e meses do ano, utilizando o algoritmo *k-means*;
- Investigar as possíveis particularidades na curva de carga do sistema elétrico estudado, no intuito de realizar a aplicação mais adequada das técnicas de filtragem de dados;
- Combinar os resultados finais de agrupamento da metodologia de pré-processamento proposta, com características de perfis de carga esperadas do sistema elétrico estudado.
- Verificar a eficácia da metodologia proposta, quanto à redução de dimensionalidade de curvas de carga em relação a dados normais, sem recurso de redução de dimensionalidade e em relação a outra técnica de redução de dimensionalidade consolidada, por exemplo, a PCA.

1.2 CONTRIBUIÇÕES

Contribuições científicas:

- Desenvolvimento de uma metodologia com novas abordagens para agrupamento de dados utilizados em análise de sistemas elétricos.
- Abordagem comparativa da potencialidade de agrupamento de curvas de carga, resultante da redução de dimensionalidade da metodologia proposta, em relação a outras ferramentas de extração de características para agrupamento.

[1] Continuidade são falhas na curva de carga, que se manifestam na forma de dados inalterados por um período de tempo considerado incomum.

[2] O termo outlier será utilizado para tratar de observações atípicas, por ser prática consagrada na literatura especializada.

Contribuição Técnica:

- Elaboração de uma metodologia para tratamento e agrupamento de curvas de carga, que poderá ser aplicada na análise de tipologias de carga de sistemas elétricos, podendo ser utilizada, junto a equipes de operação, manutenção, estudos de sistemas eletroenergéticos, dentre outros.
- Extração de características da carga a que um equipamento está submetido, viabilizando a tomada de decisões associadas à disponibilidade do mesmo num cenário de desligamento intempestivo ou programado.

1.3 ESTRUTURA DO TRABALHO

Este trabalho está dividido em cinco capítulos. No capítulo dois é apresentado o estado da arte, no qual são abordados: o problema de filtragem e agrupamento de séries temporais, a descrição dos conceitos associados a curvas de carga e suas classes típicas e um breve relato a respeito de diversos trabalhos relacionados ao tema.

No capítulo três é apresentada a metodologia proposta para filtragem e agrupamento do banco de dados do objeto de estudo, assim como seus procedimentos de verificação de eficácia de agrupamento.

No capítulo quatro são apresentados e discutidos os resultados obtidos por meio dos processos iterativos desenvolvidos para a filtragem e agrupamento das curvas de carga.

Por fim, no capítulo cinco apresentam-se as conclusões obtidas com a aplicação do método desenvolvido e as possíveis melhorias a serem realizadas em trabalhos futuros.

2. ESTADO DA ARTE

Nesse capítulo serão tratados os conceitos e aspectos teóricos, associados às ferramentas e técnicas utilizadas para o desenvolvimento da metodologia proposta.

No que diz respeito a curvas de carga, serão tratados seus conceitos, intervalos de discretização e perfis típicos.

Quanto à filtragem, serão apresentados os conceitos associados à TW e suas especificidades, assim como filtro de *hampel* e seu funcionamento. Também serão ilustrados os conceitos dos tipos de falhas de curvas de carga.

Relativo aos fundamentos para a realização de agrupamento, serão discutidos seus conceitos, redução de dimensionalidade de curvas de carga, especificidades da TW aplicada na redução de dimensionalidade de curvas de carga, algoritmo *k-means* e seus parâmetros, concluindo com conceitos e técnicas de validação de agrupamento de dados.

Também se realizou uma breve revisão bibliográfica de trabalhos associados ao tema e ferramentas, utilizados nessa pesquisa.

2.1 CURVAS DE CARGA

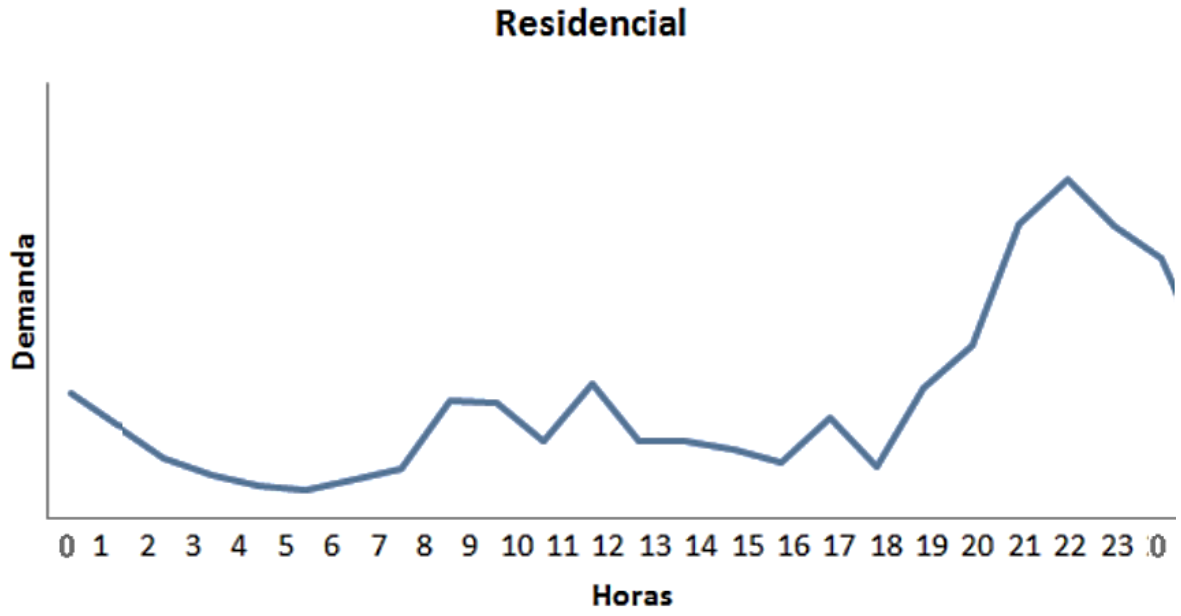
As curvas de carga demonstram o consumo de carga elétrica ao longo do tempo. Estas curvas podem ser discretizadas em pequenos intervalos, em ordem de minutos e também em grandes intervalos, em ordem de horas (DE OLIVEIRA, 2013). O período de discretização da carga pode ser dividido segundo quatro horizontes de tempo (ALMEIDA, 2013):

- Curtíssimo prazo: utilizada normalmente na operação do SEP em tempo real, para o estudo do estado dinâmico do sistema, e outras tomadas de decisões urgentes ou emergenciais;
- Curto prazo: aplicada à operação on-line ou off-line do SEP, normalmente em situações intempestivas ou programadas, e para estudar perdas de ativos nos sistemas de transmissão;
- Médio prazo: compreende um intervalo de tempo de semanas a meses. Esse horizonte de tempo pode ser considerado de extrema importância aos setores de geração, transmissão e distribuição, devido à necessidade de planejamento da operação, a fim de resguardar a continuidade da operação normal do sistema;

- Longo prazo: compreende um horizonte que varia de um a alguns anos. É indispensável aos setores de planejamento da expansão e investimentos, no que diz respeito à garantia do funcionamento seguro do SIN, bem como para o planejamento da operação.

Dependendo do tipo de consumidor as curvas de carga expressam comportamentos distintos, dentre os quais podemos citar: residenciais, comerciais, industriais ou iluminação pública (KAGAN, OLIVEIRA e ROBBA, 2005), conforme verificado na Figura 2.1, 2.2, 2.3. 2.4 e 2.5.

Figura 2.1 – Curva típica de uma carga residencial.

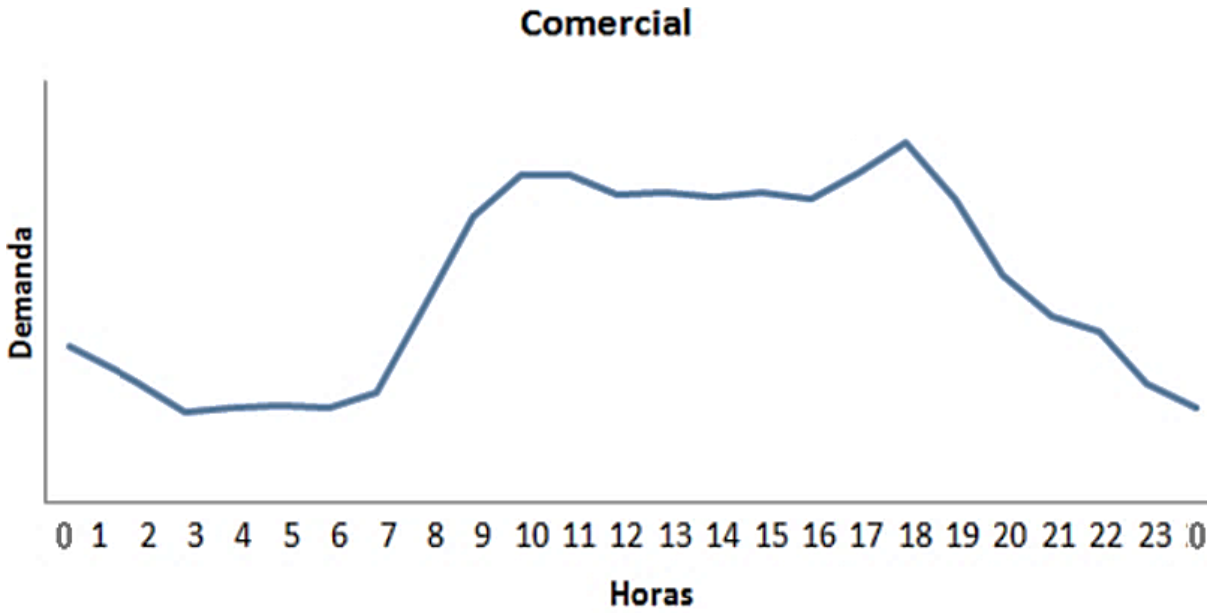


Fonte: Pessanha et.al. (2015).

Conforme verifica-se na Figura 2.1, a curva de carga residencial possui características bem definidas. Constata-se a redução gradativa da carga das 00:00h às 06:00h, seguida de uma rampa de baixa amplitude das 07:00h às 08:00h, correspondendo ao horário que ao início das atividades diárias. Ocorre outra rampa das 11:00h às 12:00, referente ao horário do almoço. Por fim, por volta das 18:00h verifica-se o aumento íngreme da carga, associada ao horário de carga máxima, concluindo com sua redução por volta das 22:00h.

Na Figura 2.2, constata-se que a carga tem valores reduzidos entre 00:00h e 07:00h. A partir de então ocorre o aumento de carga, devido o início das atividades no comércio, chegando ao seu pico de consumo por volta das 10:00h, com uma ligeira redução, em função do horário de almoço, mantendo-se praticamente constante até o segundo pico de carga às 18:00h, próximo ao final do expediente.

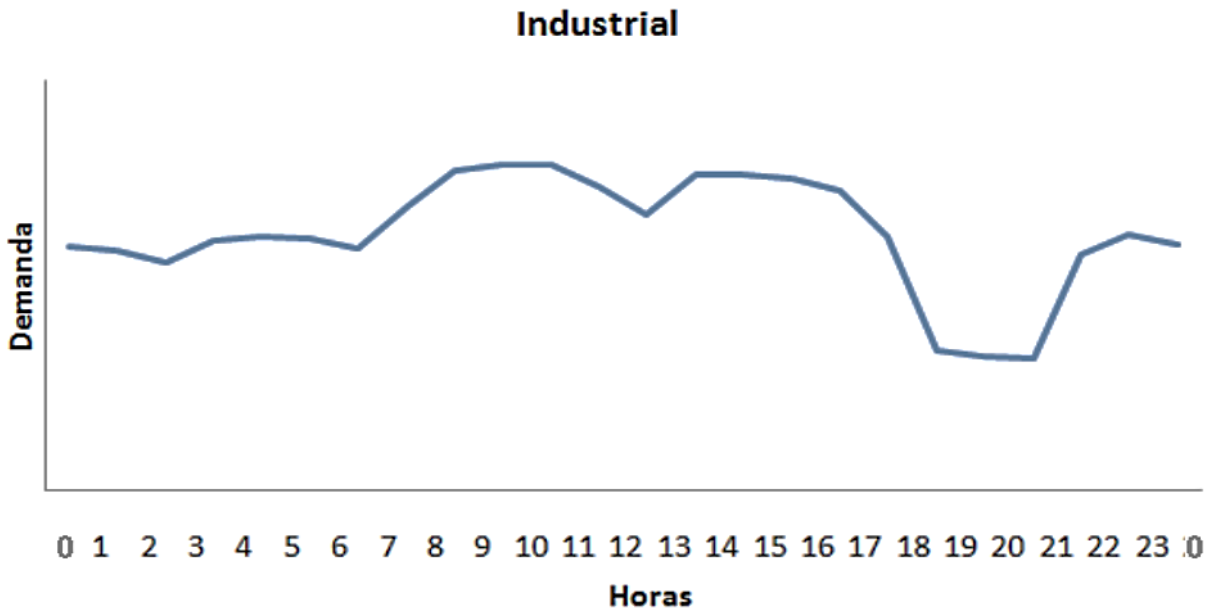
Figura 2.2 – Curva típica de uma carga comercial.



Fonte: Pessanha et.al. (2015).

A Figura 2.3 ilustra o perfil típico de uma carga predominantemente industrial. Essa curva caracteriza-se pela predominância de elevadas amplitudes de carga. No período das 00:00h às 07:00h, ocorre uma redução do valor de carga. Entretanto, a diminuição da demanda é mais predominante durante o horário de carga máxima do SIN, de 18:00h às 22h, devido às multas e restrições impostas pelas concessionárias e agente reguladores, no que se refere a esse período de consumo.

Figura 2.3 – Curva típica de uma carga industrial.



Fonte: Pessanha et.al. (2015).

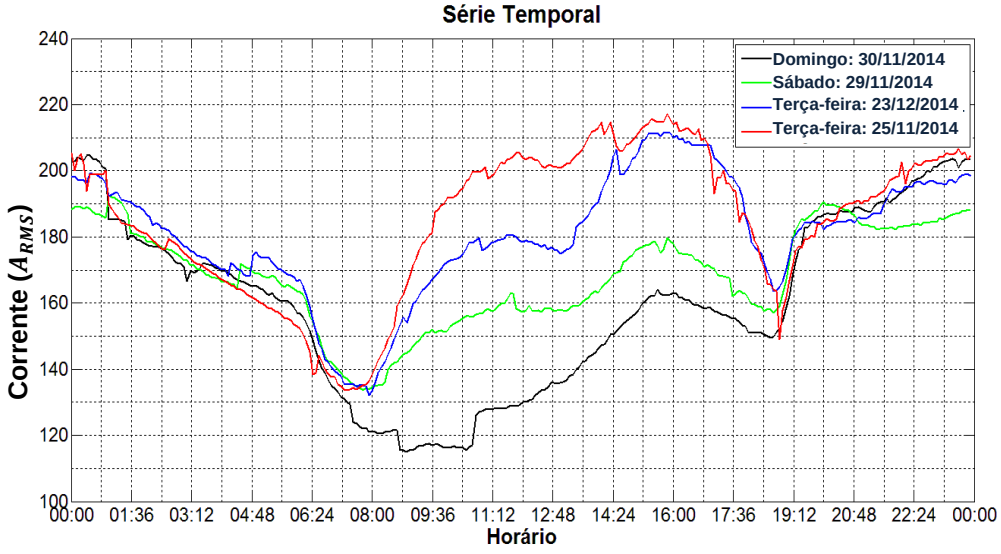
O comportamento típico da carga de iluminação pública é manter-se praticamente constante de 00:00h às 05:30h. Por volta das 07:00h ocorre o afundamento da carga a zero. Isto porque neste horário toda a iluminação pública costuma estar desligada. Uma nova rampa se inicia por volta das 17:00h e mantém o crescimento até às 19:00h, quando a iluminação pública encontrar-se-á em pleno funcionamento, Figura 2.4.

Figura 2.4 – Curva típica de iluminação pública.

Fonte: Pessanha et.al. (2015).

Conforme o período observado, as séries temporais também possuem comportamentos distintos de acordo com os dias da semana ou meses do ano, conforme pode ser exemplificado na Figura 2.5.

Figura 2.5 – Série temporal de carga em diversos dias da semana.



Fonte: Elaborada pelo autor.

Sendo assim, as formas das curvas de carga expressam tipologias predominantes ou uma combinação de tipologias.

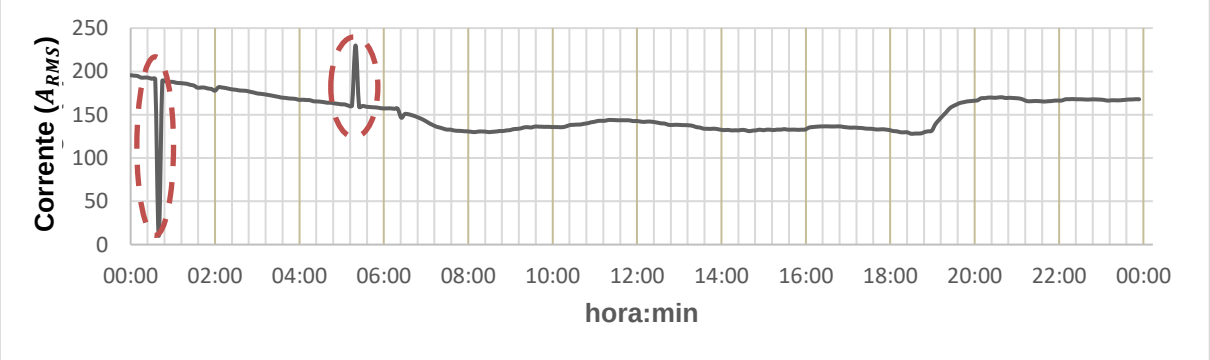
2.2 FILTRAGEM DE DADOS

A filtragem de dados consiste na identificação e correção de falhas, quando viável (DE OLIVEIRA, 2013).

As falhas identificáveis são classificadas da seguinte forma: anormalidades instantâneas, anormalidades temporárias, lacunas, e dados congelados.

As anormalidades instantâneas podem ocorrer devido à sinais indevidos, registrados pelo sistema de medição, ou oscilações instantâneas no SEP, ilustradas na Figura 2.6.

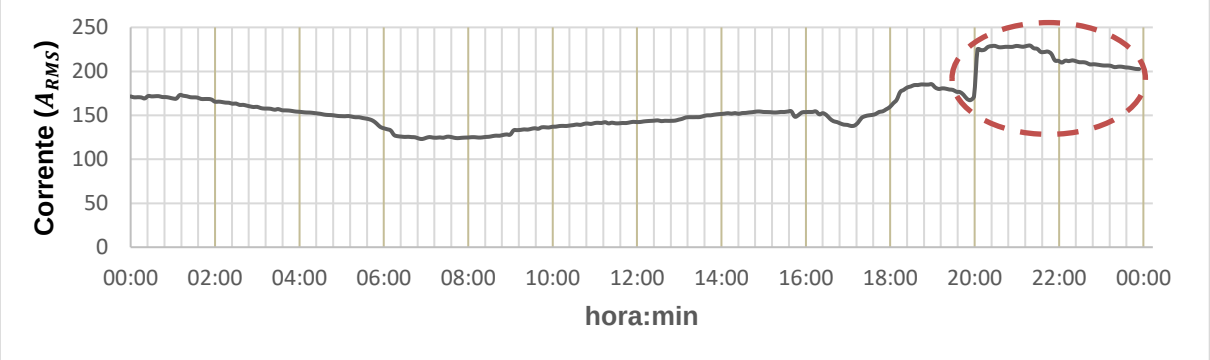
Figura 2.6 – Anormalidades instantâneas.



Fonte: Elaborada pelo autor.

Anormalidades temporárias acontecem quando há falhas no sistema elétrico, desligamento de equipamentos, retirada ou entrada de grandes blocos de carga, assim como inserção ou retirada de equipamentos de regulação de grande porte, devido circunstâncias emergenciais. Essas anormalidades são ilustradas com os deslocamentos acentuados da amplitude da curva de carga, conforme verificado na Figura 2.7.

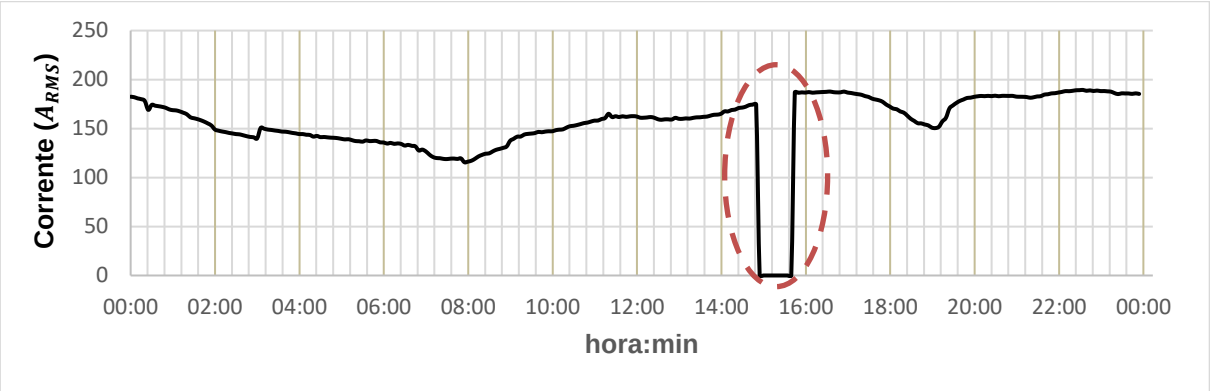
Figura 2.7 – Anormalidade temporária.



Fonte: Elaborada pelo autor.

Podem ocorrer desligamentos intempestivos ou programados nos equipamentos do SEP. Os equipamentos de medição registram esses eventos como lacunas, que por sua vez são substituídas pelo valor zero, de acordo com a Figura 2.8.

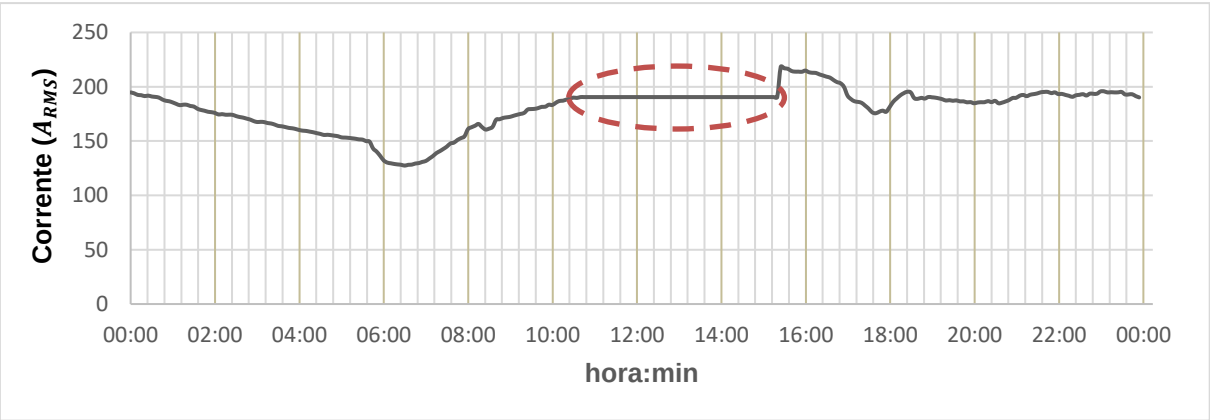
Figura 2.8 – Lacuna na curva de carga.



Fonte: Elaborada pelo autor.

Na Figura 2.9 verificam-se dados repetidos durante um intervalo de tempo atípico. Esses ocorrem na maioria das vezes, devido à falha no sistema de medição, quando novos dados não são coletados, então o sistema de medição repete o valor da última informação coletada.

Figura 2.9 – Dados congelados da grandeza medida.



Fonte: Elaborada pelo autor.

A presença das anormalidades de medição contribui para a especificação errônea do comportamento da carga, comprometendo consequentemente a acurácia do agrupamento realizado (DE OLIVEIRA, 2013). Desta forma, para reduzir os efeitos das falhas do sistema de medição e demais causas não naturais, a elaboração de qualquer metodologia de agrupamento deve ser precedida pelo tratamento dos dados históricos da carga.

Dentre as diversas ferramentas utilizadas na filtragem de dados podemos citar a TW e o filtro de *hampel* (HAMPEL, 1974).

Sendo assim, é importante que as técnicas utilizadas para tratamento de dados sejam fundamentadas em detecção e correção de dados com falhas.

2.2.1 Transformada *Wavelet*

Desenvolvida inicialmente por HAAR (1910), a *wavelet* de *Haar* permaneceu por muito tempo como a única base ortonormal de *wavelets*.

As *wavelets* fornecem a representação dos níveis de detalhe na análise de funções, constituindo uma ferramenta matemática que decompõe funções hierarquicamente, em duas componentes: uma é a aproximação grosseira e a outra representa os detalhes da função analisada. São usadas em análise funcional, em singularidades ou oscilações locais de funções, em solução de equações diferenciais, em reconhecimento de padrões, em compressão de dados, biologia, medicina, astronomia, engenharia, entre outros (ADDISON, 2002).

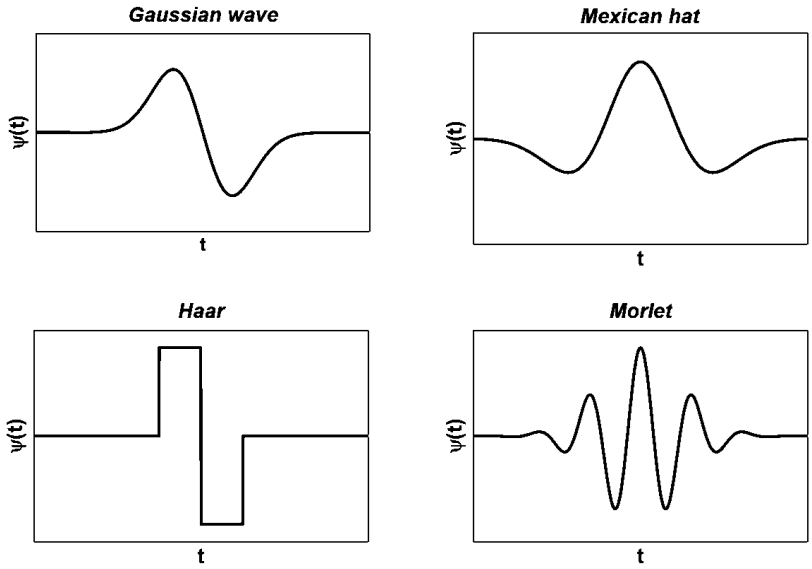
A utilização da Transformada de Fourier para analisar funções que mudem sua frequência com o tempo se torna bastante complexa, pois apesar das funções utilizadas serem no domínio da frequência, essas são aplicadas em todos os instantes. Desta forma, a análise de Fourier deve ser utilizada como ferramenta de análise de funções em que suas frequências não evoluam com o tempo, ou seja, é indicada para a análise de funções periódicas (CHUI, 1992).

Por outro lado, a TW tem por base a utilização de funções que representam pequenas formas de onda (*wavelets*), que são utilizadas para transformar um sinal em análise, através de sua convolução[3] com este sinal. Essas funções são manipuladas de duas formas, movendo-as sobre o sinal em estudo, dilatando-as ou comprimindo-as (ADDISON, 2002). Dessa forma, podem-se utilizar intervalos de tempo maiores quando se deseja informações de baixa frequência e intervalos de tempo menores quando necessitar de informações de alta frequência, ou seja, a função analisada pode ser aperiódica.

A análise *wavelet* utiliza uma função base denominada *wavelet* mãe. Na Figura 2.10 são mostrados exemplos de funções *wavelets* mãe: *Gaussian wave, Mexican hat*, *Haar* e *Morlet*.

[3] Convolução é um operador linear, que a partir de duas funções dadas resulta numa terceira, que por sua vez, mede a soma do produto dessas funções ao longo de uma região subentendida pela superposição dessas em função do deslocamento existente entre elas (CHUI, 1992).

Figura 2.10 – Formas básicas de função *wavelet.*



Adaptada de: Addison (2002).

As funções base *wavelet*s $g(t)$ devem satisfazer certas condições matemáticas para que possam originar uma família de *wavelets* (ADDISON, 2002):

1. Ser integrável:

$$\int_{-\infty}^{\infty} g(t)dt < \infty \tag{2.1}$$

2. Devem ter energia finita, a ser preservada pela análise:

$$E = \int_{-\infty}^{\infty} |g(t)|^2 dt < \infty \tag{2.2}$$

3. A condição de admissibilidade deve ser respeitada:

$$\int_{0}^{\infty} \frac{|G(f)|^2}{f}\, df < \infty \tag{2.3}$$

Onde (f) é a Transformada de *Fourier* de $g(t)$.

2.2.1.1 Transformada *Wavelet* Contínua

A Transformada *Wavelet* Contínua (TWC) de uma função $x(t)$, com relação à *wavelet* mãe $g(t)$ é dada pela Equação (2.4) (DE OLIVEIRA, 2013):

$$TWC(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) g\left(\frac{t-b}{a}\right) dt \tag{2.4}$$

Em que a é o fator escala e b é o fator de translação.

A TWC é obtida dilatando-se e transladando-se a *wavelet* mãe continuamente (ADDISON, 2002 e DE OLIVEIRA, 2013).

Sendo assim, em uma dada escala e fator de translação, o coeficiente da $TWC(a,b)$ representará como a função $x(t)$ e a *wavelet* mãe, dilatada/transladada, se assemelham. Logo, todos os coeficientes da $TWC(a,b)$ associados à função $x(t)$ em análise, correspondem a representação *wavelet* do sinal com respeito a *wavelet* mãe $g(t)$.

Por exemplo, considerando o fator de escala a de determinada *wavelet*. Com o aumento de a (encolhimento do fator de escala no tempo), as oscilações tornam-se mais rápidas e a *wavelet* exibe frequências elevadas. No entanto, quando a diminui resulta na expansão do fator de escala no tempo, logo as oscilações tornam-se mais lentas e a *wavelet* exibe baixas frequências. Isso torna a *wavelet* uma importante ferramenta na detecção de descontinuidades e *outliers*, pois com a TW componentes de alta frequência são analisadas em intervalos de tempo pequenos e componentes de baixa frequência são analisadas em intervalos de tempo longos. Essa característica da TW permite que a análise do comportamento de uma função possa ser realizada independente de suas oscilações intempestivas.

2.2.1.2 Transformada *Wavelet* Discreta

Conforme mencionado anteriormente, a função *wavelet* foi definida em termos dos parâmetros de escala e translação. Porém, na prática, os valores destes parâmetros são discretizados e esta forma é conhecida como TWD e definida como conforme a Equação (2.5) (DE OLIVEIRA, 2013).

$$TWD(m,k) = \frac{1}{\sqrt{{a_0}^m}} \sum_n x(n) g\left(\frac{k - n b_0 {a_0}^m}{{a_0}^m}\right) \tag{2.5}$$

Os parâmetros m e n são inteiros que controlam a dilatação e translação, respectivamente. A constante a_0 é um passo fixo de dilatação, e seu valor geralmente é maior que 1. Enquanto que b_0 é o parâmetro de localização, e seu valor deve ser maior do que 0. Por fim, k é inteiro e também um parâmetro inicial.

A TW é uma técnica que vem sendo bastante utilizada em SEP, pois com esta ferramenta é possível precisamente detectar o início de uma descontinuidade no sinal, quando existente, referenciando-o ao domínio do tempo, além de possibilitar uma boa extração das características do SEP estudado (JEMSE e HARBO, 2001).

2.2.1.3 Banco de Filtros da TWD Aplicados na Detecção de Falhas de Séries Temporais

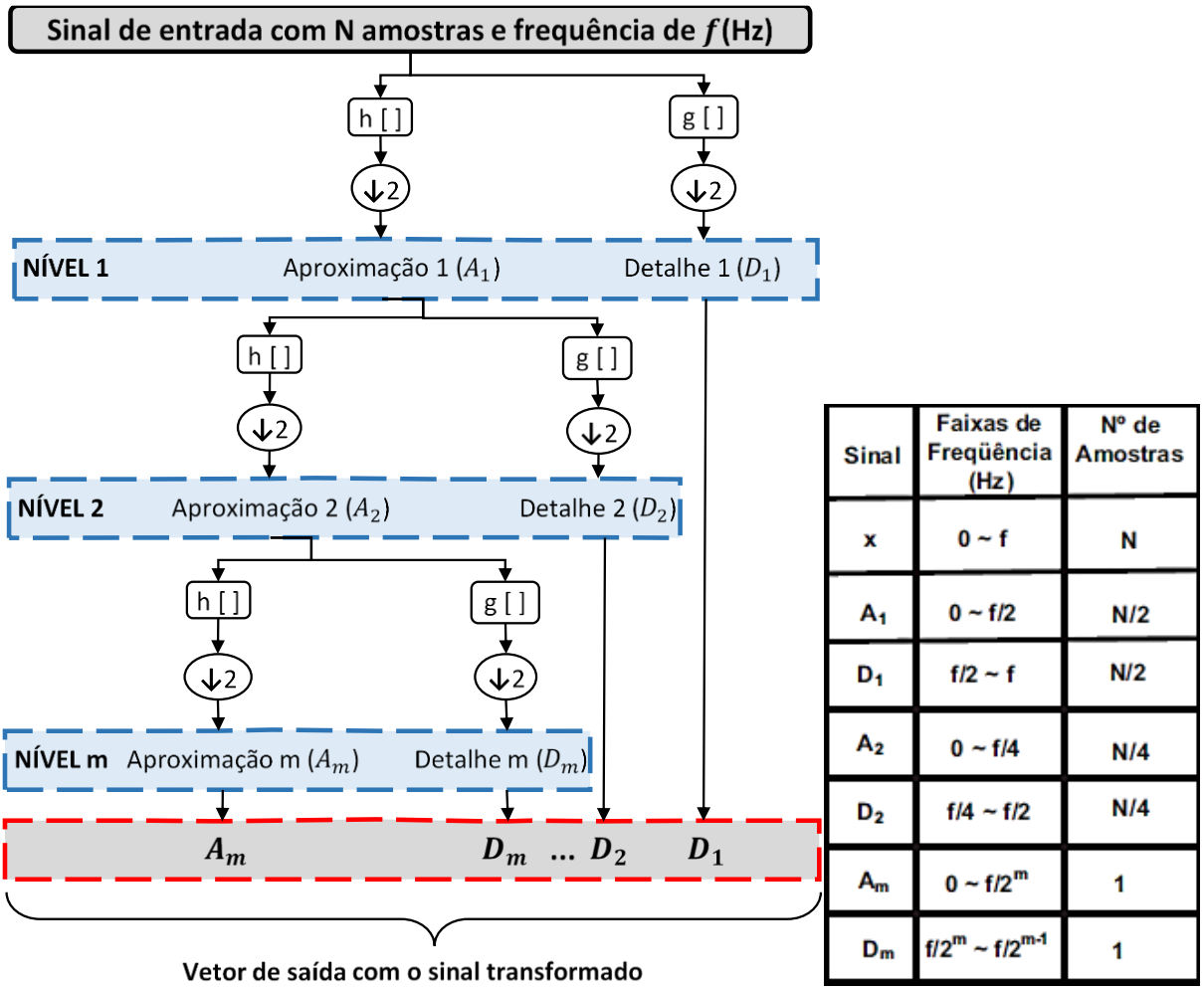
A aplicação da TW em SEP pode ser feita por meio de uma implementação de banco de filtros, uma vez que do ponto de vista prático, a TWD é um processo de filtragem digital no domínio do tempo, via convolução discreta, acompanhada do operador *downsampling* (redução) por 2 (JEMSE, HARBO, 2001).

A Figura 2.11 ilustra vários níveis de decomposição de um sinal pelo emprego da TWD. Nessa figura, um sinal é filtrado por um filtro passa-baixa (h) e por um filtro passa-alta (g), e então é aplicado o operador *downsampling* na saída de cada filtro. A saída do filtro passa-baixa representa a aproximação do sinal. Já a resposta do filtro passa-alta fornece o detalhe do sinal original. Este processo pode ser sucessivamente aplicado às aproximações para a obtenção da decomposição em diferentes níveis, sendo que em cada nível a faixa de frequência é dividida ao meio, após a passagem do sinal pelos filtros, conforme ilustrado no resultado da decomposição de curva de carga da Figura 2.12. O sinal resultante da filtragem é dado pela concatenação da aproximação com os detalhes.

A TWD de um sinal depende do filtro passa-baixa e do filtro passa-alta utilizados. Sendo assim, um parâmetro chave para trabalhar com a TW é a escolha da *wavelet* mãe que será utilizada. A literatura aponta, que a família *Daubechies* geralmente é uma boa escolha para a maioria das situações que um SEP é submetido, pois com esta, geralmente, são melhor

identificados fenômenos com decaimentos e oscilações rápidas, além de transitórios, características típicas destes eventos (BARAN e KIM, 2006). Verifica-se também que *wavelets* com poucos coeficientes são ideais para localizar fenômenos no tempo. Por outro lado, fenômenos com variações mais lentas, como afundamentos e elevações de tensão, podem ser melhor identificados por *wavelets* com maior quantidade de coeficientes.

Figura 2.11 – Banco de filtros da TWD.

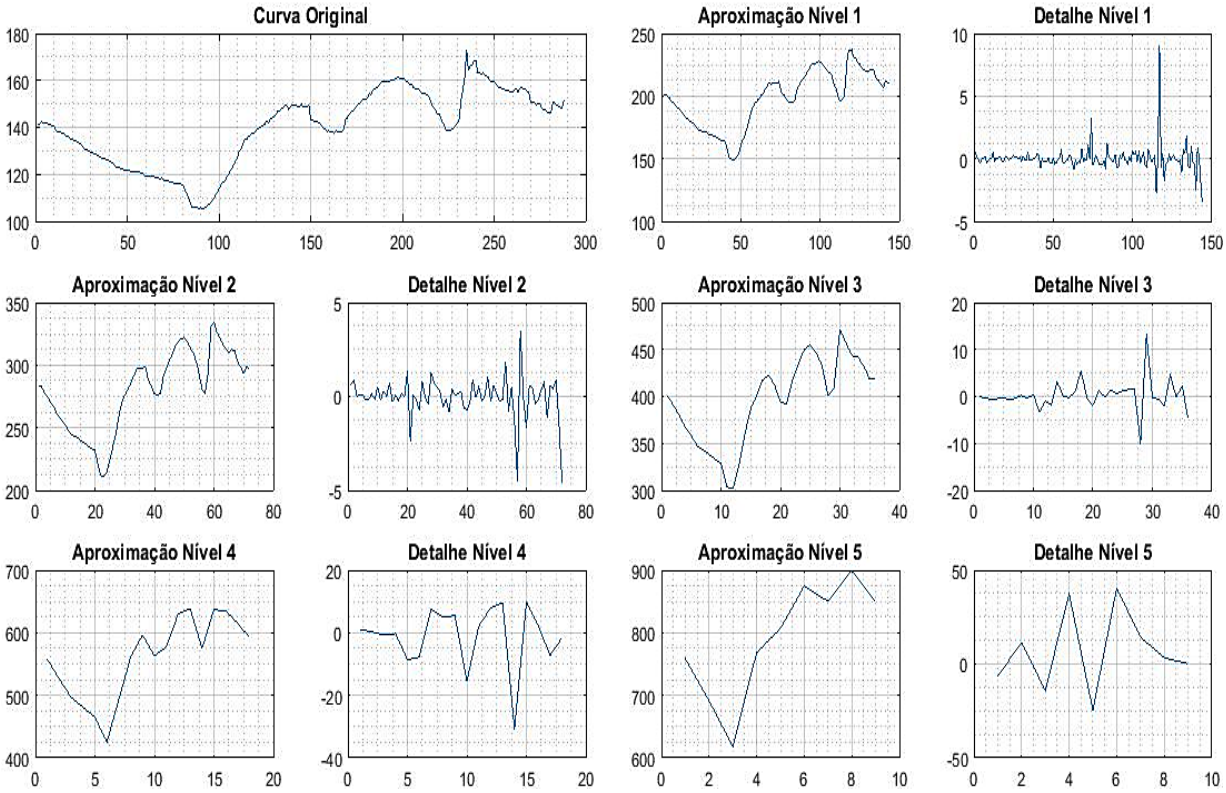


Sinal	Faixas de Freqüência (Hz)	Nº de Amostras
x	0 ~ f	N
A_1	0 ~ f/2	N/2
D_1	f/2 ~ f	N/2
A_2	0 ~ f/4	N/4
D_2	f/4 ~ f/2	N/4
A_m	0 ~ $f/2^m$	1
D_m	$f/2^m$ ~ $f/2^{m-1}$	1

Fonte: Elaborada pelo autor.

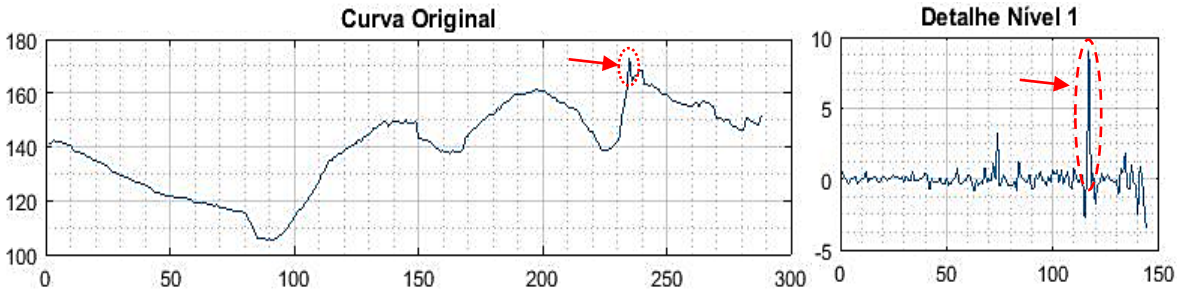
Sabendo que o sinal oriundo do filtro passa-alta (g) da Figura 2.11 fornece o detalhe do sinal estudado e que o operador *downsampling* reduz as amostras originais por 2, de acordo com a Figura 2.12. Logo, pode-se constatar por volta de qual momento está ocorrendo a alteração do sinal, conforme indicado na Figura 2.13.

Figura 2.12 – Decomposição de uma curva de carga pelo banco de filtros da TWD.



Fonte: Elaborada pelo autor.

Figura 2.13 – Curva original e detalhe nível 1 de curva de carga com *outlier*.



Fonte: Elaborada pelo autor.

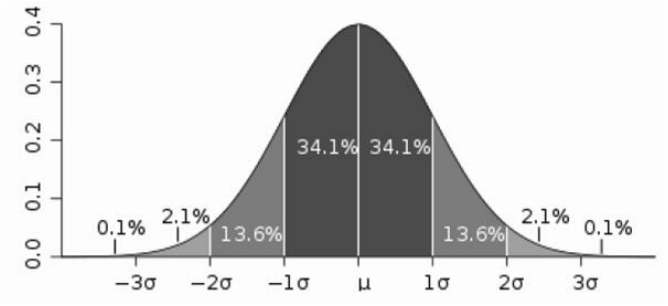
Sendo assim, torna-se evidente o detalhe do sinal da TWD como uma boa ferramenta para detecção de falhas em séries temporais.

2.2.2 Filtro de *Hampel*

O filtro de *hampel* detecta e remove *outliers* do sinal de entrada usando o identificador *hampel*, que é uma variação da regra de estatística dos três sigmas (HAMPEL, 1974). Essa regra

afirma que aproximadamente todas as amostras de uma observação numa distribuição normal, estão no intervalo de três desvios padrão da media, ou seja, aproximadamente 99,73% dos valores encontram-se numa distribuição normal (KAZMIER, 2004), conforme Figura 2.14.

Figura 2.14 – Regra dos três sigmas.

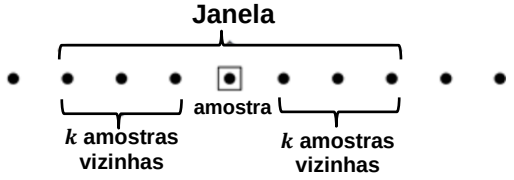


Adaptada de: Kazmier (2004).

O identificador de *hampel* é considerado como um dos identificadores mais robustos para detecção de *outliers* (KIWARE, 2009).

No filtro de *hampel* é calculada a mediana de uma janela especificada, para cada amostra do sinal de entrada, composta por *k* amostras vizinhas, *k-hampel*, em cada lado da amostra atual, conforme a Figura 2.15. Contudo, o filtro também estima o desvio padrão de cada amostra em relação à mediana de sua janela, através da mediana do desvio absoluto. Se uma amostra difere da mediana em mais do que três desvios padrão, ela é substituída pela mediana (HAMPEL, 1974).

Figura 2.15 – Janelamento filtro de *hampel*.



Fonte: Elaborada pelo autor.

2.3 AGRUPAMENTO DE SÉRIES TEMPORAIS

Agrupamento de dados é um problema conceitual fundamental na mineração de dados, o qual objetiva determinar um conjunto finito de categorias que descrevam os dados de acordo com as similaridades entre os seus objetos. A solução de um problema de agrupamento pode ajudar a resolver problemas de classificação de padrões e extração de regras (HORTA, 2013).

O agrupamento de séries temporais é uma solução importante para vários problemas em diversos campos de pesquisa, incluindo negócios, ciência médica, finanças e engenharia, sendo também uma etapa de pré-processamento. O agrupamento é considerado o mais importante problema de aprendizado supervisionado e não supervisionado. Agrupar séries temporais é particularmente vantajoso na análise exploratória de dados e extração de características de um conjunto de dados, para diversas finalidades, dentre elas a previsão da série temporal (AGHABOZORGI et al., 2014).

Resumidamente, os métodos de agrupamento de séries temporais têm por finalidade a classificação de séries temporais em grupos, de tal forma que séries temporais semelhantes sejam classificados no mesmo grupo, enquanto séries temporais de tipologias distintas sejam classificados em grupos diferentes.

Sedo assim, dado um conjunto de dados de N objetos, $D = \{F_1, F_2, \dots F_N\}$, onde F_i é uma série temporal. O processo de particionamento não supervisionado de D em $C = \{C_1, C_2, \dots C_k\}$, ocorre de tal forma, que os dados de séries temporais homogêneos são agrupados com base na similaridade de sua forma num agrupamento C_i, onde $D = \cup_{i=1}^{k}$ e $C_i \cap C_j = \emptyset$, para $i \neq j$ (AGHABOZORGI et al., 2014; LIAO, 2005).

A maioria dos algoritmos de agrupamento de dados convencionais não funcionam bem para séries temporais, em grande parte devido sua alta dimensionalidade e presença de descontinuidades e *outliers*, fazendo do agrupamento de séries temporais um grande desafio (VLACHOS, LIN e KEOGH, 2003).

2.3.1 Redução de dimensionalidade

Um dos problemas do agrupamento de grandes conjuntos de dados diz respeito a quantidade de objetos que fazem parte desse conjunto. Uma das formas de reduzir o número de objetos do banco de dados é a utilização de amostragem (KEOGH e KASETTY, 2003). No caso de séries temporais, o objetivo é encontrar uma representação com menor dimensionalidade que preserve as informações originais e descreva a forma original dos dados das séries temporais, tanto quanto possível (VLACHOS, LIN e KEOGH, 2003).

As técnicas típicas de redução de dimensionalidade incluem a Transformada de Fourier Discreta (TFD), TWD e Decomposição de Valor Singular (*Singular Value Decomposition* – SVD), baseada na PCA. Com tais técnicas, os dados ou sinal são mapeados para um pequeno

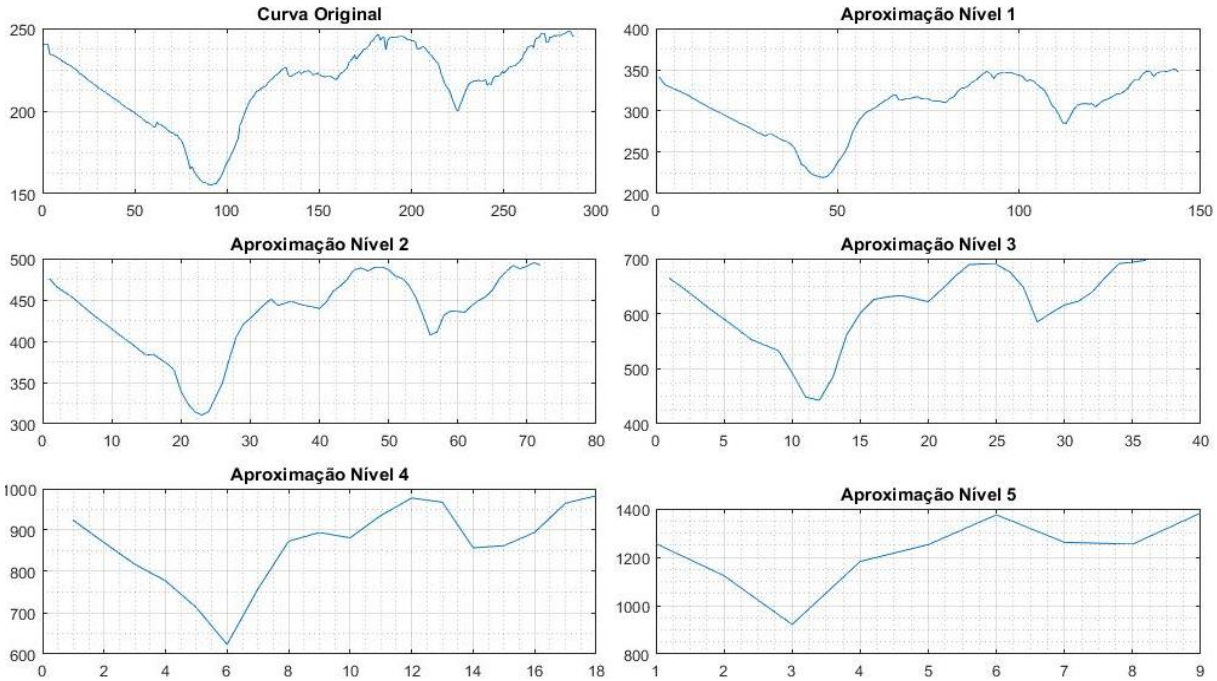
subconjunto de coeficientes que representam característica relevantes do sinal estudado (HAN, PEI e KAMBER, 2012).

Embora diversas abordagens de extração de características tenham compartilhado a capacidade de produzir uma aproximação de dimensionalidade reduzida de alta qualidade de séries temporais, as *wavelets* são únicas, na medida em que sua representação de dados é intrinsecamente multiresolução (VLACHOS, LIN e KEOGH, 2003). Nesse sentido, a TW permite a extração de características do mesmo sinal com resoluções (quantidade de amostras) diferentes, podendo ser utilizada como ferramenta de redução de dimensionalidade da série temporal.

2.3.1.1 Banco de Filtros da TWD Aplicados na Extração de Características de Séries Temporais

Conforme item 2.2.1.3, o banco de filtros da TWD fornece através do filtro passa-baixa as aproximações do sinal original em diversos níveis, com o operador *downsampling*, proporcionando diversas representações do mesmo sinal, com número de amostras reduzidas, como ilustradas na Figura 2.16.

Figura 2.16 – Curva original e aproximações dos níveis 1, 2, 3, 4 e 5 da TWD.



Fonte: Elaborada pelo autor.

As diversas resoluções oriundas do filtro passa-baixa da TWD, verificadas na Figura 2.16, podem ser utilizadas como extração das características da curva original nos diversos

níveis da TWD, fazendo dessa uma ferramenta com grande potencial, em termos de redução de dimensionalidade de séries temporais.

2.3.1.2 Análise de Componentes Principais

Há um problema inerente a objetos que possuem elevada dimensionalidade para serem visualizadas ou manipuladas. Os métodos de seleção ou extração de características de séries temporais, podem ser utilizados como uma ferramenta para a redução da dimensionalidade desses vetores (HAN, PEI e KAMBER, 2012).

A PCA é uma técnica multivariada de modelagem da estrutura de covariância (PEARSON, 1901). Trata-se de uma técnica estatística de análise multivariada que transforma linearmente um conjunto original de variáveis, inicialmente correlacionadas entre si, num conjunto menor de variáveis não correlacionadas, que contém a maior parte da informação do conjunto original (HONGYU, SANDANIELO e DE OLIVEIRA JUNIOR, 2016).

Supondo que os dados a serem reduzidos sejam séries temporais descritas por n atributos ou amostras no tempo. A PCA procura vetores ortogonais $k - dimensional$, que podem ser melhor usados para representar os dados, onde $k \leq n$. Os dados originais são então projetados em um espaço muito menor, resultando em redução de dimensionalidade. Sendo assim, a PCA "combina" as principais características do vetor estudado criando um vetor alternativo menor. Contudo, frequentemente, essa técnica revela relacionamentos que não eram suspeitos anteriormente e que permitem interpretações que normalmente não seriam possíveis (HAN, PEI e KAMBER, 2012).

Dentre as diversas técnicas baseadas na ideia de redução de dimensionalidade de dados com menor perda possível de informação a PCA é a mais difundida (HONGYU, 2015).

Depois da aplicação da PCA em um conjunto de dados, espera-se que o resultado do agrupamento, dos dados com dimensão reduzida, seja o mesmo realizado com dados com suas dimensões originais (HAN, PEI e KAMBER, 2012).

Sendo assim, a PCA pode ser utilizada em curvas de carga como entradas para análise de agrupamento que utilizam algoritmos convencionais, os quais necessitam da redução de dimensionalidade do conjunto de curvas que serão agrupadas.

2.3.2 Abordagens de Agrupamentos de Dados de Séries Temporais

O agrupamento de dados de séries temporais pode ser amplamente classificado em abordagens convencionais e abordagens híbridas. As abordagens convencionais empregadas no agrupamento de dados de séries temporais são tipicamente algoritmos de particionamento, hierárquicos ou baseados em modelo (AGHABOZORGI et al., 2014; LIAO, 2005). Por outro lado, alguns novos artigos enfatizam o aprimoramento de algoritmos convencionais, através da utilização de ferramentas que contornam as fragilidades encontradas nos algoritmos convencionais, apresentando modelos customizados (tipicamente métodos híbridos) para agrupamento de séries temporais (AGHABOZORGI et al., 2014).

No agrupamento hierárquico, uma hierarquia aninhada de objetos similares é construída com base em uma matriz de distâncias pareadas (VLACHOS, LIN e KEOGH, 2003). O agrupamento hierárquico permite uma visualização compacta de agrupamento de séries temporais, tornando esse método de agrupamento vantajoso para a verificação do resultado de agrupamento de séries temporais (HIRANO e TSUMOTO, 2005).

O agrupamento hierárquico não exige o número de grupos como um parâmetro inicial, essa característica é importante no agrupamento de séries temporais porque define o número de grupos muitas vezes complexa em problemas reais. No entanto, o agrupamento hierárquico é muito oneroso ao lidar com grandes conjuntos de dados (WANG, SMITH e HYNDMAN, 2006) devido à sua complexidade computacional quadrática. Com isso o agrupamento hierárquico é restrito a conjuntos de dados pequenos. Por outro lado, algoritmos de particionamento, como o conhecido algoritmo *k-means* (MACQUEEN, 1967), estão entre os algoritmos mais usados para agrupamento de grandes conjuntos de dados. Além do que, agrupamentos realizados com o algoritmo *k-means* são rápidos quando comparados com agrupamentos hierárquicos (BRADLEY, FAYYAD e REINA, 1998), tornando-os muito adequados para agrupamento de séries temporais.

O conceito de agrupamentos baseados em modelos considera que cada série temporal é gerada por algum tipo de modelo ou por uma mistura de distribuições de probabilidade subjacentes. As séries temporais são consideradas semelhantes quando os modelos que caracterizam séries individuais ou os resíduos remanescentes, após a montagem do modelo, são semelhantes (LIAO , 2005). Sendo assim, o agrupamento baseado em modelos pressupõe que deve haver um tipo de modelo para cada tipologia de série temporal.

No entanto, ocorrem dois inconvenientes no agrupamento baseado em modelos. Primeiramente, os parâmetros devem ser definidos e a configuração do parâmetro é baseada nas suposições do usuário, que podem ser falsas e resultar em grupos imprecisos. Segundo, o tempo de processamento de agrupamento baseado em modelos é demasiado, quando se realiza o agrupamento de grandes conjuntos de dados (ANDREOPOULOS et al., 2009).

2.3.3 Medidas de Distância de Dissimilaridade de Séries Temporais

Um item fundamental no processo de agrupamento de séries temporais é a função usada para medir sua dissimilaridade. Os dados a serem agrupados podem se manifestar de diversas formas, incluindo séries temporais de comprimento igual ou desigual, vetores de valor de característica e assim por diante (LIAO, 2005).

Várias métricas de distância para constatar dissimilaridade de séries temporais foram propostas por pesquisadores na literatura, no entanto a Distância Euclidiana (*Euclidean Distance* – ED) e a *Dynamic Time Warping* (DTW) são os métodos mais utilizados no agrupamento de séries temporais (AGHABOZORGI et al., 2014).

A ED é simples e rápida, em grande parte devido não necessitar de parâmetros, o que faz da mesma a medida de dissimilaridade mais utilizada em estudo de agrupamentos (CHAN, FU e YU, 2003), porém é sensível a *outliers* (XI et al., 2006).

Dentre as demais métricas utilizadas para verificar dissimilaridade de séries temporais, a autocorrelação tem sido proposta juntamente com uma variedade de outras medidas nos últimos anos, dentre elas, a normalização por partes e medidas probabilísticas por partes (WANG, SMITH e HYNDMAN, 2006).

Por outro lado, a DTW tem sido aplicada no agrupamento de séries temporais de comprimentos variáveis ou contendo possíveis semelhanças fora de fase. Contudo, apresenta complexidade computacional quadrática (AGHABOZORGI et al., 2014).

Pesquisa e comparação empírica de séries temporais realizada em KEOGH e KASETTY (2003), revelaram que a ED apresenta resultados mais favoráveis em relação a DTW, quando comparada com conjuntos de séries temporais de tamanhos iguais.

2.3.3.1 Distância Euclidiana

Uma medida de dissimilaridade deve satisfazer quatro condições principais, segundo HAN, PEI e KAMBER (2012):

Não negatividade: $d(i,j) \geq 0$, a distância não deve ser um número negativo.

Identidade de indiscerníveis $d(i,i) = 0$, a distância de um objeto para si mesmo deve ser 0.

Simetria: $d(i,j) = d(j,i)$, a função de distância de uma métrica de similaridade deve ser simétrica.

Desigualdade triangular: $d(i,j) \leq d(i,k) + d(k,j)$, a distância no espaço de um objeto i para um objeto j, deve ser menor ou igual a distância de i para j passando sobre outro objeto k.

Sendo assim, para $i = (x_{i1}, x_{i2}, x_{ip})$ e $j = (x_{j1}, x_{j2}, x_{jp})$, dois objetos descritos por p atributos numéricos. A ED entre os objetos i e j é definida como (HAN, PEI e KAMBER, 2012):

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2} \tag{2.6}$$

Observe que para a ED, a propriedade de não negatividade está implícita nas outras três propriedades.

2.3.4 Algoritmo *k-means*

Supondo que um conjunto de dados D contenha n objetos, esses devem ser distribuídos em k grupos, $C_1, \ldots, C_k$ dentro de D, isto é, $C_i \subset D$ e $C_i \cap C_j = \emptyset$ para $1 \leq i\ ,\ j \leq k$. Uma função objetivo é usada para avaliar o particionamento de modo que os objetos de um grupo sejam semelhantes entre si, mas diferentes de objetos em outros grupos (HAN, PEI e KAMBER, 2012). Sendo assim, a função objetivo visa alta similaridade intragrupo e baixa similaridade intergrupo.

Para a realização do processo iterativo de agrupamento, primeiramente o *k-means* seleciona aleatoriamente k objetos do conjunto de dados D, cada um dos quais representa inicialmente uma média ou centro de cada grupo (centroide). Os objetos restantes são atribuídos ao grupo mais semelhante, com base em métricas de distância e similaridade (por exemplo a

ED) entre o objeto e o centroide do grupo. O algoritmo *k-means* então melhora iterativamente a variação dentro de cada grupo, calculando a nova média, centroide do grupo, usando os objetos atribuídos ao mesmo grupo na iteração anterior. Todos os objetos são redistribuídos para os grupos a partir dos centroides atualizados. As iterações continuam até que não haja mais redistribuição de objetos em novos grupos (HAN, PEI e KAMBER, 2012; LIAO, 2005), ou relativamente poucos objetos mudem de grupos (CAMARGOS e NICOLETTI, 2015).

A complexidade do algoritmo *k-means* é $O(k.n.i.d)$, onde k é o número de grupos, n o número de pontos, i o número de iterações e d o número de atributos (CAMARGOS e NICOLETTI, 2015). Como podemos constatar, o *k-means* não é um algoritmo de complexidade quadrática, conferindo rapidez na realização de agrupamentos em relação a algoritmos de complexidade quadrática como o *k-medoids*, por exemplo.

O algoritmo *k-means* possui fragilidades, dentre elas a mais relevante trata-se da necessidade de parametrização antecipada do número de grupos para a realização dos agrupamentos. Essa limitação pode ser atenuada colocando-se o algoritmo em loop para um intervalo de diversos valores de k grupos. Metodologias de validação podem ser usadas para determinar qual valor de k é mais adequado.

O pseudocódigo do algoritmo *k-means* é apresentado a seguir:

Entradas:

k – número de grupos;

D – conjunto de dados contendo n objetos.

Saída:

Conjunto de k grupos.

Pseudocódigo:

1. Escolha arbitrariamente k objetos de D como os centros iniciais de cada grupo;
2. Repetir:
   - 2.1. redistribuir cada objeto para o grupo que há maior proximidade do objeto em relação ao centroide do grupo;
   - 2.2. calcular os novos centroides para cada grupo;
   - 2.3. caso os objetos nos grupos da iteração atual sejam diferentes da iteração anterior, voltar ao passo 2.1, até a estabilidade dos grupos.

Outra fragilidade é fato de que o *k-means* pode convergir para ótimos locais. Portanto, é importante que seja realizado no mínimo uma replicação do algoritmo, afim de verificar se a função objetivo obteve melhores resultados, com menor distância intragrupo e maior distância intergrupo, buscando ótimos globais (VLACHOS, LIN e KEOGH, 2003).

2.3.5 Validação de Agrupamentos

Validação de agrupamento é uma estratégia aplicada para verificar o melhor resultado de agrupamento realizado entre metodologias de agrupamentos diferentes ou na variação de parâmetros de uma mesma metodologia, ou seja, consiste no processo de avaliar os resultados gerados por algoritmos de agrupamentos. Para realização da validação se utiliza índices que são medidas objetivas, podendo fazer parte da estratégia de validação. As abordagens empregadas na validação de agrupamento são: validação externa, validação interna e validação relativa, que é uma especificidade da validação interna. Todas utilizam índices que aferem características da solução ou medem a similaridade entre duas soluções (CAMARGOS e NICOLETTI 2015; LIAO 2005; HAM, KAMBER e PEI, 2012; CAMPELLO, 2017).

Os índices utilizados na validação de agrupamentos são classificados em três tipos: índice interno, índice externo e índice relativo.

A validação de agrupamentos é eficaz, em relação à qualidade de agrupamentos, se satisfizer à quatro critérios essenciais (HAM, KAMBER e PEI, 2012):

- **Homogeneidade do agrupamento**. Esse critério afirma que, quanto mais puros ou similares forem os grupos de um agrupamento, melhor será o agrupamento. Em outras palavras, quanto mais similares forem os elementos de cada grupo, melhor será o resultado do agrupamento.
- **Integralidade do agrupamento**. Essa é uma consequência da homogeneidade do agrupamento, ou seja, se dois objetos pertencerem à mesma categoria de acordo com um modelo conhecido, eles devem ser atribuídos ao mesmo grupo, ou seja, um agrupamento deve atribuir objetos pertencentes à mesma categoria para o mesmo grupo.
- ***Rag Bag***. Em muitos cenários reais, há uma categoria de "*rag bag*" contendo objetos que não podem ser mesclados com outros objetos. Tal categoria é frequentemente chamada de diversos, outros, *outliers* e assim por diante. O

critério do *rag bag* afirma que a ação de colocar um objeto heterogêneo em um grupo homogêneo deve ser penalizada mais do que colocá-lo em um *rag bag*.

- **Preservação de pequenos grupos**. Se uma pequena categoria for dividida em pequenos pedaços em um agrupamento, essas pequenas partes provavelmente se tornarão ruído e, portanto, a pequena categoria não poderá ser descoberta a partir do agrupamento. O critério de preservação de pequenos grupos afirma que dividir uma pequena categoria em partes é mais prejudicial do que fazer o mesmo a uma grande categoria.

2.3.5.1 Validação Externa

Na validação externa o mais comum é a medição da similaridade entre o agrupamento gerado pelo algoritmo estudado e um agrupamento de referência, o que é feito por meio da aplicação de índices (CAMARGOS e NICOLETTI, 2015). Sendo assim, a validação externa é realizada a partir de um modelo conhecido do agrupamento ideal, ou seja, trata-se de um método supervisionado, pois mede a exatidão na qual a estrutura do agrupamento descoberta por um algoritmo de agrupamento, corresponde a uma estrutura externa ou modelo (CAMARGOS e NICOLETTI, 2015; HAM, KAMBER e PEI, 2012). Por fim, após o agrupamento sua qualidade é medida utilizando índices externos, que verificam a correspondência entre o agrupamento realizado e o modelo fornecido como referência.

2.3.5.2 Validação Interna

Na validação interna os resultados encontrados são avaliados somente com relação aos próprios dados da base, sem a auxílio de informações sobre a solução ideal (CAMARGOS e NICOLETTI, 2015; HAM, KAMBER e PEI, 2012).

A validação interna pode ser vista como uma estratégia não supervisionada pois mede a estrutura do agrupamento, em termos de coesão e isolamento, sem que seja necessário a tendência de agrupamento de um conjunto de dados, distinguindo se uma estrutura não aleatória realmente existe nos dados, pelas seguintes etapas:

- Comparar os resultados de dois diferentes conjuntos de análise de grupos para determinar qual deles é melhor e
- Determinar o número adequado de grupos.

Assim como os métodos de validação externa, os métodos de validação interna possuem suas métricas de similaridade entre objetos no conjunto de dados e são chamados índices

internos, que por sua vez são usados para medir a qualidade da estrutura de agrupamento sem relação com informações externas ou modelos (CAMPELLO, 2017; NIEVOLA 2017).

A métrica mais comum para validação interna é a Soma do Erros Quadrados (*Sum of the Squared Error* – SSE). O SSE é um índice interno usado para medir a qualidade da estrutura de agrupamento sem relação a alguma informação externa, sendo indicado para comparar dois agrupamentos ou dois grupos que utilizam o mesmo método de agrupamento com parametrizações diferentes. Nesse índice, para cada objeto do agrupamento o erro é a distância ao grupo mais próximo. Para obter o SSE, os erros são elevados ao quadrado e somados (NIEVOLA, 2017; HAM, KAMBER e PEI, 2012), conforme a Equação (2.7):

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} dist^2(m_i, x) \tag{2.7}$$

Em que:

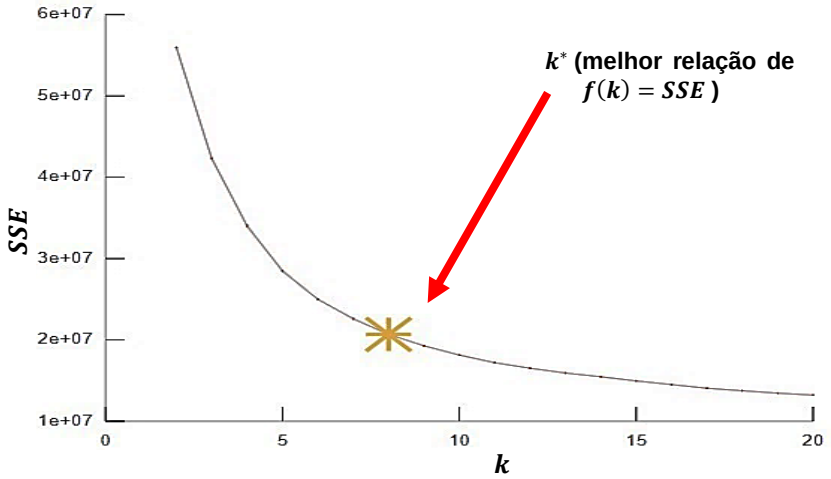
k – número de grupos;

x – objeto no grupo C_i e

m_i – centroide do grupo C_i.

A resolução do problema de determinação adequada do número de grupos k, é atacada por dezenas de pesquisadores. A maioria aceita o critério da minimização do SSE, ou seja, dado dois grupos, o grupo com melhor avaliação é aquele com o menor SSE (MIRKIN, 2011). Uma estratégia de reduzir o SSE é aumentar k, porém o melhor agrupamento deve ser aquele que resultar no menor k para o menor SSE (NIEVOLA, 2017), conforme ilustrado na Figura 2.17.

Figura 2.17 – Gráfico de valores de ***SSE*** em função do número de grupos, ***k***.



Fonte: Elaborada pelo autor.

Na Figura 2.17 pode-se verificar que, o valor com melhor relação de SSE ocorre no joelho da curva, ou seja, no ponto de saturação.

Sendo assim, o SSE no ponto de saturação pode ser utilizado para responder a uma das questões fundamentais em validação: qual é o número de grupos mais satisfatório, k^*? (CAMPELLO, 2017).

2.3.5.3 Validação Relativa

A validação relativa trata-se de uma especificidade da validação interna, com o objetivo de apontar a melhor solução de um conjunto de agrupamentos, para algoritmos e metodologias de agrupamentos diferentes (CAMARGOS e NICOLETTI, 2015; HAM, KAMBER e PEI, 2012). Para a validação relativa, são utilizados índices relativos, que comparam dois grupos ou agrupamentos diferentes. Frequentemente um índice externo ou interno é usado para esta função, sendo às vezes chamados de critérios em vez de índices. Entretanto, o critério é a estratégia geral, enquanto o índice é a medida que implementa o critério (CAMPELLO, 2017).

Existem vários índices para a validação relativa, alguns dos mais populares são: o índice de *Davies-Bouldin* (DAVIES e BOULDIN, 1979; JAIN e DUBES, 1988), o índice de *Dunn* (DUNN, 1973) e o índice da silhueta, *Silhouette Width Coefficient* (SWC) (ROUSSEEUW, 1987). Para medir o *fitness* de um agrupamento, o índice da silhueta e outros índices relativos e internos podem ser usados no ponto de saturação da curva, para encontrar o número adequado de grupos em um conjunto de dados, substituindo a soma das variâncias intragrupos (HAM, KAMBER e PEI, 2012).

Conforme VENDRAMIN, CAMPELLO e HRUSCHKA (2010), o índice da silhueta obteve resultados melhores que dezenas de outros índices relativos inclusive os citados anteriormente.

O SWC define a qualidade dos agrupamentos com base na proximidade entre os objetos de um determinado grupo, assim como, o afastamento desses objetos ao grupo mais próximo (ROUSSEEUW, 1987; HAM, KAMBER e PEI, 2012).

O processo para obtenção do SWC pode ser descrito da seguinte forma (HAM, KAMBER e PEI, 2012; KAUFMAN e ROUSSEEUW, 1990):

Para um conjunto de dados D de n objetos, suponha que D seja agrupado em k grupos, C_1, ..., C_k. Para cada objeto O pertencente a D, calculamos $a(O)$ como a distância média entre O e todos os outros objetos, O', no grupo C_i ao qual O pertence, ou seja, corresponde ao

somatório de todas as distâncias entre O e cada objeto O' do mesmo grupo de O, dividido pelo número de comparações, conforme Equação (2.8):

$$a(O) = \frac{\sum O' \in C_i, O \neq O' dist(O, O')}{|C_i| - 1}$$

(2.8)

Da mesma forma, $b(O)$ na Equação (2.9), é a distância média mínima de O para todos objetos dos grupos C_j aos quais O não pertence, sendo $O \in C_i, (1 \leq i \leq k)$. Isto é, $b(O)$ corresponde à menor média de todas as distâncias do objeto O em relação à cada um dos objetos dos outros grupos, $O' \in C_j$, sendo assim:

$$b(O) = min_{C_j: 1 \leq j \leq k, j \neq i} \left\{ \frac{\sum O' \in C_j \, dist(O, O')}{|C_i|} \right\}$$

(2.9)

O resultado do índice da silhueta do objeto analisado O será a diferença entre $b(O)$ e $a(O)$ dividido pelo maior valor entre $a(O)$ e b(O), conforme segue na Equação (2.10):

$$s(O) = \frac{b(O) - a(O)}{max\{a(O), b(O)\}}$$

(2.10)

O índice da silhueta para cada grupo, $s(C_i)$, será a média dos índices da silhueta de cada grupo C_i da Equação (2.11):

$$s(C_i) = \frac{\sum s(O), O \in C_{i,1 \leq i \leq k}}{quantidade\ de\ objetos\ de\ C_{i,1 \leq i \leq k}}$$

(2.11)

E finalmente resultando do índice da silhueta para o agrupamento, SWC, será a média dos índices da silhueta média dos grupos, que por sua vez é obtido pela Equação (2.12).

$$SWC = \frac{\sum s(C_i), 1 \leq i \leq k}{k}$$

(2.12)

O valor do coeficiente de silhueta deve estar entre -1 e 1. O valor de $a(O)$ reflete a compactação do grupo C_i ao qual O pertence. Quanto menor o valor, mais compacto é o grupo. O valor de $b(O)$ constata o grau de separação entre O e outros grupos C_j. Quanto maior $b(O)$, mais O está separado de outros grupos. Portanto, quando o valor do índice da silhueta de O se

aproxima de 1, significa que o grupo C_i contendo O está compacto e O está distante de outros grupos C_j, o que é satisfatório. No entanto, quando o valor do índice da silhueta é negativo, isto é, $b(O) < a(O)$, significando que O está mais próximo dos objetos de outro grupo, C_j, do que dos objetos no mesmo grupo, C_i e essa situação deve ser evitada.

Sendo assim, o resultado do índice da silhueta para o agrupamento, SWC, deve variar entre –1 e 1. Quanto mais próximo de 1 significa que na média os objetos estão melhor alocados no grupo. Porém, quanto mais próximo de –1, significa que para todo o agrupamento a alocação dos objetos nos grupos esta inadequada (HAM, KAMBER e PEI, 2012; PUMA-VILLANUEVA e ZUBEN, 2008; ROUSSEEUW, 1987).

2.4 TRABALHOS RELACIONADOS

Yang e Stenzel (2005) utilizaram a diferença de segunda ordem da carga para detecção dos dados errôneos através de um intervalo de confiança de -3 a +3, já que valores próximos de zero correspondem a cargas sem anormalidades. Com isso, foi possível identificar os instantes das ocorrências de erros nos dados da carga e separar uma curva de carga diária em vários segmentos contínuos. Também foi utilizado um modelo de regressão linear quadrática para avaliar a validade de cada segmento, fornecendo as estimativas da carga caso o segmento não fosse válido. Porém, os autores advertiram que esse procedimento pode gerar erros em segmentos de longa duração e restringiram essa metodologia apenas para segmentos com menos de 75 minutos, que corresponde à 5 amostras discretizadas de 15 em 15 minutos.

Em Oliveira (2013), foi proposta uma metodologia de tratamento e agrupamento de curvas de carga. O tratamento das curvas de carga foi dividido em duas etapas. Na primeira, ocorre a identificação de falhas. Para identificação de falhas por descontinuidade foram aplicadas a TWD e o aplicativo Boxplot, sendo aplicado esse último no logaritmo dos maiores valores absolutos da componente de detalhe da TWD. A segunda fase da metodologia de filtragem de dados trata da correção de falhas. Nesta fase, foi aplicado o método *Locally Weighted Regression and Smoothing Scatterplots* (LOESS) que suaviza a curva e fornece uma correção suave para os pontos descontínuos anteriores. Por fim utilizou-se do método *Hyperbolic Smoothing Clustering Method* (HSCM) para análise de agrupamentos, identificando os perfis de carga que apresentam aplicabilidade na correção de lacunas, na previsão de carga e no cálculo de tarifas.

Um dos aspectos cruciais no processamento de sinais biomédicos consiste no conhecimento associado aos ruídos que podem ocorrer nesses sinais. A coleta de dados oriundos de um eletroencefalograma (EEG) pode conter ruídos que podem imitar a atividade cerebral e são extremamente difíceis de distinguir, podendo influenciar um diagnóstico equivocadamente. Sendo assim, Rincón, Risk e Liberczuk (2012), utilizaram o filtro de *hampel* para melhorar a análise de dados em gravações de EEG, obtendo bons resultados.

Vlachos (2003) elaborou um algoritmo utilizando a propriedade multiresolução da TW, a partir de um agrupamento inicial realizado com uma representação com poucas características dos dados. Os resultados obtidos a partir deste agrupamento rápido e de poucas características são usados para inicializar um agrupamento com dados de entrada em maior nível de aproximação do sinal original. Este processo é repetido até que os resultados de agrupamento se estabilizem ou até a "aproximação" dos dados originais "não tratados". O agrupamento é dito estabilizado quando os objetos não mudam a associação da última iteração, ou quando a mudança de associação não melhora os resultados de agrupamento. A metodologia multiresolução da TW, foi utilizada como entradas no algoritmo *k-means* modificado, obtendo qualidade do agrupamento superior ao *k-means* convencional e, mesmo que o algoritmo proposto seja executado até a conclusão, o tempo tomado é menor do que o tempo praticado pelo algoritmo *k-means* convencional. No entanto, o método de validação do agrupamento se baseia em índices externos, que necessitam de modelos de agrupamentos conhecidos.

Em Aghabozorgi et al. (2014), foi proposto um algoritmo de agrupamento híbrido com base na similaridade na forma de dados de séries temporais. Os dados de séries temporais são primeiro agrupados como subgrupos com base na similaridade no tempo. Os subgrupos são expandidos baseado na similaridade na forma. Para avaliar a precisão do modelo proposto, o modelo é testado extensivamente usando conjuntos de dados de séries temporais simuladas e do mundo real, constatando que este modelo é mais preciso que outras abordagens convencionais e híbridas, assim como determina a similaridade na forma entre dados de séries temporais com baixa complexidade.

Para resolver o problema de negligência de informações após a redução de dimensionalidade em séries temporais, Lai, Chung e Tseng (2010) propuseram uma metodologia híbrida através de um método de agrupamento de dois níveis, em que tanto a série temporal inteira quanto a subsequência da série temporal são consideradas no primeiro e segundo níveis, respectivamente. Foram utilizados a transformação *Symbolic Aggregate AproxXimation* (SAX) como um método de redução de dimensão e a *Cluster Affinity Searching*

Technique (CAST) como um algoritmo de agrupamento para agrupar dados de primeiro nível. Para medir as distâncias entre os dados de séries temporais no segundo nível, a métrica de distância DTW foi usada em dados com comprimentos variados, e a ED foi utilizada em dados de igual comprimento. No entanto, o algoritmo CAST é usado duas vezes nessa abordagem, uma vez para gerar grupos iniciais e o outro para dividir cada grupo em subgrupos, com bons resultados. Contudo, essa metodologia atribuiu grande complexidade e esforço computacional a metodologia proposta.

Em Zhang et al. (2011), foi proposto um método de agrupamento híbrido utilizando uma abordagem multinível para agrupamento de séries temporais. Primeiro, os dados da série temporal são selecionados de um grupo gerado de um vizinho mais próximo. Para gerar esse grupo de séries temporais, os autores propõem uma medida de distância triangular para calcular a similaridade entre os dados das séries temporais. O armazenamento em agrupamento hierárquico é então executado nos dados das séries temporais selecionadas. Em seguida o tamanho dos dados é reduzido em aproximadamente 10%. A complexidade computacional mais próxima é $O(n^2)$, que é bastante alta. Como resultado, os autores tentam reduzir a área de pesquisa por meio do pré-agrupamento de dados usando o algoritmo *k-means* e limitando a pesquisa a cada grupo, afim de reduzir os agrupamentos gerados. No entanto, a geração dos agrupamentos em si permanece onerosa, tornando inaplicável em grandes conjuntos de dados.

Paparrizos e Gravano (2017) apresentaram o algoritmo *k-Shape* e *k-MultiShapes* (k-MS), dois novos algoritmos para agrupamento de séries temporais. O k-Shape e o k-MS dependem de um procedimento de refinamento iterativo. Como medida de distância, k-Shape e k-MS usam *Shape-Based Distance* (SBD) que se trata de uma distância baseada na forma que não utiliza parâmetros. Com base nas propriedades da SBD, foram desenvolvidos o método, *Shape Extraction* (SE) e *Multi Shapes Extraction* (MSE), para calcular centroides de grupos que são usados em cada iteração. O *k-Shape* se respalda no SE para calcular um único centroide por grupo baseado em todas as séries temporais de cada grupo. Em contraste, o *k-MS* depende do MSE para calcular vários centroides, a proximidade e a distribuição espacial de séries temporais em cada grupo. Para demonstrar a robustez da SBD, *k-Shape* e *k-MS*, realizou-se uma extensa avaliação experimental em relação a medidas de distância e métodos de agrupamento para séries temporais. A SBD, alcançou exatidão semelhante à DTW, uma medida de distância altamente precisa, porém computacionalmente onerosa, que requer ajuste de parâmetros. Para análise de agrupamento, verificou-se a eficácia do *k-Shape* e *k-MS* em relação a métodos convencionais, hierarquizados e baseados em densidade. O k-Shape superou todos os métodos,

com exceção do k-medoids com DTW, alcançando precisão semelhante. O k-MS funciona de maneira semelhante ao k-Shape, mas significativamente mais preciso.

Wang, Smith e Hyndman (2006) propuseram um método para agrupamento de séries temporais baseado em suas características estruturais estatísticas. Ao contrário de outras alternativas, esse método não agrupa objetos utilizando métrica de distância, e sim grupos baseados em características globais extraídos da série temporal. As características que são obtidas de cada série individual são inseridas em algoritmos de agrupamento, incluindo rede neural não supervisionada, mapa auto-organizáveis ou algoritmo de agrupamento hierárquico. As características extraídas, reduzem a dimensionalidade da série temporal e são muito menos sensíveis à dados ausentes, descontinuidades e *outliers*. A metodologia proposta foi testada comparando o agrupamento resultante a um conjunto de dados de séries temporais com características conhecidas. Os resultados empíricos mostram que essa abordagem é capaz de gerar agrupamentos significativos, semelhantes aos produzidos por outras metodologias.

Um novo algoritmo de agrupamento de séries temporais chamado *Haliteds* foi proposto por Silva (2013), o algoritmo tem a finalidade de realizar agrupamento em subespaços de séries temporais. É utilizada como base a técnica *Halite*, originalmente voltada para a análise de dados estáticos. Em comparação ao uso do algoritmo base em séries temporais, o novo algoritmo permite que o conhecimento obtido dos dados do passado facilite o agrupamento dos dados no presente, diminuindo o tempo de análise.

A compreensão dos padrões de consumo oriundos de energias alternativas é extremamente importante, pois tradicionalmente as análises energéticas são realizadas para grandes sistemas elétricos envolvendo regiões e nações. No entanto, com o advento das *smart grids*, o estudo do comportamento de regiões menores tornou-se uma necessidade para permitir um microgerenciamento melhor e profundo nos sistemas elétricos de potência. Sabendo disso Hernández et al. (2012) apresentam um sistema de processamento de dados para analisar padrões de consumo de energia em parques industriais, baseado na aplicação em cascata de um mapa auto-organizável e do algoritmo de agrupamento *k-means*. O sistema é validado com dados reais de carga de um parque industrial na Espanha. Os resultados da validação mostram que o sistema encontra significativas tipologias sem supervisão, ou qualquer conhecimento prévio sobre os dados.

As *smart grids* exigem métodos de previsão orientados por características de dados. Auder et al. (2018) propuseram ferramentas de agrupamento para previsão de carga de curto prazo. Foram analisados dados de consumo individual, para identificação de perfis padrões,

pois esses são essenciais no gerenciamento de energia e previsão de carga elétrica. Primeiramente foi realizada a análise de cargas no contexto industrial e residencial. Em seguida, verificou-se séries temporais hierárquicas para previsão. A estratégia consiste em decompor o sinal global e obter previsões divididas, de tal forma que a previsão resultante seja a soma das parcelas de previsões resultantes. Essa tarefa é realizada em três etapas. Na primeira etapa são identificados grande número de superconsumidores agrupando seus perfis de energia. A segunda etapa refere-se à elaboração de uma hierarquia de partições aninhadas. Por fim são selecionadas as hierarquias que minimizem determinados critérios de previsão. Utilizou-se um modelo não paramétrico para lidar com previsão e a transformada *wavelet*s afim de definir similaridades entre curvas de carga. Essa estratégia de desagregação proporcionou uma melhoria de 16% na precisão da previsão.

Conforme Lin, Wu e Su (2018), os métodos de agrupamento tradicionais não são adequados para séries temporais de curva de carga. Para agrupar os dados da curva de carga com mais precisão, aplicou-se um procedimento de similaridade aprimorada das curvas de carga, baseado na distância de *Pearson*. Esse método introduz o conceito de "ponto de alteração de tendência" e o integra à semelhança de *Pearson*. Introduzindo um peso para a distância de *Pearson*. Com base na distância ponderada de *Pearson*, é proposto um algoritmo de agrupamento de hierarquias ponderadas. Foram usados dados de curvas de carga de vários anos, para avaliação. Diversos modelos de consumo foram encontrados e analisados. Os resultados mostraram que o método proposto melhora a precisão do agrupamento de dados de carga.

Uma metodologia de dois estágios foi desenvolvida por Tsekouras, Hatziargyriou e Dialynas (2007) para a classificação de clientes de eletricidade. Essa metodologia baseia-se em métodos de reconhecimento de padrões, como *k-means*, quantização vetorial adaptativa de *Kohonen*, *fuzzy k-means* e agrupamento hierárquico, adequadamente adaptados. No primeiro estágio, as curvas de carga típicas de vários clientes são estimadas usando métodos de reconhecimento de padrões, e seus resultados são comparados usando seis medidas de adequação. No segundo estágio, a classificação dos clientes é realizada através dos métodos, medidas, e padrões de carga típicas dos clientes, obtidos no primeiro estágio. Os resultados da primeira etapa podem ser usados para previsão de carga de clientes e determinação de tarifas. Os resultados da segunda etapa fornecem informações valiosas para os fornecedores de eletricidade em mercados de energia competitivos.

Na coleta de grandes conjuntos de dados de medidas elétricas, observa-se muitas falhas. Wu et al. (2017) pesquisaram o emprego do método matemático de *hill-climbing* para

determinar os centros de agrupamentos iniciais e o número de agrupamentos. Verificou-se que o método pesquisado supera a subjetividade de determinar o número de agrupamentos e impede que a função objetivo caia no mínimo local. Em seguida, foi proposto um método de detecção de falhas, combinando o algoritmo de agrupamento *fuzzy c-means* (FCM) com o método de *hill-climbing*, para determinar o número de grupos e centroides dos agrupamentos iniciais, agrupar dados de carga e finalmente extrair a curva característica da carga. Os experimentos mostram bons resultados na detecção de falhas e correção de dados.

O agrupamento de séries temporais se mostrou eficaz no fornecimento de informações úteis em várias aplicações. Baseados nisso, Räsänen e Kolehmainen (2009) apresentam um método computacional para agrupamento de séries temporais de curvas de carga. A abordagem apresentada utilizou a extração de recursos estatísticos e seu uso em agrupamento baseado em características de dados de energia elétrica medidos em determinados horários. O agrupamento baseado em características conseguiu agrupar séries temporais, empregando apenas um conjunto de recursos estatísticos. As principais vantagens deste método foram a redução de dimensionalidade das séries temporais originais, a não sensibilização com a ausência de dados e o agrupamento de séries temporais de diferentes comprimentos. Avaliou-se o desempenho da abordagem através de dados reais, medidos a cada hora, para 1.035 clientes, durante o período de tempo de 84 dias, o agrupamento utilizando a metodologia da pesquisa resultou em curvas de carga mais precisas.

Existem diferentes aplicações de agrupamento de curvas de carga de sistemas elétricos como: análise de sistema, previsão de carga, preço e etc. Kohan, Moghaddam e Bidaki (2009) avaliaram os desempenhos de dois métodos de agrupamento, o WFA (*Weighted Fuzzy Average*) *k-means* e o MFL (*Modified Follow the Leader*), para a classificação de curvas de carga. Para avaliação e comparação dos métodos foram utilizadas as medidas de adequação, MIA (*Mean Index Adequacy*) e CDI (*Clustering Dispersion Indicator*), que por sua vez mostraram a distinção e a compactação dos grupos resultantes das respectivas metodologias. Uma característica inovadora dessa pesquisa é que foram avaliados os desempenhos de algoritmos de agrupamento com base em diferentes aplicações no sistema elétrico. Na aplicação dos métodos, em dados de uma rede de distribuição no Teerã, os resultados mostraram ser mais adequado empregar a WFA *k-means* para projetar tarifas no mercado devido aos menores valores para o MIA. Por outro lado, a MFL se mostrou mais satisfatória do que a WFA *k-means*, para nortear estudos de expansão energética.

No trabalho de Benítez et al. (2014), realizou-se a segmentação dinâmica de perfis de carga diários, ao longo dos anos de 2008 e 2009, de uma amostra de clientes residenciais. A técnica empregada realiza a classificação dos perfis de carga de consumo de energia, por meio de algoritmos de agrupamento dinâmico. A técnica utilizada se mostrou adequada, como uma ferramenta rápida, para classificar os clientes de acordo com seus padrões e tendências de consumo de energia.

Para atenuar o esforço computacional empregado no planejamento energético, algumas simplificações são necessárias, entre elas o uso de curvas de carga aproximadas ou perfis típicos de carga diária, como por exemplo: pesado, médio e leve. Sendo assim, Pessanha et al. (2018) propuseram uma metodologia para construir perfis de carga diários típicos, para os estudos de expansão e planejamento operacional. A abordagem proposta baseia-se em técnicas estatísticas de agrupamento e em Análises Exploratórias de Dados (EDA), precedidas por um procedimento de filtragem. Utilizou-se dados reais de carga horária do Sistema Interligado Brasileiro de Energia Elétrica, compreendendo um período de 5 anos. Os resultados indicaram que 3 a 5 grupos representam as tipologias de cada mês do ano.

2.5 CONCLUSÕES PARCIAIS

Neste capítulo, foram apresentadas a importância e aplicabilidade da filtragem e do agrupamento de séries temporais de carga, constituindo importantes ferramentas utilizadas na análise de carga por equipe de operação, manutenção e estudos energéticos de agentes do SEP, dentre outras. Também foram mencionados os problemas associados à filtragem de dados e posterior agrupamento das tipologias das curvas de carga, desde a classificação das falhas predominantes dos dados até as estratégias de agrupamento, redução de dimensionalidade e estratégias de validação de agrupamentos. Foram comentados alguns trabalhos relevantes na literatura que abordam o tema agrupamento de séries temporais e filtragem de dados.

De acordo com os trabalhos relacionados a maioria das aplicações utilizam *k-means*, seja isoladamente ou em conjunto com outra metodologia de agrupamento, com resultados satisfatórios de acordo com cada finalidade. Por outro lado, a extração de características é fundamental para agrupamento de séries temporais, considerando a alta dimensionalidade da mesma. Sendo assim, a multiresolução do banco de filtros da TWD faz dessa uma ferramenta ímpar, a ser utilizada tanto na fase de extração de características como na fase de filtragem. Contudo, o filtro de *hampel* é outra importante ferramenta a ser utilizada em conjunto com a

TWD na fase de detecção e tratamento de falhas, sendo estes alguns dos principais fatores que respaldam a utilização dessas ferramentas nesse trabalho.

No próximo capítulo serão apresentados os detalhes e procedimentos utilizados na metodologia adotada.

3. METODOLOGIA ADOTADA

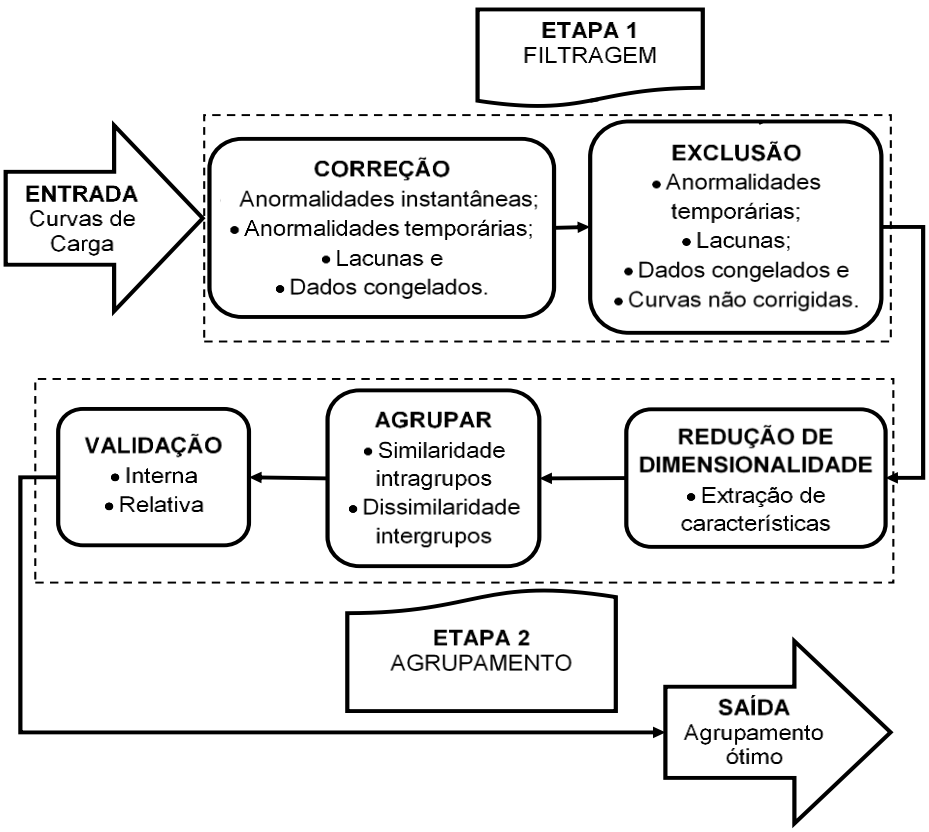
A implementação da metodologia adotada foi realizada com a utilização do software MATLAB (2016), utilizando funções pré-definidas nativas do software e outras funcionalidades.

Na primeira etapa da metodologia, realizou-se a correção das curvas de carga com anormalidade instantâneas, lacunas e dados congelados, de curtíssimo prazo, assim como, a exclusão das curvas de carga com anormalidades temporárias, lacunas e dados congelados acima do intervalo de correção satisfatória pelo filtro de *hampel*.

Na segunda etapa de pré-processamento, a metodologia proposta realizou o agrupamento das curvas de carga, verificando o número de grupos adequado e validando o melhor agrupamento, dentre uma variação de configurações de agrupamentos.

A metodologia proposta está ilustrada na Figura 3.1 e conforme verificado, realiza-se o processamento das curvas de carga históricas em duas etapas que são: Etapa 1 (filtragem) e Etapa 2 (agrupamento).

Figura 3.1 – Metodologia proposta de pré-processamento de curvas de carga.



Fonte: Elaborada pelo autor.

3.1 ETAPA 1 (FILTRAGEM)

A primeira etapa da metodologia proposta consiste na filtragem dos dados. Essa etapa compreende a correção de curvas de carga com falhas e exclusão das curvas com presença de falhas mesmo após a correção. Na correção são recuperados trechos da série temporal utilizando o filtro de hampel, enquanto que curvas de carga com falhas acima do intervalo de correção, admitido pelo filtro de *hampel*, são excluídas do banco de dados, após detecção dessas falhas com o banco de filtro TWD.

3.1.1 Correção e Exclusão de Curvas de Carga com Falhas

A correção das curvas de carga do banco de dados depende do intervalo de falhas considerado corrigível. Por outro lado, a exclusão de curvas de carga do banco de dados somente deve ser realizada para os intervalos de falhas que de fato não for possível serem corrigidas.

Sabendo que o filtro de *hampel* corrigirá as falhas de um determinado intervalo de amostras consecutivas, a questão é, qual o melhor *k-hampel* para esse filtro, *k-hampel** ?

A segunda etapa da filtragem corresponde a exclusão de todas as curvas de carga com anormalidades não corrigidas pelo filtro de *hampel*, ou seja, todas as falhas com intervalo acima do *k-hampel**. Para a identificação dessas falhas foi utilizado como parâmetro o detalhe do banco de filtros da TWD.

Sendo assim, após submeter as curvas de carga ao banco de filtros da TWD, foram consideradas portadoras de falhas todas as curvas de carga que seu detalhe, oriundo do banco de filtros da TWD, possuam valores de amplitude fora de um determinado intervalo, limiar. Contudo, surge uma nova questão: qual o limiar, intervalo ótimo, que serve como parâmetro, para indicar qual a amplitude do detalhe da TWD representa falhas que poderão estar contidas nas curvas de carga do banco de dados?

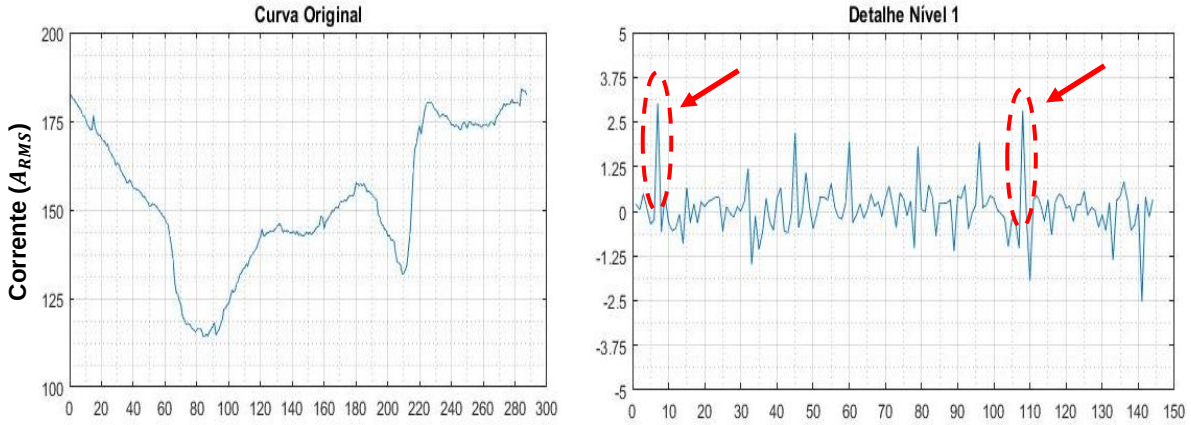
Para responder a última pergunta, realizou-se uma busca exaustiva, no intuito de verificar quais as possíveis falhas que poderiam estar contidos no banco dados e em seguida a parametrização adequada do banco de filtros da TWD.

3.1.1.1 Limiar ótimo do detalhe da TW

Verificou-se no item 2.3.1.1 que o banco de filtros da TWD pode ser aplicado como indicador de faltas no SEP, através do sinal resultante do seu filtro passa-alta, ou seja, o detalhe das altas frequências da curva original. Sendo assim, uma estratégia utilizada nessa pesquisa, para indicação de falhas na curva de carga do banco de dados, foi analisar o comportamento da curva de detalhe do banco de filtros da TWD no nível 1, já que é o nível mais próximo da curva original. Isso devido, ao aumento instantâneo na amplitude do detalhe do sinal, oriundo do banco de filtros da TWD, manifestar-se em qualquer falha da curva de carga analisada.

Na Figura 3.2 observa-se o detalhe do nível 1, do banco de filtros da TWD, oriundo de uma curva de carga sem falhas.

Figura 3.2 – Curva de carga original sem falhas e detalhe no nível 1 do banco de filtros da TWD.

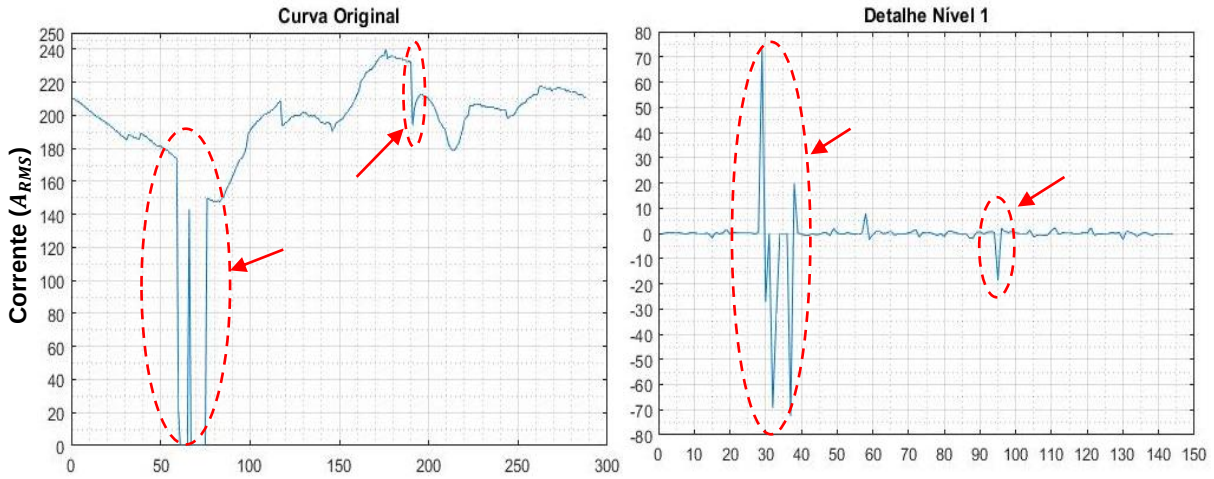


Fonte: Elaborada pelo autor.

Pode-se observar na Figura 3.2 que os maiores picos, do sinal do detalhe nível 1 do banco de filtros da TWD, não ultrapassam o intervalo de amplitude de +3,75 e -2,5. Isso se deve ao fato da curva de carga original não possuir falhas. Vale lembrar, que na Figura 3.2, os maiores picos do sinal do detalhe não correspondem a falhas. O primeiro pico, ocorrido na amostra 7 do sinal do detalhe, indicado pelo tracejado esquerdo, corresponde ao desligamento de um banco de capacitor de 69KV, conectado no lado secundário do transformador 230/69 KV referido no estudo. Contudo, o segundo pico, ocorrido na amostra 108 do sinal do detalhe, indicado pelo segundo tracejado na mesma figura, corresponde à rampa de carga que ocorre no horário de ponta, que é um comportamento normal do SEP estudado.

Por outro lado, podemos verificar na Figura 3.3 que a curva original apresenta falhas, conforme segue.

Figura 3.3 – Curva original com falhas e detalhe no nível 1 do banco de filtros da TWD.



Fonte: Elaborada pelo autor.

Na Figura 3.3 fica evidenciado que as falhas originam elevadas amplitudes, no sinal do detalhe no nível 1 do banco de filtros da TWD. Observa-se que mesmo a anormalidade instantânea mais branda, que ocorreu na amostra 191 da curva original, indicada pelo segundo círculo tracejado, originou uma amplitude no sinal do detalhe do nível 1 do banco de filtros da TWD de -18. Sendo assim, a amplitude do detalhe do nível 1 é um excelente indicativo de falha na curva de carga original. Logo o próximo passo é verificar qual o intervalo de amplitude a partir do qual se considera uma falha na curva original, ou seja, o limiar do detalhe. Para isso, se deve verificar quais as falhas presentes no banco de dados das curvas de carga.

A localização da subestação do objeto de estudo torna os dados suscetíveis ás mais diversas falhas, sejam oriundas de problemas no SEP ou dos diversos equipamentos conectados ao sistema de medição. Constatou-se que as falhas e seus intervalos se manifestaram na forma de anormalidades instantâneas, anormalidades temporárias, lacunas e dados repetidos.

Deve-se estabelecer qual *wavelet* mãe e coeficientes serão utilizados no banco de filtros, pois esses dois critérios indicam como o detalhe da curva de carga irá se comportar mediante uma falha presente na curva de carga original, sendo fundamental que o detalhe da TWD represente o momento da falta adequadamente.

De posse dos registros de todos os tipos e durações das falhas, do banco de dados, o próximo passo consistiu em verificar diversas *wavelet* mãe, coeficientes e limiares, para constatar a melhor configuração do filtro na presença de uma falha. Foram comparados os registros de falhas e respectivas durações, com os sinais do detalhe nível 1 das *wavelets* mãe e coeficientes: *Daubechies* 4, *Daubechies* 6, *Daubechies* 8, *Daubechies* 10, *Coiflet* 6, *Coiflet* 12,

Symlet 8, *Symlet* 16, *Beylkin* 18 e *Vaidyanathan* 24. Constatou-se que a *wavelet* mãe *Daubechies* com 4 coeficientes, foi aquela que melhor indicou a presença de falhas. Verificou-se que o registro de falhas indicados pelo banco de filtros da TWD, em sua quase totalidade, coincidiu com amplitudes do detalhe nível 1 fora do limiar com amplitudes de -6 à +6.

Por tudo isso, a *wavelet* mãe utilizada no banco filtros da TWD, que melhor representa a presença de falhas das curvas de carga, foi a *Daubechies* com 4 coeficientes, para um limiar do detalhe nível 1 de +6 a -6, sendo esse o melhor intervalo de amplitude, limiar*, para indicação de falhas persentes nas curvas de carga.

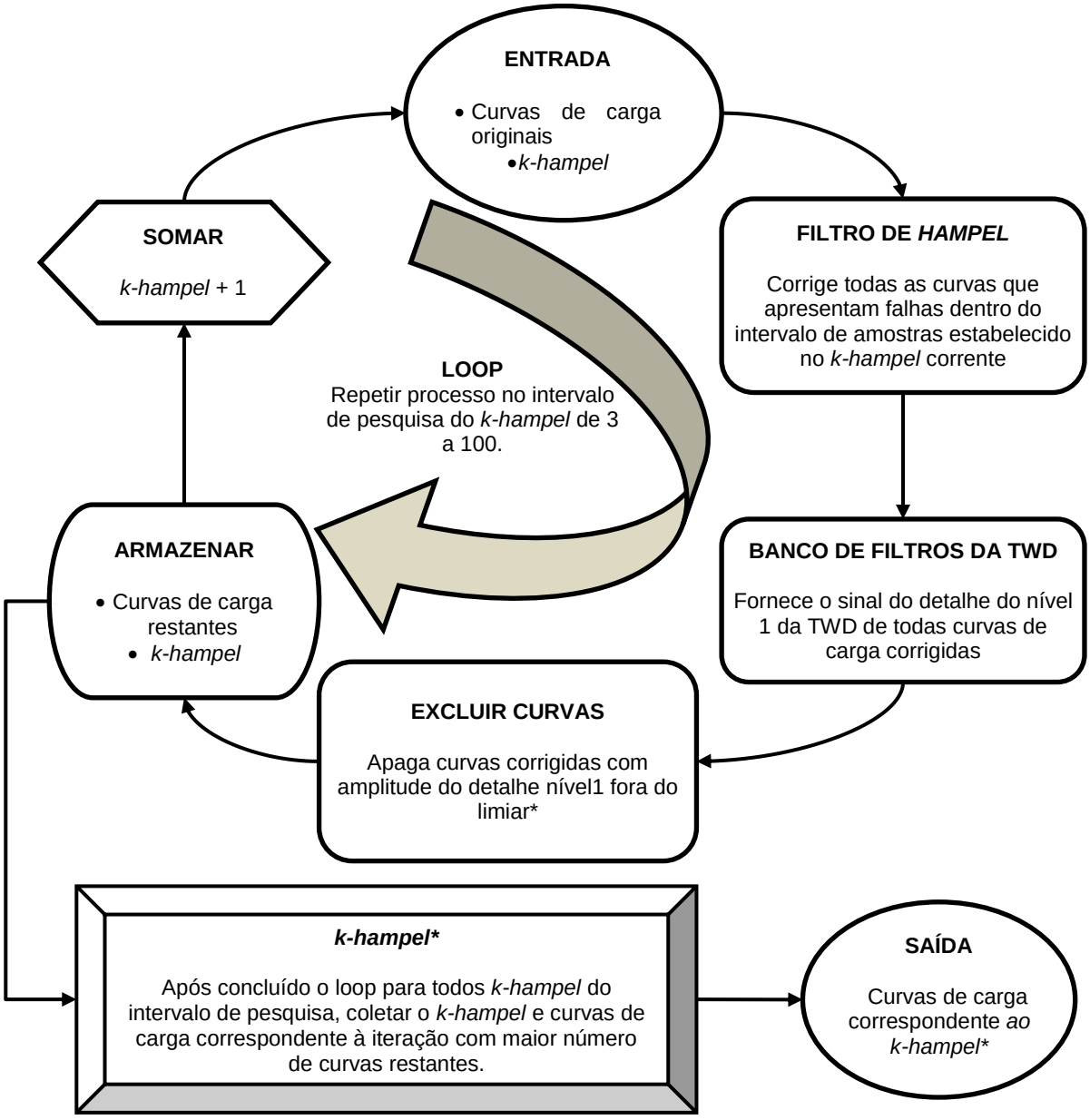
3.1.1.2 *K-hampel* ótimo, Correção e Exclusão de curvas de carga

O filtro de *hampel* é eficaz para identificar e corrigir *outliers*, de acordo com o *k-hampel* definido, porém verificou-se que para falhas de grandes intervalos de duração o filtro não é eficaz, já que se baseia na substituição de valores a partir do cálculo da mediana.

Sendo assim, para verificar o valor de *k-hampel**, foi desenvolvido um processo iterativo híbrido de correção de curvas, utilizando o filtro de *hampel* e exclusão das curvas não corrigidas, a partir do detalhe do nível 1 do banco de filtros da TWD, com referência no limiar*, Figura 3.4. O processo iterativo inicia com um *k-hampel* igual à 3, que é o mínimo para o cálculo da mediana, até o *k-hampel* de 100 amostras. Além de verificar o *k-hampel**, para esse intervalo, o processo iterativo também fornece o melhor resultado de filtragem do banco de dados com base nesse parâmetro.

Conforme Figura 3.4, todas as curvas de carga são submetidas à correção de falhas através do filtro de *hampel*, iniciando pelo intervalo inferior de pesquisa com *k-hampel* igual à três. Em seguida as curvas corrigidas são submetidas ao banco de filtros da TWD utilizando a *wavelet* mãe *Daubechies* de 4 coeficientes, sendo excluídas todas as curvas de carga que apresentarem o detalhe do nível 1 com amplitudes fora do limiar*. As curvas restantes são armazenadas, assim como o *k-hampel* correspondente da iteração corrente. Soma-se 1 ao valor do *k-hampel* e novamente reinicia-se todo o processo, até que todas as curvas de cargas restantes correspondentes a cada *k-hampel* estejam armazenadas, no *k-hampel* igual à 100. O *k-hampel* ótimo e, consequentemente, a melhor filtragem das curvas de carga, será aquele que resultar no maior número de curvas de carga restantes.

Figura 3.4 – Processo iterativo de filtragem de curvas de carga híbrido, com filtro de *hampel* e banco de filtros da TWD.



Fonte: Elaborada pelo autor.

3.2 ETAPA 2 (AGRUPAMENTO)

A segunda etapa da metodologia proposta trata do agrupamento das curvas de carga resultantes do processo de filtragem, utilizando o algoritmo *k-means*. No entanto, devido à grande dimensionalidade das curvas de carga utilizadas para realização do agrupamento, primeiramente os dados foram submetidos à redução de dimensionalidade, através da

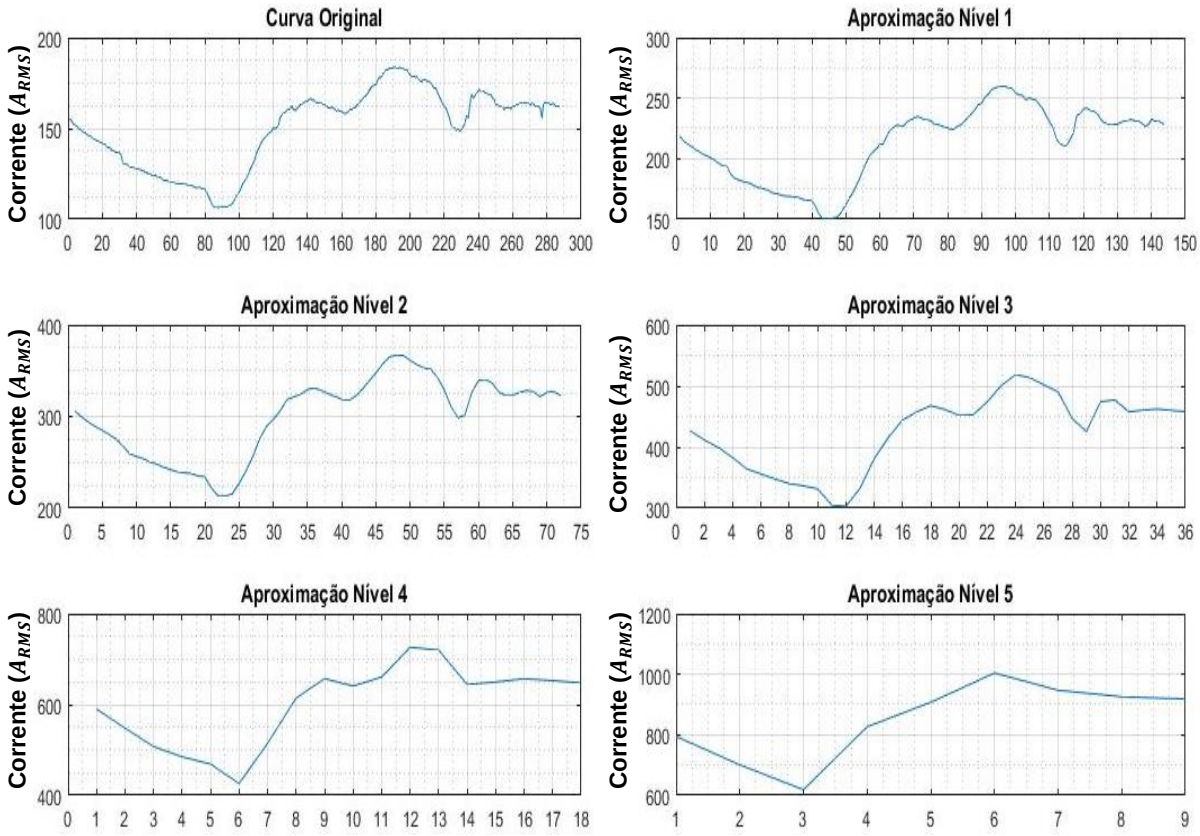
aproximação do sinal das curvas de carga utilizando o banco de filtros da TWD. Contudo, como a metodologia proposta trata da realização de agrupamentos a serem utilizados por equipes de agentes do SIN em situações reais, foi elaborado um processo de validação não supervisionado, baseado em índices internos e relativos. Esse processo de validação não necessita de modelos de agrupamentos, para a realização da verificação de eficácia.

3.2.1 Redução de dimensionalidade das curvas de carga

Conforme abordado anteriormente, séries temporais são vetores que possuem elevada dimensionalidade, tornando inadequados algoritmos de agrupamentos convencionais como o *k-means.*

Contudo, o banco de filtros da TWD produz sinais de aproximações da curva de carga original em diversos níveis. Sendo assim, utilizou-se o resultado do sinal de aproximação do banco de filtros da TWD, com *wavelet* mãe *Daubechies* de quatro coeficientes, nos diversos níveis possíveis de aproximação da curva de carga, conforme ilustrado na Figura 3.5.

Figura 3.5 – Curva de carga original e sinais de aproximação do banco de filtros da TWD nos diversos níveis.



Fonte: Elaborada pelo autor.

Na Figura 3.5 pode-se constatar que a curva de carga original possui 288 amostras, porém como o operador *downsampling* do banco de filtros da TWD realiza a divisão das amostras por 2, à medida que o nível de aproximação aumenta, o número de amostras diminui. Sendo assim, cada nível de aproximação corresponderá ao seguinte número de amostras:

- Nível 1: 144 amostras;
- Nível 2: 72 amostras;
- Nível 3: 36 amostras;
- Nível 4: 18 amostras e
- Nível 5: 9 amostras.

Por tudo isso, as amostras nos diversos níveis do banco de filtros da TWD, correspondem às características da curva de carga, já que a redução de dimensionalidade com o banco de filtros da TWD possui duas vantagens, reduzir o número de amostras da curva de carga original e obter representações distintas da mesma curva de carga com suas características mais relevantes. Logo, cada nível de aproximação do sinal das curvas de carga oriundo do banco de filtros da TWD corresponderá a um agrupamento diferente, pois expressam características diferentes das mesmas curvas de carga.

3.2.2 Agrupamento dos sinais de aproximação com *k-means*

Utilizou-se o algoritmo *k-means* para a realização do agrupamento de dados do objeto de estudo. Trata-se de um dos algoritmos de agrupamentos mais utilizado em agrupamento de séries temporais, pois é extremamente rápido, não possui complexidade computacional quadrática, sendo adequado para o grande volume de curvas de carga dessa pesquisa. O algoritmo *k-medoids* possui complexidade computacional quadrática, assim como o agrupamento hierárquico, sendo desvantajosos para o vasto banco de dados utilizado nessa pesquisa.

Apesar do algoritmo *k-means* não realizar agrupamento de dados de dimensões diferentes, esse fato não é um problema para o banco de dados de curvas de carga utilizado, já que todas as séries temporais filtradas possuem a mesma dimensão.

Anterior ao processo de agrupamento ocorre a redução de dimensionalidade das curvas de carga através do banco de filtros da TWD, conforme descrito no item 3.2.1.

Após a extração das características das curvas de carga, o próximo passo é o agrupamento das mesmas. Porém, conforme dito anteriormente, cada nível de aproximação do banco de filtros da TWD corresponde a características diferentes do mesmo conjunto de dados. Sendo assim, foram realizados agrupamentos distintos para cada nível de extração das características, iniciando da maior redução de dimensionalidade, com menos características, correspondendo a aproximação do nível 5 do banco de filtros da TWD, até a menor redução de dimensionalidade, com mais características, referindo-se a aproximação do nível 1 do banco de filtros da TWD.

Dentre as métricas de distância amplamente utilizadas em agrupamento de séries temporais, como a DTW e a ED, utilizou-se a ED devido a mesma ser simples e rápida, já que não possui complexidade computacional quadrática, como a DTW. Além do que, a ED também não necessita de parametrização, e conforme já mencionado anteriormente em Keogh e Kasetty (2003), essa métrica apresenta resultados mais favoráveis em relação a DTW, no agrupamento de séries temporais, quando comparada com conjuntos de séries temporais de tamanhos iguais, que é o caso do banco de dados do objeto de estudo.

O critério de parada do algoritmo *k-means* foi definido para no máximo 10.000 iterações.

Afim de evitar agrupamentos cujo resultados convirjam para ótimos locais, ao invés de ótimos globais, foram estabelecidas 10 replicações para o mesmo agrupamento, sendo selecionado aquele que obteve o melhor resultado da função objetivo, com menor distância intragrupo e maior distância intergrupo.

Contudo, resta o fato do *k-means* necessitar da parametrização do número de grupos k. A estratégia utilizada, para verificar o melhor número de grupos para o agrupamento do conjunto de dados pesquisado, fundamentou-se na realização de agrupamentos de um intervalo de grupos, k_2^{20}, ou seja, realizou-se variações de agrupamentos iniciando com k igual a 2 grupos até k igual a 20 grupos. Utilizou-se o critério máximo de agrupamento de 20 grupos, pois após isso verificou-se que pequenos grupos representados por comportamentos aleatórios da carga se dividem em grupos ainda menores, devido as particularidades isoladas desses comportamentos singulares da carga.

3.2.3 Validação dos agrupamentos

A validação dos agrupamentos compreende o método utilizado para verificar qual o melhor agrupamento realizado, porém como vimos anteriormente a pesquisa foi elaborada para utilização da metodologia proposta em cenários reais, que podem utilizados, dentre outras, por equipes de operação, manutenção e estudos energéticos do SEP. Sendo assim, considerou-se a situação mais complexa, ou seja, não dispor de um modelo ideal para verificação da eficácia do agrupamento realizado, correspondendo a um processo de validação não supervisionado.

Contudo, existem duas validações à serem realizadas. A primeira validação diz respeito aos agrupamentos dos k_2^{20} grupos, utilizando o mesmo nível de aproximação do banco de filtros da TWD, que diz respeito à uma validação interna e seu resultado fornece qual k grupos realiza o melhor agrupamento. A segunda validação trata dos agrupamentos que utilizaram níveis de aproximação diferentes do banco de filtros da TWD, ou seja, a partir dos melhores resultados da validação interna verifica-se qual o melhor agrupamento dentre as diferentes características extraídas do banco de dados, correspondendo ao processo de validação relativa. Sendo assim, a validação relativa fornece os dados necessários para definir o melhor agrupamento realizado.

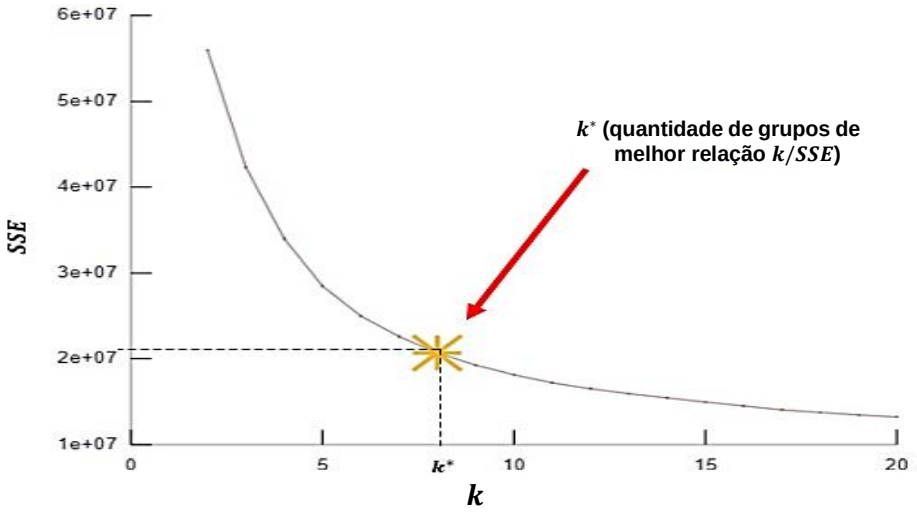
Por fim, torna-se necessário mencionar que, apesar do agrupamento ser realizado a partir das diversas características ou resoluções das curvas de carga, a validação foi realizada com as curvas de carga originais correspondentes ao resultado do agrupamento das curvas reduzidas, pois o objetivo da validação respalda-se na verificação da eficácia do agrupamento real.

3.2.3.1 Validação Interna – Ponto de saturação da curva $SSE\ vs\ k$

O índice utilizado na validação interna foi o SSE, conforme, discutido anteriormente no item 2.3.5.2, quanto menor o SSE melhor é o agrupamento. Por outro lado, a redução do SSE implica no aumento do número de grupos, fato que pode não ser satisfatório, pois segregará ainda mais pequenos grupos associados às tipologias de comportamentos aleatórios. Sendo assim, o melhor agrupamento na validação interna corresponde ao resultado da função multiobjetivo com menor SSE e menor k, ou seja, corresponde ao ponto de saturação da curva $SSE\ vs\ k$, ilustrada na Figura 3.6.

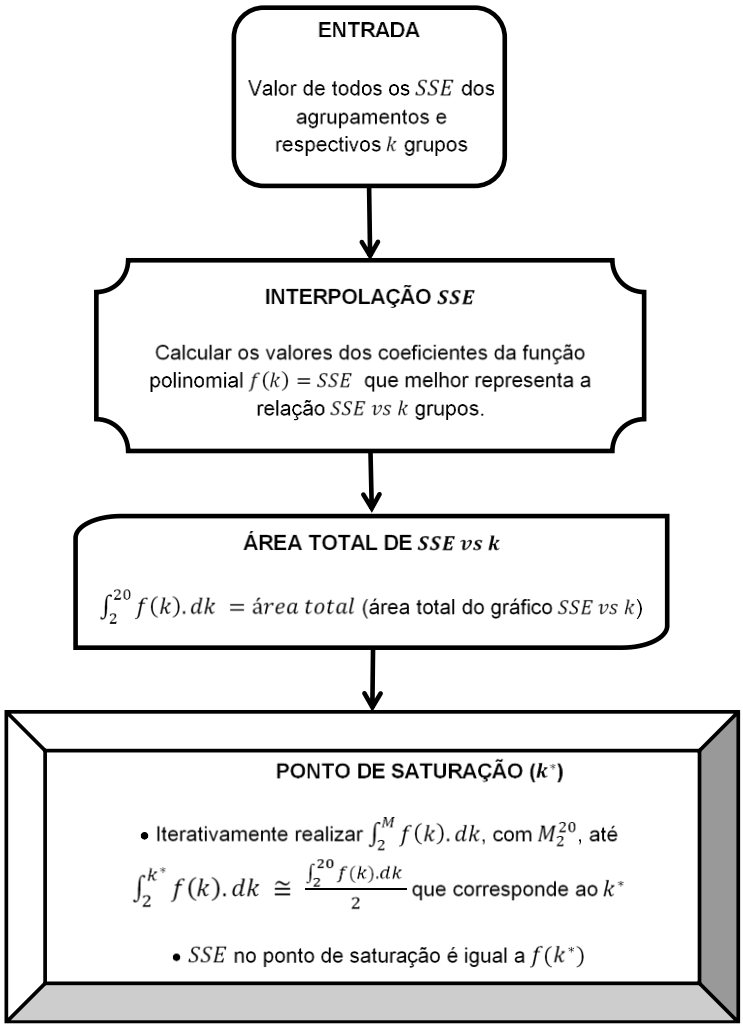
Para obter o cálculo do valor do ponto de saturação da curva da Figura 3.6, foi desenvolvido o procedimento ilustrado na Figura 3.7.

Figura 3.6 – Curva ***SSE vs k*** e ponto da melhor relação k/SSE.



Fonte: Elaborada pelo autor.

Figura 3.7 – Procedimento para localização do ponto de saturação da curva ***SSE vs k.***



Fonte: Elaborada pelo autor.

Sabendo que o gráfico da Figura 3.6 é oriundo de uma função exponencial, o procedimento descrito na Figura 3.7, para a localização do ponto de saturação (k^*, SSE), refere-se primeiramente à interpolação dos valores obtidos de SSE dos agrupamentos realizados de 2 a 20 grupos, para descobrir a função polinomial que corresponde à função exponencial $f(k) = SSE$. Verificou-se que uma função polinomial de grau sete, pode expressar adequadamente a função exponencial em questão. A partir da função polinomial obtida, o próximo passo diz respeito ao cálculo de sua integral, no intervalo de 2 a 20 grupos, $\int_2^{20} f(k).dk$, que corresponde a área total do gráfico. Por fim, o cálculo do ponto de saturação da curva consiste no processo iterativo de integração da função polinomial, para M variando de 2 a 20 grupos, $\int_2^{M} f(k).dk$. O valor de k correspondente ao ponto de saturação, k^*, será o resultado da integral anterior que estiver mais próximo da metade da área total da curva, $\int_2^{k^*} f(k).dk \cong \frac{\int_2^{20} f(k).dk}{2}$. O valor do SSE no ponto de saturação será igual ao resultado do polinômio encontrado, em função no número de grupos no ponto de saturação, $SSE = f(k^*)$. Sendo assim, a coordenada do ponto de saturação (k^*, SSE) fornece o número de grupos com melhor relação entre k e SSE, k^*, associado ao melhor agrupamento realizado com o mesmo nível de aproximação do banco de filtros da TWD da curva original.

3.2.3.2 Validação Relativa - SWC

O processo de validação relativa, diz respeito a validação dos agrupamentos oriundos do processo de validação interna. Como o banco de filtros da TWD realiza até cinco níveis de aproximações das curvas de carga originais do banco de dados do objeto de estudo, após o processo de validação interna teremos cinco agrupamentos, que representam as melhores configurações de grupos, para os respectivos níveis de aproximação da TWD. Ou seja, tratam-se dos cinco melhores agrupamentos, considerando cinco extrações de características distintas, do banco de dados do objeto de estudo.

Contudo, resta saber qual dos cinco agrupamentos, oriundos do processo de validação interna, representa a configuração mais satisfatória de grupos para todo o banco de dados. Nisso fundamenta-se o processo de validação relativa, que corresponde ao melhor agrupamento encontrado de um conjunto de agrupamentos, oriundos de extrações de características diferentes.

Como métrica para a validação relativa foi utilizado o índice da silhueta, SWC, mencionado no item 2.3.5.3, pois esse provê melhores resultados de agrupamentos, que dezenas de outros índices relativos, dentre eles os populares: índice de Davies-Bouldin e índice de Dunn.

Sendo assim, calcula-se o SWC de cada agrupamento oriundo do processo de validação interna, correspondente ao ponto de saturação em k^* de cada extração de característica do banco de dados, que por sua vez correspondem aos melhores agrupamentos dos níveis de aproximação do banco de filtros da TWD. O agrupamento mais satisfatório será aquele que resultar no maior SWC.

Ainda convém lembrar, que o resultado do maior SWC deve estar entre –1 e 1. Quanto mais próximo de 1 significa que, na média as curvas de carga originais estão melhor alocadas em seus respectivos grupos. Por outro lado, quanto mais próximo de –1 significa que, a alocação das curvas nos grupos e em todo o agrupamento está cada vez mais insatisfatória.

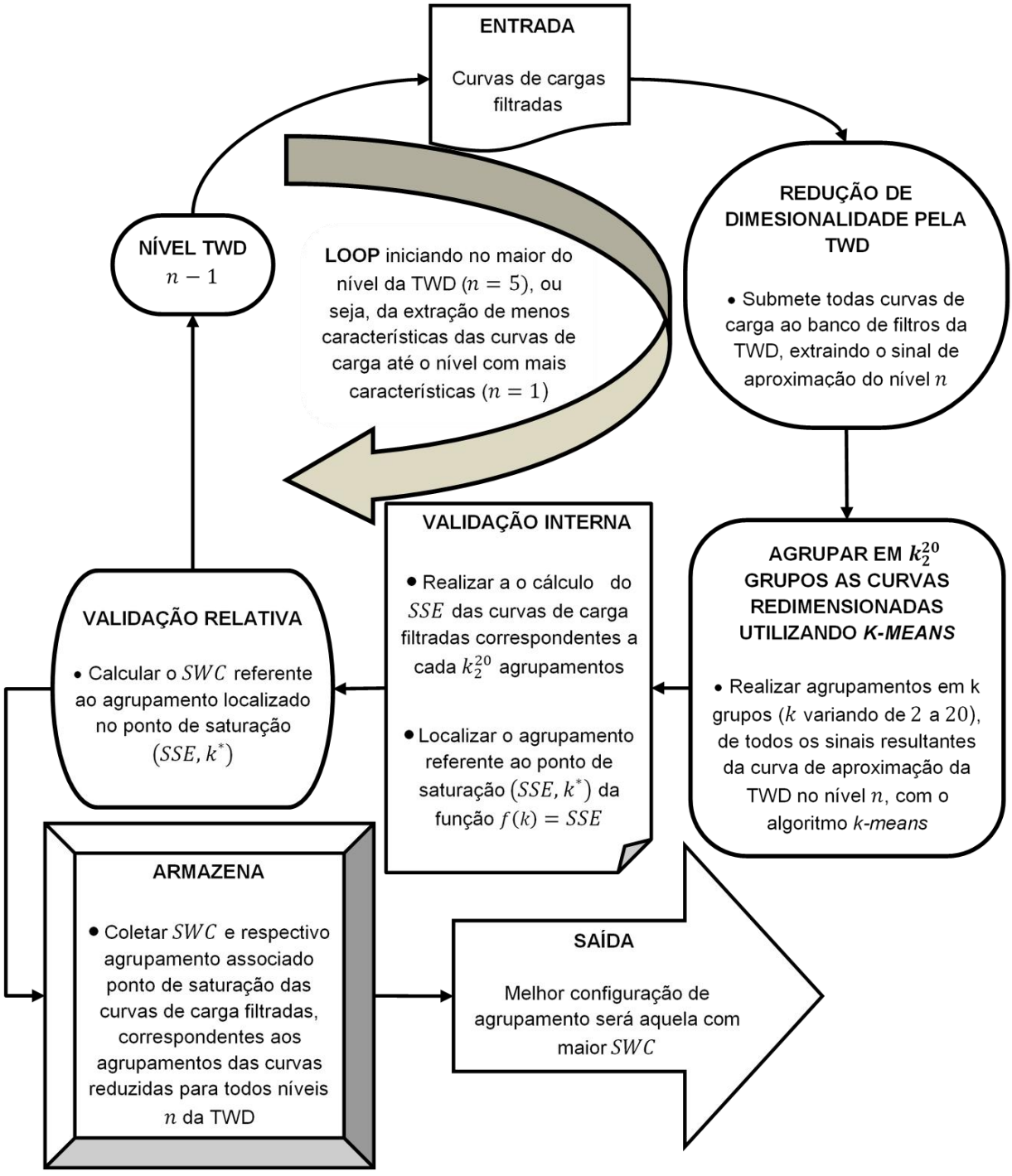
3.2.4 Processo iterativo de Agrupamento e Validação

Todos os procedimentos de agrupamento e validação descritos anteriormente, foram concentrados no processo iterativo ilustrado na Figura 3.8.

O processo iterativo de agrupamento, ilustra a proposta de agrupamento híbrido desse trabalho, utilizando banco de filtros da TWD, algoritmo de agrupamento *k-means* e o processo de validação interno e relativo, concentrados no mesmo procedimento.

O resultado final de agrupamento, indicado na saída do procedimento da Figura 3.8, fornece o resultado do melhor agrupamento das curvas de carga filtradas e agrupadas dentre os k_2^{20} grupos, entre todas as cinco características extraídas do banco de dados, que são correspondentes aos agrupamentos, das curvas reduzidas, realizados para as diferentes características extraídas da curva de aproximação do banco de filtros da TWD nos seus diferentes níveis.

Figura 3.8 – Procedimento iterativo para agrupamento e validação.



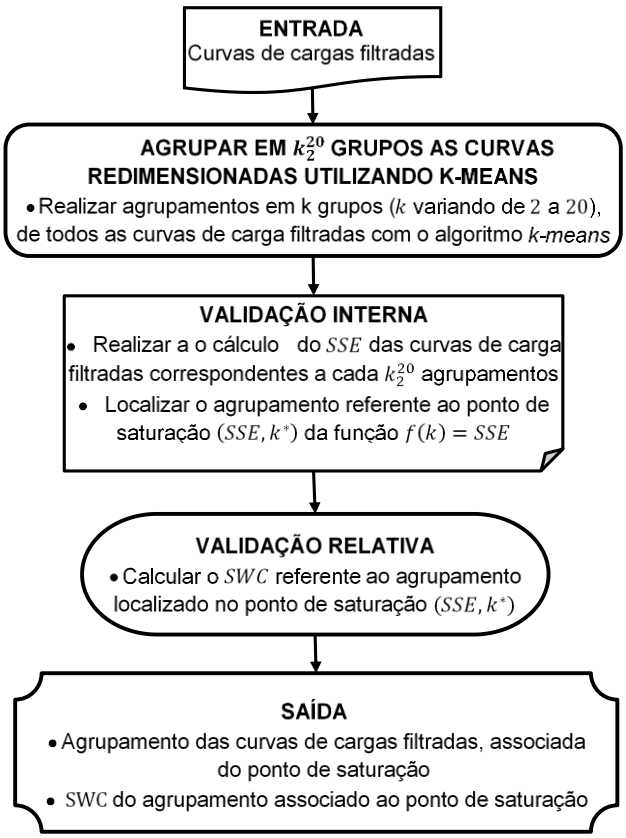
Fonte: Elaborada pelo autor.

3.3 VERIFICAÇÃO DE EFICÁCIA DA ETAPA DE AGRUPAMENTO DA METODOLOGIA PROPOSTA

Como a estratégia de validação da metodologia proposta utilizou de métricas de validação interna e validação relativa, que utilizam respectivamente índices internos (SSE) e índices relativos (SWC) consolidados, o ponto determinante do agrupamento dessa metodologia corresponde à redução de dimensionalidade. Esse fato ocorre devido o agrupamento proposto ser realizado considerando as extrações de características das curvas de carga. No entanto, por mais que o algoritmo de agrupamento seja consistente, a redução de dimensionalidade obtida pelos diversos níveis de aproximação das curvas de carga, oriundos do banco de filtros da TWD, fornece características diferentes de um mesmo banco de dados, resultando em agrupamentos distintos.

Verificou-se o resultado do agrupamento da metodologia proposta, em relação ao agrupamento sem redução de dimensionalidade, para constatar a eficácia da ferramenta de extração de características proposta, conforme a Figura 3.9.

Figura 3.9 – Procedimento para agrupamento sem redução de dimensionalidade.



Fonte: Elaborada pelo autor.

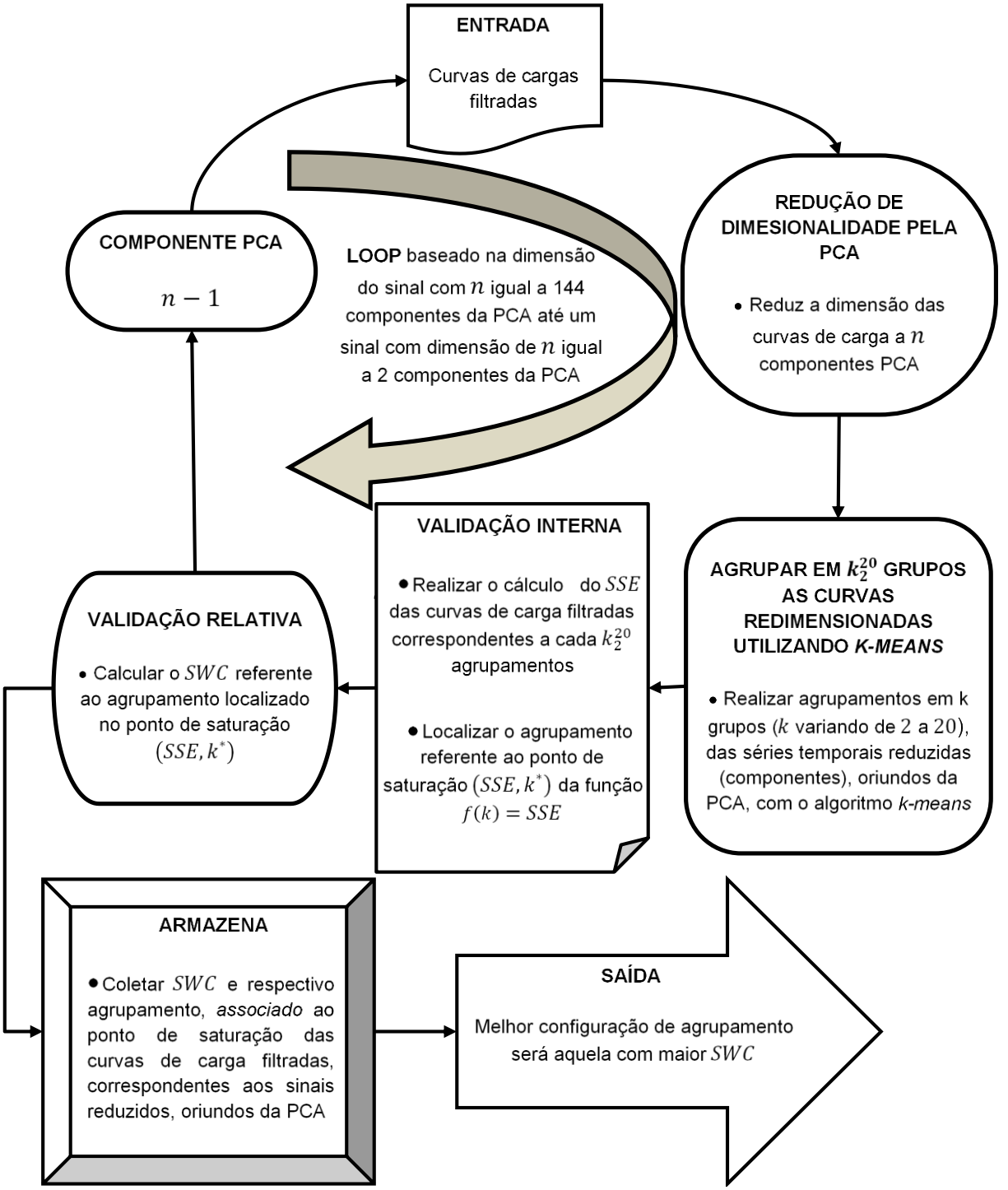
A Figura 3.9 ilustra o procedimento para o agrupamento sem o recurso de redução de dimensionalidade, fornecendo em sua saída o agrupamento das curvas de cargas filtradas, associada do ponto de saturação, assim como o SWC desse agrupamento, para posterior verificação de eficácia em relação a metodologia proposta.

Outra importante verificação foi realizada utilizando a redução de dimensionalidade utilizando a PCA, na metodologia proposta, em substituição ao banco de filtros da TWD, para realizar a análise comparativa de eficácia das duas ferramentas, conforme a Figura 3.10. A PCA trata de uma técnica típica de redução de dimensionalidade, bastante eficaz, utilizada para a redução de dimensão de séries temporais. Conforme HONGYU (2016), dentre as diversas técnicas baseadas na ideia de redução de dimensionalidade de dados com menor perda possível de informação a PCA é a mais difundida. Na redução de dimensionalidade utilizando PCA, o vetor de componentes oriundo da série temporal poderá ter qualquer dimensão, desde o mínimo de duas componentes, até o máximo de componentes igual ao número de amostras da série temporal analisada.

Para realizar uma comparação de eficácia adequada entre o banco de filtros da TWD e a PCA, o limite inferior e superior do loop no banco de filtros deveria ter o mesmo intervalo, de 9 amostras á 144 amostras. No entanto, deve-se destacar que enquanto no banco de filtros da TWD os detalhes do sinal assumem dimensões de 144 amostras no nível 1 até 9 amostras no nível 5, permitindo apenas cinco extrações de características diferentes da curva de carga, a PCA permite diversas extrações de características, com dimensões variando em uma amostra apenas.

Sendo assim, para o intervalo de extração de características do procedimento da Figura 3.10 a variação de redução de dimensionalidade da PCA pode variar em uma amostra, nas extrações de características da curva de carga, conferindo mais representações de características diferentes da mesma curva de carga em relação ao banco de filtros da TWD. Em outras palavras, a PCA pode variar de 2 a 288 amostras, a sua representação de características das curvas de carga do banco de dados. No entanto, como a intenção da extração de características traduz-se na redução de dimensionalidade, não se realizou uma comparação da dimensão da PCA maior que a dimensão extraída do banco de filtros da TWD, que corresponde à 144 amostras.

Figura 3.10 – Substituição do banco de filtros da TWD pela PCA, no procedimento iterativo para o agrupamento e validação.



Fonte: Elaborada pelo autor.

3.4 CONCLUSÕES PARCIAIS

Neste capítulo foram tratadas das estratégias e ferramentas associadas a metodologia proposta, que consiste na filtragem e agrupamento das curvas de carga do objeto de estudo.

Na filtragem foi elaborado uma metodologia híbrida composta pelo filtro de *hampel* e o banco de filtros da TWD. Já para o agrupamento elaborou-se uma metodologia híbrida composta pelo banco de filtros da TWD e o algoritmo *k-means*.

Sabendo que cada nível de aproximação do sinal das curvas de carga oriundo do banco de filtros da TWD, correspondem a um agrupamento diferente, pois expressam características diferentes das mesmas curvas de carga, foi elaborada uma estratégia e validação híbrida utilizando validação interna e relativa, que funciona dentro da metodologia de agrupamento proposta.

Foi elaborado um procedimento para utilização da metodologia sem redução de dimensionalidade, a fim de constatar se de fato a redução de dimensionalidade da metodologia proposta confere as vantagens esperadas, no resultado do agrupamento final.

Para consolidar a metodologia proposta e comparar sua eficácia, foi elaborado um procedimento considerando a PCA como ferramenta de redução de dimensionalidade, já que este é uma ferramenta de redução de dimensionalidade consolidada, além de fornecer maiores extração de características de uma mesma série temporal, em relação ao sinal do detalhe do banco de filtros da TWD.

No próximo capítulo serão apresentados os resultados da metodologia proposta nas etapas de filtragem, agrupamento e ferramentas associadas.

4. APRESENTAÇÃO DOS RESULTADOS

A apresentação dos resultados consiste na verificação do desempenho de duas etapas centrais da metodologia adotada, que são a filtragem e o agrupamento de dados. Dentro de cada uma dessas etapas também são abordados os resultados das ferramentas utilizadas, como o filtro de *hampel* e o banco de filtros da TWD, na etapa de filtragem, assim como, o banco de filtros da TWD, o algoritmo *k-means* e a metodologia de validação, na etapa de agrupamento.

4.1 SISTEMA ELÉTRICO ADOTADO

Para contribuir com a contextualização dos resultados da carga estudada, será exposto o cenário regional e local, associado ao sistema elétrico referido.

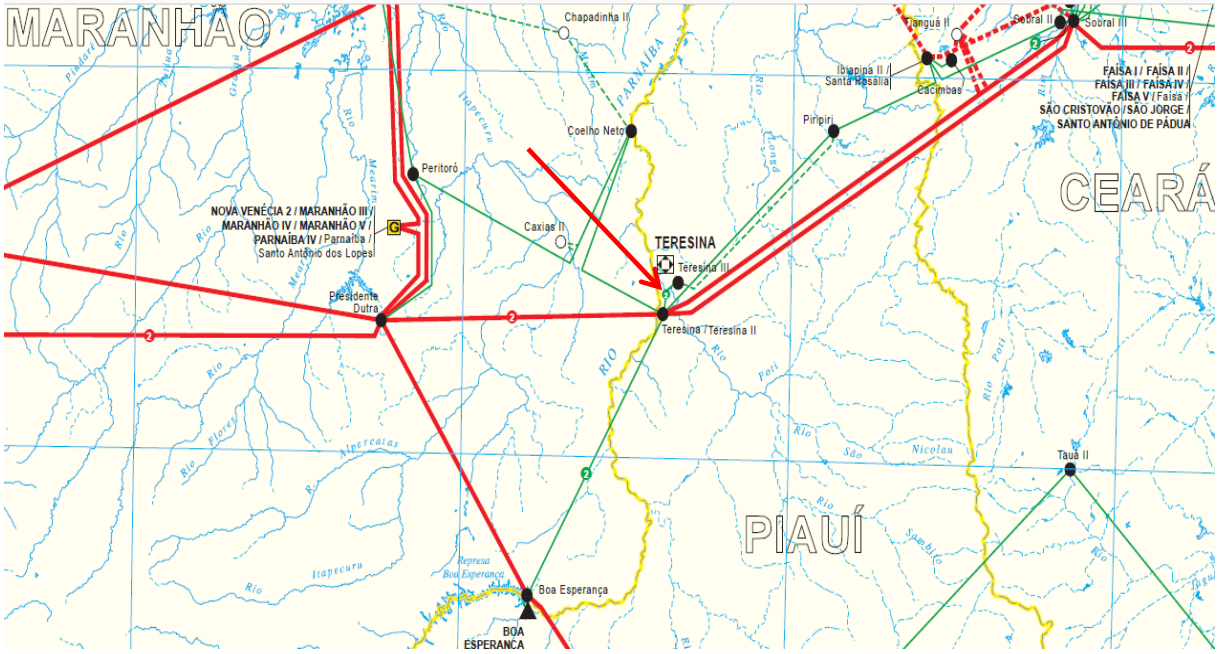
O sistema elétrico associado ao objeto de estudo trata-se de um sistema real.

Optou-se pelo estudo da corrente elétrica monitorada num transformador de potência, devido esse ser um dos equipamentos mais importantes no aspecto de suprimento de cargas numa subestação. Definiu-se como critérios de seleção do transformador, sua similaridade de configuração em relação a maioria das instalações, viabilizando a utilização da metodologia utilizada nesse trabalho em outras subestações.

Convencionou-se a referência de carga, como sendo a magnitude da corrente elétrica RMS do lado primário, de um dos transformadores de potência de 100MVA, entre os barramentos de 230KV e 69KV da subestação Teresina (TSA), da Companhia Hidrelétrica do São Francisco (CHESF), que é uma subestação da Rede de Operação[4], Figura 4.1. Esse transformador está, normalmente, em paralelo à outros quatro transformadores de igual potência.

[4] Rede de Operação: união da Rede Básica, da Rede Complementar, das usinas despachadas centralizadamente (usinas classificadas na modalidade de operação como Tipo I ou Tipo II-A, conforme critérios e sistemática estabelecidos no Módulo 26: Modalidade de operação de usinas) e das instalações de transmissão de energia elétrica destinadas a interligações internacionais conectadas à Rede Básica. Rede Básica: instalações de transmissão pertencentes ao SIN, classificadas segundo regras e condições estabelecidas pela Agência Nacional de Energia Elétrica – ANEEL. Rede Complementar: rede fora dos limites da Rede Básica, cuja operação afeta a otimização energética do SIN ou os parâmetros de avaliação do desempenho elétrico em instalações e equipamentos da Rede Básica, que levem a condições operativas fora dos critérios estabelecidos nos Procedimentos de Rede (ONS, 2017).

Figura 4.1 – Subestação Teresina no contexto regional.



Adaptada de: ONS (2018b).

A subestação TSA possui o maior número de pontos de conexão para as concessionárias de distribuição, no estado do Piauí. No setor de 13,8KV, derivam alimentadores que estão conectados diretamente à rede de distribuição da cidade de Teresina-PI. O setor de 69 KV, provê fornecimento de energia elétrica de grande parte do estado do Piauí e cidades do Maranhão. Do setor de 230KV, derivam linhas de transmissão para as subestações de Boa Esperança (BEA), Teresina Dois (TSD), Piripiri (PRI) e Coelho Neto (CNO). Possui recursos de regulação através de bancos de capacitores em todos os níveis de tensão e reatores nos setores de 13,8 KV e 230 KV (ONS, 2018a).

Por tudo isso, a escolha dos dados da corrente elétrica do transformador referido, é uma escolha estratégica, já que esse transformador supre carga de concessionárias de distribuição dos estados do Piauí e Maranhão, com as mais diversas tipologias. Além do que, está localizado em um ponto do SEP susceptível a diversos intempéries e peculiaridades, pois está conectado próximo à pontos de interligações regionais do SIN, conforme Figura 4.1. Isso promove uma verificação da robustez da metodologia proposta.

4.2 RESULTADOS DA ETAPA DE FILTRAGEM

Conforme a Figura 3.4, o procedimento iterativo da etapa de filtragem é composto por correção de curvas de carga, com filtro de *hampel*, detecção de curvas não corrigidas, com

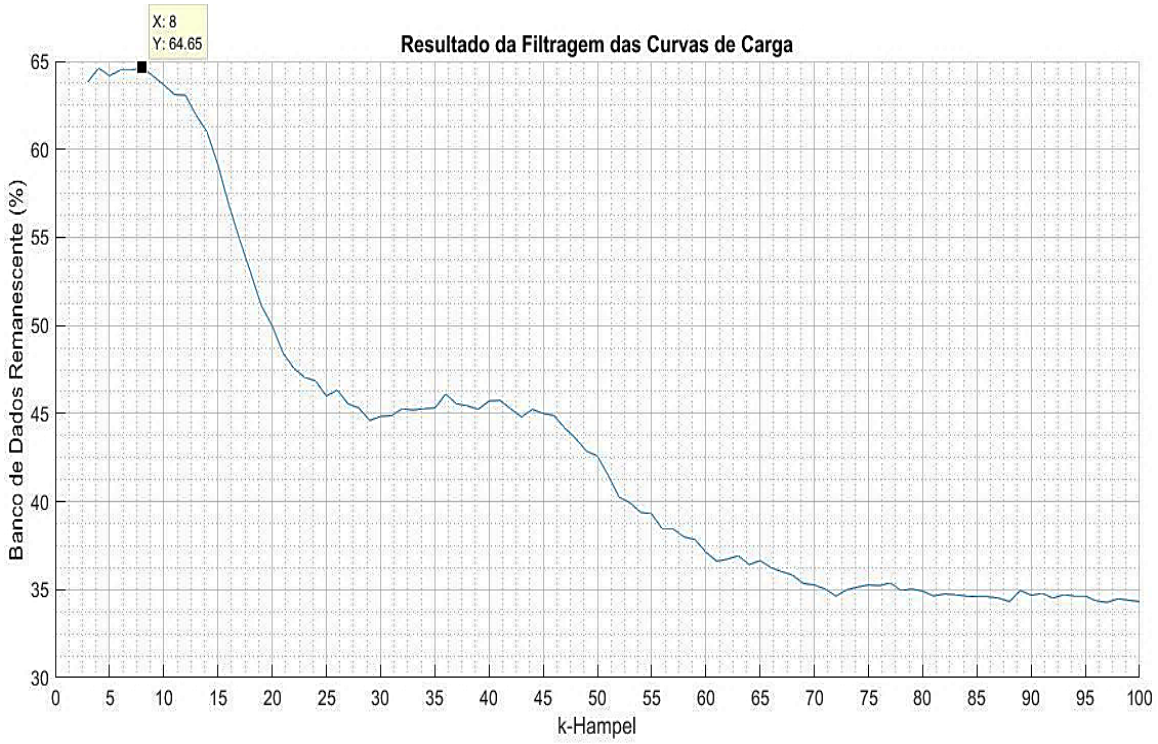
banco de filtros da TWD e posterior exclusão das mesmas. Sendo assim, serão abordados os resultados do processo iterativo para a correção e exclusão de curvas de carga com falhas.

4.2.1 Resultado do processo iterativo de filtragem das curvas de carga

No item 3.1 foi abordado que a melhor iteração do processo de filtragem será aquela que corrigir mais curvas e excluir menos curvas de carga, ou seja, a melhor iteração é aquela que, após o processo de filtragem, resultar no maior número possível de curvas de carga filtradas no banco de dados. Contudo essa condição de otimalidade coincide com o *k-hampel**, que é o fator de correção do filtro de *hampel*. Vale lembrar, que o fator de exclusão para as curvas de carga, foi definido como um valor de amplitude do sinal de detalhe do nível 1, das curvas de carga submetidas ao banco de filtros da TWD, fora do intervalo de +6 a -6.

Sendo assim, podemos observar na Figura 4.2, o resultado das curvas de cargas remanescentes, corrigidas e filtradas, do banco de dados do objeto de estudo, à medida que ocorrem iterações de *k-hampel* no intervalo de 3 a 100.

Figura 4.2 – Gráfico do resultado do banco de dados percentual remanescente após filtragem.



Fonte: Elaborada pelo autor.

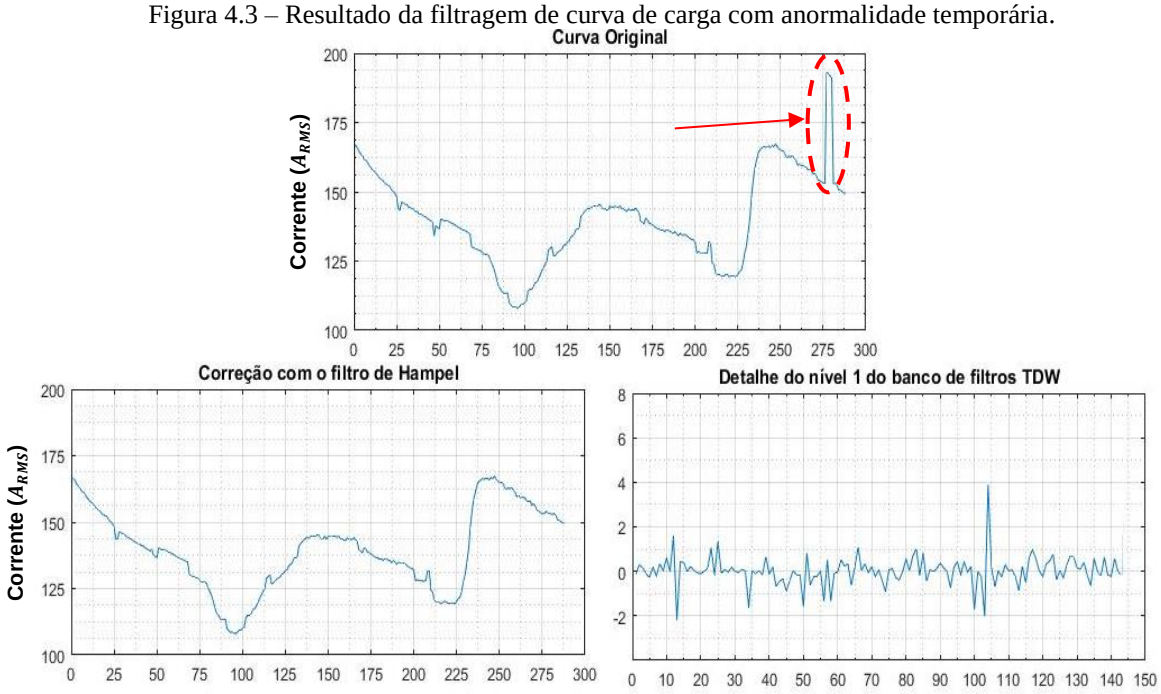
A Figura 4.2 ilustra o resultado de cada iteração do processo de filtragem, no intervalo de 3 a 100 *k-hampel* amostras. Observa-se que o processo de filtragem obtém os melhores

resultados de correção nas iterações iniciais, quando o *k-hampel* varia de três a oito. No entanto, a partir de um valor de *k-hampel* igual a nove amostras, verifica-se uma redução vertiginosa do banco de dados remanescente, à medida que o valor do *k-hampel* aumenta. Isso indica que o banco de filtros da TWD detectou muitas curvas de carga fora do limiar*, no detalhe do nível 1, que não foram corrigidas adequadamente pelo filtro de *hampel*, sendo excluídas sucessivamente.

Sendo assim, o melhor processo de filtragem corresponde a um menor percentual de exclusão de curvas de carga, do banco de dados, em função da detecção de falhas pelo banco de filtros da TWD, significando uma maior correção de curvas de carga através do filtro de *hampel*. Portanto, o melhor resultado de filtragem refere-se às curvas de carga corrigidas com k-hampel* igual a oito amostras, que por sua vez corresponde a um total de 1673 curvas de carga, ou seja, aproximadamente 64,65% da quantidade de curvas do banco de dados original.

Podemos verificar a eficácia do processo iterativo de filtragem na Figura 4.3. Nessa figura a curva de carga original possui uma anormalidade temporária equivalente a cinco amostras consecutivas. Como o *k-hampel** corresponde a oito amostras, a anormalidade referida encontra-se dentro da janela de correção do filtro de *hampel*, evidenciada pela correção da curva de carga com filtro de *hampel*. No detalhe do nível 1do banco de filtros da TWD, pode-se constatar que não há amplitudes do sinal do detalhe fora do limiar*, [6,-6]. Isso indica que após a correção da curva de carga com o filtro de *hampel*, todas as falhas foram corrigidas.

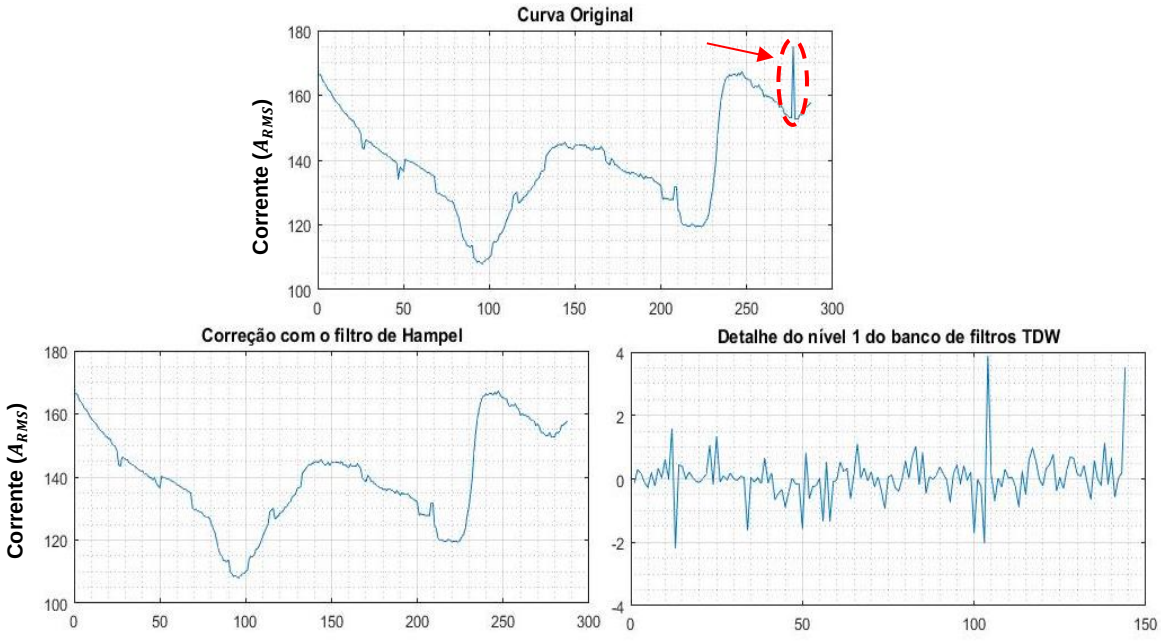
Figura 4.3 – Resultado da filtragem de curva de carga com anormalidade temporária.



Fonte: Elaborada pelo autor.

Outro exemplo do resultado do processo iterativo de filtragem está ilustrado na Figura 4.4. Nessa figura está indicada a ocorrência de uma anormalidade instantânea na curva de carga original. Como o *k-hampel** corresponde a oito amostras, qualquer anormalidade instantânea estará dentro da janela de correção do filtro de *hampel*. Na correção da curva de carga com filtro de *hampel*, verifica-se que a anormalidade instantânea foi detectada e corrigida.

Figura 4.4 – Resultado da filtragem de curva de carga com anormalidade instantânea.



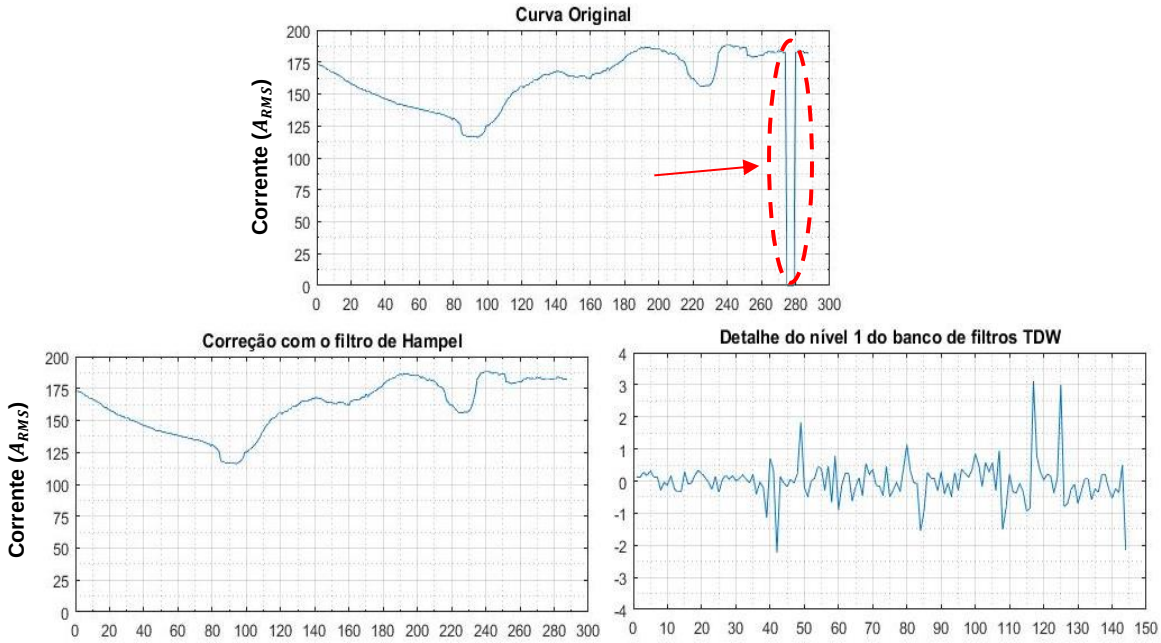
Fonte: Elaborada pelo autor.

Ainda na Figura 4.4, não são constatados valores do sinal no detalhe do nível 1 fora do limiar*, resultando em uma curva de carga tratada e pronta para ser utilizada na etapa de agrupamento da metodologia proposta.

Na Figura 4.5, verifica-se a ocorrência de uma lacuna equivalente a cinco amostras consecutivas na curva de carga original. A lacuna referida está dentro da janela de correção do *k-hampel**, fato demonstrado pelo resultado da correção da curva de carga com filtro de *hampel*. Não se constatam valores de amplitude do sinal no detalhe do nível 1 fora do limiar*, na Figura 4.5, correspondendo à uma correção satisfatória da curva de carga.

Contudo, conforme verificado na Figura 4.2, ocorreram exclusões de curvas de carga do banco de dados, após serem submetidas à correção de falhas com o filtro de *hampel*, sugerindo que nem todas as falhas foram corrigidas satisfatoriamente, conforme pode-se observar no resultado da filtragem da Figura 4.6.

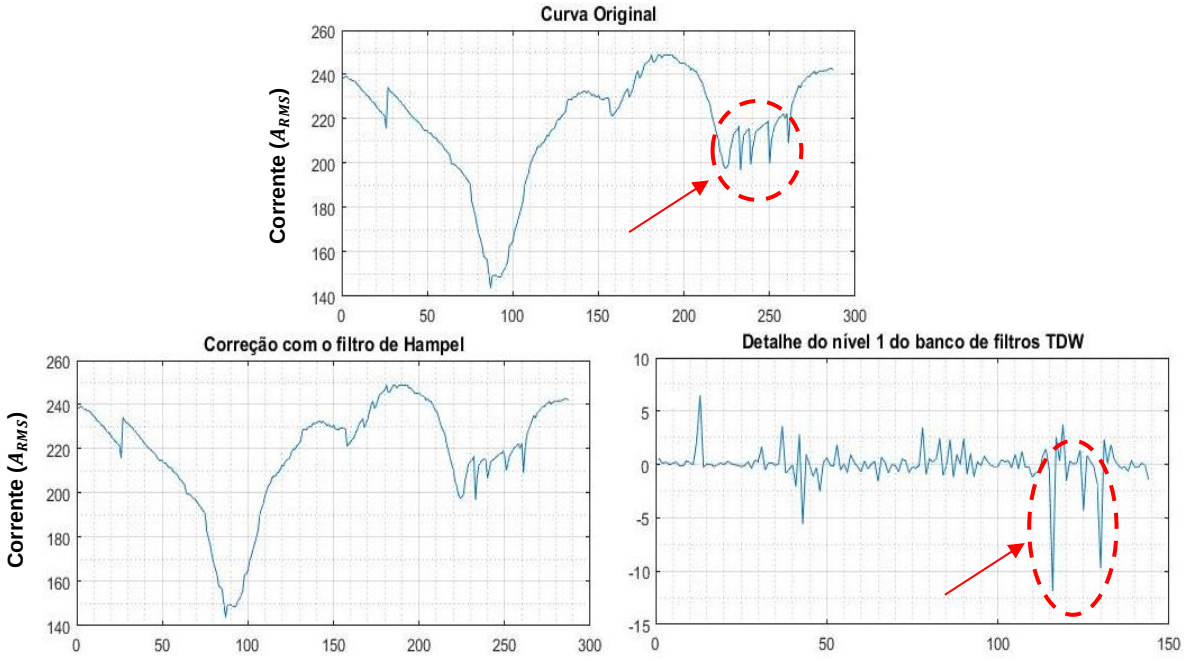
Figura 4.5 – Resultado da filtragem de curva de carga com lacuna.



Fonte: Elaborada pelo autor.

Verifica-se na Figura 4.6, a ocorrência de uma anormalidade temporária na curva de carga original, com um período equivalente a 32 amostras consecutivas ou 02:40h de duração. Essa ocorrência está fora da janela de correção, pois a curva de carga original é praticamente a mesma, após ser submetida ao filtro de *hampel*, significando que os dados não foram corrigidos satisfatoriamente.

Figura 4.6 – Resultado da filtragem de curva de carga com anormalidade temporária.



Fonte: Elaborada pelo autor.

A anormalidade temporária na curva de carga original, da Figura 4.6, está fora da janela de correção, pois observa-se que após ser submetida ao filtro de *hampel* a curva de carga continua com as oscilações indicadas na curva original. Por esse motivo, o sinal do detalhe do nível 1 do banco de filtros da TWD, apresenta amplitudes do sinal inferiores à -6, significando que os dados não foram corrigidos satisfatoriamente pelo filtro de *hampel*.

Sendo assim, por mais que o filtro de *hampel* não realize uma correção satisfatória de curvas de carga com falhas, o banco de filtros da TWD detecta a falha indicando qual curva de carga deve ser excluída do banco de dados, conferindo robustez à metodologia de filtragem proposta.

4.3 RESULTADOS DA ETAPA DE AGRUPAMENTO

No resultado da etapa de agrupamento foi avaliado o desempenho da metodologia proposta para o agrupamento das curvas de carga filtradas. Na verificação de eficácia realizou-se o agrupamento das curvas de carga filtradas sem extração de características e utilizando PCA. Para isso, foram registrados os resultados do SSE e respectivo k^*, assim como o SWC e o tempo de processamento. Esse tempo refere-se ao período compreendido desde o início da etapa de agrupamento até a validação.

4.3.1 Resultado do agrupamento das curvas de carga filtradas da metodologia proposta e demais procedimentos de comparação

O processo iterativo da Figura 3.8, representa as tarefas de agrupamento e validação da metodologia proposta. O maior SWC dentre as diversas tipologias corresponde ao melhor agrupamento resultante desse processo iterativo, conforme pode-se verificar na Tabela 4.1 e Figura 4.7. O melhor agrupamento resultante corresponde ao agrupamento associado à redução de dimensionalidade no nível 4 do banco de filtro da TWD.

Vale lembrar que a redução de dimensionalidade através do banco de filtros da TWD representa extrações de características distintas, traduzindo-se em diferentes representações do mesmo banco de dados. Sendo assim, pode-se concluir que a melhor extração de características dentre as cinco possíveis, pelo banco de filtro da TWD, corresponde a aproximação do nível 4, que por sua vez possui o maior SWC (0,2479175).

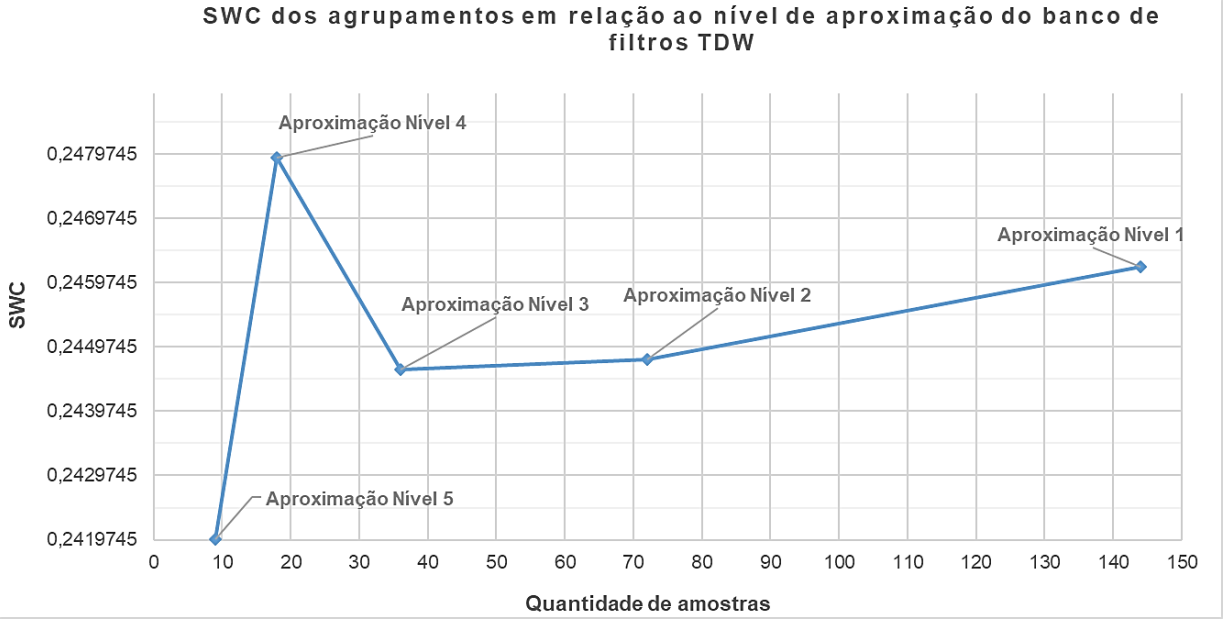
Tabela 4.1 – Resultado do agrupamento e validação através do processo iterativo da metodologia proposta.

Referencial de agrupamento	Discretização	k^*	SSE	SWC	Tempo (s)
Aproximação Nível 5 da TWD	9 amostras	8	41878464,16	0,2419845	36,58636
Aproximação Nível 4 da TWD	**18 amostras**	**8**	**41870823,17**	**0,2479175**	**36,08893**
Aproximação Nível 3 da TWD	36 amostras	8	41862073,82	0,2446269	36,09320
Aproximação Nível 2 da TWD	72 amostras	8	41855167,98	0,2447829	36,62774
Aproximação Nível 1 da TWD	144 amostras	8	41899534,33	0,2462119	37,48785

Fonte: Elaborada pelo autor.

Pode ser constatado na Tabela 4.1, que o melhor agrupamento corresponde a oito grupos em todas as extrações de características. Outro resultado importante do melhor agrupamento, refere-se ao fato de que o mesmo foi obtido, com uma redução de dimensionalidade de 288 amostras, na curva original filtrada, para 18 amostras, no sinal da aproximação do nível 4 da TWD, que corresponde a 6,25% da dimensão do sinal original e o melhor tempo de processamento, 36,08893 segundos.

Figura 4.7 – Resultado do SWC dos agrupamentos com base na aproximação do banco de filtros da TWD.



Fonte: Elaborada pelo autor.

A Figura 4.7 ilustra um comportamento peculiar associado ao resultado da validação relativa dos agrupamentos das curvas de carga, SWC. Verifica-se a redução do índice de validação relativa, dos agrupamentos baseados na aproximação do banco de filtros, do nível 1 da TWD, maior extração de características, ao nível 3 da TWD, menor extração de

características, indicando uma melhor qualidade em função do maior número de características. No entanto, o agrupamento baseado no nível 4 de aproximação da TWD, que possui menos características que os demais citados anteriormente, ocorre o maior SWC, indicando o melhor agrupamento e, consequentemente, a melhor extração de características. Contudo, o resultado da aproximação do nível 5 indica uma grande redução do SWC, indicando que as características extraídas das curvas de carga não foram suficientes, em quantidade ou qualidade para representar o banco de dados do objeto de estudo.

4.3.2 Resultado do agrupamento das curvas de carga filtradas sem redução de dimensionalidade

Realizou-se o agrupamento do banco de dados do objeto de estudo, sem utilizar o recurso da redução de dimensionalidade, conforme o procedimento da Figura 3.9. O resultado do SWC do agrupamento resultante segue na Tabela 4.2.

Tabela 4.2 – Resultado do agrupamento e validação utilizando as curvas de carga filtradas sem redução de dimensionalidade.

Referencial de agrupamento	Discretização	k^*	SSE	SWC	Tempo (s)
Curva de Carga Sem Redução	288 amostras	8	41899534,33	0,2462119	38,13740

Fonte: Elaborada pelo autor.

Observa-se que o número ótimo de grupos, k^*, da Tabela 4.2, é o mesmo da Tabela 4.1, ou seja, oito grupos. Esse fato confirma que a extração de características do banco de filtros da TWD na metodologia proposta, independentemente do nível de aproximação, não perdeu informações necessárias que caracterizam as tipologias existentes no banco de dados do objeto de estudo.

Tanto o SWC resultante do agrupamento sem redução de dimensionalidade, da Tabela 4.2, quanto o SWC do agrupamento baseado na aproximação do nível 1 da Tabela 4.1, possuem o mesmo índice da silhueta, igual à 0,2462119, significando que se obteve a mesma qualidade de agrupamento.

Somente o SWC do agrupamento baseado na aproximação nível 5, da Tabela 4.1, foi superior ao SWC resultante do agrupamento sem redução de dimensionalidade, da Tabela 4.2. Indicando melhor qualidade para o agrupamento resultante sem redução de dimensionalidade.

Contudo, os agrupamentos baseados na aproximação dos níveis 2, 3 e 4, do banco de filtros, obtiveram valores de SWC superiores ao agrupamento sem redução de dimensionalidade, com SWC máximo de 0,2479175, indicando a melhor qualidade de agrupamento obtida.

Outro fator relevante refere-se ao tempo de processamento do agrupamento sem redução de dimensionalidade, pois esse é superior a todos os tempos de processamento obtidos em função dos diversos níveis do banco de filtros da TWD, da metodologia proposta.

Sendo assim, exceto para o agrupamento baseado na aproximação do nível 5 do banco de filtros, pode-se afirmar uma maior eficácia do agrupamento das curvas de carga filtradas da metodologia proposta, em relação ao agrupamento sem redução de dimensionalidade. Essa eficácia refere-se ao índice de validação relativa de qualidade de agrupamento, SWC, e/ou ao tempo de processamento da etapa de agrupamento.

4.3.3 Resultado do agrupamento com redução de dimensionalidade através de PCA

Sabendo que a redução de dimensionalidade constitui um fator importante no resultado do agrupamento, torna-se essencial que seja verificada a eficácia do agrupamento da metodologia proposta, utilizando a redução de dimensionalidade das curvas de carga através de outra técnica consolidada na literatura. Para tanto, realizou-se o agrupamento das curvas de carga da metodologia proposta, com base em dados redimensionados através da PCA, conforme a Figura 3.10. Sendo assim, o maior SWC dentre as diversas tipologias constitui o melhor agrupamento resultante, conforme podemos verificar na Tabela 4.3 e Figura 4.8.

O melhor resultado corresponde às curvas de carga agrupadas com base no agrupamento das séries reduzidas, em 24 e 25 componentes de PCA, já que correspondem aos maiores SWC, 0,2467192, entre todos os agrupamentos realizados. No entanto, o agrupamento das séries reduzidas com 24 componentes de PCA, obtiveram tempo de processamento menor do que o agrupamento baseado em 25 componentes de PCA.

Tabela 4.3 – Resultado do agrupamento e validação utilizando redução de dimensionalidade através da PCA.

Referencial de agrupamento	Discretização	k^*	SSE	SWC	Tempo (s)
Componentes PCA	144 amostras	8	41865849	0,2403127	36,97802
Componentes PCA	143 amostras	8	41865849	0,2403127	36,23761

Referencial de agrupamento	Discretização	k^*	SSE	SWC	Tempo (s)
Componentes PCA	142 amostras	8	41865849	0,2403127	37,19206
Componentes PCA	141 amostras	8	41865849	0,2403127	36,96419
Componentes PCA	140 amostras	8	41865849	0,2403127	36,06907
Componentes PCA	139 amostras	8	41865849	0,2403127	36,20366
Componentes PCA	138 amostras	8	41865849	0,2403127	36,36304
Componentes PCA	137 amostras	8	41865849	0,2403127	36,27712
Componentes PCA	136 amostras	8	41865849	0,2403127	36,29373
Componentes PCA	135 amostras	8	41865849	0,2403127	36,44931
Componentes PCA	134 amostras	8	41865849	0,2403127	36,50051
Componentes PCA	133 amostras	8	41865849	0,2403127	35,99982
Componentes PCA	132 amostras	8	41865849	0,2403127	36,08054
Componentes PCA	131 amostras	8	41865849	0,2403127	36,18211
Componentes PCA	130 amostras	8	41865849	0,2403127	36,37853
Componentes PCA	129 amostras	8	41865849	0,2403127	36,13285
Componentes PCA	128 amostras	8	41865849	0,2403127	36,87654
Componentes PCA	127 amostras	8	41865849	0,2403127	36,29706
Componentes PCA	126 amostras	8	41865849	0,2403127	36,64001
Componentes PCA	125 amostras	8	41865849	0,2403127	36,30342
Componentes PCA	124 amostras	8	41865849	0,2403127	36,4174
Componentes PCA	123 amostras	8	41865849	0,2403127	36,28926
Componentes PCA	122 amostras	8	41865849	0,2403127	37,0864
Componentes PCA	121 amostras	8	41865849	0,2403127	37,08282
Componentes PCA	120 amostras	8	41865849	0,2403127	36,32001
Componentes PCA	119 amostras	8	41865849	0,2403127	36,08381
Componentes PCA	118 amostras	8	41865849	0,2403127	36,64317
Componentes PCA	117 amostras	8	41865849	0,2403127	37,00109
Componentes PCA	116 amostras	8	41865849	0,2403127	36,52095
Componentes PCA	115 amostras	8	41865849	0,2403127	36,38434
Componentes PCA	114 amostras	8	41865849	0,2403127	36,25439
Componentes PCA	113 amostras	8	41865849	0,2403127	36,22366
Componentes PCA	112 amostras	8	41865849	0,2403127	36,43764
Componentes PCA	111 amostras	8	41865849	0,2403127	36,80852
Componentes PCA	110 amostras	8	41865849	0,2403127	36,45601
Componentes PCA	109 amostras	8	41865849	0,2403127	36,24465
Componentes PCA	108 amostras	8	41865849	0,2403127	36,48559
Componentes PCA	107 amostras	8	41865849	0,2403127	36,82503
Componentes PCA	106 amostras	8	41865849	0,2403127	36,58263
Componentes PCA	105 amostras	8	41865849	0,2403127	36,12536
Componentes PCA	104 amostras	8	41865849	0,2403127	36,14853
Componentes PCA	103 amostras	8	41865849	0,2403127	36,13364
Componentes PCA	102 amostras	8	41865849	0,2403127	36,05522
Componentes PCA	101 amostras	8	41865849	0,2403127	36,56879
Componentes PCA	100 amostras	8	41865849	0,2403127	36,35678
Componentes PCA	99 amostras	8	41865849	0,2403127	36,32755

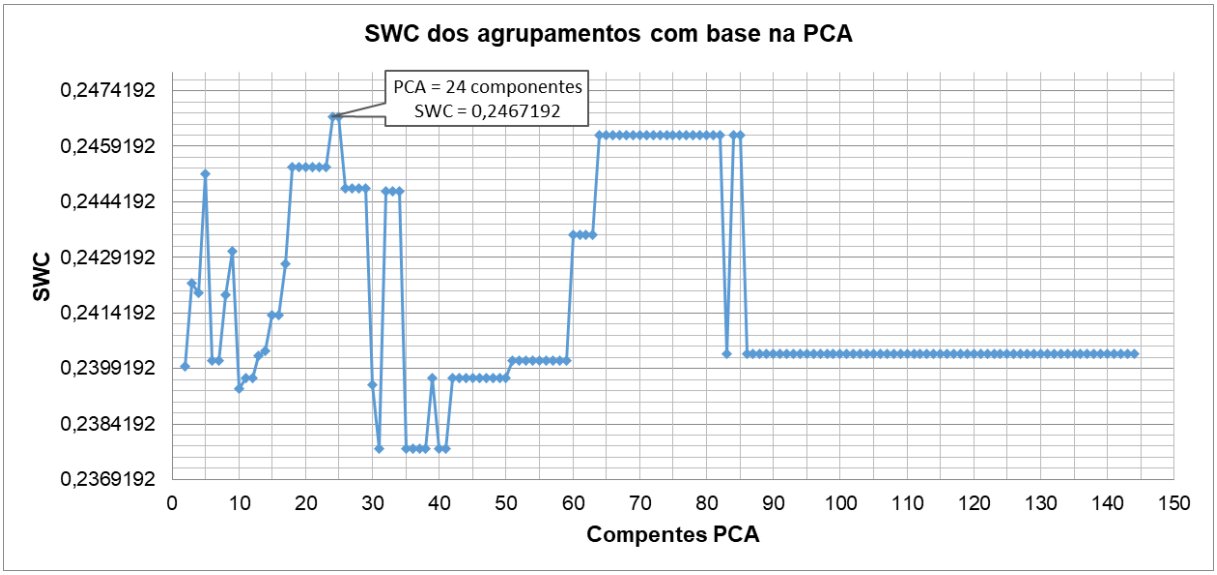
Referencial de agrupamento	Discretização	k^*	SSE	SWC	Tempo (s)
Componentes PCA	98 amostras	8	41865849	0,2403127	36,20219
Componentes PCA	97 amostras	8	41865849	0,2403127	36,45261
Componentes PCA	96 amostras	8	41865849	0,2403127	40,69142
Componentes PCA	95 amostras	8	41865849	0,2403127	44,68064
Componentes PCA	94 amostras	8	41865849	0,2403127	41,07697
Componentes PCA	93 amostras	8	41865849	0,2403127	42,26313
Componentes PCA	92 amostras	8	41865849	0,2403127	41,01474
Componentes PCA	91 amostras	8	41865849	0,2403127	47,39966
Componentes PCA	90 amostras	8	41865849	0,2403127	49,30859
Componentes PCA	89 amostras	8	41865849	0,2403127	45,86195
Componentes PCA	88 amostras	8	41865849	0,2403127	44,33004
Componentes PCA	87 amostras	8	41865849	0,2403127	62,16028
Componentes PCA	86 amostras	8	41865849	0,2403127	38,54928
Componentes PCA	85 amostras	8	41899534	0,2462119	37,07482
Componentes PCA	84 amostras	8	41899534	0,2462119	37,07396
Componentes PCA	83 amostras	8	41865849	0,2403127	36,3052
Componentes PCA	82 amostras	8	41899534	0,2462119	36,24239
Componentes PCA	81 amostras	8	41899534	0,2462119	36,1788
Componentes PCA	80 amostras	8	41899534	0,2462119	36,62171
Componentes PCA	79 amostras	8	41899534	0,2462119	36,55679
Componentes PCA	78 amostras	8	41899534	0,2462119	40,19418
Componentes PCA	77 amostras	8	41899534	0,2462119	35,92267
Componentes PCA	76 amostras	8	41899534	0,2462119	36,14691
Componentes PCA	75 amostras	8	41899534	0,2462119	36,48313
Componentes PCA	74 amostras	8	41899534	0,2462119	36,28746
Componentes PCA	73 amostras	8	41899534	0,2462119	35,95114
Componentes PCA	72 amostras	8	41899534	0,2462119	35,98566
Componentes PCA	71 amostras	8	41899534	0,2462119	36,52598
Componentes PCA	70 amostras	8	41899534	0,2462119	36,19498
Componentes PCA	69 amostras	8	41899534	0,2462119	36,19235
Componentes PCA	68 amostras	8	41899534	0,2462119	36,83339
Componentes PCA	67 amostras	8	41899534	0,2462119	36,58022
Componentes PCA	66 amostras	8	41899534	0,2462119	36,34102
Componentes PCA	65 amostras	8	41899534	0,2462119	36,54334
Componentes PCA	64 amostras	8	41899534	0,2462119	36,48046
Componentes PCA	63 amostras	8	41855947	0,2435297	36,27599
Componentes PCA	62 amostras	8	41855947	0,2435297	36,94777
Componentes PCA	61 amostras	8	41855947	0,2435297	36,12947
Componentes PCA	60 amostras	8	41855947	0,2435297	37,09755
Componentes PCA	59 amostras	8	41938910	0,2401312	36,19012
Componentes PCA	58 amostras	8	41938910	0,2401312	35,94537
Componentes PCA	57 amostras	8	41938910	0,2401312	37,15665
Componentes PCA	56 amostras	8	41938910	0,2401312	36,94852
Componentes PCA	55 amostras	8	41938910	0,2401312	36,35721

Referencial de agrupamento	Discretização	k^*	SSE	SWC	Tempo (s)
Componentes PCA	54 amostras	8	41938910	0,2401312	36,03577
Componentes PCA	53 amostras	8	41938910	0,2401312	36,59294
Componentes PCA	52 amostras	8	41938910	0,2401312	36,77419
Componentes PCA	51 amostras	8	41938910	0,2401312	36,11305
Componentes PCA	50 amostras	8	41943207	0,2396582	36,46406
Componentes PCA	49 amostras	8	41943207	0,2396582	36,10528
Componentes PCA	48 amostras	8	41943207	0,2396582	36,22252
Componentes PCA	47 amostras	8	41943207	0,2396582	36,91125
Componentes PCA	46 amostras	8	41943207	0,2396582	36,8039
Componentes PCA	45 amostras	8	41943207	0,2396582	36,59186
Componentes PCA	44 amostras	8	41943207	0,2396582	36,45802
Componentes PCA	43 amostras	8	41943207	0,2396582	36,06003
Componentes PCA	42 amostras	8	41943207	0,2396582	36,09731
Componentes PCA	41 amostras	8	41948836	0,2377535	36,09093
Componentes PCA	40 amostras	8	41948836	0,2377535	36,52029
Componentes PCA	39 amostras	8	41943207	0,2396582	36,29428
Componentes PCA	38 amostras	8	41948836	0,2377535	36,15641
Componentes PCA	37 amostras	8	41948836	0,2377535	36,15797
Componentes PCA	36 amostras	8	41948836	0,2377535	37,6164
Componentes PCA	35 amostras	8	41948836	0,2377535	37,28418
Componentes PCA	34 amostras	8	41867116	0,2447134	38,3932
Componentes PCA	33 amostras	8	41867116	0,2447134	36,28629
Componentes PCA	32 amostras	8	41867116	0,2447134	36,44835
Componentes PCA	31 amostras	8	41948836	0,2377535	36,44899
Componentes PCA	30 amostras	8	41922772	0,2394744	36,29282
Componentes PCA	29 amostras	8	41855168	0,2447829	36,38284
Componentes PCA	28 amostras	8	41855168	0,2447829	36,19456
Componentes PCA	27 amostras	8	41855168	0,2447829	36,19288
Componentes PCA	26 amostras	8	41855168	0,2447829	36,2545
Componentes PCA	25 amostras	8	41898354	0,2467192	40,67428
Componentes PCA	**24 amostras**	**8**	**41898354**	**0,2467192**	**37,80176**
Componentes PCA	23 amostras	8	41856214	0,2453622	39,87964
Componentes PCA	22 amostras	8	41856214	0,2453622	40,3753
Componentes PCA	21 amostras	8	41856214	0,2453622	36,27111
Componentes PCA	20 amostras	8	41856214	0,2453622	37,00949
Componentes PCA	19 amostras	8	41856214	0,2453622	35,9512
Componentes PCA	18 amostras	8	41856214	0,2453622	36,14003
Componentes PCA	17 amostras	8	41899261	0,2427474	36,12156
Componentes PCA	16 amostras	8	41910662	0,2413501	38,66031
Componentes PCA	15 amostras	8	41910662	0,2413501	38,94482
Componentes PCA	14 amostras	8	41938572	0,2403995	36,67663
Componentes PCA	13 amostras	8	41945203	0,2402546	46,69313
Componentes PCA	12 amostras	8	41940022	0,2396645	42,40434
Componentes PCA	11 amostras	8	41940022	0,2396645	40,90926

Referencial de agrupamento	Discretização	k^*	SSE	SWC	Tempo (s)
Componentes PCA	10 amostras	8	41923294	0,2393612	38,10522
Componentes PCA	9 amostras	8	41897659	0,2430861	38,23818
Componentes PCA	8 amostras	8	41860171	0,2418957	40,0413
Componentes PCA	7 amostras	8	41862674	0,2401368	39,79755
Componentes PCA	6 amostras	8	41862674	0,2401368	39,79755
Componentes PCA	5 amostras	8	41863837	0,2451649	36,2734
Componentes PCA	4 amostras	8	41879627	0,2419646	37,21253
Componentes PCA	3 amostras	8	41894756	0,2422342	37,20841
Componentes PCA	2 amostras	8	41913324	0,2399859	44,63586

Fonte: Elaborada pelo autor.

Figura 4.8 – Resultado do SWC dos agrupamentos com base na PCA.



Fonte: Elaborada pelo autor.

O ponto de saturação associado ao melhor agrupamento, k^*, corresponde a oito grupos em todas as diversas representações reduzidas das curvas de carga, de 144 a 2 componentes de PCA, indicando que as tipologias de curvas de carga de todo banco de dados são melhor representadas em oito grupos distintos.

O tempo de processamento do agrupamento baseado em 24 componentes de PCA foi de 37,80176 segundos, sendo superior ao tempo de processamento médio geral calculado, de 37,58117 segundos, significando que a redução de dimensionalidade utilizando a PCA não implica necessariamente no menor tempo de processamento de agrupamento.

4.3.4 Resumo dos resultados dos agrupamentos com a TWD, PCA e sem redução de dimensionalidade

Por fim, segue a Tabela 4.4 que resume os melhores resultados das Tabelas 4.1, 4.2 e 4.3, para fins de verificação de eficácia da metodologia proposta.

Tabela 4.4 – Resultado dos melhores agrupamentos, considerando dimensionalidades diferentes para verificação de eficácia da metodologia proposta.

Referencial de agrupamento	Amostras	Redução (%)	k^*	SWC	Tempo (s)
Aproximação Nível 4 da TWD	18	93,75	8	0,2479175	36,088933
Curva de Carga sem Redução	288	0,00	8	0,2462119	38,13740
Componentes PCA	24	91,67	8	0,2467192	37,80176

Fonte: Elaborada pelo autor.

Conforme a Tabela 4.4, verifica-se que a metodologia proposta resulta na melhor eficácia de agrupamento, em relação aos demais procedimentos de agrupamento com variações da estratégia de redução da dimensionalidade. A metodologia proposta resultou no maior SWC, através da menor representação de dimensionalidade, com o menor tempo de processamento, conferindo velocidade e qualidade à metodologia de agrupamento proposta.

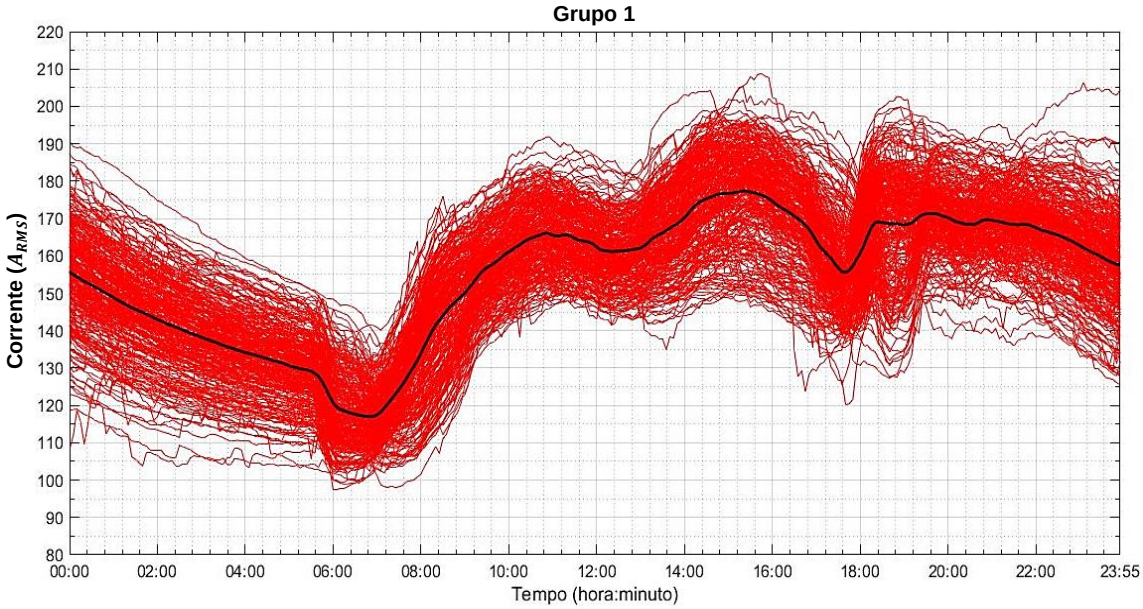
4.3.5 Tipologias dos grupos referentes ao agrupamento resultante da metodologia proposta

Na Tabela 4.1, observou-se que o melhor agrupamento da metodologia proposta é composto por 8 grupos de curvas filtradas, equivalentes aos agrupamentos realizados com as curvas reduzidas através da aproximação no nível 4, do banco de filtros da TWD. Esses grupos devem representar as tipologias mais significativas, encontradas no banco de dados do objeto de estudo.

Apresentam-se a seguir as curvas de carga de cada grupo oriundo do agrupamento resultante da metodologia proposta.

A Figura 4.9 ilustra as curvas de carga associadas ao grupo 1. As curvas desse grupo possuem certa homogeneidade em torno do seu centroide, curva preta, conferindo um agrupamento satisfatório.

Figura 4.9 – Curvas de carga resultantes do grupo 1.



Fonte: Elaborada pelo autor.

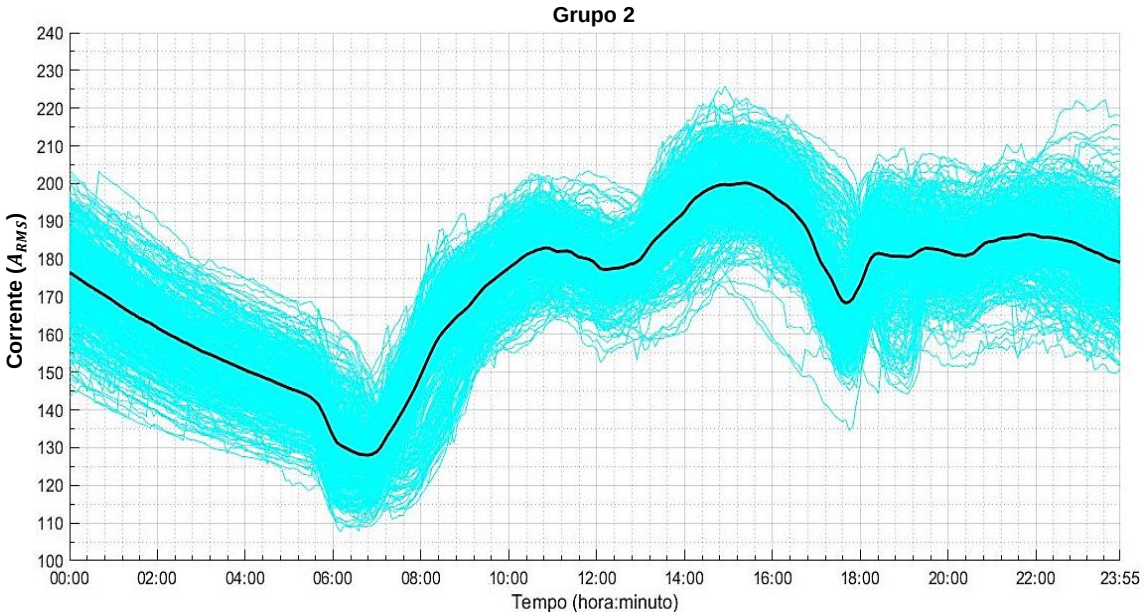
Verifica-se que o grupo 1 possui tanto as características de cargas comerciais quanto residenciais, verificados no item 2.2. Das 07:00h às 18:00h ocorrem dois picos de carga por volta das 11:00h e 15:00h, com elevada demanda no restante do período, caracterizando a predominância comercial da carga. A classe residencial torna-se evidente, no declive da curva de carga de 00:00h às 07:00h e rampa de carga de 18:00 às 21:00h, mantendo-se em um valor de elevada amplitude até cerca de 22:00h, quando inicia a redução de carga com novo declive.

Na Figura 4.10 verifica-se o comportamento das curvas de carga associadas ao grupo 2. Constata-se que as curvas de carga do grupo 2, possuem maior homogeneidade em torno do seu centroide que o grupo 1, resultando em um agrupamento bem definido.

O grupo 2 possui características de carga similares ao grupo 1, com predominância residencial de 00:00h às 07:00h e de 18:00h às 23:55h, assim como preeminência comercial das 07:00 às 18:00h.

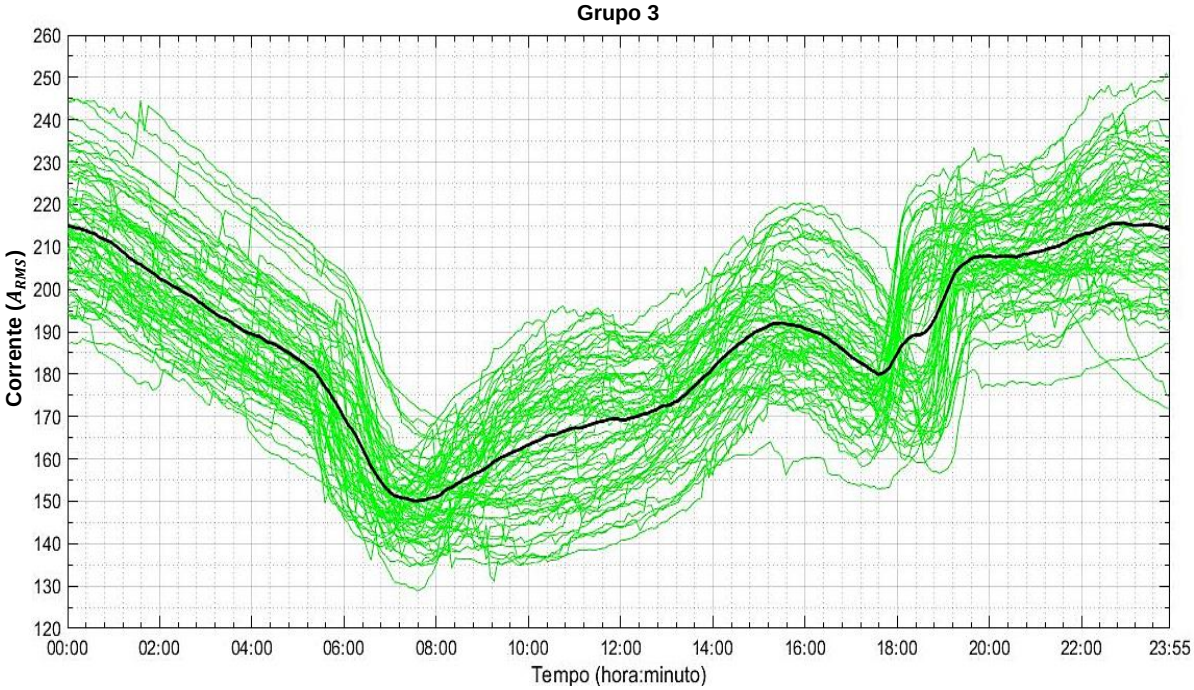
Na tipologia das curvas de carga do grupo 3, da Figura 4.11, observa-se uma redução de carga das 00:00h às 07:00h. Das 07:00h às 15:00h ocorre uma rampa, seguida de nova redução de carga até às 18:00h, prosseguindo com nova rampa de carga até meados das 20:00h, quando a carga permanece sem maiores alterações significativas até o final do dia.

Figura 4.10 – Curvas de carga resultantes do grupo 2.



Fonte: Elaborada pelo autor.

Figura 4.11 – Curvas de carga resultantes do grupo 3.



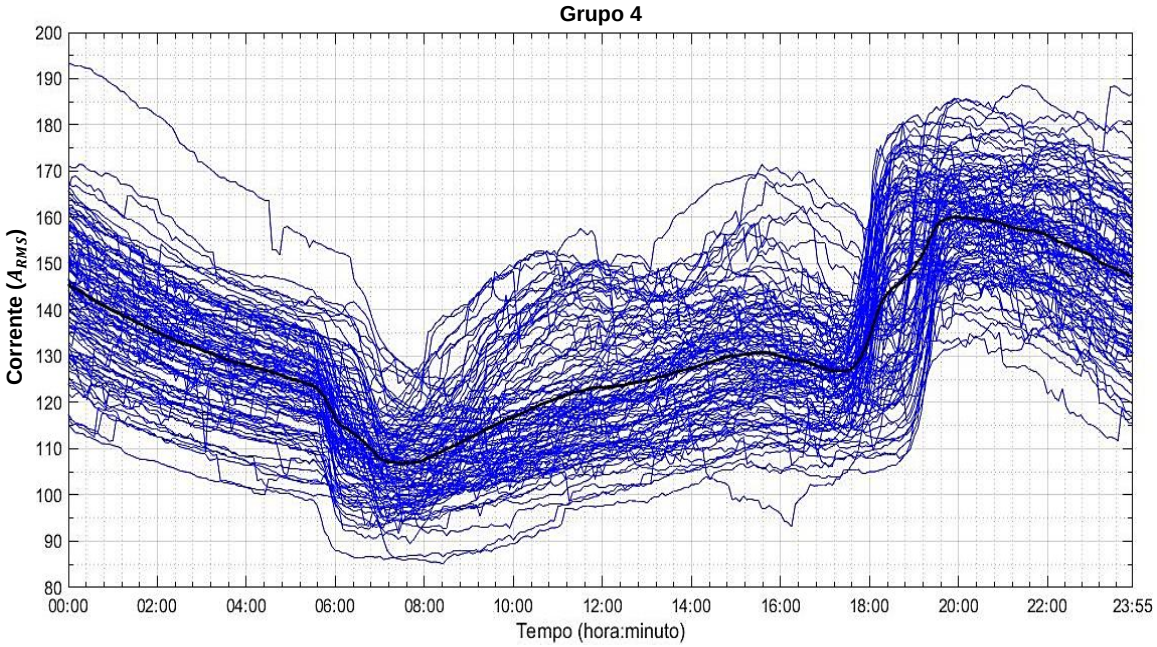
Fonte: Elaborada pelo autor.

Dentre as descrições da tipologia das curvas de carga do grupo 3, verifica-se que ocorrem dois vales de carga bem definidos, por volta das 07:00 e 18:00h, respectivamente, além do que entre esse intervalo não se verifica uma rampa de carga íngreme. Essas características

indicam que tanto cargas residenciais como iluminação pública, fazem parte das classes de carga predominantes desse grupo, com elevado consumo durante a tarde e noite.

A Figura 4.12 não ilustra um grupo de curvas de carga homogêneo como os demais, porém representa um comportamento similar em torno do seu centroide. O grupo 4 descreve características oriundas de classes de carga residenciais e iluminação pública, semelhante ao grupo 3, pois ocorrem dois vales de carga bem definidos, às 07:00 e 18:00h respectivamente, além do que não ocorrem rampas associadas a outros perfis típicos de carga nesse mesmo intervalo.

Figura 4.12 – Curvas de carga resultantes do grupo 4.

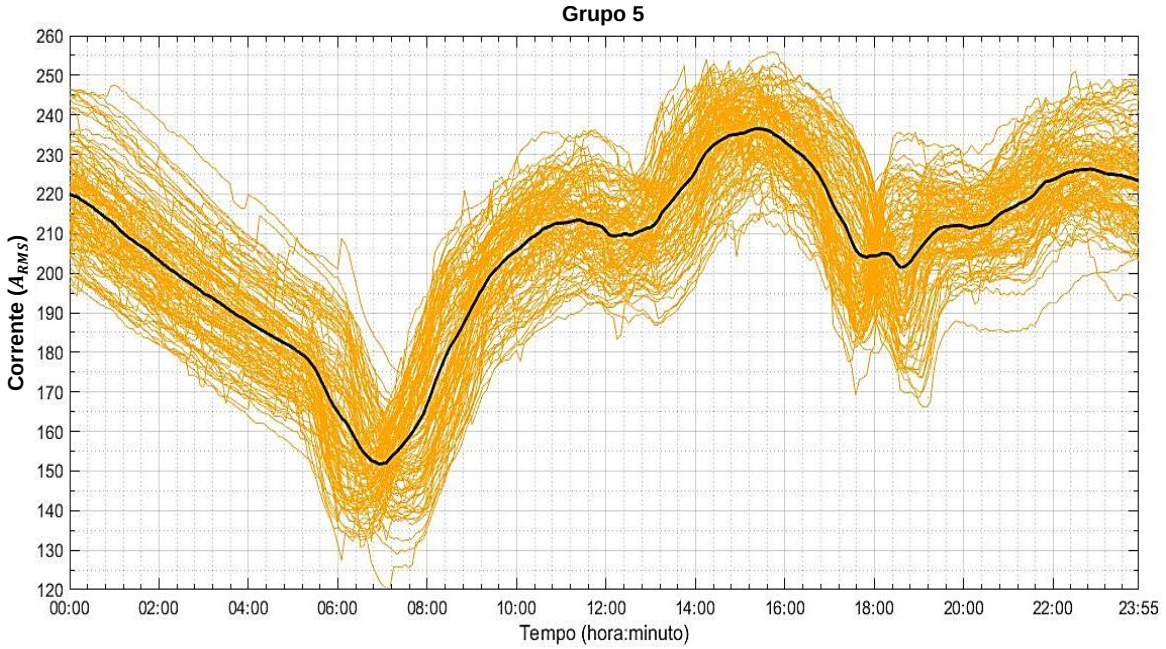


Fonte: Elaborada pelo autor.

A tipologia das curvas de carga do grupo 5 pode ser verificada na Figura 4.13. Verifica-se boa homogeneidade em torno da média, curva preta, resultando em um agrupamento bem característico.

Semelhante aos grupos 1 e 2, as características mais relevantes da tipologia apresentada no grupo 5, refere-se ao fato de que esse grupo possui características residenciais de 00:00h às 07:00h e de 18:00 às 23:55h, assim como um perfil de consumo comercial das 07:00 às 18:00h.

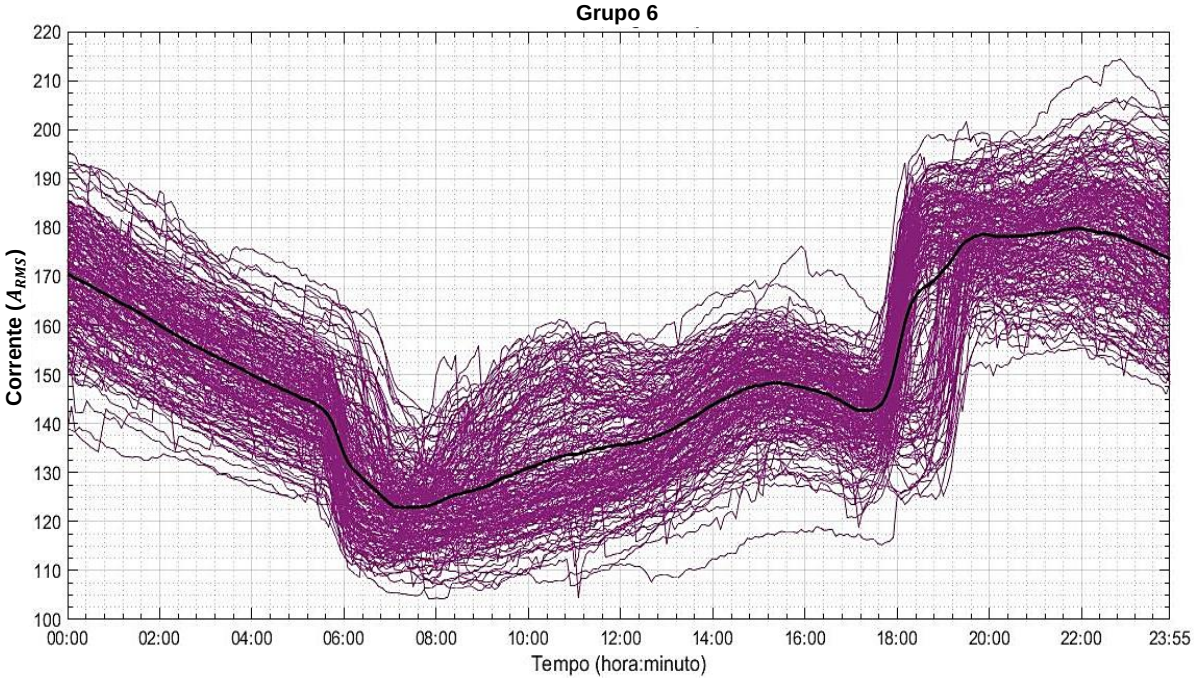
Figura 4.13 – Curvas de carga resultantes do grupo 5.



Fonte: Elaborada pelo autor.

As curvas de carga da tipologia presente no grupo 6, são ilustradas na Figura 4.14, apresentado um grupo de curvas de carga homogêneo. Esse grupo representa um comportamento similar das curvas de carga em torno da média, resultando em um agrupamento satisfatório.

Figura 4.14 – Curvas de carga resultantes do grupo 6.



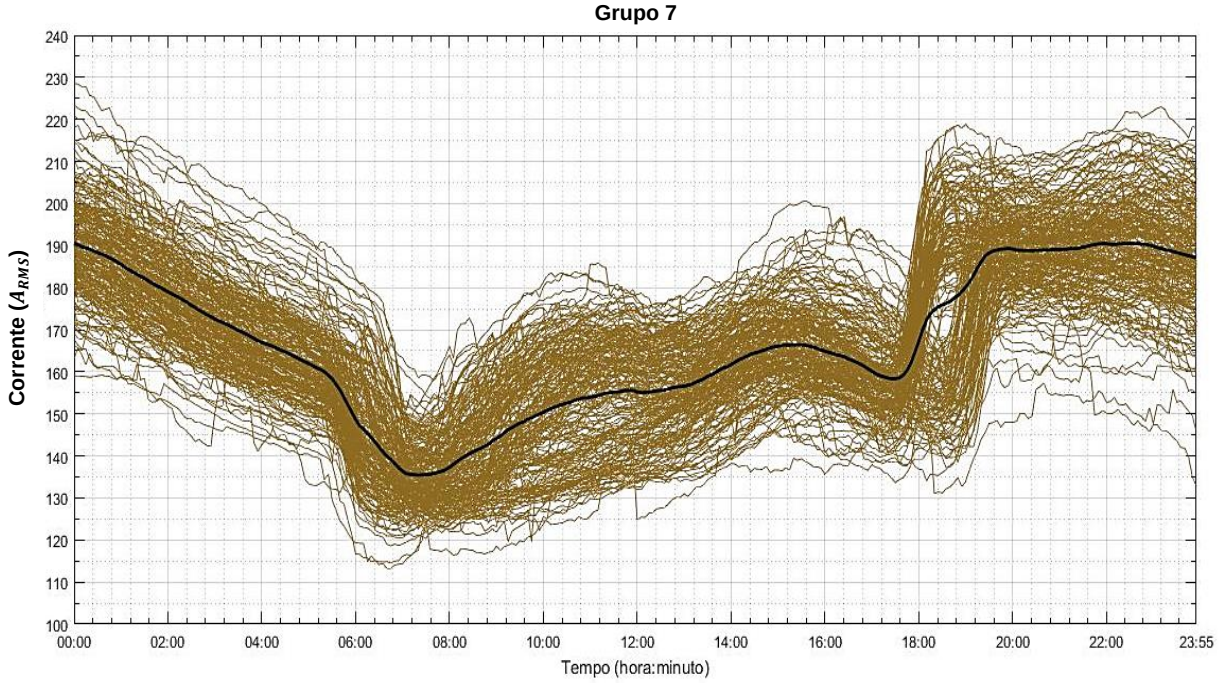
Fonte: Elaborada pelo autor.

Similar aos grupos 3 e 4, o grupo 6 descreve características oriundas de classes de carga residenciais e iluminação pública, pois ocorrem dois vales de carga bem definidos, às 07:00 e 18:00h, respectivamente, além do que não ocorrem rampas preeminentes nesse mesmo intervalo.

A tipologia característica das curvas de carga do grupo 7, pode ser verificada na Figura 4.15. Constata-se que as curvas de carga estão agrupadas homogeneamente em torno do centroide.

O grupo 7 descreve características próprias de carga residenciais e iluminação pública, tal qual os grupos 3, 4 e 6, já que às 07:00 e 18:00h ocorrem dois vales de carga bem definidos respectivamente, porém não ocorrem rampas de carga relevantes nesse intervalo.

Figura 4.15 – Curvas de carga resultantes do grupo 7.



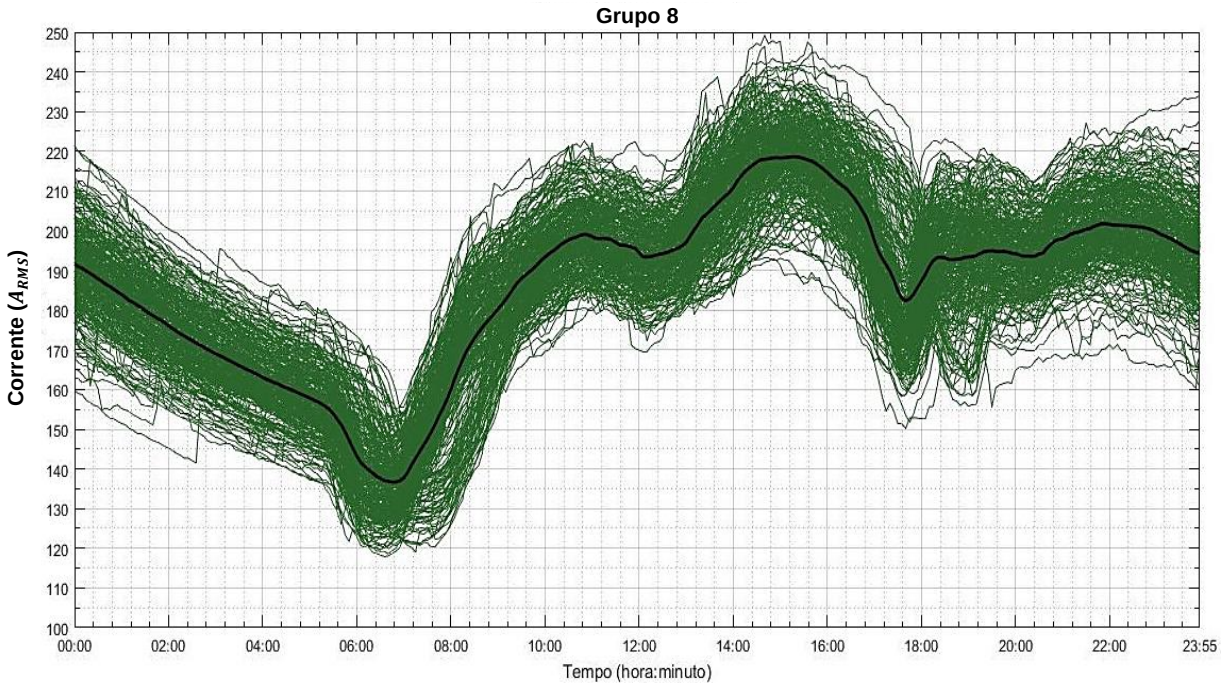
Fonte: Elaborada pelo autor.

Por fim, a tipologia das curvas de carga do grupo 8 está ilustrada na Figura 4.16, apresentado um grupo bastante homogêneo em torno do seu centroide, resultando em uma tipologia bem definida e com comportamento semelhante aos grupos 1, 2 e 5.

A tipologia do grupo 8 possui perfil residencial de 00:00h às 07:00h e comercial das 07:00 às 18:00h, já que nesse último período as amplitudes permanecem bastante elevadas com pequena variação. O fato do patamar das amplitudes elevarem-se das 18:00h às 19:00h,

reduzindo sua carga a partir das 22:00h, caracterizam nova predominância de classe de cargas com perfil residencial.

Figura 4.16 – Curvas de carga resultantes do grupo 8.



Fonte: Elaborada pelo autor.

Sendo assim, constata-se que dentre os perfis de carga dos grupos resultantes acima, prevalecem cargas residenciais e comerciais, concordando com trechos das características de cargas típicas do item 2.2.

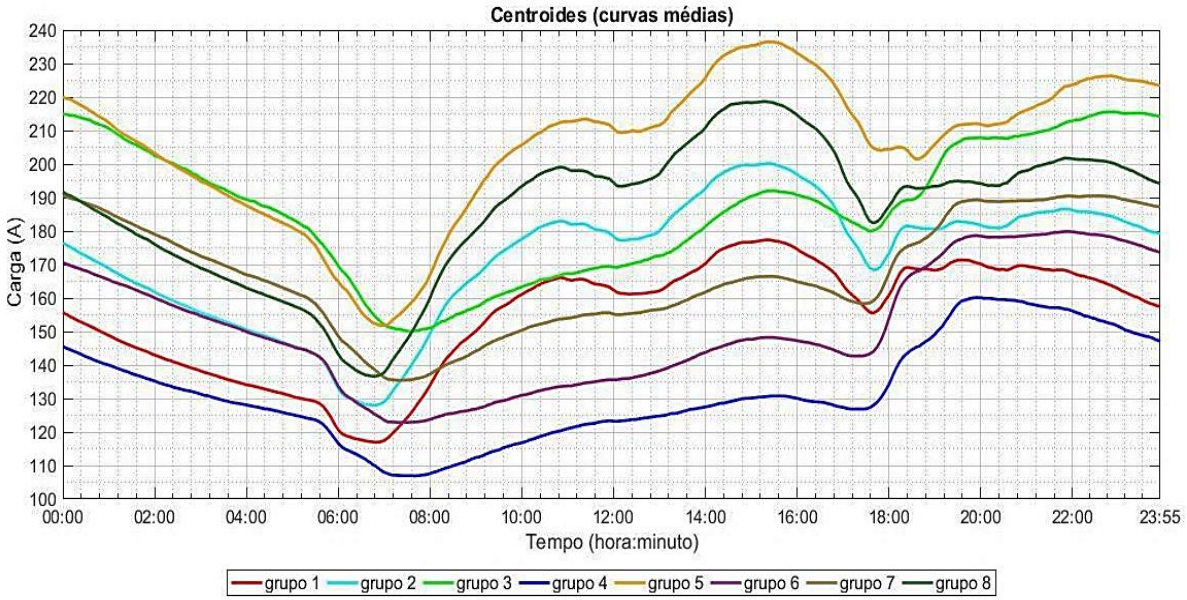
Apesar dos centroides dos grupos manifestarem similaridades nos seus comportamentos, esses diferem nas amplitudes das curvas de carga. Para ilustrar essas distinções e similaridades entre as curvas médias das tipologias resultantes, a Figura 4.17 ilustra os centroides dos grupos que constituem o melhor agrupamento realizado pela metodologia proposta.

Nos grupos resultantes da Figura 4.17, existem dois pontos de intersecção predominantes, para a maioria das características extraídas do banco de dados, ocorrendo por volta das 07:00h e das 18:00h, que são os pontos associados aos vales de carga mais predominantes do SEP estudado. Essa peculiaridade divide todo o banco de dados em três segmentos de 00:00h às 07:00h, das 07:00 às 18:00h e das 18:00 às 23:55h.

As similaridades associadas aos perfis dos centroides das curvas de carga, dividem o banco de dados basicamente em duas tipologias predominantes. O primeiro perfil é representado pelos grupos 1, 2, 5 e 8, referente às cargas predominantemente residenciais e

comerciais. Enquanto, o segundo refere-se aos grupos 3, 4, 6 e 7, que são de cargas predominantemente residenciais e iluminação pública.

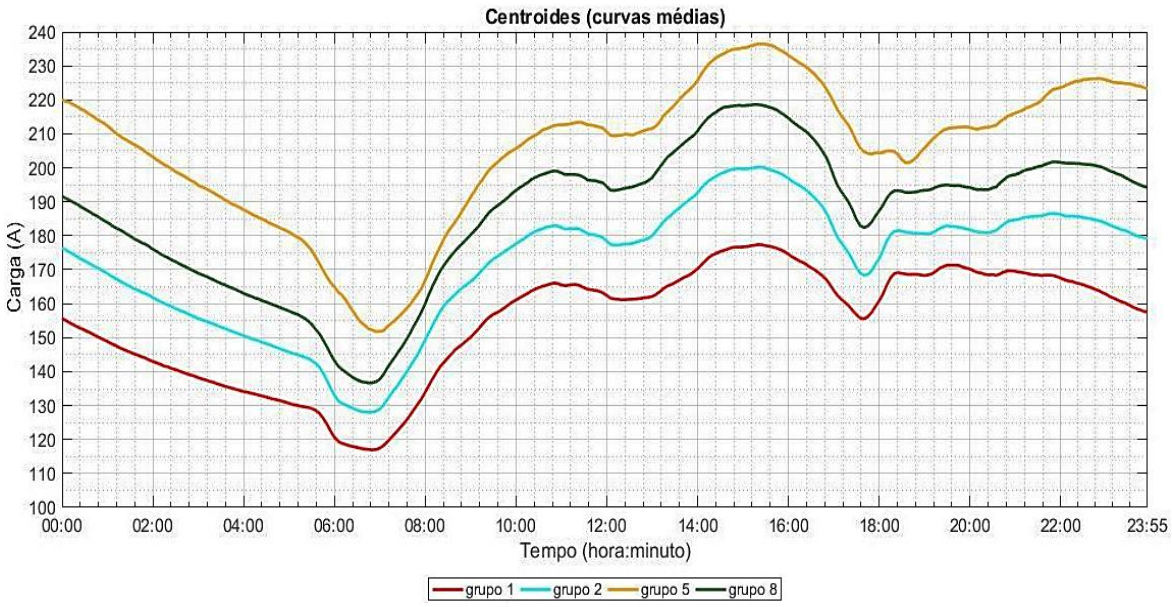
Figura 4.17 – Centroides das curvas de carga do agrupamento resultante da metodologia proposta.



Fonte: Elaborada pelo autor.

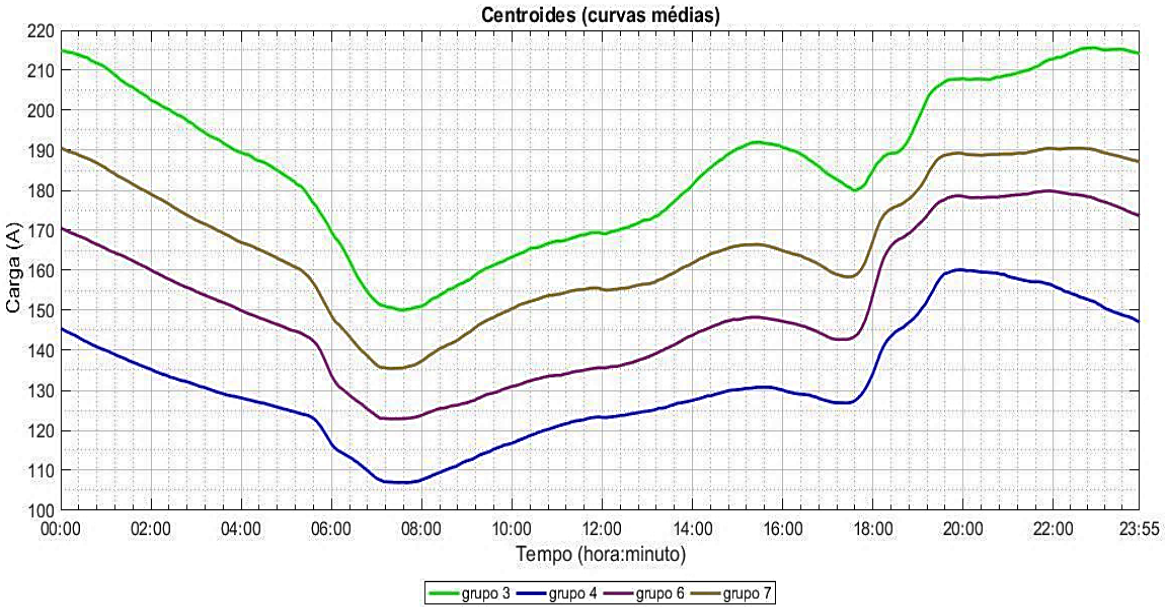
Por outro lado, os grupos similares na tipologia de carga são diferentes em termos de amplitude, conforme verifica-se nas Figura 4.18 e 4.19.

Figura 4.18 – Centroides das curvas de carga do agrupamento resultante, com mesma tipologia, predominantemente comercial e residencial.



Fonte: Elaborada pelo autor.

Figura 4.19 – Centroides das curvas de carga do agrupamento resultante com mesma tipologia, predominantemente residencial e iluminação pública.



Fonte: Elaborada pelo autor.

Sendo assim, observa-se que a metodologia proposta agrupou as curvas de carga em duas características predominantes, a primeira refere-se à grupos predominantemente de cargas comerciais e residenciais, conforme a Figura 4.18, enquanto que a segunda característica diz respeito às cargas predominantemente residenciais e iluminação pública. Contudo, verifica-se que tanto na Figura 4.18 como na Figura 4.19, o que diferencia os grupos com mesmas caraterísticas é sua amplitude de carga, demonstrando de fato que ocorrem oito tipologias distintas.

4.3.6 Resultado dos agrupamentos em função dos dias da semana e meses do ano

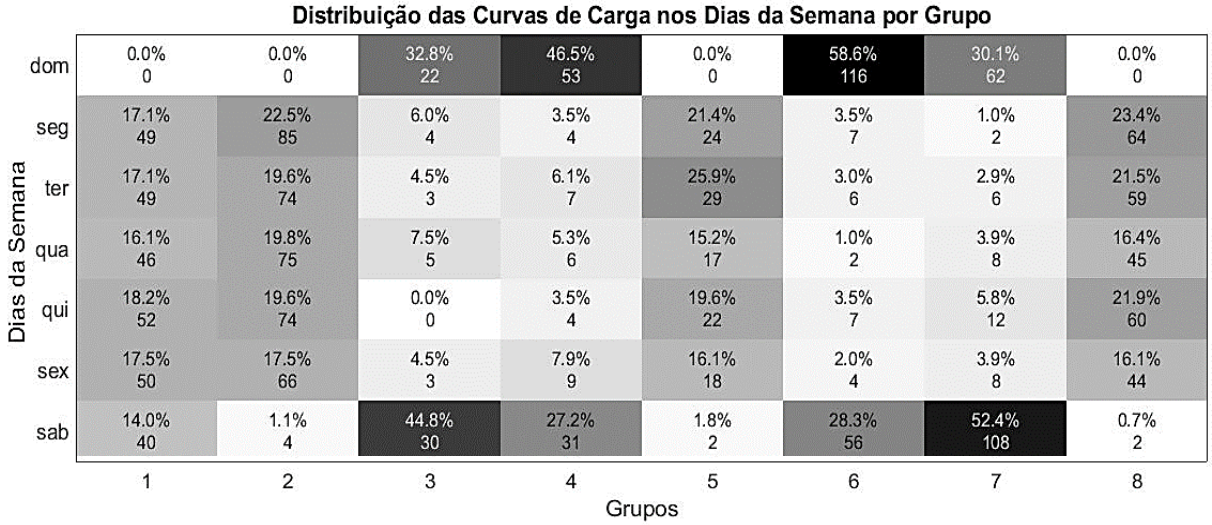
Os grupos resultantes do agrupamento das curvas de carga, da metodologia proposta, possuem características bem definidas, que denotam predominância comercial e residencial ou perfil de consumo residencial. Todavia há diferenças de tipologias dentro dessas duas características das classes de cargas predominantes, ilustradas nas Figuras 4.18 e 4.19, manifestando-se através de diferentes amplitudes.

O consumo das cargas de um SEP sofre influência dos dias da semana e meses do ano. Sendo assim, realizou-se o levantamento da distribuição das curvas de carga dos referidos

grupos, em função dos dias da semana e meses do ano, para verificar as particularidades de cada grupo resultante da metodologia de agrupamento proposta.

A Tabela 4.5, ilustra a relação das curvas de carga distribuídas nos dias da semana com cada grupo resultante da metodologia de agrupamento.

Tabela 4.5 – Distribuição das curvas de carga em dias da semana por grupo.



Fonte: Elaborada pelo autor.

Para viabilizar a análise de predominância das curvas de carga das Tabelas 4.5 e 4.6, em seus respectivos grupos, meses ou dias da semana, definiu-se como predominantes, os elementos das colunas com percentual igual ou superior à metade do elemento com maior predominância. Por exemplo, na coluna do grupo 5 da Tabela 4.5 o maior número de curvas de carga ocorre na terça-feira, que corresponde a 25,9% das curvas do grupo, logo serão considerados os dias da semana com percentual de curvas de carga maior ou igual à 12,95%. Sendo assim, o grupo 5 trata de um grupo com predominância de dias úteis.

Conforme constata-se na Tabela 4.5, os grupos 2, 5 e 8, são associados aos dias úteis da semana. Esse fato corrobora com as tipologias dos grupos 2, 5 e 8, da Figura 4.18, pois tratam-se de cargas predominantemente comerciais, durante o período de 07:00h às 18:00h, que é o horário de expediente comercial e perfil predominantemente residencial no restante do dia, durante os dias úteis da semana,

Outra constatação da Tabela 4.5 refere-se ao grupo 1, que corresponde aos dias úteis da semana e o sábado. Considerando o que expressa a tipologia do grupo 1 na Figura 4.18, o mesmo possui características de classes de cargas comerciais e residenciais, com

predominância comercial de 07:00 às 18:00h e residencial no restante do dia, assim como os grupos 2, 5 e 8.

Diferentemente dos grupos 2, 5 e 8, que se referem apenas aos dias úteis da semana, o grupo 1 também possui curvas de carga referentes ao sábado, ilustrando uma peculiaridade no agrupamento resultante.

Por fim, os grupos 3, 4, 6 e 7, são associados aos fins de semana. Isso está de acordo com os resultados da Figura 4.19, que evidenciam os grupos 3, 4, 6 e 7, como cargas com comportamento predominantemente residencial e iluminação pública.

Os fatores mencionados anteriormente confirmam a divisão das tipologias em duas formas de onda básicas. A primeira forma de onda está associada aos grupos 1, 2, 5 e 8, que manifestam elevadas amplitudes de carga no período de 07:00 às 18:00, correspondendo aos dias úteis e no caso do grupo 1 também considera sábados atípicos com mesmo comportamento de carga mencionado. A segunda forma de onda predominante refere-se aos fins de semana típicos, representados pelos grupos, 3, 4, 6 e 7.

Contudo, mesmo para formas de ondas médias similares, ocorrem variações de suas amplitudes. Para analisar tal fato, segue abaixo a Tabela 4.6, que demonstra a relação das curvas de carga distribuídas nos meses do ano e cada grupo resultante da metodologia de agrupamento.

Tabela 4.6 – Distribuição das curvas de carga nos meses do ano por grupo.

Distribuição das Curvas de Carga nos Meses do ano por Grupo

Mês	1	2	3	4	5	6	7	8
jan	12.6% 36	8.2% 31	6.0% 4	19.3% 22	4.5% 5	11.1% 22	11.2% 23	4.4% 12
fev	12.2% 35	4.5% 17	3.0% 2	20.2% 23	2.7% 3	7.6% 15	9.2% 19	5.1% 14
mar	17.5% 50	8.2% 31	0.0% 0	14.0% 16	0.0% 0	12.1% 24	3.4% 7	6.2% 17
abr	13.3% 38	9.5% 36	1.5% 1	13.2% 15	1.8% 2	8.6% 17	6.8% 14	5.1% 14
mai	10.1% 29	10.1% 38	0.0% 0	7.0% 8	0.9% 1	10.1% 20	9.7% 20	13.1% 36
jun	7.7% 22	15.9% 60	0.0% 0	1.8% 2	0.0% 0	9.6% 19	10.7% 22	10.2% 28
jul	14.3% 41	18.8% 71	0.0% 0	11.4% 13	0.0% 0	13.1% 26	5.3% 11	7.7% 21
ago	6.3% 18	10.6% 40	9.0% 6	1.8% 2	12.5% 14	8.1% 16	5.8% 12	9.5% 26
set	0.7% 2	3.7% 14	13.4% 9	0.9% 1	18.8% 21	5.6% 11	9.2% 19	15.0% 41
out	0.3% 1	2.1% 8	25.4% 17	3.5% 4	24.1% 27	3.0% 6	6.3% 13	8.8% 24
nov	1.0% 3	4.8% 18	22.4% 15	3.5% 4	19.6% 22	7.1% 14	9.2% 19	9.5% 26
dez	3.8% 11	3.7% 14	19.4% 13	3.5% 4	15.2% 17	4.0% 8	13.1% 27	5.5% 15

Grupos

Fonte: Elaborada pelo autor.

Quanto à distribuição das curvas de carga nos grupos 1, 2, 5 e 8, da Tabela 4.6, constata-se que, o grupo 1 é predominante nos meses de janeiro à maio e julho, enquanto que o grupo 2 predomina nas curvas dos meses de abril a agosto. Já o grupo 5 prevalece nos meses de agosto a dezembro, e o grupo 8 é preeminentemente nos meses de maio a novembro.

Por outro lado, quanto aos grupos 3, 4, 6 e 7 da Tabela 4.6, observa-se que o grupo 3 é preponderante nos meses de setembro a dezembro, o grupo 4 prevalece nos meses de janeiro à abril e julho, enquanto que o grupo 6 predomina nos meses de janeiro à agosto e novembro, sendo o grupo 7 preponderante nos meses de abril à junho, setembro e novembro a fevereiro.

Sendo assim, verifica-se que os grupos se manifestam ao longo dos meses em blocos consecutivos, ou blocos fragmentados.

Sabe-se que a temperatura é um dos fatores que condicionam o consumo de carga e que a mesma varia durante os meses do ano de acordo com as estações climáticas. Conforme o Centro de Previsões do Tempo e Estudos Climáticos ligado ao Instituto Nacional de Pesquisas Espaciais (CPTEC/INPE), segue abaixo, na Tabela 4.7, as temperaturas históricas da cidade de Teresina, onde está instalado o transformador associado ao banco de dados desse estudo.

Tabela 4.7 – Temperaturas históricas máximas, nas estações do ano na cidade de Teresina.

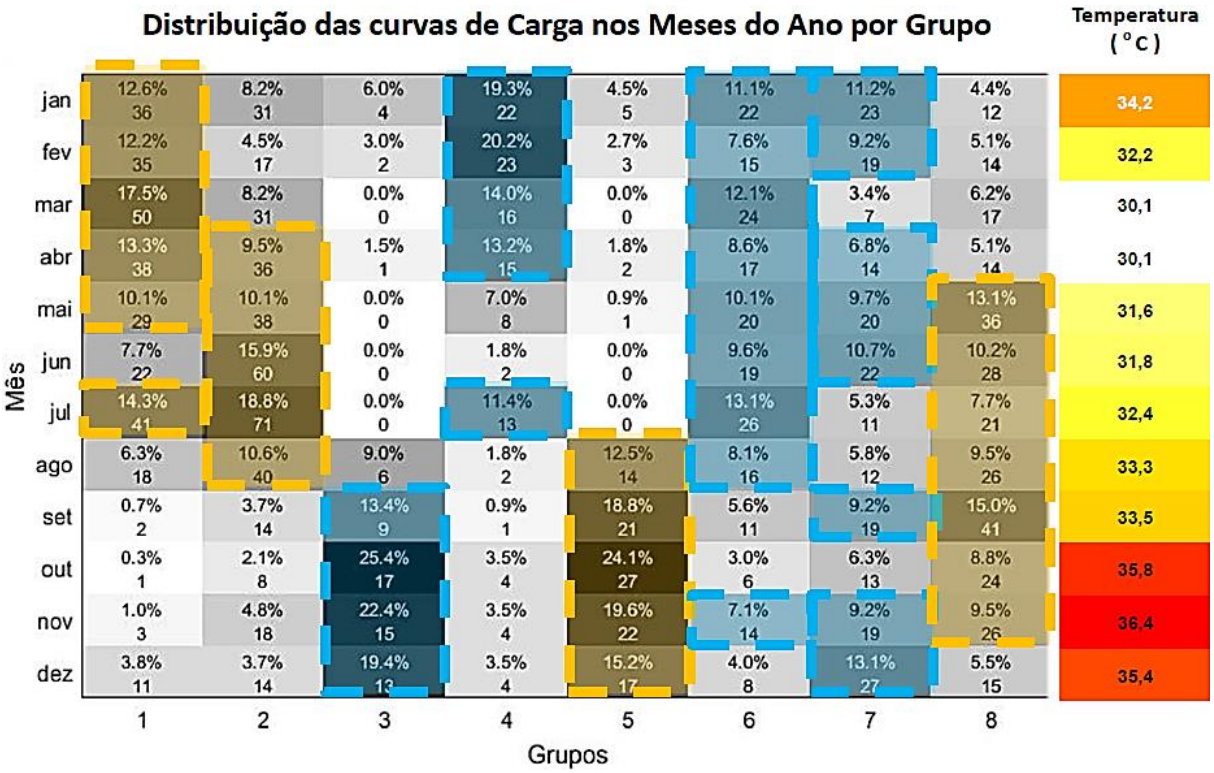
Estações Climáticas	Meses do Ano	Temperatura (ºC)
Verão	Dezembro	34,2
	Janeiro	32,2
	Fevereiro	30,1
Outono	Março	30,1
	Abril	31,6
	Maio	31,8
Inverno	Junho	32,4
	Julho	33,3
	Agosto	33,5
Primavera	Setembro	35,8
	Outubro	36,4
	Novembro	35,4

Fonte: CPTEC/INPE (2018).

Com isso, constata-se que as concentrações das curvas de carga nos meses do ano estão associadas às temperaturas históricas, conforme Tabela 4.8. Sendo assim, verifica-se que o deslocamento de amplitude dos grupos, com características de classes de cargas semelhantes ilustrados nas Figuras 4.18 e 4.19, ocorrem devido à variação de temperatura nos meses do ano.

Quanto mais concentrado o grupo estiver em torno de meses com temperaturas históricas elevadas, maior será a amplitude da curva média que descreve o grupo.

Tabela 4.8 – Disposição dos grupos de curvas de carga nos meses do ano, de acordo com as temperaturas históricas.



Fonte: Elaborada pelo autor.

Portanto, a densidade de curva de carga nos meses do ano, para os grupos que possuem centroides com forma de onda predominantemente residencial e comercial, da Figura 4.18 são:

- Grupo 1 (amplitude menor) - predomina nos meses de janeiro à maio e julho, logo está presente no verão, outono e inverno, porém é mais concentrado no outono, estação do ano com menores temperaturas.
- Grupo 2 (amplitude média inferior) - prevalece nos meses de abril a agosto e está diluído nas estações do outono e inverno, porém mais concentrado no inverno, estação do ano com temperaturas superiores ao outono.
- Grupo 5 (amplitude superior) - prepondera nos meses de agosto a dezembro, logo está presente no inverno, primavera e verão, contudo está muito concentrado na primavera, estação do ano com as maiores temperaturas.
- Grupo 8 (amplitude média superior) - predomina nos meses de maio a novembro, logo está diluído nos meses do outono, inverno e primavera, com

concentração dividida nas estações do inverno e primavera, resultando em amplitudes entre os grupos 2 e 5.

Quanto aos grupos referentes aos centroides com forma de onda predominantemente residencial e iluminação pública, da Figura 4.19, a concentração de curvas de carga são:

- Grupo 3 (amplitude superior) - predomina nos meses de setembro a dezembro, logo está presente na primavera e verão, porém é mais concentrado na primavera, estação do ano com maiores temperaturas.
- Grupo 4 (amplitude inferior) - prevalece nos meses de janeiro à abril e julho, logo está diluído nas estações do verão, outono e inverno, porém está mais concentrado no outono, estação do ano com menores temperaturas.
- Grupo 6 (amplitude média inferior) - prepondera nos meses de janeiro à agosto e novembro, logo está presente em todas as estações do ano. Contudo, sua concentração está dividida nas estações do outono e inverno, que são as duas estações com menores registros históricos de temperaturas, resultando em amplitudes entre os grupos 4 e 7.
- Grupo 7 (amplitude média superior) - predomina nos meses de abril à junho, setembro e novembro a fevereiro, logo está diluído em todas as estações do ano, maior concentração no verão, estação do ano com temperaturas imediatamente inferiores à primavera, resultando em amplitudes entre os grupos 3 e 6.

Por tudo isso, os grupos resultantes da metodologia proposta descrevem oito tipologias distintas bem definidas em termos de cargas típicas, dias da semana e meses do ano.

4.4 CONCLUSÕES PARCIAIS

Nesse capítulo foram apresentados os resultados gerais referentes à filtragem e agrupamento da metodologia proposta. Verificou-se a eficácia da filtragem e sua robustez, assim como a caracterização do agrupamento resultante, em grupos bem definidos, de acordo com as classes de cargas e regime de consumo predominantes.

Quanto à eficácia de filtragem, ficou claro que a metodologia proposta possui mecanismo de bloqueio para curvas de carga não corrigida, já que após a correção das mesmas, através do filtro de *hampel*, verifica-se que a ocorrência de falhas no restante das curvas de carga filtradas, foram detectadas pelo banco de filtros da TWD.

No que se refere ao agrupamento, a metodologia proposta demonstrou ser eficaz, já que os oito grupos resultantes tiveram características bem definidas no que diz respeito ao tipo de carga presente no banco de dados. Além do que, a metodologia caracterizou categoricamente as diferenças presentes nas cargas com comportamento similar.

Outro importante resultado da metodologia proposta, foi que essa aglutinou pequenos grupos dentro de grupos mais similares, afim de evitar grupos muito reduzidos, próprios de eventos aleatórios, como verificado com a inclusão dos sábados atípicos, dentro do grupo de dias úteis.

Quanto a comparação através de métricas de validação de agrupamento, das características extraídas das curvas de cargas filtradas, oriundas das curvas do banco de filtros da TWD, a metodologia proposta mostrou-se mais eficaz que o agrupamento sem redução de dimensionalidade, assim como, mais satisfatória que a extração de características utilizando PCA. Todos esses fatores consistem na eficácia da estratégia de agrupamento não supervisionado, elaborada para a metodologia proposta.

No capítulo seguinte serão tratadas as conclusões gerais desse trabalho e apresentadas as propostas de trabalhos futuros.

5. CONCLUSÕES

O agrupamento de dados é um dos principais problemas na mineração de dados, sendo utilizado em diversos campos de pesquisa, incluindo negócios, ciência médica, finanças e engenharia. Agrupar curvas de carga é uma valiosa ferramenta na análise de SEP, já que com essa pode-se realizar a classificação de tipologias e extração de características do comportamento de sistemas elétricos.

Considerando falhas, *outliers* e a alta dimensionalidade inerente à series temporais de carga de um SEP real, foi proposta uma metodologia dividida em duas etapas. Na primeira etapa, foi desenvolvido um procedimento iterativo de filtragem híbrido, utilizando o filtro *hampel*, como ferramenta para a correção de curvas de carga e o sinal do detalhe do banco de filtros da TWD, para a identificação e exclusão de curvas de carga com falhas após o processo de correção. Na segunda etapa, elaborou-se um segundo procedimento iterativo, que utiliza o sinal de aproximação do banco de filtros da TWD, para a redução de dimensionalidade da curva de carga filtrada e o algoritmo de agrupamento *k-means*, para realizar o agrupamento das curvas de carga reduzidas. Contudo, embutido ao processo iterativo da etapa de agrupamento, ocorre a validação dos agrupamentos das curvas de cargas filtradas com dimensão original, baseados no resultado do agrupamento com redução de dimensionalidade.

5.1 CONCLUSÕES DA ETAPA DE FILTRAGEM

O processo iterativo da etapa de filtragem realiza a correção das curvas de carga com falhas satisfatoriamente para um intervalo de falhas de até 40 minutos, com 8 amostras, pois o filtro de hampel utiliza dados da própria curva para a correção. Acima desse período a correção pode tornar-se viável com técnicas de predição, utilizando dados históricos do banco de dados da carga.

O banco de filtros da TWD mostrou-se muito eficaz, como retaguarda para falhas de medição acima do intervalo de correção admitido pelo filtro de hampel. Isso se deve a utilização da *wavelet Daubechies*, selecionada após testes. Confirmando a indicação dessa *wavelet* mãe pela comunidade acadêmica, para sinalização de fenômenos com decaimentos e oscilações.

Além da indicação de curvas com falhas, o sinal do detalhe do banco de filtros da TWD também contribuiu como parâmetro de calibração do *k-hampel*, do filtro de *hampel*, através do

recurso de redundância do processo iterativo da etapa de filtragem, por meio da busca de otimalidade da melhor correção em função do trecho corrigido.

Constatou-se após investigação no sistema de medição, que o elevado número de curvas de carga excluídas, 35,35%, do banco de dados original, deve-se principalmente ao fato do banco de dados conter, dentre as falhas características, predominantemente, lacunas de grande duração, oriundas de falhas intermitentes nos equipamentos de aquisição de dados. Sendo assim, a metodologia proposta, no processo iterativo desenvolvido para a filtragem das curvas de carga do banco de dados, apresenta resultados satisfatórios para a correção e limpeza de falhas do banco de dados, assim como, também serve como indicativo para verificação de problemas associados aos equipamentos envolvidos, desde a medição até a aquisição dos dados.

5.2 CONCLUSÕES DA ETAPA DE AGRUPAMENTO

A metodologia proposta resultou no agrupamento com o maior SWC, através da menor representação de dimensionalidade, com o menor tempo de processamento, em relação aos demais procedimentos de agrupamentos verificados, com ou sem estratégia de redução dimensionalidade, conferindo velocidade e qualidade à metodologia de agrupamento proposta.

No resultado do agrupamento, baseado no nível 5 de aproximação da TWD, contatou-se menor SWC e o maior tempo de processamento em relação ao agrupamento baseado no nível 4. Isso pode ter ocorrido devido à redução demasiada da curva de carga subtrair importantes características, que representam a tipologia da curva de carga sob análise, provocando um esforço computacional maior, para estabilizar ou resultar num agrupamento.

A PCA possui a vantagem de representar mais características de um mesmo banco de dados, já que a mesma pode representar curvas de carga por qualquer quantidade inteira de amostras, a partir de duas amostras até o tamanho da série temporal original. No entanto, o banco de filtros da TWD, que pode representar as curvas de carga, do banco de dados do objeto de estudo, por no máximo cinco representações distintas, forneceu aquela que extraiu as melhores características gerais do banco de dados. Conferindo à extração de características da metodologia proposta qualidade e precisão.

O fato do agrupamento no ponto de saturação resultar em 8 grupos, com redução de dimensionalidade ou não, indica que a parametrização utilizada no algoritmo *k-means* aponta para um ótimo global.

A metodologia proposta classificou pelo menos duas formas de onda bem definidas presentes no banco de dados, associadas às cargas residenciais e comerciais dos dias úteis e residenciais e iluminação pública dos finais de semana. Além de cumprir um dos importantes objetivos de agrupamentos de dados, que diz respeito à busca da aglutinação de eventos aleatórios em grupos aproximados, que ocorreram em grupos de sábados com formas de onda aproximadas à dias úteis evitando grupos muito reduzidos.

O deslocamento de amplitude de cada curva de carga média, ilustrou adequadamente a concentração dos grupos em cada estação do ano, que por sua vez estão associadas às temperaturas históricas.

Embora o procedimento proposto disponha de validação não supervisionada, esse realizou agrupamentos distintos e bem definidos, com características que representam desde o tipo de carga, dia e meses de consumo predominantes, fornecendo valiosas informações à respeito do banco de dados. Isso atende aos interesses ao redor da análise de curvas de carga, podendo ser utilizado em diversas áreas de aplicação.

Esse trabalho proposto, pode ser utilizada em demais grandezas de equipamentos do SEP, seja para verificar a eficácia dos sistemas de aquisição de dados de medição, ou para explorar características inerentes ao funcionamento de equipamentos ou sistemas elétricos. Além do que, fornece tipologias que expressam características que podem servir de insumo na tomada de decisões, nas áreas de operação de sistemas eletroenergéticos, programação de intervenções, manutenção, estudos energéticos de SEP, dentre outras. Por tudo isso, a metodologia de filtragem e agrupamento de curvas de carga atende aos interesses ao redor da análise proposta, além de possuir um vasto campo de aplicação em SEP.

Essas evidências conferem aos procedimentos desenvolvidos de filtragem, agrupamento e validação, robustez e qualidade, atendendo aos objetivos gerais, específicos, contribuições científicas e técnicas, dessa pesquisa.

5.3 TRABALHOS FUTUROS

Como proposta de trabalhos futuros, a metodologia proposta pode ser melhorada no aspecto de filtragem das curvas de carga, pois as falhas superiores à 40 minutos são excluídas do banco de dados, o que pode ser um problema natural em determinados períodos do ano, como em chuvas ou queimadas. Sendo assim, pode-se utilizar métodos baseados nos dados

históricos, dentre outros a previsão de curva de carga, para realizar substituição de trechos com falhas no banco de dados.

Pesquisar técnicas que possam ser utilizadas para obter limiares adaptativos, quanto ao valor de amplitude do sinal do detalhe do banco de filtros da TWD, utilizados como critério de indicação de falhas presentes nas curvas de carga. Outra linha de pesquisa constitui, definir um limiar que seja adequado para indicar as falhas considerando uma curva de carga normalizada.

Quanto à metodologia de agrupamento, propõem-se realizar, com a metodologia atual, o agrupamento por trechos das curvas ao invés do agrupamento da curva completas. O referencial pode ser a média dos pontos de intersecção dos centroides, por setores, resultantes do agrupamento da metodologia atual, que ocorreram em média às 07:00h e 18:00h. Sendo assim, os agrupamentos seriam realizados, por exemplo, das 00:00h às 07:00h, das 07:00h às 18:00h e das 18:00h às 23:55h.

Pretende-se utilizar árvore de decisão agregando, características de temperatura, mês, dia da semana, ocorrência de feriado, entre outras, ao resultado do agrupamento da metodologia proposta. Dentre as diversas aplicações, planeja-se utilizar essas informações na etapa de pré-processamento em previsão de carga. Espera-se que as tipologias resultantes indiquem os padrões a serem utilizados na predição, proporcionando uma melhores acurácia e tempo de processamento da metodologia utilizada na previsão de carga.

Apesar do custo computacional da DTW, também se pretende explorar a realização de agrupamentos da metodologia proposta com essa medida de dissimilaridade. Para verificar a eficácia de agrupamento desta em relação a ED, já que a DTW permite a medida de distância de vetores fora de fase com bons resultados.

Por fim, outra proposta de agrupamento refere-se à utilização de dados estatísticos, para a redução de dimensionalidade na metodologia de agrupamento atual.

Sendo assim, espera-se que com a melhora da etapa de filtragem, menos dados sejam excluídos. Com as propostas de melhoria ou alteração da metodologia de agrupamento, maior precisão seja conferida aos resultados dos grupos, apresentando melhores tipologias. E em função de melhores características utilizadas para o agrupamento, o tempo de processamento seja reduzido, assim como, sejam obtidas melhores tipologias resultantes.

REFERÊNCIAS BIBLIOGRÁFICAS

ADDISON, P. **The Illustrated Wavelet Transform Handbook: Introductory Theory and Applications in Science, Engineering, Medicine and Finance**. Institute of Physics Publishing, 2002.

AGHABOZORGI, S.; YING W. T.; HERAWAN, T.; JALAB, H. A.; SHAYGAN, M. A.; JALALI, A. A hybrid algorithm for clustering of time séries data based on affinity search technique. **The Scientific World Journal**. 2014.

ALMEIDA, V. A. **Previsão Carga Através de Modelos Neuro-Fuzzy**. Projeto de Graduação. UFRJ - Escola Politécnica. Rio de Janeiro, 2013.

ANDREOPOULOS B.; AN A.; WANG X.; SCHROEDER M. A roadmap of clustering algorithms: finding a match for a biomedical application, **Briefings in Bioinformatics**, v. 10, n. 3, p. 297–314, 2009.

AUDER, B.; CUGLIARI, J.; GOUDE, Y.; POGGI, J. M. Scalable Clustering of Individual Electrical Curves for Profiling and Bottom-Up Forecasting. **Energies**, v. 11, n. 7, p. 1893, 2018.

BARAN, M.; J. KIM. A Classifier for Distribution Feeder Overcurrent Analysis. ***Power Delivery, IEEE Transactions*** *21*(1), 456–462. 2006.

BENÍTEZ, I.; QUIJANO, A.; DÍEZ, J. L.; Delgado, I. Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers. **International Journal of Electrical Power & Energy Systems**, v. 55, p. 437-448, 2014.

BRADLEY, P., S.; FAYYAD U.; REINA, C., Scaling clustering algorithms to large databases, in **Proceedings of the 4th International Conference on Knowledge Discovery & Data Mining** (KDD ’98), p. 9–15, 1998.

CAMARGOS, R. C.; NICOLETTI, M. C. Algoritmos Aglomerativos de Agrupamento Baseados em Teoria de Matrizes. **Workshop de Computação da FACCAMP**. Campo Limpo Paulista, Volume: 2. 2015.

CAMPELLO, R. J. G. B. **Análise de Agrupamento de Dados. Validação de Agrupamento: Parte I**. 54 slides. Disponível em: < https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=2ahUKEwi53OWZvZ_dAhWGiZAKHc6eDhkQFjAAegQIABAC&url=http%3A%2F%2Fwww.cin.ufpe.br%2F~jbr%2Fquestao_1%2FValidacao_Rand.pdf&usg=AOvVaw2Zb0TeAF46ALwNZgMWxf_q>. Acessado em 30 de dezembro de 2017.

CENTRO DE PREVISÃO DE TEMPO E ESTUDOS CLIMATOLÓGICOS (CPTEC). **Estações do Ano.** Disponível em <http://clima1.cptec.inpe.br/estacoes/#>. Acessado em 25 de abril de 2018.

CHAN F. K.-P.; FU A. W.-C.; YU C., “Haar *wavelet*s for efficient similarity search of time-séries: with and without time warping”, **IEEE Transactions on Knowledge and Data Engineering**, v. 15, n. 3, p. 686–705, 2003.

CHUI, C. K. **An introduction to *wavelet*s**. San Diego: Academic Press,1992.

DAVIES, D. L.; BOULDIN, D. W. A cluster separation measure. **IEEE transactions on pattern analysis and machine intelligence**, n. 2, p. 224-227, 1979.

DE OLIVEIRA, L. A. A. **TRATAMENTO DE DADOS DE CURVAS DE CARGA VIA ANÁLISE DE AGRUPAMENTOS E TRANSFORMADA *WAVELET*S**. 2013. Tese de Doutorado. Universidade Federal do Rio de Janeiro.

DUNN, J. C. A fuzzy relative of the ISODATA process and its use in Detecting Compact Well-Separated Clusters. **Cybernetics and Systems**. n. 3, p. 32-57, 1973.

HAAR, A. Zur theorie der orthogonalen funktionensysteme. **Mathematische Annalen**, v. 69, n. 3, p. 331-371, 1910.

HAMPEL, F. R. The influence curve and its role in robust estimation. **Journal of the american statistical association**, v. 69, n. 346, p. 383-393, 1974.

HAN, J.; PEI, J.; KAMBER, M. **Data mining: concepts and techniques**. Elsevier, 2012.

HERNÁNDEZ, L.; BALADRÓN, C.; AGUIAR, J.; CARRO, B.; SÁNCHEZ-ESGUEVILLAS, A. Classification and clustering of electricity demand patterns in industrial parks. **Energies**, v. 5, n. 12, p. 5215-5228, 2012.

HIRANO S.; TSUMOTO S., Empirical comparison of clustering methods for long time-series databases, **in Active Mining**, v. 3430, p. 268–286, 2005.

HONGYU, K. **Comparação do GGE-biplot ponderado e AMMI-ponderado com outros modelos de interação genótipo × ambiente**. 2015. 155p. Tese (Doutorado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2015.

HONGYU, K.; SANDANIELO, V. L. M.; DE OLIVEIRA JUNIOR, G. J. Análise de Componentes Principais: resumo teórico, aplicação e interpretação. **E&S Engineering and Science**, v. 5, n. 1, p. 83-90, 2016.

HORTA, D. **Algoritmos e técnicas de validação em agrupamento de dados multi-representados, agrupamento possibilístico e bi-agrupamento**. 2013. Tese de Doutorado. Universidade de São Paulo.

JAIN, A. K.; DUBES, R. C. **Algorithms for clustering data**. New Jersey: Prentice Hall, 1988.

JEMSE, A.; HARBO, A., C. **A Ripples in Mathematics: The Discrete *Wavelet* Transform**. New York: Springer-Verlag. 2001.

KAGAN, N.; DE OLIVEIRA, C. C. B.; ROBBA, E. J. **Introdução aos sistemas de distribuição de energia elétrica**. Edgard Blücher, 2005.

KAUFMAN, L.; ROUSSEEUW, P. J. Partitioning around medoids (program pam). **Finding groups in data: an introduction to cluster analysis**, p. 68-125, 1990.

KAZMIER, L. J. **Schaum's outline of business statistics**. [S.l.]: McGraw-Hill, 2004.

KEOGH, E; KASETTY, S. On the need for time series data mining benchmarks: A survey and empirical demonstration. In **Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, Edmonton, Alberta, Canada, p. 102–111. 2003.

KIWARE, S. S. **Detection of outliers in time series data**. 89 f. Master's Theses - Marquette University, 2009.

KOHAN, N.; MOGHADDAM, M. P.; BIDAKI, S. M. Evaluating performance of WFA K-means and modified follow the leader methods for clustering load curves. **In: Power Systems Conference and Exposition, 2009. PSCE'09. IEEE/PES.** IEEE, 2009. p. 1-5.

LAI, C.,-P., P.; CHUNG, P.,-C., C.; TSENG, V., S., A novel two-level clustering method for time series data analysis, **Expert Systems with Applications**, v. 37, n. 9, p. 6319–6326, 2010.

LIAO, T., W. Clustering of time séries data—a survey. **Pattern recognition**, v. 38, n. 11, p. 1857-1874, 2005.

LIN, J.; VLACHOS, M.; KEOGH E.; GUNOPULOS, D., Iterative incremental clustering of time series, **in Advances in Database Technology—EDBT 2004**, p. 106–122, 2004.

LIN, R.; WU, B.; SU, Y. An Adaptive Weighted Pearson Similarity Measurement Method for Load Curve Clustering. **Energies**, v. 11, n. 9, p. 2466, 2018.

MACQUEEN J., Some methods for classification and analysis of multivariate observations, in **Proceedings of the 5th Berkeley Symposium Mathematical Statistics and Probability**, v. 1, p. 281–297, 1967.

MATLAB R2016. Copyright 1984–2016 by the MathWorks, Inc.

MIRKIN, B. Choosing the number of clusters. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 1, n. 3, p. 252-260, 2011.

MORRIS B.; TRIVEDI M., Learning trajectory patterns by clustering: experimental studies and comparative evaluation, in **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition** (CVPR ’09), p. 312–319, 2009.

NIEVOLA, J. C. **Análise de Agrupamentos**. 99 slides. Disponível em: < https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=2ahUKEwizyfbHwZ_dAhWLFZAKHbtrBGUQFjAAegQIABAC&url=http%3A%2F%2Fwww.ppgia.pucpr.br%2F~fabricio%2Fftp%2FAulas%2FMestrado%2FIA%2FNievola%2FMD%2FMD-06-Agrupamento.pdf&usg=AOvVaw1pnXCbpupeIyf06LUu2JDe>. Acessado em 17 de dezembro de 2017.

OPERADOR NACIONAL DE SISTEMAS ELÉTRICOS (ONS). **IO-OI.NE.TSA, Revisão 36 de 31 de julho de 2018a.** Procedimentos Sistêmicos para a Operação da SE Teresina. Disponível em < http://ons.org.br/_layouts/download.aspx?SourceUrl=http://ons.org.br%2FMPO%2FDocumento%20Normativo%2F3.%20Instru%C3%A7%C3%B5es%20de%20Opera%C3%A7%C3%A3o%20-%20SM%2010.21%2F3.7.%20Opera%C3%A7%C3%A3o%20de%20Instala%C3% A7%C3%B5es%2F3.7.3.%20Nordeste%2F3.7.3.6.%20%C3%81rea%20230%20kV%20Oeste%2FIO-OI.NE.TSA_Rev.36.docx >. Acessado em 04/08/2018.

OPERADOR NACIONAL DE SISTEMAS ELÉTRICOS (ONS). **Mapa Geoelétrico - Rede de Operação - Brasil - Horizonte 2023. Emissão em 20 de julho de 2018b.** <http://ons.org.br/_layouts/download.aspx?SourceUrl=http://ons.org.br/Mapas/MapaGeoeletrico_RededeOperacao_Brasil_2018.pdf>. Acessado em 30 de julho de 2018.

OPERADOR NACIONAL DE SISTEMAS ELÉTRICOS (ONS). **Submódulo 23.2, Revisão 2016.12 de 01/01/2017**. Critérios para Definição das Redes do Sistema Interligado Nacional. Disponível em <http://www.ons.org.br/_layouts/download.aspx?SourceUrl=http://www.ons.org.br%2FProcedimentosDeRede%2FM%C3%B3dulo%2023%2FSubm%C3%B3dulo%2023.2%2FSubm%C3%B3dulo%2023.2%202016.12.pdf>. Acessado em 05 de março de 2018.

PAPARRIZOS, John; GRAVANO, Luis. Fast and accurate time-series clustering. **ACM Transactions on Database Systems (TODS)**, v. 42, n. 2, p. 8, 2017.

PEARSON, K. Principal Components Analysis. ***The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science***, v. 6, n. 2, p. 559, 1901.

PESSANHA, J. F. M.; MELA, A. C. G.; JUSTINO, T. C.; MACEIRA, M. E. P. Combining Statistical Clustering Techniques and Exploratory Data Analysis to Compute Typical Daily Load Profiles-Application to the Expansion and Operational Planning in Brazil. **In: 2018 IEEE International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)**. IEEE, 2018. p. 1-6.

PESSANHA, J. F. M.; XAVIER, V. L.; AMARAL, M. R. S.; LAURENCEL, L. C. L Construindo Tipologias de Curvas de Carga com o Programa R. **Pesquisa Operacional para o Desenvolvimento**. 7. 29, 2015.

PUMA-VILLANUEVA, W. J.; ZUBEN, F. J. V. Índices de validação de agrupamentos. **Universidade Estadual de Campinas. Unicamp. Faculdade de Engenharia Elétrica e de Computação. FEEC**, 2008.

RANI, S.; SIKKA G., Recent techniques of clustering of time series data: a survey, **International Journal of Computational and Applied**, v. 52, n. 15, p. 1–9, 2012.

RÄSÄNEN, T.; KOLEHMAINEN, M. Feature-based clustering for electricity use time series data. **In: International Conference on Adaptive and Natural Computing Algorithms**. Springer, Berlin, Heidelberg, 2009. p. 401-412.

RINCÓN, A. Q.; RISK, M.; LIBERCZUK, S. Preprocesamiento de EEG con Filtros Hampel. **IEEE Latin American Transactions**, 2012.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of computational and applied mathematics**, v. 20, p. 53-65, 1987.

SILVA, M. C. A., **Aplicação de um Sistema *Fuzzy* para Classificação de Opinião em Diferentes Domínios**. 2013. Dissertação de Mestrado. Universidade Federal do Rio de Janeiro.

TSEKOURAS, G. J.; HATZIARGYRIOU, N. D.; DIALYNAS, E. N. Two-stage pattern recognition of load curves for classification of electricity customers. **IEEE Transactions on Power Systems**, v. 22, n. 3, p. 1120-1128, 2007.

VENDRAMIN, L.; CAMPELLO, R. J. G. B.; HRUSCHKA, E. R. Relative clustering validity criteria: A comparative overview. **Statistical analysis and data mining: the ASA data science journal**, v. 3, n. 4, p. 209-235, 2010.

VLACHOS, M.; LIN, J.;KEOGH, E., A *wavelet*-based anytime algorithm for k-means clustering of time series, **in Proceedings of the Workshop on Clustering High Dimensionality Data and Its Applications**, p. 23–30, 2003.

WANG, X.; SMITH, K.; HYNDMAN R., Characteristic-based clustering for time series data, **Data Mining and Knowledge Discovery**, v. 13, n. 3, p. 335–364, 2006.

WU, Z.; DONG, X.; LIU, Z.; KONG, X.; ZENG, Y.; HU, Q. A.; CHEN, Y. Power system bad load data detection based on an improved fuzzy C-means clustering algorithm. **In: Power & Energy Society General Meeting, 2017 IEEE**. IEEE, 2017. p. 1-5.

XI, A.; KEOGH, E.; SHELTON, C.; WEI, L. and RATANAMAHATANA, C, A. "Fast time series classification using numerosity reduction", in **Proceedings of the 23rd International Conference on Machine Learning (ICML '06)**, p. 1033–1040, 2006.

YANG, J.; STENZEL, J. Historical load curve correction for short-term load forecasting, **7th International Power Engineering Conference**, 29 Nov – 02 Dec, Singapore, 2005.

ZHANG, X.; LIU, J.; DU, Y.; LV, T., A novel clustering method on time series data, **Expert Systems with Applications**, v. 38, n. 9, p. 11891–11900, 2011.