Revista Brasileira de Geografia Física

ISSN:1984-2295

Homepage:https://periodicos.ufpe.br/revistas/rbgfe

Uso do Algoritmo PARAFAC-EM para Preenchimento de Falhas em Séries Temporais de Velocidade do Vento na Região Nordeste do Brasil

Emerson Mariano da Silva[1], Israel Luís Rodrigues Pinto[2]

[1]Mestrado Profissional em Climatologia/Universidade Estadual do Ceará (UECE)/Fortaleza-Ceará/Brasil, emerson.mariano@uece.br, ORCID: https://orcid.org/0000-0002-9131-9820; [2]israelluisrpinto@gmail.com, ORCID: https://orcid.org/0009-0003-7389-6311.

Artigo recebido em 07/04/2024 e aceito em 30/10/2024

R E S U M O

Apresenta-se neste estudo a avaliação do desempenho de um método de preenchimento de falhas em séries temporais de dados meteorológicos, em particular para séries de dados de velocidade do vento, usando-se o algoritmo PARAFAC-EM. Foram usados dados de médias horárias de velocidade do vento obtidos das plataformas de coleta de dados dos municípios de Acopiara e Sobral no Estado do Ceará/Brasil, totalizando 8.640 registros para cada ano em cada região. Em seguida foram induzidas falhas nas séries de dados de 10%, 30% e 50%, que foram preenchidas com os modelos PARAFAC-EM a fim de determinar o *rank* de melhor resultado. O desempenho do método proposto foi validado através do cálculo de métricas estatísticas e aplicação do teste t de Student. Os resultados obtidos mostraram que o método proposto com *rank* = 2 apresentou o melhor desempenho em estimar as falhas nas séries de dados, com correlação estatística classificadas como moderadas no preenchimento das falhas de 10%, 30% e 50% para a região de Acopiara/CE, e de fortes a muito forte para a região de Sobral/CE, com nível de significância da de 99%.

Palavras-chave: Preenchimento de falhas, velocidade do vento, Algoritmo PARAFAC-EM.

Use of the PARAFAC-EM Algorithm for Data Imputation in Wind Speed Time Series in the Northeast Region of Brazil

A B S T R A C T

This study evaluates the performance of a method for filling gaps in meteorological data time series, for wind speed data series, using the PARAFAC-EM algorithm. Hourly average wind speed data obtained from data collection platforms in the municipalities of Acopiara and Sobral in the state of Ceará/Brazil were used, totalling 8,640 records for each year in each region. Next, gaps were induced in the data series of 10%, 30% and 50%, which were filled in with PARAFAC-EM models to determine the rank with the best result. The performance of the proposed method was validated by calculating statistical metrics and applying Student's t-test. The results obtained showed that the proposed method with rank = 2 had the best performance in estimating the gaps in the data series, with statistical correlation classified as moderate in filling the gaps of 10%, 30% and 50% for the Acopiara/CE region, and from strong to very strong for the Sobral/CE region, with a significance level of 99%.

Keywords: Data imputation, wind speed, PARAFAC-EM algorithm.

Introdução

As séries temporais de dados meteorológicos têm grande importância para o entendimento e a modelagem adequada das variáveis atmosféricas. A obtenção destas informações em superfície é realizada por instrumentos convencionais ou por equipamentos eletrônicos, com frequência de medição que depende do uso dos dados. Assim, um problema encontrado nas séries de dados meteorológicos são as falhas, falta de informações que podem ser devido a erros de medição ou por problemas nos instrumentos usados para as medições destas variáveis (Bonfante et al., 2013; Giovanella et al., 2021; Ruezzene et al., 2021; Weerakody et al., 2021; Miao et al., 2021).

Estudos publicados na literatura afirmam que a ocorrência de falhas em leituras de variáveis meteorológicas em estações de superfície pode comprometer a consistência das séries históricas, inviabilizando ou prejudicando sua utilização para

compreensão e modelagem da variabilidade espaço temporal destas variáveis (Brubacher et al., 2020; Giovanella et al., 2021; Ahn et al., 2022).

Weerakody et al. (2021) afirmam que as falhas em séries temporais são cada vez mais frequentes nos sistemas multi-sensores, e que essas falhas limitam seriamente a capacidade de análise destes dados.

Assim, determinar o método adequado para preencher as falhas de uma série de dados meteorológicos pode ser uma questão bastante delicada, pois a utilização de métodos inadequados pode levar a conclusões erradas sobre o conjunto de dados a ser usados nas análises (Rubin, 1996; Afrifa-Yamoah et al., 2020; Sciscenko et al., 2022; Cai et al., 2023).

Desta forma, é importante que as falhas encontradas nas séries de dados das variáveis meteorológicas sejam preenchidas com valores próximos ao que seria medido para proporcionar a obtenção de análises mais confiáveis.

Realizar o preenchimento das falhas de uma série de dados consiste em estimar valores através do comportamento de uma variável com base em históricos de observações. A partir dos anos 80 surgiram diversas técnicas estatísticas para imputar valores satisfatórios aos dados ausentes nestas séries observadas (Nunes et al., 2009; Bonfante et al., 2013).

Estas técnicas de preenchimento de falhas podem ser aplicadas nos dados meteorológicos, tais como a utilização da média estatística dos dados observados antes da falha, contudo, o uso deste método não permite que a série de dados represente completamente a realidade das medições, pois as falhas podem ser contínuas e uso deste método requer que se tenha dados próximos as falhas para preenchê-las.

Outro método amplamente usado para enfrentar este problema é o emprego de regressão linear, neste método o preenchimento das falhas é realizado através da equação de uma reta ajustada ao conjunto dos dados a ser preenchido e ao conjunto dos dados da variável explicativa (Oliveira et al., 2010). No entanto, para que esse método obtenha resultados satisfatórios é necessário que a variável explicativa tenha correlação estatística classificada como forte com a série de dados a ser preenchida as falhas (Lira et al., 2011; Lira et al., 2014).

Nos últimos anos tem-se observado o uso de redes neurais e de inteligência artificial para cumprir esta tarefa de preenchimento de falhas, com resultados promissores (Giovanella et al., 2021; Ruezzene et al., 2021; Weerakody et al., 2021; Miao et al., 2021).

Nesse contexto, considerando a importância desta temática para os estudos e análises das séries das variáveis meteorológicas, observa-se a necessidade da avaliação de métodos que possam auxiliar a realização do preenchimento de falhas em séries de dados meteorológicos, em particular neste estudo é apresentado um método como alternativa para o preenchimento de falhas em séries temporais de dados de velocidade do vento que são usadas nas análises diagnósticas e previsão do potencial eólico em diversas regiões.

Daí, o objetivo deste estudo é avaliar o desempenho do uso do Algoritmo PARAFAC-EM para o preenchimento de falhas em séries de velocidade do vento obtidas nas Plataformas de Coletas de Dados (PCD) da Fundação Cearense de Meteorologia e Recursos Hídricos (FUNCEME) nos municípios de Acopiara e Sobral no estado do Ceará (Brasil).

Preenchimento de falhas em séries temporais

Zhang (2003) denomina de imputação de dados o processo de preenchimento de falhas em séries temporais, que consiste em substituir os dados que faltam em uma série temporal por valores satisfatórios que proporcionem a aplicação de métodos estatísticos tradicionais de análises.

O autor menciona que a utilização da imputação de dados em séries temporais com falhas evita a complexidade causada pela presença da falta desses valores, no entanto, a imputação desses valores insere incertezas que devem ser consideradas adequadamente, pois os dados imputados não são os observados e medidos. Assim destaca três fontes destas incertezas: (a) as incertezas associadas com a modelagem da

distribuição conjunta das variáveis respostas com as indicadoras de falhas, (b) as incertezas associadas com a amostragem dos dados a serem imputados tendo conhecimento dos dados observados e dos parâmetros do modelo de imputação e (c) as incertezas associadas aos valores dos parâmetros do modelo de imputação.

Em Donders et al. (2006) os métodos de imputação foram classificados em imputação simples e múltipla. Essa classificação é frequentemente encontrada nos artigos publicados sobre essa temática nos trabalhos de Baraldi e Enders (2010) e Ramli et al. (2013). Segundo estes autores a imputação simples consiste na aplicação de um conjunto de técnicas em que o valor em falta é imputado por um único valor aparentemente adequado para essa reposição, como por exemplo a imputação pela média dos valores da série temporal ou dos valores vizinhos, a imputação por regressão linear e a imputação por regressão estocástica.

Segundo Junninen et al. (2004) o desempenho do método de imputação de valores não depende apenas da quantidade de falhas (valores ausentes ou incoerentes), depende também das características dos mecanismos geradores de falhas, mecanismos que foram classificados por Donders et al. (2006) em: ausentes completamente aleatórios (Missing Completely at Random – MCAR), ausentes não aleatórios (missing not at random – MNAR) e ausentes aleatórios (missing at random – MAR).

Segundo Camargos et al. (2011) o termo mecanismo e´ usado nestes estudos no contexto como sinônimo de função de distribuição de probabilidades dos dados ausentes. Os autores citam como exemplo as falhas que em geral, ocorrem em pesquisas com séries temporais de dados do nível de renda das famílias (Donders et al., 2006) e em pesquisas sobre o Índice de Massa Corporal – IMC.

O problema das falhas nas séries temporais de dados meteorológicos é recorrente e exige esforços científicos para enfrentá-lo, assim, vários métodos de imputação de dados tem sido testado para se determinar a melhor aplicação para cada uma das variáveis atmosféricas usadas em análises e aplicações em diversas áreas do conhecimento.

Encontra-se na literatura a aplicação de diversos métodos matemáticos e estatísticos, classificados como usuais para o preenchimento de falhas em séries temporais, em particular nas séries históricas de precipitação e de evapotranspiração, que são destinadas aos estudos na área de climatologia e de recursos hídricos (Afrifa-Yamoah et al., 2020; Brubacher et al., 2020; Zhang e Thorburn, 2021; Corbo et al., 2024; Tavares et al. (2024).

As redes neurais e a inteligência artificial também são ferramentas usados para cumprir esta tarefa, encontram-se na literatura diversos estudos de aplicação também em séries temporais de variáveis climáticas (Giovanella et al., 2021; Ruezzene et al., 2021; Weerakody et al., 2021; Miao et al., 2021; Silva Junior et al., 2024).

Em adição, se encontram na literatura estudos que usam a decomposição tensorial (PARAFAC) como métodos para preenchimentos de falhas em séries temporais de variáveis ambientais (Ooba et al.; 2006; Bonfante et al., 2013; Turova et al., 2021; Sciscenko et al., 2022; Cai et al., 2023).

Assim, ressalta-se a importância e os esforços científicos dispensados a resolução dos problemas associados a presença de falhas nas series temporais em diversas áreas do conhecimento, em particular para os estudos climatológicos, onde se posiciona os estudos dos regimes de vento, objeto deste estudo.

Tensores e Decomposição Tensorial PARAFAC

Um tensor é definido em matemática como uma matriz multidimensional de dados em N modos, que podem ser unidimensionais (vectores), bidimensionais (matrizes), tridimensionais e N-dimensionais, sendo os dois últimos mais frequentemente designados por tensores (kolda e bader, 2009).

A figura a seguir mostra uma representação pictórica de um tensor tridimensional, ou um tensor de modo-3 com dimensões I ×J×K (Kolda e Bader, 2009).

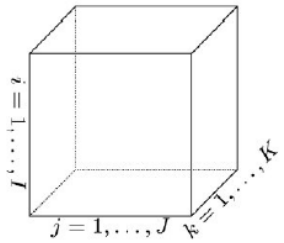

Figura 1 - Representação pictórica de um tensor de modo-3 com dimensões *i*×*j*×*k* (Kolda e Bader, 2009).

As decomposições tensoriais são originalmente atribuídas a Hitchcock (1927) e a ideia de modelo multimodo é atribuída a Cattel (1944). No entanto, estes conceitos receberam pouca atenção até ao trabalho de Tucker na década de 1960 e de Carroll, Chang e Harshman na década de 1970, todos trabalhos associados ao domínio da psicometria.

Adicionalmente, de acordo com Kolda e Bader (2009), autores como Appellof e Davidson são creditados como os primeiros a utilizar decomposições tensoriais em quimiometria, por volta do ano de 1981.

Durante muito tempo, as aplicações das decomposições tensoriais estiveram praticamente restritas aos domínios da psicometria e da quimiometria, sendo populares nestes domínios ao ponto de ter sido publicado um livro em 2004. No entanto, desde o final da década de 1990, as áreas de aplicação das decomposições tensoriais diversificaram-se significativamente. Novas áreas de aplicação que podem ser mencionadas incluem o processamento de sinais, a visão computacional, a neurociência, a extração de dados, entre outras.

A decomposição PARAFAC (*Parallel Fator Analysis*) descreve o tensor X, modo-3, de dimensões *i*×*j*×*k* como uma soma de uma quantidade finita *R* de tensores de modo-1, como mostra a Figura 2.

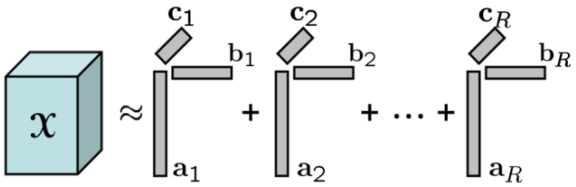

Figura 2 - Representação pictórica da Decomposição PARAFAC (Kolda e Bader, 2009).

A quantidade finita de tensores de modo-1 que geram o tensor a partir da Decomposição PARAFAC, denotada por R, é chamada de modo do tensor e, no contexto da análise tensorial, é a quantidade mínima de tensores de modo-1 necessária para gerar o tensor X (Kolda e Bader, 2009) e pode ser determinada de acordo com a seguinte relação (equação 1).

$$rank(\mathcal{X}) \leq \min\{iI, jJ, kK\} \tag{1}$$

Em linguagem matemática, na Decomposição PARAFAC, os elementos x_{ijk} do tensor podem ser escritos pela equação escalar abaixo:

$$x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} \tag{2}$$

Em que: *i*=1,...,I, *j*=1,...,J e *k*=1,...,K, e os valores a_{ir}, b_{jr} e c_{kr} são os elementos das matrizes fatoriais A (*I*×*R*), B (*J*×*R*) e C (*K*×*R*) que precisam ser estimados por meio de um algoritmo (Tomasi e Bro, 2006).

Um dos algoritmos utilizados para calcular as matrizes de fatores da decomposição PARAFAC é o método dos mínimos quadrados alternados (ALS). O algoritmo PARAFAC-ALS estima iterativamente essas matrizes de modo a minimizar o erro ou a função de perda.

$$f\left(A^{(s)}, B^{(s)}, C^{(s)}\right) = \left\| \mathcal{X}^{(s)}_{(I \times JK)} - A^{(s)}(C^{(s)} \diamond B^{(s)}) \right\|_F^2 \tag{3}$$

Em que $\|*\|_F$ denota o Método de Frobenius.

Portanto, o tensor pode ser escrito, nas suas formas de desdobramento, de acordo com as seguintes equações:

$$\mathcal{X}_{(I \times JK)} = A(C \odot B)^T \tag{4}$$

$$\mathcal{X}_{(I \times JK)} = A(C \odot B)^T \tag{5}$$

$$\mathcal{X}_{(K \times IJ)} = C(B \odot A)^T \tag{6}$$

Em seguida apresentam-se os passos para calcular a decomposição PARAFAC utilizando o algoritmo PARAFAC-ALS:

- ✓ Inicializar as matrizes B e C de forma aleatória;
- ✓ A partir das matrizes B e C dadas, calcular a matriz A utilizando a equação 7.

$$A = \mathcal{X}_{(I \times JK)}(C \odot B)^T(B^TB * C^TC)^\dagger \quad (7)$$

- ✓ A partir de A e de C calcular a matriz B utilizando a equação 8.

$$B = \mathcal{X}_{(J \times IK)}(C \odot A)^T(A^TA * C^TC)^\dagger \quad (8)$$

- ✓ A partir de B e de A calcular a matriz C utilizando a equação 9.

$$C = \mathcal{X}_{(K \times IJ)}(B \odot A)^T(A^TA * B^TB)^\dagger \quad (9)$$

- ✓ Se a equação 2 convergir para o valor mínimo, parar; se não convergir, repetir os passos 2 a 4 até se atingir a convergência.

Em que $x_{(I \times JK)}$, $x_{(J \times IK)}$ e $x_{(K \times IJ)}$ são as formas matriciais do tensor (Tomasi e Bro, 2006). Os símbolos "⊙" e "*" representam respetivamente os produtos de Khatri-Rao e Hadamard, e o símbolo "$A^\dagger$" é a pseudo inversa de Moore-Penrose da matriz A (Kolda e Bader, 2009).

O algoritmo PARAFAC-ALS não pode ser aplicado diretamente a séries de dados incompletos, ou seja, dados com valores omissos (Tomasi e Bro, 2005). Para tornar possível a sua aplicação a séries de dados incompletos, o algoritmo PARFAC-ALS tem de ser modificado.

Para aplicar o PARAFAC-ALS no caso de séries de dados incompletos deve-se imputar e reimputar os valores em falta e em seguida aplicar o PARAFAC-ALS para imputar novamente esses valores em falta, em um processo iterativo até se atingir a convergência (Acar et al., 2010). Neste processo as equações 7, 8 e 9 são aplicadas não sobre o tensor original $\mathcal{X}$, mas sobre o tensor $\widehat{\mathcal{X}}^{(s)}$ definido como:

$$\widehat{\mathcal{X}}^{(s)} = \mathcal{X} * \mathcal{M} + (1 - \mathcal{M}) * \mathcal{Y}^{(s)} \quad (10)$$

Neste caso, $\mathcal{X}$ é o tensor em que os valores em falta são substituídos por zeros, 1 representa um tensor de elementos iguais a 1, $\mathcal{Y}^{(s)}$ é um tensor gerado pelas matrizes A, B e C calculadas na iteração (s - 1) e M é um tensor de ponderação, conforme a equação 11.

$$m_{ijk} = \begin{cases} 0, & se\ x_{ijk}\ for\ falha \\ 1, & se\ x_{ijk}\ for\ dado \end{cases} \quad (11)$$

O procedimento acima é descrito como uma abordagem alternada, classificado como *Expectation Maximization* (*EM*), em que os valores em falta são imputados utilizando os fatores (A, B e C) determinados na iteração (s - 1) (Acar et al., 2010).

Tomasi e Bro (2005) afirmam que este método é concebido com base na estrutura de máxima verossimilhança e é dividido em duas etapas. Na etapa E, a expetativa condicional da função de máxima verossimilhança é calculada através dos dados observados e dos parâmetros estimados, o que equivale ao cálculo da função de perda (Equação 12).

$$f\left(A^{(s)}, B^{(s)}, C^{(s)}\right) = \left\| \hat{x}^{(s)}_{(I \times JK)} - A^{(s)}(C^{(s)} \diamond B^{(s)}) \right\|_F^2 \quad (12)$$

O passo M é o cálculo dos novos parâmetros, que corresponde ao cálculo das matrizes A, B e C utilizando as equações 7 a 9. Assim, o algoritmo PARAFAC-ALS com as modificações mencionadas será aqui designado por PARAFAC-EM, e os passos para a sua aplicação são:

- ✓ Inicializar A, B e C aleatoriamente e calcular $\mathcal{Y}^{(0)}_{(I\ x\ JK)}$ usando a equação 3.
- ✓ Calcular $\widehat{\mathcal{X}}^{(s)}$ usando a equação 10.
- ✓ Decomponha $\widehat{\mathcal{X}}^{(s)}$ para obter A, B e C usando as equações 7 a 9.

- ✓ Calcule $\left|y^{(s)}_{(I\ x\ JK)}\right.$ usando a equação 3.
- ✓ Se a equação 12 convergir para o valor mínimo, pare; se não convergir, repita os passos 2 a 4 até atingir a convergência.

Material e métodos

O Algoritmo PARAFAC-EM foi aplicado para preencher lacunas nas séries de velocidade do vento obtidas das PCDs da FUNCEME instaladas nos municípios de Acopiara e Sobral no Estado do Ceará no Brasil, mostradas na Figura 3.

O critério de dispor de, pelo menos, dois anos de dados completos e sem erros deve ser atendido para que os dados possam ser organizados em uma estrutura tensorial e para que os resultados do método aplicado para preencher os erros induzidos não fossem influenciados pela ausência de dados pré-existentes.

A estrutura tensorial escolhida para organizar os dados de cada PCD foi a de tensores de dimensões 720 ×12 ×2, em que a primeira camada frontal é a matriz dos dados observados no ano 1 (2005) e a segunda camada frontal é a matriz dos dados observados no ano 2 (2006) (Figura 4).

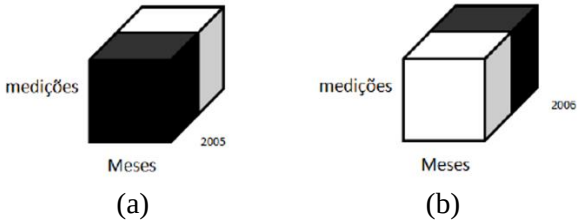

Figura 4 - Base de dados como um tensor de modo-3. (a) A primeira camada frontal corresponde às medições efetuadas em 2005. (b) A segunda camada corresponde às medições efetuadas em 2006.

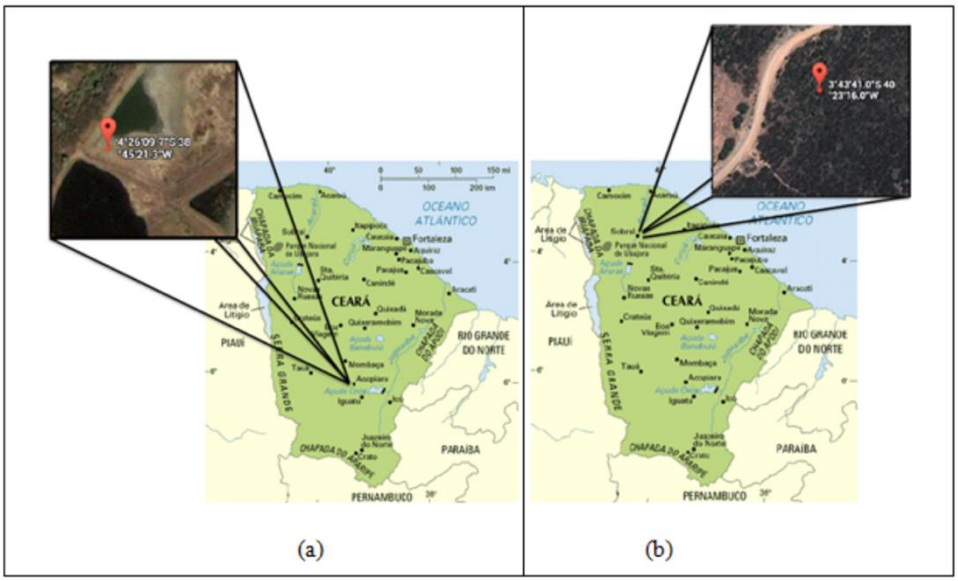

Figura 3 – Mapa do Estado do Ceará (Brasil) com a localização das PCDs dos municípios de (a) Acopiara (04˚26’9,7”; 38˚45’21”) e (b) Sobral (3˚40’58”; 40˚23’16”).

O desempenho do algoritmo PARAFAC-EM no preenchimento de falhas de dados de vento foi avaliado utilizando os parâmetros estatísticos: Coeficiente de Correlação de Pearson (*r*), Índice de Concordância ou Índice de Wilmot (*d*), Erro Absoluto Médio (*MAE*), Erro Não Absoluto Médio (*MNAE*) e Erro Quadrático Médio (*MSE*). Estas métricas estatísticas foram calculadas para cada par de dados retirados/estimados. Além disso, foi aplicado o teste *t* de Student para obter

os níveis de significância estatística (90%, 95% e 99%).

De acordo com Larson e Farber (2010), o coeficiente de correlação é uma medida da relação linear entre duas variáveis (Equação 13).

$$r = \frac{1}{N}\sum_{i=1}^{N}\frac{X_i Y_i}{\sigma_x \sigma_y} \quad (13)$$

Em que *X* e *Y* representam os conjuntos de dados que estão a ser correlacionados, X_i e Y_i são as i-ésimas observações, σ_x e σ_y são os desvios-padrão de *X* e *Y*, e *N* é o número de observações nos conjuntos de dados.

O coeficiente de correlação estatística pode variar entre -1 e 1. Assim, $r = 1$ indica que todos os pares de dados (X_i e Y_i) se encontram numa linha reta com um declive positivo, $r = -1$ indica que todos os pares de dados se encontram numa linha reta com um declive negativo, e $r = 0$ indica que não existe uma relação linear entre os pares de dados ou que não existe qualquer relação entre os dados (Devore, 2006).

A interpretação do coeficiente de correlação baseou-se nas definições dos intervalos apresentados na Tabela 1.

Tabela 1 - Interpretação do coeficiente de correlação de Pearson (r).

Intervalos de valores	Definição
$0.00 \leq \lvert r \rvert \leq 0.19$	Correlação muito fraca
$0.20 \leq \lvert r \rvert \leq 0.39$	Correlação fraca
$0.40 \leq \lvert r \rvert \leq 0.69$	Correlação moderada
$0.70 \leq \lvert r \rvert \leq 0.89$	Correlação forte
$0.90 \leq \lvert r \rvert \leq 1.00$	Correlação muito forte

De acordo com Lima et al. (2012), não basta calcular o coeficiente de correlação, é necessária a aplicação de um teste de significância estatística para se obter o verdadeiro grau de relação entre essas variáveis. Assim, neste estudo foi utilizado o teste *t* de Student, calculado através da equação 14 (Huang e Paes, 2009), que serviu de referência para o cálculo do coeficiente de correlação crítico (r_c), mostrado na equação 15.

$$t_0 = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \quad (14)$$

$$r_c = \sqrt{\frac{t^2}{(N-2)+t^2}} \quad (15)$$

O teste *t* de Student foi aplicado a amostras de 1.728, 5.184 e 8.640 pares de dados retirados/estimados, respetivamente, valores que correspondem a 10%, 30% e 50% do total de dados de cada estação meteorológica usada no estudo.

Os pares de dados observados/estimados resultaram em, respetivamente, N - 2 = 1.726, 5.182 e 8.638 graus de liberdade. Assim, usando uma tabela de pontos de percentil da distribuição t, conforme descrito em Montgomery e Runger (2009), os valores de $t_{\alpha,v}$ e r_c para esses graus de liberdade são:

- ✓ No nível de significância de 99% (erro de 1% - α = 0,01): $t_{\alpha,v}$ = 2,326 e r_c = 0,257;
- ✓ No nível de significância de 95% (erro de 5% - α = 0,05): $t_{\alpha,v}$ = 1,645 e r_c = 0,185;
- ✓ No nível de significância de 90% (erro de 10% - α = 0,1): $t_{\alpha,v}$ = 1,282 e r_c = 0,143.

De acordo com Montgomery e Runger (2009), o valor de t_0 é utilizado para decidir se se rejeita a hipótese nula (H_0) de que a correlação entre as variáveis é nula. Assim, se $|t_0| > t_{\alpha,v}$, a hipótese nula é rejeitada, e conclui-se que o coeficiente de correlação tem significância estatística.

Adicionalmente, para determinar o nível de significância de r, este deve ser comparado com o rc calculado para cada nível de significância; ou seja, se *r* for maior que 0,257, terá uma significância estatística de 99%; se for menor que 0,257 e maior que 0,185, terá um nível de significância de 95%; e se for menor que 0,185 e maior que 0,143, terá uma significância estatística de 90% (Lima et al., 2012).

A variabilidade dos dados estimados em relação aos dados observados foi obtida através de

medidas estatísticas de dispersão (variância e desvio padrão). Segundo Crespo (2009), essas medidas descrevem a variabilidade entre os valores em torno da média aritmética (Equações 16 e 17).

$$\sigma^2 = \frac{\sum(x_i - \vec{x})^2}{N} \quad (16)$$

$$\sigma = \sqrt{\frac{\sum x_i^2}{N} - \left(\frac{\sum x_i}{N}\right)^2} \quad (17)$$

O índice de concordância de Wilmot (*d*) é um índice que varia de 0 a 1, e quanto mais próximo estiver do valor unitário menor será a amplitude do erro (Righi et al., 2012). O cálculo desse índice é feito através da equação 18, onde *X* e *Y* são, respetivamente, as médias dos conjuntos de dados *X* e *Y*.

$$d = 1 - \frac{\sum_{i=1}^{N}(X_i - Y_i)^2}{\sum_{i=1}^{N}\left(|X_i - \vec{X}| + |Y_i - \vec{Y}|\right)^2} \quad (18)$$

Os índices de concordância obtidos foram classificados com base nos intervalos descritos na Tabela 2.

Tabela 2 - Classificação do índice de concordância de Wilmot (d).

Intervalos de valores	Definição
< 0.4	Péssimo
0.41 a 0.5	Ruim
0.51 a 0.6	Razoável
0.61 a 0.65	Mediano
0.66 a 0.75	Bom
0.76 a 0.85	Muito bom
> 0.85	Ótimo

O *MAE* é uma média dos erros absolutos entre os valores estimados (Y_i) e os valores observados (X_i), pode ser utilizado como uma medida de tendência, sendo uma média simples dos erros entre os valores estimados e observados (DEVORE, 2006). Os cálculos desses erros e do *MSE* são obtidos através das equações 19, 20 e 21. Em que *N* é o número total de pares de valores observados e estimados.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|Y_i - X_i| \quad (19)$$

$$MNAE = \frac{1}{N}\sum_{i=1}^{N}(Y_i - X_i) \quad (20)$$

$$MSE = \left[\frac{1}{N}\sum_{i=1}^{N}(Y_i - X_i)^2\right]^{\frac{1}{2}} \quad (21)$$

Para determinar o *rank* do algoritmo PARAFAC-EM o método proposto foi aplicado repetidamente na mesma situação, variando apenas o rank, que assumiu todos os valores inteiros no intervalo [2, 24]. Em cada aplicação foram calculados os parâmetros estatísticos descritos anteriormente, e assim, o *rank* foi determinado como o valor que obteve os maiores valores de correlação e concordância.

Procedimento de indução de falhas nas séries de dados.

Para avaliar a qualidade dos dados imputados pelo algoritmo PARAFAC-EM para o preenchimento das falhas nas séries temporais utilizadas neste estudo, foi necessário gerar falhas artificiais (indução), processo que envolve a remoção de dados da série original através de um algoritmo, para que estas possam ser preenchidas e comparadas com os dados imputados pelo algoritmo de preenchimento de falhas.

Assim, neste estudo, as falhas foram consideradas como valores em falta na série (gaps) e assumidas como tendo um padrão de distribuição aleatório ao longo da série temporal, o que significa que assumem um padrão de falhas distribuídas aleatoriamente ao longo do tempo.

O algoritmo de indução de falhas desenvolvido gera falhas na série de dados com base em percentagens de falhas predefinidas, que neste estudo são 10%, 30% e 50%.

Os passos utilizados para induzir estas falhas são:

- ✓ Seleção da percentagem de dados a remover da matriz original;

- ✓ Determinação do número de linhas (I) e colunas (J) da matriz de dados X;
- ✓ Seleção aleatória de um número inteiro i, tal que $1 \leq i \leq I$, e de um número inteiro j, tal que $1 \leq j \leq I$;
- ✓ Remoção do elemento X(i,j) da matriz de dados e atribuição de uma posição com a entrada *NaN*;
- ✓ Se os dados na posição selecionada no passo (3) já tiverem sido removidos, repetir este passo até encontrar dados que ainda não tenham sido removidos;
- ✓ Repetir os passos (3) e (4) até que o número de dados removidos da matriz de dados seja equivalente à percentagem de dados selecionada.

É importante observar que se a percentagem de falhas escolhida corresponder a um número fracionário do total de dados, o algoritmo descrito acima deve remover da matriz de dados um total de dados igual ao número inteiro mais próximo da quantidade de falhas correspondente.

Resultados e discussões

Na Figura 5 são apresentados os gráficos com os índices de correlação e de concordância estatística entre as séries de dados de velocidade do vento obtidas na PCD de Acopiara (Ceará/Brasil), com as falhas induzidas (10%, 30% e 50% do total), e as séries estimadas pelo modelo PARAFAC-EM usado no preenchimento das falhas.

Observa-se que o modelo PARAFAC-EM de *rank* 2 preencheu as falhas induzidas com um conjunto de dados estimados que apresenta correlação estatística moderada ($0,6 \leq r \leq 0,7$) e índice de concordância classificado como muito bom ($0,7 \leq d \leq 0,85$). Assim, modelo de *rank* 2 se mostrou menos sensível às variações na quantidade de falhas induzidas nas séries de dados a serem preenchida.

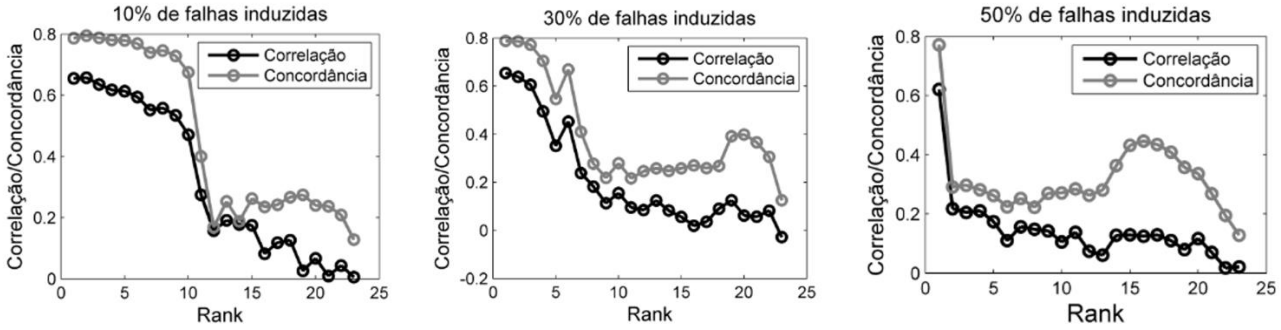

Figura 5 – Correlação e concordância estatística em função do rank para as séries de médias horárias de velocidade do vento com falhas induzidas (10%, 30% e 50%).

De acordo com a Tabela 3, a correlação estatística obtida com o preenchimento da série com 10% de falhas induzidas foi de 0,65 e com 30% de falhas foi de 0,65. Para 50% de falhas induzidas, obteve-se 0,62. O índice de concordância variou entre 0,78 e 0,77 para as séries com 10%, 30% e 50% de falhas induzidas, respectivamente.

Os valores do EMA (Tabela 3) indicam que o módulo da diferença entre os valores das séries de dados com falhas e estimados variou de 0,89 a 0,95, com menor valor (0,89) para a série com 20% de falhas e o maior valor (0,94) para a série de 50% falhas induzidas.

Um ponto importante a ser considerado sobre as métricas estatísticas é que estas medidas estão relacionadas com a unidade de medida da variável em estudo, que neste caso é a velocidade do vento, medida em m/s (metro por segundo). Assim, é possível afirmar que no preenchimento da série com 10% de falhas induzidas o erro estimado é em média de ±0,90 m/s, que para 30% de falhas o erro é em média de ±0,89 m/s e para 50% de falhas induzidas o erro médio é de 0,94 m/s. E, que estes erros representam baixos

valores, considerando que os valores das séries de velocidade do vento da PCD de Acopiara (2005 e 2006) variam de 0 a 8,38 m/s.

Os valores do EMNA (Tabela 3) indicam que no preenchimento das falhas induzidas nas séries de dados (10%, 30% e 50%) o modelo PARAFAC-EM de *rank* 2 estimou dados com média aproximadas da média dos dados retirados de cada série de velocidade do vento.

Tabela 3 – Métricas estatísticas de avaliação do desempenho do modelo PARAFAC-EM de *rank* 2 no preenchimento das falhas induzidas nas séries de médias horárias de velocidade do vento na PCD de Acopiara (Ceará/Brasil) (Total de dados 17.280; *99% de significância – Teste t de Student).

Falhas induzidas		**Métricas Estatísticas**				
Total de dados retirados		***r***	***d***	***EMA***	***EMNA***	***EQM***
10%	1728	0,65*	0,78	0,90	0,04	1,14
30%	5184	0,65*	0,78	0,89	-0,01	1,15
50%	8640	0,62*	0,77	0,94	-0,01	1,21

A Figura 6 (a, b e c) mostram os diagramas de dispersão entre os dados retirados da série (falhas induzidas) e os correspondentes estimados com o modelo PARAFAC-EM (rank 2). Observa-se que os dados estimados para preencher as falhas de 10% e 30% não apresentaram valores menores que o limite inferior da série (zero) e nem maiores que o limite superior da série original (8,38 m/s). No entanto, no preenchimento das falhas de 50% os dados estimados apresentaram valores maiores que o limite superior da série original.

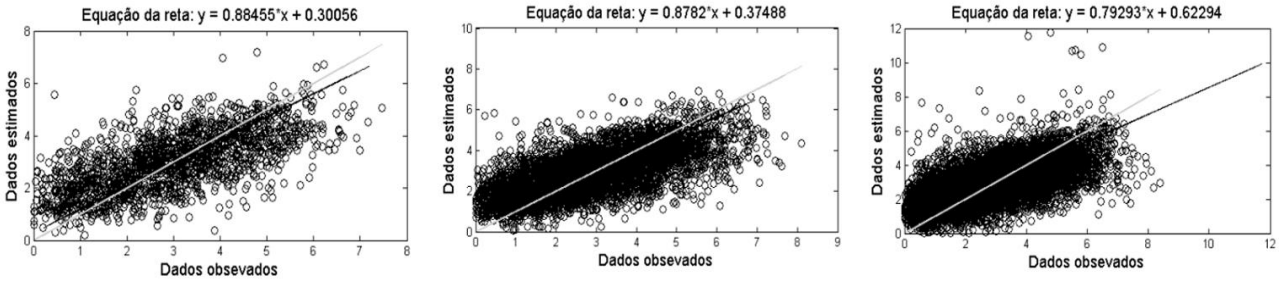

Figura 6 – Dispersão entre os dados retirados (falhas induzidas) e os estimados, (a) 10%, (b) 30% e (c) 50%.

A Tabela 4 mostra a comparação entre os limites superior e inferior dos conjuntos de dados retirados da série e dos conjuntos de dados estimados pelo modelo PARAFAC-EM (*rank* 2) no preenchimento das falhas. A Tabela 5 mostra os valores médios e do desvio-padrão deste conjunto de dados. Esses parâmetros foram usados para avaliar a variabilidade dos dados estimados pelo método e compará-los com a variabilidade dos dados retirados da série.

Os resultados mostrados nas tabelas indicam que a variabilidade dos dados estimados é menor que a variabilidade dos dados retirados da série em todas as porcentagens de falhas preenchidas. Em adição, observa-se que o desvio-padrão dos dados estimados apresenta valores iguais a 1,16 m/s, 1,12 m/s e 1,19 m/s, para as séries com 10%, 30% e 50%, respectivamente. Esses valores indicam que a variabilidade dos dados estimados foi maior para 50% de falhas.

Tabela 4 – Valores limites das séries de dados retirados e simulados.

Falhas induzidas	Dados Retirados		Dados Estimados	
	Mínimo	Máximo	Mínimo	Máximo
10%	0,005	7,46	0,21	7,18
30%	0,005	8,11	0,08	6,93
50%	0,005	8,38	0,003	11,73

Tabela 5 – Valores limites das séries de dados retirados e simulados.

Falhas induzidas	Média		Desvio Padrão	
	Dados Retirados	Dados Estimados	Dados Retirados	Dados Estimados
10%	2,95	3,00	1,50	1,11
30%	2,99	2,97	1,51	1,12
50%	2,98	2,98	1,52	1,19

As Figuras (7, 8 e 9) a seguir mostram as médias diárias de velocidade do vento calculadas a partir das médias horárias com as séries originais (linhas pretas) e preenchida com dados simulados (linhas cinzas). No cálculo destas médias diárias também foram usados os dados que não foram retirados da série original no processo de indução de falhas.

Observa-se que qualitativamente as séries (original e preenchidas) apresentam concordância ao longo tempo. Características confirmadas pelo cálculo das métricas estatísticas (r, d, EMA, EMNA) apresentadas na Tabela 6.

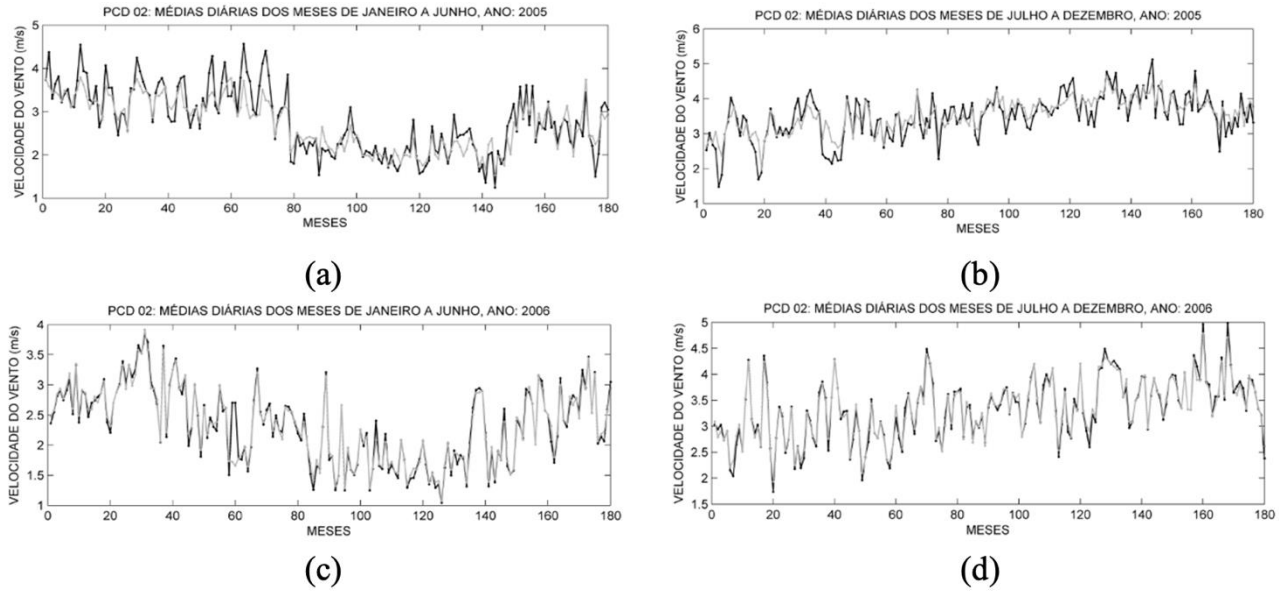

Figura 7 – Médias diárias da velocidade do vento (2005 e 2006) calculadas a partir das séries de velocidade do vento, originais (linhas pretas) e preenchidas com 10% de falhas induzidas (linhas cinzas).

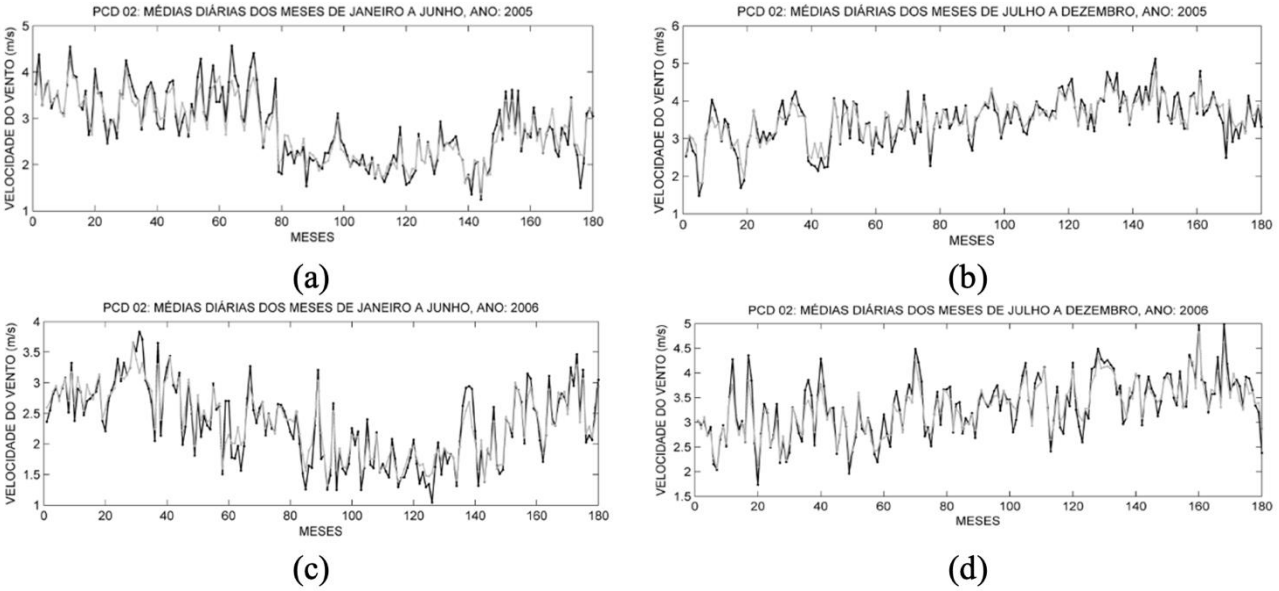

Figura 8 – Médias diárias da velocidade do vento (2005 e 2006) calculadas a partir das séries de velocidade do vento, originais (linhas pretas) e preenchidas com 30% de falhas induzidas (linhas cinzas).

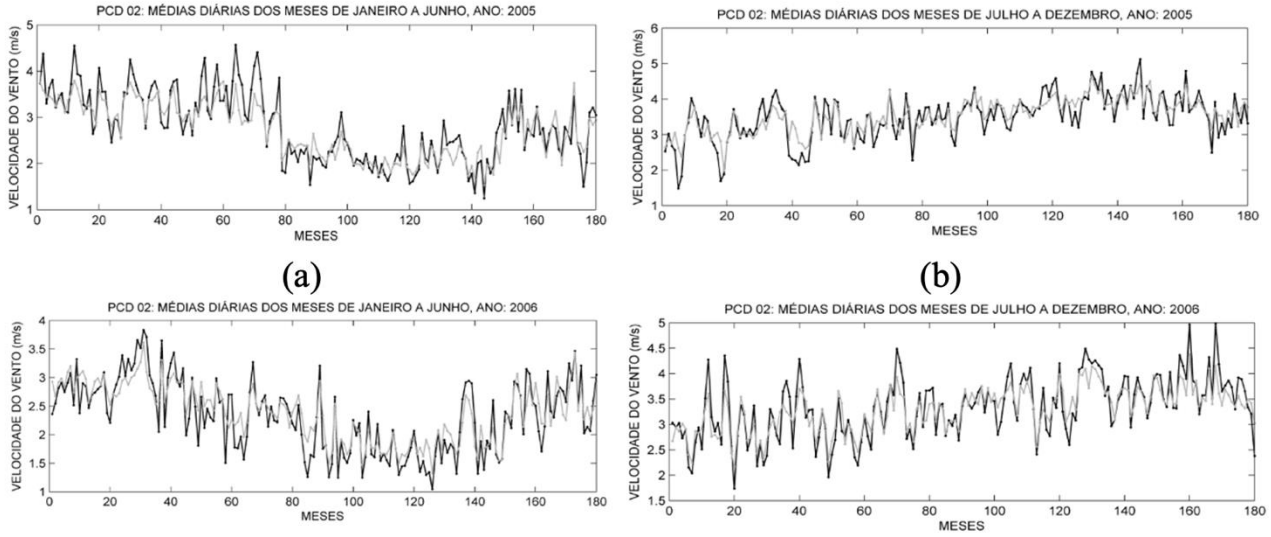

Figura 9 – Médias diárias da velocidade do vento (2005 e 2006) calculadas a partir das séries de velocidade do vento, originais (linhas pretas) e preenchidas com 50% de falhas induzidas (linhas cinzas).

Os valores do coeficiente de correlação estatística (r) calculados entre os valores das séries originais e preenchidas são classificados de fortes ($0.70 \leq | r | \leq 0.89$) a muito forte ($0.90 \leq | r | \leq 1.00$). Para a série com 50% das falhas induzidas é classificado como forte, e para as séries de 10% e 30% são classificados como muito forte.

Os valores dos índices de concordância (d) obtidos são classificados como ótimos (> 0.85) para todas as porcentagens de falhas induzidas na série original (Tabela 6). Os valores do *EMA* indicam erros de 0,11 m/s para o caso de 10% de falhas, de 0,28 m/s para o caso de 30% de falhas e de 0,48 m/s para o caso de 50% de falhas induzidas.

Tabela 6 – Métricas estatísticas calculadas para os valores médios de velocidade de vento entre as séries originais e com falhas preenchidas (Acopiara/CE) (Total de dados 1.440; *99% de significância – Teste t de Student).

Falhas induzidas	*r*	*d*	*EMA*	*EMNA*	*EQM*
10%	0,96*	0,98	0,11	0,004	0,40
30%	0,91*	0,95	0,28	0,002	0,64
50%	0,82*	0,90	0,48	0,0003	0,87

Resultados obtidos para a região de Sobral no Estado do Ceará (Brasil)

O teste da metodologia de preenchimento de falha nos dados da PCD de Sobral (Figura 1) mostra que o modelo PARAFAC-EM (*rank* 2 e 3) preencheu as falhas (10%, 30% e 50%) com correlações estatísticas classificadas como fortes (0, 7 ≤ *r* ≤ 0, 9), mostradas na Figura 10. Tendo o modelo com *rank* 2 apresentado ligeiramente melhor desempenho, atestado através das métricas estatísticas que são mostradas na Tabela 7.

O índice de concordância entre os dados retirados da série e os correspondentes dados estimados (*rank* 2 e 3) apresentam valores classificados como ótimos (0,85≤ *d* ≤1,00), variando de 0,90 a 0,88, respectivamente, de 10% a 50% das falhas induzidas na série de dados original (Tabela 7). Os valores do EMA indicam erros de ±0,76 m/s na série com 10% de falhas, de ±0,80 m/s para 30% de falhas e ±0,83 m/s para a série com 50% de falhas induzidas na série original. Valores considerados baixos em relação aos limites dos valores desta série (0 a 8,46 m/s).

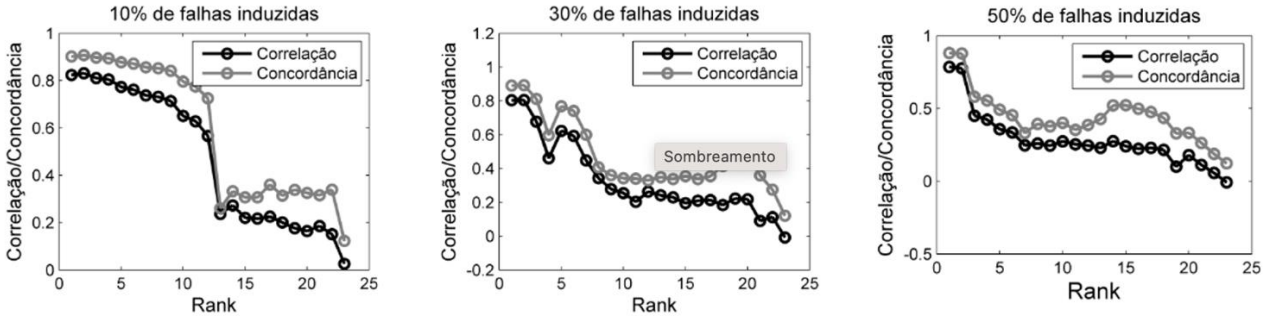

Figura 10 – Correlação e concordância estatística em função do rank para as séries de médias horárias de velocidade do vento com falhas induzidas (10%, 30% e 50%).

Tabela 7 – Métricas estatísticas de avaliação do desempenho do modelo PARAFAC-EM de rank 2 no preenchimento das falhas induzidas nas séries de médias horárias de velocidade do vento na PCD de Sobral (Ceará/Brasil) (Total de dados 17.280; *99% de significância – Teste t de Student).

Falhas induzidas		Métricas Estatísticas				
Total de dados retirados		*r*	*d*	*EMA*	*EMNA*	*EQM*
10%	1728	0,82*	0,90	0,76	0,01	0,98
30%	5184	0,80*	0,89	0,80	-0,01	1,04
50%	8640	0,78*	0,88	0,83	-0,02	1,08

O cálculo do desvio padrão indica que a variabilidade dos dados estimados é menor que a variabilidade dos dados retirados da série original, em todas as quantidades de falhas induzidas.

Apresenta valores respectivamente iguais a 1,50 m/s (10%), 1,51 m/s (30%) e 1,52 m/s (50% de falhas induzidas) que indicam

variabilidade maior para a série com 50% de falhas induzidas (Tabela 8).

Tabela 8 – Métricas estatísticas entre as séries de dados retirados e simulados.

Falhas induzidas	Média		Desvio Padrão	
	Dados Retirados	Dados Estimados	Dados Retirados	Dados Estimados
10%	2,82	2,83	1,73	1,50
30%	2,84	2,83	1,74	1,51
50%	2,80	2,77	1,72	1,52

Em relação a comparação entre as médias diárias calculadas a partir das séries de médias horárias de velocidade do vento e preenchidas pelo método proposto, mostradas nas figuras (11, 12 e 13) a seguir, observa-se que as curvas acompanham o perfil sazonal do vento para a região.

As métricas mostram os coeficientes de correlação (*r*) classificados como muito forte ($0.90 \leq |\, r \,| \leq 1.00$) e índices de concordância (*d*) classificados como ótimos (> 0.85) para todas as porcentagens de falhas induzidas na série original.

Os valores do EMA indicam erros de 0,08 m/s (10%), de 0,24 m/s (30%) e de 0,41 m/s para o caso de 50% de falhas induzidas na série original (Tabela 9).

Tabela 9 – Métricas estatísticas calculadas para os valores médios de velocidade de vento entre as séries originais e com falhas preenchidas (Sobral/CE) (Total de dados 1.440; *99% de significância – Teste t de Student).

Falhas induzidas	*r*	*d*	*EMA*	*EMNA*	*EQM*
10%	0,98*	0,99	0,08	0,005	0,40
30%	0,94*	0,97	0,24	0,002	0,64
50%	0,89*	0,94	0,41	0,0003	0,87

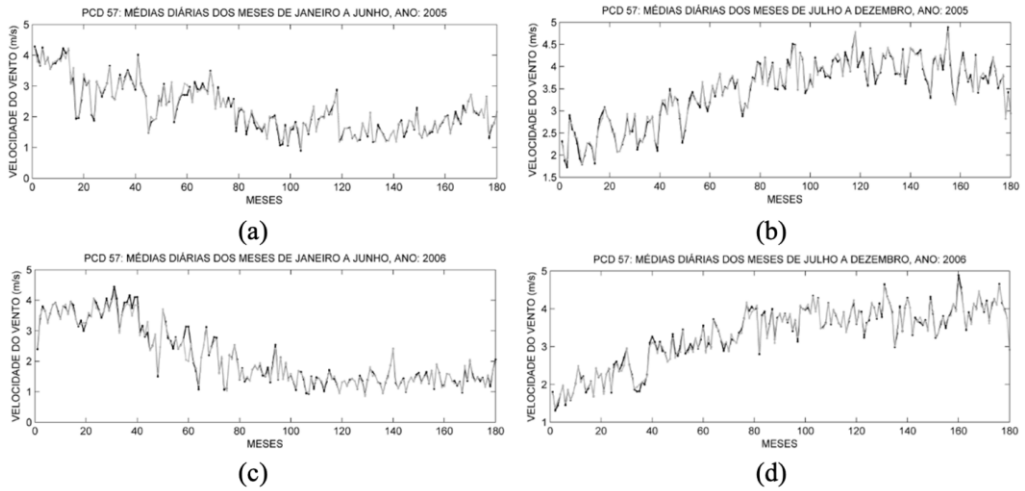

Figura 11 – Médias diárias da velocidade do vento (2005 e 2006) calculadas a partir das séries de velocidade do vento (PCD de Sobral/CE), originais (linhas pretas) e preenchidas com 10% de falhas induzidas (linhas cinzas).

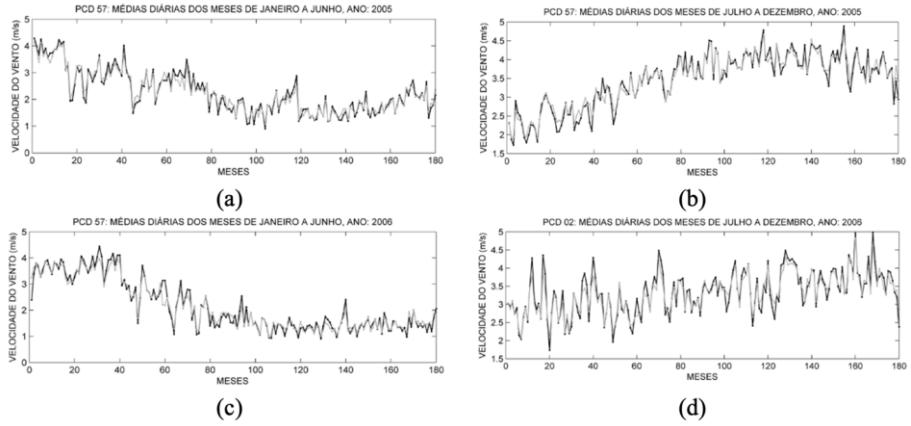

Figura 12 – Médias diárias da velocidade do vento (2005 e 2006) calculadas a partir das séries de velocidade do vento (PCD de Sobral/CE), originais (linhas pretas) e preenchidas com 30% de falhas induzidas (linhas cinzas).

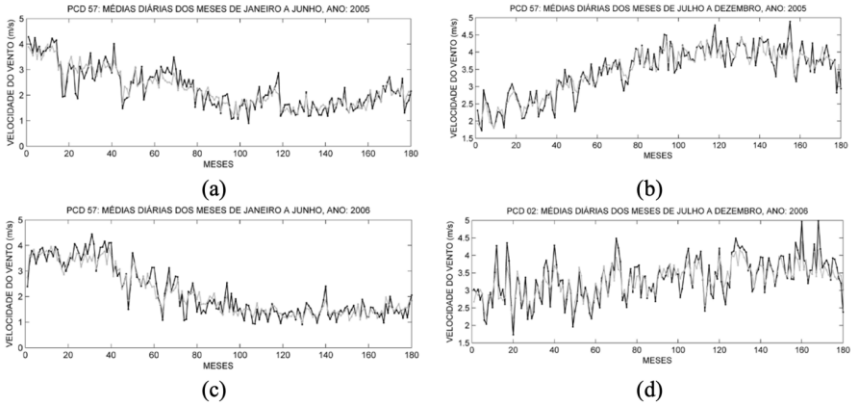

Figura 13 – Médias diárias da velocidade do vento (2005 e 2006) calculadas a partir das séries de velocidade do vento (PCD de Sobral/CE), originais (linhas pretas) e preenchidas com 50% de falhas induzidas (linhas cinzas).

Análise dos resultados obtidos e comparação com outros métodos de imputação de falhas

Os resultados da aplicação do método proposto para os preenchimentos das falhas (10%, 30% e 50%) das séries das médias horárias de velocidade do vento, obtidas nas PCDs das regiões de Acopiara e Sobral no Estado do Ceará (Brasil), mostraram que para falhas aleatoriamente distribuídas o melhor desempenho foi obtido com duas componentes do modelo PARAFAC-EM (*rank* 2).

Menciona-se que esse desempenho foi atestado através do cálculo de métricas estatísticas que indicam a dispersão entre as séries dados preenchidos e as séries originais (sem a indução de falhas), assim os coeficientes de correlação obtidos são classificados de moderado a forte e os índices de concordância de muito bom a ótimo.

O cálculo das demais métricas estatísticas (*EMA*, *EMNA* e *EQM*) na comparação das duas séries de dados (originais e preenchidas) também mostraram resultados satisfatórios em todos os casos analisados, os *EMA* não ultrapassam a amplitude das séries originais, as médias dos dados estimados são próximas das médias dos correspondentes dados retirados da série original, assim, observa-se concordância entre as variabilidades das séries de originais e das preenchidas pelo método proposto.

Comparado a outros métodos de imputação de dados em séries temporais que são usados em diversas áreas do conhecimento, como por exemplo os métodos matemáticos e estatísticos, as redes neurais e a inteligência artificial, usados nos estudos climatológicos e de recursos hídricos (Afrifa-Yamoah et al., 2020; Brubacher et al., 2020; Giovanella et al., 2021;

Miao et al., 2021; Ruezzene et al., 2021; Weerakody et al., 2021; Zhang e Thorburn, 2021; Corbo et al., 2024; Silva Junior et al., 2024; Tavares et al., 2024). E, os métodos de decomposição tensorial (PARAFAC) que são usados nos estudos ambientais (Ooba et al.; 2006; Bonfante et al., 2013; Sciscenko et al., 2022; Turova et al., 2021; Cai et al., 2023). Se observa que estes resultados mostram potencial de uso semelhante ao apontados nos estudos supracitados, especificamente para uso em séries temporais de velocidade do vento que são usados nos diagnósticos e nas estimavas de potencial eólico em diversas regiões do planeta.

Conclusões

O uso do algoritmo PARAFAC-EM no preenchimento de falhas em séries temporais apresenta potencial vantagem de não depender do tipo de variável, por ser um algoritmo iterativo que depende apenas da própria série de dados, da condição inicial das matrizes fatores (A, B e C) que podem ser aleatórias, e do *rank* que deve ser determinado previamente.

Daí, conclui-se que para aplicações em preenchimento de falhas em séries temporais de velocidade do vento, em que a estrutura da série possa ser montada como um tensor (dimensões $720 \times 12 \times 2$) o método proposto pode ser usado de forma satisfatória, com desempenho similar a outros métodos (matemáticos, estatísticos, redes neurais e inteligência artificial) que foram usados em estudos publicados na literatura.

Conclui-se também que a metodologia usada neste estudo pode contribuir para auxiliar tanto na avaliação dos recursos eólicos como na projeção destes recursos nas regiões em que haja a ausência de dados observados consistentes, bem como na validação de resultados de simulações desta variável que são usadas para prospecção de áreas com potencial de geração de energia eólica.

Por fim, menciona-se que a metodologia usada neste estudo pode contribuir para melhoria da qualidade das séries de velocidade do vento, proporcionando assim a realização de estudos mais confiáveis que possam contribuir com o planejamento energético da Região Nordeste do Brasil.

Referências

Acar, E.; Dunlavy, D.M.; Kolda, T.G.; Morup, M. 2010. Scalable Tensor Factorizations with Missing Data. Chemometrics and Intelligent Laboratory Systems, v.106, p. 41-46. https://epubs.siam.org/doi/pdf/10.1137/1.9781611972801.61.

Afrifa-Yamoah, E., Mueller. U.A., Taylor, S.M., Fisher, A.J. 2020. Missing data imputation of high-resolution temporal climate time series data. Meteorol Appl., 27, e1873. https://doi.org/10.1002/met.1873.

Ahn, H., Sun, K., Kim, K.P. 2022. Comparison of Missing Data Imputation Methods in Time Series Forecasting. Computers, Materials & Continua, 70 (1), 767-779. https://doi.org/10.32604/cmc.2022.019369.

Baraldi, A.N.; Enders, C.K. 2010. An introduction to modern missing data analyses. Journal of School Psychology, v.48, p.5-37. DOI: 10.1016/j.jsp.2009.10.001.

Bonfante, A.G.; Ventura, T.M.; Oliveira, A.G.; Marques, H.O.; Oliveira, R.S.; Martins, C.A.; Figueiredo, J.M. 2013. Uma abordagem computacional para preenchimento de falhas em dados micrometeorológicos. Revista Brasileira de Ciências Ambientais, v.27, p.61-70.

Brubacher, J. P., Oliveira, G. G., Guasselli, L. A. 2020. Preenchimento de Falhas e Espacialização de Dados Pluviométricos: Desafios e Perspectivas. Revista Brasileira de Meteorologia, 35(4), 615–629. https://doi.org/10.1590/0102-77863540067.

Cai, M., Tian, Y., Li, A., Li, Y., Han, Y., Ji, W., Zhou, Q., Li, J., Li, W. (2023). Unraveling the evolution of dissolved organic matter in full-scale A/A/O wastewater treatment process using size exclusion chromatography with PARAFAC and nonnegative matrix factorization analysis, Water Research, vol. 235,119879. ISSN 0043-1354, https://doi.org/10.1016/j.watres.2023.119879.

Camargos, V.P.; César, C.C.; Caiaffa, W.T.; Xavier, C.C.; Proietti, F.A. 2011. Imputação múltipla e análise de casos completos em modelos de regressão logística: uma avaliação prática do impacto das perdas em covariáveis. Caderno de Saúde Pública, v.27, n.12, p.2299-2313.https://doi.org/10.1590/S0102311X2011001200003.

Corbo, A.R., Santos, S.A.F., Bertolino, A.V.F.A., Pinto, A.B.S. 2024. Técnicas individuais e combinadas para preenchimento de falhas em dados diários de precipitação no município de

São Gonçalo (RJ). Revista Brasileira de Climatologia, 35(20), 401–427. https://doi.org/10.55761/abclima.v35i20.1739.

Devore, J. L. 2006. Probabilidade e Estatística para Engenharia e Ciência. Thomson Pioneira, São Paulo, Brasil.

Donders, A.R.T. et al. Review: 2006. A gentle introduction to imputation of missing values. Journal of Clinical Epidemiology, v.59, p.1087-1091. 10.1016/j.jclinepi.2006.01.014.

Giovanella, T.H., Oliveira, F.C., Marchi, V. A. A., Tluszcz, J. 2021. Desempenho de Métodos de Preenchimento de Falhas em Dados de Evapotranspiração de Referência para Região Oeste do Paraná. Revista Brasileira de Meteorologia, 36(3), 415–422. https://doi.org/10.1590/0102-77863630001.

Harshman, R.A. 1970. Foundations of Parafac procedure: Models and conditions of an "explanatory" multi-model facor analysis. UCLA, Working papers in phonetics. pp.1-84.

Hitichcock, F.L. The Expression of a Tensor or a Polyadic as Sum of Products. 1927. Journal of Mathematics and Physics, vol. 6, pp. 164-189.

Huang, G.; Paes, A. T. 2009. Posso usar o teste t de Student quando preciso comparar três ou mais grupos? Einstein: Educação Continuada em Saúde, v.7, n.2, p.63-64.

Junninen, H.; Niska, H.; Tuppurainem, K.; Ruuskanen, J. 2004. Methods for imputation of missing values in air quality data sets. Atmospheric Environment, v.38, p.2895-2907. DOI: 10.1016/j.atmosenv.2004.02.026.

Kolda, T.; Bader, B. 2009. Tensor Decompositions and Applications. SIAM Publication Library, vol.51, n.3, p.455-500. DOI: https://doi.org/10.1137/07070111X.

Larson, R.; Farber, B. 2010. Estatística Aplicada. Prentice Hall, 4˚Edição, São Paulo: Pearson.

Lira, M.A.T.; Silva, E.M.; Brabo, J.M. 2011. Estimativas dos Recursos Eólicos no Litoral Cearense usando a Teoria da Regressão Linear. 2007. Revista Brasileira de Meteorologia, vol.26, n.3, p.349-366. DOI: https://doi.org/10.1590/S010277862011000300003.

Lira, M.A.T.; Silva, E.M.; Alves, J.M.B.; Veras, G.V.O. 2014. Estimation of wind resources in the coast of Ceará, Brazil, using the linear regression theory. Renewable & Sustainable Energy Reviews, v.39, p.509-529. DOI: https://doi.org/10.1016/j.rser.2014.07.097.

Lima, F.J.P.; Cavalcanti, E.P.; Souza, E.P.; Silva, E.M. 2012. Evalution of the Wind Power in the State of Paraíba Using the Mesoscale Atmospheric Model Brazilian Developments on the Regional Atmospheric Modelling System. ISRN Renewable Energy. Article ID 847356. https://doi.org/10.5402/2012/847356.

Miao, X., Wu, Y., Wang, J., Gao, Y., Mao, X., & Yin, J. 2021. Generative Semi-supervised Learning for Multivariate Time Series Imputation. Proceedings of the AAAI Conference on Artificial Intelligence, 35(10), 8983-8991. https://doi.org/10.1609/aaai.v35i10.17086.

Mongomery, D. C.; Runger, G.C. 2009. Estatística aplicada e probabilidade para engenheiros. LTC, Rio de Janeiro.

Nunes, L.N.; Kluck, M.M.; Fachel, J.M.G. 2009. Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos. Cad. Saúde Pública, vol. 25, n.2, pp.268-278.https://doi.org/10.1590/S0102-311X2009000200005.

Oliveira, L.F.C.; Fioreze, A.P.; Medeiros, A.M.M.; Silva. M.A.S. 2010. Comparação de metodologias de preenchimento de falhas de séries históricas de precipitação anual. Revista Brasileira de Engenharia Agrícola e Ambiental, v.14, n.11, p.1186-1192. DOI: https://doi.org/10.1590/S141543662010001100008.

Ooba, M.; Hirano, T.; Mogami, J.; Hirata, R. 2006. Comparisons of Gap-Filling Methods for Carbon Flux Dataset: A Combination of a Genetic Algorithm and an Artificial Neural Network. Ecological Modelling, vol.198, n.3-4, pp.473-486. 10.1016/j.ecolmodel.2006.06.006.

Ramli, M.N.N.; Yahaya, A.S.; Ramli, N.A.; Yusof, N.F.F.M.; Abdullah, M.M.Ab. 2013. Roles of Imputation Methods for Filling the Missing Values: A Review. Advances in Environmental Biology, v.7, n.12, p.3861-3869.

Rubin D.B. 1996. Multiple Imputation after 18+ years. Journal of the American Statistical Association, v.91, p.473-489. DOI: https://doi.org/10.2307/2291635.

Ruezzene, C.B., Miranda, R.B., Tech, A.R.B., Mauad, F.F. 2021. Preenchimento de Falhas em Dados de Precipitação através de Métodos Tradionais e por Inteligência Artificial. Revista Brasileira de Climatologia, 29, 177–204. https://ojs.ufgd.edu.br/rbclima/article/view/1518.

Sciscenko, I., Arques, A., Micó, P., Mora, M., García-Ballesteros, S. 2022. Emerging applications of EEM-PARAFAC for water treatment: a concise review, Chemical

Engineering Journal Advances, vol. 10, 100286. ISSN 2666-8211. https://doi.org/10.1016/j.ceja.2022.100286.

Silva Junior, A. A., Gomes, R.S.R., Musis, C.R., Novais, J.W.Z., Maionchi, D., Figueiredo, J.M. 2024. Preenchimento de falhas em séries temporais da temperatura do ar: uma comparação entre modelos de Machine Learning. Revista Brasileira de Climatologia, 35(20), 362–377. https://doi.org/10.55761/abclima.v35i20.17649.

Tavares, P.S., Pilotto, I.L., Chou, S.C., Souza, S.A., Fonseca, L.M.G. Chagas, J. D. 2024. A Dataset of High-Resolution Climate Change Projections Over South America with Bias Correction. Derbyana, ISSN 2764-1465, 45: e821, 2024. DOI: 10.69469/derb.v45.821.

Tomasi, G.; Bro, R. 2005. Parafac and missing values. Sicence Direct, Chemometrics and Intelligent Laboratory Systems, v.75, n.2, p.163-180. 10.1016/j.chemolab.2004.07.003.

Turova, P., Styles, I., Timashev, V., Kravets, K., Grechnikov, A., Lyskov, D., Samigullin, T., Podolskiy, I., Shpigun, O., Stavrianidi, A. (2021). Unsupervised methods in LC-MS data treatment: Application for potential chemotaxonomic markers search, Journal of Pharmaceutical and Biomedical Analysis, vol. 206, 114382. ISSN 0731-7085, https://doi.org/10.1016/j.jpba.2021.114382.

Weerakody, P.B., Wong, K.W., Wang, G., Ela, W. 2021. A review of irregular time series data handling with gated recurrent neural networks, Neurocomputing, 441 (161-178). https://doi.org/10.1016/j.neucom.2021.02.046.

Zhang, P. 2003. Multiple imputation: Theory and method. International Statistical Review, v.71(3), 581-592. https://www.jstor.org/stable /1403830.

Zhang, Y., Thorburn, P.J. 2021. A dual-head attention model for time series data imputation, Computers and Electronics in Agriculture, vol.189, 106377. ISSN 0168-1699. https://doi.org/10.1016/j.compag.2021.106377.