Explorando causalidade na seleção de variáveis para previsão de séries temporais multivariadas

Patrícia O. Lucas
Laboratório Machine Intelligence and Data Science (MINDS)
Programa de Pós-Graduação em Engenharia Elétrica
Universidade Federal de Minas Gerais (UFMG)
Av. Antônio Carlos 6627, 31270-901, Belo Horizonte, Brasil
Instituto Federal do Norte de Minas Gerais (IFNMG)
Salinas, Brasil
0000-0002-7334-8863

Eduardo M. A. M. Mendes
Departamento de Eletrônica
Universidade Federal de Minas Gerais (UFMG)
Av. Antônio Carlos 6627, 31270-901, Belo Horizonte, Brasil
0000-0002-3267-3862

Frederico G. Guimarães
Laboratório Machine Intelligence and Data Science (MINDS)
Departamento de Engenharia Elétrica
Universidade Federal de Minas Gerais (UFMG)
Av. Antônio Carlos 6627, 31270-901, Belo Horizonte, Brasil
0000-0001-9238-8839

***Resumo*—A previsão de séries temporais desempenha um papel fundamental em diversos setores, fornecendo insights valiosos para o planejamento e tomada de decisões em várias áreas de aplicação. A seleção de características em séries temporais é, por sua vez, um componente crucial no *pipeline* de aprendizado de máquina. Ao selecionar cuidadosamente um conjunto apropriado de variáveis, os modelos de aprendizado de máquina podem aumentar a precisão da previsão, reduzir o tempo de processamento e melhorar a interpretabilidade. Neste estudo, comparamos o método de descoberta causal PCMCI (*Peter and Clark Momentary Conditional Independence*) com outros métodos de seleção de características não causais, como Correlação, Algoritmos Genéticos e regularização LASSO. Três conjuntos de dados do mundo real são usados nos experimentos de acordo com três resultados principais: tamanho do gráfico gerado (número de variáveis selecionadas), tempo de execução e precisão da previsão em diferentes horizontes de 1, 3 e 5 passos à frente. Os resultados mostram a capacidade do PCMCI em gerar modelos de previsão mais enxutos, proporcionando maior interpretabilidade e reduzindo a suscetibilidade ao sobre-ajuste.**

***Palavras-chaves*—Séries temporais, previsão, seleção de características, descoberta causal, PCMCI.**

I. Introdução

A previsão de séries temporais desempenha um papel crucial em vários campos, fornecendo informações valiosas para o planejamento e a tomada de decisão no domínio da aplicação. Ao entender padrões históricos, organizações podem antecipar tendências futuras, identificar anomalias e tomar decisões baseadas em dados para melhorar a eficiência, a lucratividade e o desempenho geral em diversos setores.

Um modelo de previsão utiliza as observações passadas (*lags*) para extrair conhecimento de uma ou mais séries temporais, a fim de fornecer previsões. Devido à estrutura temporal desses dados, as séries temporais são consideradas dados de alta dimensão. Nas séries temporais multivariadas, onde várias variáveis são observadas simultaneamente, cada variável adicional contribui para o aumento da dimensão dos dados. No entanto, lidar com dados de alta dimensão continua sendo um desafio, pois requer um alto custo computacional e tempo de processamento [1].

A seleção de características, também conhecida como seleção de *lags*, desempenha um papel fundamental no pipeline de aprendizado de máquina para previsão de séries temporais. Essa etapa tem como objetivo reduzir a dimensionalidade dos dados, removendo características irrelevantes e redundantes. Ao realizar uma seleção apropriada de *lags*, é possível aprimorar a precisão dos modelos, reduzir o tempo de treinamento e obter modelos mais simples [2], [3].

Os métodos de seleção de características podem ser divididos em três categorias principais: filtro, *wrapper* e embutido [2]. Na análise de séries temporais, esses três métodos, bem como os híbridos, foram explorados.

Métodos de filtro são frequentemente usados em trabalhos baseados em correlação para seleção de características. Por exemplo, em [4], um método baseado em correlação foi usado para prever os preços do petróleo bruto. Da mesma forma, em [5], a correlação foi empregada na previsão de carga de trabalho para data centers, enquanto [3] testou quatro métodos correlacionais na previsão de carga elétrica.

A seleção de características por meio de métodos *wrapper* é tratada como um problema de otimização. Em [6], uma abordagem de computação evolutiva multi-objetivo foi aplicada para selecionar características em séries temporais multivariadas. Em [7], a seleção de características foi incorporada à otimização de hiperparâmetros do modelo Fuzzy Times Series (FTS) usando algoritmos genéticos. Outros trabalhos como [8]

e [9] propuseram um método de seleção híbrido que combina as melhores características dos métodos *wrapper* e filtro.

Em relação aos métodos embutidos, o *Random Forest* (RF) e o método de regressão e regularização LASSO foram utilizados em [10] e [11], respectivamente, para realizar a seleção de características em séries temporais.

Com o crescente volume de dados de alta dimensão, os métodos de seleção de filtro ganharam destaque devido à sua rápida velocidade de processamento, independência de modelos de previsão e capacidade de lidar com o sobre-ajuste (*overfitting*) [12]. Além disso, a seleção de recursos causais tornou-se um tópico de pesquisa emergente em aprendizado de máquina [13]. Esses métodos podem fornecer uma compreensão mais profunda do processo subjacente de geração de dados, o que pode melhorar o poder explicativo e a robustez dos modelos de previsão [14].

O método de Causalidade de Granger é amplamente utilizado na análise de séries temporais para estimar associações causais defasadas no tempo usando modelos autorregressivos [13]. Ele se baseia na ideia de que uma série temporal X causa outra série temporal Y se a inclusão do histórico de X melhora a previsão de Y em comparação com um modelo que usa apenas o histórico de Y. Exemplos de trabalhos que usam métodos baseados em Causalidade de Granger para seleção de características em séries temporais são: [15], [16] e [17].

Outro método é o *Peter and Clark Momentary Conditional Independence* (PCMCI) [13] que se baseia na inferência causal. Ele combina testes de independência condicional com um algoritmo de descoberta causal para estimar redes causais a partir de conjuntos de dados de séries temporais com resultados promissores [18]–[20]. Este método demonstrou um poder de detecção significativamente maior em comparação com outros métodos, como o algoritmo LASSO, *Peter and Clark* (PC) e Causalidade de Granger [13]. Em [21], por exemplo, os autores compararam o método baseado em Causalidade de Granger com o PCMCI em dados sintéticos de séries temporais multivariadas variando de 6 a 81 variáveis e argumentaram que o PCMCI pode produzir resultados mais precisos. Além disso, em [22] o PCMCI com métodos de seleção de características não causais (correlação, seleção aleatória e regressão *Random Forest*) foram comparados para prever ciclones tropicais. Os resultados também indicaram que o PCMCI selecionou um número reduzido de variáveis em relação aos outros modelos.

Assim, neste artigo o método PCMCI é explorado para seleção de características em séries temporais multivariadas. Comparamos o PCMCI com outros métodos de seleção de características não causais, incluindo filtros baseados em correlação, *wrappers* com algoritmos genéticos e métodos embutidos com LASSO. Os resultados são analisados em termos de (i) tamanho do grafo gerado (número de características selecionadas), (ii) tempo de execução e (iii) precisão da previsão em 1, 3 e 5 passos à frente. Três conjuntos de dados do mundo real são usados para validar os resultados.

O restante deste trabalho está organizado da seguinte forma. A Seção 2 fornece uma breve revisão sobre seleção de características e descoberta causal em séries temporais. Na Seção 3, descrevemos a metodologia de comparação proposta. A seção 4 descreve os experimentos, resultados e discussões. Por fim, a Seção 5 contém algumas conclusões e possibilidades para trabalhos futuros.

II. Seleção de características

A seleção de características é um processo essencial na construção de modelos de previsão, no qual um subconjunto representativo e relevante de características é selecionado. Esse subconjunto deve ser parcimonioso, ao mesmo tempo em que oferece o melhor desempenho para o modelo [1], [3]. Esse processo tornou-se indispensável em aplicações que lidam com dados de alta dimensão, como séries temporais multivariadas, pois permite reduzir a dimensionalidade dos dados e eliminar características irrelevantes e redundantes. Um conjunto de características menor resulta em treinamento mais rápido e menor complexidade do modelo de previsão [2].

No entanto, encontrar o melhor subconjunto de características não é uma tarefa simples devido aos extensos espaços de busca envolvidos. Nos modelos de previsão de séries temporais, valores defasados são utilizados como entrada, o que significa que a dimensionalidade do problema aumenta à medida que mais dados históricos são considerados, especialmente em séries temporais multivariadas.

Os métodos de seleção de características podem ser categorizados em: filtro, *wrapper* e embutidos [23]. Eles são descritos abaixo.

- Filtro: classifica cada característica com base em alguma medida estatística e, em seguida, seleciona-as com base em um limiar. Esses métodos são independentes dos modelos de previsão e, portanto, computacionalmente menos exigentes.
- Wrapper: as características são classificadas de acordo com a precisão fornecida pelo modelo de previsão. Esses métodos tendem a ter um desempenho melhor, uma vez que levam em consideração a hipótese do modelo treinado no espaço de características. No entanto, eles são computacionalmente mais caros [1].
- Embutido: a seleção de características é realizada como parte do procedimento de aprendizado, o que os torna específicos para um único modelo.

A. Descoberta causal

Um grafo causal $G = (V, E)$ é uma representação gráfica que descreve as relações causais de um sistema. Em outras palavras, uma aresta direcionada $X \longrightarrow Y$ estabelece uma relação de causa e efeito entre a causa X e seu efeito Y. Dado um conjunto $\mathcal{G}$ que consiste em todos os grafos definidos pelas variáveis V de um conjunto de dados D, o problema da descoberta causal é encontrar o grafo G^* que seja uma possível explicação para D [24].

Quando um grafo satisfaz a suposição de fidelidade[1], o cobertor ou envoltória de Markov (*Markov Blanket*) (MB) de

[1]A suposição de fidelidade requer que todas as independências condicionais observadas surjam da estrutura gráfica causal [13].

uma variável X no grafo é único e composto por seus pais (causas diretas), filhos (efeitos diretos) e cônjuges (outros pais dos filhos). De acordo com a Condição Causal de Markov [25], todas as outras variáveis que não pertencem ao MB são probabilisticamente independentes quando a variável em questão é condicionada em seu MB. Portanto, o MB de X é o subconjunto ideal mínimo de recursos de X a ser usado na seleção de características [12].

Considerando o problema de seleção de *lags* em séries temporais, o MB de uma variável é reduzido apenas aos seus pais, já que valores no tempo t não podem causar valores no tempo $t-1$. Ou seja, o condicionamento apenas nos pais de uma variável é suficiente para estabelecer independência condicional e identificar links espúrios [13].

Em teoria, os subconjuntos de características obtidos por métodos de seleção causal estão mais próximos do MB de uma variável em comparação com métodos não causais. Como resultado, eles podem melhorar a capacidade explicativa dos modelos de previsão e torná-los mais robustos. Isso ocorre porque os relacionamentos causais implicam no mecanismo subjacente dos dados, o que faz com que esses relacionamentos sejam persistentes em diferentes ambientes [14].

B. Descoberta causal em séries temporais multivariadas

Nesta seção será discutido o método PCMCI desenvolvido por [13]. O PCMCI é um método de descoberta causal que gera um grafo causal diretamente de séries temporais multivariadas. O PCMCI consiste de duas etapas: primeiro, usa o algoritmo PC_1 para identificar os pais $\hat{\mathcal{P}}(X_t^j)$ para todas as variáveis da série temporal $X_t^j \in X_t^1, \cdots, X_t^N$, e segundo, aplica o teste de independência condicional momentânea (MCI) para testar ligações indiretas $X_{t-\tau}^i \longrightarrow X_t^j$.

O PC_1 é um algoritmo que usa testes iterativos de independência para descoberta de conjuntos de Markov. Para cada variável X_t^j é feita a inicialização dos pais preliminares $\hat{\mathcal{P}}(X_t^j) = (X_{t-1}, X_{t-2}, \cdots, X_{t-\tau_max})$. Primeiro, testes de independência incondicionais são aplicados para remover $X_{t-\tau}^i$ de $\hat{\mathcal{P}}(X_t^j)$, caso a hipótese nula $X_{t-\tau}^i \perp\!\!\!\perp X_t^j$ não for rejeitada a um nível de significância α_{PC}. Os pais preliminares são classificados por seu valor absoluto estatístico de teste.

A Figura 1 ilustra o PC_1 para duas variáveis X_1 e X_3 onde a intensidade da cor representa o valor absoluto estatístico de teste das variáveis dependentes (cores escuras indicam valores maiores). Os nós cinza representam as variáveis independentes. Em seguida, testes de independência condicional $X_{t-\tau}^i \perp\!\!\!\perp X_t^j | \mathcal{L}$ são executados, onde $\mathcal{L}$ são os pais mais fortes em $\hat{\mathcal{P}}(X_t^j) \setminus X_{t-\tau}^i$ e os pais independentes são removidos de $\hat{\mathcal{P}}(X_t^j)$. Dessa forma, o PC_1 converge para apenas algumas condições relevantes que incluem os pais causais com alta probabilidade (rosa escuro/azul escuro) e potencialmente alguns falsos positivos (setas pontilhadas).

Na segunda etapa, o MCI usa os pais estimados pelo PC_1 para identificar causas indiretas. No exemplo da Figura 1, as condições $\hat{\mathcal{P}}(X_t^3)$ são suficientes para estabelecer a independência condicional para testar $X_{t-2}^1 \longrightarrow X_t^3$. Além disso, os pais defasados de $\hat{\mathcal{P}}(X_1^{t-2})$ são inseridos como condições adicionais e são responsáveis por manter a taxa de falsos positivos a um nível esperado.

O método PCMCI baseia-se nas suposições de suficiência causal[2], na Condição Causal de Markov e na suposição de fidelidade. Também não assume efeitos causais contemporâneos e assume estacionariedade [13]. O PCMCI possui complexidade polinomial no número de variáveis N e τ_{max}. No pior caso, onde o grafo está completamente conectado, a complexidade computacional do estágio de seleção de condição PC_1 para N variáveis equivale a $N^3\tau_{max}^2$. A etapa MCI envolve ainda testes $N^2\tau_{max}$ (para $\tau > 0$). Portanto, o pior caso de complexidade computacional total no número de variáveis é polinomial e dado por $N^3\tau_{max}^2 + N^2\tau_{max}$.

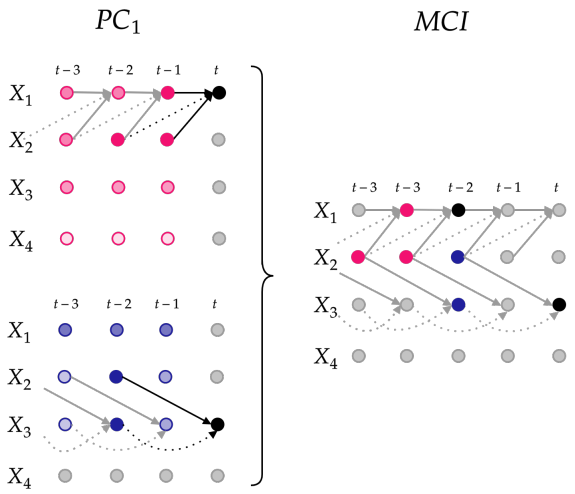

Figura 1. Ilustração do método PCMCI.

III. METODOLOGIA

Nesta seção apresentamos a metodologia usada para explorar o método de descoberta causal PCMCI e compará-lo com outros métodos na seleção de *lags* de variáveis de séries temporais multivariadas.

A Figura 2 ilustra, de modo geral, o fluxo de processos realizados nesse estudo. Os algoritmos foram implementados usando uma interface *Multiple Input Single Output* (MISO), onde múltiplas entradas indica o número de variáveis que compõem a série temporal, ou simplesmente (X_1, X_2, X_3). A saída é representada como X_1, a ser predita. A série temporal multivariada é usada para seleção de características. Com os *lags* selecionados, um grafo é gerado e usado para organizar os dados de entrada para o modelo de previsão Perceptron Multicamadas (MLP). O método *Grid Search* foi usado para otimização de hiperparâmetros (OHP) do modelo MLP.

Os hiperparâmetros do PCMCI são: α_{PC}, τ_{max} e o teste de independência condicional. O α_{PC} assume o papel de um parâmetro de regularização. Usando níveis muito altos, como

[2]Implicando que todos os fatores comuns estão entre as variáveis observadas.

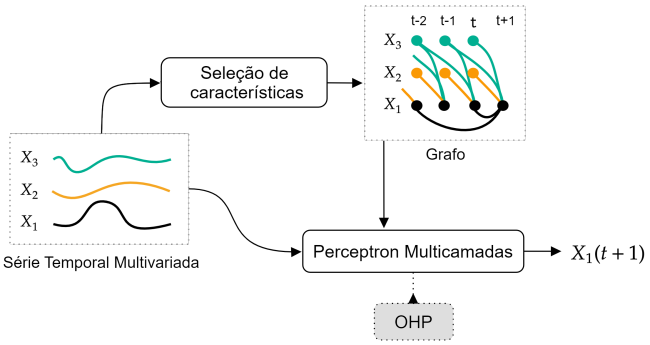

Figura 2. Ilustração do fluxo de processos realizados na metodologia de comparação dos métodos de seleção de características.

$\alpha_{PC} = 1$, nenhum pai é removido. Isso significa que todas as variáveis são selecionadas como pais. Neste estudo, foi usado $\alpha_{PC} = 0.1$, valor otimizado usando para seleção o Critério de Informação de Akaike descrito em [13]. O τ_{max} é o *lag* máximo que será usado na geração do grafo. Levando-se em consideração o custo computacional e o número de defasagens significativamente autocorrelacionadas das séries temporais em estudo, foi utilizado um $\tau_{max} = 10$. O teste de independência condicional escolhido foi o linear.

O método baseado em correlação, chamado aqui de Correlacional, foi implementado usando autocorrelação e correlação cruzada, selecionando os *lags* com correlação significativamente diferente de 0 com nível de confiança de 99%. Para medir a significância da correlação cruzada, aplicou-se um modelo substituto.

Para melhor compreensão, veja o exemplo da Figura 2 com X_1 e X_2: primeiro, a correlação cruzada para cada ponto de X_1 e X_2 ($C_{1,2}$) foi calculada. Como a ideia é verificar se $C_{1,2}$ é diferente de 0, usou-se X_1 para gerar 1000 séries permutadas diferentes, mas preservando suas propriedades estatísticas individuais e removendo qualquer correlação cruzada entre elas. Em seguida, calculou-se a correlação cruzada de X_1 com cada uma das séries permutadas gerando uma série temporal de distribuições de correlações (D). Perceba que a média dessas distribuições deve ser 0. Por fim, um teste estatístico foi aplicado para verificar se $C_{1,2}$ é significativamente diferente de 0 com nível de confiança de 99% usando D.

Como abordagem *wrapper*, implementou-se o algoritmo genético (GA). Um GA opera sobre um conjunto de vetores (cromossomos) de soluções, denominado população. Cada cromossomo (indivíduo) é composto por *lags* selecionados aleatoriamente. O GA apresenta 3 operadores principais para melhorar os indivíduos da população inicial: (1) operador que seleciona indivíduos promissores para participarem do cruzamento (seleção), (2) operador que combina soluções para encontrar indivíduos mais bem adaptados (cruzamento) e (3) operador que cria perturbações em alguns indivíduos para aumentar a exploração do espaço de busca e escapar de ótimos locais (mutação) [26]. Para o AG, foi utilizado o seguinte: validação cruzada K-*fold* com $k = 5$, tamanho da população = 80, probabilidade de cruzamento = 0.5, probabilidade de mutação = 0.2 e número de gerações = 100. Os indivíduos foram avaliados usando regressão linear.

O LASSO é uma técnica de regularização que adiciona o valor absoluto dos coeficientes dos pesos como termo de penalidade [11]. Esse modelo realiza seleção de características automaticamente, gerando vários coeficientes com peso zero, ou seja, que são ignorados pelo modelo. O LASSO foi implementado usando regressão linear com constante α que multiplica o termo L_1 igual a 0.1. O grafo foi gerado selecionando-se os *lags* cujos coeficientes fossem maiores que zero.

Ambos GA e LASSO, são avaliados em termos de RMSE normalizado (NRMSE) usando um modelo diferente daquele usado para gerar o grafo. O principal objetivo é investigar se as características selecionadas por esses métodos, que são dependentes de modelos, podem ser aplicadas com sucesso em modelos diferentes. O NRMSE foi calculado como na equação (1), onde y_{max} e y_{max} são os valores máximo e mínimo do conjunto de dados de teste.

$$NRMSE = \frac{\sqrt{\sum_{t=0}^{n}(y(t) - \hat{y}(t))^2)}}{y_{max} - y_{min}} \quad (1)$$

IV. Experimentos

O objetivo dos experimentos foi avaliar o desempenho do método de descoberta causal PCMCI na seleção de *lags* para previsão de séries temporais multivariadas. O desempenho foi medido através do tempo de execução, tamanho do grafo gerado (número nós) e da média do NRMSE retornado pelo modelo MLP.

Para realizar uma análise comparativa, usou-se três bases de dados das áreas financeira, aplicações IoT (internet das coisas) e clima. Os testes Augmented Dickey-Fuller (ADF) e Kwiatkowski-Phillips-Schmidt-Shin (KPSS) foram empregados com nível de confiança de 95% para determinar estacionariedade. A diferenciação foi aplicada em casos de comportamento não estacionário.

Os dados foram divididos em duas partes: a primeira para seleção de características (SC) e OHP e a segunda para treinamento e teste do modelo MLP. A normalização no intervalo [0,1] foi executada antes da aplicação do algoritmo MLP. Em seguida, foi dividido em 5 janelas para avaliar o desempenho do modelo em diferentes segmentos da série temporal. Detalhes sobre os dados são mostrados na Tabela I.

Tabela I
Descrição e configuração dos experimentos para cada base de dados.

		Número de amostras		
Base de dados	**Variáveis**	**SC / OHP**	**Treino**	**Teste**
DOW JONES	6	4000	1500	500
CASA	28	10000	1900	500
EVAPOTRANSPIRAÇÃO	7	3500	1000	500

A. Bases de dados

A base de dados DOW JONES[3] contém o índice médio diário que reflete aproximadamente a situação do mercado financeiro dos EUA, compreendendo uma combinação dos 30 títulos mais importantes do mercado de ações. Além do índice médio, outras variáveis dessa base de dados são os preços de abertura, alta, baixa, fechamento e volume para cada dia útil de 1985 a 2017.

CASA[4] é um conjunto de dados de previsão de consumo de energia de eletrodomésticos. O dados incluem medições de temperatura e umidade coletadas por uma rede de sensores sem fio (WSN) de várias áreas de uma casa de baixo consumo de energia na Bélgica, informações meteorológicas de uma estação meteorológica próxima e uso de energia registrado de aparelhos e luminárias. Os dados dos aparelhos de energia foram obtidos medindo continuamente (a cada 10 minutos) por 137 dias.

A base de dados EVAPOTRANSPIRAÇÃO foi usada para previsão de evapotranspiração de referência no Brasil. Os dados foram extraídos de [27] nas coordenadas: latitude: $-19.46°C$, longitude: $-44.25°C$ e cobrem o período de 2000 a 2019 de dados diários referente a temperatura máxima, temperatura mínima, radiação solar, umidade relativa, velocidade do vento, precipitação e evapotranspiração de referência.

B. Experimentos computacionais

Todos os experimentos foram implementados e testados com Python 3 usando os pacotes open source Scikit-Learn [28] e Tigramite[5].

Para promover a transparência e a reprodutibilidade dos resultados do teste, o código-fonte e os conjuntos de dados estão disponíveis em https://bit.ly/CausalFeatureSelection.

C. Resultados

Nesta seção são apresentados os resultados dos experimentos que comparam os métodos PCMCI, Correlacional, GA e LASSO.

Os resultados da Tabela II e Tabela III correspondem à etapa de seleção de recursos e fornecem informações sobre o tamanho dos grafos gerados pelos modelos, bem como o tempo de execução necessário para gerá-los .

Tabela II
TAMANHO DO GRAFO GERADO PELOS MÉTODOS PCMCI, CORRELACIONAL, GA, E LASSO PARA TODAS AS BASES DE DADOS DO ESTUDO.

Bases de dados	PCMCI	Correlacional	GA	Lasso
DOW JONES	09	59	29	31
CASA	75	183	126	05
EVAPOTRANSP.	8	66	28	11

A respeito do tamanho do grafo, o método PCMCI gerou os menores grafos, com exceção da base de dados CASA, onde o LASSO apresenta um grafo significativamente menor em relação a todos os métodos. Percebe-se que o método Correlacional gerou os maiores grafos para todos as bases de dados. O GA, apesar de ter gerado um grafo menor que o LASSO em DOW JONES, nas outras bases de dados apresentou grafos bem maiores que o LASSO e PCMCI.

Tabela III
TEMPO DE EXECUÇÃO EM SEGUNDOS DA GERAÇÃO DOS GRAFOS PELOS MÉTODOS PCMCI, CORRELACIONAL, GA, E LASSO PARA TODAS AS BASES DE DADOS DO ESTUDO.

Bases de dados	PCMCI	Correlacional	GA	Lasso
DOW JONES	6.1	1.8	71.8	10.4
CASA	313.9	7.2	1542.4	208.0
EVAPOTRANSP.	15.9	1.6	58.8	10.5

Em termos de tempo de execução, o método Correlacional apresenta tempos bem abaixo que os demais métodos. O PCMCI e LASSO apresentam tempos mais próximos, com o LASSO obtendo tempos um pouco melhores nas bases de dados CASA e EVAPOTRANSPIRAÇÃO. Já o GA apresenta os piores tempos de execução.

Por fim, a Tabela IV apresenta os resultados da média do NRMSE (%) e o desvio padrão da etapa de previsão de 1, 3 e 5 passos à frente. É possível observar que para o conjunto de dados DOW JONES o MLP apresenta uma diferença significativa apenas na previsão de 1 passo à frente. O GA apresentou maior NRMSE. Para o conjunto de dados CASA, os resultados são mais variados. No geral, PCMCI e GA têm valores médios de NRMSE mais baixos. Esses resultados são seguidos por LASSO e Correlacional com valores médios de NRMSE mais altos. Os resultados para o conjunto de dados EVAPOTRANSPIRAÇÃO exibem um padrão semelhante ao DOW JONES, com médias NRMSE variando apenas na previsão de 1 passo à frente, e LASSO mostrando o pior resultado, enquanto os outros métodos têm médias estatisticamente semelhantes.

D. Discussão

O objetivo dos experimentos foi demonstrar o desempenho do método de descoberta causal PCMCI para seleção de características em três séries temporais multivariadas: financeira (DOW JONES), aplicação IoT (CASA) e climática (EVAPOTRANSPIRAÇÃO). Os resultados são comparados com outros métodos não causais analisando três resultados principais: o tamanho do grafo gerado, o tempo de execução e a precisão das previsões de 1, 3 e 5 passos à frente.

Ao apresentar os resultados do tamanho do grafo gerado pelos métodos, observou-se que o PCMCI consegue gerar conjuntos de características significativamente menores que os demais métodos testados. Esse resultado corrobora com os resultados apresentados em [22]. Nesse estudo os autores avaliaram a seleção de características para previsão da intensidade de ciclones tropicais. Percebe-se também que as característica selecionadas pelo PCMCI são relevantes ao apresentar NRMSE estatisticamente iguais aos demais modelos. Esse resultado mostra a capacidade do PCMCI de gerar modelos

[3]Disponível em: https://github.com/PYFTS/pyFTS/tree/master/pyFTS/data

[4]Disponível em: https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction

[5]https://jakobrunge.github.io/tigramite/

Tabela IV
NRMSE (%) PARA HORIZONTES DE PREVISÃO DE 1, 3 E 5 DAS BASES DE DADOS EM ESTUDO.

Bases de dados	Horizonte de previsão	PCMCI	Correlacional	GA	Lasso
DOW JONES	**1**	1.48 ± 0.81	1.71 ± 1.05	2.40 ± 1.30	5.84 ± 1.61
	3	13.51 ± 2.42	13.60 ± 2.79	13.66 ± 2.76	13.28 ± 2.84
	5	13.42 ± 2.44	13.87 ± 2.74	13.73 ± 2.56	13.29 ± 2.77
CASA	**1**	0.99 ± 0.76	1.71 ± 0.84	1.70 ± 0.66	18.50 ± 3.34
	3	12.87 ± 3.75	19.49 ± 5.06	13.45 ± 3.38	17.90 ± 3.34
	5	17.68 ± 2.82	23.14 ± 3.11	17.02 ± 3.10	17.49 ± 3.54
EVAPOTRANSPIRAÇÃO	**1**	0.34 ± 0.19	1.59 ± 0.28	0.93 ± 0.10	4.35 ± 0.93
	3	16.56 ± 1.42	16.25 ± 1.52	15.71 ± 1.34	15.50 ± 1.20
	5	17.16 ± 1.73	16.95 ± 1.90	16.28 ± 1.26	15.83 ± 1.26

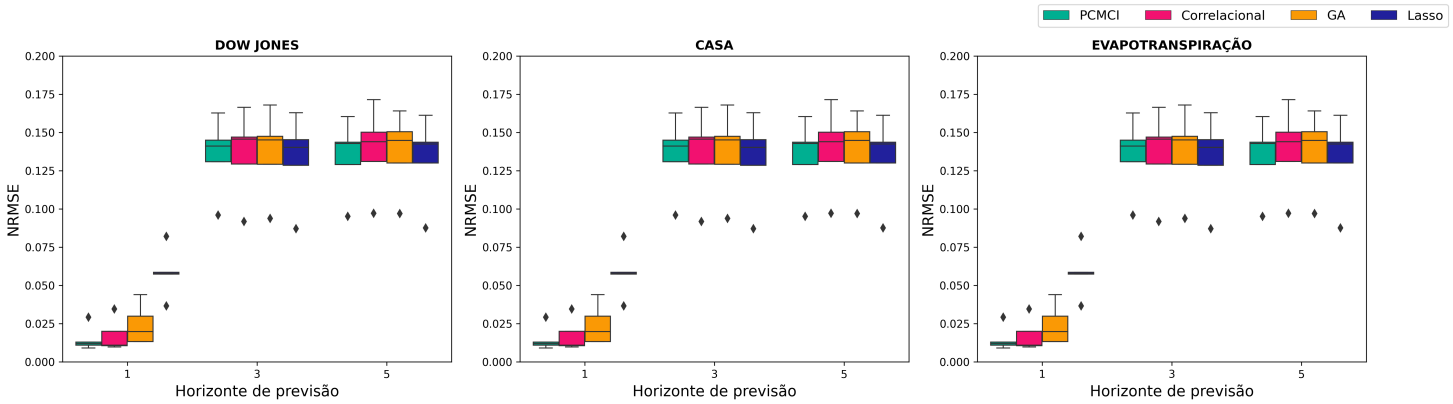

Figura 3. NRMSE para horizontes de previsão de 1, 3 e 5 usando os métodos PCMCI, Correlacional, GA e Lasso para as bases de dados analisadas.

de previsão mais parcimoniosos e consequentemente mais interpretáveis e menos suscetíveis a *overfitting*.

Vale ressaltar também que o LASSO gerou um conjunto muito reduzido de características no dataset CASA. No entanto, o NRMSE apresentado foi estatisticamente maior que dos demais modelos. O PCMCI mostrou também que, para esse contexto de datasets com até 28 variáveis, o tempo de execução é aceitável já que ficou muito próximo ao tempo de execução do método LASSO usando regressão linear.

Por fim, percebeu-se que um conjunto maior de características não gerou modelos mais precisos e no caso do método Correlacional, piorou os resultados na base de dados CASA. Isso pode indicar a criação de relações espúrias sobre os dados.

V. CONCLUSÃO

O objetivo deste estudo foi investigar o método de descoberta causal PCMCI para a seleção de características em séries temporais multivariadas, comparando-o com métodos não causais. Os resultados indicaram que o PCMCI foi capaz de gerar um conjunto de características até 8 vezes menor do que os demais métodos, sem comprometer a precisão do modelo em diferentes horizontes de previsão. Isso demonstra a capacidade competitiva do PCMCI em relação às abordagens mais conhecidas na área de seleção de características em séries temporais, como métodos baseados em correlação. O uso do PCMCI possibilita a geração de modelos mais enxutos e compreensíveis nos *pipelines* de previsão.

Para pesquisas futuras, seria interessante explorar a robustez desse método examinando o grafo causal gerado e verificando se a precisão do modelo permanece consistente em diferentes ambientes. Por exemplo, nos casos em que as relações causais estabelecidas pelo método realmente implicam o mecanismo subjacente que gerou os dados. Além disso, seria válido testar o método com outros modelos de previsão e em conjuntos de dados adicionais do mundo real para validar sua eficácia em cenários mais amplos.

AGRADECIMENTOS

Este trabalho foi financiado pelas agências brasileiras (i) Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Grant no. 312991/2020-7 e 310788/2021-8; (ii) Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) por meio do Programa de Excelência Acadêmica (PROEX) e (iii) Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), Auxílio nº. APQ-01779-21. Laboratório MINDS – https://minds.eng.ufmg.br/.

REFERÊNCIAS

[1] Z. M. Hira and D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data," *Advances in Bioinformatics*, vol. 2015, 2015.

[2] R. Sheikhpour, M. A. Sarram, S. Gharaghani, and M. A. Z. Chahooki, "A Survey on semi-supervised feature selection methods," *Pattern Recognition*, vol. 64, pp. 141–158, 4 2017.

[3] I. Koprinska, M. Rana, and V. G. Agelidis, "Correlation and instance based feature selection for electricity load forecasting," *Knowledge-Based Systems*, vol. 82, pp. 29–40, 7 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0950705115000714

[4] S. Karasu, A. Altan, S. Bekiros, and W. Ahmad, "A new forecasting model with wrapper-based feature selection approach using multi-objective optimization technique for chaotic crude oil time series," *Energy*, vol. 212, p. 118750, 12 2020.

[5] J. Xue, F. Yan, R. Birke, L. Y. Chen, T. Scherer, and E. Smirni, "PRACTISE: Robust prediction of data center time series," *Proceedings of the 11th International Conference on Network and Service Management, CNSM 2015*, pp. 126–134, 12 2015.

[6] F. Jiménez, J. Palma, G. Sánchez, D. Marín, M. D. Francisco Palacios, and M. D. Lucía López, "Feature selection based multivariate time series forecasting: An application to antibiotic resistance outbreaks prediction," *Artificial Intelligence in Medicine*, vol. 104, p. 101818, 4 2020.

[7] P. C. L. Silva, P. d. O. e Lucas, H. J. Sadaei, and F. G. Guimaraes, "Distributed Evolutionary Hyperparameter Optimization for Fuzzy Time Series," *IEEE Transactions on Network and Service Management*, pp. 1–1, 3 2020.

[8] Z. Hu, Y. Bao, T. Xiong, and R. Chiong, "Hybrid filter–wrapper feature selection for short-term load forecasting," *Engineering Applications of Artificial Intelligence*, vol. 40, pp. 17–27, 4 2015.

[9] T. Niu, J. Wang, H. Lu, W. Yang, and P. Du, "Developing a deep learning framework with two-stage feature selection for multivariate financial time series forecasting," *Expert Systems with Applications*, vol. 148, p. 113237, 6 2020.

[10] J. Guo, H. Sun, and B. Du, "Multivariable Time Series Forecasting for Urban Water Demand Based on Temporal Convolutional Network Combining Random Forest Feature Selection and Discrete Wavelet Transform," *Water Resources Management*, vol. 36, no. 9, pp. 3385–3400, 7 2022. [Online]. Available: https://link.springer.com/article/10.1007/s11269-022-03207-z

[11] G. R. Nitta, B. Y. Rao, T. Sravani, N. Ramakrishiah, and M. BalaAnand, "LASSO-based feature selection and naïve Bayes classifier for crime prediction and its type," *Service Oriented Computing and Applications*, vol. 13, no. 3, pp. 187–197, 9 2019. [Online]. Available: https://link.springer.com/article/10.1007/s11761-018-0251-3

[12] K. Yu, L. Liu, and J. Li, "A Unified View of Causal and Non-causal Feature Selection," 2018.

[13] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, "Detecting and quantifying causal associations in large nonlinear time series datasets," *Science Advances*, vol. 5, no. 11, pp. 4996–5023, 11 2019. [Online]. Available: https://www.science.org/doi/10.1126/sciadv.aau4996

[14] K. Yu, X. Guo, L. Liu, J. Li, H. Wang, Z. Ling, and X. Wu, "Causality-based Feature Selection," *ACM Computing Surveys (CSUR)*, vol. 53, no. 5, 9 2020. [Online]. Available: https://dl.acm.org/doi/10.1145/3409382

[15] Y. Sun, J. Li, J. Liu, C. Chow, B. Sun, and R. Wang, "Using causal discovery for feature selection in multivariate numerical time series," *Machine Learning 2014 101:1*, vol. 101, no. 1, pp. 377–395, 7 2014. [Online]. Available: https://link.springer.com/article/10.1007/s10994-014-5460-1

[16] Y. Hmamouche, A. Casali, and L. Lakhal, "A Causality Based Feature Selection Approach for Multivariate Time Series Forecasting," p. 1467523, 5 2017. [Online]. Available: https://hal.science/hal-01467523https://hal.science/hal-01467523/document

[17] M. Dong and Y. Kluger, "GEASS: Neural causal feature selection for high-dimensional biological data," in *The Eleventh International Conference on Learning Representations*, 2 2023. [Online]. Available: https://openreview.net/forum?id=aKcS3xojnwY

[18] M. Noorbakhsh, C. Connaughton, and F. A. Rodrigues, "Discovering causal factors of drought in Ethiopia," *ACM International Conference Proceeding Series*, pp. 72–78, 9 2020. [Online]. Available: https://arxiv.org/abs/2009.07955v1

[19] S. Uereyen, F. Bachofer, and C. Kuenzer, "A Framework for Multivariate Analysis of Land Surface Dynamics and Driving Variables—A Case Study for Indo-Gangetic River Basins," *Remote Sensing*, vol. 14, no. 1, p. 197, 1 2022. [Online]. Available: https://www.mdpi.com/2072-4292/14/1/197/htmhttps://www.mdpi.com/2072-4292/14/1/197

[20] M. Rademaker, I. M. Smallegange, and A. Van Leeuwen, "Causal links between North Sea fish biomass trends and seabed structure," *Marine Ecology - Progress Series*, vol. 677, pp. 129–140, 10 2021. [Online]. Available: https://doi.org/10.3354/meps13845

[21] S. Peterson, "Comparison of Lasso Granger and PCMCI for Causal Feature Selection in Multivariate Time Series," Ph.D. dissertation, The University of Arizona, 2022. [Online]. Available: http://hdl.handle.net/10150/665005

[22] S. G. S., T. Beucler, F. I.-H. Tam, M. S. Gomez, J. Runge, and A. Gerhardus, "Selecting Robust Features for Machine Learning Applications using Multidata Causal Discovery," 4 2023. [Online]. Available: https://arxiv.org/abs/2304.05294v3

[23] X. He, K. Zhao, and X. Chu, "AutoML: A Survey of the State-of-the-Art," *Knowledge-Based Systems*, vol. 212, 8 2021. [Online]. Available: http://arxiv.org/abs/1908.00709http://dx.doi.org/10.1016/j.knosys.2020.106622

[24] A. Zanga, E. Ozkirimli, and F. Stella, "A Survey on Causal Discovery: Theory and Practice," *International Journal of Approximate Reasoning*, vol. 151, pp. 101–129, 12 2022.

[25] F. Eberhardt, "Introduction to the foundations of causal discovery," *International Journal of Data Science and Analytics*, vol. 3, no. 2, pp. 81–91, 3 2017. [Online]. Available: https://link.springer.com/article/10.1007/s41060-016-0038-6

[26] G. Wu, R. Mallipeddi, and P. N. Suganthan, "Ensemble strategies for population-based optimization algorithms – A survey," *Swarm and Evolutionary Computation*, vol. 44, pp. 695–711, 2 2019.

[27] A. C. Xavier, C. W. King, and B. R. Scanlon, "Daily gridded meteorological variables in Brazil (1980-2013)," *International Journal of Climatology*, vol. 36, no. 6, pp. 2644–2659, 10 2015. [Online]. Available: https://doi.org/10.1002/joc.4518

[28] F. Pedregosa, V. Michel, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, J. Vanderplas, D. Cournapeau, F. Pedregosa, G. Varoquaux, A. Gramfort, B. Thirion, O. Grisel, V. Dubourg, A. Passos, M. Brucher, M. Perrot andEdouardand, a. Duchesnay, and F. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [Online]. Available: http://scikit-learn.sourceforge.net.