UMA METODOLOGIA PARA TRATAMENTO DE DADOS DE CURVAS DE CARGA BASEADA EM TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL

Victor Andrade de Almeida

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Elétrica, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Elétrica.

Orientadores: Luiz Pereira Calôba

José Francisco Moreira Pessanha

Rio de Janeiro

Fevereiro de 2017

UMA METODOLOGIA PARA TRATAMENTO DE DADOS DE CURVAS DE CARGA BASEADA EM TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL

Victor Andrade de Almeida

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Examinada por:

Prof. Luiz Pereira Calôba, D. Ing.

Prof. José Francisco Moreira Pessanha, D.Sc.

Prof.ª Mariane Rembold Petraglia, Ph.D.

Prof. Fernando Luiz Cyrino de Oliveira, D.Sc.

RIO DE JANEIRO, RJ - BRASIL

FEVEREIRO DE 2017

Almeida, Victor Andrade de

Uma Metodologia para Tratamento de Dados de Curvas de Carga Baseada em Técnicas de Inteligência Artificial/ Victor Andrade de Almeida – Rio de Janeiro: UFRJ/COPPE, 2017.

XIII, 92 p.: il.; 29,7 cm.

Orientadores: Luiz Pereira Calôba

José Francisco Moreira Pessanha

Dissertação (mestrado) – UFRJ/ COPPE/ Programa de Engenharia Elétrica, 2017.

Referências Bibliográficas: p. 87-90.

1. Tratamento de Dados. 2. Classificação de Curvas de Carga. 3. Inteligência Artificial. I. Calôba, Luiz Pereira *et al*. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Título.

Agradecimentos

Agradeço a Deus, por me iluminar em cada escolha que fiz ao longo da vida e por me cercar de tantas pessoas incríveis que me ajudam a aprender e crescer como profissional e como ser humano.

Aos meus pais, Nizeuma e Fernando, por todo o amor e cuidado dedicado a mim durante toda a minha vida. Se cheguei até aqui, foi graças ao esforço, dedicação e muitas vezes renúncia aos próprios sonhos. Não aos meus, mas aos seus. Obrigado por tudo.

À minha irmã, Viviane, por servir de exemplo de empenho nos estudos e trilhar o difícil caminho da engenharia antes de mim. Tudo ficou mais fácil porque eu sabia que era possível.

Ao professor, amigo e orientador José Francisco. Muito obrigado por acreditar em mim e me ensinar tanto. Obrigado pelo apoio, pelo incentivo e confiança em todas as fases da minha carreira.

Ao professor e orientador Luiz Pereira Calôba pela gentileza, disponibilidade e contribuição tanto na solidificação dos conhecimentos transmitidos durante as suas disciplinas quanto na colaboração para a realização desta dissertação.

Aos professores Mariane Petraglia e Fernando Cyrino pela participação na banca examinadora e pelas contribuições para a conclusão deste trabalho.

Ao CEPEL, como instituição, por me admitir ainda como estagiário de graduação, me proporcionando o acesso a uma ampla estrutura física e intelectual na busca do conhecimento. Este apoio possibilitou que eu me tornasse engenheiro, bolsista de mestrado, pesquisador e hoje, mestre.

Ao Programa de Engenharia Elétrica da COPPE/UFRJ, obrigado pela oportunidade, pela estrutura e pelos conhecimentos adquiridos.

Aos amigos do CEPEL, Adhara, Diego, Felipe, Hugo, Juan, Lívia e Priscilla. Obrigado pelo bom humor, pelo incentivo e companhia de todos vocês.

Finalmente, agradeço a minha namorada, esposa, amiga e companheira de jornada. Aline, esta conquista é nossa. Mais uma entre as muitas que certamente virão. Você é um exemplo de amor e persistência.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

UMA METODOLOGIA PARA TRATAMENTO DE DADOS DE CURVAS DE CARGA BASEADA EM TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL

Victor Andrade de Almeida

Fevereiro/2017

Orientadores: Luiz Pereira Calôba
José Francisco Moreira Pessanha

Programa: Engenharia Elétrica

A qualidade dos dados é fundamental na previsão de curto prazo da carga. Não raro, os dados de carga apresentam valores aberrantes, descontinuidades e lacunas resultantes da operação anormal do sistema elétrico ou falhas e problemas no sistema de medição. A presença de dados corrompidos prejudica a especificação dos modelos de previsão de carga e consequentemente afeta a qualidade das previsões obtidas. Portanto, a construção de um modelo de previsão de carga deve ser precedida por uma etapa de tratamento dos dados. Nesta dissertação apresenta-se uma metodologia baseada em métodos estatísticos e de inteligência artificial para tratamento de dados de carga. Ao longo da dissertação apresentam-se os métodos utilizados e como cada um deles é empregado na identificação e correção dos principais tipos de erros frequentemente encontrados nos dados de carga. Adicionalmente, experimentos computacionais foram conduzidos com dados de carga provenientes do Sistema Interligado Nacional (SIN) com a finalidade de avaliar a capacidade da metodologia proposta em limpar e recuperar os padrões originais de curvas de carga corrompidas. Nos experimentos realizados as curvas de carga foram corrompidas artificialmente por meio de simulação estatística e posteriormente tratadas pela metodologia proposta. Os resultados alcançados mostram a boa aderência das curvas de carga resultantes do processo de limpeza de dados aos respectivos perfis originais não corrompidos.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

A METHODOLOGY FOR THE TREATMENT OF LOAD CURVE DATA BASED ON ARTIFICIAL INTELLIGENCE TECHNIQUE

Victor Andrade de Almeida

February/2017

Advisors: Luiz Pereira Calôba

José Francisco Moreira Pessanha

Department: Electrical Engineering

Data quality is critical in the short-term load forecasting. Frequently, load data show aberrant values (outliers), discontinuities, and gaps (missing data) caused by the abnormal operation of the electrical system or failures and problems in the measurement system. The presence of corrupted data impairs specification of load forecasting models and consequently affects the quality of predictions obtained. Therefore, the construction of a load prediction model must be preceded by a data processing step. This dissertation presents a methodology based on statistical methods and artificial intelligence for the treatment of load data. Throughout the dissertation are presented the methods used and how each of them is employed in the identification and correction of the main types of errors frequently found in the load data. In addition, computational experiments were conducted with load data from the National Interconnected System in order to evaluate the ability of the proposed methodology to clean and recover the original patterns of corrupted load curves. In the experiments performed the load curves were artificially corrupted by means of statistical simulation and later treated by the proposed methodology. The results show the good adherence of the load curves resulting from the data cleaning process to their original uncorrupted profiles. Computational experiments were conducted with real data from the National Interconnected System (SIN) to evaluate the ability of the proposed methodology to clean load data and recover the original patterns of corrupted load curves. In the experiments, the load curves were artificially corrupted and then filtered by the proposed methodology. The results show the good adherence of the load curves resulting from the data cleaning process to their original uncorrupted profiles.

Sumário

Lista de Figuras

Lista de Tabelas

1. INTRODUÇÃO

Uma informação fundamental para a operação econômica e segura de um sistema elétrico de potência é a previsão de curto prazo da demanda por energia elétrica. As previsões de curto prazo incluem projeções da ponta diária(valor máximo da demanda em cada dia) e as previsões de demanda em horizontes até 168 horas(1 semana) à frente, em bases horárias ou com resolução temporal de 10, 15 e 30 minutos [1].

A previsão de carga alimenta sofisticados sistemas computacionais que suportam a tomada de decisões operativas destinadas ao controle da frequência e tensão da rede elétrica em seus níveis nominais, condição essencial para a manutenção do equilíbrio entre a geração e a demanda. Por exemplo, previsões de carga em bases horárias para um dia à frente constituem informações básicas para a programação diária da operação do Sistema Interligado Nacional [2]. Na Tabela 1.1 apresenta-se o cronograma para envio da previsão de carga para a elaboração da programação diária da operação eletroenergética. Tais previsões de carga são úteis em análises de contingência conduzidas *off-line* com a finalidade de detectar condições operativas futuras que representem risco para o sistema elétrico. A partir destas análises o operador pode preparar antecipadamente as ações corretivas necessárias para a operação segura do sistema [1].

Tabela 1.1 - Cronograma para envio das previsões de carga para a programação diária da operação

Dia de elaboração da programação	Carga prevista para o dia	Carga prevista para o dia (antecipação da programação)
2ª feira	3ª feira	4ª feira
3ª feira	4ª feira	5ª feira
4ª feira	5ª feira	6ª feira
5ª feira	6ª feira e sábado	domingo
6ª feira	domingo e 2ª feira	3ª feira

Fonte: ONS - Consolidação da previsão de carga para a programação diária da operação eletroenergética e para a programação de intervenções em instalações da rede de operação

Na operação em tempo real, os erros na previsão de carga têm impacto na segurança e economicidade da operação do sistema elétrico. Se, por um lado, previsões superestimadas tendem a elevar os custos operacionais do sistema com a necessidade de

uma maior reserva operativa, por outro, previsões subestimadas contribuem para reduzir a reserva operativa do sistema, comprometendo a sua segurança.

A previsão de carga em um horizonte de curto prazo ou *short-term load forecasting* (STLF) constitui uma área de intensa pesquisa e conta com um inventário de técnicas que evolui continuamente, conforme se pode constatar na sequência cronológica iniciada por Gross & Galiana [1] e seguida em Liu et al [3], Lotufo & Minussi [4], Alfares & Nazeeruddin [5] e Hahn et al [6]. A evolução técnica reflete a busca incessante por métodos capazes de gerar previsões mais precisas e que permitam fazer um melhor uso dos recursos disponíveis e, portanto, contribuam para a operação otimizada do sistema elétrico. Dentre as metodologias mais utilizadas, destacam-se os métodos estatísticos e os métodos baseados em inteligência computacional e *machine learning*, em especial as redes neurais artificiais, a logica *fuzzy* ou nebulosa e as máquinas de vetor de suporte – SVM (*Suport Vector Machine*).

Independentemente da metodologia utilizada e da forma como esta é empregada, a construção de um modelo de previsão baseia-se no comportamento da carga no passado e das suas relações com outras variáveis explicativas que têm influência na demanda, como por exemplo, a temperatura. Portanto, para que o modelo utilizado tenha uma boa capacidade preditiva é imprescindível que os dados tenham a maior qualidade possível. No caso ideal, os dados devem estar livres de erros e perturbações provocadas por problemas no sistema de medição ou transmissão de dados, curtos-circuitos, falha de equipamentos e quaisquer outras causas não naturais que afetam a trajetória da carga. No entanto, a realidade está longe da situação ideal. Consequentemente, a presença de erros no conjunto de dados compromete o ajuste de qualquer modelo de previsão, o que contribui para a menor acurácia das previsões.

Para reduzir os efeitos de erros no sistema de medição e demais causas não naturais sobre a identificação dos modelos de previsão, a construção de qualquer modelo deve ser precedida pelo tratamento dos dados históricos da carga com a finalidade de corrigir ou atenuar os diferentes tipos de erros encontrados nos dados brutos. As técnicas de tratamento de dados também devem ser aplicadas aos registros históricos das variáveis explicativas consideradas no modelo de previsão.

O tratamento de dados de carga, denominado aqui por filtragem de dados, consiste em identificar e corrigir os diferentes tipos de erros verificados nos registros de carga (dados brutos). Dentre os erros mais comuns estão a presença de observações aberrantes ou *outliers* (Figura 1.1) [7], lacuna de dados (Figura 1.2) [7] e descontinuidades na curva de carga (Figura 1.3) [7].

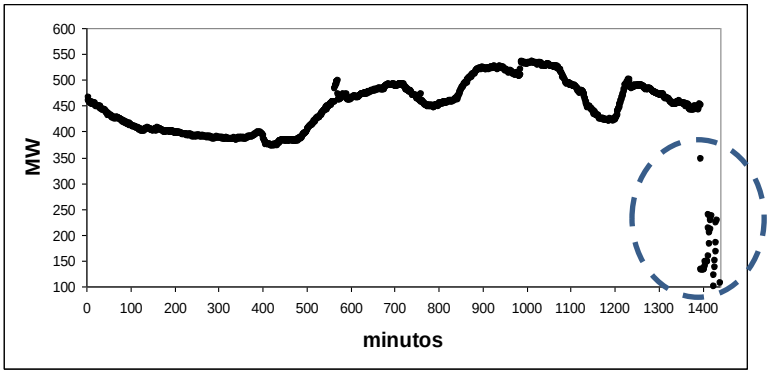


Figura 1.1 – Observações aberrantes ou *outliers*

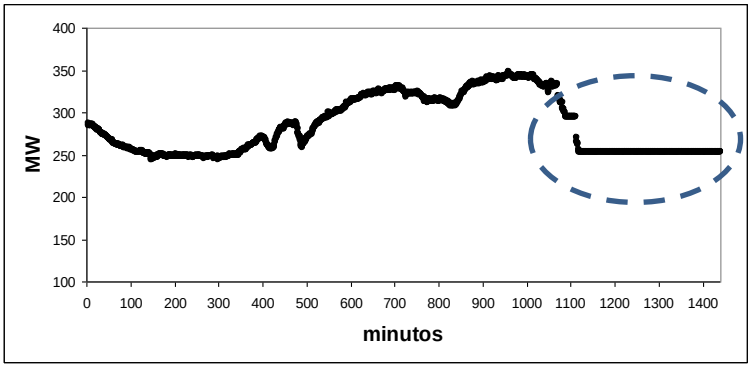


Figura 1.2 – Lacuna de dados

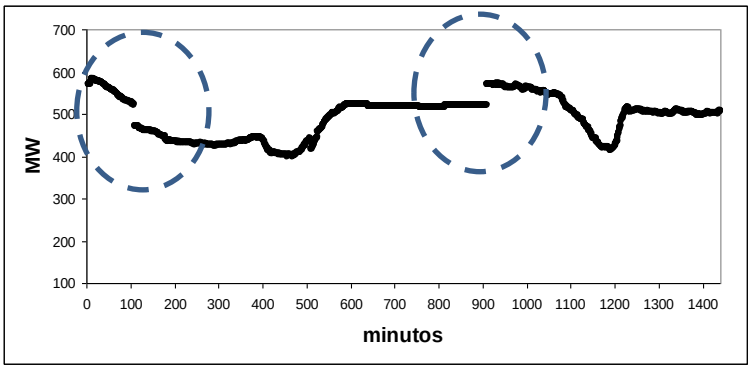


Figura 1.3 – Descontinuidades na curva

1.1.REVISÃO DA LITERATURA

Não raro, o tratamento de dados constitui-se na primeira etapa de qualquer análise estatística pois, frequentemente, os dados estão sujeitos a erros de medição ou coleta. As estatísticas descritivas, a análise exploratória de dados (*boxplot*) [8], o método de Grubbs [9] e Chauvenet [10] estão entre os primeiros e mais frequentes métodos utilizados na árdua tarefa de tratamento de dados. Contudo, na análise de séries

temporais, estas técnicas não são suficientes, sobretudo na análise de dados de carga em alta frequência (séries com resolução temporal horária, de meia hora ou com menor intervalo de integração), pois neste caso há necessidade de lidar com a tendência, vários regimes sazonais, feriados e dias especiais(feriados e outras datas com padrão de consumo diferenciados). Assim, as propostas para o tratamento de dados de carga recorrem ao uso de metodologias em que diversos métodos são alinhados com a finalidade de identificar e tratar os problemas presentes nos dados, conforme ilustrado por Suarez-Farinas et al. [11], Yang & Stenzel [12], Guirelli [13], Xiaoxing & Caixin [14], Grigoras et al. [15], Guan et al. [16], Chen et al [17], Qu et al. [18], Tang et al. [19], Gupta et al. [20], Guo et al. [21], Pessanha et al. [22] e Oliveira [23]. Vale destacar a reduzida quantidade de trabalhos com propostas para o tratamento de dados de carga quando contrastada com a numerosa quantidade de trabalhos sobre modelos de previsão de carga.

Em Suarez-Farinas et al. [11] apresenta-se uma metodologia para filtragem de séries temporais, com aplicação em series de alta frequência. A metodologia proposta utiliza um modelo linear dinâmico juntamente com o Fator de Bayes na identificação de outliers e o método Spline Cúbico Suavizado na correção dos valores faltantes.

Yang & Stenzel [12] ressalta a importância da qualidade dos dados para uma boa previsão de curto-prazo e sugere o uso da série de diferenças de segunda ordem na detecção de dados aberrantes e descontinuidades. A correção dos erros nos dados de carga é realizada com o auxílio da análise de agrupamentos (K Means) e da suavização LOWESS.

Guirelli [13] apresenta três alternativas para a filtragem de dados aberrantes: o método LOWESS, a filtragem através do uso de Transformadas de Fourier e, por último, a filtragem dos dados utilizando a Transformada *Wavelet*.

Xiaoxing & Caixin [14] propõe um modelo de correção de dados dinâmico e inteligente baseado em técnicas de mineração de dados. O trabalho combina a análise de agrupamentos *fuzzy* e de Kohonen com redes neurais de base radial para a identificação de problemas e tratamento dos dados. Grigoras [15] apresenta uma comparação de desempenho entre os métodos de vizinhos mais próximos (*k-Nearest Neighbors* ou kNN´s) e o *fuzzy clustering method* (FCM) no tratamento de dados faltantes.

Guan et al. [16] propõe um método baseado em redes neurais *wavelets*, que decompõe a carga em varias componentes de frequência, executa a transformação apropriada a cada componente e reconstrói o sinal através de uma rede neural.

Chen et al. [17] apresenta dois métodos de suavização para a correção automática de dados corrompidos ou faltantes, o método *B-Spline* e a suavização por *Kernel*. Qu et al. [18] propõe um método de tratamento de dados de carga que utiliza o método CURE de agrupamento para identificação de erros e o método da média móvel exponencialmente ponderada para a correção dos dados.

Tang et al. [19] apresenta uma nova abordagem para análise de dados de curva de carga que identifica padrões periódicos nos dados e reorganiza os dados a fim de facilitar a sua correção. Em Gupta et al. [20] é apresentado um estudo detalhado de várias metodologias utilizadas na detecção de *outliers*. Guo et al. [21] descrevem um método para detecção de *outliers* em dados de curvas de carga no qual emprega-se o estimador de Nadaraya Watson para suavizar os dados de curva de carga e obter um modelo da tendência e sazonalidade da série de carga. Os desvios dos registros de carga em relação ao modelo da série permite identificar os *outliers*. De forma resumida, dada uma curva de carga com alguma indicação de defeito, Guo et al. [21] procuram perfis semelhantes ao longo do histórico. Caso o perfil analisado seja similar à maioria dos demais perfis no histórico, então o perfil avaliado não tem *outliers*, caso contrário o perfil avaliado tem dados aberrantes que devem ser corrigidos. Na avaliação da similaridade Guo et al. [21] não recomenda o uso da distância euclidiana e emprega uma métrica denominada *Longest Common Sub-Sequence* (LCSS).

Já Pessanha et al. [22] propõe uma metodologia para filtragem de dados capaz de remover *outliers* e descontinuidades, bem como preencher lacunas de dados (*missing data*). A metodologia proposta pelos autores foi implementada no módulo de tratamento de dados do sistema de previsão de carga em uso pelo Operador Nacional do Sistema (ONS) [24]. Na metodologia proposta pelos autores, a identificação e remoção das observações aberrantes e descontinuidades é realizada pela aplicação sequencial de uma série de técnicas estatísticas, em especial, técnicas de análise exploratória de dados (histogramas e *boxplots*), regressão não paramétrica LOWESS (*Locally weighted scatterplot smoothing*) [25], intervalos de confiança e algoritmo DBSCAN (*Density Based Spatial Clustering of Applications with Noise*) [26]. Já o preenchimento das lacunas de dados é realizado com o auxílio da análise de agrupamentos FCM (F*uzzy Clustering Method*) [27] com a finalidade de identificar, no conjunto de dados não corrompidos, o padrão temporal da carga mais adequado ao preenchimento de uma lacuna.

Oliveira [23] apresenta uma metodologia para o tratamento de dados de carga na qual são empregadas *wavelets* na identificação de falhas nos registros de carga e o algoritmo de análise de agrupamentos *Hyperbolic Smoothing Clustering Method* (HSCM) no reconhecimento dos perfis típicos da curva de carga.

1.2.OBJETIVO

A presente dissertação tem por objetivo propor uma metodologia alternativa para a filtragem automática de dados de carga, baseada em um modelo que combina o uso de técnicas estatísticas e de inteligência artificial.

A identificação das observações aberrantes, descontinuidades e lacunas de dados segue as linhas gerais da metodologia proposta por Pessanha et al [22] e consiste na aplicação de técnicas estatísticas para análise exploratória de dados e uso dos métodos LOWESS e DBSCAN.

A maior contribuição desta dissertação reside na proposta de empregar o classificador Naive-Bayes [28, 29] e o Mapa de Kohonen [30, 31], uma rede neural não supervisionada, na identificação de padrões adequados ao preenchimento de lacunas nos dados de carga corrompidos. Vale destacar que as duas técnicas supracitadas destinam-se a resolver dois problemas distintos, frequentemente encontrados na análise de dados: o classificador Naive-Bayes realiza uma classificação supervisionada (*classification*), enquanto o Mapa de Kohonen realiza uma classificação não supervisionada (*clustering*).

Adicionalmente, a dissertação descreve uma abordagem de simulação de falhas em dados de curvas de carga para avaliar a metodologia proposta. A partir de dados de carga não corrompidos, as simulações introduziram aleatoriamente alguns dos problemas mais frequentemente observados nos registros de carga. Os perfis de carga artificialmente corrompidos pela simulação foram tratados pela metodologia proposta com a finalidade de verificar se as correções propostas são compatíveis com os registros originalmente observados. Assim, os dados simulados são filtrados utilizando a metodologia proposta e então comparados com os dados reais, de forma a se obter um índice de desempenho do modelo. Para ilustrar a metodologia proposta foram realizados experimentos computacionais com dados de carga reais provenientes do Sistema Interligado Nacional (SIN).

Em seguida, a mesma estratégia de avaliação baseada em simulações foi empregada com a finalidade de comparar a metodologia proposta com outras alternativas para o preenchimento de lacunas de dados.

1.3.ESTRUTURA

A presente dissertação de mestrado foi organizada em cinco capítulos. No Capítulo 1 foi introduzido o problema da filtragem de dados de carga, assim como foi realizada a apresentação de uma revisão bibliográfica sobre o tema e o objetivo da dissertação.

Na sequência, no Capítulo 2 apresenta-se uma base teórica que contempla as técnicas estatísticas e as técnicas de inteligência artificial utilizadas na metodologia proposta e que são necessárias tanto no processo de identificação quanto na correção do problema.

Em seguida, no Capítulo 3, apresenta-se uma visão geral da metodologia proposta para a filtragem dos dados de carga, acompanhada de uma descrição detalhada da finalidade de cada um de seus módulos.

A análise exploratória de dados conduzida com dados reais de carga do SIN e os resultados das simulações realizadas nos experimentos computacionais são descritos no Capítulo 4.

Por fim, no Capítulo 5 são resumidas as principais conclusões a respeito da metodologia proposta e possíveis aplicações em trabalhos futuros.

2. MÉTODOS PARA ANÁLISE DE DADOS

O tratamento ou filtragem dos dados de carga envolve o uso de diversas técnicas estatísticas e de inteligência artificial na análise de grandes quantidades de dados. Neste capítulo serão abordados os métodos utilizados pela metodologia proposta.

2.1 SUAVIZAÇÃO LOWESS

O LOWESS (*Locally weighted regression scatter plot smoothing*) [25] é um método de regressão local não paramétrico, que obtém uma curva suavizada através de sucessivas regressões lineares em uma determinada vizinhança. Para cada instante de tempo da série observada, cada ponto pertencente a sua região de vizinhança é ponderado de forma que pontos mais próximos recebem pesos maiores. Então, uma estimativa suavizada para este instante de tempo é calculada através do método dos mínimos quadrados ponderados. Para encontrar a curva suavizada basta percorrer todos os pontos da série, movendo a janela de vizinhança em questão. A seguir, tem-se uma descrição das etapas do algoritmo LOWESS:

1) Seja y um vetor contendo um perfil diário de carga e x um contador de tempo, por exemplo, para resolução temporal de 1 minuto tem-se que $x = \{1,2,3,...,1440\}$.

2) Para cada instante de tempo $x_0 \in x$ identifique os k instantes x_i ($i=1,k$) na vizinhança de x_0 e denote este conjunto por $N(x_0)$. O tamanho k da janela de vizinhança é um parâmetro de entrada do LOWESS.

3) Calcule a maior distância entre x_0 e o ponto x_i dentro da janela $N(x_0)$:

$$\Delta(x_0) = máximo_{x_i \in N(x_0)} \|x_0 - x_i\|$$

4) Pondere cada par (x_i,y_i), x_i em $N(x_0)$ com base na função tricúbica [25]:

$$peso_i(x_0) = W\left(\frac{\|x_0 - x_i\|}{\Delta(x_0)}\right), \text{ onde } W(u) = \begin{cases} (1-u^3)^3 & 0 \le u < 1 \\ 0 & \text{caso contrário} \end{cases}$$

5) Aplique o estimador de mínimos quadrados ponderados ao conjunto de observações que pertencem à vizinhança $N(x_0)$ para obter uma estimativa $\hat{y}$ de y no ponto x_0.

6) Repita os passos de 3 a 6 para cada instante de tempo no vetor x.

Na Figura 2.1 pode-se visualizar a função $W(u)$ utilizada para a ponderação das observações na vizinhança de diferentes instantes de tempo x_0. Vale destacar que outras funções de ponderação podem ser utilizadas no lugar da função tricúbica. Em cada função de ponderação ilustrada na Figura 2.1 foi considerada uma janela de 120 minutos.

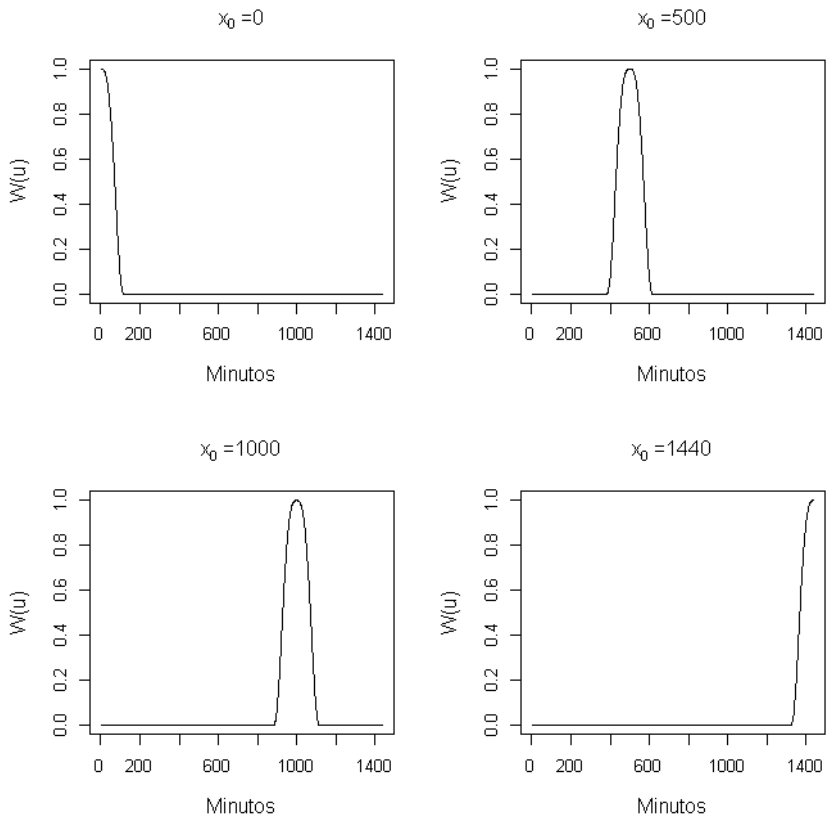


Figura 2.1 – Função W(u) utilizada para ponderar as observações na vizinhança de x_0.

Na Figura 2.2 tem-se a ilustração um ponto x_i qualquer da curva diária e sua respectiva região de vizinhança. Apenas os pontos nesta região são contabilizados na estimativa do ponto suavizado. Neste caso, a janela de vizinhança é de uma hora, ou seja, 60 pontos.

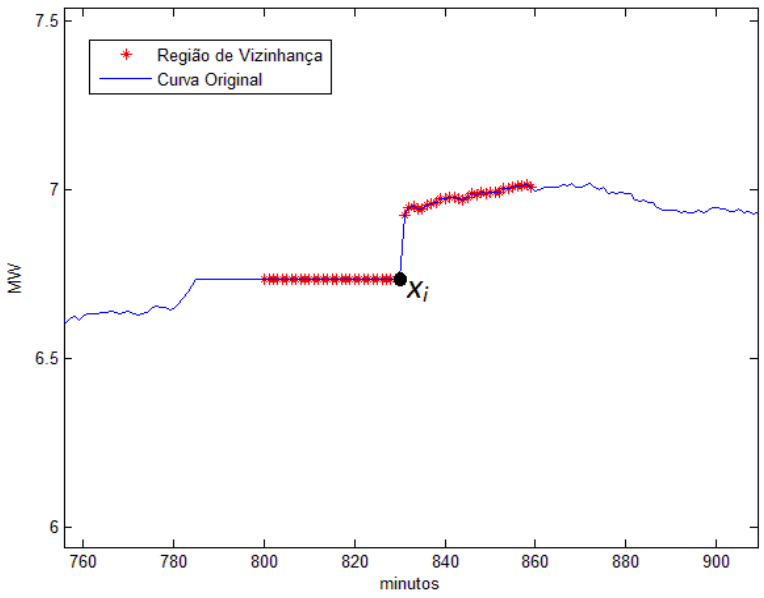


Figura 2.2 – Região de vizinhança de um ponto na curva de carga diária.

O resultado da aplicação do procedimento LOWESS na suavização de perfis diários de carga é ilustrado na Figura 2.3. Neste caso, a curva diária possui discretização de 1 minuto, totalizando 1440 pontos. Na Figura 2.3 é possível visualizar a curva original, a curva suavizada resultante de uma janela de vizinhança com 60 pontos e a curva suavizada resultante uma janela de vizinhança com 120 pontos.

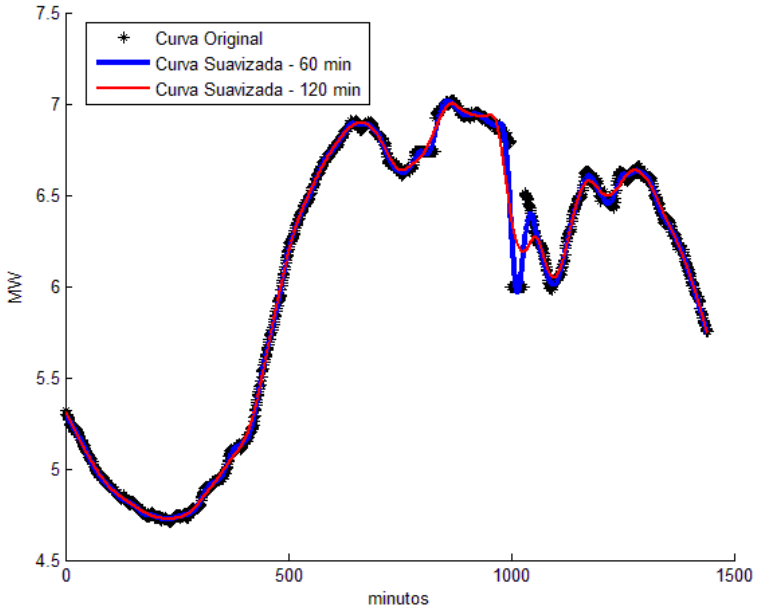


Figura 2.3 – Suavização Lowess em curva de carga diária.

2.2 MAPAS AUTO-ORGANIZÁVEIS DE KOHONEN

As redes neurais artificiais (RNAs) são modelos matemáticos inspirados nos princípios de funcionamento dos neurônios biológicos e na estrutura do cérebro, i.e., são técnicas computacionais que se assemelham a estrutura neural dos organismos inteligentes e adquirem conhecimento através de experiências [31]. O conhecimento absorvido através de exemplos é armazenado em pesos sinápticos que interligam os neurônios. A Figura 2.4 apresenta um diagrama esquemático do primeiro um neurônio artificial proposto por McCulloch e Pitts, na década de 40 [32].

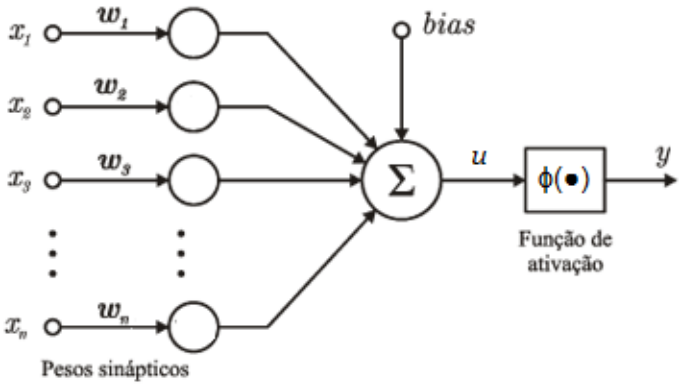


Figura 2.4 – Neurônio Artificial clássico.

No diagrama da Figura 2.4 é possível identificar os principais componentes de um neurônio artificial:

- As sinapses ou elos de conexão, que são caracterizadas por um peso próprio. Basicamente, um sinal x_i na entrada é multiplicado pelo seu respectivo peso w_i;
- Um somador que serve para adicionar os sinais de entrada multiplicados pelos seus respectivos pesos;
- Uma função de ativação que pode limitar a amplitude de saída do neurônio, restringindo o intervalo de amplitude do sinal de saída a um valor finito.

O neurônio artificial apresentado na Figura 2.4 possui uma saída y que pode ser escrita segundo (2.1):

$$y = \phi\left(\sum_{i=1}^{n} x_i w_i + b\right) \tag{2.1}$$

Sendo:

x_i – sinais de entrada do neurônio;

w_i – pesos sinápticos do neurônio;

b – *bias*;

ϕ (•) – função de ativação;

y – sinal de saída do neurônio.

O *bias* é um parâmetro externo que serve basicamente para aumentar o grau de liberdade dos ajustes dos pesos. No entanto, ele pode ser incorporado ao neurônio como se fosse uma das entradas do sistema. Isto é possível adaptando (2.1) para (2.2):

$$y = \phi\left(\sum_{i=0}^{n} x_i w_i\right) \tag{2.2}$$

Fixando x_0=1 e w_0 = b, pode-se reformular o modelo do neurônio artificial para o da Figura 2.5:

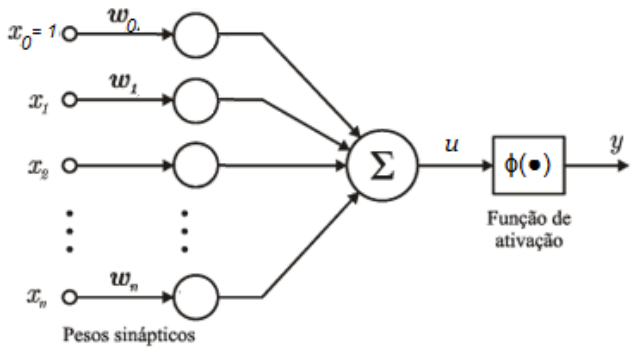


Figura 2.5 – Neurônio Artificial com Bias incorporado.

A função de ativação pode apresentar características lineares ou não lineares, de forma a determinar a saída de um neurônio a partir do seu potencial de ativação [11]. Dentre os tipos de função de ativação pode-se destacar a função linear, a função degrau, a função sigmoide e a função gaussiana. Estes quatros tipos de função de ativação são vistos na Figura 2.6:

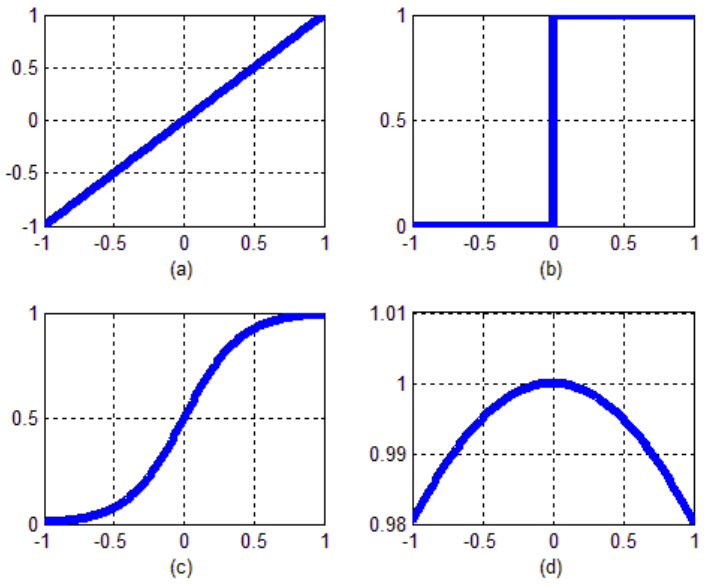


Figura 2.6 - Funções de ativação : (a) Função Linear. (b) Função Degrau. (c) Função Sigmoidal. (d) Função Gaussiana

As várias funções de ativação tornam as RNAs capazes de lidar com problemas lineares ou não lineares. As RNAs também possuem grande capacidade de generalização e adaptabilidade, i.e, elas são capazes de resolver problemas que não fazem parte da base de conhecimento fornecida durante o seu treinamento.

As grandes dificuldades do uso das redes neurais residem na escolha da topologia da rede, que não segue uma teoria exata e acontece de forma empírica, e a na interpretação dos conceitos aprendidos pelas RNAs, pois estes são codificados em seus pesos formando uma espécie de caixa preta.

As redes neurais artificiais podem ser classificadas em relação à maneira pela qual elas adquirem conhecimento: redes supervisionadas e redes não supervisionadas [31].

- Redes supervisionadas – A rede é treinada utilizando pares de treinamento que contêm vetores com valores de entrada e saída. Um vetor de entrada é inserido à rede e a saída correspondente é calculada e comparada com o respectivo sinal de saída desejado, gerando um sinal de erro. O objetivo do algoritmo de treinamento consiste em ajustar os pesos sinápticos de tal forma a minimizar este erro entre a saída gerada pela rede e a saída que se deseja. Portanto, em classificadores supervisionados, as classes desejadas são informadas ao modelo, e este é responsável por classificar as entradas.

- Redes não supervisionadas – É um tipo de aprendizado onde não existe
um vetor de saída desejado. São utilizados vetores de entrada para que a rede identifique padrões e agrupe esses padrões semelhantes em classes (clusters). Portanto, em classificadores não supervisionados, as classes não são informadas, sendo o modelo responsável por determinar as classes e classificar as entradas, ou seja, a RNA executa uma análise de agrupamentos (*cluster analysis*).

Os mapas auto-organizáveis de Kohonen são um exemplo de redes neurais não supervisionadas. Neste tipo de rede os neurônios estão organizados em uma grade ou reticulado, normalmente unidimensional ou bidimensional, como mostrado na Figura 2.7.

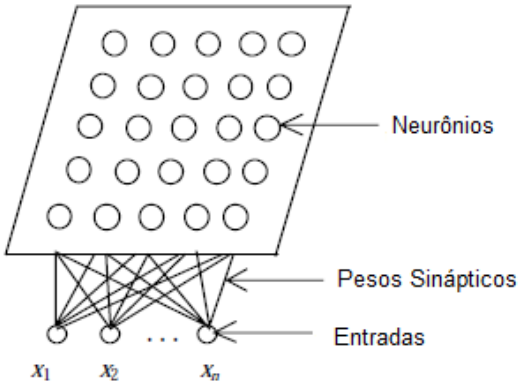


Figura 2.7 - Mapa de Kohonen Bidimensional.

Os mapas de Kohonen fazem parte de um conjunto de redes neurais que são baseadas em modelos de competição. Este nome é devido ao fato dos neurônios competirem entre si para saber quem gera o maior valor de saída, ou seja, qual o vetor de pesos é mais semelhante ao padrão de entrada apresentado. Esta medida é geralmente calculada através da distância entre os exemplos de entrada e os pesos sinápticos dos neurônios. Uma das alternativas é utilizar a distância euclidiana, dada por (2.3):

$$d_{kij} = \sqrt{\sum_{j=1}^{n} (x_k - w_{ij})^2} \tag{2.3}$$

Sendo:

d_{kij} – distância euclidiana;

x_k - exemplos de entrada;

w_{ij} – pesos sinápticos que ligam as entradas aos neurônios da grade de saída;

n – numero de entradas;

Ao longo do processo de treinamento da rede, dado um exemplo de entrada, o neurônio vencedor será aquele que apresentar menor distância em relação ao exemplo. Uma vez encontrado o neurônio vencedor, a atualização dos pesos sinápticos pode ocorrer de duas maneiras: modelo “vencedor leva tudo” ou modelo “vencedor leva parte”. No modelo “vencedor leva tudo” apenas os pesos sinápticos do neurônio vencedor serão atualizados. A regra de atualização do neurônio vencedor é dada por (2.4):

$$w_{ij}(novo) = w_{ij}(antigo) + \alpha.\Delta.\left(x_j - w_{ij}(antigo)\right) \quad (2.4)$$

Sendo:

w_{ij} – pesos sinápticos que ligam as entradas ao neurônio vencedor;

x_j – exemplos de entrada;

α – taxa de aprendizagem;

Δ – função de vizinhança adotada;

No modelo "vencedor leva parte", além do neurônio vencedor, os neurônios vizinhos a ele também tem seus pesos sinápticos ajustados. Neste caso, é necessário estabelecer um critério para definir o limite desta vizinhança. Uma opção é definir este limite em função da distância de um neurônio até o neurônio vencedor através de uma função gaussiana, por exemplo, como visualizado na Figura 2.8 e dada por:

$$\Delta = e^{\left(\frac{-d_{ij}^2}{2\sigma^2}\right)} \quad (2.5)$$

Sendo:

d_{ij} - distância do neurônio i ao neurônio j.

σ – desvio padrão da gaussiana.

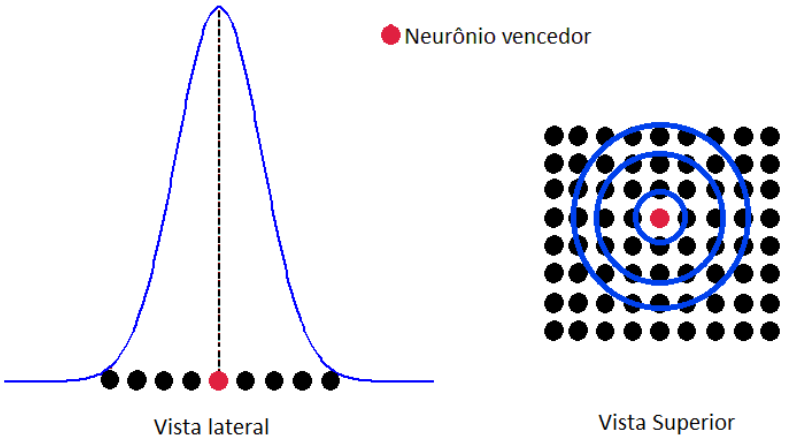


Figura 2.8 – Função de vizinhança gaussiana.

Na fase de treinamento é necessário reduzir a região de vizinhança à medida que aumentam o número de iterações, como observado na Figura 2.9. Com isso, espera-se obter uma organização do mapa topológico, ou seja, os padrões detectados por um

determinado neurônio estarão relacionados com a posição do neurônio na grade. Há ainda uma fase de ajuste fino do mapa topológico, onde se utiliza uma taxa de aprendizagem baixa e uma região de vizinhança pequena [31].

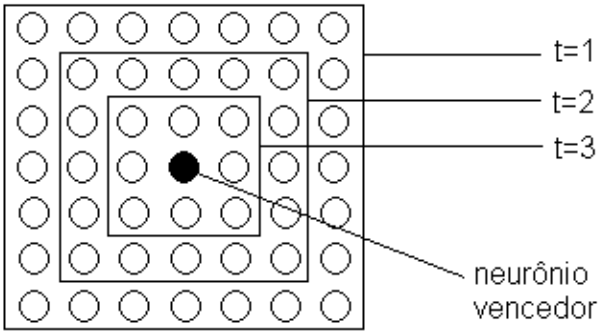


Figura 2.9 - Redução gradual da região de vizinhança desde o instante t=1 até o instante t=3.

2.3 CLASSIFICADOR NAIVE BAYES

Trata-se de um classificador probabilístico baseado no teorema de Bayes, sendo denominado ingênuo (*naive*) por assumir que, para uma dada classe qualquer, a presença ou a ausência de uma característica independe da presença ou da ausência de outras características [28]. Portanto, o classificador *Naive Bayes* assume que as característica são condicionalmente independentes. Apesar desta premissa simplista, o classificador apresenta excelente desempenho em várias tarefas de classificação.

Em uma situação geral do problema de classificação, os objetos podem ser classificados em uma classe entre m classes possíveis $C_1, C_2, \dots, C_m$. Cada objeto é descrito por p características codificadas por meio de variáveis categóricas x_i, $i = 1, \dots, p$, em um vetor $X = (x_1, x_2, \dots, x_p)$. A proposta do *Naive Bayes* é classificar cada objeto na classe com a maior probabilidade a posteriori condicionada a X. Ou seja, o classificador associa o objeto com características X à classe C_k se, e somente se, $P(C_k|X) > P(C_j|X)$ para $\forall$ k $\neq$ j = 1,...,m. A classe C_k para a qual $P(C_k|X)$ é maximizada denomina-se hipótese de máxima a posteriori.

Pelo Teorema de Bayes a probabilidade a posteriori $P(C_j|X) = P(C_j|x_1, x_2, \dots, x_p)$ pode ser expressa em função da probabilidade a priori do objeto pertencer à classe C_j, $P(C_j)$, da verossimilhança $P(x_1, x_2, \dots, x_p|C_j)$ e da evidência $P(x_1, x_2, \dots, x_p)$:

$$P(C_j|x_1, x_2, \dots, x_p) = \frac{P(x_1, x_2, \dots, x_p|C_j) * P(C_j)}{P(x_1, x_2, \dots, x_p)} \qquad (2.6)$$

O denominador em (2.5) é constante, pois depende apenas das variáveis categóricas $x_1, x_2, \dots, x_p$, cujos valores são conhecidos. Logo, o denominador é um fator de escalonamento que pode ser ignorado para fins de classificação. Já o produto no numerador é a probabilidade conjunta $P(x_1, x_2, \dots, x_p, C_j)$ e que pode ser expressa da seguinte maneira após sucessivas aplicações do Teorema da Probabilidade Condicional (2.6):

$$P(C_j) * P(x_1|C_j) * P(x_2|C_j, x_1) * P(x_3|C_j, x_1, x_2) * \dots * P(x_p|C_j, x_1, x_2, \dots, x_{p-1}) \quad (2.7)$$

A estimativa da probabilidade conjunta $P(x_1, x_2, \dots, x_p, C_j)$, em geral, é difícil, pois demanda uma amostra muito grande quando o vetor X envolve muitas variáveis. Para contornar esta dificuldade o estimador *Naive Bayes* assume que cada variável $x_i, i = 1,\ p$ é condicionalmente independente das demais variáveis do vetor de características X, dada uma classe C_j. Por meio desta simplificação, obtém-se a seguinte aproximação para a probabilidade conjunta (2.7):

$$P(C_j) * P(x_1|C_j) * P(x_2|C_j) * P(x_3|C_j) * \dots * P(x_p|C_j) \quad (2.8)$$

Assim, a probabilidade a posteriori é proporcional ao produto da probabilidade a priori $P(C_j)$ e das probabilidades condicionais $P(x_i|C_j), i = 1, \dots, p$ (2.8):

$$P(C_j|x_1, x_2, \dots, x_p)\ \alpha\ P(x_1|C_j) * P(x_2|C_j) * \dots * P(x_p|C_j) * P(C_j) \quad (2.9)$$

A premissa de independência condicional das características simplifica o problema e permite a construção de um classificador computacionalmente simples e útil quando a dimensionalidade do espaço de características é elevada. Contudo, a aproximação não deve ser utilizada se o objetivo consistir em calcular probabilidades.

A simplicidade do classificador *Naive Bayes* implica no reduzido número de parâmetros que devem ser calculados: as probabilidades a priori de cada classe e as probabilidades condicionais de cada atributo, dada uma classe. Por exemplo, para o caso com m classes e em que cada variável categórica $x_i, i = 1, \dots, p$tenha até r categorias possíveis, o total de probabiliades a serem calculadas é igual a $(m - 1) + prm$.

A probabilidade a priori de cada classe corresponde à frequência relativa da classe no conjunto de dados de treinamento. Já as probabilidades condicionais $P(x_i|C_i)\forall_i = 1, ..., p$ e $\forall_j = 1,...,m$ podem ser calculadas a partir das instâncias do conjunto de treinamento por meio das frequências relativas (2.9):

$$P(x_i|C_j) = \frac{f_{ij}}{f_j} \tag{2.10}$$

Onde f_j é a frequência absoluta da classe C_j e f_{ij} é a frequência absoluta das instâncias na classe C_j e com o i-ésimo atributo igual a x_i, ambas calculadas a partir dos dados no conjunto de treinamento.

Em posse do modelo probabilístico, o classificador pode então calcular as probabilidades a posteriori de cada classe para um determinado vetor $X = (x_1, x_2, \dots, x_p)$ e classificá-lo na classe com a maior probabilidade a posteriori.

Um exemplo básico da utilização do classificador *naive bayes* [28] é o da decisão se uma pessoa vai ou não jogar tênis em um determinado dia, baseando-se nas informações de dias anteriores. Nos 10 dias utilizados para o treinamento do classificador são consideradas as condições do tempo (sol, nublado ou chuvoso), a temperatura (alta, normal ou baixa), umidade (alta ou normal), vento (forte ou fraco) e por último se a pessoa jogou tênis naquele dia. As informações estão apresentadas na Tabela 2.1,

Tabela 2.1 - Características dos dias passados – Classificador Naive Bayes.

DIA	TEMPO	TEMPERATURA	UMIDADE	VENTO	JOGOU TENIS
D1	SOL	ALTA	ALTA	FRACO	NÃO
D2	SOL	ALTA	ALTA	FORTE	NÃO
D3	NUBLADO	ALTA	ALTA	FRACO	SIM
D4	CHUVA	NORMAL	ALTA	FRACO	SIM
D5	CHUVA	BAIXA	NORMAL	FRACO	NÃO
D6	CHUVA	BAIXA	NORMAL	FORTE	NÃO
D7	NUBLADO	BAIXA	NORMAL	FORTE	SIM
D8	SOL	NORMAL	ALTA	FRACO	NÃO
D9	SOL	BAIXA	NORMAL	FRACO	SIM
D10	CHUVA	NORMAL	NORMAL	FRACO	SIM

A partir do histórico, pode-se avaliar se a pessoa irá jogar tênis em um dia qualquer, com base nas mesmas informações. Por exemplo: em um dia de sol, com temperatura baixa, umidade alta e vento forte, a pessoa irá jogar tênis? Para chegar a uma resposta, deve-se estimar as probabilidades através da frequência de cada característica no conjunto de treinamento, como demonstrado a seguir:

$$P(sim) = \frac{5}{10} = 0.5$$
$$P(não) = \frac{5}{10} = 0.5$$

$$P(sol|sim) = \frac{1}{5} = 0.2$$
$$P(sol|não) = \frac{3}{5} = 0.6$$

$$P(temperatura\ baixa|sim) = \frac{2}{5} = 0.4$$
$$P(temperatura\ baixa|não) = \frac{2}{5} = 0.4$$

$$P(umidade\ alta|sim) = \frac{2}{5} = 0.4$$
$$P(umidade\ alta|não) = \frac{3}{5} = 0.6$$

$$P(vento\ forte|sim) = \frac{1}{5} = 0.2$$
$$P(vento\ forte|não) = \frac{2}{5} = 0.4$$

Finalmente, as probabilidades do sim e do $não$ são obtidas através dos seguintes produtórios:

sim:
$P(sim)\ .\ P(sol|sim)\ .\ P(temperatura\ baixa|sim)\ .\ P(umidade\ alta|sim)\ .$
$P(vento\ forte|sim)$ = 0.0032

$não$:
$P(não)\ .\ P(sol|não)\ .\ P(temperatura\ baixa|não)\ .\ P(umidade\ alta|não)\ .$
$P(vento\ forte|não)$ = 0.0288

Portanto, como o produto $não$ obteve maior valor de saída, o classificador *naive bayes* prevê que a pessoa não irá jogar tênis neste dia.

2.3 DBSCAN

Proposto por Ester et al. [33] em 1996, o DBSCAN (*Density Based Spatial Clustering of Applications with Noise*) é um algoritmo para análise de agrupamentos

com uma abordagem baseada em densidade. Nesta abordagem, os agrupamentos ou clusters são regiões de alta densidade separadas por regiões de baixa densidade [26].

Sem perda de generalidade, considere um conjunto de objetos caracterizados por p atributos (x_1,...,x_p). Assim, cada objeto pode ser representado por um ponto no espaço R^p.

Dado um conjunto de pontos, o DBSCAN classifica-os em três categorias: centrais (*core-points*), pontos de limite (*border points*) e ruído (*noise points*). As três categorias são ilustradas na Figura 2.10 na qual se considerou MinPts = 7.

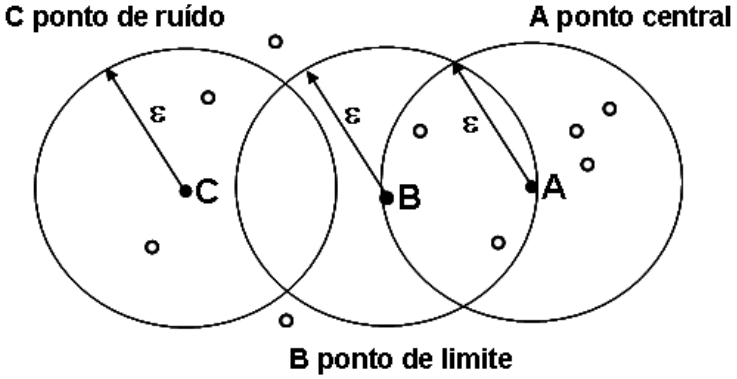


Figura 2.10 - Pontos central, de limite e ruído.

Um ponto *p* é classificado como central se em uma esfera de raio ε centrada no objeto *p* (ε-vizinhança de *p*) há pelo menos MinPts objetos dentro da esfera, incluindo o ponto *p*. A ε-vizinhança de um ponto p, denotada por Nε (*p*), é o conjunto de pontos q tal que a distância (p,q) < ε. Os parâmetros ε e MinPts são especificados a priori.

Por sua vez, um ponto *q* é um ponto de limite se for alcançável a partir de um ponto central *p*. Todo ponto não alcançável a partir de outro ponto central ou de limite é classificado como sendo um *outlier*.

Se *p* é um ponto central, ele forma um cluster com todos os pontos na sua ε-vizinhança. Assim, cada cluster tem pelo menos 1 ponto central e quaisquer dois pontos centrais com distância menor ou igual a ε são colocados no mesmo cluster. Qualquer ponto de limite que está na ε-vizinhança de um ponto central é colocado no mesmo cluster do ponto central. Os pontos não pertencentes às ε-vizinhanças dos pontos centrais são classificados como ruído.

Resumindo, a execução do DBSCAN pode ser realizada por meio do seguinte algoritmo:

1) Classifique todos os pontos como centro, de limite ou de ruído.

2) Adicione uma aresta entre todos os pontos de centro cujas ε-vizinhanças se sobreponham.

3) Torne cada grupo de pontos de centro conectados um *cluster*.

4) Atribua cada ponto de limite a um dos *clusters* de pontos de centro.

Conforme indicado no algoritmo acima, os pontos de ruído não são classificados em nenhum dos *clusters*, logo os pontos de ruído podem ser interpretados como sendo *outliers*.

3. METODOLOGIA PROPOSTA

Neste capítulo é apresentada a metodologia proposta para a filtragem de dados de carga. Primeiro, é exibida uma visão geral do modelo. Depois, cada etapa do modelo é descrita separadamente em cada seção.

3.1 FLUXOGRAMA DA METODOLOGIA PROPOSTA

O fluxograma da metodologia proposta pode ser visualizado na Figura 3.1. Primeiro, os dados são aquisitados através de um sistema externo de medição de curvas de carga diárias. Depois o modelo verifica se há curvas com lacunas ou dados aberrantes (*outliers*) através da construção de histogramas, aplicação do método LOWESS e cálculo das diferenças de primeira ordem.

As curvas que não possuem lacunas de dados ou dados aberrantes (*outliers*) são então classificadas utilizando a técnica de mapas auto-organizáveis de Kohonen (SOM). O centroide de cada classe de curvas agrupadas é identificado como sendo um perfil típico de curva diária.

Após a identificação dos perfis típicos, o próximo passo consiste em calcular as probabilidades condicionais a priori das curvas classificadas em cada classe(*cluster*) em relação a cada uma das varíaveis(feriados, horário de verão, dia da semana e mês do ano). De posse destas probabilidades condicionais e dos perfis típicos, o classificador Naive Bayes é utilizado para fazer uma associação probabilística de cada uma das curvas do histórico(mesmo aquelas que não possuem lacuna de dados ou outliers) a um dos perfis típicos identificados.

Na sequência, tem-se a correção das curvas que apresentaram lacunas de dados identificadas na primeira etapa. As curvas são corrigidas utilizando as estimativas fornecidas pelo perfil típico mais provável.

Em seguida o modelo identifica e corrige as sequências de dados atípicas, que são aquelas que apesar de não conterem lacunas ou dados aberrantes, apresentam um comportamento diferente do esperado. Este processo se dá através da construção de intervalos de confiança e do uso de *boxplots*.

Por fim, é realizada a identificação e a correção de possíveis descontinuidades nas curvas de carga por meio da projeção dos segmentos da curva avaliada sobre o seu perfil típico mais provável.

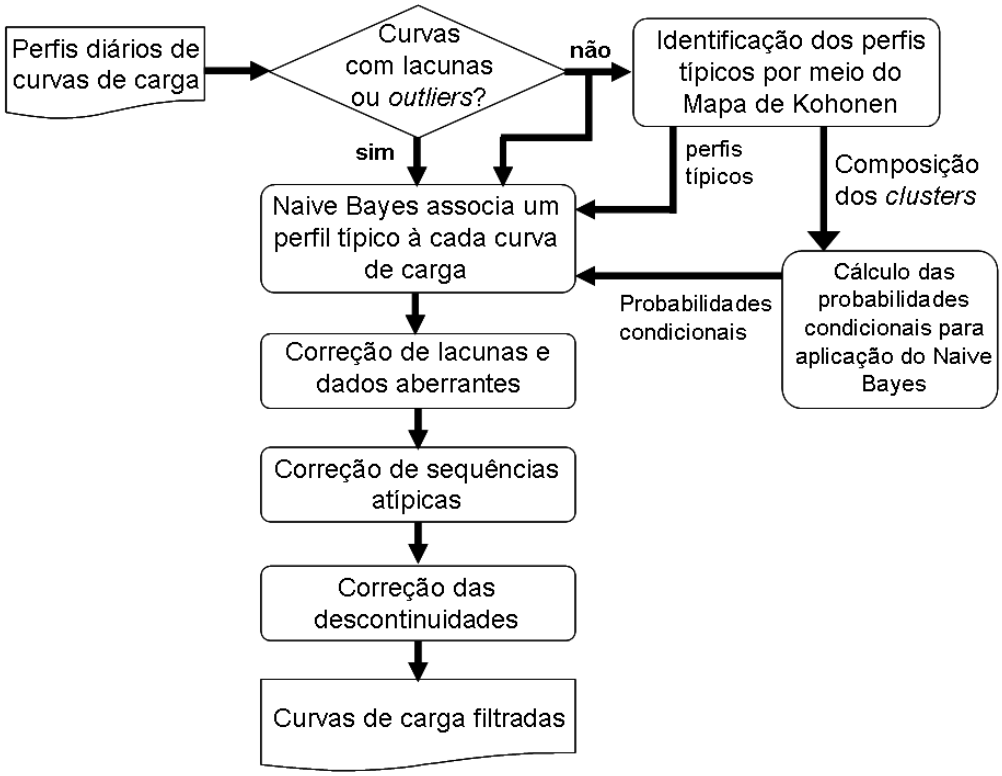


Figura 3.1 – Metodologia proposta para o tratamento dos dados de carga.

3.2 IDENTIFICAÇÃO DE CURVAS COM OBSERVAÇÕES ABERRANTES E LACUNAS DE DADOS

3.2.1– TRATAMENTO DE CURVAS COM DADOS ABERRANTES

Nesta primeira etapa constrói-se a distribuição de frequências para cada curva de carga diária e procura-se pela existência de classes com frequências nulas, pois a presença de tais classes sugere a possibilidade de valores aberrantes e descontinuidades nos registros de carga. Por exemplo, considere a curva de carga ilustrada na Figura 3.2 [7] com valores aberrantes no final do dia. A distribuição de frequência e o histograma das demandas desta curva de carga são apresentados na Figura 3.3 [7] e na Tabela 3.1, respectivamente.

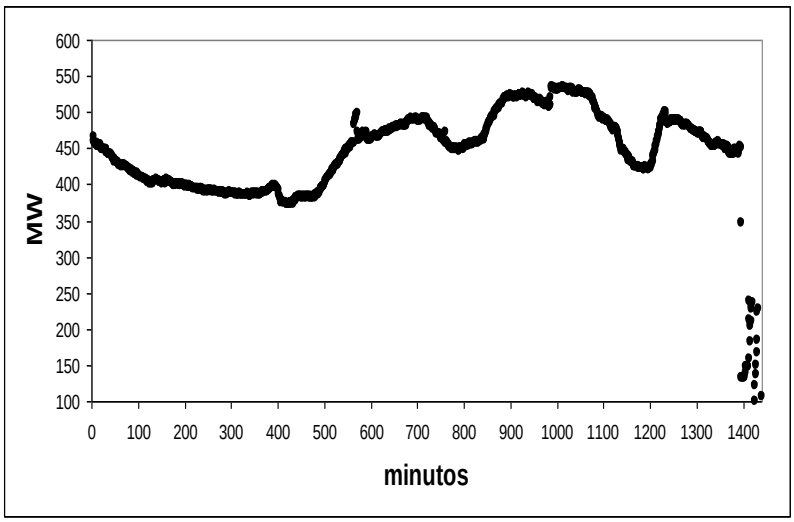


Figura 3.2 – Curva de carga com *outliers*.

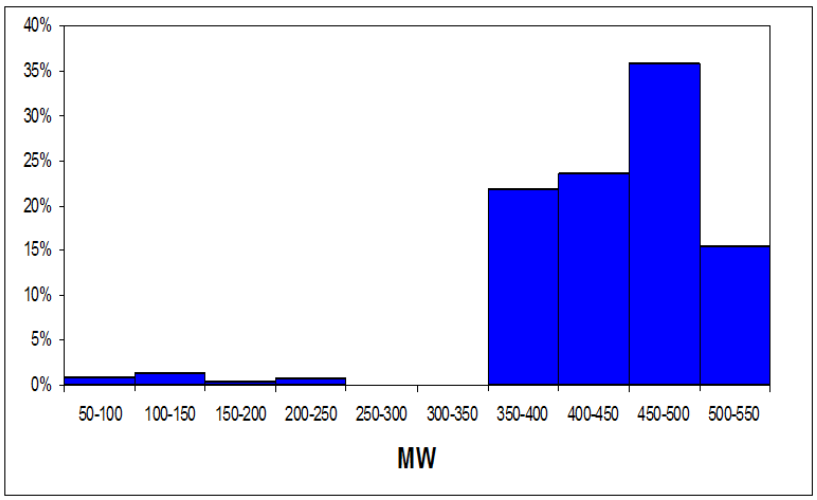


Figura 3.3 – Distribuição de frequências da demanda de uma curva de carga.

A presença de classes de demanda com frequência nula, conforme ilustrado na Figura 3.3 e na Tabela 3.1, sugere a existência de descontinuidades e de observações aberrantes na curva de carga. Estes fatos são observados na nuvem de pontos no lado direito da curva na Figura 3.2. Adicionalmente, categorias com apenas uma observação, por exemplo, a categoria com demanda entre 300 e 350 MW, pode ser entendida como um ruído e deve ter seu valor corrigido para o valor da demanda imediatamente anterior. Caso a observação seja referente ao primeiro minuto da curva diária, o valor é corrigido para o valor da demanda no minuto seguinte.

Tabela 3.1 - Distribuição de frequência da demanda.

Classe de demanda (MW)	Frequência absoluta
50 a 100	13
100 a 150	21
150 a 200	5
200 a 250	9
250 a 300	0
300 a 350	1
350 a 400	313
400 a 450	340
450 a 500	517
500 a 550	221
Total	1440

A existência de uma classe com frequência nula indica que os valores da demanda podem ser agrupados em duas categorias, neste caso: demandas abaixo de 250 MW e demandas acima de 300 MW, conforme ilustrado na Tabela 3.1. A categoria com demandas abaixo de 250 MW contém 48 observações (3,3% dos 1440 minutos), enquanto a categoria com demandas acima de 300 MW contém 1392 observações (96,7%). Cada categoria é formada pela agregação de classes adjacentes com frequências não nulas. Neste caso as classes foram agrupadas em duas categorias, pois a distribuição de frequências tem apenas uma classe com frequência nula. No caso em que todas as classes apresentam frequências não nulas a agregação das classes produz uma única categoria.

Como a maior parte das observações concentra-se na categoria com demanda acima de 300 MW, então se conclui que a categoria com demandas abaixo de 250 MW é formada por observações aberrantes e que devem ser corrigidas. Após algumas análises conduzidas com uma amostra de medições considerou-se que categorias com menos de 144 observações (10% do total de observações da curva de carga diária com resolução temporal de 1 minuto) são consideradas categorias atípicas. Logo, as demais classes são classificadas como normais.

A correção das observações classificadas nas categorias com demandas abaixo de 250 MW consiste em substituir os respectivos valores pela média das observações classificadas como normais, neste caso são as demandas acima de 300 MW. A correção das observações classificadas na categoria atípica (demandas menores que 250 MW) gera uma sequência de valores constantes em substituição ao conjunto de pontos atípicos, conforme ilustrado na Figura 3.4. Esta sequência de valores constantes será corrigida em etapas mais avançadas da filtragem. Vale ressaltar que o tamanho da grade do histograma em MW é um parâmetro de entrada do tratamento de dados.

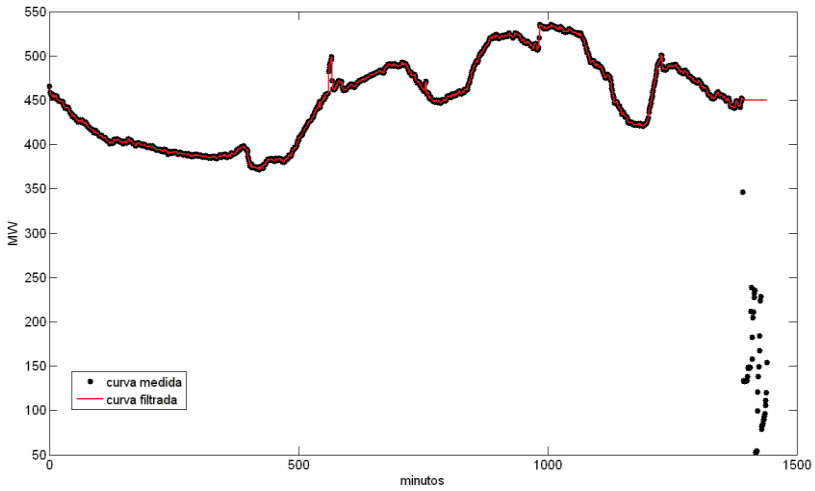


Figura 3.4 - Curva medida e curva filtrada.

3.2.2 – IDENTIFICAÇÃO DE CURVAS COM LACUNAS DE DADOS

Além das observações aberrantes, as curvas de carga também podem apresentar lacunas de observações (*missing data*) em função das falhas esporádicas do sistema de medição. Para evitar as lacunas de observações durante uma falha, o sistema de medição passa a repetir a cada instante o último valor medido da demanda instantânea até que o sistema seja restabelecido. Este procedimento produz curvas de carga com sequencias de valores constantes, conforme ilustrado na Figura 3.5 [7]e na Figura 3.6 [7].

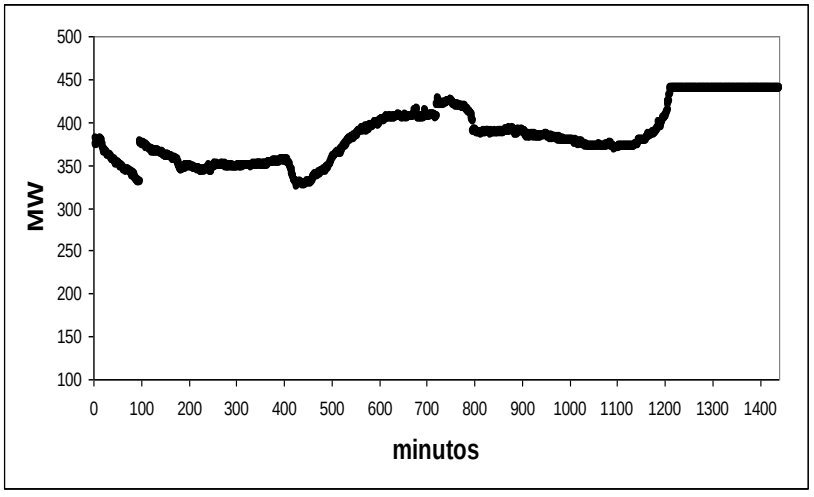


Figura 3.5 - Curva com sequências de valores constantes.

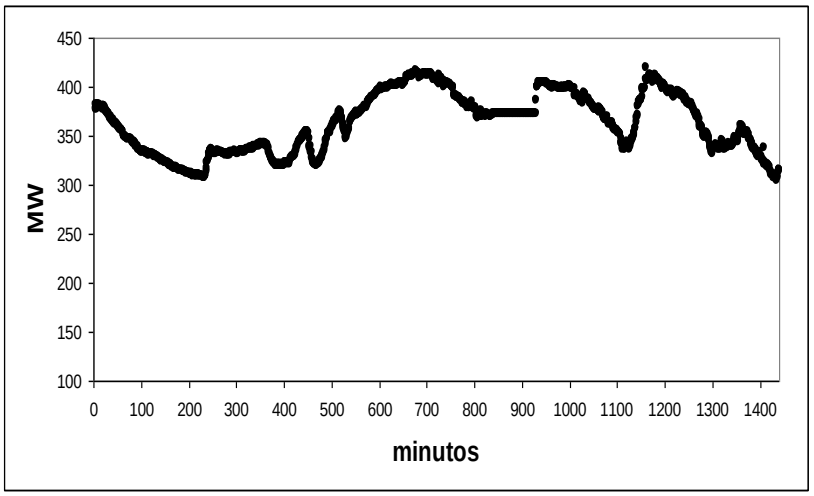


Figura 3.6 - Curva com sequências de valores constantes.

A detecção de sequências de valores constantes inicia-se pelo cálculo das primeiras diferenças em cada curva de carga. Seja $L(t)$ a demanda no instante t, então a primeira diferença é dada por $\Delta L = L(t) - L(t\text{-}1)$. A título de ilustração considere a curva de carga ilustrada na Figura 3.6 e a respectiva série de diferenças na Figura 3.7 [7]. A presença de sequências de valores constantes produz sequências de valores nulos na série das diferenças. Então longas sequências de diferenças nulas indicam que a repetição de valores da demanda instantânea deve-se a uma falha do sistema de medição, enquanto sequências curtas indicam que a repetição da demanda instantânea deve ser encarada como uma situação normal. Seguindo a mesma estratégia adotada em [7] considerou-se como lacuna de dados qualquer sequência de diferenças nulas com duração superior a 10 minutos.

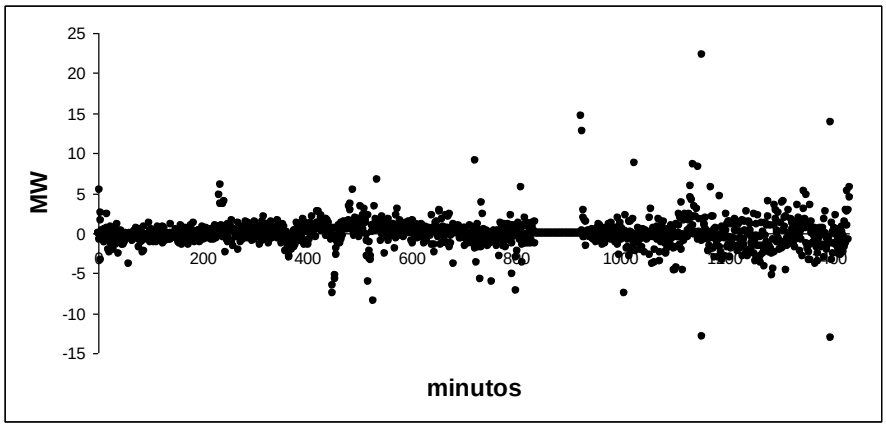


Figura 3.7 - Primeiras diferenças da curva de carga.

O procedimento descrito acima não é suficiente para identificar sequências de valores quase constantes, nas quais as primeiras diferenças assumem magnitudes próximas a zero, conforme ilustrado na Figura 3.8 [7] e suas respectivas diferenças na Figura 3.9 [7]. Um aprimoramento do procedimento acima consiste em calcular as

primeiras diferenças da curva de carga normalizada pela demanda média diária da própria curva de analisada.

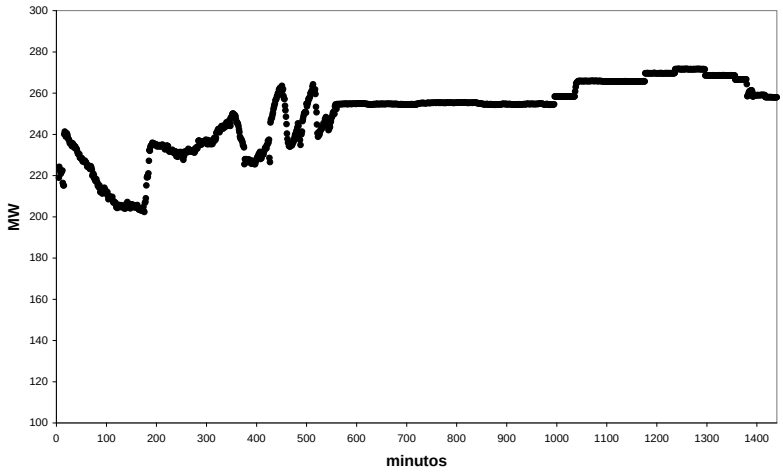


Figura 3.8 - Curva de carga com sequências de demanda quase constantes.

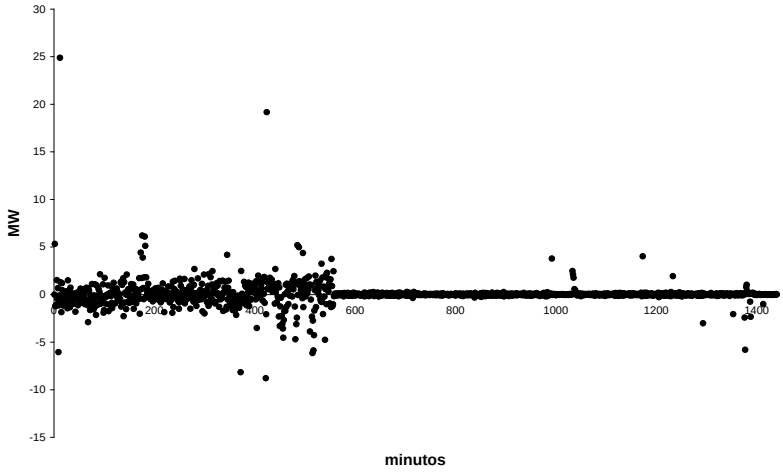


Figura 3.9 - Diferenças de primeira ordem com valores quase constantes.

Para detectar as sequências de valores quase constantes admite-se um intervalo de comprimento Δ, de tal forma que diferenças absolutas menores que Δ são contabilizadas como sendo diferenças nulas. Então se aplica a mesma lógica descrita anteriormente, ou seja, sequências com valores quase nulos por mais de 10 minutos na série das diferenças indicam que a repetição de valores da demanda instantânea deve-se a uma falha do sistema de medição. Ao longo do trabalho foi adotado um Δ=0.0001. O tamanho do intervalo é um parâmetro de entrada do tratamento de dados.

3.3 IDENTIFICAÇÃO DOS PERFIS TÍPICOS DA CURVA DE CARGA

Em função da sazonalidade da demanda por energia elétrica, o perfil da curva de carga diária apresenta um comportamento bastante regular, de maneira que é possível identificar perfis típicos da curva de carga para cada um dos sete dias da semana, feriados e os respectivos dias adjacentes, pois nestes dias a curva de carga tem um perfil diferenciado. Em função desta regularidade, os perfis típicos da curva de carga constituem uma informação valiosa para a correção dos eventuais erros na curva de carga.

A identificação dos perfis típicos é realizada por meio da aplicação de um algoritmo de análise de agrupamentos ou *cluster analysis* [34, 35] em uma amostra de medições de curvas de carga. Vale destacar que as curvas de carga com dados aberrantes e com lacunas de dados identificados pelos procedimentos descritos na Seção 3.2 não participam da definição dos perfis típicos.

A identificação dos perfis típicos é realizada por meio de algoritmos de análise de agrupamentos ou *cluster analysis*. Tais algoritmos classificam as curvas de carga da amostra em grupos ou *clusters* de forma que curvas com perfis semelhantes são classificadas no mesmo grupo, enquanto curvas com perfis diferentes são classificadas em grupos distintos. Há uma variedade de técnicas para análise de agrupamentos que podem ser empregadas na classificação das curvas de carga [34, 35]. Na metodologia proposta, adotou-se a metodologia dos mapas auto-organizáveis de Kohonen (*Self-organized Map - SOM*), pois esta abordagem gera como resultado um mapa topologicamente ordenado em que se pode facilmente visualizar os perfis típicos e as semelhanças ou diferenças entre eles.

A finalidade da análise de agrupamentos é identificar uma estrutura natural de agrupamento dos dados. No SOM, a estrutura natural de agrupamentos é identificada através de uma medida de similaridade entre as curvas, conforme descrito na Seção 2.2.

Como o objetivo é agrupar as curvas com perfis semelhantes em um mesmo *cluster*, a análise de agrupamentos é aplicada nas curvas de carga normalizadas pelas respectivas demandas médias. Os centroides em cada *cluster* são adotados como perfis típicos de curva de carga. Na Figura 3.10 e na Figura 3.11 são apresentados dois exemplos de agrupamentos identificados pelo método SOM. Na primeira, pode-se visualizar curvas com perfis de finais de semana ou feriados. Já na segunda, pode-se visualizar curvas com perfis de dias úteis. As curvas em destaque são os perfis típicos. A Figura 3.12 apresenta um exemplo de mapa topológico das curvas de carga classificadas utilizando a técnica de Kohonen, no qual as curvas de carga em feriados e finais de semana aparecem do lado esquerdo do mapa, enquanto as curvas de carga de dias úteis aparecem do lado direito do mapa. A dimensão do mapa, i.e., o número de *clusters* utilizado para o agrupamento, é um parâmetro de entrada do tratamento de dados.

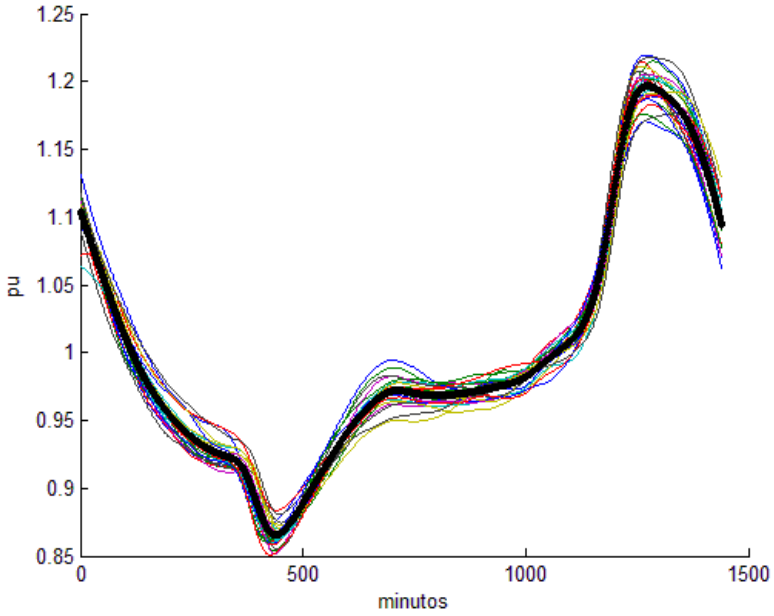


Figura 3.10 - Exemplo de agrupamento de curvas de carga.

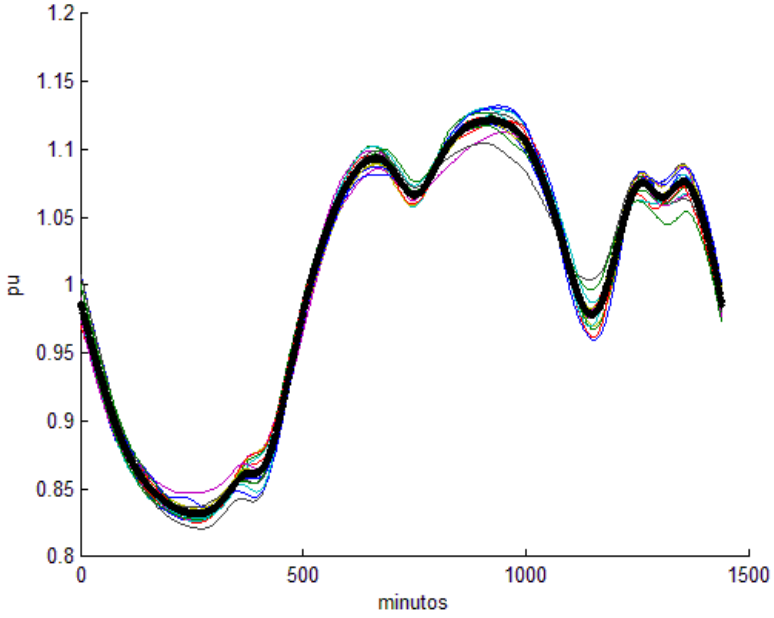


Figura 3.11 - Exemplo de agrupamento de curvas de carga.

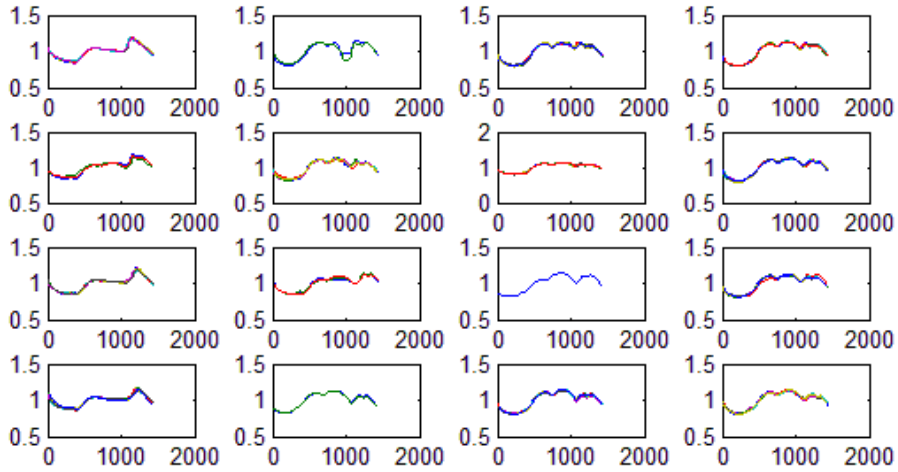


Figura 3.12 - Exemplo de mapa topológico das curvas de carga.

3.4 ASSOCIAÇÃO DE CADA CURVA A UM PERFIL TÍPICO

Nesta etapa da metodologia proposta deve-se associar cada curva de carga diária, inclusive as curvas com problemas detectados pelos procedimentos descritos na Seção 3.2, ao seu respectivo perfil típico mais provável. O perfil típico mais provável é obtido através do classificador Naive Bayes, descrito na Seção 2.3. Para isso, são utilizadas as seguintes características relativas a cada curva diária:

- A curva diária é um feriado?
- A curva diária pertence ao horário de verão?
- Dia da semana da curva diária.
- Mês da curva diária.

Inicialmente, em cada *cluster*, devem ser contabilizadas as respectivas frequências relativas ou probabilidades condicionais P(característica|*cluster*) das características supracitadas, bem como a participação do *cluster* no conjunto de curvas de carga, P(*cluster*). Essas probabilidades são guardadas em forma matricial, como pode ser visualizado na Tabela 3.2. A partir destas probabilidades calculam-se as probabilidades condicionais a posteriori, P(*cluster*|característica), em cada cluster de maneira similar ao exemplo jogador de tênis da Seção 2.3. As probabilidades mostradas na Tabela 3.2 são apenas uma parte de um exemplo de classificação e devem ser consideradas apenas para fins de ilustração da associação probabilística.

Tabela 3.2 - Visão parcial da tabela de probabilidades condicionais em cada *cluster*.

Probabilidades	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
P(*cluster*)	0.0058	0.0130	0.0101	0.0029	0.0130
P(feriado)	1.0000	0.0000	0.0714	0.2500	0.0000
P(hverão)	1.0000	1.0000	1.0000	0.2500	0.0000
P(domingo)	0.2500	1.0000	0.9286	0.0000	0.0000
P(segunda-feira)	0.0000	0.0000	0.0000	0.2500	0.0000
P(terça-feira)	0.2500	0.0000	0.0000	0.2500	0.0556
P(quarta-feira)	0.2500	0.0000	0.0714	0.0000	0.3889
P(quinta-feira)	0.1250	0.0000	0.0000	0.2500	0.2222
P(sexta-feira)	0.0000	0.0000	0.0000	0.2500	0.3333
P(sábado)	0.1250	0.0000	0.0000	0.0000	0.0000
P(janeiro)	0.5000	0.3333	0.0714	0.0000	0.0000
P(fevereiro)	0.0000	0.0556	0.0000	0.2500	0.0000
P(março)	0.0000	0.0000	0.0000	0.0000	0.2222
P(abril)	0.0000	0.0000	0.0000	0.0000	0.3333
P(maio)	0.0000	0.0000	0.0000	0.0000	0.0000
P(junho)	0.0000	0.0000	0.0000	0.2500	0.0000
P(julho)	0.0000	0.0000	0.0000	0.5000	0.0000
P(agosto)	0.0000	0.0000	0.0000	0.0000	0.0000
P(setembro)	0.0000	0.0000	0.0000	0.0000	0.1111
P(outubro)	0.0000	0.0000	0.2143	0.0000	0.3333
P(novembro)	0.0000	0.0556	0.7143	0.0000	0.0000
P(dezembro)	0.5000	0.5556	0.0000	0.0000	0.0000

Agora, como exemplo, pode-se estimar o perfil típico mais provável para um dia qualquer que possua as características visualizadas na Tabela 3.3. Para tal, deve-se calcular a probabilidade do Dia 1 pertencer a todos os *clusters* e escolher aquele de maior valor (classificação pelo critério da máxima probabilidade a posteriori). Dentre os 5 *clusters* apresentados, o Dia 1 apresenta a maior probabilidade de pertencer ao *Cluster* 3, como demonstrado nos cálculos abaixo:

Tabela 3.3 – Características do Dia 1 para estimação de perfil mais provável.

	Feriado	Horário de Verão	Dia da semana	Mês
Dia 1	Não	Sim	Domingo	Novembro

P(Dia1) = P(*cluster*) x [1 – P(feriado)] x P(hverão) x P(domingo) x P(novembro)
Cluster 1 -> P(Dia1) = 0.0058 x [1 – 1] x 1 x 0.25 x 0 = 0
Cluster 2 -> P(Dia1) = 0.013 x [1 - 0] x 1 x 1 x 0.0556 = 0.000723
Cluster 3 -> P(Dia1) = 0.0101 x [1 – 0.074] x 1 x 0.9286 x 0.7143 = 0.006221
Cluster 4 -> P(Dia1) = 0.0029 x [1 – 0.25] x 0.25 x 0 x 0 = 0
Cluster 5-> P(Dia1) = 0.013 x [1 – 0] x 0 x 0 x 0 = 0

3.5 CORREÇÃO DE CURVAS COM OBSERVAÇÕES ABERRANTES E LACUNAS DE DADOS

O procedimento de preenchimento de lacunas de dados consiste em substituir as demandas observadas nestas sequências pelas correspondentes estimativas fornecidas pelo perfil típico similar expresso em MW. Na Figura 3.13 e na Figura 3.14 são apresentados alguns exemplos de preenchimento de lacunas de dados. Mais especificamente, o preenchimento de uma lacuna é realizado de tal forma que a taxa de crescimento da demanda durante o intervalo da lacuna seja a taxa de crescimento verificada no respectivo perfil típico.

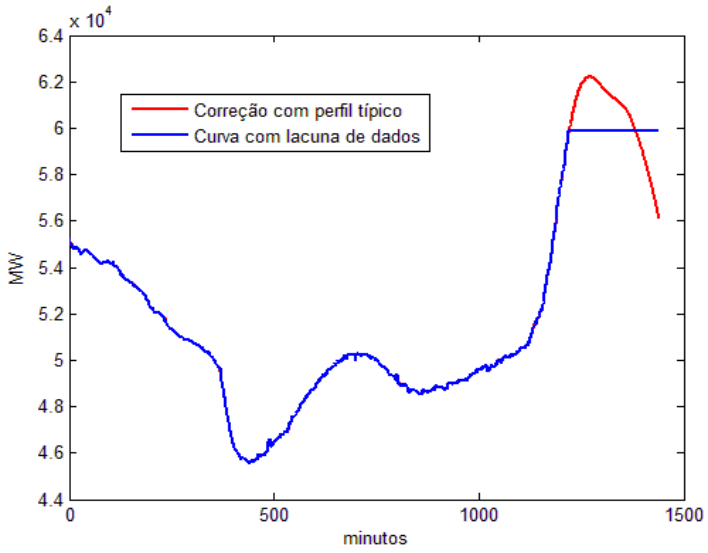


Figura 3.13 - Exemplo da correção das sequências de observações constantes.

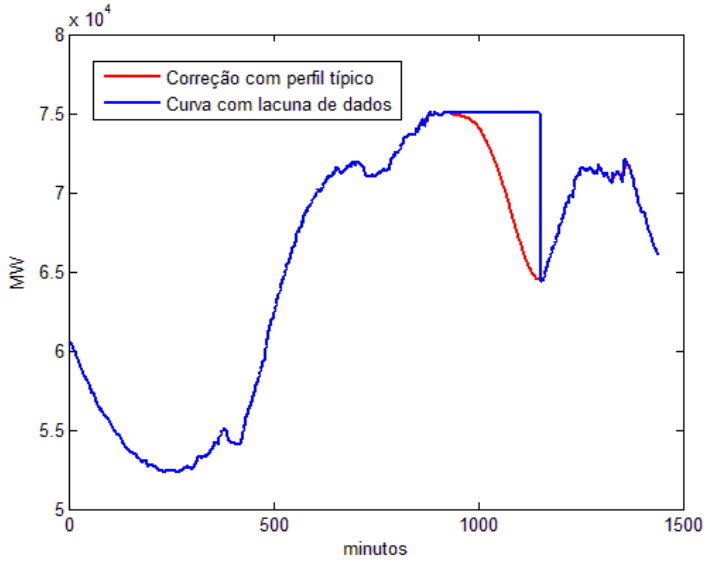


Figura 3.14 - Exemplo da correção das sequências de observações constantes.

3.6 CORREÇÃO DE CURVAS COM SEQUÊNCIAS ATÍPICAS

Cada medição de curva de carga pode ser entendida como sendo a manifestação de um perfil típico identificado pela análise de agrupamentos adicionada de ruídos aleatórios. Assim, cada medição está associada com um determinado perfil típico, ou seja, as curvas medidas aderem razoavelmente aos perfis típicos.

A comparação da curva de carga com o perfil típico permite identificar segmentos na curva medida com comportamentos diferentes do esperado, ou seja, permite detectar erros de medição. Naturalmente, nem todo desvio em relação ao padrão estabelecido pelo perfil típico significa que seja um erro, pois é natural que existam flutuações na demanda. Assim, é necessário dispor de um critério de decisão capaz de separar as flutuações normais dos erros.

Uma forma de visualizar a aderência entre as curvas medidas e os respectivos perfis típicos consiste em comparar as respectivas distribuições de frequências dos valores da demanda. Curvas de carga similares também devem apresentar distribuições de frequência semelhantes, conforme ilustrado pelos histogramas na Figura 3.16 e na Figura 3.17, que representam a distribuição de frequência das curvas da Figura 3.15.

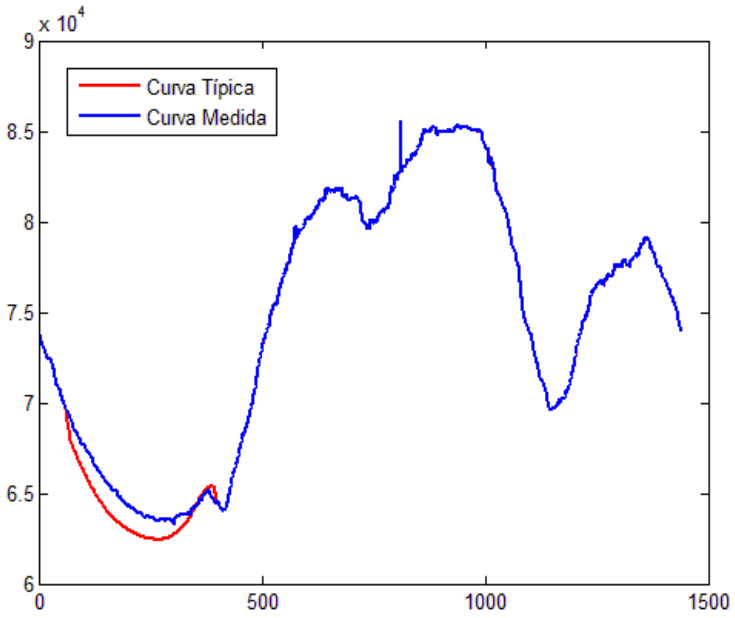


Figura 3.15 - Curva sem sequência de valores atípicos.

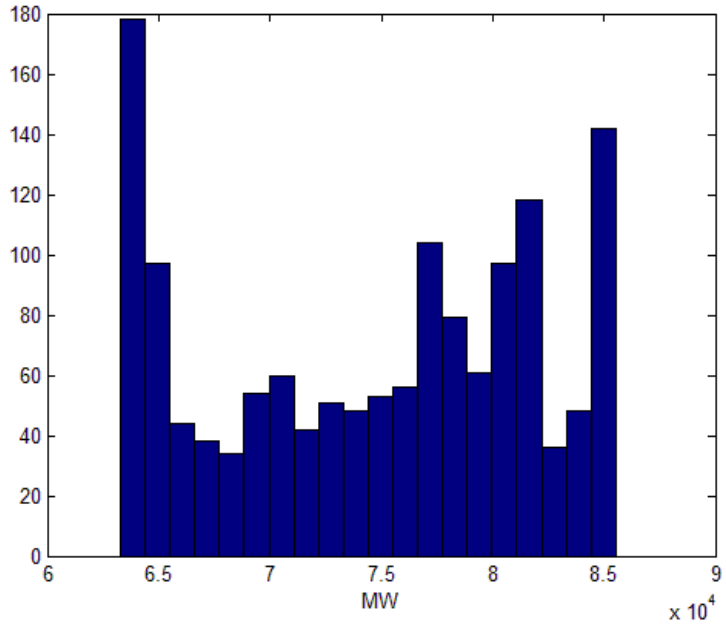


Figura 3.16 - Distribuição de frequências do perfil típico.

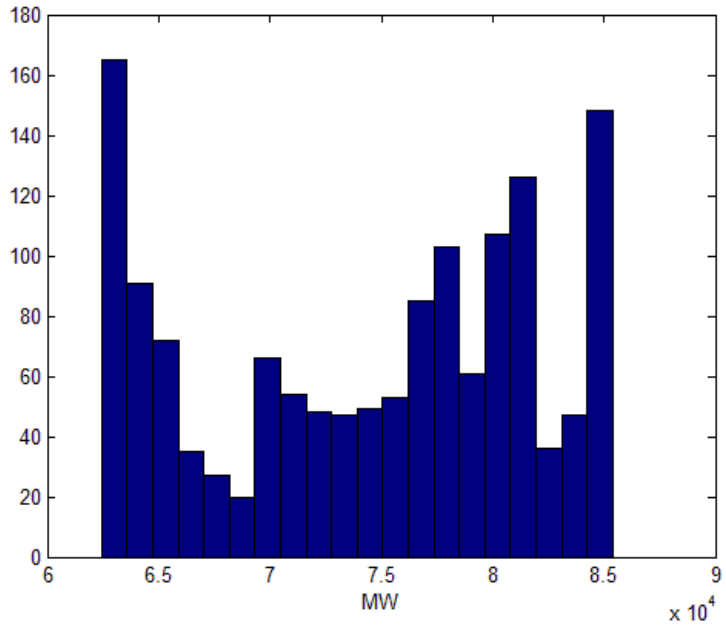


Figura 3.17 - Distribuição de frequências da curva medida.

No entanto, se a curva medida apresenta um comportamento diferenciado do perfil típico durante um determinado período do dia, as respectivas distribuições de frequência podem ser bem diferentes. A Figura 3.18 ilustra um exemplo de curva que possui uma sequência de valores atípicos e, portanto, que não foram detectados pelos

métodos aplicados anteriormente. Em seguida, pode-se visualizar na Figura 3.19 e na Figura 3.20 os histogramas das curvas exibidas na Figura 3.18.

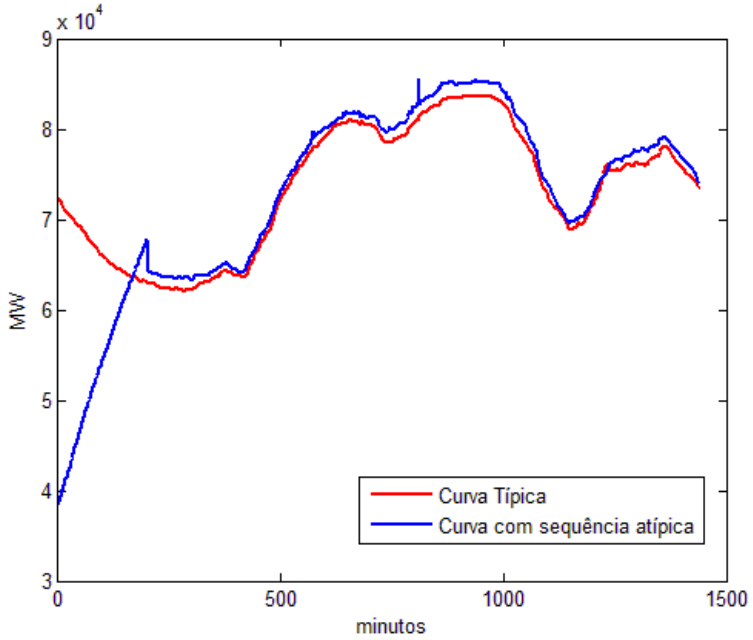


Figura 3.18 – Curva com sequência de valores atípicos.

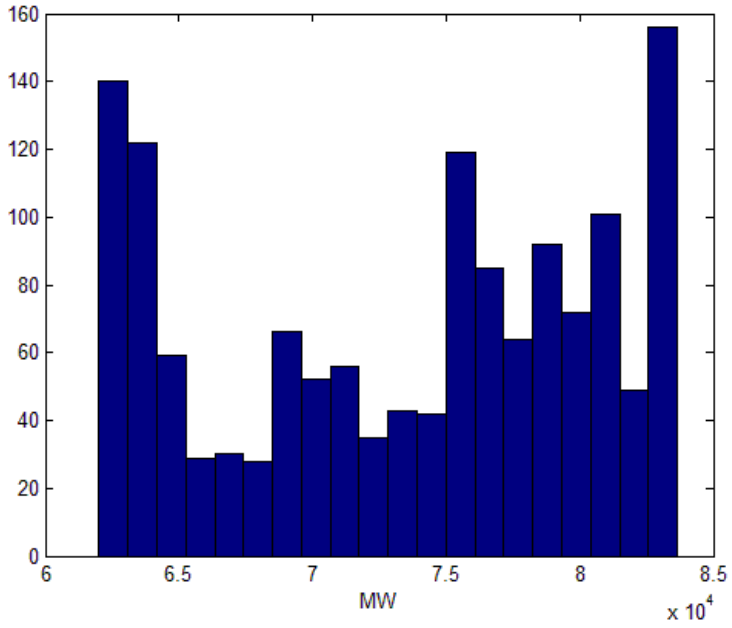


Figura 3.19 – Distribuição de frequências do perfil típico.

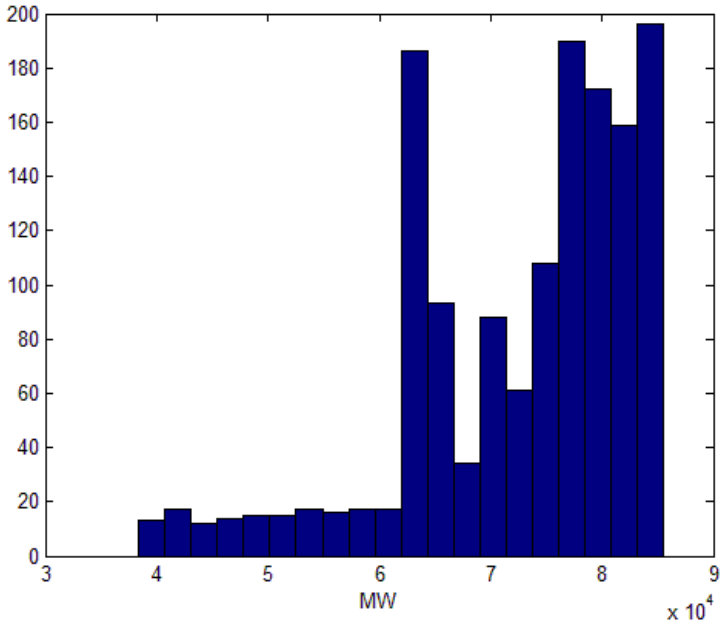


Figura 3.20 - Distribuição de frequências da curva com sequência atípica.

Como a intenção é identificar medições com grandes discrepâncias em relação aos padrões estabelecidos pelos perfis típicos, o modelo avalia a semelhança das distribuições por meio da comparação dos *boxplots* da curva medida e do perfil típico. Por exemplo, para o caso das curvas de carga apresentadas na Figura 3.15 observa-se a sobreposição entre os respectivos *boxplots* na Figura 3.21 e, portanto, a curva medida é semelhante ao perfil típico e não apresenta segmentos atípicos.

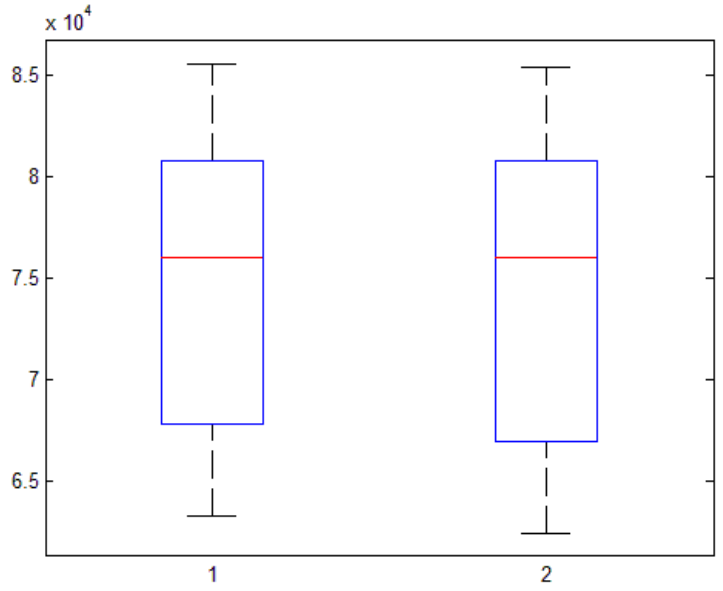


Figura 3.21 - *Boxplots* da curva medida e do perfil típico.

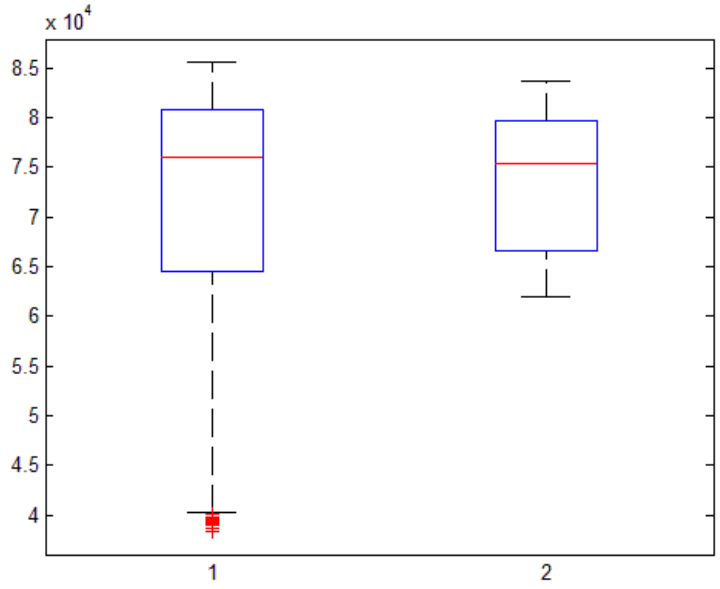


Figura 3.22 - *Boxplots* da curva com sequência atípica e do perfil típico.

Para o caso da curva ilustrada na Figura 3.18, observa-se uma sobreposição imperfeita entre os *boxplots*, caracterizada por observações fora do intervalo delimitado pelas cercas inferior e superior apenas no *boxplot* da curva medida, conforme indicado pelos símbolos (-) na Figura 3.22. Ou seja, neste caso o *boxplot* da curva medida é diferente do *boxplot* do perfil típico.

O *boxplot* é um recurso frequentemente utilizado na detecção de observações aberrantes, mas nesta aplicação a simples presença de observações fora do intervalo definido pelas cercas inferior e superior não configura observações aberrantes. Por exemplo, para as curvas da Figura 3.23, os *boxplots* apresentados na Figura 3.24 apresentam observações fora dos intervalos definidos pelas respectivas cercas inferior e superior, porém as curva medida e o perfil típico são aderentes e nenhum deles apresenta observações aberrantes.

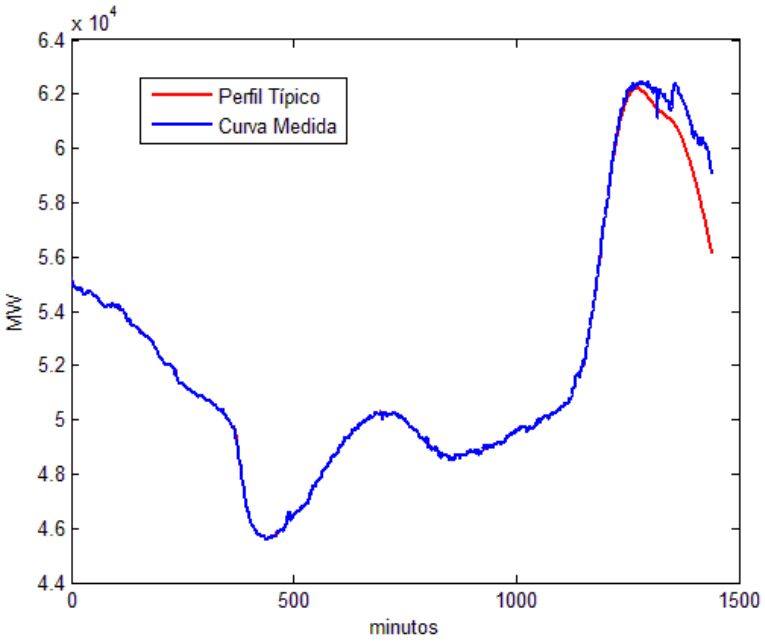


Figura 3.23 – Curva sem sequência de valores atípicos.

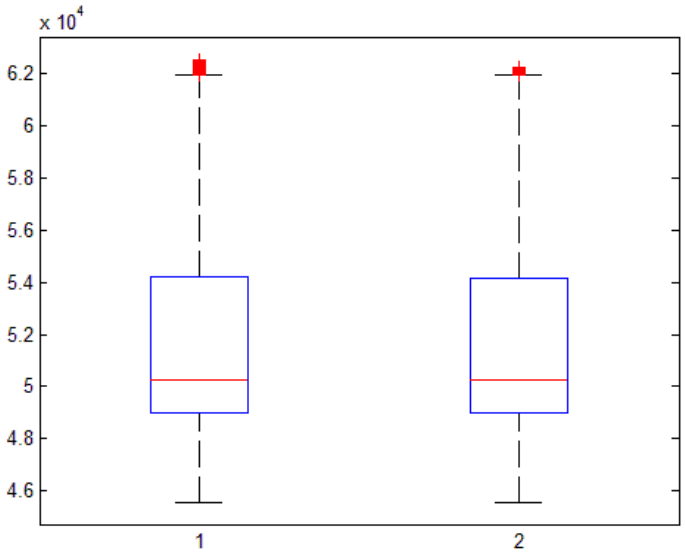


Figura 3.24 - *Boxplots* da curva sem sequência atípica e do perfil típico.

Com base nestas considerações, uma curva de carga diária é considerada atípica se as seguintes condições são verificadas:

- O *boxplot* da curva medida apresenta observações fora do intervalo definido pelas cercas inferior e superior.

- O *boxplot* do respectivo perfil típico NÃO apresenta observações fora do intervalo definido pelas cercas inferior e superior.

Para uma curva considerada atípica devem ser construídos intervalos de confiança. A construção dos intervalos de confiança inicia-se com o cálculo dos desvios entre a curva medida $L(t)$ e a curva típica $Y(t)$, conforme (3.1)

$$Desvio(t) = L(t) - Y(t) \tag{3.1}$$

Em seguida calcula-se o desvio padrão (DP) dos desvios obtidos por (3.1) e intervalo de confiança para cada instante de tempo, através do limite superior (LS) e limite inferior (LI) dados por (3.2) e (3.3), respectivamente. Na Figura 3.25 são visualizados os intervalos de confiança da curva da Figura 3.18. O valor de 3.5 vezes o desvio padrão dos desvios é um parâmetro de entrada do modelo. No entanto, ao longo do trabalho este valor foi escolhido e fixado de acordo com observado em Louzada et. al [36] na elaboração de gráficos de controle da média no contexto do controle de qualidade.

$$LS(t) = Y(t) + 3.5\ DP \tag{3.2}$$

$$LI(t) = Y(t) - 3.5\ DP \tag{3.3}$$

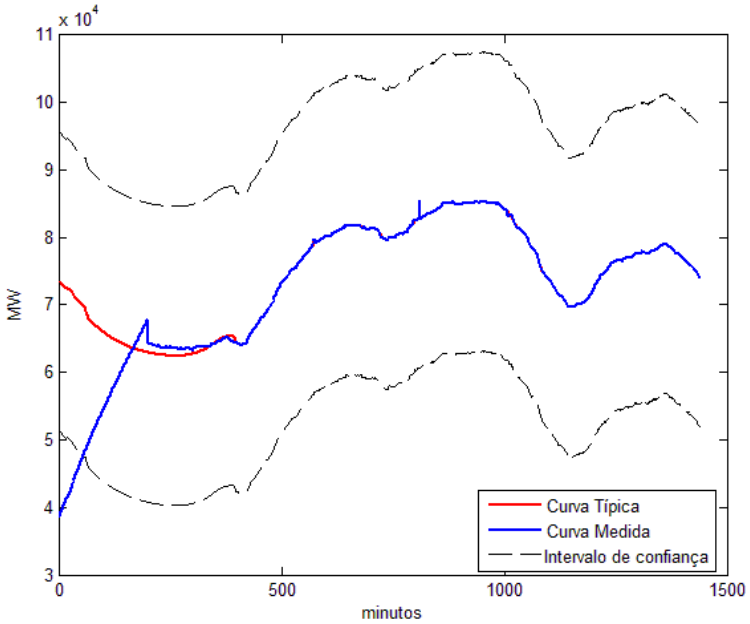


Figura 3.25 – Intervalos de confiança de curva com sequência atípica.

O procedimento de correção é aplicado somente nas curvas atípicas identificadas pelo critério da comparação das distribuições de frequência. Basicamente, o procedimento de correção consiste em verificar os instantes de tempo em que os valores da demanda da curva medida estão fora dos intervalos de confiança calculados. As observações localizadas acima ou abaixo dos limites dos intervalos de confiança são

substituídas pelo valor correspondente do perfil típico. A seguir, na Figura 3.26 é apresentada a correção da curva de carga medida da Figura 3.18.

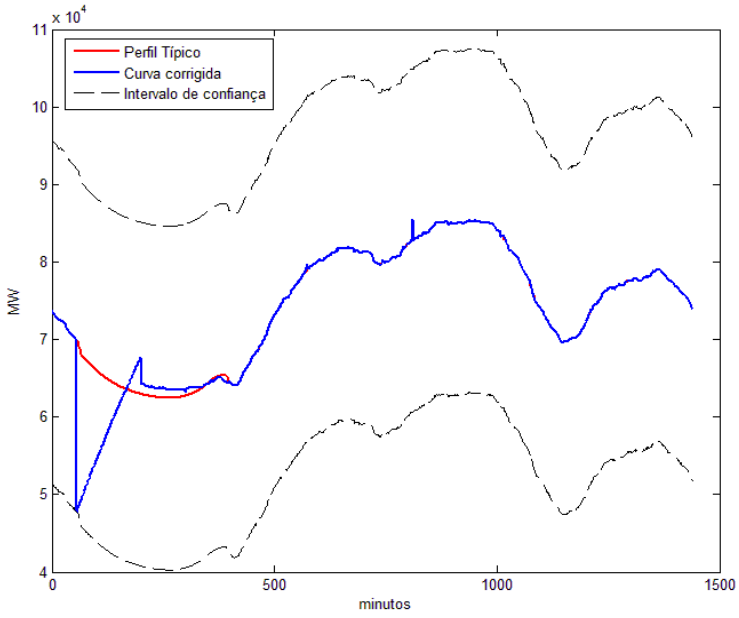


Figura 3.26 – Curva com sequência atípica corrigida.

No entanto, é possível observar que a curva ainda apresenta observações atípicas para as primeiras horas do dia. Então, a fim de tornar a correção mais efetiva, todo o procedimento de correção é aplicado novamente nas curvas corrigidas até que nenhuma curva seja considerada atípica. No caso da curva da Figura 3.18 foram necessárias 3 iterações para alcançar o critério de parada e obter uma correção satisfatória. A segunda iteração desta correção é mostrada na Figura 3.27 e Figura 3.28. A curva corrigida após a terceira iteração é mostrada na Figura 3.29 Finalmente, a Figura 3.30 compara a curva original medida e a curva obtida após o procedimento descrito nesta seção.

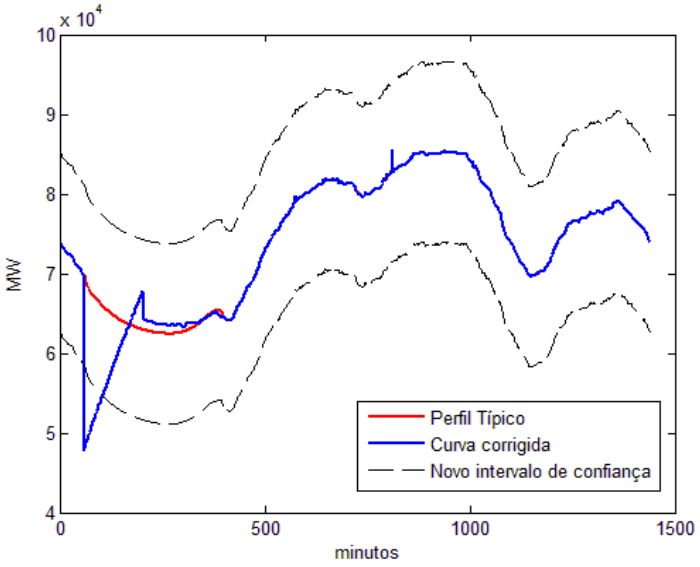


Figura 3.27 – Segunda iteração da correção de sequência atípica.

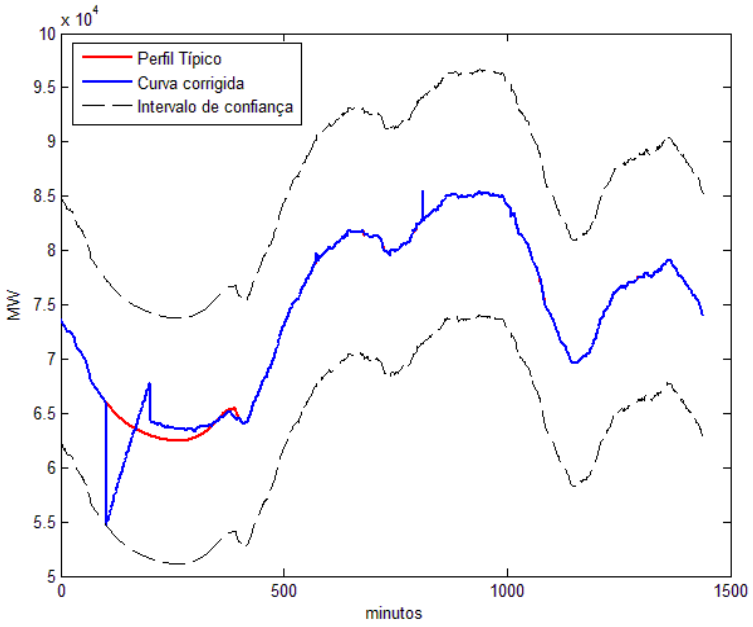


Figura 3.28 – Curva com sequência atípica corrigida após a segunda iteração.

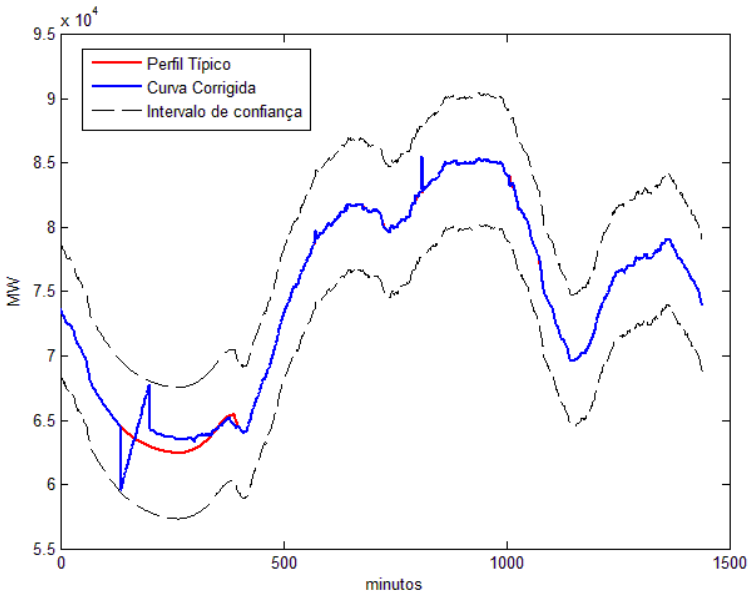


Figura 3.29 - Curva com sequência atípica corrigida após a terceira iteração.

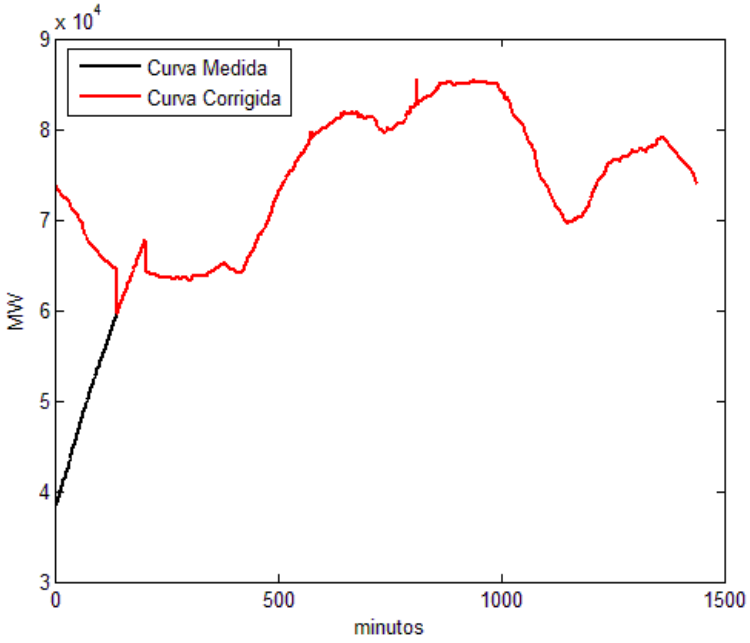


Figura 3.30 – Comparação entre curva original e curva após o processo.

3.7 CORREÇÃO DE DESCONTINUIDADES NAS CURVAS DE CARGA

O processo de correção de descontinuidades nas curvas de carga tem início com a aplicação do algoritmo de suavização LOWESS, descrito na seção 2.1, em cada uma das curvas medidas. A curva $Y(t)$, suavizada pela aplicação do LOWESS é uma referência para a comparação com a curva de carga medida $L(t)$. A partir do desvio

padrão (DP) das diferenças entre as duas curvas, são calculados os limites dos intervalos de confiança da demanda em cada instante de tempo. Os limites superior e inferior são dados por (3.4) e (3.5), respectivamente. A Figura 3.31 [7] Figura 3.31 - Curva de carga medida e limites do intervalo de confiança.ilustra um exemplo de curva diária e seus limites de intervalo de confiança.

$$LS(t) = Y(t) + 3.5\,DP \tag{3.4}$$

$$LI(t) = Y(t) - 3.5\,DP \tag{3.5}$$

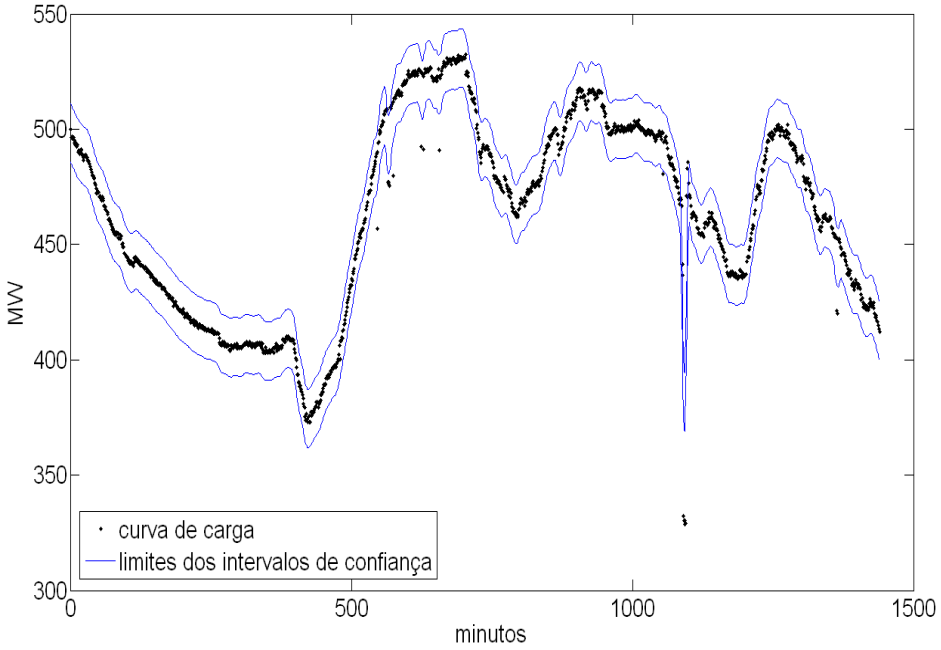


Figura 3.31 - Curva de carga medida e limites do intervalo de confiança.

Em um primeiro momento, para cada instante de tempo é feita a verificação se a demanda medida pertence ao intervalo de confiança. No caso da demanda estar fora dos limtes, o valor é considerado atípico e substituído pelo valor da demanda imediatamente anterior, como visto na Figura 3.32 [7].

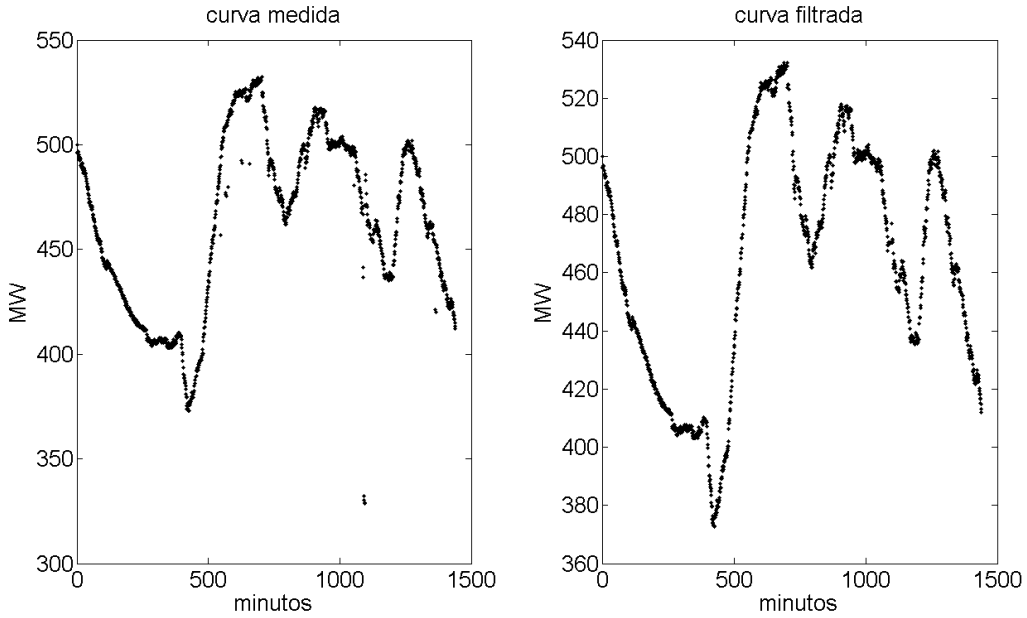


Figura 3.32 - Curvas de carga antes e após a filtragem das observações aberrantes.

Em seguida, são tomadas as diferenças de primeira ordem da curva de carga $L(t)$ para os instantes de tempo $t = 1$ até $t = 1440$, segundo (3.6).

$$\Delta L(t) = L(t) - L(t-1) \tag{3.6}$$

As Figura 3.33 [7] e Figura 3.34 [7] mostram duas curvas de carga distintas e os seus respectivos diagramas de dispersão da série de diferenças ΔL. Os diagramas representam a relação entre $\Delta L(t)$ e $\Delta L(t-1)$. A primeira, da Figura 3.33, é uma curva continua; isto é, sem descontinuidades. Já a segunda curva, da Figura 3.34, apresenta descontinuidades ao longo de sua trajetória. Considerando a forma elíptica observada nos diagramas, é possível admitir que o vetor ΔL possui uma distribuição normal bivariada $\Delta L \sim N_2(\mu, \Sigma)$ [8]. Com isso, o quadrado da distância de Mahalanobis tem distribuição qui-quadrado com 2 graus de liberdade, ou seja, $(\Delta L - \mu)^T \Sigma^{-1} (\Delta L - \mu) \sim x_2^2$. [37].

Fazendo $(\Delta L - \mu)^T \Sigma^{-1} (\Delta L - \mu) \sim x_2^2(\alpha)$, onde $x_2^2(\alpha)$ é o quantil da distribuição qui-quadrado que deixa uma probabilidade de α à sua direita, obtém-se a equação da elipse que delimita o contorno de probabilidade 1- α. Portanto, a probabilidade de um ponto qualquer do diagrama de dispersão pertencer ao contorno de probabilidade 1- α é dada por (3.7).

$$\begin{aligned} P(\Delta L \in contorno) &= P((\Delta L - \mu)^T \Sigma^{-1} (\Delta L - \mu) \leq x_2^2(\alpha)) \\ &= 1 - \alpha \end{aligned} \tag{3.7}$$

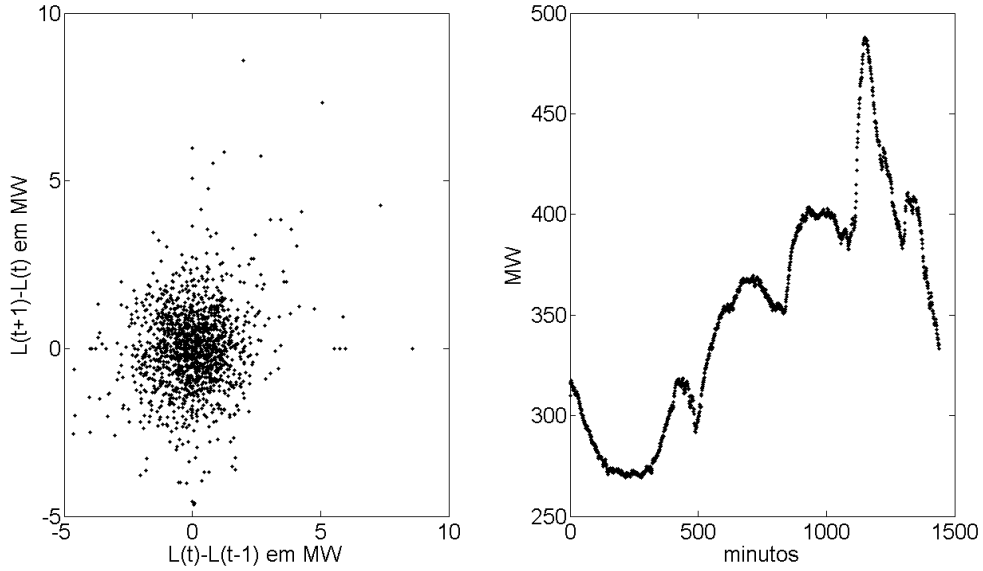


Figura 3.33 – Curva e diagrama de dispersão – sem descontinuidade.

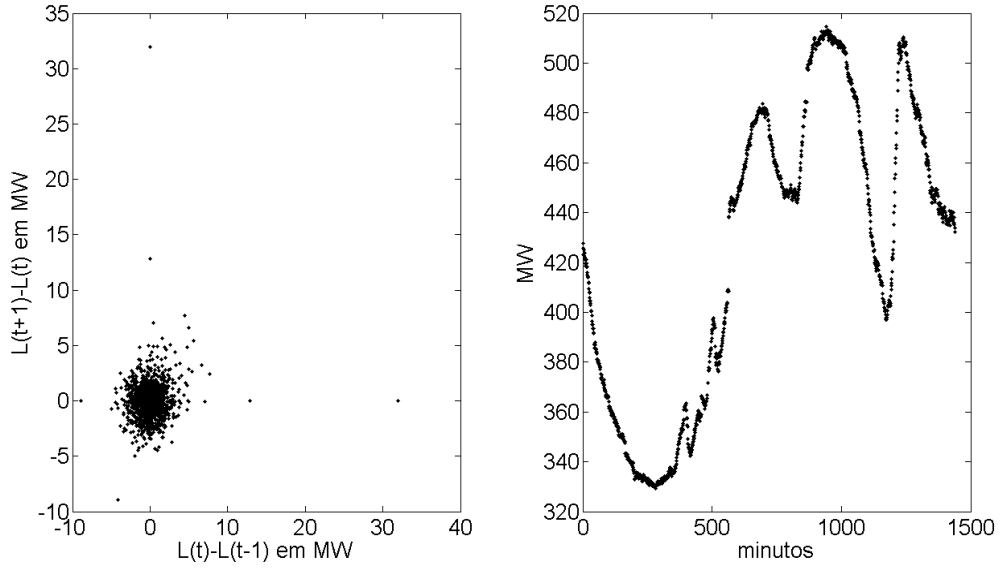


Figura 3.34 - Curva e diagrama de dispersão – com descontinuidade.

Assim, para uma dada probabilidade α, um determinado vetor ΔL é considerado normal se $(\Delta L - \mu)^T \Sigma^{-1} (\Delta L - \mu) \leq x_2^2(\alpha)$, caso contrário, o vetor é classificado como anormal. Na Figura 3.35 [7] apresenta-se um exemplo de diagrama de dispersão com quadrados indicando os pontos que não pertencem ao contorno de probabilidade de 90% ($\alpha = 10\%$).

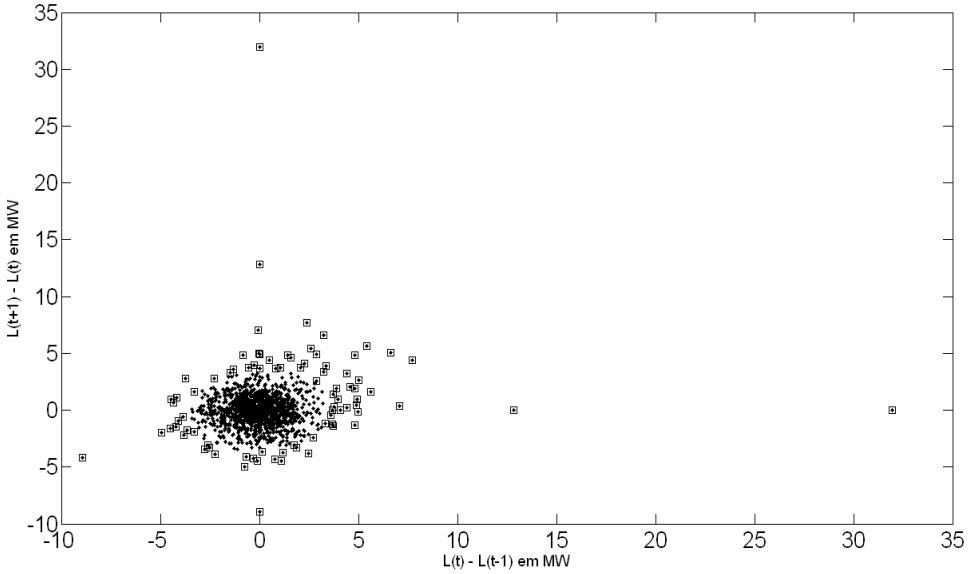


Figura 3.35 - Pontos fora do contorno de probabilidade de 90%.

No entanto, para uma curva de carga ser considerada descontínua é necessário que o respectivo diagrama de dispersão apresente alguma anomalia. Neste trabalho, as anomalias são os pontos fora do contorno de probabilidade que estejam isolados ou acompanhados de no máximo um determinado número δ de outros pontos anormais dentro de uma vizinhança circular de raio φ [26]. Assim, uma forma de detectar as anomalias consiste em aplicar o algoritmo DBSCAN [26]. Por exemplo, considerando um raio φ=10 MW e uma vizinhança com até δ=5 pontos, as anomalias encontradas no diagrama de dispersão da Figura 3.35 correspondem aos pontos indicados na Figura 3.36 [7]. Os parâmetros φ e δ foram escolhidos após uma sequência de testes e são parâmetros de entrada do tratamento de dados.

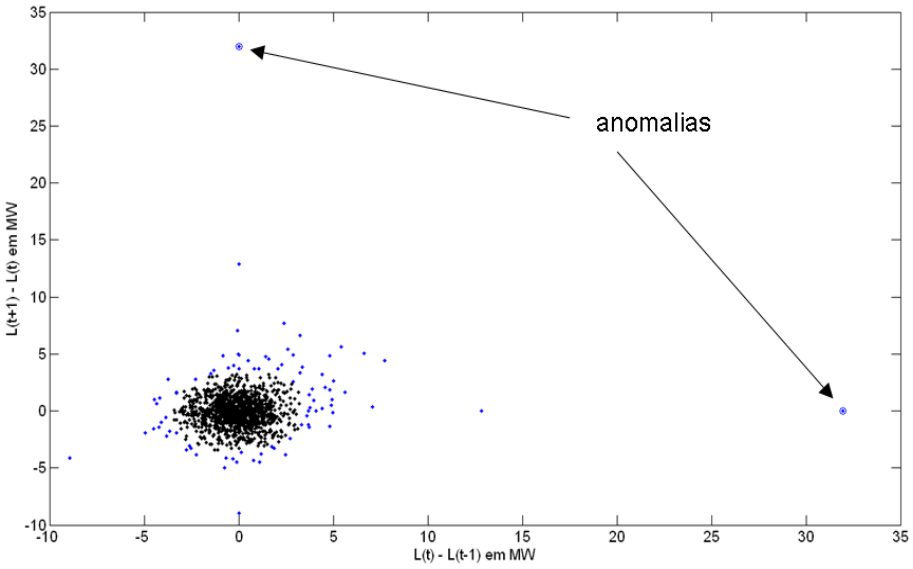


Figura 3.36 - Anomalias no diagrama de dispersão.

Por meio do procedimento descrito é possível identificar os instantes de tempo em que ocorrem as descontinuidades e, desta forma, separar uma curva de carga em dois ou mais segmentos contínuos, conforme ilustrado na Figura 3.37 e na Figura 3.38.

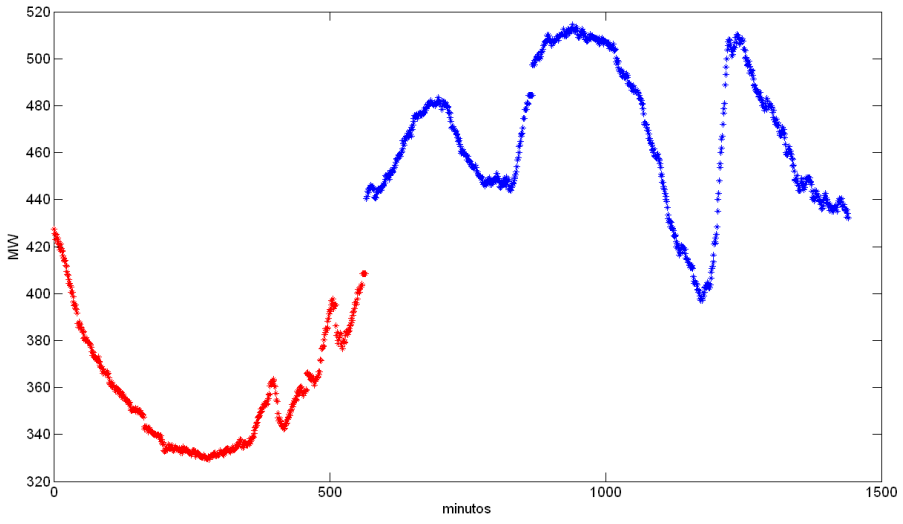


Figura 3.37 – Segmentação de curva de carga em dois segmentos contínuos.

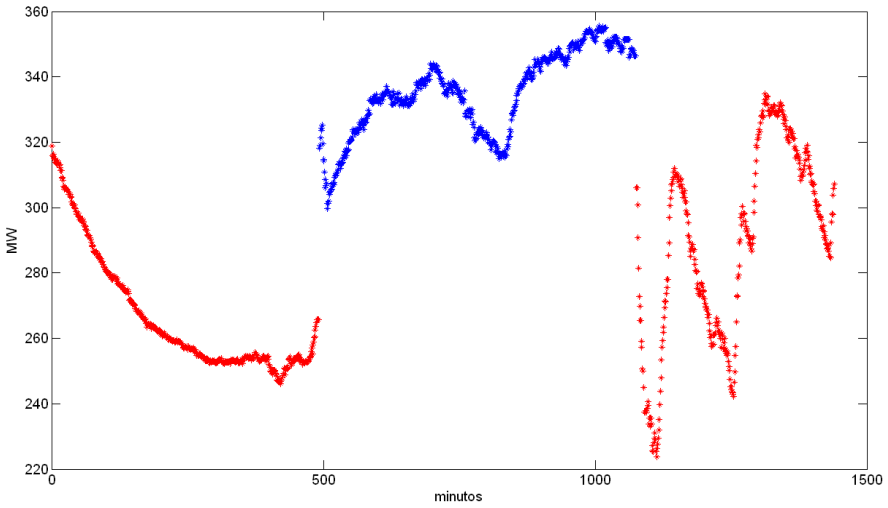


Figura 3.38 - Segmentação de curva de carga em três segmentos contínuos.

Para realizar a correção das descontinuidades identificadas nas curvas de carga, admitem-se como referências os perfis típicos identificados na seção 3.3. Primeiro, cada curva de carga $L(t)$ e respectivo perfil típico associado $Tipo(t)$ são normalizados, conforme (3.8) e (3.9).

$$L_N(t) = \frac{L(t) - mínimo(L(t))}{maximo\big(L(t)\big) - \ mínimo(L(t)} \tag{3.8}$$

$$Tipo_N(t) = \frac{Tipo(t) - mínimo(Tipo(t))}{maximo\big(Tipo(t)\big) - \ mínimo(Tipo(t)} \tag{3.9}$$

A Figura 3.39 ilustra um exemplo de curva com descontinuidades e seu perfil típico associado, ambos normalizados. É possível notar a boa aderência entre os valores normalizados da curva de carga e o respectivo perfil típico. A correção das descontinuidades, consiste em projetar os segmentos da curva de carga normalizada sobre a curva do perfil típico associado (também normalizado), de forma que a projeção minimize a soma dos quadrados dos desvios entre as duas curvas, segundo (3.10), sendo cada segmento da curva de carga projetado independentemente dos demais segmentos. A Figura 3.40 [7] mostra o resultado da projeção dos segmentos de uma curva de carga sobre o seu perfil típico associado.

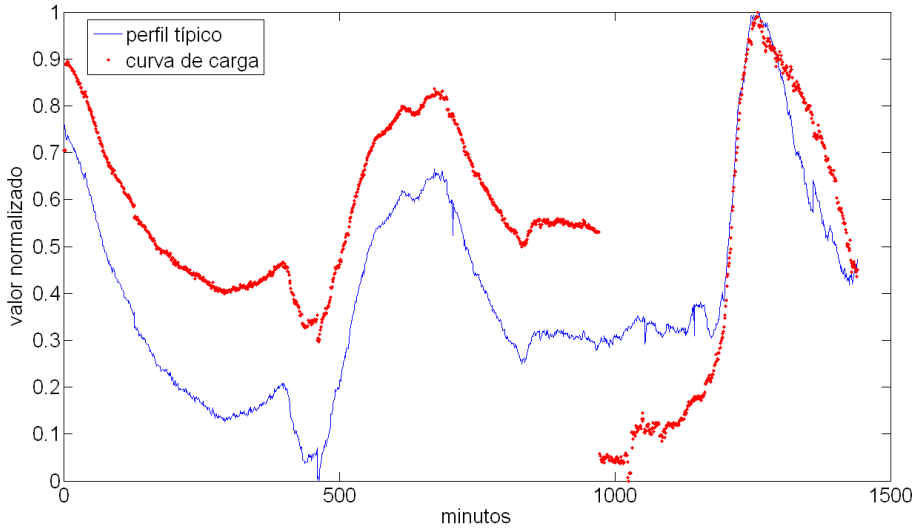


Figura 3.39 – Curva de carga com descontinuidade e seu perfil típico associado.

$$L_N^*(i,t) = L_N(i,t) + \frac{1}{N_i}\sum_{j=1}^{N_i}\left(Tipo_N(i,t) - L_N(i,t)\right) \tag{3.10}$$

Sendo:

- N_i é o total de observações contidos no i-ésimo segmento.
- $L_N(i,t)$ é o vetor com as demandas normalizadas da curva de carga durante o intervalo de tempo abrangido pelo i-ésimo segmento.
- $Tipo_N(i,t)$ é o vetor com as demandas normalizadas do perfil típico durante o intervalo de tempo abrangido pelo i-ésimo segmento.
- $L_N^*(i,t)$ é o vetor com valores projetados da curva de carga no perfil típico durante o intervalo de tempo abrangido pelo i-ésimo segmento.

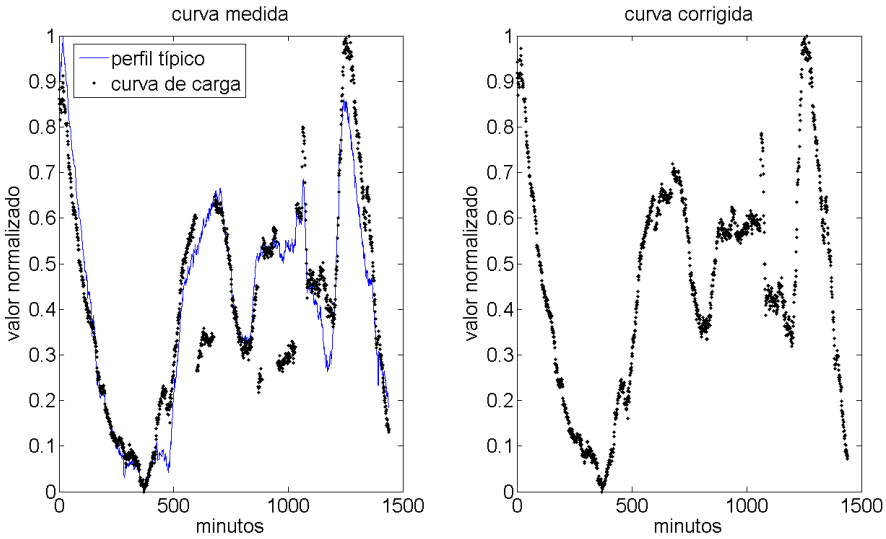


Figura 3.40 - Projeção dos segmentos da curva de carga no perfil típico.

Na sequência, para obter a curva de carga filtrada em MW ($L^*(t)$),basta desfazer a normalização da curva de carga, segundo (3.11). Por fim, as curvas filtradas são suavizadas através do processo LOWESS, descrito na seção 2.1. A Figura 3.41 [7], Figura 3.42 [7] e a Figura 3.43 [7] apresentam três exemplos de curvas medidas e as respectivas curvas corrigidas após o processo descrito nesta seção.

$$L^*(t) = mínimo\big(L(t)\big) + L_N^*(t) * [\,maximo\big(L(t)\big) - mínimo(L(t))] \qquad (3.11)$$

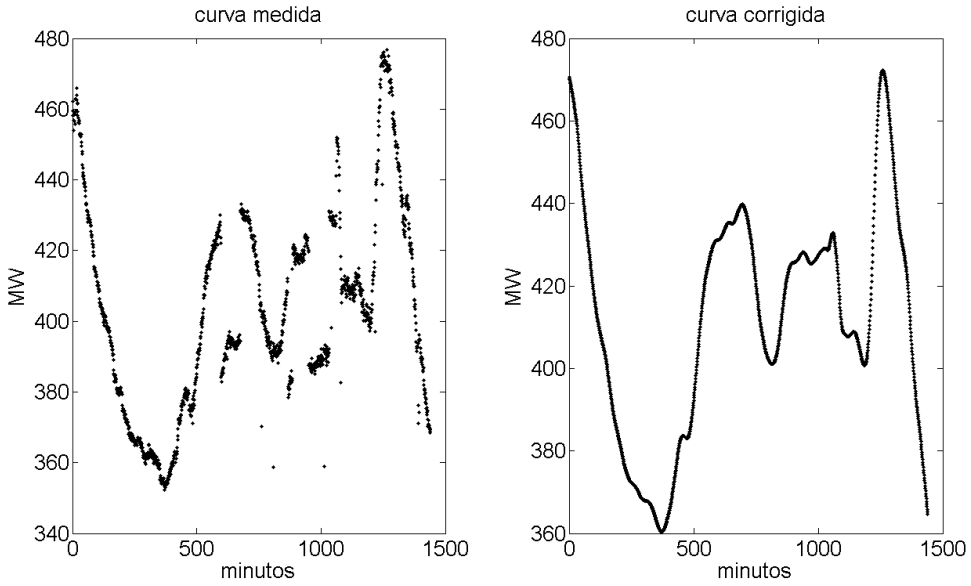


Figura 3.41 – Curva medida e curva corrigida.

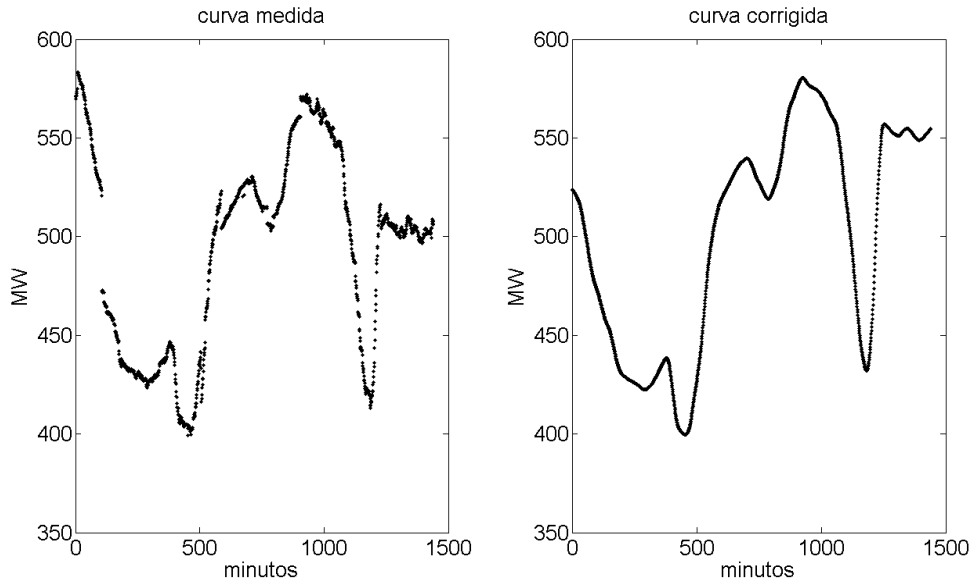


Figura 3.42 - Curva medida e curva corrigida.

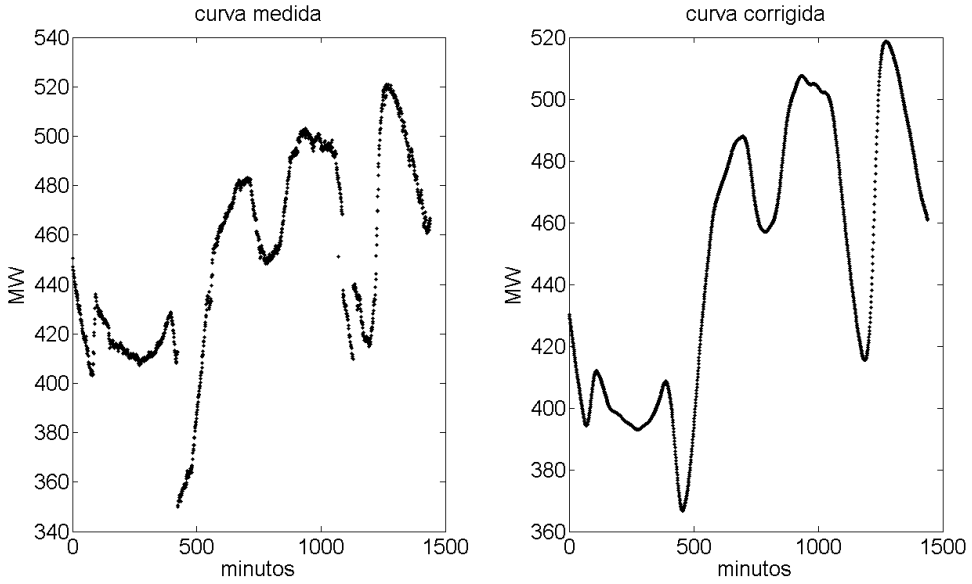


Figura 3.43 - Curva medida e curva corrigida.

4. RESULTADOS

Neste capítulo apresentam-se os resultados oriundos de experimentos computacionais que foram formulados com a finalidade de avaliar e demonstrar a habilidade da metodologia proposta no tratamento dos perfis de carga. Inicialmente, faz-se uma análise exploratória de um conjunto de dados de carga e na sequência apresentam-se os experimentos computacionais nos quais se inserem falhas aleatórias nos dados de carga por meio de simulações e verificam-se as correções sugeridas pela metodologia proposta.

4.1. ANÁLISE EXPLORATÓRIA DE DADOS

Os dados utilizados durante a elaboração e testes do modelo são provenientes do Sistema Interligado Nacional (SIN). Os dados são aquisitados com discretização de minuto a minuto e, portanto, cada curva diária possui um total de 1440 pontos. No total, foram utilizadas curvas de carga ao longo de um período de 5 anos; dentre esses, 4 anos foram utilizados para o treinamento do Mapa de Kohonen e do classificador Naive Bayes e 1 ano para os testes que servem para avaliar a performance da metodologia proposta. No total foram utilizadas 1461 curvas no período de treinamento e 365 curvas no período de validação.

Inicialmente, as curvas foram separadas de acordo com o dia da semana para uma primeira visualização. A Figura 4.4 ilustra as curvas diárias e suas médias para cada dia da semana. É possível observar uma grande semelhança entre as curvas relativas aos dias úteis (segunda-feira, terça-feira, quarta-feira, quinta-feira e sexta-feira), como era de se esperar. Analogamente, também há uma correspondência entre as curvas relativas aos finais de semana (sábado e domingo).

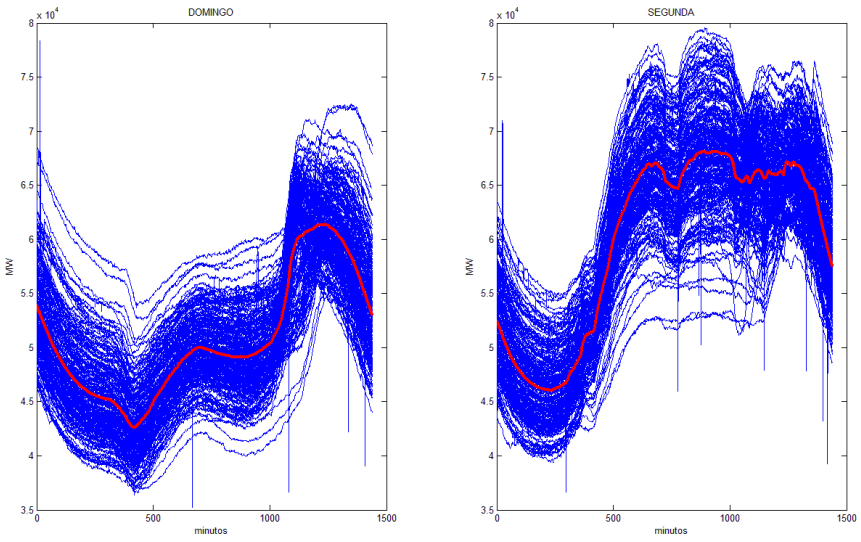


Figura 4.1 - Curvas de carga de domingo e segunda-feira.

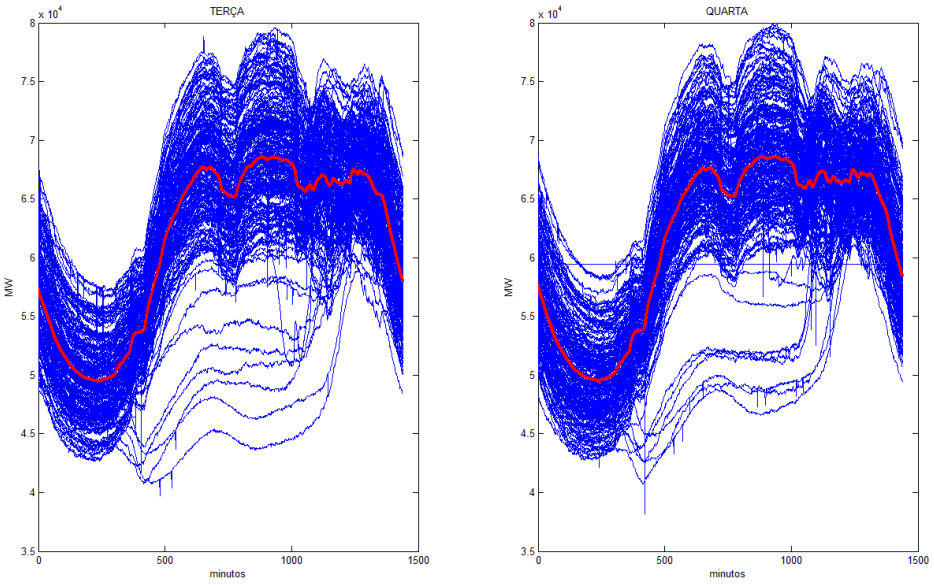


Figura 4.2 - Curvas de carga de terça-feira e quarta-feira.

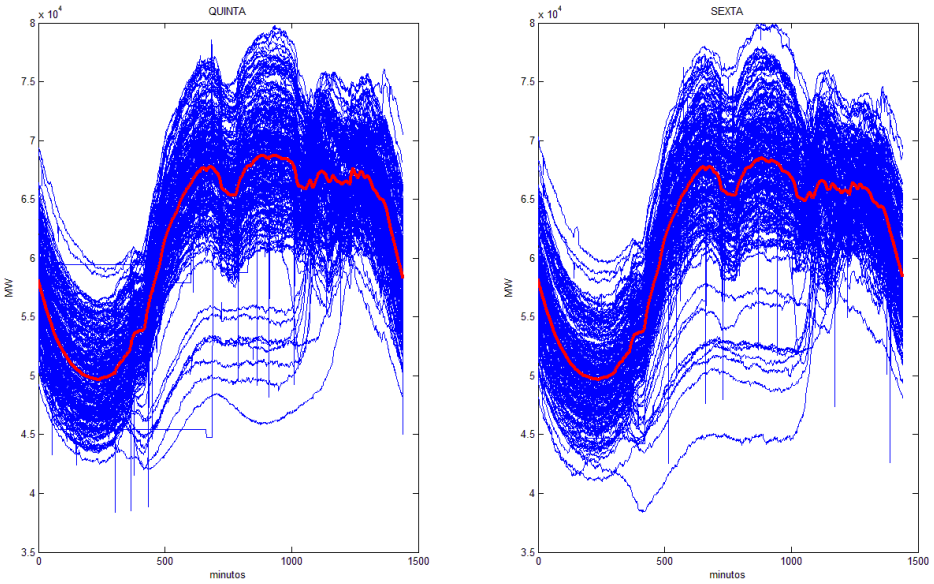


Figura 4.3- Curvas de carga de quinta-feira e sexta-feira.

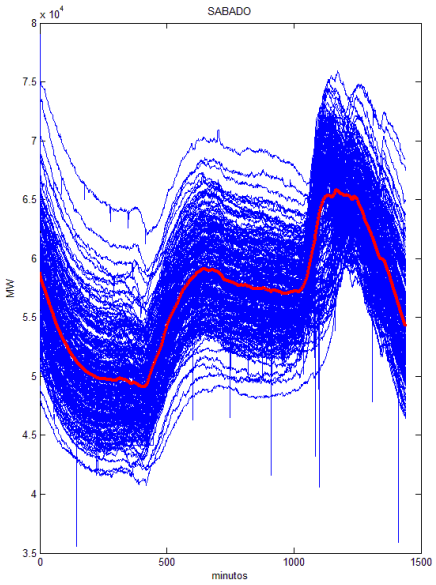


Figura 4.4 – Curvas de carga de sábado.

Uma observação mais cuidadosa revela que para um mesmo dia da semana existem curvas que possuem formas bem diferentes. Essas diferenças são explicadas, em parte, por informações de alta relevância quando se trata de demanda de energia elétrica – o horário de verão e feriados. A Figura 4.5 traz uma comparação entre a curva média do período correspondente ao horário de verão e a curva média do período que não corresponde ao horário de verão, i.e., o horário oficial brasileiro. É possível observar uma diferença bastante significativa no formato das curvas, inclusive na diferença da localização do pico de demanda de energia elétrica.

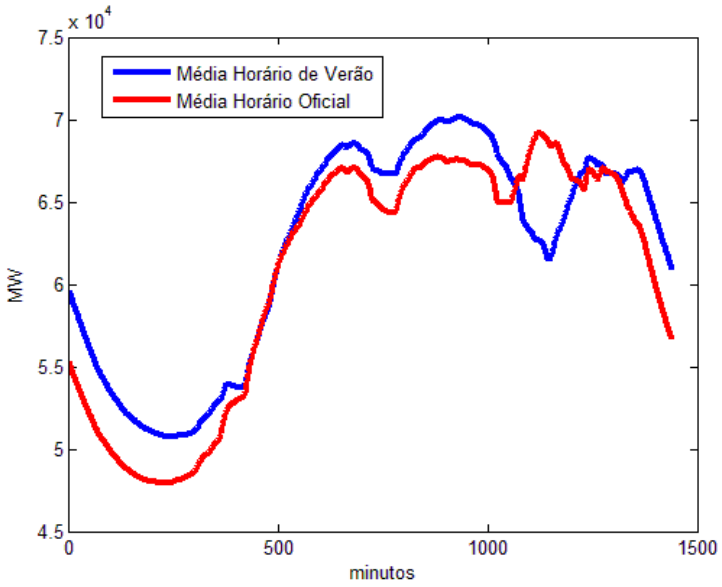


Figura 4.5 – Curvas médias do horário de verão e do horário oficial.

A Figura 4.6 ilustra a comparação entre a curva média dos dias úteis e a curva média dos feriados para o período de dados analisado. É possível notar uma grande semelhança entre a carga média dos feriados e um dia de final de semana.

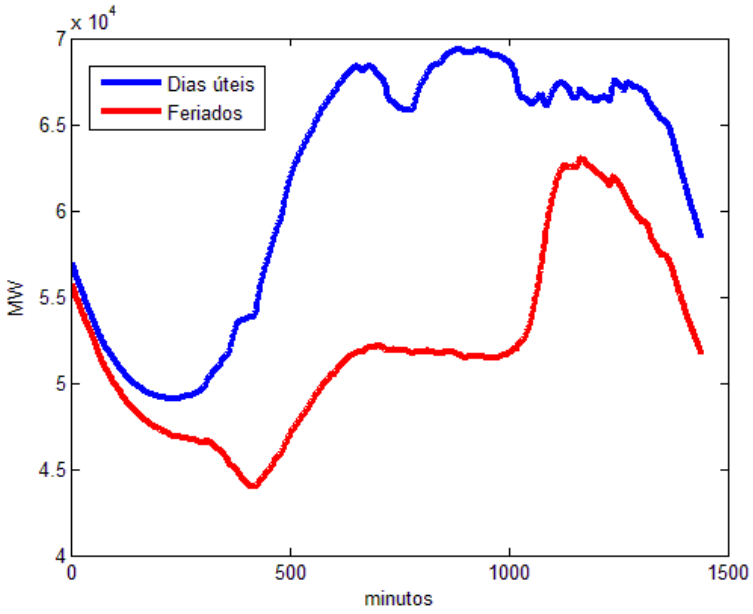


Figura 4.6 – Curvas médias dos dias úteis e dos feriados.

4.2. EXPERIMENTOS COMPUTACIONAIS

Os experimentos computacionais foram realizados em duas etapas. A primeira etapa, denominada por etapa de treinamento, consiste em gerar uma base de perfis típicos a partir dos dados do conjunto de treinamento, bem como um mapeamento das probabilidades condicionais em cada perfil típico pelo Naive Bayes. A base de perfis típicos e as respectivas probabilidades condicionais são os insumos para a segunda parte dos experimentos computacionais, cujo objetivo consiste em avaliar a qualidade das respostas fornecidas pela metodologia proposta para o tratamento de dados.

Na segunda etapa, denominada por etapa de avaliação, as curvas do conjunto de teste são artificialmente corrompidas com a adição dos resultados da simulação dos dos erros mais frequentemente verificados nos registros de carga(*outliers*, lacunas de dados e descontinuidades). As curvas artificialmente corrompidas são analisadas pela metodologia proposta e então pode-se avaliar a capacidade da metodologia de tratamento de dados em recuperar os perfis de carga originais livres dos erros simulados. O diagrama da Figura 4.7 ilustra a integração entre etapas do experimento computacional. A Tabela 4.1 apresenta os parâmetros utilizados durante a aplicação da metodologia proposta.

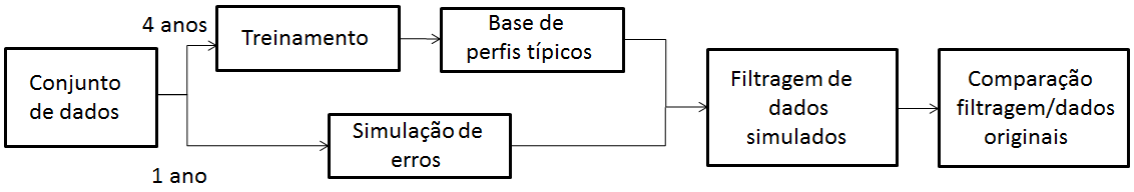


Figura 4.7 – Diagrama com a sequencia para os testes.

Tabela 4.1 – Parâmetros utilizados para aplicação do modelo.

Numero de agrupamentos para o SOM	100
Tolerância no calculo de diferenças de 1ª ordem na identificação de segmentos constantes	0.0001
Tamanho da janela para a filtragem pelo Local Linear Regression (Lowess)	10 minutos
Número de desvios-padrão para construção de intervalos de confiança	3.5
Raio φ para identificação de descontinuidades (DBSCAN)	10 MW
Número mínimo de pontos δ para identificação de descontinuidades (DBSCAN)	5

4.2.1 – ETAPA DE TREINAMENTO

O processo de treinamento do modelo tem início com a apresentação das curvas relativas aos 4 anos de dados mencionados anteriormente ao procedimento de filtragem proposto. Então, os dados são processados por cada etapa de filtragem descrita no Capítulo 3. Os resultados de cada etapa de filtragem são apresentados nas subseções seguintes.

4.2.1.1 – SEPARAÇÃO DE CURVAS COM LACUNAS OU DADOS ABERRANTES

A aplicação do tratamento de dados inicia-se com a identificação das curvas com dados aberrantes (*outliers*), conforme a metodologia apresentada na seção 3.2. Do total de 1461 curvas analisadas, 122(aproximadamente 8%) curvas apresentaram valores aberrantes. Estas curvas são ilustradas na Figura 4.8. No entanto, não é possível identificar todas as curvas devido à diferença de magnitude entre os *outliers* encontrados.

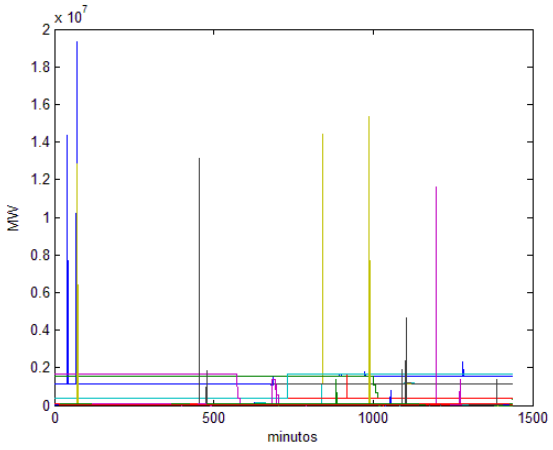


Figura 4.8 – Curvas com outliers.

Posteriormente, faz-se a identificação das curvas que apresentam lacunas de dados, conforme a metodologia apresentada na seção 3.2. Do total de 1461 curvas analisadas, 75(aproximadamente 5%) curvas apresentaram o problema. Estas curvas são ilustradas na Figura 4.9.

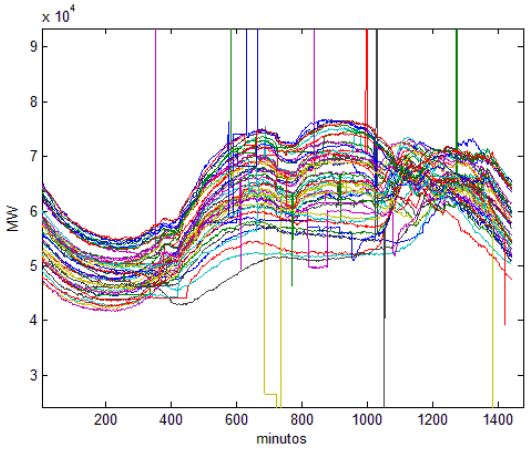


Figura 4.9 - Curvas com lacunas de dados.

É importante ressaltar que algumas curvas apresentaram *outliers* e lacunas de dados. Assim, o número total de curvas filtradas nesta etapa foi de 174(aproximadamente 12%). Estas curvas não participam da análise de agrupamentos, responsável pela geração de uma base de perfis típicos para filtragem, e são separadas para correção posterior.

4.2.1.2 – IDENTIFICAÇÃO DE PERFIS TÍPICOS

Nesta etapa foram utilizadas todas as 1287 curvas que não foram filtradas na etapa anterior; ou seja, são as curvas que não apresentam *outliers* ou lacunas de dados. Para identificação dos perfis típicos de curva de carga foi utilizada uma Rede Neural

Auto-Organizável de Kohonen, descrita na Seção 2.2. Inicialmente, foi arbitrado o número máximo de 100 *clusters*. A clusterização foi realizada com o auxílio da ferramenta *Neural Network Tools*, do Matlab, que é mostrada na Figura 4.10. A disposição dos neurônios e o número de curvas em cada neurônio após a classificação podem ser visualizados na Figura 4.11.

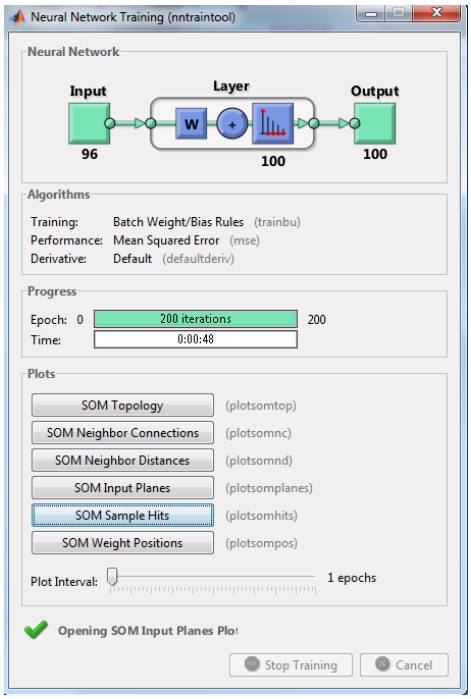


Figura 4.10 – Ferramenta NNTool- Matlab.

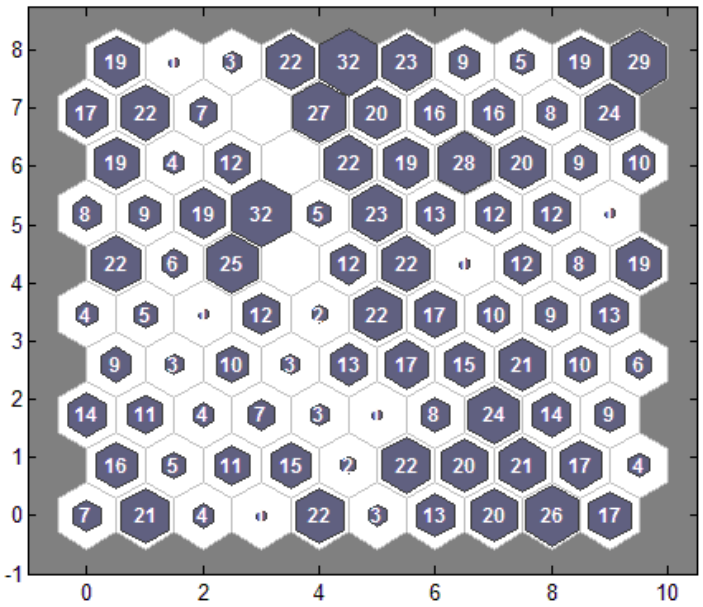


Figura 4.11 – Topologia da rede e número de curvas por cluster.

Uma característica interessante das Redes Auto-Organizáveis de Kohonen é a capacidade de organizar os objetos (neste caso, as curvas de carga) em um mapa topologicamente ordenado [38], i.e., objetos semelhantes são posicionados próximos uns dos outros no mapa. Esta propriedade pode ser percebida na Figura 4.12, que mostra o centroide do *cluster* 1 e de seus vizinhos com até 2 neurônios de distância (*clusters* 2,3,11,12,13,21 e 22). Vale possível notar que todas as curvas médias apresentam grande semelhança e parecem curvas típicas de feriados ou finais de semana. De forma análoga, para o *cluster* 78 e sua região de vizinhança com até 2 neurônios de distância, na Figura 4.13 também é possível observar uma grande semelhança entre as curvas médias de cada *cluster*. Os 100 *clusters* podem ser visualizados na Figura 4.14 e no Apêndice I, assim como as curvas classificadas em cada um deles. É possível observar clusters que não possuem curvas. Esta situação não acarreta nenhum tipo de problema ao processo, uma vez que as probabilidades relativas a este cluster serão sempre nulas. É interessante perceber que há basicamente duas grandes regiões distintas que separam as curvas mais semelhantes ao centroide do *cluster* 1 das curvas mais semelhantes ao centroide do *cluster* 78. Esta separação também é visualizada na Figura 4.15, que apresenta a distância entre os pesos de cada neurônio em relação aos seus vizinhos. As regiões vermelhas são aquelas em que ocorre o maior salto de distância em relação a *clusters* vizinhos, e portanto, são regiões de fronteira em relação às tipologias.

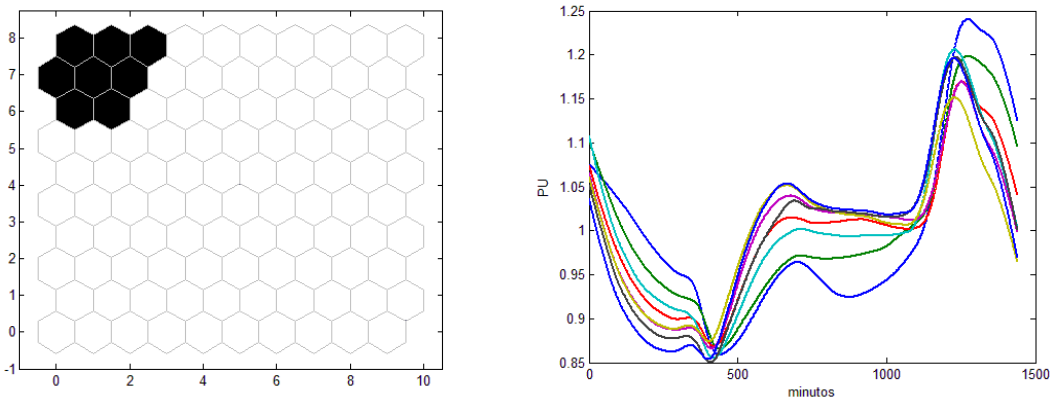


Figura 4.12 – Região de vizinhança do Cluster 1 e seus centroides.

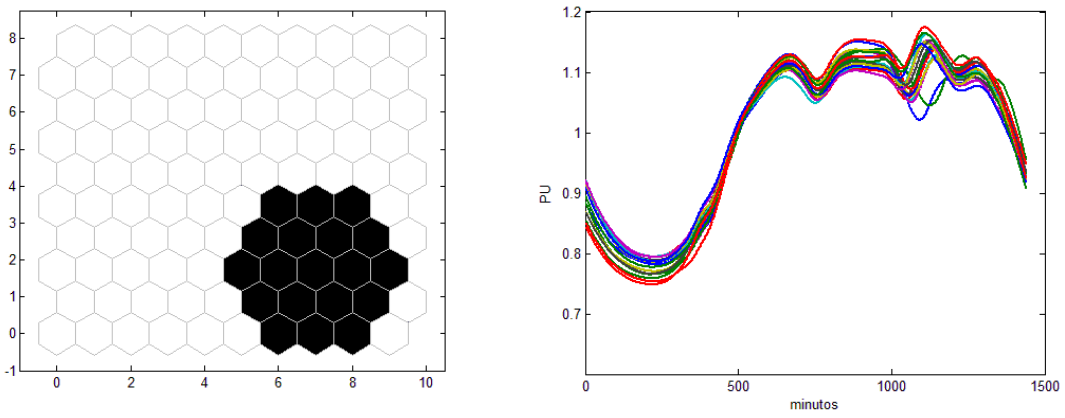


Figura 4.13 - Região de vizinhança do Cluster 78 e seus centroides.

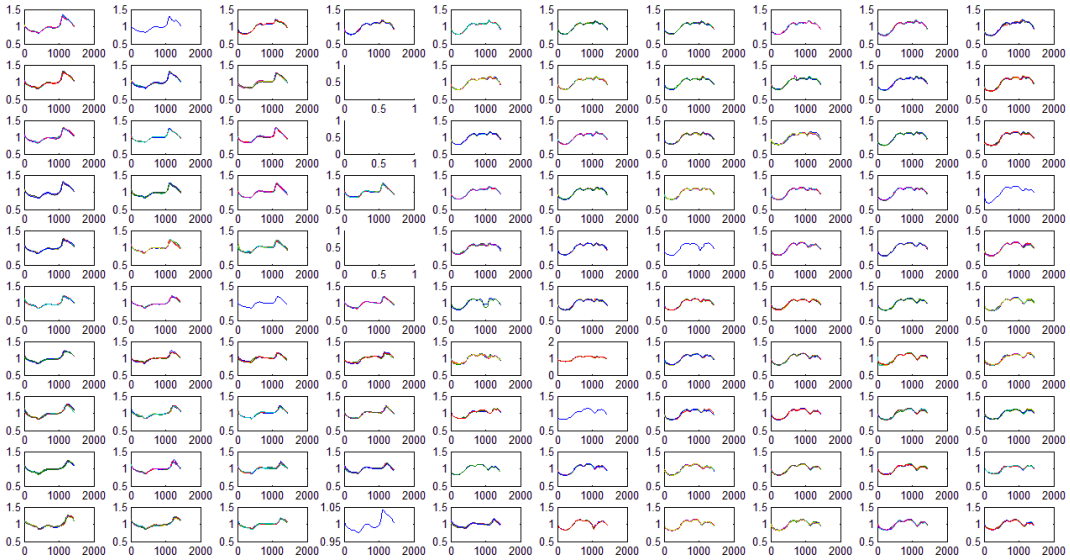


Figura 4.14 – Mapa de Kohonen com todos os 100 clusters.

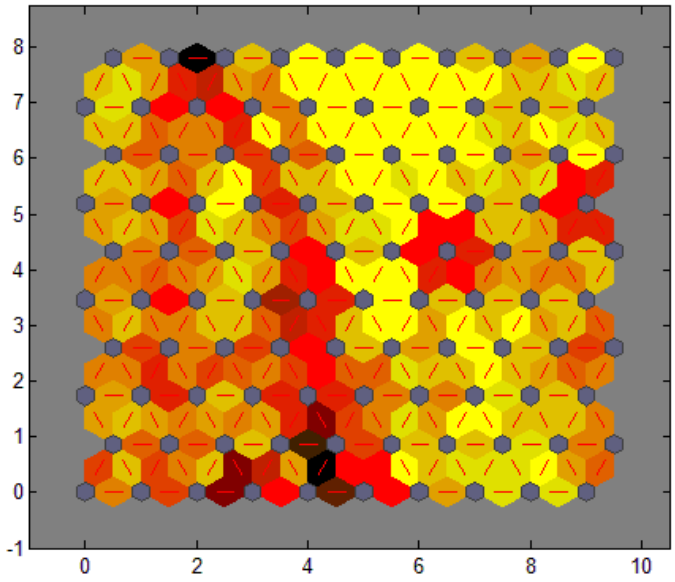


Figura 4.15 – Distância entre os pesos entre um neurônio e sua vizinhança.

Após a obtenção dos perfis típicos com o SOM, é necessário calcular as probabilidades condicionais em cada um dos *clusters*. As probabilidades serão usadas pelo Classificador Naive Bayes, descrito na Seção 2.3, para definir qual perfil típico será utilizado na correção das curvas defeituosas. Em cada *cluster* deve ser contabilizada a frequência de feriados, dias de horário de verão, dias de cada dia da semana (domingo a sábado), e dias de cada mês do ano (janeiro a dezembro). Na sequência, estas frequências calculadas devem ser divididas pelo total de curvas em cada *cluster*. Com isso, têm-se as probabilidades condicionadas de cada característica (feriado, horário de verão...), dado um *cluster* P(característica| *cluster*). O resultado desse cálculo é armazenado em forma de tabela, que é parcialmente visualizada na Figura 4.16. No total, são 100 linhas, onde cada linha representa um *cluster*. A primeira coluna representa a probabilidade de uma curva qualquer pertencer a este *cluster*, e é calculada simplesmente dividindo o número de curvas classificadas no *cluster* pelo número total de curvas utilizadas no treinamento. As demais colunas representam probabilidades condicionais de cada uma das características mencionadas: feriado, horário de verão, dia da semana (1 coluna para cada dia) e mês do ano (1 coluna para cada mês), dado que as curvas pertencem ao *cluster*, totalizando 22 colunas.

```
0.0054 1.0000 1.0000 0.2857 0.0000 0.2857 0.2857 0.1429 0.0000 0.0000 0.4286 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.5714
0.0163 0.0000 1.0000 1.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.4286 0.3333 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.2381
0.0031 0.2500 1.0000 0.0000 0.0000 0.2500 0.0000 0.0000 0.0000 0.7500 0.0000 0.2500 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.7500
0.0008 0.0000 0.0000 1.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 1.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
0.0171 0.0000 1.0000 0.0000 0.0000 0.0455 0.0455 0.0000 0.0000 0.9091 0.3182 0.5000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.1818
0.0023 0.0000 0.0000 0.0000 0.0000 0.3333 0.0000 0.3333 0.3333 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.3333 0.6667 0.0000 0.0000 0.0000 0.0000 0.0000
0.0101 0.0000 0.0000 0.0000 0.0000 0.1538 0.1538 0.3077 0.3846 0.0000 0.0000 0.4615 0.5385 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
0.0155 0.0000 1.0000 0.0000 0.0000 0.0500 0.2500 0.2000 0.5000 0.0000 0.2000 0.0500 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0500 0.2000 0.5000
0.0202 0.0000 1.0000 0.0000 0.0000 0.3077 0.1923 0.1538 0.3462 0.0000 0.5385 0.4615 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
0.0132 0.0000 1.0000 0.0000 0.0000 0.0588 0.1176 0.2941 0.5294 0.0000 0.3529 0.6471 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
                                                                                    .
                                                                                    .
0.0225 0.0000 0.0000 0.0000 0.9655 0.0000 0.0345 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.2069 0.3448 0.4483 0.0000 0.0000 0.0000 0.0000 0.0000
```

Figura 4.16 – Probabilidades relativas a cada cluster.

4.2.1.3 – CORREÇÃO DE LACUNAS E DADOS ABERRANTES

Nesta etapa são corrigidas as curvas que foram separadas na primeira etapa da filtragem e não participam do processo de *clusterização*. Como mencionado anteriormente, as lacunas de dados são substituídas pelo segmento da curva sugerido pelo perfil típico mais provável. A Figura 4.17 mostra um exemplo de correção de curva com lacuna, que teve o perfil típico da Figura 4.18 como sugestão de trajetória da curva. Já a Figura 4.19 mostra um exemplo de correção de curva com *outlier* detectado. O modelo substitui o ponto com valor aberrante pelo valor normal mais próximo.

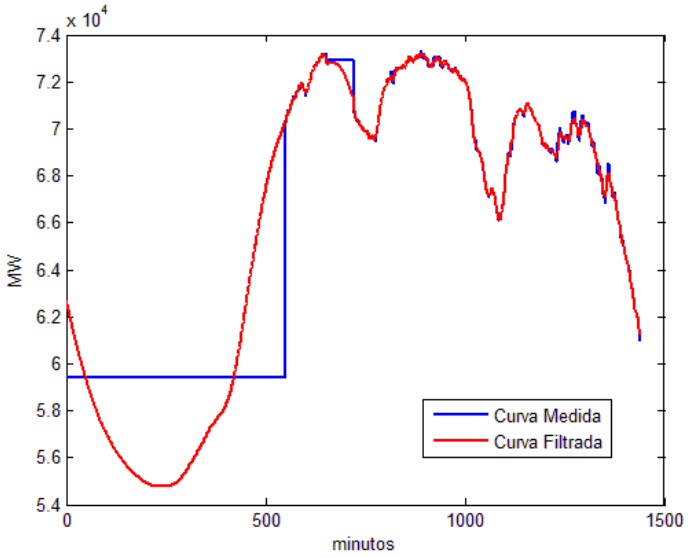


Figura 4.17 – Filtragem de curva com lacuna de dados.

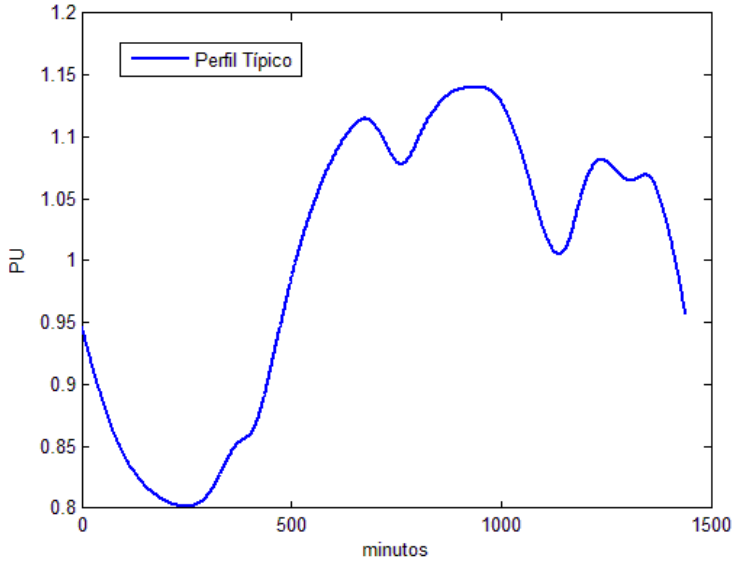


Figura 4.18 – Perfil típico utilizado na filtragem da curva com lacuna de dados.

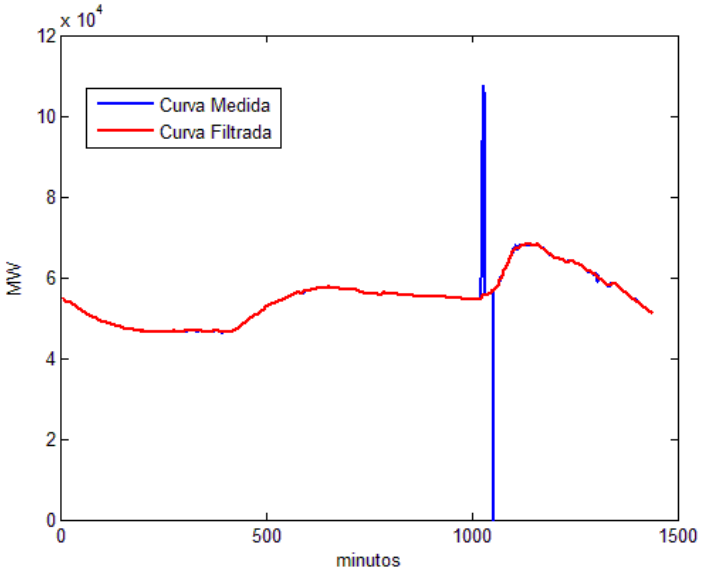


Figura 4.19 – Filtragem de curva com outlier.

4.2.1.4 – CORREÇÃO DE SEQUÊNCIAS ATÍPICAS

Nesta etapa da filtragem são tratadas as curvas em que não foram detectados os problemas de dados aberrantes ou lacunas de dados, mas que possuem discrepâncias em relação ao seu perfil típico associado. A Figura 4.20 ilustra um exemplo de curva com sequência atípica. Neste caso, toda a curva é considerada anormal, pois a curva apresenta uma magnitude muito acima do valor esperado a partir do seu perfil típico. Na Figura 4.21 é possível visualizar o perfil típico utilizado na identificação e correção da curva medida.

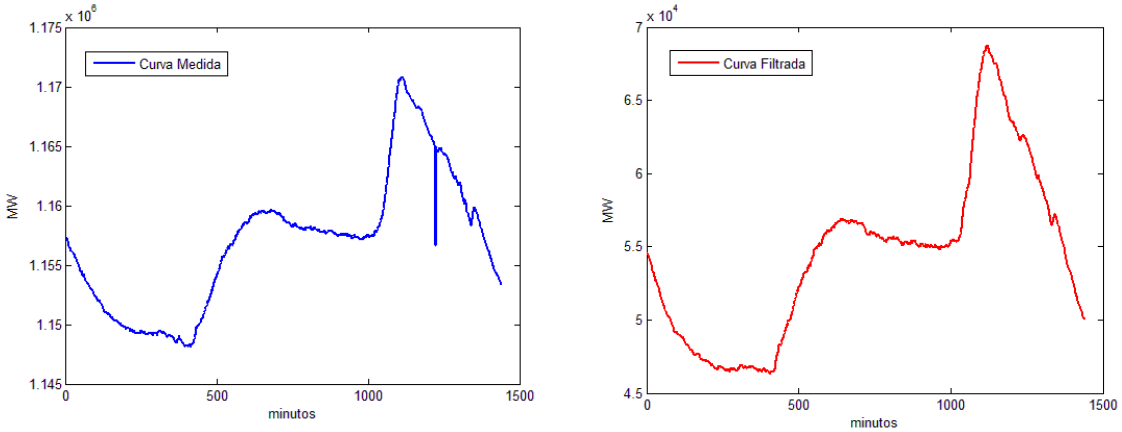


Figura 4.20 – Curva medida com sequência atípica e curva corrigida.

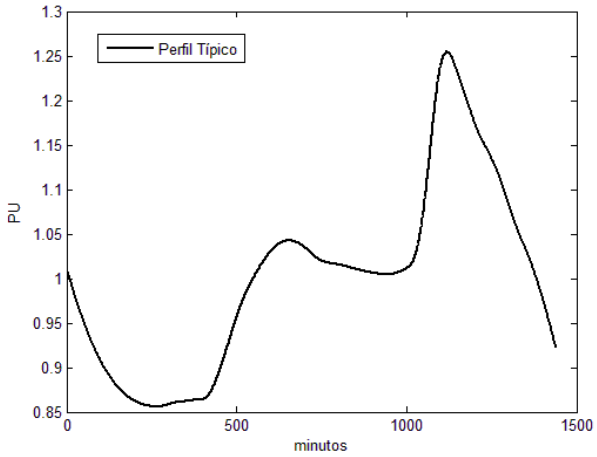


Figura 4.21 – Perfil típico utilizado na identificação e correção de curva com sequência atípica.

4.2.1.5 – CORREÇÃO DE DESCONTINUIDADES

Nesta última etapa de filtragem são corrigidas as curvas que apresentam um ou mais pontos de descontinuidade em sua trajetória. A Figura 4.22 ilustra um exemplo de curva com descontinuidade que, por acaso, também é considerada uma sequência atípica. Para corrigir o problema, a curva é comparada com o perfil típico da Figura 4.23.

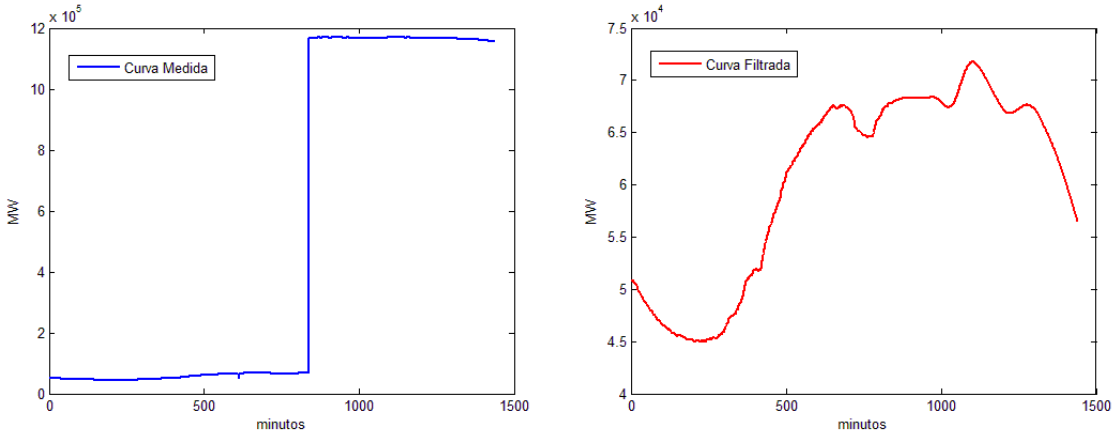


Figura 4.22 - Curva medida com descontinuidade e curva corrigida.

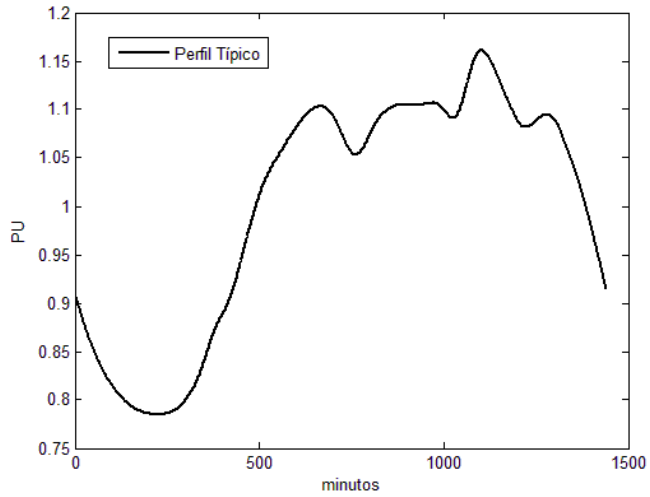


Figura 4.23 - Perfil típico utilizado na correção de curva com descontinuidade.

4.2.1.6 – OUTROS EXEMPLOS

A Figura 4.24 ilustra outros casos de filtragens realizadas na etapa de treinamento. Estas filtragens foram realizadas nas curvas que não participaram da análise de agrupamentos. Com estes exemplos, percebe-se a presença real dos erros abordados neste trabalho e a importância do tratamento destes dados.

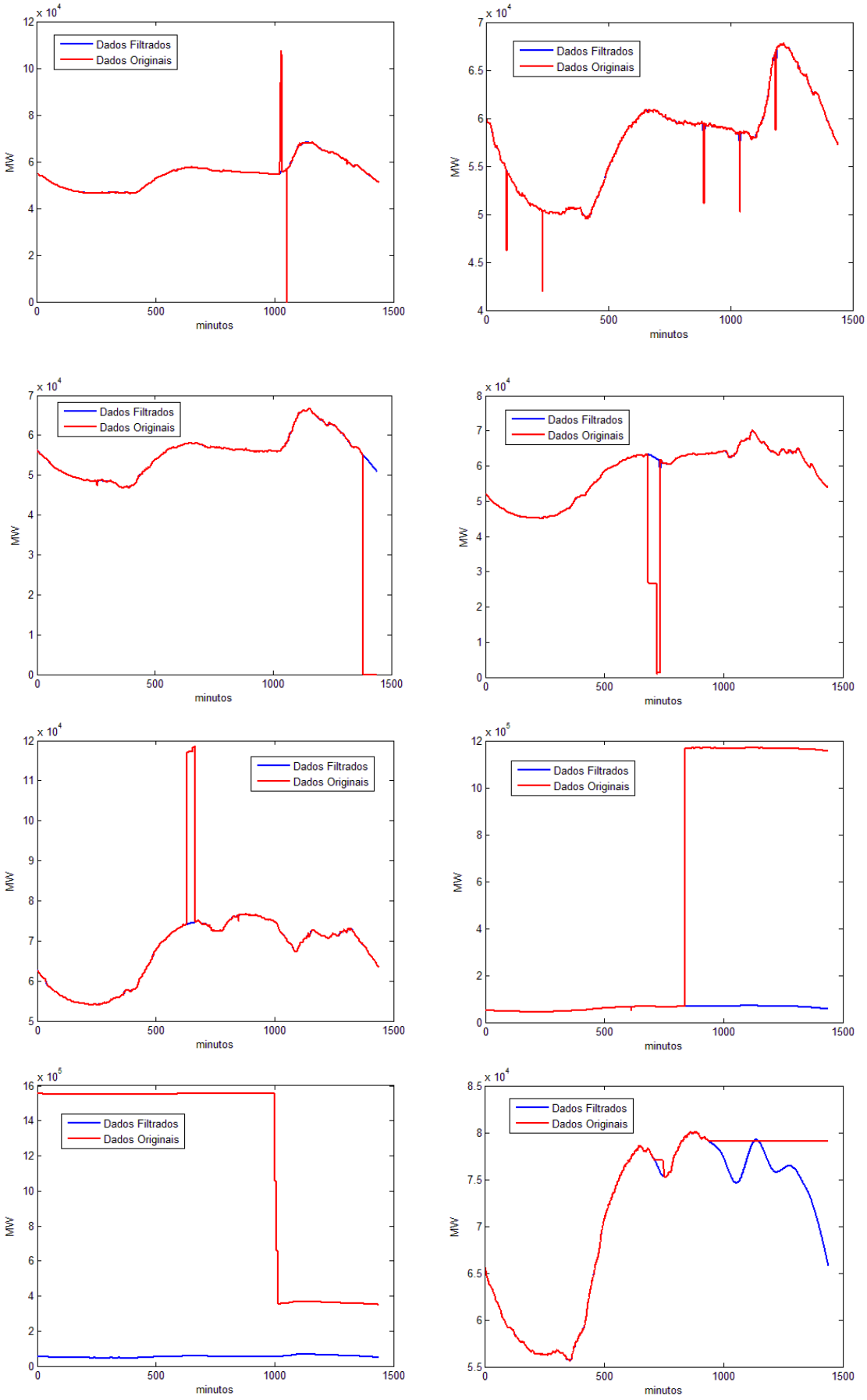


Figura 4.24 – Exemplos de filtragens de curvas reais durante a etapa de treinamento.

4.2.2 – ETAPA DE AVALIAÇÃO

Na etapa de treinamento foram gerados os perfis típicos e calculadas as probabilidades condicionais para a aplicação do classificador Naive Bayes. Ainda na etapa de treinamento foram tratadas algumas curvas de carga e os resultados preliminares já mostram o potencial da metodologia proposta.

Já na fase de avaliação, os erros mais observados no conjunto de curvas de carga durante o treinamento são simulados no conjunto de 365 curvas (1 ano de dados), não utilizadas no treinamento. Na sequência, as curvas com erros simulados são tratadas conforme a metodologia proposta e os resultados são comparados com os dados originais, i.e, antes de semrem corrompidos artificialmente. Assim, pode-se avaliar a qualidade da filtragem.

4.2.2.1 – SIMULAÇÃO DE ERROS

Os erros introduzidos aos dados originais são de três tipos diferentes: *outliers*, lacuna de dados e descontinuidades. Para a simulação de dados aberrantes sorteou-se aleatoriamente as curvas que iriam receber *outliers*. Depois, para cada curva, foram sorteados, também de maneira aleatória, os instantes de tempo da curva em que seriam inseridos os dados aberrantes. Por fim, os *outliers* foram gerados seguindo uma distribuição normal em torno da curva original e adicionados à mesma. Esse processo é ilustrado na Figura 4.25.

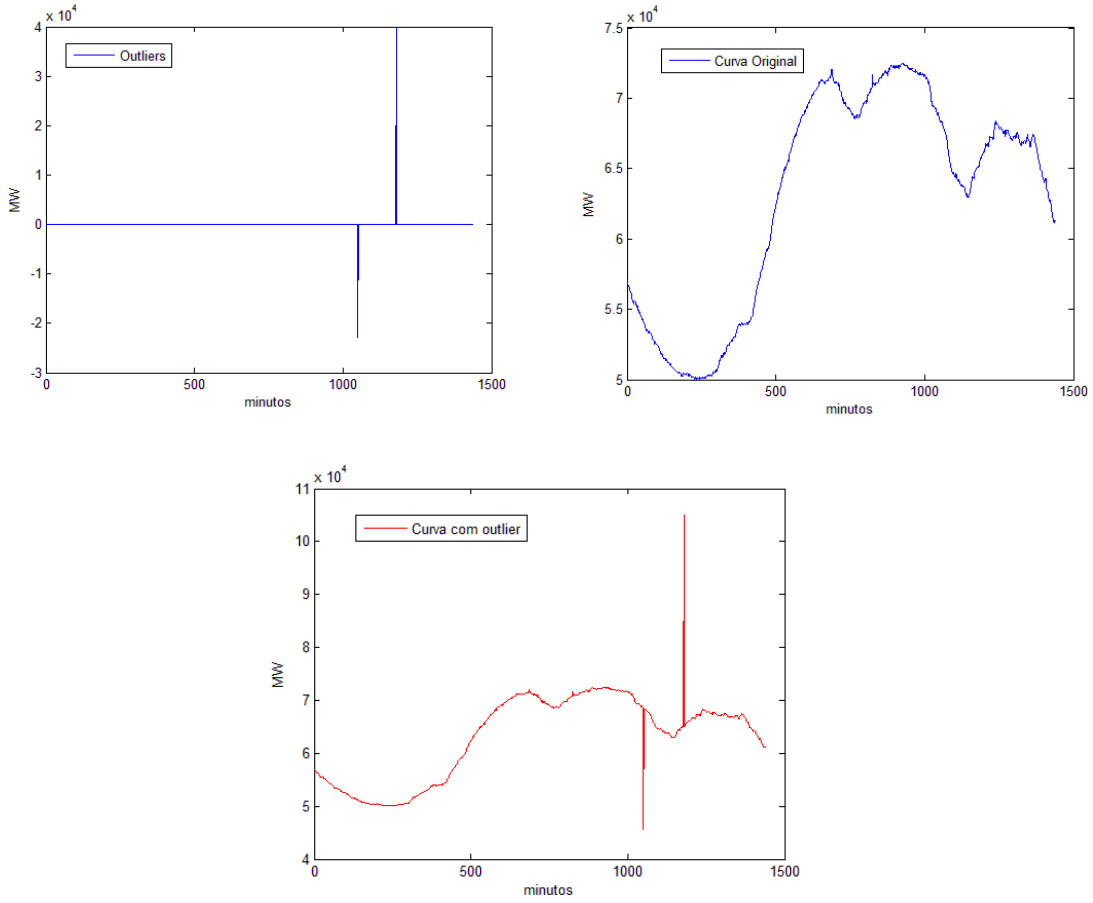


Figura 4.25 – Outliers simulados e inseridos na curva de carga original.

De maneira análoga, para a inserção de lacunas de dados, as curvas modificadas são sorteadas aleatoriamente. Há ainda o sorteio do instante de tempo inicial da lacuna de dados e a sua duração. A duração da lacuna de dados foi limitada a 576 pontos, i.e., 40% dos 1440 pontos de um perfil diário. A Figura 4.26 ilustra um exemplo de uma lacuna de dados de 323 minutos de duração inserida no instante 716 de uma curva diária.

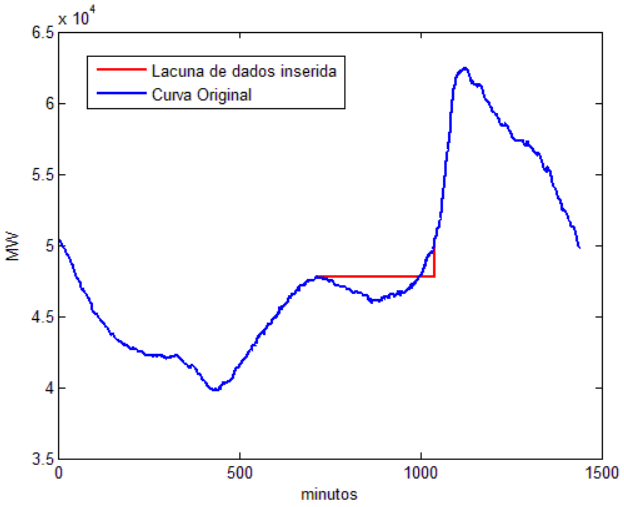


Figura 4.26 – Lacuna de dados inserida artificialmente em uma curva de carga.

Por fim, são inseridas as descontinuidades nas curvas de carga. Para isso, são sorteadas aleatoriamente as curvas que irão apresentar descontinuidade, o instante da descontinuidade e se a curva de carga é deslocada para cima ou para baixo. Nesta simulação foi adotado o máximo de uma descontinuidade por curva. Depois, o intervalo selecionado no sorteio é multiplicado por um valor aleatório com distribuição normal. A Figura 4.27 ilustra um exemplo de descontinuidade inserida artificialmente em uma curva de carga.

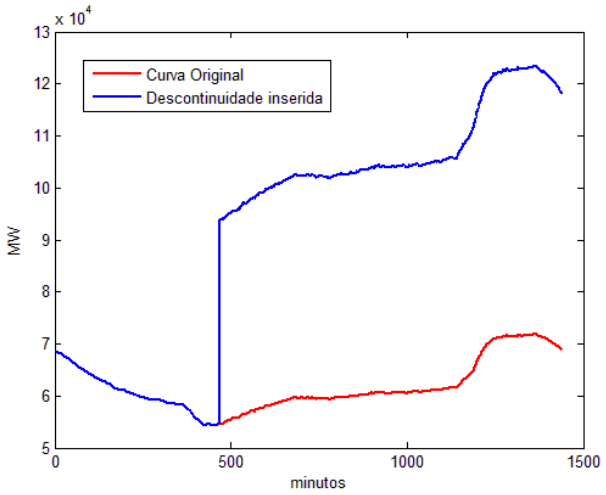


Figura 4.27 - Descontinuidade inserida artificialmente em uma curva de carga.

4.2.2.2 – FILTRAGEM DOS DADOS SIMULADOS

Os dados simulados, ou seja, aqueles em que foram introduzidos erros de maneira artificial são agora filtrados utilizando a metodologia proposta. Para uma melhor avaliação, as lacunas de dados, as observações aberrantes e as descontinuidades foram inseridas no mesmo conjunto de dados, porém de forma independente. Isto é, primeiro tem-se o conjunto de dados original com a simulação das lacunas de dados. Depois, o mesmo conjunto de dados originais com a simulação de *outliers.* Por ultimo, o conjunto de dados originais com a simulação de descontinuidades.

Nesta simulação, foram inseridas lacunas de dados em todas as 365 curvas referentes a 1 ano de dados. A Figura 4.28 e a Figura 4.29 ilustram exemplos da filtragem realizada utilizando a metodologia proposta. No lado esquerdo, as curvas em azul representam os dados originais sem qualquer tipo de tratamento, as curvas em preto representam os dados que tiveram as lacunas de dados introduzidas artificialmente e as curvas em vermelho representam os resultados da filtragem realizada pela metodologia proposta. No lado direito a curva em vermelho representa a curva típica adotada na correção da respectiva curva. É possível verificar a boa aderência das curvas filtradas em relação às curvas originais em lacunas de diferentes durações e instantes iniciais, além dos diferentes tipos de curvas tratados (feriados, finais de semana ou dias úteis).

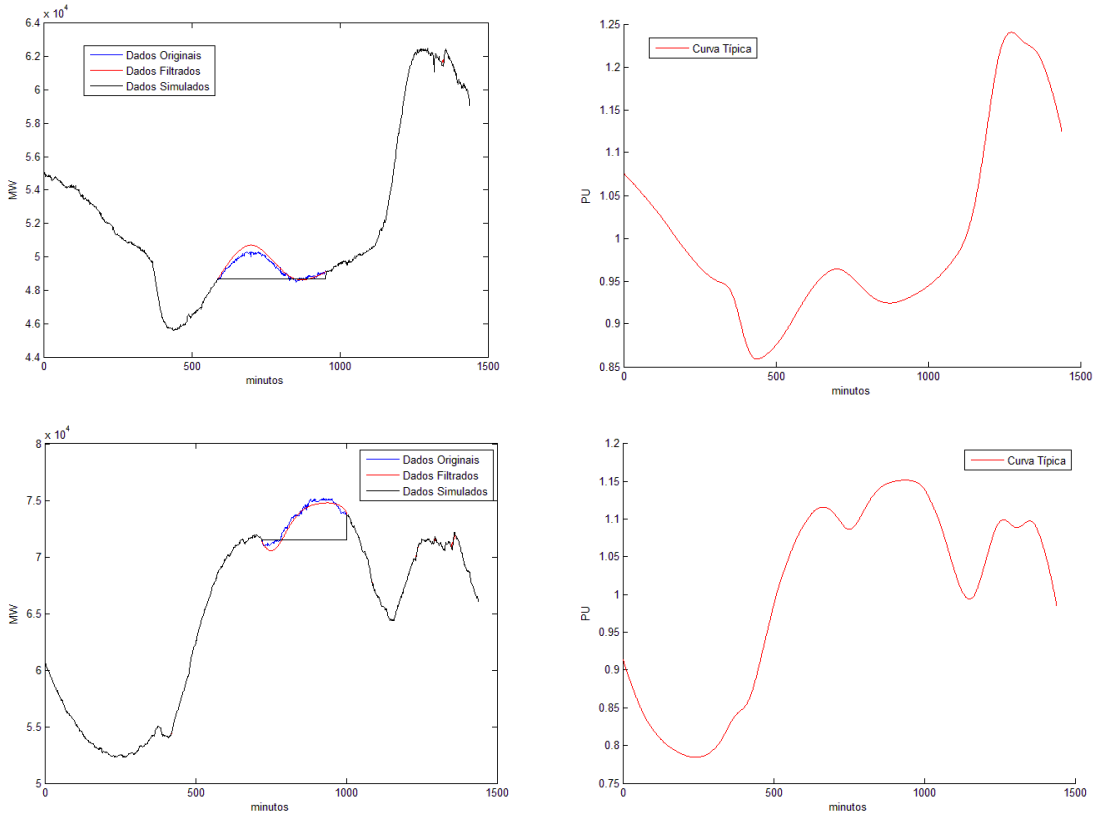


Figura 4.28 – Comparação entre os dados originais, lacunas simuladas e dados filtrados.

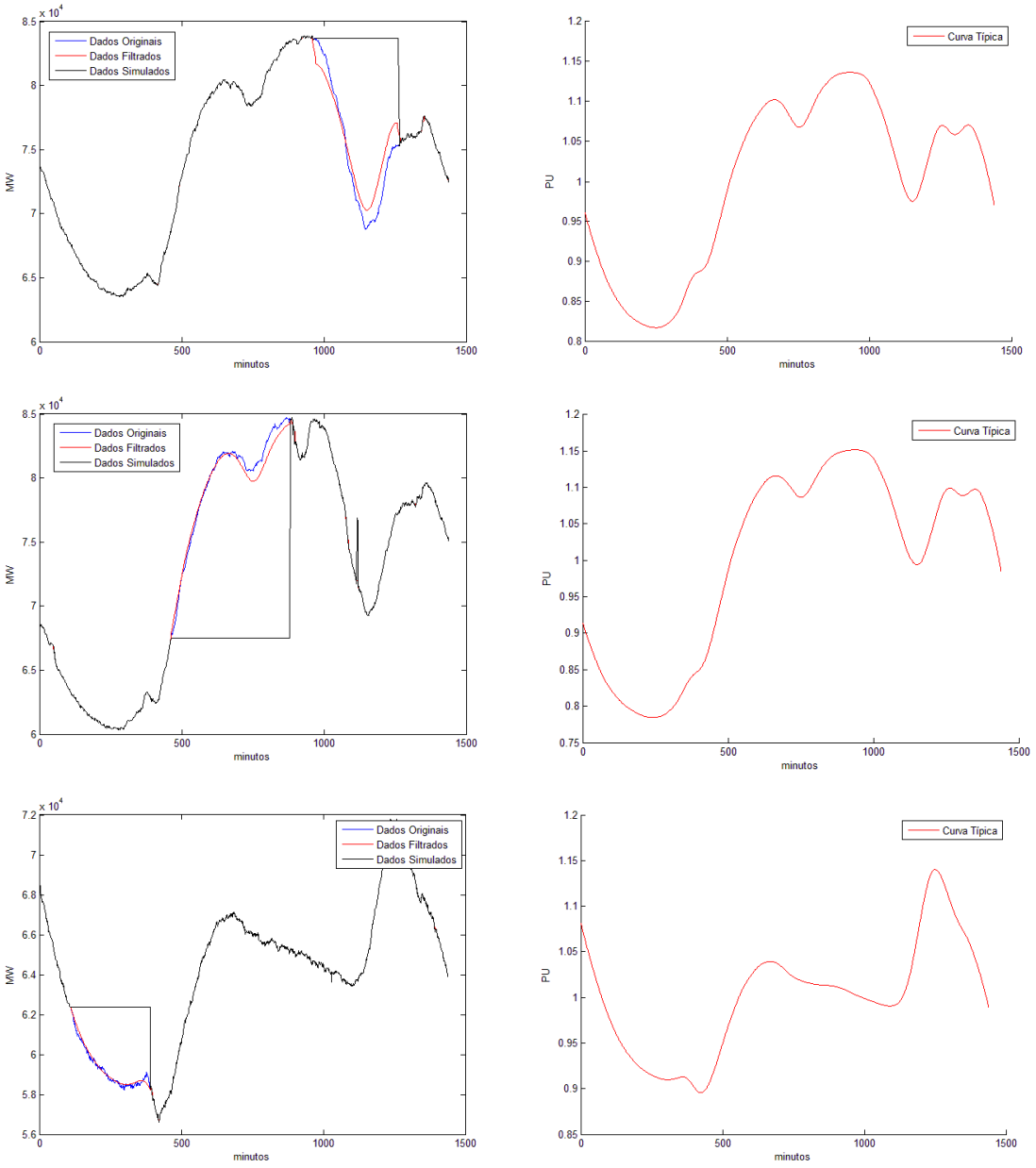


Figura 4.29 – Comparação entre os dados originais, lacunas simuladas e dados filtrados.

A Figura 4.30 ilustra o resultado da filtragem de *outliers* por meio da metodologia proposta. O lado esquerdo contém as curvas com os *outliers* simulados artificialmente, enquanto o lado direito apresenta as curvas originais em azul e as filtradas em vermelho. Novamente, é possível observar que em todos os casos a metodologia proposta foi capaz de tratar este tipo de erro.

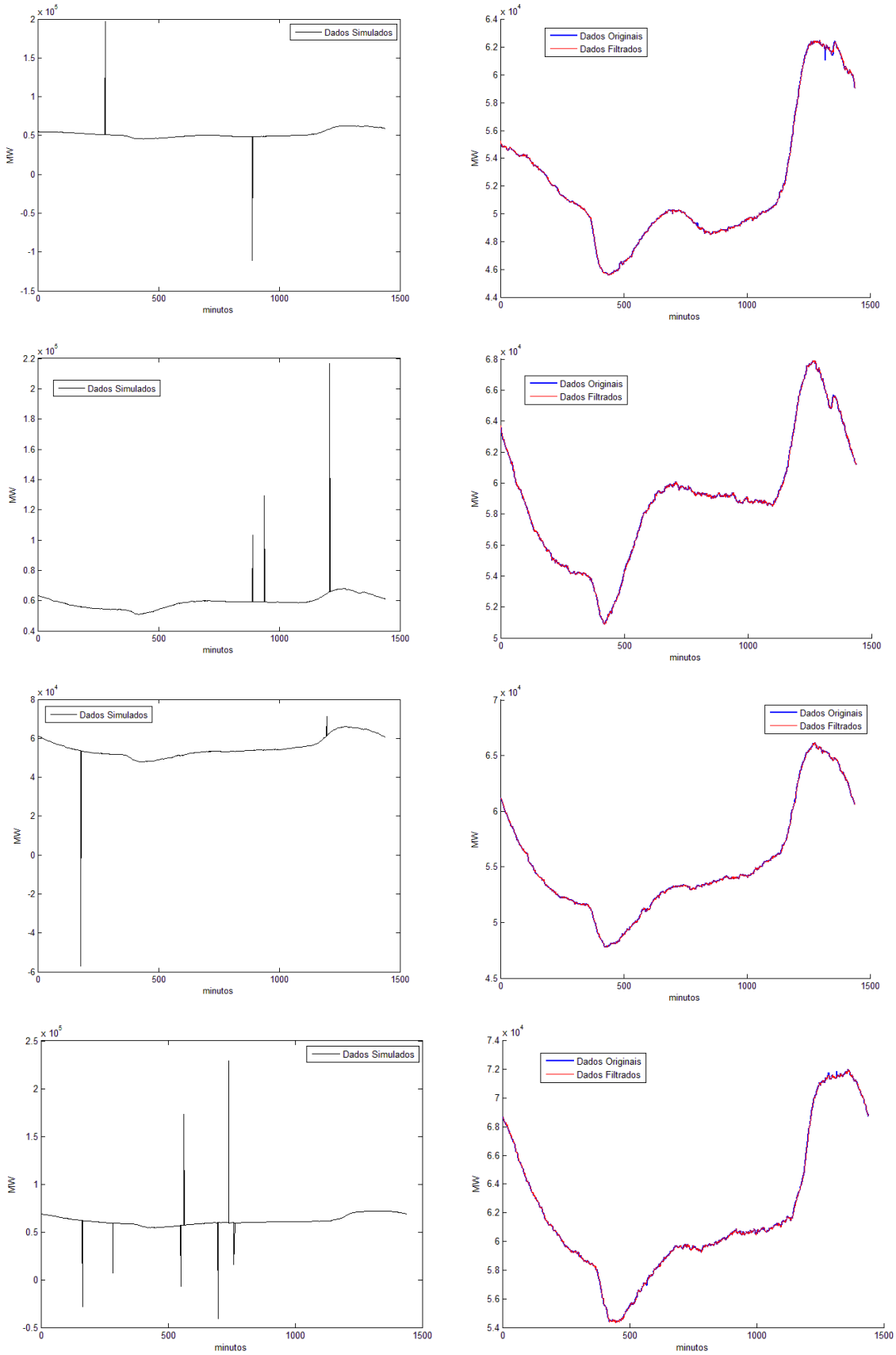
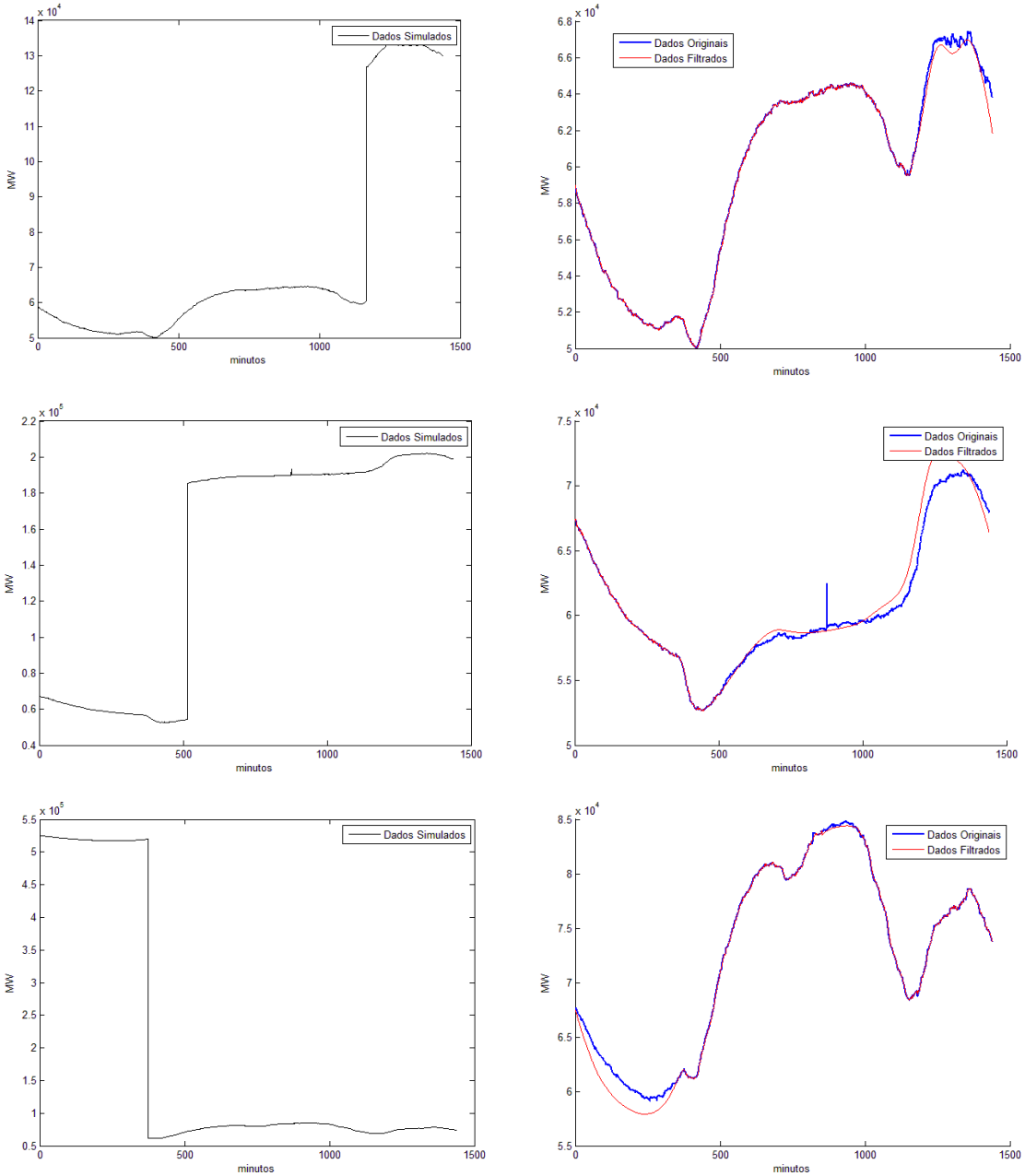


Figura 4.30 - Comparação entre os dados originais, outliers simulados e dados filtrados.

A Figura 4.31 ilustra os resultados da filtragem de descontinuidades através da metodologia proposta. No lado esquerdo estão as descontinuidades que foram inseridas de maneira artificial. No lado direito, em azul, estão as curvas originais às quais foram adicionadas as descontinuidades e, em vermelho, as curvas filtradas. Pode-se observar uma razoável aderência das curvas filtradas às curvas originais. Também é possível considerar, ainda que de maneira preliminar, que quanto maior o período deslocado da curva de carga, mais difícil é para o modelo se aproximar da curva original. A Figura 4.32 ilustra mais exemplos de filtragens de descontinuidades simuladas.



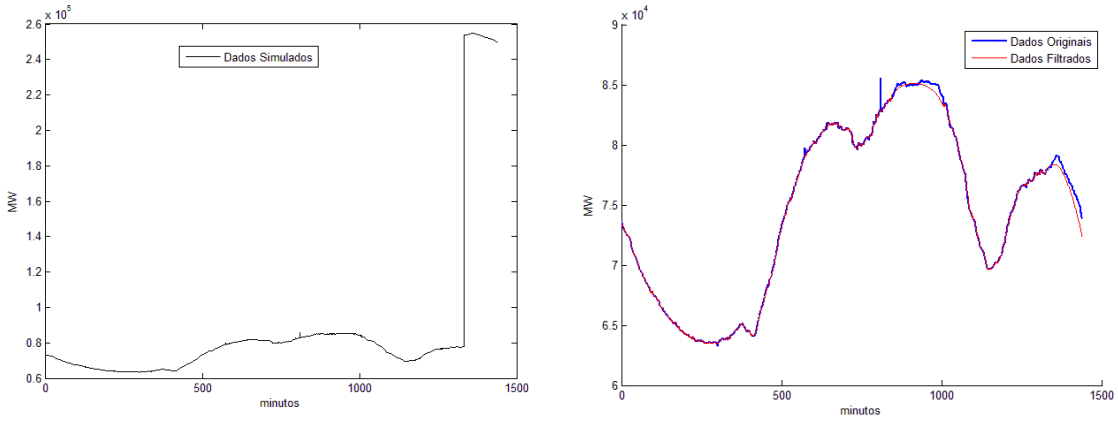


Figura 4.31 - Comparação entre os dados originais, descontinuidades simuladas e dados filtrados.

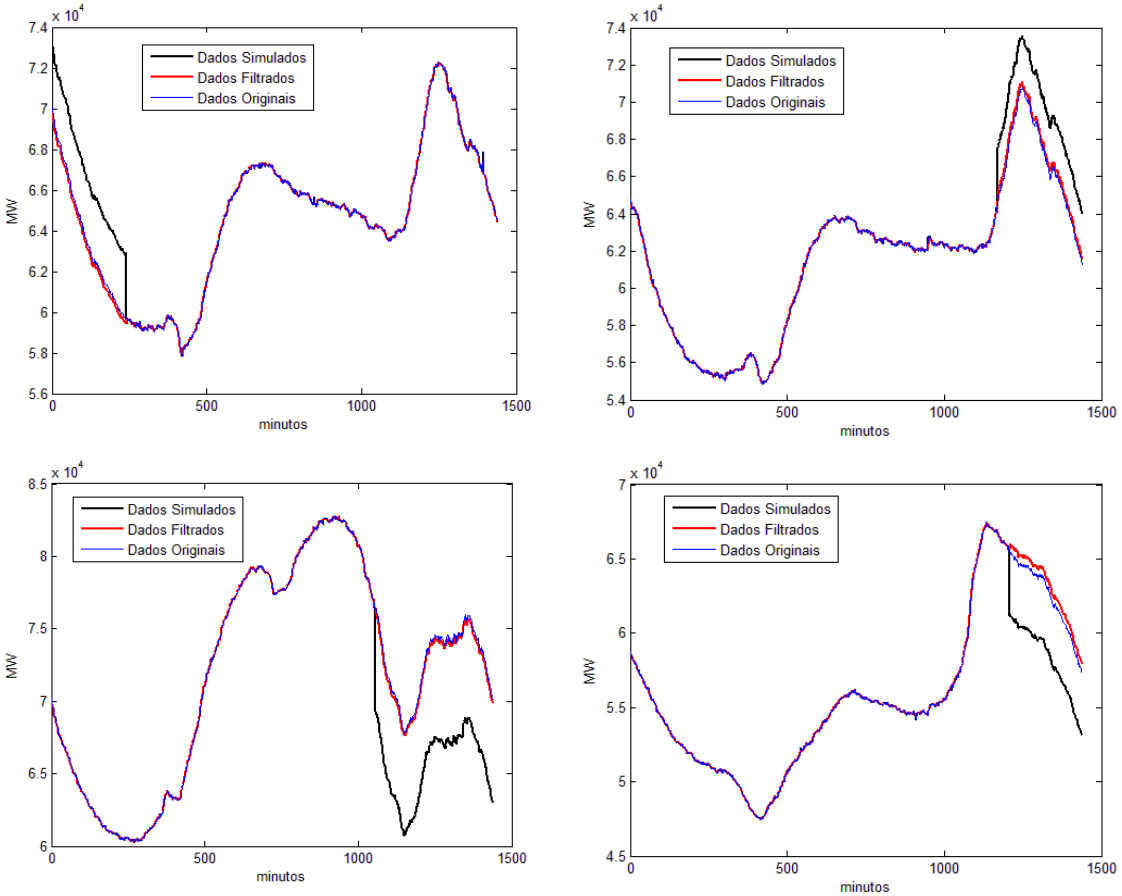


Figura 4.32 – Outros exemplos de filtragens de descontinuidades em dados simulados.

4.2.2.3 – DESEMPENHO DA FILTRAGEM DE DADOS SIMULADOS

Para a avaliação do desempenho na filtragem de *outliers*, foi adotada a métrica utilizada em Akouemo & Povinelli [39]. Nela, os valores corrigidos dos pontos identificados como *outliers* são comparados com o valor original. A Tabela 4.2 apresenta os resultados obtidos na filtragem dos *outliers*. No total foram introduzidos 815 dados aberrantes, sendo 100% identificados pelo modelo. Deste total, o desvio entre o valor filtrado e o valor original foi menor que 1% em 98,9% dos casos.

Tabela 4.2 – Desempenho da filtragem de *outliers*.

Número de outliers introduzidos	Número de outliers detectados	Desvios < 1%	Desvios < 2%	Desvios < 5%
815	815 (100%)	806 (98,9%)	809 (99,26%)	815 (100%)

Para a avaliação do desempenho na filtragem de lacuna de dados e descontinuidades, foi adotado o índice de desempenho conhecido como MAPE (*Mean Absolute Percentage Error*) (4.1):

$$MAPE = \frac{\sum_{i=1}^{n} \frac{|e_i - o_i|}{o_i}}{n} \qquad (4.1)$$

Em que:
e_i - valor corrigido pelo modelo;
o_i - valor observado na série histórica original;
n - número de observações.

O MAPE identifica o percentual médio que a filtragem se distanciou da curva original. Para as curvas com lacuna de dados e descontinuidades simuladas, os desvios foram calculados apenas nos fragmentos que apresentam estes tipos de erros, o que possibilita uma melhor sensibilidade na interpretação dos índices de cada curva filtrada.

Para ter uma base de comparação, os índices também foram calculados para uma metodologia alternativa de filtragem, em Pessanha et al. [22], que propõem a utilização do algoritmo de clusterização FCM (*Fuzzy C-Means*) [40] para o tratamento de dados de carga. Além disso, esta metodologia alternativa utiliza uma técnica [41] mais simples na associação de uma curva de carga qualquer a um perfil típico, na qual cada curva é classificada no *cluster* com centroide mais próximo com base na distância. Na Tabela 4.3 pode-se visualizar um quadro comparativo dos métodos utilizados na metodologia proposta e na metodologia alternativa. A metodologia proposta utiliza a combinação dos métodos de Mapas de Kohonen e Naive Bayes, enquanto que a metodologia alternativa utiliza a combinação dos métodos FCM e menor distância euclidiana.

Tabela 4.3 - Métodos utilizados em cada modelo.

Função	Metodologia Proposta	Metodologia Alternativa
Agrupamento de curvas	Mapa de Kohonen	FCM
Associação aos perfis típicos	Naive Bayes	Distância Euclidiana

A Tabela 4.4 apresenta os erros médios (MAPE) na filtragem de curvas com lacunas de dados das duas metodologias a partir dos mesmos dados originais e simulados. Os índices estão separados por tipo de erro: lacuna de dados e descontinuidades.

Tabela 4.4 - Erros MAPE obtidos em cada modelo para cada tipo de erro.

Tipo de erro	Kohonen-Naive Bayes	FCM-Distância Euclidiana
Lacunas de dados	1,04%	1,95%
Descontinuidades	2,40%	2,88 %

Em relação ao tratamento de lacunas de dados, é possível visualizar na Figura 4.33 o histograma de erros da filtragem realizada pela metodologia proposta, que utiliza a combinação dos métodos de Kohonen e Naive Bayes. A distribuição de erros apresenta média de -68 MW e desvio padrão de 967 MW. Já na Figura 4.34 tem-se o histograma de erros da filtragem realizada pela metodologia alternativa, que utiliza a combinação dos métodos de FCM e Distância Euclidiana. Na metodologia alternativa, a distribuição de erros apresenta média de -284 MW e desvio padrão de 2430 MW. A Tabela 4.5 apresenta o percentual de ocorrências de erros em diversas faixas de valores para cada modelo. Nela, é possível observar que em 81% dos casos os erros entre os valores filtrados e os valores originais foram menores que 1% para a metodologia proposta, tendo este um desempenho superior em relação à metodologia alternativa. Isto já era esperado devido à menor dispersão da sua distribuição de frequências.

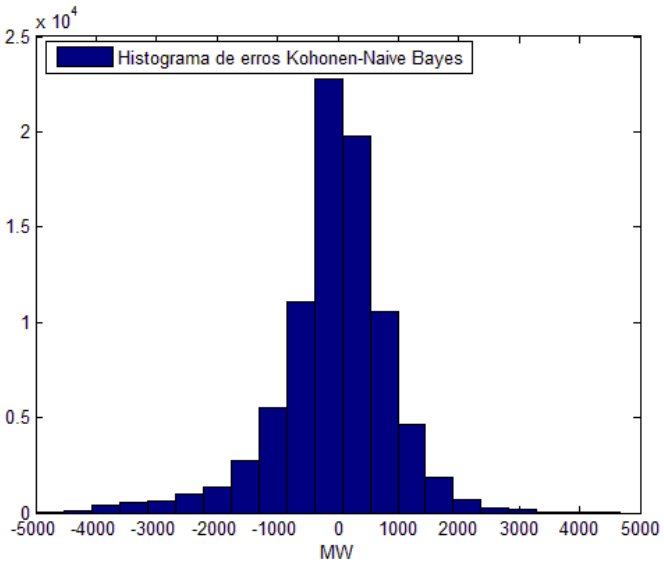


Figura 4.33 – Distribuição de frequências de erros da metodologia proposta – Lacuna de Dados.

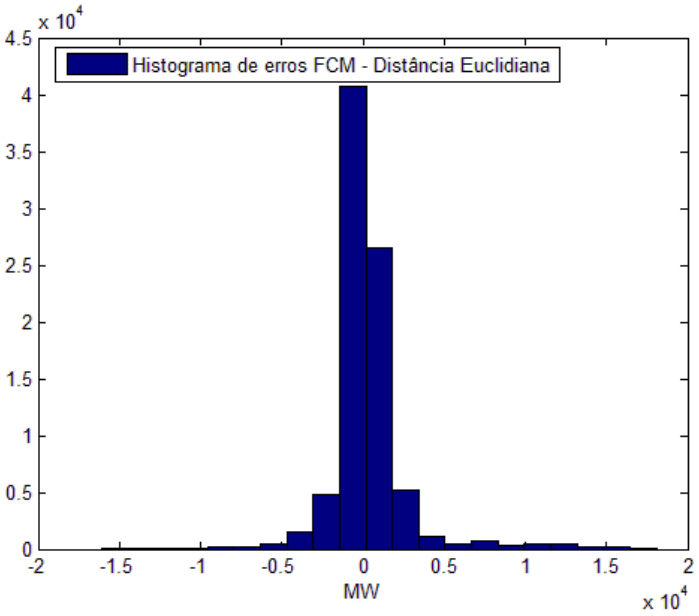


Figura 4.34 - Distribuição de frequências de erros da metodologia alternativa – Lacuna de Dados.

Tabela 4.5 - Distribuição de erros de cada modelo por faixa percentual – Lacuna de Dados.

Modelo	Erro < 1%	Erro < 2%	Erro < 5%
Kohonen-Naive Bayes	81,1%	94,9%	99,9%
FCM – D. Euclidiana	73,8%	84,4%	95,2%

Na Figura 4.35 é possível visualizar a série de erros percentuais de filtragem obtida pelas duas abordagens avaliadas. No modelo Kohonen-Naive Bayes, o erro máximo foi de 7,4%. Enquanto que no modelo FCM-Distância Euclidiana o erro máximo foi de 27,0%. Ao analisar este elevado erro da metodologia alternativa, fica evidente uma grande vantagem da metodologia proposta. O modelo Kohonen-Naive Bayes possui a capacidade de absorver informações relativas ao tipo de dia da curva a ser filtrada. A Figura 4.36 ilustra um bom exemplo de como estas informações podem afetar a filtragem. Nela, tem-se a curva original em azul, a curva simulada em preto, a curva filtrada pela metodologia proposta em vermelho e a curva filtrada pela metodologia alternativa em verde. É possível observar que há um erro de interpretação da metodologia alternativa sobre qual perfil típico utilizar para preencher a lacuna de dados, o que fica comprovado ao analisar na Figura 4.37 o perfil típico utilizado por cada metodologia.

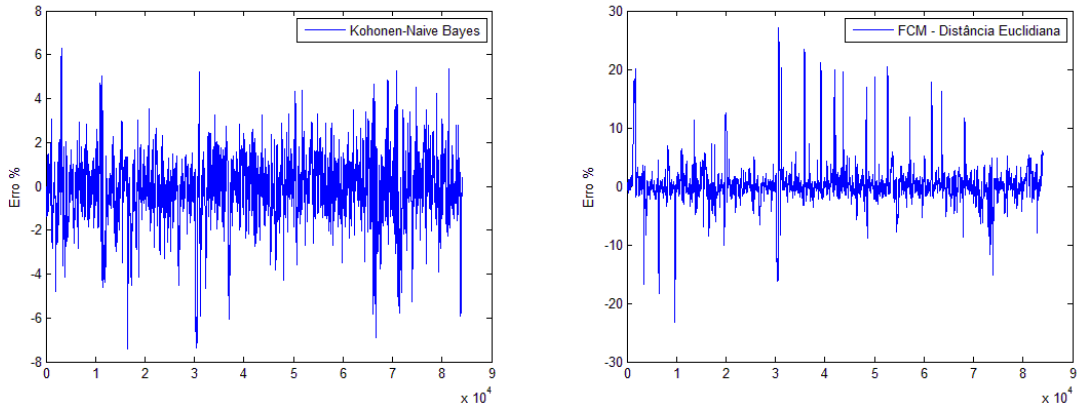


Figura 4.35 – Série de erros relativos da metodologia proposta e da metodologia alternativa – Lacuna de dados.

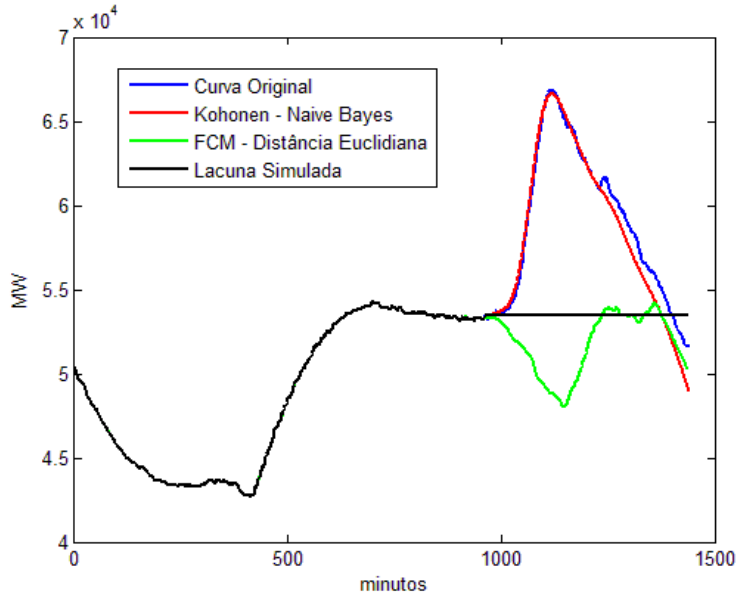


Figura 4.36 – Comparação entre as curvas tratadas pelas duas metodologias avaliadas e a curva original.

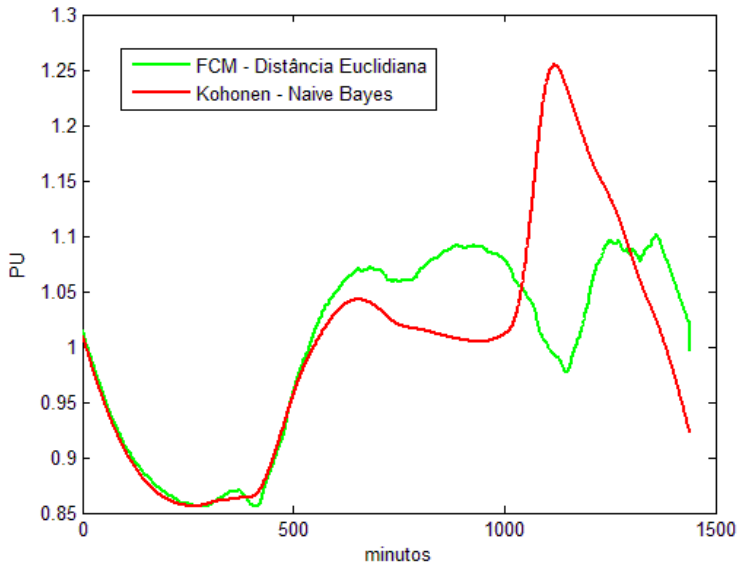


Figura 4.37 – Perfil típico utilizado por cada metodologia.

Em relação ao tratamento de descontinuidades de dados, é possível visualizar na Figura 4.38 o histograma de erros da filtragem realizada pela metodologia proposta, que utiliza a combinação dos métodos de Kohonen e Naive Bayes. A distribuição de erros apresenta média de -214 MW e desvio padrão de 3095 MW. Na Figura 4.39 é possível visualizar o histograma de erros da filtragem realizada pela abordagem alternativa, que utiliza a combinação dos métodos de FCM e Distância Euclidiana. A distribuição de erros apresenta média de 98 MW e desvio padrão de 3306 MW. A Tabela 4.6 apresenta o percentual de ocorrências de erros em diversas faixas de valores para cada modelo. Nela, é possível observar que em 73,8% dos casos os erros entre os valores filtrados e os valores originais foram menores que 1% para a metodologia proposta, tendo este um desempenho ligeiramente superior em relação à metodologia alternativa.

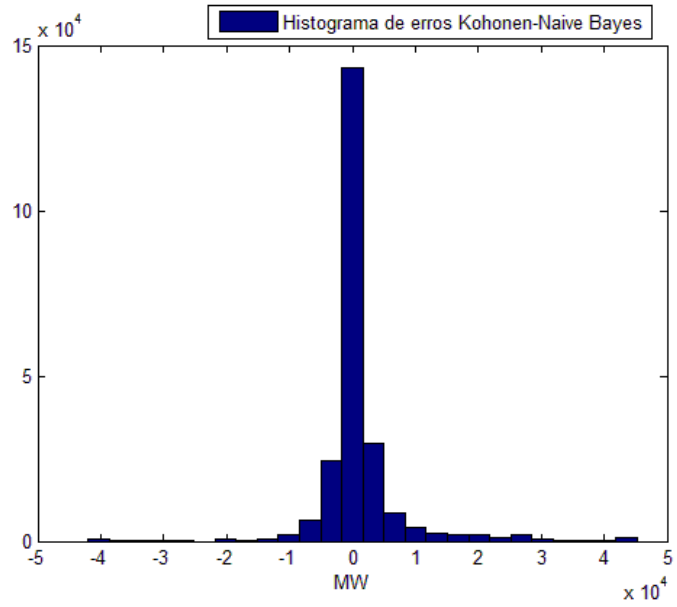


Figura 4.38 - Distribuição de frequências de erros da metodologia proposta – Descontinuidades.

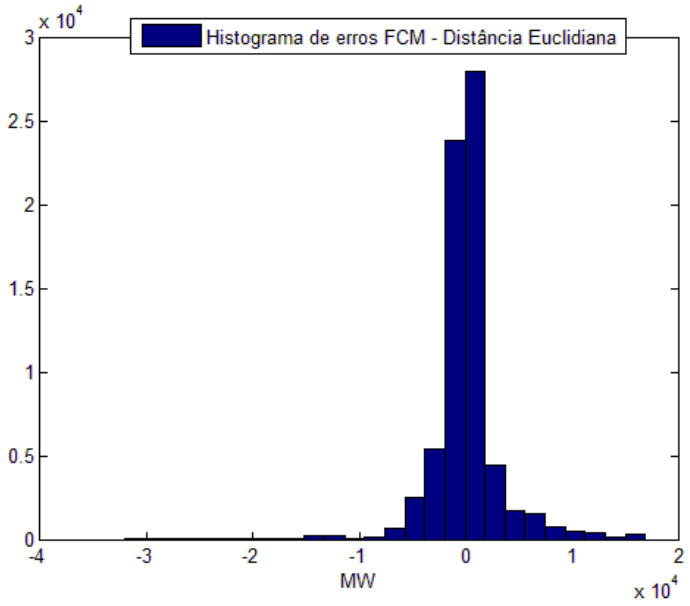


Figura 4.39 - Distribuição de frequências de erros da metodologia alternativa – Descontinuidades.

Tabela 4.6 - Distribuição de erros de cada modelo por faixa percentual – Descontinuidades.

Modelo	Erro < 1%	Erro < 2%	Erro < 5%
Kohonen-Naive Bayes	73,8%	85,3%	94,9%
FCM – D. Euclidiana	72,2%	82,0%	90,7%

Na Figura 4.40 é possível visualizar a série de erros percentuais de filtragem obtida pelos dois modelos. No modelo Kohonen-Naive Bayes, o erro máximo foi de 54,2%. Enquanto que no modelo FCM-Distância Euclidiana o erro máximo foi de 53,8%. Novamente, uma grande vantagem da metodologia proposta é a melhor associação da curva diária a um determinado perfil típico. A Figura 4.41 ilustra vários exemplos de filtragens de descontinuidades em que a metodologia proposta realiza uma correção satisfatória, enquanto que filtragem realizada pela abordagem alternativa diverge muito da curva original justamente pela escolha equivocada do perfil típico utilizado para a correção da curva diária. A Figura 4.42 ilustra outros exemplos de filtragens de descontinuidades simuladas, comparando a filtragem realizada por cada metodologia.

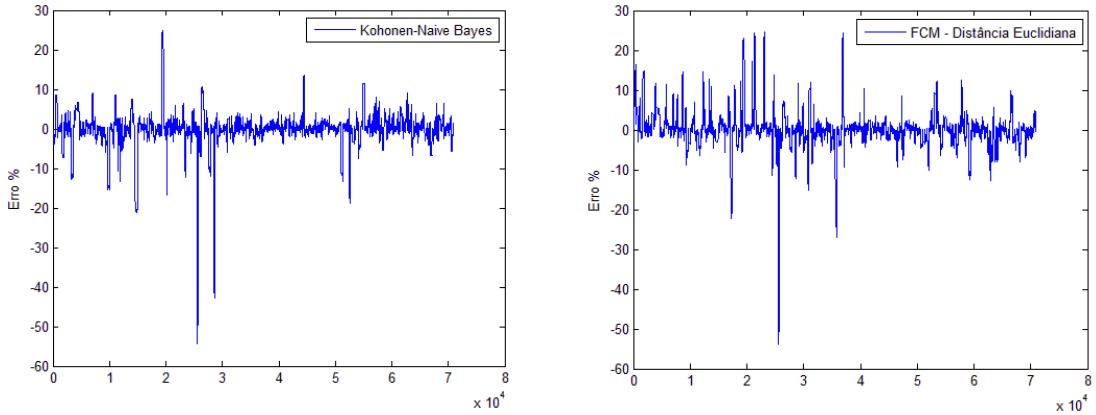


Figura 4.40 - Série de erros da metodologia proposta e da metodologia alternativa – Descontinuidades.

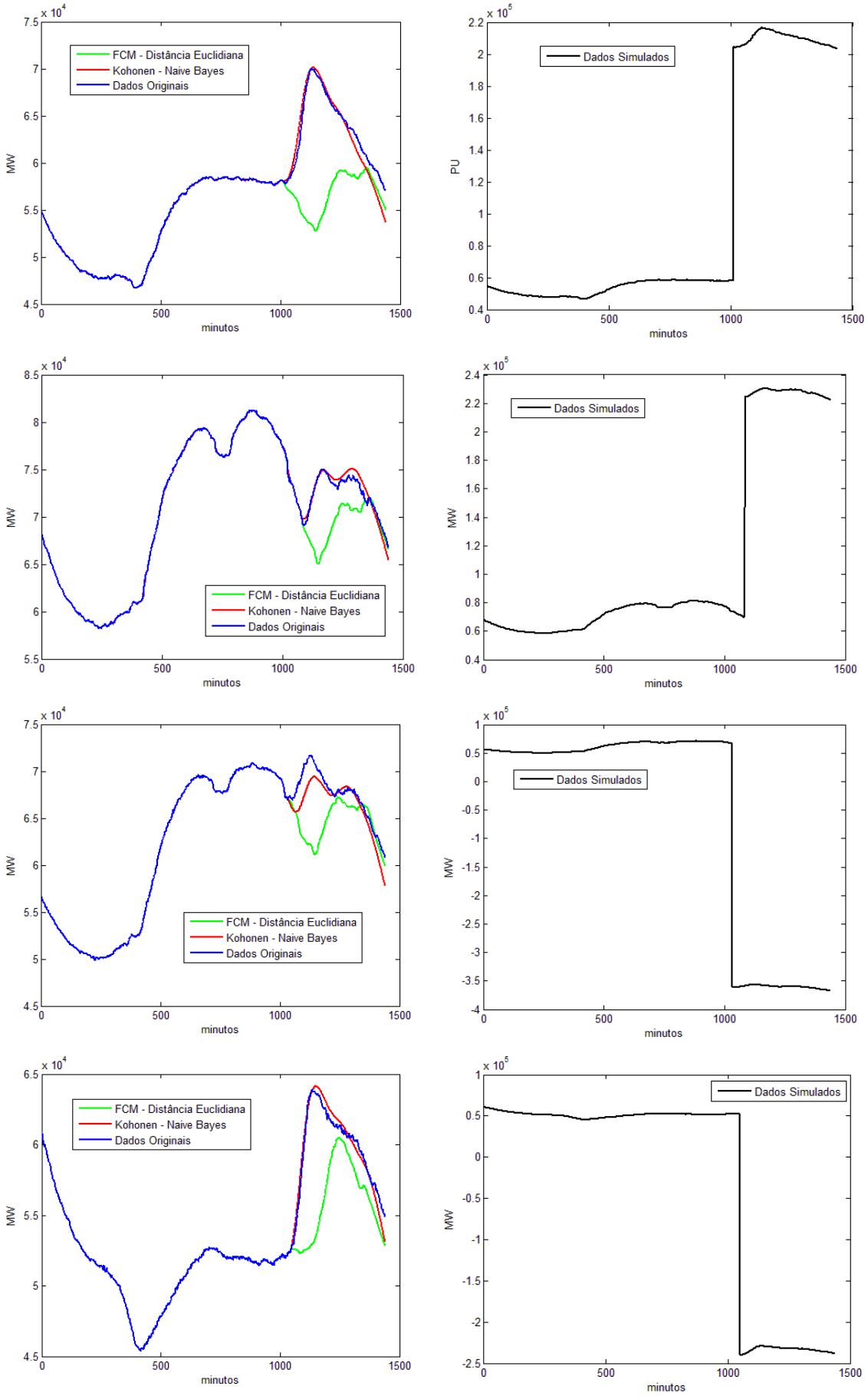


Figura 4.41 - Comparação entre filtragens, curva original e dados simulados.

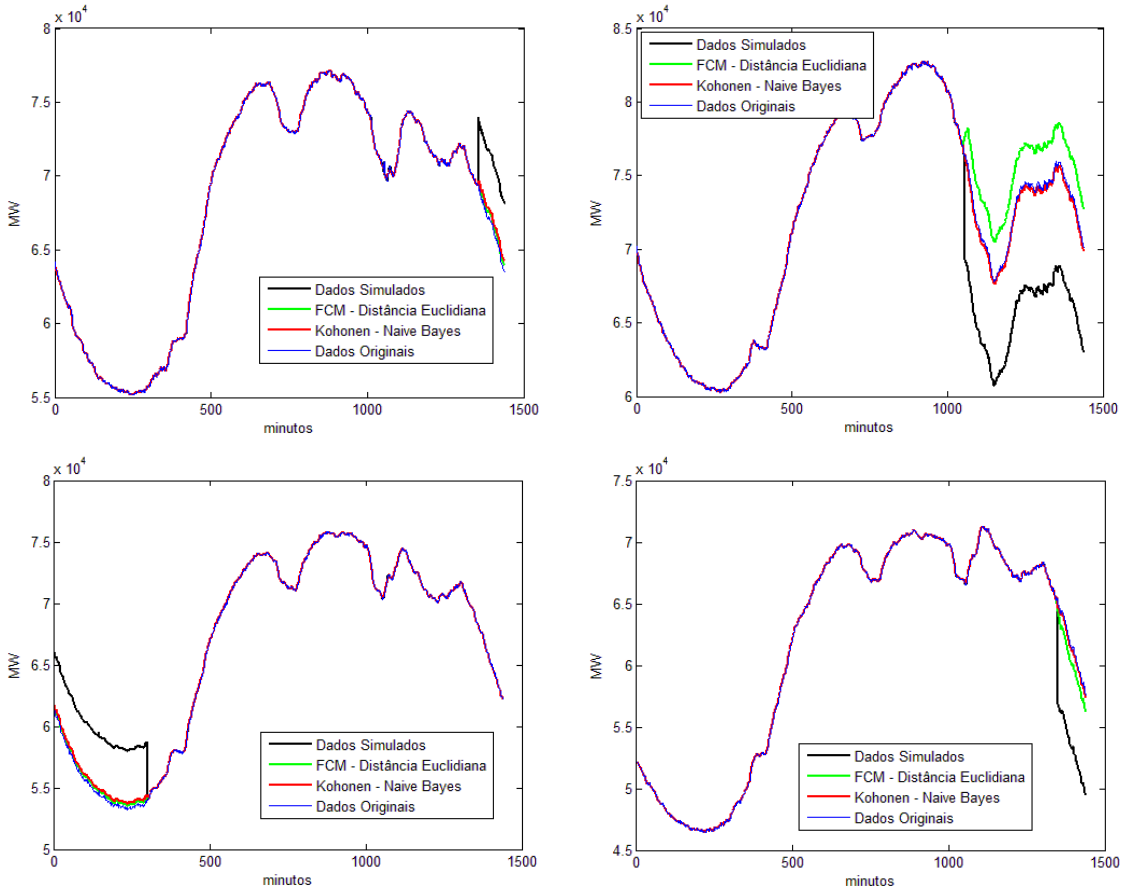


Figura 4.42 – Outros exemplos da filtragem de descontinuidades em dados simulados.

Após realizar os testes e avaliar o desempenho da filtragem em cada tipo de erro de maneira individual, foi realizada uma simulação em que as curvas poderiam apresentar qualquer um dos erros abordados, e até mesmo, todos eles. A Figura 4.43 e Figura 4.44 ilustram exemplos das filtragens realizadas nesta simulação. Como esperado, pelo fato da metodologia tratar separadamente cada tipo de erro, a inclusão de vários tipos de erros em uma mesma curva não influenciou de maneira significativa o tratamento dos dados, de modo que a metodologia adotada manteve o bom desempenho obtido nas simulações anteriores.

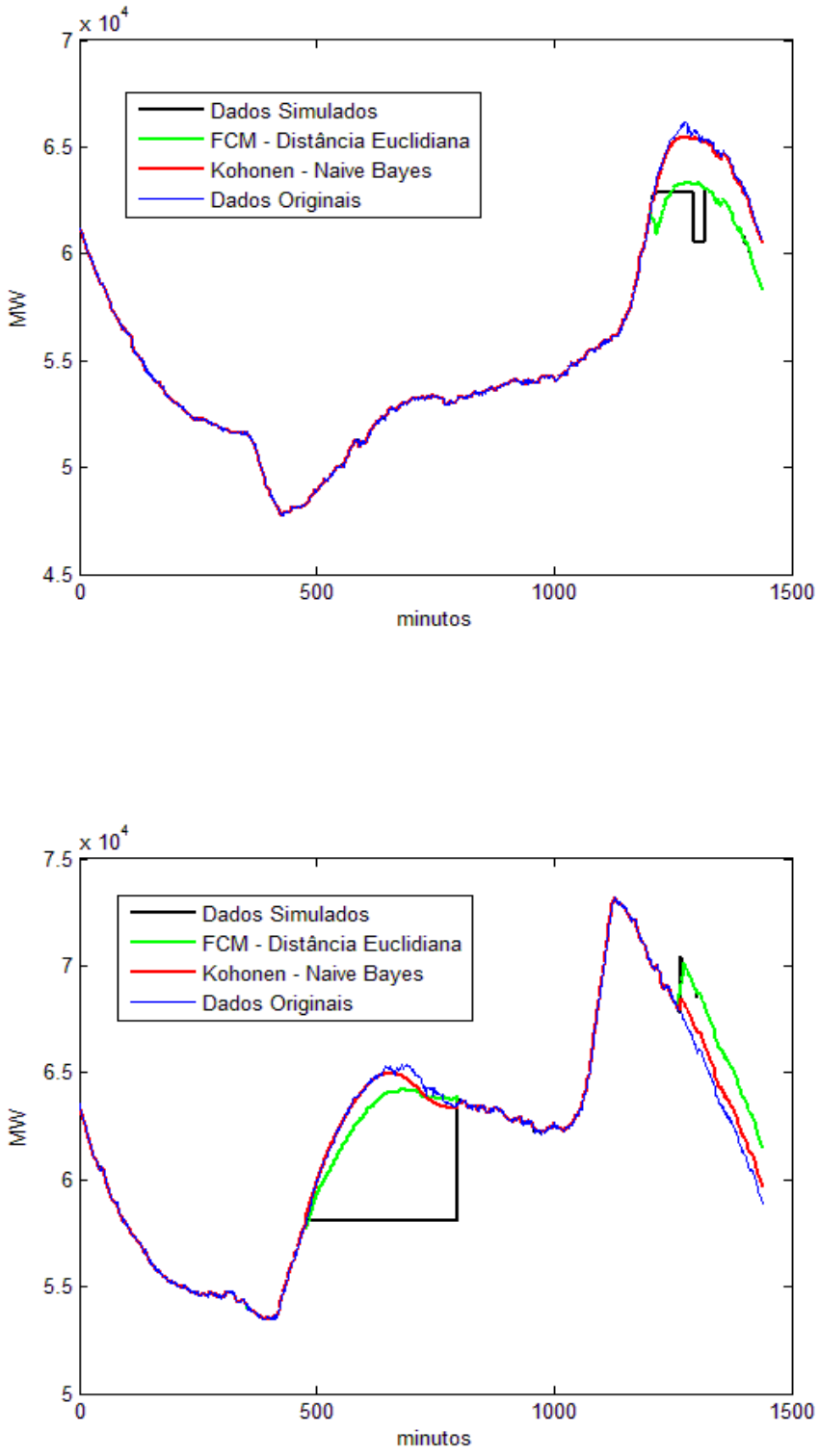


Figura 4.43 - Filtragens em simulações dos três erros em uma mesma curva.

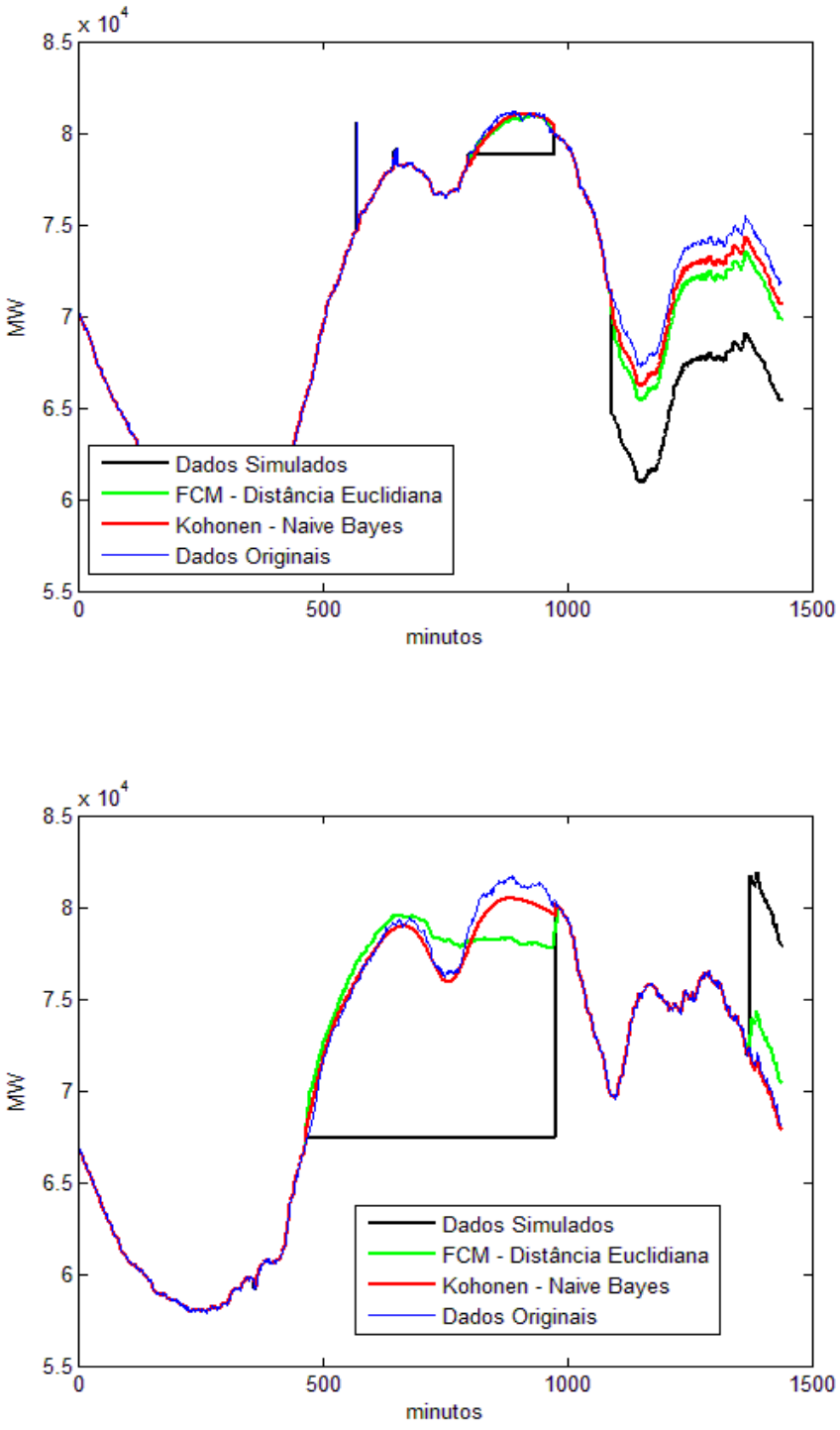


Figura 4.44 – Filtragens em simulações dos três erros em uma mesma curva.

5. CONCLUSÕES

A previsão da carga elétrica tem um papel decisivo na operação econômica e segura de um sistema elétrico de potência e a precisão destas previsões está diretamente ligada à qualidade dos dados de carga. No entanto, eventos inerentes à operação do sistema elétrico, como curtos-circuitos ou falhas de equipamentos, ou perturbações provocadas por problemas no sistema de medição, como falha de medidores ou transmissão de dados, são relativamente frequentes e podem comprometer a qualidade dos dados e, portanto, a capacidade preditiva de qualquer modelo de previsão de carga. Deste modo, os dados históricos da carga elétrica devem ser tratados previamente antes de serem utilizados como insumos para modelos de previsão, com o objetivo de minimizar a presença de erros.

A presente dissertação apresentou uma metodologia para o tratamento de curvas de cargas que utiliza várias técnicas de inteligência artificial. Dentre elas, o Mapa Auto-Organizável de Kohonen, que realiza uma classificação não supervisionada (análise de agrupamentos) dos dados e possibilita uma visualização das curvas de carga. Destaca-se também a utilização do Classificador Naive-Bayes, um classificador supervisionado que utiliza as probabilidades condicionais de atributos qualitativos, como por exemplo, os feriados e o horário de verão, para associar cada curva de carga verificada à um perfil típico. Ainda há a técnica LOWESS, um método de regressão local não paramétrico utilizado na suavização das curvas de carga, e o DBSCAN, um algoritmo para análise de agrupamentos com uma abordagem baseada em densidade que serviu de inspiração para detecção de pontos de descontinuidade nas curvas de carga.

A metodologia proposta constitui uma sequência de aplicação dos métodos citados juntamente com o uso de técnicas estatísticas de análise de dados como histogramas de frequência de dados, *boxplots*, diagramas de dispersão e intervalos de confiança.

Adicionalmente, para avaliar o desempenho da metodologia proposta foram elaborados experimentos computacionais organizados em duas grandes etapas: treinamento e avaliação. Na etapa de treinamento, um grande conjunto de dados de carga é analisado para a criação de uma base de perfis típicos que representam o comportamento da carga elétrica nas diferentes circunstâncias possíveis. Na sequência, os perfis típicos são probabilisticamente associados a cada dos perfis de carga analisados. Na etapa de avaliação, um outro conjunto de dados é artificialmente acrescido dos erros frequentemente observados nos dados aquisitados do sistema de medição. Assim, os dados artificiais são filtrados utilizando a metodologia proposta e os resultados são comparados com os dados originais. O objetivo consiste em avaliar a capacidade da metodologia proposta em recuperar o padrão original das curvas corrompidas.

Os resultados dos experimentos computacionais foram avaliados com base na métrica de desempenho MAPE, que representa o desvio médio percentual absoluto entre

as curvas filtradas e as curvas originais. Adicionalmente, como base de comparação, utilizou-se uma metodologia alternativa de tratamento de dados, que emprega o método *Fuzzy Clustering Means* (FCM) no agrupamento dos dados e geração de perfis típicos e os associa a cada uma das curvas utilizando apenas a distância euclidiana entre ambos.

Os resultados foram apresentados separadamente de acordo com cada tipo de erro. Primeiro, foram exibidos o desempenho do modelo na detecção e correção de *outliers* ou dados aberrantes. O modelo foi capaz de detectar 100% dos *outliers* gerados aleatoriamente, sendo o desvio entre valor corrigido e o valor original foi menor que 1% em aproximadamente 99% dos casos.

Posteriormente, foi avaliado o desempenho do modelo no preenchimento de lacunas de dados. O modelo apresentou erro médio percentual de cerca de 1% no conjunto de dados analisado, com o desvio entre a curva filtrada e a curva original sendo menor que 1% em aproximadamente 81% do tempo.

Finalmente, foi avaliado o desempenho do modelo na correção de descontinuidades nas curvas de carga. O modelo apresentou erro médio percentual de cerca de 2,4% no conjunto de dados analisado, com o desvio entre a curva filtrada e a curva original sendo menor que 1% em aproximadamente 74% do tempo.

Já no preenchimento de lacunas de dados e correção de descontinuidades a metodologia proposta apresentou um desempenho ligeiramente superior. A abordagem alternativa alcançou um erro médio percentual de aproximadamente 2,0% e 2,9% no tratamento de lacunas e descontinuidades, respectivamente, sendo o desvio entre a curva filtrada e a curva original menor que 1% em cerca de 74% e 72% do tempo. A principal diferença notada entre as duas abordagens é a capacidade da metodologia proposta de levar em consideração as características pertencentes a cada um dos dias analisados no período. Portanto, ela é capaz de sugerir perfis típicos mais adequados aos dias úteis, feriados, finais de semana e dias de horário de verão. Também são levados em consideração o dia da semana e o mês do ano da curva a ser corrigida. Isto fornece um grande ganho em muitas situações no preenchimento de lacunas e correção de descontinuidades. A melhor capacidade de associação dos perfis típicos deve-se a escolha do Classificador Naive-Bayes, sendo o método o maior responsável pelo melhor desempenho da metodologia proposta. Por outro lado, o Mapa Auto-Organizável de Kohonen e o FCM tiveram desempenhos equivalentes no agrupamento das curvas. No entanto, uma grande vantagem do SOM é a organização topológica dos *clusters*, de forma que é possível visualizar e analisar a forma na qual são organizados os perfis típicos e os pesos sinápticos da rede.

Ao analisar os índices de desempenho e os gráficos que ilustram alguns exemplos da filtragem, é possível conferir o bom desempenho da metodologia proposta no tratamento dos erros produzidos artificialmente na simulação. Com isso, conclui-se que a metodologia proposta está habilitada para o tratamento dos erros em casos reais, servindo de ferramenta na limpeza dos dados que podem ser introduzidos em um modelo de previsão de carga. De fato, a metodologia proposta substituiu a metodologia

alternativa, descrita em Pessanha et. al [22], no módulo de tratamento de dados do sistema de previsão de carga em uso pelo ONS e vem apresentando bons resultados.

Os resultados do experimento computacional mostram que a metodologia proposta conseguiu recuperar o perfil original da curva de carga que foi corrompido pelas falhas simuladas nos dados. Embora melhorias sejam sempre possíveis, os resultados alcançados mostram que a metodologia proposta é bastante promissora.

É importante ressaltar que uma boa filtragem dos dados está intimamente ligada à quantidade de dados históricos disponíveis para criação da base de dados (perfis típicos). De maneira semelhante, também é possível deduzir que quanto maior a qualidade dos dados brutos, mais eficaz será o tratamento executado pelo modelo.

Como sugestão de trabalhos futuros pode-se citar a aplicação da metodologia proposta, ou parte dela, no tratamento de registros anemométricos e de geração de energia eólica [42, 43, 44]. Outra possível aplicação é na identificação de subsequências atípicas de séries temporais [45], por exemplo, na análise de perturbações em sistemas elétricos, onde um dos interesses é a identificação de perturbações causadas por eventos nos componentes do sistema (desligamentos, faltas, religamentos, etc.) [46] e no tratamento de dados provenientes de unidades de medição fasorial (*Phasor Measurement Unit* – PMU) [47]. Adicionalmente, o Mapa de Kohonen e o classificador *Naive Bayes* também podem ser úteis na previsão de séries temporais [48], por exemplo, na construção de um previsor de carga semelhante ao algoritmo LBF (*Labeled Based Forecasting*) [49, 50] que explora as similaridades entre os perfis diários de carga [51].

6. REFERÊNCIAS

[1] G. Gross e F. Galiana, "Short-Term Load Forecasting," *Proceedings of IEEE,* vol. 75, pp. 1558-1573, Dez 1987.

[2] H. Chipp, Estrutura da Operação do Sistema Interligado Nacional, In: Nery, E. mercados e Regulação de Energia Elétrica, Editora Interciência, 2012.

[3] K. Liu, S. Subbarayan, R. Shoults, M. Manry, C. Kwan, F. Lewis e J. Naccarino, "Comparison of Very Short-Term Load Forecasting Techniques," *IEEE Transactions on Power Systems,* vol. 11, 1996.

[4] A. Lotufo e C. Minussi, Electric power systems load forecasting: a survey, Budapest: Powertech, 1999.

[5] H. K. Alfares e M. Nazeeruddin, "Electric Load Forecasting: Literature Survey And Classification Of Methods," *International Journal of Systems Science,* vol. 33, pp. 23-34, 2002.

[6] H. Hahn, S. Meyer-Nieberg e S. Pickl, "Electric load forecasting methods: Tools for decision making," *European Journal of Operational Research,* p. 902–907, 2009.

[7] J. Pessanha, T. Justino e M. Maceira, "Metodologia de tratamentos dos dados de carga elétrica," Rio de Janeiro,Cepel, 2011,Relatório Técnico 42907/2011.

[8] V. Hodge e J. Austin, "A survey of Outlier Detection Methodologies," *Artificial Intelligence Review 22,* pp. 85-126, 2004.

[9] F. Grubbs, "Procedures for Detecting Outlying Observations in Samples," *Technometrics,* vol. 11, pp. 1-21, 1969.

[10] W. Chauvenet, A Manual of Spherical and Practical Astronomy Vol II, 5th ed., 1960, pp. 474-566.

[11] M. S. Farinas, R. L. Sousa e R. C. Souza, "A Methodology to Filter Time Series: Application to Minute-By-Minute Electric Load Series," *Pesquisa Operacional,* vol. 24, pp. 355-371, 2004.

[12] J. Yang e J. Stenzel, "Historical load curve correction for short-term load forecasting," *7th International Power Engineering Conference,* 2005.

[13] C. Guirelli, "Previsão da carga de curto prazo de áreas elétricas através de técnicas de inteligência artificial," Tese de Doutorado, Escola Politécnica da Universidade

de São Paulo, São Paulo, 2006.

[14] Z. Xiaoxing e S. Caixin, "Dynamic intelligent cleaning model of dirty electric load data," *Energy Conversion and Management,* vol. 49, pp. 564-569, 2008.

[15] G. Grigoras, G. Cartina, E. Bobric e C. Barbulesc, "Missing data treatment of the load profiles in distribution networks," em *Bucharest Power Tech Conference*, Bucharest, 2009.

[16] C. Guan, P. Luh, M. Coolbeth, L. Michel, Y. Zhao, Y. Chen, C. J. Manville, P. Friedland e S. Rourke, "Very short-term load forecasting: multilevel wavelet neural networks with data pre-filtering," em *IEEE Power & Energy Society General Meeting*, Storrs, 2009.

[17] J. Chen, W. Li, A. Lau, J. Cao e K. Wang, "Automated load curve data cleansing in power systems," em *IEEE Transactions on Smart Grid*, 2010.

[18] Z. Qu, Y. Wang, C. Wang, N. Qu e J. Yan, "A Data Cleaning Model for Electric Power Big Data," vol. 121, pp. 405-411, 2016.

[19] G. Tang, K. Wu, J. Lei, Z. Bi e J. Tang, "From Landscape to Portrait: A New Approach for Outlier Detection in Load Curve Data," *IEEE Transactions on Smart Grid,* vol. 5, 2014.

[20] M. Gupta, J. Gao, C. Aggarwal e J. Han, "Outlier Detection for Temporal Data: A Survey," *IEEE Transactions on Knowledge and Data Engineering,* vol. 26, 2014.

[21] Z. Guo, W. Li, A. Lau, T. Inga-Rojas e K. Wang, "Detecting X-Outliers in load curve data in power systems," *IEEE Transations on Power Systems,* vol. 27, n. 2, pp. 875-844, May 2012.

[22] J. F. Pessanha, T. C. Justino e M. E. Maceira, "Metodologia para Filtragem de Registros de Carga," em *Xii Simpósio de Especialistas em Planejamento da Operação e Expansão Elétrica*, 2012.

[23] L. Oliveira, "Tratamento de dados de curvas de carga via análise de agrupamentos e transformada wavelet," Tese de de Doutorado, Rio de Janeiro, UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2013.

[24] J. Pessanha e V. Almeida, "Manual do usuário do programa Reger-filtro," Relatório Técnico 29550/2016, Centro de Pesquisas de Energia Elétrica, 2016.

[25] W. Martinez e A. Martinez, Computational Statistics Handbook with Matlab, 1 ed., Chapman and Hall/CRC, 2002.

[26] P. Tan, M. Steinbach e V. Kumar, Introdução ao Data Mining Mineração de Dados, Rio de Janeiro: Editora Ciência Moderna, 2009.

[27] J. Jang, C. Sun e E. Mizutani, Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence, Prentice Hall, 1997.

[28] I. Witten e E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed., Morgan Kaufmann, 2005.

[29] J. Foreman, Data Smart: using data science to transform information into insight, John Willey & Sons, 2013.

[30] T. Masters, Practical Neural Network Recipes in C++, Morgan Kaufmann Publishers, 1993.

[31] S. Haykin, Redes Neurais: Princípio e Práticas, Artmed, 1999.

[32] Z. Kovács, Redes Neurais Artificiais, Fundamentos e Aplicações, 2 ed., São Paulo: Edição Acadêmica, 1996.

[33] M. Ester, H. Kriegel, J. Sander e X. Xu, "A density based algorithm for discovering clusters in large spatial database with noise," em *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, Portland, Oregon, August, 1996.

[34] J. Pessanha, R. Velasquez, . A. Melo e R. Caldas, "Técnicas de Cluster Analysis na Cosntrução de Tipologias de Curvas de Carga," em *XV Seminário Nacional de Distribuição de Energia Elétrica*, Salvador, 2002.

[35] J. Pessanha e L. Laurencel, "Clustering Eletric Load Curves: The Brazilian experience," em *Workshop France-Brésil sur la fouille de données*, Recife, 2009.

[36] F. Louzada, C. Diniz, P. Ferreira e E. Ferreira, Controle Estatísticos de Processos - Uma abordagem prática para cursos de Engenharia e Administração, Rio de Janeiro: LTC, 2013.

[37] R. Johnson and D. Wichern, Applied multivariate statistical analysis, Pearson, 2002.

[38] D. Chantelou, G. Hébrail e C. Muller, "Visualizing 2665 electric power load curves on a single A4 sheet of paper," em *International Conference on Intelligent Systems Applications to Power Systems*, Orlando, 1996.

[39] H. N. Akouemo e R. J. Povinelli, "Time Series outlier detection and imputation," em *PES General Meeting Conference & Exposition IEEE*, 2014.

[40] E. Tsao, J. Bezdek e N. Pal, "Fuzzy Kohonen clustering networks," *Pattern Recognition,* vol. 27, pp. 757-764, 1994.

[41] S. Press, Applied Multivariate Analysis: using Bayesian and frequentist methods of inference, New York: Dover Publication Inc, 1982.

[42] J. Pessanha, V. Castellani, T. Justino, D. Jardim e M. Maceira, "Maceira Uma metodologia para tratamento de dados anenométricos," em *X Congresso Brasileiro de Inteligência Computacional*, Fortaleza, 2011.

[43] G. Xiaoli, Y. Ying, W. Ling, Q. Zhaoyang e W. Yongwen, "Wind data proprocessing algorithm based on extracting isolated points," *International Journal of Multimedia and Ubiquitous Engineering,* vol. 10, pp. 53-62, 2015.

[44] L. Zheng, W. Hu e Y. Min Raw, "Wind Data Preprocessing: A Data-Mining Approach," *IEEE Transactions on Sustainable Energy,* vol. 6, pp. 11-19, January 2015.

[45] E. Keogh, J. Lin, S. Lee e H. Van Herle, "Fiding the most unusual time series subsequences: algorithms and applications," *Knowledge and Information Systems,* vol. 11, pp. 1-27, 2006.

[46] J. Oliveira, S. Filho, M. Rodrigues, S. Diniz, A. Miranda, D. Medeiros e M. Batalha, "Análise de perturbações utilizando uma ferramenta WEB para tratamento de registros oscilográficos," em *XIII SEMINÁRIO TÉCNICO DE PROTEÇÃO E CONTROLE*, BRASÍLIA, 2016.

[47] R. Hafen, T. Gibson, K. Van Dam e T. Critchlow, "Power Grid Data Analytics with R and Hadoop, In: Y.Zhao, Y.Cen," em *Data Mining Applications with R*, Academic Press, 2014.

[48] S. Thomassey e M. Happiette, "A neural clustering and classification system for forecasting of new apparel items," *Applied Soft Computing 7*, pp. 1177-1187, 2007.

[49] F. Álvarez, A. Troncoso, J. Riquelme e J. Aguillar-Ruiz, "LBF: A Labeled-Based Forecasting Algorithm and its Application to Electricity Price Time Series," *8th IEEE International Conference on Data Mining*, December 2008.

[50] G. Hébrail, "Practical Data Mining in a Large Utility Company," *Quaderns d'Estadistica i Investigacio Operatia,* vol. 25, pp. 509-520, 2001.

[51] V. Almeida, J. Pessanha e A. T.M.L., "Uso combinado de redes neurais artificiais e lógica fuzzy na previsão de carga para programação diária da operação," em *XXIII Seminário Nacional de Produção e Transmissão de Energia Elétrica*, Foz do iguaçu, 2015.

APÊNDICE I