Pesquisa de Sistemas Elétricos de Energia 191 (2021) 106885

Listas de conteúdo disponíveis em ScienceDirect

Pesquisa de Sistemas Elétricos de Energia

página inicial do jornal: www.elsevier.com/locate/epsr

Revisão de métodos de pré-processamento para séries temporais voláteis univariadas em aplicações de sistemas de potência

Kumar Gaurav Ranjana , B Rajanarayan Prustyb,ÿ , Debashisha Jenac

[a] Instituto Nacional de Ciência e Tecnologia, Berhampur, Índia
[b] Instituto de Tecnologia Vellore, Vellore, Índia
[c] Instituto Nacional de Tecnologia Karnataka, Surathkal, Índia

INFORMAÇÕES DO ARTIGO

Palavras-chave:
Falso outlier
Detecção e correção de outliers
Pré-processando
Verdadeiro valor atípico
Dados voláteis

RESUMO

A detecção de outliers e a correção de séries temporais conhecidas como pré-processamento, desempenham um papel vital na previsão em sistemas de energia. Pesquisas rigorosas sobre esse tema foram feitas nas últimas décadas e ainda estão em andamento. Neste artigo, é feito um levantamento detalhado dos diferentes métodos de pré-processamento, e os métodos de pré-processamento existentes são categorizados. Além disso, a capacidade de pré-processamento de cada método é destacada. Os métodos bem estabelecidos de cada categoria aplicáveis a dados univariados são analisados criticamente e comparados com base em sua capacidade de pré-processamento. A análise dos resultados inclui a aplicação de métodos bem estabelecidos a séries temporais voláteis frequentemente utilizadas em aplicações de sistemas de potência. Geração fotovoltaica, potência de carga e séries temporais de temperatura ambiente (limpa e bruta) de diferentes etapas de tempo coletadas de vários locais/zonas meteorológicas são consideradas para comparação baseada em índice e baseada em gráficos entre os métodos bem estabelecidos. O impacto da mudança nos valores do(s) parâmetro(s) crucial(s) e resolução temporal dos dados no desempenho dos métodos também é elucidado neste artigo. Os prós e contras dos métodos são discutidos juntamente com o escopo para improvisação.

1. Introdução

Nos últimos anos, a integração de gerações renováveis com sistemas de energia elétrica tem aumentado o interesse devido aos benefícios ambientais e econômicos. A previsão de variáveis de entrada constitui a base para a previsão de regime permanente do sistema elétrico [1-3]. As entradas dos estudos acima incluem potência de carga, gerações renováveis, variáveis meteorológicas, como temperatura ambiente, velocidade do vento, etc. A geração de informações sintéticas de alta qualidade para tais entradas de séries temporais históricas do mundo real por meio de técnicas de pré-processamento de dados é imperativa na obtenção de maior precisão na previsão [3-7]. O pré-processamento de dados é seguido pela seleção de entrada para reduzir a incerteza de previsão [8-11]. Os algoritmos de aprendizado de máquina ganharam popularidade em vários campos, como a previsão de séries temporais, principalmente modelando as relações não lineares entre as variáveis. A série temporal dessas entradas sendo altamente volátil (volatilidade em uma série temporal refere-se ao fenômeno em que a variância condicional varia ao longo do tempo) e muitas vezes vulnerável a outliers apresenta vários desafios ao processo de pré-processamento. Um ponto de dados em uma série temporal é considerado um outlier se (a) for consideravelmente diferente dos pontos restantes no conjunto de dados (não seguir a tendência e o padrão periódico) [12], (b) estimular a suspeita, ou seja, faltam dados [13], (c) está consideravelmente longe de sua vizinhança [14], ou (d) aparece em uma região de baixa densidade [14]. O processo de detecção desses pontos de dados é conhecido como detecção de outliers ou detecção de anomalias. Se a detecção de discrepâncias for seguida por correção de discrepâncias, isso é chamado de pré-processamento de dados. O pré-processamento tem uma vasta área de aplicações e é extenuante listar todas elas.
No entanto, as áreas onde o uso do pré-processamento de dados é indispensável são destacadas abaixo:

a) *Previsão:* O pré- processamento aumenta a precisão dos modelos de previsão que usam dados históricos para prever potência de carga, geração fotovoltaica, velocidade do vento, temperatura ambiente, etc. para planejamento de expansão do sistema de energia, planejamento operacional e operação em tempo real [15]. b) *Redes de sensores:* Os dados coletados de vários sensores sem fio apresentam várias características únicas. Nesses casos, a detecção de outliers ajuda na detecção de sensores defeituosos ou eventos incomuns, como intrusão [14]. c) *Análise de cuidados de saúde e diagnóstico médico:* A detecção de padrões anormais nos dados de exames médicos dos pacientes ajuda no diagnóstico adequado da condição e permite que os médicos tomem as medidas adequadas [14]. d) *Detecção de fraude:* A detecção de outlier é necessária para detectar o uso fraudulento de cartões de crédito e outras informações críticas [16]. e) Detecção de *intrusão: A detecção* de outlier auxilia na detecção de intrusão, que se refere à detecção de atividades maliciosas (invasões, penetrações,

ÿ Autor correspondente.
E-mail: brprusty@ieee.org (BR Prusty).

https://doi.org/10.1016/j.epsr.2020.106885 Recebido
em 21 de março de 2020; Recebido em formulário revisado em 4 de outubro de 2020; Aceito em 8 de outubro de 2020
Disponível online em 23 de outubro de 2020

ou outras formas de abuso de computador) em um sistema relacionado a computadores de uma perspectiva de segurança [17]. f) *Logs de dados e logs de processos:* a detecção de discrepâncias orienta as empresas a obter informações ocultas em seus sites, o que reduz o custo e o esforço posteriormente. Técnicas automatizadas de detecção de outliers são usadas para detectar padrões incomuns em logs de grande volume [18]. g) *Processamento de imagem:* A detecção de outliers é necessária para detectar as alterações ou regiões anormais em uma imagem estática. Os domínios sob este incluem reconhecimento de dígitos, imagem mamográfica, imagens de satélite, vigilância por vídeo e espectroscopia [14].

Além das áreas acima, outras aplicações de detecção e correção de valores discrepantes incluem finanças, texto e mídia social, ciências da terra, internet das coisas (IoT), etc.

Pesquisas rigorosas sobre o pré-processamento de séries temporais voláteis foram feitas nas últimas seis décadas e ainda estão em andamento. Em 2005, foi feita a primeira tentativa de comparar os métodos de pré-processamento existentes [19].
Mas, até então, um número limitado de métodos de pré-processamento estava disponível. De 2009 a 2016, vários trabalhos elegantes foram produzidos para comparar os diferentes métodos de pré-processamento [20-27]. Embora muitos métodos tenham sido incluídos na revisão, poucos métodos importantes não foram completamente analisados ou foram ignorados [15,18,20,21], ou uma categoria particular de métodos de pré-processamento [21,22,24,27] foi enfocada. Em 2019, foi proposto um estudo meticuloso que categoriza um grande número de métodos de pré-processamento e destaca seus méritos e lacunas [14].
No entanto, poucos métodos foram excluídos da análise. A capacidade de pré-processamento de todos os métodos, comparação entre os métodos e aplicabilidade de diferentes métodos a diferentes dados não é elucidada. O impacto de valores de parâmetros significativos e o passo de tempo dos dados no desempenho dos métodos de pré-processamento não foram analisados. Todas as lacunas listadas acima são abordadas neste artigo.

As contribuições significativas deste artigo são as seguintes:

- Várias abordagens de detecção e correção de valores discrepantes são categorizadas, e a capacidade de pré-processamento de cada método é destacada.
- Métodos bem estabelecidos de cada categoria (aplicável apenas a dados univariados) são escolhidos para a análise, e seus algoritmos são explicados em detalhes.
- Os métodos bem estabelecidos são aplicados a várias séries temporais, e seus desempenhos são analisados e comparados criticamente.
- O impacto de valores de parâmetros significativos no desempenho dos métodos bem estabelecidos é elucidado.
- O efeito da resolução temporal dos dados no bem estabelecido

o desempenho dos métodos é elaborado.

Este artigo tem como objetivo visualizar as diferenças entre os métodos de pré-processamento de forma lúcida e orientar um pesquisador iniciante na escolha da melhor forma de pré-processamento de acordo com o tipo de dado e sua aplicação.

O artigo está organizado da seguinte forma: Na Seção 2, os métodos de pré-processamento existentes são categorizados. Métodos de pré-processamento bem estabelecidos são explicados na Seção 3, e a base para comparação desses métodos é discutida na Seção 4. Análise de resultados, incluindo uma comparação detalhada entre diferentes métodos em várias plataformas, méritos e lacunas de cada método e escopo para improvisação , está incluído na Seção 5. As observações finais são fornecidas na Seção 6.

2. Métodos de pré-processamento existentes

Neste artigo, métodos de pré-processamento explicitamente aplicáveis a séries temporais são considerados para análise. Existem também métodos de pré-processamento que se aplicam a dados que não são séries temporais [28]. Os métodos de detecção e correção de valores discrepantes existentes podem ser categorizados em quatro tipos principais, definidos como abaixo:

a) *Métodos baseados em estatística:* Nesses tipos de métodos, os pontos de dados são modelados usando uma distribuição estocástica e os outliers são detectados com base na relação com o modelo de distribuição. Sob a suposição de que a distribuição se ajusta ao conjunto de dados, os outliers são aqueles pontos que não concordam ou não estão em conformidade com o modelo subjacente para os dados.

i) Métodos paramétricos: Esses métodos baseados em estatísticas assumem que as séries temporais seguem um modelo de distribuição paramétrica específico ou exibem algumas características. ii) Métodos não paramétricos: Esses tipos de métodos baseados em estatística não assumem que as séries temporais seguem um modelo de distribuição paramétrico específico ou exibem quaisquer características.

b) *Métodos baseados em densidade:* Os métodos de detecção de valores discrepantes baseados em densidade trabalham com o princípio central de que um valor discrepante pode ser encontrado em uma região de baixa densidade. Simultaneamente, assume-se que os não-outliers (também conhecidos como inliers/pontos de dados genuínos) aparecem em bairros densos.
Eles comparam as densidades dos pontos locais com as densidades de seus vizinhos locais. Os métodos baseados em densidade usam mecanismos sofisticados para modelar a discrepância (grau de ser discrepante) dos pontos de dados.
Normalmente, envolve a investigação das densidades locais do ponto em estudo e seus vizinhos mais próximos. A métrica de outlierness de um ponto de dados é normalmente uma proporção de sua densidade em relação às densidades médias de seus vizinhos mais próximos, tornando-a relativa. Os métodos baseados em densidade mostram uma capacidade de modelagem mais robusta, mas, ao mesmo tempo, exigem computação cara.

c) *Métodos baseados em distância:* Esses métodos detectam outliers pelo cálculo das distâncias entre pontos, onde os pontos de dados distantes de seu vizinho mais próximo são considerados outliers. Os out liers baseados em distância também podem ser definidos, dada a medida de distância do espaço de características, como: i) Pontos situados dentro da distância *d* , mas com menos de *p* diferente amostras.
ii) Os *n* pontos mais distantes do k- ésimo vizinho mais próximo. iii) Os *n* principais pontos com a maior distância média aos *k* vizinhos mais próximos.

As abordagens de detecção de valores discrepantes baseados em distância são moderadas e dimensionam bem para grandes conjuntos de dados com dimensionalidade média a alta. Eles são flexíveis, computacionalmente eficientes e tendem a ter uma base robusta.

a) *Métodos baseados em* cluster: As técnicas baseadas em cluster geralmente contam com o uso de métodos de cluster para descrever o comportamento dos dados. Para isso, os pequenos clusters com significativamente menos pontos de dados do que outros são rotulados como outliers.

Os métodos de pré-processamento existentes são categorizados e mostrados cronologicamente na Tabela 1. Sua capacidade de correção de outliers é mencionada com ela. Além disso, a tabela também contém informações sobre quaisquer improvisações disponíveis nesses métodos e a capacidade de correção de valores discrepantes em suas versões improvisadas. Vale a pena notar que os métodos baseados em estatísticas e baseados em distância se aplicam a séries temporais univariadas. Em contraste, métodos baseados em densidade e baseados em cluster são apropriados apenas para séries temporais multivariadas. As vantagens e desvantagens dos métodos de pré-processamento existentes estão listadas na Tabela 2.

3. Métodos de pré-processamento bem estabelecidos

Em meio aos inúmeros métodos de pré-processamento, as técnicas bem estabelecidas de cada categoria são analisadas em detalhes. Os métodos baseados em conjunto de dados SWP e retrato são métodos estatísticos paramétricos para pré-processamento. Eles são aplicados em vários artigos devido às suas capacidades de pré-processamento decentes [34-36,38]. O método baseado em SWP é amplamente utilizado devido à sua capacidade de rastrear a tendência e a sazonalidade das séries temporais melhor do que outros algoritmos. O método baseado em conjunto de dados de retrato oferece uma nova maneira de visualizar dados, retratando os dados ocultos

Tabela 1

Classificação dos métodos de pré-processamento existentes.

Categoria	Ano Método original	Versão improvisada do OCC		OCC
Paramétrico baseado em estatística	1989 baseado em ARMA [29,30]	Não	–	–
	2005 baseado no teorema de Chebyshev [31]	Não	Agregação simbólica integrada baseada em aproximação [32]	Sim
	2008 baseado em ARMASel [33]	Sim	–	–
	2010 Com base na previsão de janela deslizante (SWP) [34-36]	Sim	–	–
	2010 Baseado em estimativa de densidade do kernel [37]	Não	–	–
	Baseado em conjunto de dados 2014 Portrait [38]	Sim	–	–
	2014 baseado em ARIMAX [39]	Sim	–	–
	2016 Com base em características atípicas [40]	Sim	–	–
	2016 baseado em classificador Bayesiano [41]	Não	–	–
	Detector Bernoulli 2016 [42]	Não	–	–
	2017 baseado em modelo ARX e ANN [43]	Sim	–	–
	Baseado em estimador M não paramétrico de 1964 Huber-skip [44,45]	Não	–	–
	2005 Baseado na diferença de segunda ordem [46]	Sim	Agregação simbólica integrada baseada em aproximação [32] Sim	
	2010 B-spline com base em suavização [47]	Sim	–	–
	2010 Esqueleto baseado em pontos [48]	Não	–	–
	2011 baseado em suavização do kernel [49]	Sim	–	–
	2011 Robusto NP-SC [50]	Sim	–	–
	2011 LS-SV [51]	Não	–	–
	2014 Baseado em progressão aritmética [52]	Não	–	–
	2018 baseado em DRM [53]	Sim	–	–
	2018 baseado na janela de tempo [54]	Sim	–	–
	2019 Autoencoder profundo probabilístico [55]	Sim	–	–
Baseado em densidade	2000 Fator de outlier local (LOF) [3,56]	Nenhum fator de outlier baseado em conectivo (COF) [57]		Não
			LOcal Correlação Integral (LOCI) baseado [58]	Não
			Outlierness influenciado (INFLO) [59]	Não
			Computação LOF distribuída (DLC) [60]	Não
	2004 Fator de densidade relativa (RDF) [61]	Não	–	–
	2007 Baseado em funções de densidade do kernel [62]	Sem RKOF [63]		Não
			Baseado em estimativa de kernel ponderado [64]	Não
			Baseado em KDE [65]	Não
			Baseado em densidade de kernel adaptável [66]	Não
			QUELOS [67]	Não
	2009 Probabilidades discrepantes locais (LoOP) [68]	Não	–	–
	2009 Fator outlier baseado em resolução (ROF) [69]	Não	Fator de outlier de janela dinâmica (DWOF) [70]	Não
	Pontuação Outlier Baseada em Densidade Relativa de 2017 (RDOS) [71] Não		–	–
Baseado em distância	Métodos de Grubbs de 2011 [72]	Nenhum método de Grubbs modificado [73]		Não
	2014 Com base nos vizinhos mais próximos [74]	Não baseado em kNN [75]		Não
			kNN adaptado [76]	Não
	2015 LDOF [77]	Não	–	–
Baseado em cluster	2007 D-stream [78] 2008 k-	Não	–	–
	means clustering baseado [79,80]	Sem kNN-SVM [81]		Não
			baseado em k-medoids [82]	Não
			Baseado em kNN inspirado em MST [83]	Não
			kNN-FSVM [84]	Não
	2000 Floresta de isolamento [85-87]	Não Floresta de isolamento estendida [88] Floresta de		Não
			isolamento baseada em k-means [89]	Não
			SSO-IF [90]	Não
	2009 SDStream [91]	Não	–	–
	2010 Algoritmo de detecção de outlier inteligente (IODA) [92] Não		–	–
	2012 Atributos de ponderação em clusters [93]	Não	–	–

Nota: OCC significa "capacidade de correção de outlier".

características como tendência e sazonalidade, potencializando o desempenho da abordagem. O método baseado em suavização B-spline [47] é um método estatístico não paramétrico usado por muitos pesquisadores como referência para comparar seus métodos. Essa abordagem pode pré-processar qualquer série temporal volátil sem fazer qualquer suposição sobre a distribuição ou periodicidade dos dados. O método baseado em k-nearest neighbor (kNN) [75] é um dos métodos baseados em distância mais amplamente utilizados para detecção de outliers, devido à sua facilidade de aplicação a vários tipos de dados sem conhecer a distribuição ou características dos dados. Além dos méritos mencionados acima, selecionar esses valores de parâmetro é menos tedioso do que outros métodos.

Os métodos de pré-processamento baseado em SWP, baseado em dados de retrato e baseado em suavização B-spline e o método de detecção de outlier baseado em kNN são elaborados respectivamente nas Seções 3.1, 3.2, 3.3 e 3.4. As abordagens acima são explicadas considerando uma série temporal uni , variável $X = \{x_1, x_2, \ldots, x_n\}$ de comprimento n.

3.1. Abordagem baseada em previsão de janela deslizante

É um método baseado em previsão que usa o vizinho mais próximo capa de dados [34-36]. Considerando uma série temporal X, o kÿ o mais próximo bairro $n_i^{k'}$ de largura *2k'* de um ponto de dados x_i em um instante de tempo *i* é definido como,

$$n_i^{k'} = \{x_{i-2k'}, x_{i-2k'+1}, \ldots, x_{i-1}\}, \quad (1)$$

onde $n_i^{k'}$ é uma janela unilateral que considera *2k'* pontos de dados antes de x_i.

Uma janela de ambos os lados que considera kÿ pontos de dados antes de x_i e kÿ pontos de dados após x_i é dado como

$$n_i^{k'} = \{x_{i-k'}, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{i+k'}\}. \quad (2)$$

Em geral, a janela de um lado é preferível à janela de ambos os lados no caso de detecção de valores discrepantes, pois os pontos de dados k' após x_i podem ter valores discrepantes não detectados e não corrigidos. O valor previsto de x_i, ou seja, x_i', com base na vizinhança mais próxima unilateral $n_i^{k'}$ é definido como

Tabela 2

Vantagens e desvantagens dos métodos de pré-processamento existentes.

Paramétrico baseado em estatística	Vantagens	• Matematicamente admissível e com alta velocidade de desempenho após a construção do modelo. • Eficiente e é possível interpretar os outliers em sentido real. • A maioria dos métodos é capaz de dar uma abordagem para corrigir outliers.
	Desvantagens	• O conhecimento sobre a distribuição dos dados é necessário a priori. Os dados podem não se encaixar precisamente no pressuposto distribuição, o que resultará em modelagem inadequada e degradação do desempenho. • Normalmente, não aplicável a dados multivariados do ponto de vista de custo computacional. • Além de todas as vantagens dos
Vantagens não paramétricas		métodos paramétricos, eles não fazem nenhuma suposição sobre a distribuição dos dados. Portanto, nenhuma informação prévia sobre a distribuição dos dados é necessária para modelagem e pré-processamento.
	Desvantagens	• Não é adequado para conjuntos de dados multivariados e grandes, pois envolvem cálculos intensivos.
Baseado em densidade	Vantagens	• Não dependa da distribuição assumida dos dados para calcular a estimativa de densidade, que é usada posteriormente para detectar discrepâncias. • Capaz de ter um bom desempenho em dados multivariados e detectar discrepâncias próximas a regiões densas. • Requer muito menos otimização de parâmetros para um desempenho ideal.
	Desvantagens	• Métodos complicados e computacionalmente intensivos. • Não aplicável a dados univariados. • Nenhuma abordagem para corrigir outliers.
Baseado em distância	Vantagens	• Simples de entender, bem como de compreender. • Não dependa da distribuição assumida para ajustar os dados.
	Desvantagens	• Incapaz de corrigir outliers. • Não aplicável a dados multivariados. • Não é preferível para o pré-processamento de grandes conjuntos de
Baseado em cluster	Vantagens	dados. • Independente da distribuição dos dados. • Capaz de detectar outliers em dados multivariados.
	Desvantagens	• Não aplicável a dados univariados, nem capaz de corrigir outliers. • Não é possível detectar discrepâncias que ocorrem no cluster, pois é contra sua suposição. • Os métodos são de natureza binária; portanto, não fornece nenhuma indicação quantitativa da discrepância de um ponto de dados.

$$x'_{eu} = \frac{\sum_{kj=1}^{2k'} w_{ij} x_{ij}}{\sum_{kj=1}^{2k'} w_{ij}}, \tag{3}$$

onde w_{ij} é o peso do ponto de dados x_{ij}, que é inversamente proporcional à distância entre os pontos xi e x_{ij}, que é dado como

$$w_{ij} = \frac{1}{|x_i - x_{ij}|}. \tag{4}$$

Quanto maior a distância entre os pontos de dados *xi* e x_{ij}, menor é o peso w_{ij}, indicando que a dependência de xion x_{ij} é insignificante e vice-versa. Com base na previsão em (3), o intervalo de confiança previsto (PCI) para *xi* é dado como,

$$\text{PCI} = x_i' \pm \left(t_{\alpha/2,2k'-1} \times s \sqrt{1 + \frac{1}{2k'}} \right), \tag{5}$$

onde $t_{\alpha/2,2k'-1}$ é 100 (1 − α)% do valor crítico do percentil para o distribuição t de Student com $2k' - 1$ graus de liberdade e *s* é o desvio padrão da amostra de vizinhança $n_i^{k'}$.

Se *xi* não estiver dentro do PCI, ele será tratado como um valor discrepante. O ponto de dados corrompido *xi* é substituído pelos dados corrigidos, que é o valor previsto x_i' usando (3). Finalmente, a janela é atualizada para o próximo ponto de dados. Este processo de detecção e correção continua até que o último ponto de dados da série temporal seja verificado. O fluxograma para pré-processamento usando SWP é dado na Fig. 1.

Begin: Input the given univariate time-series X and select a window width $2k'$.
for $i = 2k'+1$ until $i = n$
do
- Compute the k'^{th} nearest neighborhood $n_i^{k'}$ using (1).
- Compute the predicted value x_i' of x_i using (3).
- Compute the PCI for x_i using (5).

if x_i is not within the PCI
Replace the data point x_i with its predicted value x_i'.
end if
end for

Fig. 1. Algoritmo para pré-processamento baseado em SWP.

3.2. Abordagem baseada em conjunto de dados do Portrait

O método [38] aplica uma nova técnica de visualização de dados, denominada retrato, nos dados, analisando os padrões periódicos neles e reorganiza os dados para facilitar a análise. Os dados de retrato são compostos de pontos de dados que se enquadram nos intervalos de tempo correspondentes, ou seja, o conjunto de dados de retrato básico (BPD) *pj* é construído como

$$p_{xij}(T_{ij}) = x_{ij} + j, \quad T_{ij} = 1, 2, \cdots N, \quad j = 0, 1, \cdots N - 1, \tag{6}$$

onde *T* é o período fundamental que é calculado usando Fast Fourier Transform (FFT).

O vetor característico do conjunto de dados de retrato básico *pj* é definido como

$$[v_{ij}(\theta), j], \tag{7}$$

onde $\bar{y}_j$ e M_j representam a mediana e o desvio absoluto mediano (MAD) de *pj*, respectivamente.

A semelhança entre dois BPDs *pj*, $p_{j'}$ com vetores característicos v_j, j', respectivamente, é definido como

$$s_{jj'} = \begin{cases} \infty, & v_j = v_{j'} \\ 1/\|v_j - v_{j'}\|_2, & \text{por outro lado.} \end{cases} \tag{8}$$

Vários BPDs com vetores de características semelhantes são mesclados em um conjunto de dados de retrato virtual (VPD) para acelerar o pré-processamento. Depois que os dados de paisagem são convertidos em conjuntos de dados de retrato, a região atípica de cada conjunto de dados de retrato pode ser calculada de acordo com o tipo de distribuição que o conjunto de dados segue e seu comprimento. Como não é aconselhável realizar testes estatísticos em dados, que são afetados por outliers, para encontrar a distribuição, vários casos potenciais para detecção de outliers são considerados.

Caso 1: Para um conjunto de dados de retrato seguindo a distribuição normal $N(\bar{y}, \ddot{y}2)$ e para qualquer coeficiente de confiança ÿ, $0 < \ddot{y} < 1$, a região ÿ-outlier é dada como

$$\text{fora}_{\alpha}(\bar{y}, \ddot{y}^2) = \left\{ x : |x - \bar{y}| > z_{1-\frac{\alpha}{2}} \ddot{y} \right\}, \tag{9}$$

onde $\bar{y}$ e $\ddot{y}$ são a média e o desvio padrão do conjunto de dados de retrato, respectivamente e $z_{1-\alpha/2}$ é o percentil $100(1/2)$ de $N(0, 1)$. Como $\bar{y}$ e $\ddot{y}$ são sensíveis a outliers, a mediana e o MAD são usados em vez disso

de média e desvio padrão.

Caso 2: Para um conjunto de dados de retrato seguindo a distribuição gama *G(ÿ, ÿ)* com parâmetro de forma *ÿ* e parâmetro de escala ÿ, sua região ÿ-outlier é dada como

$$\text{fora}(\beta,\phi) = \left\{ x \in F : x < F^{-1}(\beta,\phi) \text{ ou } x > F^{-1}_{1-\frac{\alpha}{2}}(\beta,\phi) \right\} \quad (10)$$

onde F^{-1} é a função de distribuição cumulativa inversa de *G(ÿ, ÿ)*, e $F^{-1}_{1-\alpha/2}$ é o 100 (1 × − α/ 2) percentil de *G (ÿ, ÿ)*.

Caso 3: Quando o conjunto de dados é pequeno em comprimento, o box plot é usado para definir a região discrepante e é dado como

$$\text{fora}(Q_1, Q_3) = \{x \in Q : x < Q_1 - k \times \text{IQR ou } x > Q_3 + k \times \text{IQR}\}, \quad (11)$$

onde *ÿ* é um índice de significância, Q1 e Q3 são os percentis 25 e 75 , respectivamente, e a diferença (intervalo quartil $Q_3 - Q_1$) é chamado de inter

Um ponto de dados situado na região atípica do conjunto de dados de retrato correspondente é tratado como um atípico. Em seguida, o outlier é substituído por um valor aceitável do conjunto de dados correspondente, ou seja, o valor médio para os Casos 1 e 3 e o valor de ÿ/ÿ para o Caso 2. O fluxograma para pré-processamento baseado em dados de retrato é dado na Fig. 2.

3.3. Abordagem baseada em suavização de B-spline

O método [47] é baseado na modelagem de padrões inerentes de uma série temporal para detectar a presença de outliers. Os dados podem ser modelados como

$$X_m = m(t) + \varepsilon, \quad (12)$$

onde *m(t)* é a função subjacente e ÿ é o termo de erro que se assume ser normalmente e independentemente distribuído com média zero e variância constante ÿ2

.

A principal tarefa é encontrar uma estimativa adequada da função *m(t),* ou seja, $\hat{m}(t)$, usando os dados coletados. Com base na estimativa de *m(t)*, intervalo de confiança para cada ponto de dados é calculado. Os pontos de dados fora do intervalo de confiança são marcados como outliers e são substituídos pelo valor estimado $\hat{m}(t)_i$.

Para estimar *m(t),* a suavização de B-spline usa um sistema de funções de base que consiste em um conjunto de funções de base conhecidas e

Begin: Input the given univariate time-series X .
- Calculate the fundamental period T using FFT.
- Construct the BPDs using (6).
- Construct the characteristic vector v_j for each BPD using (7).
- Construct the similarity matrix s' using (8).
- Merge the similar BPDs into minimum number of VPDs.

for $i = 1$ until $i =$ total number of VPDs
do
- Calculate the outlier region for the i^{th} VPD using (9), (10) or (11).

for $j = 1$ until $j =$ length of i^{th} VPD
do
if j^{th} data point lies in the outlier region
Replace the data point with an acceptable value of i^{th} VPD
end if
end for
end for

Fig. 2. Algoritmo para pré-processamento baseado em dados de retrato.

Begin: Select a time-series X .
- Construct the basis function vector $\vec{\phi}$.
- Construct Φ using (14).
- Calculate the coefficient vector $\hat{c}$ using (15).
- Calculate the fitted value vector $\hat{X}$ using (16).
- Calculate the degrees of freedom df using (18).
- Calculate the square of predicted error $\left(\hat{E}_i\right)^2$ using (19), (20) and (21).

for $i = 1$ until $i = n$
do
- Select a data point x_i .
- Calculate the confidence interval for x_i using (22).

if x_i is not within the confidence interval
Replace x_i with its fitted value $\hat{x}_i$.
end if
end for

Fig. 3. Algoritmo para pré-processamento baseado em suavização de B-spline.

Begin: Select a time-series X .
- Prepare the embedding matrix X_{em} as in (23).
- Calculate the SED of each window from all other windows using (24), except for near-in-time window(s).
- Assign smallest SED of a window as the outlier index of that window.
- Calculate the threshold value of outlier index.

for $e = 1$ until $e = n - L' + 1$
do
- Select a window r_e .
- Compare its outlier index OI_e with the threshold value OI^{α} .

if $OI_e > OI^{\alpha}$
Mark all in data points in the r_e as outliers.
end if
end for

Fig. 4. Algoritmo para detecção de valores discrepantes baseados em kNN.

é expresso em forma vetorial como

$$\hat{m}(t) = c^{\rightarrow} \cdot \vec{\varphi}, \quad (13)$$

onde $c^{\rightarrow} = (c_1, \ldots, c_b)$ é o vetor coeficiente, $\vec{\varphi} = (\varphi_1, \varphi_2, \ldots, \varphi_b)$ é o vetor de funções de base *eb* denota o número de funções de base consideradas. Para estimar os coeficientes *c* a partir dos dados, uma matriz *n* × *b* é definida como

$$\Phi = \begin{bmatrix} \varphi_1(t_1) & \varphi_2(t_1) & \cdots & \varphi_b(t_1) \\ \varphi_1(t_2) & \varphi_2(t_2) & \cdots & \varphi_b(t_2) \\ \vdots & \vdots & & \vdots \\ \varphi_1(t_n) & \varphi_2(t_n) & \cdots & \varphi_b(t_n) \end{bmatrix}, \quad (14)$$

Tabela

3 Significado da(s) função(ões)/parâmetro(s) crucial(s) de diferentes métodos.

Método	Função(ões)/parâmetro(s) crucial(is)	Significado	
Baseado em SWP	Largura da janela *(2kÿ)*	Ele decide o número de pontos de dados de entrada fornecidos ao modelo para o pré-processamento de valores discrepantes. A largura inadequada da janela levará a uma previsão abaixo/superior.	
Baseado em dados de retrato	Período de tempo fundamental *(T)*	Ele orienta a construção de conjuntos de dados de retratos primários. O período impreciso resultará em conjuntos de dados de retrato instáveis.	
B-spline com base em suavização	Vetor de função de base (( )t) $\vec{\phi}$	É a alma do método. Do ajuste de curva ao pré-processamento de valores discrepantes, depende de quão bem a(s) função(ões) base(s) é(em) escolhida(s). Todo o resto é derivado dele.	
baseado em kNN	Comprimento da janela (Lÿ)	Ele decide a vizinhança de um ponto de dados. A semelhança entre as janelas aumenta com o comprimento da janela.	Ambos são usados para decidir o valor limite do índice outlier.

Tabela 4

Descrição abrangente das informações de tempo e amostras de dados de vários casos em consideração.

Caso não.	Dados	Local/zona meteorológica	Notação	Tipo de dados	Informações de tempo	Passo de tempo	Tamanho da amostra
1	geração fotovoltaica	Parkesburg, EUA	fotovoltaica	Cru	Meio-dia, 2012–2013	Diário	730
2	Potência de carga	Anchorage, EUA	PD1	Limpar \ limpo	Janeiro de 2004	De hora em hora	744
3		Centro-Norte do Texas, EUA	PD2	Cru	Janeiro de	De hora em hora	744
4			PD3	Cru	2012 10h, 2012-2013	Diário	730
5	Temperatura ambiente	Berhampur, Índia	TAMB1	Cru	Janeiro de 2010	De hora em hora	744

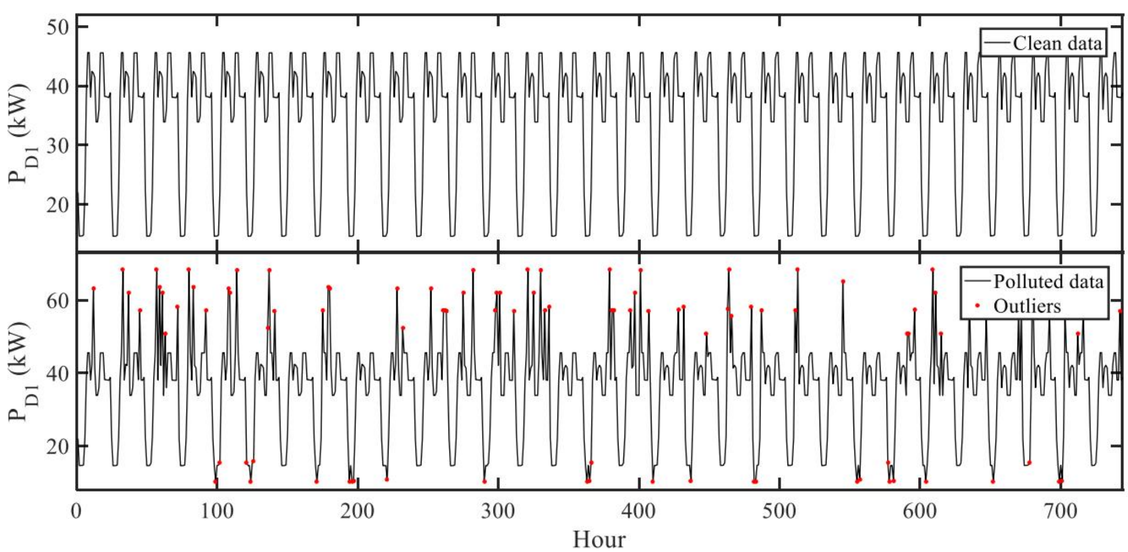

Fig. 5. Dados de potência de carga univariada limpos e poluídos.

onde $[\Phi]_{ij} = \phi_j(t_i)$ representa o valor do *j* o função de base no tempo de.

A estimativa do vetor coeficiente *c* é obtida como

$$\hat{c} = (\Phi^T\Phi + \lambda R)^{-1}\Phi^T X, \quad (15)$$

onde ÿ é o parâmetro de suavização, $R = \int D^2 \vec{\phi}(t) D^2 \vec{\phi}^T(t)\, dt$, e $D^2(.)$ é o operador derivativo de segunda ordem.

O vetor de valor ajustado $\hat{X}$ é calculado por

$$\hat{X} = \Phi\hat{c}, \quad (16)$$

ou

$$\hat{X} = SX, \quad (17)$$

onde $S = \Phi(\Phi^T\Phi + \lambda R)^{-1}\Phi^T$. É chamada de matriz hat, pois converte o vetor variável dependente *X* em seu valor ajustado $\hat{X}$.

O número de graus de liberdade do ajuste é o traço da matriz do chapéu

$$df = \text{traço}(S), \quad (18)$$

onde trace(.) calcula a soma dos elementos diagonais de uma matriz.

O quadrado do erro previsto estimado é dado por

$$\widehat{EP}^2 = \text{MSE} + (\widehat{E}(\hat{x}(t_i)))^2, \quad (19)$$

onde MSE é o erro quadrático médio calculado como

$$\text{MSE} = \frac{1}{n - df}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2 \quad (20)$$

e $(\widehat{E}(\hat{x}(t_i)))^2$ é a variância amostral para o ajuste no tempo t_i e é dado por elementos diagonais da matriz Var[$\hat{X}$]

$$\text{Var}[\hat{X}] = \text{MSE} \times S \times S^T. \quad (21)$$

Para um determinado nível de significância ÿ, a confiança de 100 (1 − α)% no intervalo no tempo *ti* é calculada como

$$\left[\hat{x}_i - Z_{1-\alpha/2} \times \widehat{EP},\ \hat{x}_i + Z_{1-\alpha/2} \times \widehat{EP}\right], \quad (22)$$

onde $Z_{1-\alpha/2}$ é o 100 (1 − α/2) percentil do normal padrão distribuição.

Qualquer ponto de dados fora do intervalo de confiança é tratado como um outlier. O fluxograma para pré-processamento de dados usando suavização de B-spline

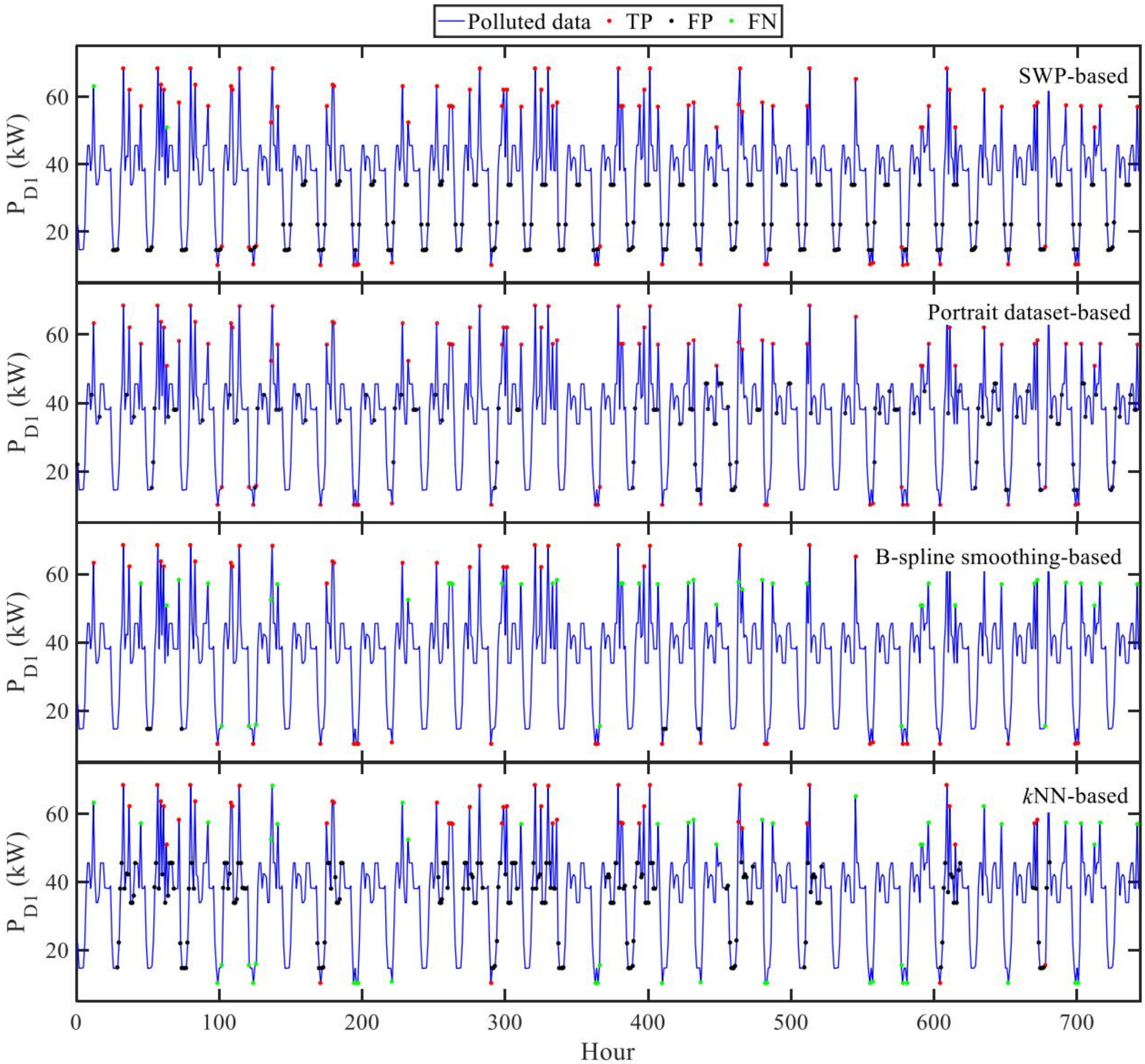

Fig. 6. Desempenho de detecção de outlier quando aplicado a dados poluídos univariados.

Tabela 5

Comparação de desempenho de detecção de outlier quando aplicado a dados poluídos univariados.

Índices	Conjunto de dados Portrait baseado em SWP	Sediada	B-spline alisamento baseado	baseado em kNN
Nº de TP	98	100	56	50
Nº de FP	208	133	7	187
Nº de FN	2	0	44	50
P	0,3203	0,4292	0,8889	0,2110
R	0,98	1	0,56	0,5
F-medida	0,4828	0,6006	0,6871	0,2967

é como mostrado na Fig. 3.

3.4. abordagem baseada no vizinho k-mais próximo

O método segrega toda a série temporal em vários grupos ou janelas [75]. A série temporal *X* pode ser dividida em janelas como

$$X_{em} = \begin{bmatrix} x_1 & x_2 & \cdots & x_{L} \\ x_2 & x_3 & \cdots & x_{L+1} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n-L+1} & x_{n-L+2} & \cdots & x_n \end{bmatrix}, \quad (23)$$

onde *Xem* é a matriz de incorporação, a linha *re* denota o *e* º janela e *L* denota o número de pontos de dados em cada janela.

Cada janela do *Xem* é então comparada com as outras janelas, usando o quadrado da distância euclidiana (SED) da seguinte forma:

$$d_r^2(e, f) = \sum_{j=1}^{L} \left(x_{e+j-1} - x_{f+j-1}\right)^2 \quad (24)$$

Assim, um valor de índice *discrepante OIe* para o *e* º a janela *re* é determinada, que é o k- ésimo menor SED entre *re* e todas as outras janelas, exceto para a(s) janela(s) próxima(s) de *re.* As janelas near-in-time de *re* são aquelas que têm pelo menos um ponto de dados em comum com *re.* Uma vez que cada janela atinge seu valor de índice de outlier, é necessário determinar um valor limite para detecção de outlier com base na ob

sequência de índices outliers. O valor limite do índice outlier *Olÿ* é tomado como ÿº valor mais alto de $\{OI_e\}_{e=1}^{n-L+1}$, onde ÿ é o inteiro mais próximo de $\times(n-L+1)$. Janelas com índice de outlier superior ao valor limite são consideradas como outliers. Apesar das proficiências dos métodos discutidos acima, a precisão de cada método varia muito com o valor de seu(s) parâmetro(s) crucial(is), portanto, a otimização do(s) parâmetro(s) crucial(is) é de extrema importância. O fluxograma para detecção de valores discrepantes baseados em kNN é dado na Fig. 4.

A precisão de cada um dos métodos explicados acima depende de parâmetros cruciais específicos, cujo significado é elucidado na Tabela 3.

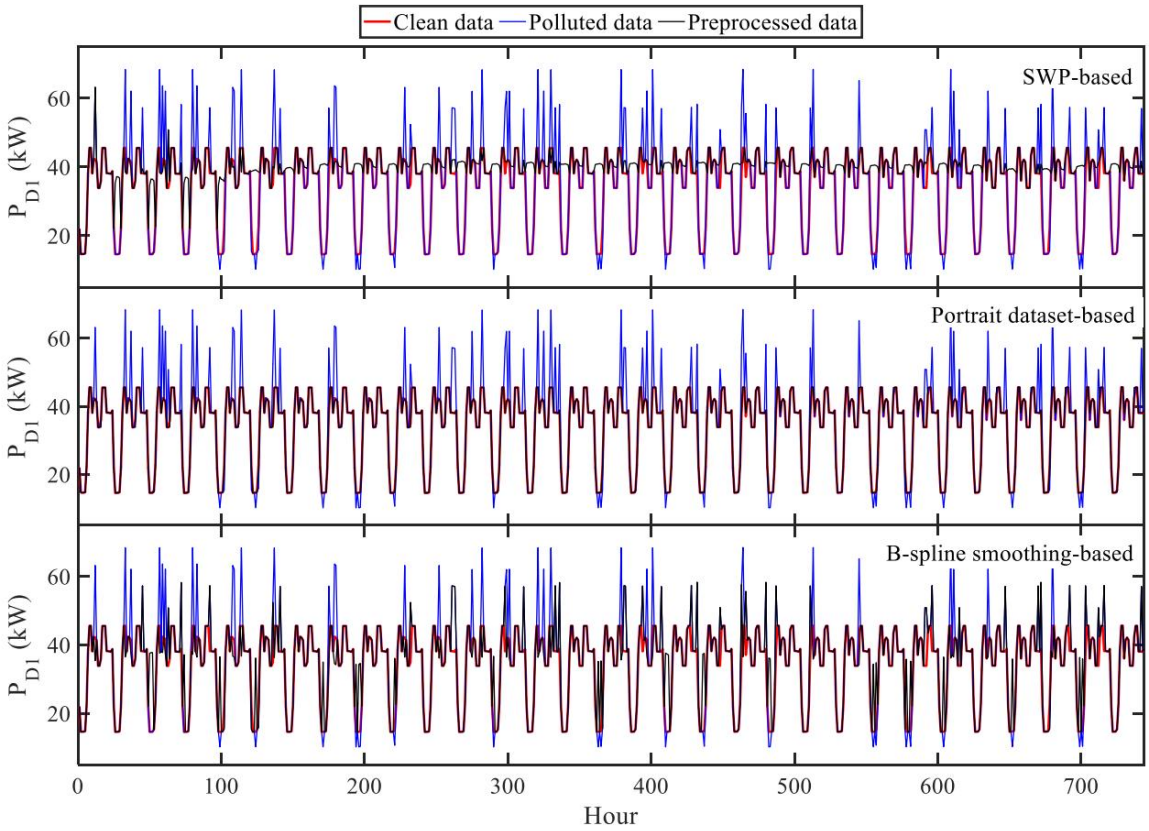

Fig. 7. Comparação de desempenho de correção de outlier quando aplicada a dados poluídos.

Além dos parâmetros interpretados na tabela, o desempenho dos métodos seletivos bem estabelecidos e aplicáveis apenas a dados univariados é afetado pelo coeficiente de confiança. O coeficiente de confiança influencia parcialmente o intervalo de confiança, que é o intervalo no qual um ponto de dados é tratado como normal/genuíno.

4. Base para comparação de métodos de pré-processamento

Os métodos de pré-processamento existentes na literatura diferem entre si em suas capacidades de pré-processamento. Portanto, a base para comparar esses métodos deve ser diferente. Neste artigo, medidas separadas são discutidas para comparar as capacidades de detecção e correção de valores discrepantes dos diferentes métodos de pré-processamento. Duas suposições de tipo de dados distintas são consideradas: dados limpos e dados brutos.
O primeiro assume que o conjunto de dados está livre de outliers, enquanto o segundo, na maioria dos casos, não faz tal suposição.

4.1. Base para comparação da capacidade de detecção de valores discrepantes dos métodos

4.1.1. Para dados limpos

Este artigo segue o padrão geral para comparação entre métodos baseados em sua capacidade de detecção de valores discrepantes, como frequentemente usado por muitos autores [38,47]. Primeiramente, para este tipo de dados, os outliers são introduzidos manualmente e, em seguida, o método é aplicado aos dados e seu desempenho é analisado. Três indicadores usados são definidos em:

a) Verdadeiro positivo (TP), ou seja, os pontos de dados identificados corretamente como out liers pelo método. b) Falso positivo (FP), ou seja, os pontos de dados genuínos marcados como outliers pelo método. c) Falso negativo (FN), ou seja, os outliers que não são identificados pelo

método.

Com base nesses indicadores, são definidas três bases de comparação:

1) Precisão (P): É a porcentagem de outliers detectados corretamente, sobre o número total de pontos de dados marcados como outliers pelo método.
   Quanto maior o valor P de um método, melhor é seu desempenho em relação a outros métodos.
2) Recall (R): É o percentual de outliers detectados corretamente pelo método em relação ao número total de outliers presentes no conjunto de dados. O método com o maior valor de recall é considerado o melhor.
3) F-measure (F): É a média harmônica de precisão e recall. Sua significância está na comparação de métodos quando dois métodos apresentam valores P e R contraditórios. Por exemplo, sejam 'A' e 'B' os dois métodos de pré-processamento. 'A' tem maior precisão do que 'B', enquanto 'B' tem maior recall do que 'A.' Neste caso, o método com maior F-measure é considerado melhor.

4.1.2. Para dados brutos

Em um cenário prático, o conhecimento prévio sobre os outliers não está disponível em sua maioria. Portanto, quando aplicadas a dados limpos, as medidas usadas para comparar os métodos não podem ser aplicadas para comparar os métodos quando aplicadas a dados brutos. Portanto, uma comparação de métodos baseada em índices (como na Seção 4.1.1) não é possível. Além disso, esses índices fornecem apenas uma comparação numérica entre os métodos e, em última análise, falham em destacar o motivo da falha ou do sucesso de um método. O conhecimento dessas razões pode orientar um pesquisador iniciante na seleção de um método adequado. Portanto, quando aplicado a esses dados, a eficácia de um método é julgada apenas pela análise de sua abordagem e desempenho específico dos dados. Além disso, como as informações sobre outliers não são conhecidas a priori, a única maneira de verificar se um ponto de dados é um outlier verdadeiro ou não é verificar os pontos de dados próximos e ver se ele segue características semelhantes.

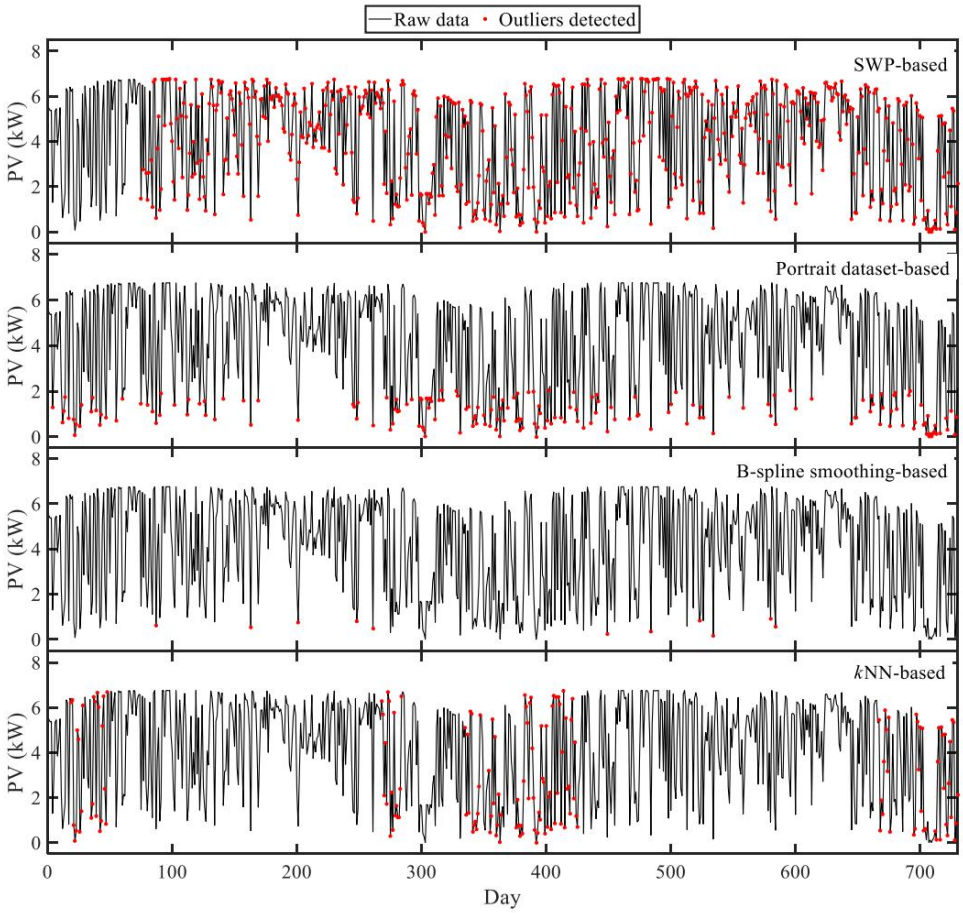

Fig. 8. Comparação da capacidade de detecção de outliers quando aplicada a séries temporais de geração FV.

4.2. Base para comparação da capacidade de correção de valores discrepantes dos métodos

Dentre o grande número de métodos de pré-processamento, poucos conseguem corrigir os outliers detectados. Portanto, não existe um padrão generalizado para comparar a capacidade de correção de valores discrepantes de um método. Assim, este artigo usa uma abordagem gráfica para examinar a capacidade de correção de outliers de diferentes métodos aplicados. De acordo com a abordagem gráfica, os pontos de dados identificados como outliers, após a correção, devem seguir a tendência e sazonalidade do conjunto de dados. Os métodos de pré-processamento bem estabelecidos são comparados na seção subsequente usando essas medidas como base de comparação.

5. Análise de resultados

5.1. Conjuntos de dados

Os dados de geração fotovoltaica são coletados de uma instalação fotovoltaica no telhado nos EUA [94]. Os dados de potência de carga são coletados de um restaurante em Anchorage, EUA [95], e da zona climática North-Central do Texas, EUA [96]. Os dados de temperatura ambiente são coletados de Berhampur, Índia [97]. Uma descrição detalhada das informações de tempo e do tamanho da amostra é elaborada na Tabela 4. Os vários casos da Tabela 4 são usados na análise dos resultados.

5.2. Comparação de desempenho de pré-processamento quando aplicado a dados limpos

O objetivo da análise nesta seção é comparar o desempenho de métodos bem estabelecidos quando aplicados a dados limpos. Primeiramente, a análise é feita para métodos bem estabelecidos aplicáveis a séries temporais univariadas (dados limpos, ou seja, Caso 2 da Tabela 4). Os dados são mostrado na Fig. 5. Observando os dados da Fig. 5, fica evidente que os dados estão limpos (livres de outliers). Portanto, para comparação de desempenho usando os índices definidos na Seção 4.1.1, 100 valores discrepantes são introduzidos nos dados e agora são denominados "dados poluídos". Isso é feito manipulando 100 pontos de dados selecionados aleatoriamente. O valor de um ponto de dados é aumentado em 50% se for maior que o valor médio; caso contrário, o valor do ponto de dados é reduzido em 30%. Os dados poluídos também são mostrados na Fig. 5.

Os métodos bem estabelecidos aplicáveis a séries temporais univariadas, conforme elucidado na Seção 3, são aplicados a dados poluídos, e seu desempenho de detecção de valores discrepantes é comparado na Fig. 6 e na Tabela 5. Cada método é testado com uma faixa de valores de parâmetros significativos , e o conjunto de valores dos parâmetros é escolhido no qual o desempenho do método é ótimo. Os resultados da comparação mostram que a abordagem baseada em suavização de B-spline tem um desempenho melhor do que todos os outros métodos. Ele detecta um número comparativamente menor de valores atípicos verdadeiros, mas também detecta apenas alguns valores atípicos falsos devido à sensibilidade da curva ajustada em relação aos valores atípicos presentes nos dados. Como todos os pontos de dados, incluindo valores discrepantes, são usados para ajustar uma curva, o ajuste insuficiente/sobreajuste é a principal desvantagem dessa abordagem. Os desempenhos de detecção e correção de valores discrepantes, dependendo da curva ajustada, são afetados devido ao ajuste inferior/superior. Considerando que, com base em conjunto de dados de retrato, detecta todos os outliers introduzidos nos dados e muitos outliers falsos. O método marca muitos pontos de dados genuínos como discrepantes. Baseia-se na suposição de que os dados em um conjunto de dados de retrato são semelhantes, o que geralmente não é o caso de dados voláteis. A suposição leva à detecção de pontos de dados genuínos em um conjunto de dados de retrato como valores discrepantes. A abordagem baseada em SWP fornece resultados comparáveis aos de uma abordagem baseada em conjunto de dados de retrato, mas marca um número maior de pontos de dados genuínos como discrepantes. O método falha por causa de sua abordagem de correção de valores discrepantes. A abordagem baseada em SWP usa um tipo de modelo AR para prever o valor do ponto de dados

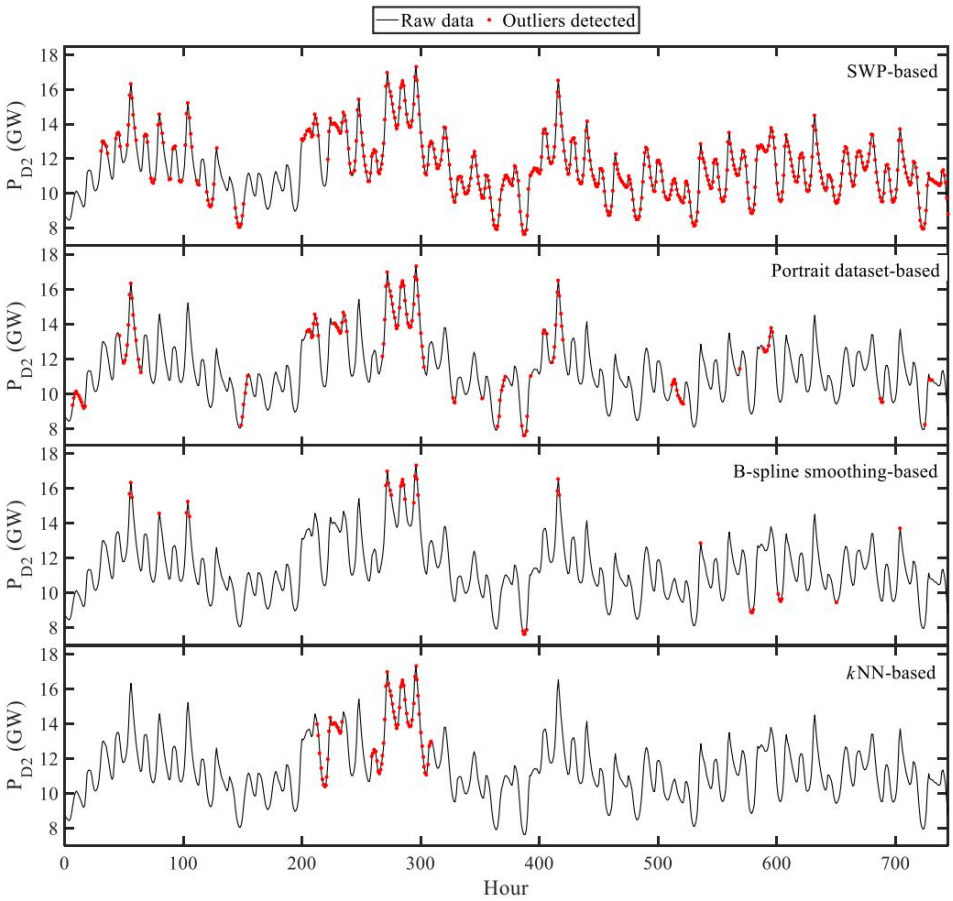

Fig. 9. Comparação da capacidade de detecção de outlier quando aplicada a séries temporais de potência de carga.

em cada instante de tempo, o que só é aplicável a dados estacionários. A previsão imprecisa leva ao cálculo de um intervalo de confiança inadequado, que resulta na detecção de um ponto de dados genuíno como um valor discrepante, sendo substituído pelo valor previsto incorreto. O ponto de dados é usado para prever os próximos pontos de dados. O erro se propaga e, após um certo instante de tempo, o método marca todos os pontos de dados como discrepantes [98].

A abordagem baseada em kNN tem um desempenho abaixo da média. A precisão do método diminui com o aumento do efeito de volatilidade [98]. Supõe-se que as janelas contendo valores discrepantes sejam distintas das demais, o que não é o caso dos dados voláteis. Além disso, essa abordagem não pode detectar valores discrepantes em instantes de tempo específicos; em vez disso, ele marca vários pontos de dados (janela que contém valores discrepantes) como discrepantes. Isso resulta na detecção de um alto número de falsos valores discrepantes.

O desempenho de correção de outlier é comparado na Fig. 7.

É evidente na Fig. 7 que a abordagem baseada em conjunto de dados de retrato segue a tendência dos resultados de comparação anteriores (mostrados na Fig. 6 e na Tabela 5) e tem um desempenho melhor do que outros métodos. Os dados pré-processados desviam-se menos dos dados limpos e seguem o padrão periódico dos dados limpos, mesmo depois de detectar um grande número de falsos valores discrepantes, porque os dados limpos são menos voláteis e os dados em um conjunto de dados em retrato são semelhantes. O desempenho da abordagem baseada em conjunto de dados de retrato é diferente quando o método é aplicado a dados altamente voláteis.

Os outliers, após a correção usando uma abordagem baseada em suavização B-spline, também seguem os dados limpos. Mas, devido à capacidade dessa abordagem de detectar todos os valores discrepantes, os dados gerais pré-processados se desviam dos dados limpos. Considerando que a abordagem baseada em SWP falha completamente no pré-processamento dos dados, e os dados após o pré-processamento são incapazes de seguir o padrão periódico. A partir dos resultados da comparação de detecção e correção de valores discrepantes acima, é perceptível que o conjunto de dados de retrato abordagem baseada supera os outros métodos quando os dados são menos voláteis e limpos.

Todas as análises nesta seção são feitas considerando dados limpos menos voláteis. Mas, o desempenho dos métodos bem estabelecidos varia quando são aplicados a dados brutos mais voláteis. Essa análise é feita na seção seguinte.

5.3. Comparação de desempenho de pré-processamento quando aplicado a dados brutos

Como o objetivo desta seção é analisar o desempenho dos métodos bem estabelecidos quando aplicados a dados brutos voláteis, os Casos 1, 3 e 5 da Tabela 4 são usados para as análises. Como as informações sobre outliers não são conhecidas a priori nos dados brutos, uma comparação baseada em índice não é possível. Os métodos são aplicados a várias séries temporais, com diferentes resoluções temporais, e seus desempenhos são comparados observando os resultados gráficos. A capacidade de detecção de valores discrepantes de métodos bem estabelecidos quando aplicados a diferentes dados brutos é elucidada nas Figs. 8, 9 e 10.

A partir da aplicação de diferentes métodos bem estabelecidos em vários tipos de conjuntos de dados nas Figs. 8, 9 e 10, é evidente que a abordagem baseada em suavização de B-spline tem um desempenho melhor do que outros métodos. Embora detecte um número menor de valores atípicos verdadeiros, ele marca um número muito menor de pontos de dados genuínos como valores atípicos. Considerando que, abordagens baseadas em conjunto de dados de retrato e baseadas em kNN detectam um grande número de falsos outliers junto com os verdadeiros outliers. A aplicação da abordagem baseada em SWP para qualquer tipo de dados falha após um determinado instante de tempo e marca todos os pontos de dados como discrepantes. A razão para tal desempenho dos métodos já é explicada em detalhes na Seção 5.2.

Para comparar a capacidade de correção de outliers de métodos bem estabelecidos quando aplicados a dados brutos voláteis, eles são aplicados a vários

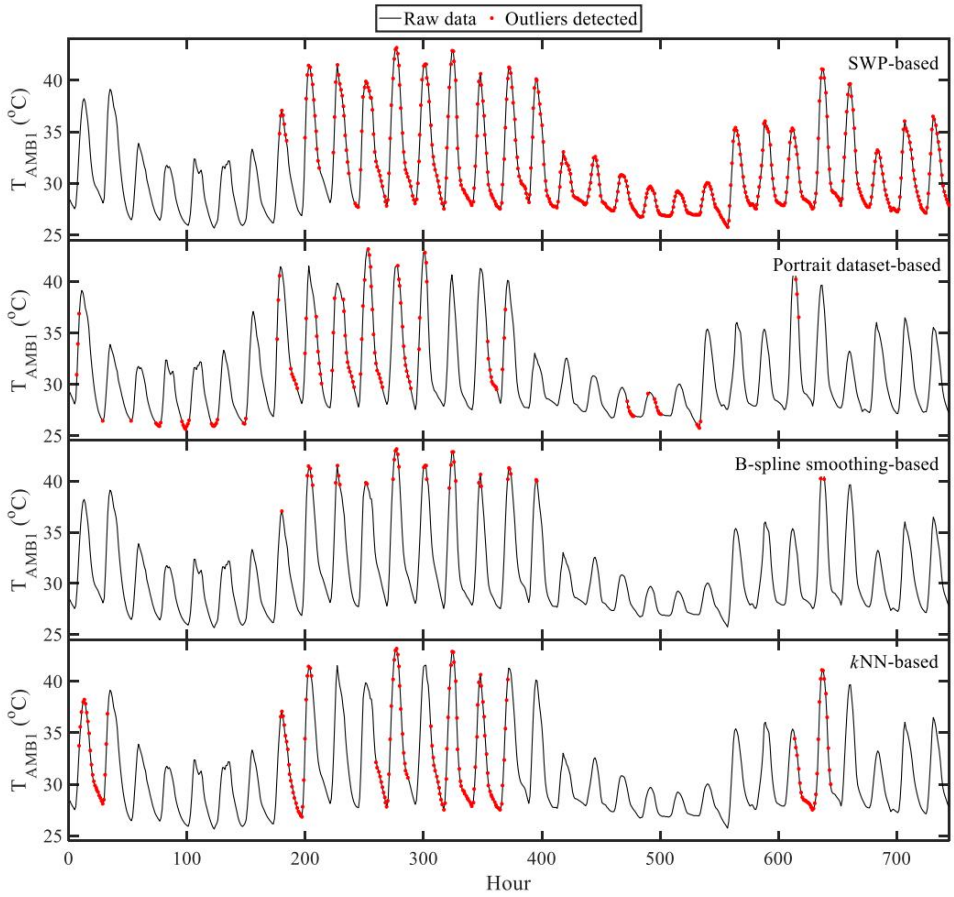

Fig. 10. Comparação da capacidade de detecção de valores discrepantes quando aplicada a dados de temperatura ambiente.

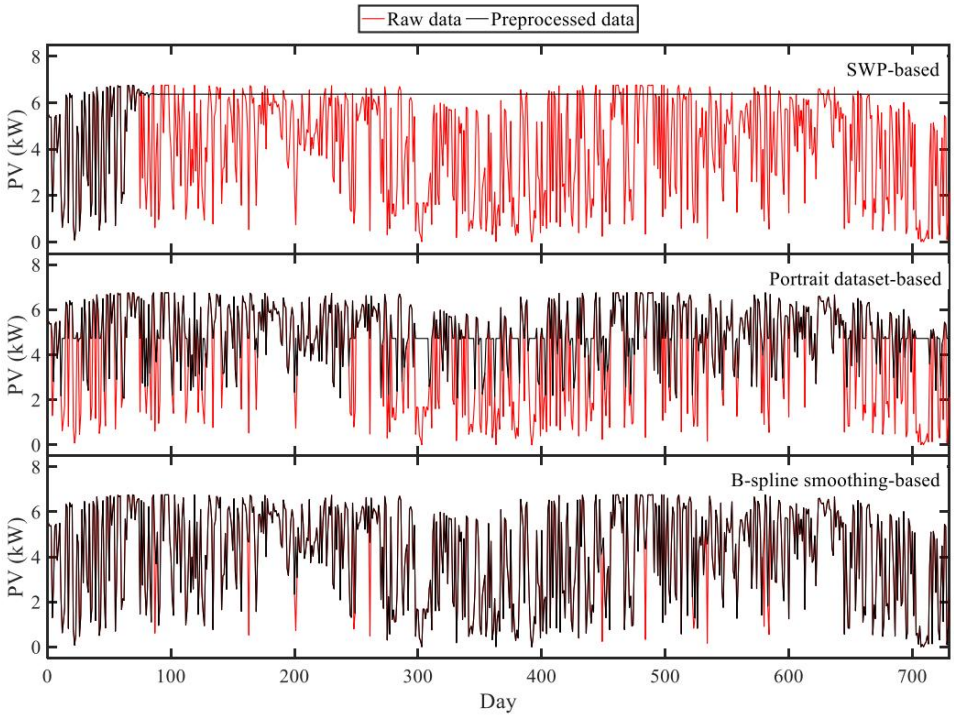

Fig. 11. Comparação da capacidade de correção de outlier quando aplicada aos dados de geração fotovoltaica.

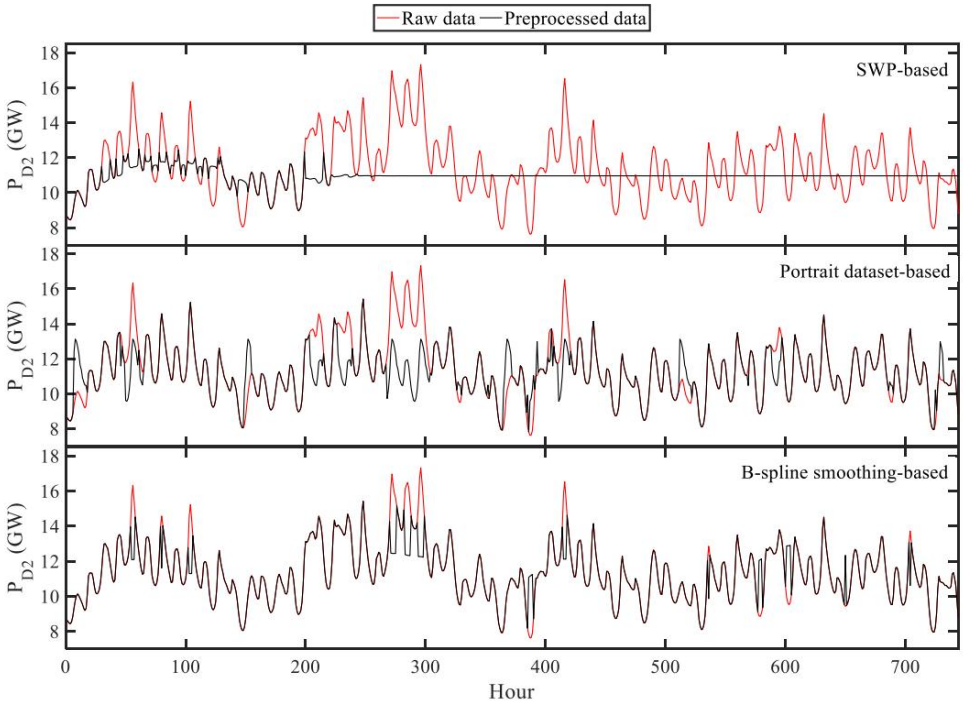

Fig. 12. Comparação da capacidade de correção de outlier quando aplicada a dados em dados de potência de carga.

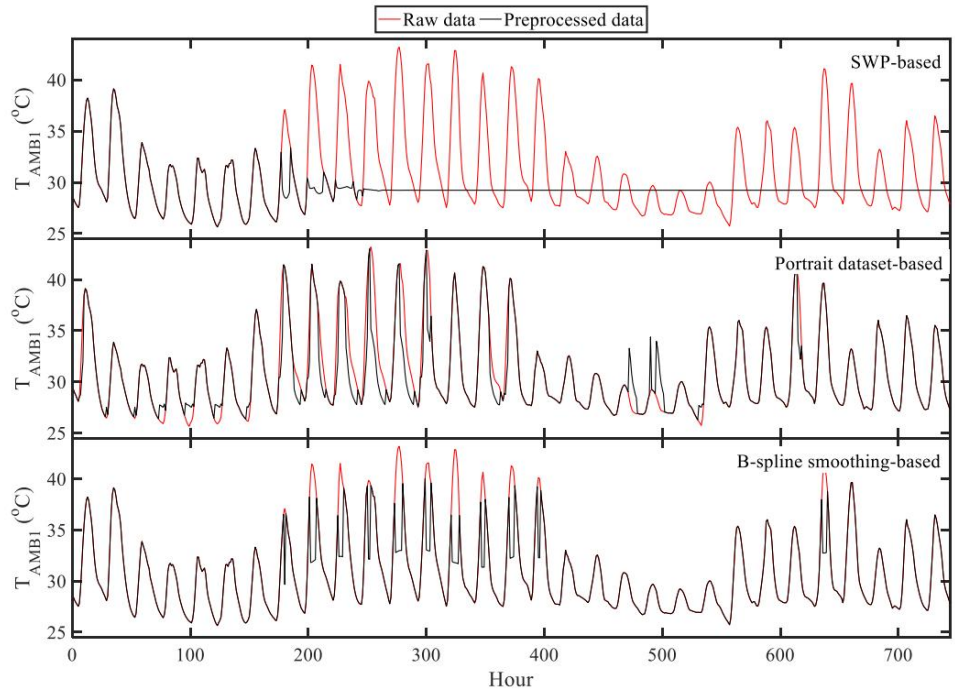

Fig. 13. Comparação da capacidade de correção de outlier quando aplicada a dados de temperatura ambiente.

dados brutos, e os resultados são analisados nas Figs. 11, 12 e 13.

Fica claro a partir da análise acima que a abordagem baseada em suavização de B-spline supera outros métodos quando aplicada a dados brutos voláteis.
Os dados após o pré-processamento seguem a tendência e o padrão sazonal dos dados brutos. A vantagem de usar essa abordagem é que o método depende da curva ajustada, que pode ser controlada pelo usuário.

Por outro lado, a abordagem baseada em conjunto de dados de retrato é incapaz de manter as características dos dados brutos intactas após o pré-processamento. A principal razão para o fracasso da abordagem é a suposição de que os dados em um conjunto de dados de retrato são estáveis. Mas, em dados voláteis, dados em um conjunto de dados de retrato têm tendência e padrões periódicos. Devido a isso, muitos falsos valores discrepantes são detectados e os dados após a correção não conseguem rastrear a tendência e os padrões periódicos dos dados brutos. A abordagem baseada em SWP falha após um certo instante de tempo devido à propagação do erro de previsão, que afeta a detecção e correção de valores discrepantes. Portanto, os dados após o pré-processamento tornam-se constantes após um certo instante de tempo.

Das análises detalhadas nas Seções 5.2 e 5.3, vale a pena notar os seguintes pontos importantes. A abordagem baseada em SWP não pode pré-processar nenhum dado volátil, pois falha após um certo instante de tempo em todos os tipos de dados. O método baseado em conjunto de dados de retrato funciona bem no caso

Tabela 6

Efeito na largura da janela no desempenho da abordagem baseada em SWP.

Índices	Largura da janela *(2kÿ)* 12 *k* = 2	216k =	220 *mil* =
Nº de TP	99	99	98
Nº de FP	613	234	208
Nº de FN	1	1	2
P	0,1390	0,2973	0,3203
R	0,99	0,99	0,98
F-medida	0,2438	0,4573	0,4828

Tabela 7

Efeito de VPDs no desempenho da abordagem baseada em conjunto de dados de retrato.

Índices	6 VPDs	12 VPDs	24 VPDs
Nº de TP	84	69	100
Nº de FP	130	61	133
Nº de FN	16	31	0
P	0,3925	0,5307	0,4292
R	0,84	0,69	1
F-medida	0,5350	0,5999	0,6006

de dados menos voláteis, onde os dados em um conjunto de dados de retrato são semelhantes. Ainda assim, a precisão diminui quando aplicada a dados altamente voláteis. O método baseado em suavização de B-spline é melhor do que outros métodos em todos os aspectos, independentemente dos dados serem brutos ou voláteis. O método baseado em kNN também não é aconselhável para dados voláteis. Como ele assume a janela contendo valores atípicos, ele detecta muitos valores atípicos falsos e faz apenas metade do pré-processamento, pois não é capaz de corrigir valores atípicos.

O(s) parâmetro(s) crucial(is), resolução de tempo e volatilidade dos dados impactam significativamente a detecção de discrepâncias e o desempenho de correção dos métodos bem estabelecidos. Esses efeitos são elaborados em Seção 5.4.

5.4. Impacto da mudança de valores de parâmetros cruciais e resolução de tempo no pré-processamento

Os desempenhos dos métodos bem estabelecidos são discutidos na Seção 5.2 e 5.3. As razões para seu desempenho observado e o efeito do parâmetro são explicados em detalhes nesta seção.

5.4.1. Abordagem baseada em previsão de janela deslizante

O efeito de parâmetros cruciais no desempenho da abordagem baseada em SWP é inferido, considerando o Caso 2 da Tabela 4. O método é aplicado aos dados selecionando diferentes valores de largura de janela, e os resultados são mostrados na Tabela 6. É evidente a partir a mesa que janela largura tem um impacto significativo no desempenho do método. A razão está no modelo que é usado pela abordagem baseada em SWP. Ele usa um tipo de modelo AR para prever o valor dos pontos de dados em cada instante, e a largura da janela é a ordem do modelo. Portanto, uma largura de janela adequada estabelece uma relação errada entre os dados, prevendo um valor inadequado de dados a cada instante de tempo. Devido à previsão imprecisa, um intervalo de confiança errado é calculado, o que resulta na detecção de pontos de dados genuínos como valores discrepantes. Os valores previstos inadequados substituem os valores discrepantes detectados e os dados pré-processados não conseguem seguir a tendência e os padrões sazonais dos dados originais. Este fenômeno torna-se mais proeminente em dados voláteis, como evidenciado pelos resultados das Seções 5.2 e 5.3.

5.4.2. Abordagem baseada em conjunto de dados de retrato

O efeito de VPDs no desempenho do método é demonstrado usando o Caso 2 da Tabela 4. O método é aplicado aos dados mesclando conjuntos de dados de retrato semelhantes em diferentes números de VPDs (selecionados aleatoriamente) e os resultados são comparados em Tabela 7. Depreende-se da Tabela 7 que o número de VPDs tem um efeito significativo no desempenho do método. Como o método assume os dados dentro de um conjunto de dados de retrato para ser estável, um conjunto de dados de retrato semelhante deve ser mesclado para o desempenho ideal do método. Mas o método detecta um grande número de falsos valores discrepantes em qualquer dado; os dados em um conjunto de dados de retrato não são estáveis no caso de dados voláteis. Sua capacidade de correção é melhor do que outros métodos quando aplicada a dados menos voláteis. No entanto, a sua aplicação a dados voláteis, como é evidente nas Figs. 11 e 12 indicam que os dados pré-processados não podem seguir as características dos dados brutos. O método funciona bem apenas com um conjunto de dados de retrato estável (não volátil), o que geralmente não é o caso.

5.4.3. Abordagem baseada em suavização de B-spline

Esse método funciona melhor do que outros métodos na maioria dos casos. O método depende exclusivamente da curva ajustada. A curva ajustada resulta das funções de base utilizadas, que estão na mão do usuário. A única lacuna que este método enfrenta é a sensibilidade da curva ajustada para os valores discrepantes presentes nos dados. O caso 4 da Tabela 4 é considerado para a análise para demonstrar essa sensibilidade. O método é aplicado aos dados e, em seguida, dez valores discrepantes são introduzidos manualmente a partir do dia 96-105. O método é aplicado aos dados novamente e os resultados são comparados na Fig. 14. É evidente a partir da figura que a curva B-spline ajustada é sensível aos outliers presentes nos dados, pois o método não detecta todos os outliers quando aplicado a dados manipulados, conforme detectado anteriormente em dados brutos. A inclusão de todos os pontos de dados (out liers e pontos de dados genuínos) para ajustar a curva aos dados é a causa raiz desse problema. A forma da curva muda para ajustar a curva a todos os pontos de dados. Devido à mudança na forma da curva, alguns outliers vão

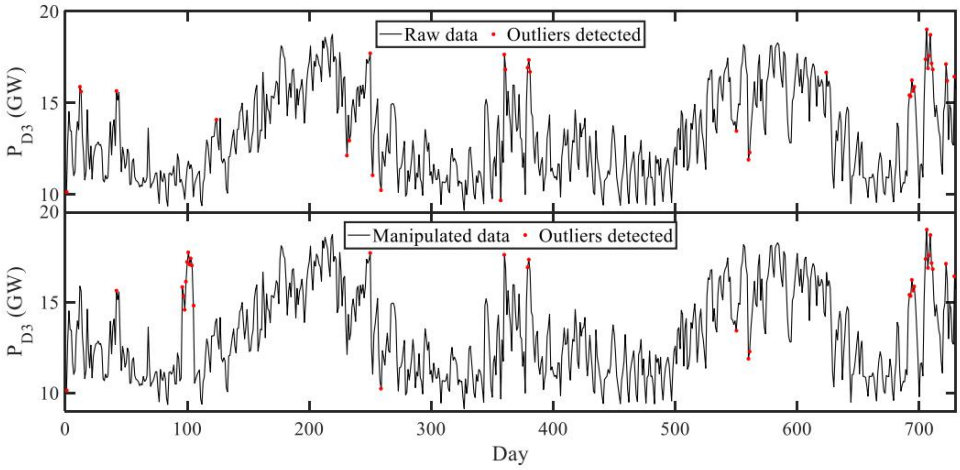

Fig. 14. Sensibilidade da curva B-spline ajustada para os outliers presentes nos dados.

Tabela 8.

Efeito das funções de base no desempenho da abordagem baseada em suavização de B-spline.

Índices	Funções básicas			
	Termos de Fourier		Função aleatória	
	Conjunto de candidatos sensatos	Seleção aleatória Cúbica		Quadrático
Nº de TP Nº	56	16	39	30
de FP Nº	7	0	41	50
de FN 44 P		84	61	70
	0,8889	1	0,4875	0,325
R	0,56	0,16	0,39	0,30
F-medida	0,6871	0,2759	0,4333	0,3333

Tabela 9

Efeito de parâmetros cruciais no desempenho da abordagem baseada em kNN.

Índices	$k = 3$			$k = 5$		
	$L' = 12$	$L' = 16$	$L' = 24$	$L' = 12$	$L' = 16$	$L' = 24$
Nº de TP Nº	35	33	31	31	34	27
de FP	113	118	128	106	121	113
Nº de FN 65		67	69	69	66	73
P	0,2365	0,2185	0,195	0,2263	0,2194	0,1929
R	0,35	0,33	0,31	0,31	0,34	0,27
F-medida	0,2823	0,2629	0,2394	0,2616	0,2667	0,2250

Tabela 10

Efeito das normas L1 e L2 no desempenho da abordagem baseada em kNN quando aplicada a dados poluídos.

Índices	Norma L1	Norma L2
Nº de TP	39	35
Nº de FP	108	111
Nº de FN P	61	65
	0,2653	0,2397
R	0,39	0,35
F-medida	0,3158	0,2846

Tabela 11.

Revisão sobre a aplicabilidade de métodos de pré-processamento quando aplicados a séries temporais voláteis.

Coeficiente de confiança	Índices	Baseado em conjunto de dados Portrait baseado em SWP	B-spline alisamento baseado	baseado em kNN
0,05	Nº de TP 98 Nº de FP 208 Nº de FN 2	100	23	35
		133	0	111
		0	77	65
	P 0,3203	0,4292	1	0,2397
	R 0,98	1	0,23	0,35
	F-medida 0,4828	0,6006	0,374	0,2846
0,10	Nº de TP 99 Nº de FP 229 Nº de FN 1 P	100	56	50
		143	7	187
		0	44	50
	0,3018	0,4115	0,8889	0,2110
	R 0,99	1	0,56	0,5
	F-medida 0,4626	0,5831	0,6871	0,2967
0,20	Nº de TP 99	100	92	58
	Nº de FP 348 Nº de FN 1	180	102	270
		0	8	42
	P 0,2215	0,3571	0,4742	0,1768
	R 0,99	1	0,92	0,58
	F-medida 0,3620	0,5263	0,6259	0,2710

não detectado, uma vez que a detecção de valores discrepantes é baseada apenas na curva ajustada. No entanto, a capacidade de correção de outliers é melhor do que outros métodos, independentemente do efeito de volatilidade dos dados.

Além da sensibilidade da curva ajustada para os outliers presentes nos dados, o desempenho do método também depende das funções de base utilizadas para o ajuste da curva. O Caso 2 da Tabela 4 é escolhido para mostrar essa dependência, e o método é aplicado usando diferentes funções de base. Como os dados são periódicos (assim como outros dados usados na análise dos resultados), o método proposto em [7] é aplicado a cada semana da série temporal considerada para selecionar um conjunto candidato sensato de componentes de frequência para capturar as sazonalidades. O método é primeiramente implementado usando termos de Fourier correspondentes aos componentes de frequência dominante (semanal e diário). Posteriormente, ele é implementado usando termos de Fourier correspondentes aos componentes de frequência selecionados aleatoriamente (semanal), e usando funções aleatórias como funções de base, e os resultados são comparados na Tabela 8.

É evidente na Tabela 8 que o desempenho é sensível às funções de base usadas para ajustar uma curva. A seleção de qualquer função aleatória/componente de frequência produz bons resultados. Assim, para quaisquer dados periódicos específicos, os componentes de frequência dominantes devem ser revelados para se ajustarem a uma curva que capture as sazonalidades dos dados e obtenha bons resultados de pré-processamento.

5.4.4. Abordagem baseada em k-vizinhos mais

próximos O caso 2 da Tabela 4 é considerado para a análise, e o método é aplicado aos dados definindo diferentes valores dos parâmetros cruciais.
O efeito de parâmetros cruciais sobre o desempenho do método é elucidado na Tabela 9. Além disso, para ilustrar o efeito da norma de desempenho do método, ele é implementado utilizando a norma L1 ao invés da norma L2 convencional. Os resultados da comparação são mostrados na Tabela 10.

Depreende-se da Tabela 9 que o parâmetro *k* não tem papel significativo na determinação do desempenho. Em contraste, o comprimento da janela (L') tem um impacto substancial no desempenho do método. A Tabela 10 mostra que o desempenho do método é ligeiramente elevado ao utilizar a norma L1 para calcular a distância entre as viúvas, pois a norma L1 é mais robusta em relação à norma L2. Como a norma L2 é utilizada nos artigos base, o método é implementado utilizando a mesma em todas as análises dos resultados. Além disso, o método é incapaz de detectar valores discrepantes em um determinado instante de tempo. Ele marca um grupo de dados como valores discrepantes, incluindo pontos de dados genuínos, conforme evidenciado nas Seções 5.2 e 5.3. Além disso, a incapacidade do método de corrigir outliers o torna o método menos preferido dentre os outros métodos bem estabelecidos.

5.4.5. Efeito do coeficiente de confiança O

desempenho dos métodos bem estabelecidos é afetado pelo coeficiente de confiança. O caso 2 da Tabela 4 é considerado para mostrar isso

efeito, e os métodos são aplicados aos dados definindo diferentes valores de coeficiente de confiança. Os resultados são comparados na Tabela 11.

É evidente na Tabela 11 que o coeficiente de confiança influencia

o desempenho do respectivo método bem estabelecido, bem como o método proposto. Em outras palavras, o intervalo no qual um ponto de dados é tratado como normal/genuíno, ou seja, o intervalo de confiança, é parcialmente dependente do coeficiente de confiança. Além do coeficiente de confiança, o intervalo de confiança depende de outros fatores específicos do método (por exemplo, na abordagem baseada em SWP, o intervalo de confiança para um ponto de dados também depende de seu valor previsto, o desvio padrão de sua vizinhança e a largura da janela escolhida), que são explicadas na Seção 3.2. Portanto, o grau com que o coeficiente de confiança afeta o desempenho de um método varia de método para método. Um coeficiente de confiança inadequado resulta em intervalos de confiança imprecisos, o que diminui ainda mais o número de outliers verdadeiros e aumenta o número de outliers falsos, sendo detectados. Portanto, um coeficiente de confiança adequado deve ser escolhido para o desempenho ideal de um método, que constrói um intervalo de confiança pontual para detectar valores discrepantes.

Além do(s) parâmetro(s) crucial(is), os métodos bem estabelecidos também são sensíveis ao efeito de volatilidade dos dados, que é explicado em detalhes na Tabela 12.

Tabela 12

Revisão da aplicabilidade de métodos bem estabelecidos quando aplicados a séries temporais voláteis.

Métodos	Sensibilidade do método ao efeito da volatilidade
Baseado em SWP	A abordagem de correção de outlier falha quando aplicada a dados voláteis. O erro se propaga, levando à falha completa do pré-processamento após um determinado período. Portanto, este método não se aplica a séries temporais voláteis.
Baseado em dados de retrato	Pressupõe que os dados em um conjunto de dados em retrato sejam estáveis. Portanto, esse método falha quando aplicado a séries temporais voláteis à medida que a variação dentro do conjunto de dados de retrato aumenta.
B-spline com base em suavização	O método se aplica a dados voláteis, desde que a curva seja adequadamente ajustada usando uma função de base adequada. O desempenho do método não se degrada com o aumento do efeito de volatilidade dos dados.
baseado em kNN	Assume que a janela que contém outlier(s) será diferente de todas as outras janelas. Portanto, esse método falha quando aplicado a séries temporais voláteis, pois as janelas genuínas são muito diferentes umas das outras em uma série temporal volátil.

Tabela 13

Méritos, lacunas e possibilidades de improvisação de métodos bem estabelecidos.

Baseado em SWP	Méritos	• Não usa nenhum parâmetro estatístico afetado pela presença de outliers. • A série temporal não volátil não perde sua tendência
	Lacunas	inerente e padrão sazonal após o pré-processamento. • Falha ao pré-processar uma série temporal volátil. • Nenhuma regra generalizada sugerida para a otimização do parâmetro crucial. • É necessário um caminho para a seleção ideal da largura da janela. • É necessária uma melhor abordagem de correção de
	Alcance	valores discrepantes. • Não usa nenhum parâmetro estatístico afetado pela presença de outliers. • Capaz de visualizar as características ocultas como tendência e sazonalidade. • A série temporal não volátil não perde sua tendência inerente e padrão sazonal após o pré-processamento. • Usa parâmetros estatísticos que
Com base no conjunto de dados do Portrait	Méritos	são afetados pela presença de outliers. • Diminuição da precisão com aumento da volatilidade das séries temporais. • Algoritmo pouco claro para mesclar conjuntos de dados de retrato semelhantes. • A abordagem de correção de valores discrepantes falha com o aumento da variação dentro dos conjuntos de
	Lacunas	dados de retrato. • Algoritmo para mesclar conjuntos de dados de retrato semelhantes. • Uma melhor abordagem de correção de outliers é necessária no caso de séries temporais voláteis. • Não usa nenhum parâmetro estatístico afetado pela presença de outliers. • As séries temporais não voláteis e voláteis não perdem sua tendência inerente e padrão sazonal após o pré-processamento. • A seleção de funções de base apropriadas e o número de funções de base a serem usadas são as principais preocupações. • Under-fitting e over-fitting levam, respectivamente, à detecção de falsos outliers, ignorando um genuíno outlier.
	Alcance	• Processo demorado e alto poder computacional são necessários para sua aplicação em grandes dados. • Como é utilizado modelo de regressão não paramétrica, que não considera a forma estrutural pré-definida dos dados, a curva ajustada é sensível à presença de outliers. • A sensibilidade da curva ajustada para os outliers presentes nos dados precisa ser reduzida. • Não requer conhecimento prévio do tipo de distribuição que uma série temporal apresenta. • Capaz
B-spline com base em suavização	Méritos	de detectar outliers em uma série temporal não volátil. • Incapaz de corrigir discrepâncias, portanto, faz apenas metade do trabalho. • Não é possível detectar valores discrepantes em um determinado instante de tempo. • O algoritmo de detecção de valores discrepantes falha quando aplicado a séries temporais voláteis. • Nenhuma regra generalizada sugeriu otimizar os parâmetros cruciais. • É necessária uma abordagem de correção de valores discrepantes. • É preciso encontrar uma maneira de
	Lacunas	detectar discrepâncias em um determinado instante de tempo.
	Alcance	
baseado em kNN	Méritos	
	Lacunas	
	Alcance	

5.5. Méritos, lacunas e possibilidades de improvisação

Os méritos, lacunas e escopos de improvisação para as abordagens de detecção e correção de valores discrepantes bem estabelecidos são discutidos na Tabela 13.

6. Conclusão

Um levantamento rigoroso da pesquisa em pré-processamento de dados é realizado categorizando vários métodos de pré-processamento. O conceito de out lier, as aplicações de pré-processamento e as principais categorias de métodos de pré-processamento são elaborados de forma compreensível. Muitos métodos de pré-processamento existentes são analisados, categorizados e a capacidade de pré-processamento de cada técnica é destacada. As técnicas bem estabelecidas de cada categoria são elaboradas com uma explicação detalhada de seu algoritmo. Dois tipos diferentes de dados, ou seja, os dados limpos menos voláteis e os dados brutos altamente voláteis, são considerados para a análise dos resultados. Diferentes bases são definidas para comparar os métodos bem estabelecidos quando aplicados a dados limpos e dados brutos. Os métodos bem estabelecidos são aplicados a séries temporais univariadas (limpas e brutas) com diferentes resoluções de tempo. Os desempenhos das técnicas bem estabelecidas são analisados criticamente com base em sua capacidade de detecção e correção de outliers. O impacto da mudança em importantes valores de parâmetro(s) e resolução temporal dos dados sobre o desempenho dos métodos seletivos também são elucidados. Por fim, destacam-se os méritos, lacunas e possibilidades de improvisação das técnicas já consagradas.

A partir das análises detalhadas dos diferentes métodos bem estabelecidos quando aplicados aos dados limpos menos voláteis e aos dados brutos mais voláteis, as inferências significativas são as seguintes:

1) A abordagem baseada em SWP funciona bem com dados limpos menos voláteis em comparação com os dados brutos mais voláteis. Ele tem uma precisão menor de 0,3203, mas um recall comparativamente alto de 0,98. Devido à sua dependência do modelo de previsão do tipo AR, ele falha quando aplicado a séries temporais altamente voláteis. Sua capacidade de pré-processamento é altamente dependente da largura da janela e do coeficiente de confiança, como evidenciado nas análises. Portanto, é necessária uma seleção ideal da largura da janela, juntamente com uma melhor técnica de correção de valores discrepantes.
2) A abordagem baseada em conjunto de dados de retrato tem um desempenho decente quando aplicada a dados menos voláteis. Apesar de muitos falsos positivos, verifica-se uma recordação perfeita de 1 no caso-2 considerado. No entanto, ele não pode refazer os padrões nos dados após a correção de valores discrepantes, quando aplicado aos dados brutos mais voláteis. É afetado pelo número de VPDs. Devido à suposição de dados estáveis em um conjunto de dados de retrato; retratos semelhantes precisam ser mesclados para um desempenho ideal. U

algoritmo adequado para isso é essencial. Além disso, é necessário melhorar a técnica de correção de outliers com dados voláteis.

3) A abordagem de suavização B-spline tem um desempenho relativamente melhor do que os outros métodos bem estabelecidos em dados. Possui a maior medida F de 0,6871, entre todos os métodos comparados, o que indica seu melhor desempenho. Ele fornece o menor número de falsos positivos, ao contrário das abordagens baseadas em conjunto de dados de retrato e baseadas em kNN. Este método é sensível aos outliers presentes nos dados e à função base utilizada para o ajuste da curva.
Às vezes, resulta em ajuste insuficiente/superior da curva de regressão não paramétrica, levando a detecções falsas.
4) A abordagem baseada em kNN só pode detectar outliers e não pode corrigi-los. Entre os diferentes métodos bem estabelecidos, tem o pior desempenho quando aplicado aos dados limpos, conforme indicado pelos menores valores de precisão, recall e F-measure, ou seja, 0,2110, 0,5 e 0,2967 respectivamente. Ele detecta o menor número de valores discrepantes verdadeiros com muitos falsos positivos. Devido à sua sensibilidade em relação ao comprimento da janela e à norma de distância, não tem bom desempenho e não é aconselhável para dados voláteis. Ele precisa de algoritmos para detecção de discrepâncias em instantes de tempo específicos e também para correção de discrepâncias.

Declaração de Interesse Concorrente

Os autores declaram que não conhecem interesses financeiros concorrentes ou relações pessoais que possam ter influenciado o trabalho relatado neste artigo.

Referências

[1] A. Bracale, P. Caramia, G. Carpinelli, AR Di Fazio, P Varilone, Uma abordagem bayesiana para uma previsão de estado estacionário de curto prazo de uma rede inteligente, IEEE Trans. Rede Inteligente 4 (4) (2013) 1760–1771.

[2] BR Prusty, D Jena, Uma avaliação de risco acima do limite do sistema de energia integrado PV usando fluxo de carga probabilístico baseado em modelagem de incerteza instantânea multi-tempo, Renew. Energia 116 (2018) 367-383.

[3] L. Zheng, W. Hu, Y Min, Pré-processamento de dados de vento bruto: uma abordagem de mineração de dados, Trans. IEEE Sustentar. Energia 6 (1) (2015) 11–19.

[4] J. An, Z. Bie, X. Chen, B. Hua, S Liu, Um método generalizado de pré-processamento de dados para previsão de energia eólica, 2013 IEEE Power & Energy Society General Meeting, 2013, pp. 1–5.

[5] JY Wang, Z. Qian, H. Zareipour, D Wood, Avaliação de desempenho de módulos fotovoltaicos com base na estimativa diária de geração de energia, Energy 165 (2018) 1160–1172.

[6] BR Prusty, D Jena, Etapas de modelagem de incerteza para estado estacionário probabilístico Analysis, Applications of Computing, Automation and Wireless Systems in Electrical Engineering, Springer, Cingapura, 2019, pp. 1169–1177.

[7] BR Prusty, D. Jena, Uma temperatura baseada em modelo probabilístico espaço-temporal fluxo de carga probabilístico aumentado considerando gerações PV, Int. Trans. Elétrico. Sistema de Energia 29 (5) (2019) e2819 https://doi.org/10.1002/2050-7038.2819.

[8] JS Park, CG Park, KE Lee, Detecção simultânea de outliers e seleção de variáveis via modelo de regressão baseado em diferenças e seleção de variáveis de busca estocástica, Commun. Estado. Aplic. Métodos 26 (2) (2019) 149-161.

[9] DD Prastyo, FS Nabila, MH Lee, N. Suhermi, SF Fam, VAR e seleção de recursos baseada em GSTAR em regressão vetorial de suporte para previsão espaço-temporal multivariada , Conferência Internacional sobre Computação Suave em Ciência de Dados, Cingapura, Springer, 15 de agosto de 2018, pp. 46–57.

[10] JL Speiser, ME Miller, J. Tooze, E Ip, Uma comparação de métodos de seleção de variáveis florestais aleatórias para modelagem de previsão de classificação, Expert Syst. Aplic. 134 (2019) 93-101.

[11] Y. Jiang, Y. Wang, J. Zhang, B. Xie, J. Liao, W Liao, detecção de outlier e seleção de variável robusta através do método LAD-LASSO ponderado penalizado, J. Appl. Estado. (2020) 1-13.

[12] HawkinsD.M.Identificação de outliers. Monografias sobre Probabilidade e Estatística Aplicadas; 1980.

[13] EQ McCallum, Bad Data handbook: Mapeando o mundo dos problemas de dados, O'Reilly Media, Sebastopol, CA, 2012.

[14] H. Wang, MJ Bah, M. Hammad, Progresso em técnicas de detecção de valores discrepantes: uma pesquisa, IEEE Access 7 (2019) 107964–108000.

[15] X. Dong, L. Qian, L Huang, Uma abordagem de aprendizagem baseada em ensacamento baseada na CNN para previsão de carga de curto prazo em redes inteligentes, IEEE SmartWorld, Inteligência e Computação Ubíqua, Computação Avançada e Confiável, Computação e Comunicações Escaláveis , Nuvem e Computação de Big Data, Internet das Pessoas e Inovação em Cidades Inteligentes (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), San Francisco, CA, 2017, pp. 1–6.

[16] S. Ghosh, L Reilly D, Detecção de fraude de cartão de crédito com uma rede neural, Proceedings of the 27th Annual Hawaii International Conference on System Science, Los Alamitos, CA, 1994, p. 3.

[17] D. Dasgupta, N. Majumdar, detecção de outlier em dados multidimensionais usando algoritmo de seleção negativa, Anais da Conferência IEEE sobre Computação Evolutiva, Havaí, 2002, pp. 1039-1044.

[18] S. Ghanbari, AB Hashemi, C Amza, detecção de anomalia com reconhecimento de palco por meio de pontos de log de rastreamento, Proceedings of the 15th International Middleware Conference, 2014, pp. 253–264.

[19] HJ Escalante, Uma comparação de algoritmos de detecção de outliers para aprendizado de máquina, Proceedings of the International Conference on Communications in Computing, 2005, pp. 228–237.

[20] ZhangJ.Advancements of outlier detection: a survey. Trans. ICST. Informe Escalável. Syst.2013;13(1):1–26.

[21] D. Dave, T Varma, Uma revisão de vários métodos estatísticos para detecção de valores discrepantes, Int. J. Computação. Sci. Eng. Tecnol. 5 (2) (2014) 137-140.

[22] BhosaleS.V. Uma Pesquisa: Detecção de Outliers em Streaming de Dados Usando Clustering Approached. IJCSIT) Int. J. Comp. Sci. Informar. Tech.2014;5:6050–6053.

[23] V. Chandola, A. Banerjee, V Kumar, detecção de anomalias: uma pesquisa, ACM Comput. Sobreviv. (CSUR) 41 (3) (2009) 1–58.

[24] JohansenS., NielsenB.Algoritmos de detecção de outliers para séries temporais de mínimos quadrados re agressão. Disponível em SSRN 2510281; 2014.

[25] M. Gupta, J. Gao, CC Aggarwal, J Han, detecção de outlier para dados temporais: uma pesquisa, IEEE Trans. Conhecimento Eng. de Dados 26 (9) (2013) 2250-2267.

[26] ZA Bakar, R. Mohemad, A. Ahmad, MM Deris, Um estudo comparativo para técnicas de detecção de outliers em mineração de dados, conferência IEEE 2006 sobre cibernética e sistemas inteligentes, 2006, pp. 1–6.

[27] M. Goldstein, S Uchida, Uma avaliação comparativa de algoritmos de detecção de anomalias não supervisionadas para dados multivariados, PLoS ONE 11 (4) (2016) e0152173.

[28] PJ Rousseeuw, I. Ruts, JW Tukey, The bagplot: a boxplot bivariado, Am. Estado. 53 (4) (1999) 382-387.

[29] B. Abraham, A Chuang, detecção de valores discrepantes e modelagem de séries temporais, Technometrics 31 (2) (1989) 241-248.

[30] GM Ljung, Sobre detecção de valores discrepantes em séries temporais, JR Stat. Soc. Série B (Metodológico) 55 (2) (1993) 559-567.

[31] BG Amidan, TA Ferryman, SK Cooley, detecção de dados discrepantes usando o Teorema de Chebyshev, 2005 IEEE Aerospace Conference, 2005, pp. 3814-3819.

[32] M Yue, Um método integrado de detecção de anomalias para dados de previsão de carga sob ataques cibernéticos, 2017 IEEE Power & Energy Society General Meeting, 2017, pp. 1–5.

[33] PM Broersen, ARMASel para detecção e correção de outliers em dados estocásticos univariados, IEEE Trans. Instrumento Medir. 57 (3) (2008) 446-453.

[34] HillD.J., MinskerB.S.Detecção de anomalias na transmissão de dados de sensores ambientais: uma abordagem de modelagem orientada por dados. Ambiente. Modelo. Softw.2010;25(9):1014–1022.

[35] YuY., ZhuY., LiS. Detecção de valores discrepantes de séries temporais com base na previsão de janela deslizante. Matemática. Problema Eng.2014; 1–14.

[36] L. Ma, X. Gu, B Wang, Correção de outliers em séries temporais de temperatura com base na previsão de janela deslizante na rede de sensores meteorológicos, Informação 8 (2) (2017) 60.

[37] VK Samparthi, HK Verma, detecção Outlier de dados em redes de sensores sem fio usando estimativa de densidade do kernel, Int. J. Computação. Aplic. 5 (7) (2010) 28–32.

[38] G. Tang, K. Wu, J. Lei, Z. Bi, J Tang, Da paisagem ao retrato: uma nova abordagem para detecção de outlier em dados de curva de carga, IEEE Trans. Rede Inteligente 5 (4) (2014) 1764–1773.

[39] HN Akouemo, RJ Povinelli, Detecção e imputação de outliers de séries temporais, Assembleia Geral IEEE PES 2014 | Conferência e Exposição, 2014, pp. 1–5.

[40] YeX., LuZ., QiaoY., MinY., O'MalleyM. Identificação e correção de outliers em dados de energia de séries temporais de parques eólicos. Trans. IEEE Pancada. Syst.2016;31(6):4197–4205.

[41] HN Akouemo, RJ Povinelli, Detecção de anomalias probabilísticas em dados de séries temporais de gás natural , Int. J. Previsão. 32 (3) (2016) 948-956.

[42] Harlé F., Chatelain F., Gouy-Pailler C., AchardS. Modelo Bayesiano para detecção de múltiplos pontos de mudança em séries temporais multivariadas. Trans. IEEE Processo de Sinal. 2016;64(16):4351–4362.

[43] AkouemoH.N., PovinelliR.J.Data melhorando em séries temporais usando modelos ARX e ANN. Trans. IEEE Pancada. Syst.2017;32(5):3352–3359.

[44] HuberP.J. Estimativa robusta de um parâmetro de localização. Ana. Matemática. Estado. 1964;35:73101.

[45] PJ Huber, EM Ronchetti, Robust Statistics, Wiley, Nova York, 2009.

[46] J. Yang, J Stenzel, Correção da curva de carga histórica para previsão de carga de curto prazo, 2005 International Power Engineering Conference, 2005, pp. 1–40.

[47] J. Chen, W. Li, A. Lau, J. Cao, K Wang, limpeza de dados de curva de carga automatizada em sistemas de energia, IEEE Trans. Rede Inteligente 1 (2) (2010) 213–221.

[48] A. Zhang, W. Zhong, W Liu, Detecção de padrões discrepantes em registros de chamadas com base em pontos de esqueleto, 2010 Conferência Internacional sobre Inteligência Artificial e Inteligência Computacional, 2 2010, pp. 145–149.

[49] GuoZ., LiW., LauA., Inga-RojasT., WangK. Detecção de X-outliers em dados de curva de carga em sistemas de potência. Trans. IEEE Pancada. Syst.2011;27(2):875–884.

[50] Mateos G., Giannakis G.B. Regressão não paramétrica robusta via controle de esparsidade com aplicação para limpeza de dados de curva de carga. Trans. IEEE Processo de Sinal. 2011;60(4):1571–1584.

[51] L. Fang, M. Zhi-zhong, Um método de detecção de outlier online para tempo de controle de processo série, 2011 Conferência Chinesa de Controle e Decisão (CCDC), 2011, pp. 3263–3267.

[52] KKLB Adikaram, MA Hussein, M. Effenberger, T Becker, detecção de valores discrepantes método de regressão linear baseado na soma da progressão aritmética, Sci. Mundo J. (2014).

[53] LUO Jian, HONG Tao, YUE Meng, detecção de anomalias em tempo real para previsão de carga de muito curto prazo, J. Mod. Pancada. Sistema Energia Limpa 6 (2) (2018) 235–243.

[54] C. Jeenaunta, KD Abeyrathna, MS Dilhani, SW Hnin, PP Phyo, Série temporal

detecção de outlier para previsão de demanda de carga de eletricidade de curto prazo, Int. Sci. J. Eng. Tecnologia (ISJET) 2 (1) (2018) 37–50.

[55] Y. Lin, J Wang, autoencoder profundo probabilístico para detecção e reconstrução de valores discrepantes de medição de sistema de energia, IEEE Trans. Rede Inteligente 11 (2) (2020) 1796–1798.

[56] M. Breunig, H. Kriegel, R. Ng, J Sander, LOF: Identificando out liers locais baseados em densidade, Anais da Conferência Internacional 2000 ACM SIGMOD sobre gerenciamento de dados, Dallas, 29 2000, pp. 93– 104.

[57] J. Tang, Z. Chen, AWC Fu, DW Cheung, Aumentando a eficácia do outlier de proteções para padrões de baixa densidade, Conferência Pacífico-Ásia sobre Descoberta de Conhecimento e Mineração de Dados, 2002, pp. 535–548.

[58] S. Papadimitriou, H. Kitagawa, PB Gibbons, C Faloutsos, LOCI: fast outlier de proteção usando a integral de correlação local, Proceedings of the 19th International Conference on Data Engineering, Los Alamitos, EUA, IEEE Computer Society Press, 2003, pp. 315–326.

[59] W. Jin, AKH Tung, J. Han, W Wang, Ranking de outliers usando vizinhança simétrica relacionamento de borhood, Proceedings of the 10th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining (PAKDD'06), 2006, pp. 577–593.

[60] M. Bai, X. Wang, J. Xin, G Wang, Um algoritmo eficiente para detecção de valores discrepantes baseados em densidade distribuída em big data, Neurocomputing 181 (2016) 19–28.

[61] D. Ren, B. Wang, W Perrizo, RDF: um método de detecção de valores discrepantes baseados em densidade usando representação vertical de dados, Conferência Internacional sobre Mineração de Dados (ICDM'04), 2004, pp. 503-506.

[62] LJ Latecki, A. Lazarevic, D Pokrajac, Detecção de valores discrepantes com funções de densidade do kernel , Anais da 5ª Conferência Internacional sobre Aprendizado de Máquina e Mineração de Dados em Reconhecimento de Padrões, 2007, pp. 61–75.

[63] J. Gao, W. Hu, ZM Zhang, X. Zhang, O Wu, RKOF: detecção de outlier local baseada em kernel robusto, Conferência Pacífico-Ásia sobre Descoberta de Conhecimento e Mineração de Dados, 2011, pp. 270–283.

[64] M. Pavlidou, G Zioutas, detector de valores discrepantes de densidade do kernel, Tópicos em não paramétricos Estatísticas, Springer, Nova York, NY, 2014, pp. 241–250.

[65] ZhengZ., JeongH.Y., HuangT., ShuaJ. Detecção de outliers baseada em Kde em fluxos de dados distribuídos em rede multimídia. Multim. Ferramenta Appl.2016;1–19.

[66] L. Zhang, J. Lin, R Karim, detecção de anomalia baseada em densidade de kernel adaptável para sistemas não lineares, Knowl. Sistema baseado. 139 (2018) 50-63.

[67] X. Qin, L. Cao, EA Rundensteiner, S Madden, estimativa de densidade do kernel escalável com base na detecção de valores discrepantes locais em grandes fluxos de dados, EDBT (2019) 421–432.

[68] H. Kriegel, P. Kröger, E. Schubert, A Zimek, LoOP: probabilidades discrepantes locais, Anais da 18ª Conferência da ACM sobre Gestão da Informação e do Conhecimento (CIKM'09), Hong Kong, China, 2009, pp. 1649–1652.

[69] H. Fan, OR Za¨yane, A. Foss, J Wu, Fator outlier baseado em resolução: detectando os n pontos de dados mais distantes em dados de engenharia, Knowl. Inf. Sistema 19 (1) (2009) 31–51.

[70] R. Momtaz, N. Mohssen, MA Gowayyed, DWOF: um outlier robusto baseado em densidade abordagem de detecção, Conferência Ibérica sobre Reconhecimento de Padrões e Análise de Imagem, 2013, pp. 517–525.

[71] B. Tang, H He, Uma abordagem baseada em densidade local para detecção de valores discrepantes, Neurocomputação 241 (2017) 171-180.

[72] T Jin, Pesquisa sobre método de modelagem reversa para unidade de energia térmica, Escola de Energia, Energia e Engenharia Mecânica, Pequim, North China Eleetric Power University, 2011Vol. Doutor.

[73] M. Qi, Z. Fu, F Chen, método de detecção de valores atípicos de múltiplos pontos de medição de parâmetros em unidades de usinas de energia, Appl. Term. Eng. 85 (2015) 297-303.

[74] IM Cecílio, JR Ottewill, J. Pretlove, NF Thornhill, Método dos vizinhos mais próximos para detectar distúrbios transitórios em processos e sistemas eletromecânicos, J. Controle de Processo 24 (9) (2014) 1382–1393.

[75] L. Cai, NF Thornhill, S. Kuenzel, BC Pal, Detecção em tempo real de distúrbios do sistema de energia com base na análise do vizinho mais próximo $ K $, IEEE Access 5 (2017) 5631-5639.

[76] Y. Djenouri, A. Belhadi, JCW Lin, A Cano, k-vizinhos mais próximos adaptados para detectando anomalias no fluxo de tráfego espaço-temporal, IEEE Access 7 (2019) 10015–10027.

[77] S. Mohseni, A Fakharzade, Uma nova abordagem de detecção de outlier baseada em distância local para dados fuzzy por métrica de vértice, 2015 2ª Conferência Internacional sobre Engenharia e Inovação Baseada em Conhecimento (KBEI), 2015, pp. 551–554.

[78] Y. Chen, L Tu, agrupamento baseado em densidade para dados de fluxo em tempo real, os anais da 13ª conferência internacional ACM SIGKDD sobre descoberta de conhecimento e mineração de dados. KDD '07, Nova York, NY, EUA, 2007, pp. 133–142.

[79] Al-ZoubiBelalM. Uma abordagem eficaz baseada em agrupamento para detecção de valores discrepantes. EUR. J. Sci. Res.2009;28(2):310–316.

[80] P. Yang, B Huang, algoritmo de detecção de valores discrepantes baseado em KNN em grande conjunto de dados, 2008 Workshop Internacional sobre Tecnologia Educacional e Treinamento e Workshop Internacional de 2008 sobre Geociência e Sensoriamento Remoto, 1 2008, pp. 611–613.

[81] S. Xu, C. Hu, L. Wang, G Zhang, Máquinas vetoriais de suporte baseadas no algoritmo K vizinho mais próximo para detecção de outliers em RSSFs, 2012 8th International Conference on Wireless Communications, Networking and Mobile Computing, 2012, pp. 1–4.

[82] H. Rizk, S. Elgokhy, AM Sarhan, Um algoritmo híbrido de detecção de valores discrepantes baseado em particionamento de clustering e medidas de densidade, 2015 Décima Conferência Internacional sobre Engenharia de Computação e Sistemas (ICCES), 2015, pp. 175–181.

[83] X. Wang, XL Wang, Y. Ma, DM Wilkes, um outlier rápido baseado em kNN inspirado em MST método de detecção, Inf. Sistema 48 (2015) 89-112.

[84] J. Liu, E Zio, KNN-FSVM para detecção de falhas em trens de alta velocidade, 2018 IEEE Conferência Internacional sobre Prognóstico e Gestão da Saúde (ICPHM), 2018, pp. 1–7.

[85] FT Liu, KM Ting, ZH Zhou, Isolation forest, 2008 Oitava IEEE International Conference on Data Mining, 2018, pp. 413–422.

[86] LinZ., LiuX., ColluM. Previsão de energia eólica com base em dados SCADA de alta frequência, juntamente com floresta de isolamento e redes neurais de aprendizado profundo. Int. J. Eleito. Pancada. Energy Syst.2020;118- 105835.

[87] ShenL., DuH., LiuS., ChenS., QiaoL., LiuS., LiJ. Monitoramento de outlier em tempo real para diagnóstico de falhas em transformadores de potência com base em floresta isolada. Em IOP Conf. Série Mate. Sci. Eng.2020;715(1):012033.

[88] HaririS., KindM.C., BrunnerR.J. Floresta de isolamento estendida. 2018;pré-impressão arXiv arXiv:1811.02141.

[89] P. Karczmarek, A. Kiersztyn, W. Pedrycz, E Al, floresta de isolamento baseada em K-Means, Knowl. Sistema baseado. (2020) 105659.

[90] ChaithanyaP.S., PriyangaS., PravinrajS., SriramV.S.SSO-IF: an Outlier Detection Abordagem para Detecção de Intrusão em Sistemas SCADA. Em Inventar. Comum. Computar. Tech.2020;921-929.

[91] J. Ren, R Ma, fluxos de dados baseados em densidade agrupados em janelas deslizantes, Conferência Internacional sobre Sistemas Fuzzy e Descoberta de Conhecimento (FSKD), 5 2009, pp. 248–252.

[92] WeekleyR.A., GoodrichR.K., CornmanL.B. Um algoritmo para classificação e detecção de out lier de dados de séries temporais. J. Atmos. Oceano. Tech.2010;27(1):94–107.

[93] ToshniwalD.A estrutura para detecção de outliers em fluxos de dados em evolução ponderando atributos em clustering. Procedia Tech.2012;6:214–222.

[94] Dados de geração fotovoltaica horária. [Conectados]. Disponível: <https://www.pvoutput.org>.

[95] Consumo de carga horária. [Conectados]. Disponível:<https://openei.org/datasets/files/961/pub>.

[96] Dados de carga horária. [Conectados]. Disponível: <http://www.ercot.com/gridinfo/load/carga_hist>.

[97] Dados de temperatura horária. [Conectados]. Disponível:<https://maps.nrel.gov/nsrdb visualizador>.

[98] KG Ranjan, BR Prusty, D Jena, Comparação de dois métodos de limpeza de dados como aplicado a séries temporais voláteis, 2019 International Conference on Power Electronics Applications and Technology in Present Energy Scenario (PETPES), Mangalore, Índia, 2019, pp. 1–6.