Imputação de dados ausentes através de redes neurais recorrentes no monitoramento de integridade estrutural

Luiz Sousa
Instituto de Tecnologia
Universidade Federal do Pará
Marabá, Brasil
luiz.sousa@unifesspa.edu.br

Adam Santos
Instituto de Geociências e Engenharias
Universidade Federal do Sul e Sudeste do Pará
Marabá, Brasil
adamdreyton@unifesspa.edu.br

João Weyl
Instituto de Tecnologia
Universidade Federal do Pará
Belém, Brasil
jweyl@ufpa.br

***Resumo*— Um problema comum em grandes conjuntos de dados é a informação ausente, seja por falha nos sensores de captura, perda no transporte, ou outra situação que culmine com a perda de dados. Diante desta situação, é frequente que a decisão do pesquisador seja desconsiderar os dados ausentes, removê-los do conjunto, no entanto, essa exclusão pode gerar inferências que não são válidas, principalmente se os dados que permanecem na análise são diferentes daqueles que foram excluídos. Para lidar com este problema em conjuntos de dados de monitoramento de integridade estrutural (*Structural health monitoring* - SHM), este trabalho faz uso de redes neurais recorrentes *Gated Recurrent Units* (GRU) e *Long Short-Term Memory* (LSTM), para realizar a tarefa de imputação de dados ausentes. Em uma etapa anterior à imputação, foi realizada a amputação artificial dos dados, assumindo o mecanismo de dados ausentes *Missing Completely at Random* (MCAR), em percentuais de 25, 50 e 75%. As técnicas de imputação foram avaliadas com o uso da métrica *Mean Absolute Percentage Error* (MAPE). Posteriormente, foi aplicada a etapa de detecção de dano, as bases imputadas foram submetidas aos algoritmos *Mahalanobis Square Distance* (MSD) e *kernel principal component analysis* (kPCA) a fim de se obter as taxas de erros T1 e T2 detectadas. A partir dos resultados obtidos, foi possível observar que o uso da LSTM na imputação dos dados, alcançou resultados melhores que a GRU em todas as taxas de amputação, este melhor desempenho pode também ser notado na etapa de detecção de dano, onde as bases imputadas por LSTM alcançam melhores resultados de detecção de erros T1 e T2.**

***Palavras-Chave*— Dados ausentes, imputação, SHM, redes neurais recorrentes, LSTM, GRU, detecção de dano, MSD, kPCA.**

I. Introdução

As sociedades modernas são fortemente dependentes de sistemas estruturais e mecânicos, como aeronaves, pontes, sistemas de geração de energia, plataformas de petróleo, edifícios e sistemas de defesa. Muitos desses sistemas existentes estão atualmente chegando ao fim da vida útil de projeto original. Manter sua operação segura e confiável é crucial para garantir o bem-estar das pessoas, além de proteger os investimentos [1].

O termo monitoramento de integridade estrutural (*Structural health monitoring* - SHM) se refere ao processo de implementação de uma estratégia de detecção de danos para infraestrutura aeroespacial, civil ou mecânica. Esse processo envolve a observação de uma estrutura ao longo do tempo usando medições dinâmicas espaçadas periodicamente, a extração de características sensíveis a dano dessas medições e a análise estatística dessas características para determinar o estado atual de integridade do sistema. O SHM pode ser usado para fornecer, em tempo quase real, informações confiáveis sobre o desempenho do sistema estrutural [2].

O monitoramento de integridade estrutural faz-se necessário, pois, todas as estruturas feitas pelo homem têm vida útil finita e começam a se degradar assim que são colocadas em serviço. Processos como corrosão, fadiga, erosão, desgaste e sobrecargas os degradam até que não sejam mais adequados para o uso pretendido [3].

A eficácia de aplicações SHM depende muito da precisão e confiabilidade dos dados adquiridos. No entanto, medições com dados incompletos, ausentes ou corrompidos aleatoriamente, ocorrem frequentemente neste contexto. A ausência de dados pode estar relacionada com o mau funcionamento do sensor, erro de instrumentação, falha na transmissão ou gravação do dado, etc. Essas incidências corrompem significativamente as medições de vibração da estrutura [4]. Erros ocasionados pela perda de dados afetam a análise da estrutura e impactam a tomada de decisão subsequente [5].

Dados ausentes são um problema comum na maioria dos domínios de pesquisa científica, como biologia, medicina, ciências climáticas e outros. Eles introduzem um elemento de ambiguidade na análise de dados e podem afetar as propriedades dos estimadores estatísticos, como médias, variações ou porcentagens, resultando em conclusões enganosas. Uma variedade de técnicas foram propostas para substituir valores ausentes por previsão estatística, este processo é geralmente referido como imputação de dados ausentes [6].

O presente trabalho tem como foco a comparação entre métodos de imputação de dados no contexto de SHM, através de Redes Neurais Recorrentes (RNR), bem como a aplicação de técnicas de detecção de dano nas bases imputadas. A avaliação dos resultados da etapa de imputação é realizada através da métrica MAPE. Para avaliar o comportamento das bases imputadas na detecção de dano são aplicados os algoritmos MSD e kPCA, a escolha destes se justifica pela característica de análise linear e não-linear dos dados presentes

no conjunto. MSD é uma técnica voltada para conjuntos com características lineares, onde parte-se do pressuposto de que os dados estudados possam estar correlacionados. O uso do kPCA se dá em casos em que se está interessado nas componentes principais no espaço de características, que é não-linearmente relacionado com as variáveis de entrada originais, sendo esta última a característica mais encontrada em problemas de SHM.

Este trabalho está organizado da seguinte forma, a seção II aborda trabalhos relacionados à tarefa de imputação de dados ausentes em bases de SHM. A seção III apresenta os detalhes da etapa de desenvolvimento do trabalho, processo de implementação e técnicas utilizadas. A seção IV discute os resultados obtidos. As considerações finais e trabalhos futuros são apresentados na seção V.

II. Trabalhos Relacionados

O tema de imputação de dados ausentes no contexto de SHM representa um campo relativamente novo, há trabalhos que abordam esta linha de pesquisa focados na área de redes de sensores sem fio. Esses trabalhos propõem técnicas para lidar com esse problema no âmbito da transmissão de dados e a perda que pode ocorrer durante a comunicação entre os sensores.

A. Imputação de dados no contexto de SHM

A pesquisa de Bao et al. [5] propõe uma nova aplicação do uso de amostragem compressiva (*Compressive sampling* - CS) para a recuperação de dados perdidos em uma rede de sensores sem fio usada em SHM. O estudo foi realizado com dados de sensores instalados na ponte Jinzhou West Bridge no Centro Aquático Nacional em Beijing. Primeiramente, os dados originais a serem transmitidos são transformados em um sinal sintético e uma porcentagem dos dados transformados é considerada perdida durante a transmissão sem fio. O sinal original então é reconstruído a partir dos dados sintéticos utilizando a técnica de otimização l_1. O autor aponta conseguir uma boa precisão de recuperação dos dados nos domínios de tempo, frequência e fase. O erro de reconstrução aumenta conforme o crescimento da taxa de perda de dados, demonstrando a importância do problema de dados ausentes.

Zou et al. [7] incorporou um algoritmo de recuperação de dados baseado em CS em sensores inteligentes sem fio. A exemplo do trabalho de Bao et al. [5], o autor chega à conclusão de que o desempenho da reconstrução dos dados cai com o aumento da taxa de dados ausentes.

O trabalho de Luo et al. [8] categoriza mecanismos de geração de dados ausentes em monitoramento de integridade estrutural, em um cenário com uso de sensores de rede sem fio. O autor cita três fatores principais que podem resultar na perda de dados durante o processo de monitoramento: obstáculos durante a transmissão dos dados, falha de energia e falha dos instrumentos de monitoramento. De acordo com esses fatores, ocorre a classificação dos dados ausentes também em três categorias. O autor propõe abordagens diferentes para cada uma delas visando a reconstrução dos dados perdidos.

A pesquisa de Chen et al. [9] implementa um novo método indireto de regressão de distribuição-para-distribuição (*distribution-to-distribution regression* - DDR) para imputação de dados ausentes em SHM. A proposta do autor aborda o problema de dados ausentes com foco nas distribuições de probabilidade, que podem variar de acordo com o segmento de tempo da amostra obtida. Em um trabalho mais recente, Chen et al. [10] apresenta uma atualização em seu método proposto.

Ainda no campo da perda de sinal na transmissão de dados no SHM com uso de redes de sensores sem fio, Fan et al. [11] aponta em sua pesquisa que os dados de vibração coletados com altos índices de perda dificilmente poderão ser utilizados em uma análise. Para tratar esse problema, o autor propõe uma abordagem baseada em redes neurais convolucionais (CNN) para recuperar os dados de vibração perdidos durante a transmissão.

Os trabalhos que tratam de recuperação de dados no contexto de SHM, citados nesta seção, propõem novas técnicas e não realizam comparação entre as técnicas existentes de imputação, tal comparação encontrada em trabalhos de outras áreas do conhecimento, como abordado na seção a seguir.

B. Imputação de dados

Peter Schmitt e Jonas Mandel [6] realizaram um estudo comparativo entre seis técnicas de imputação de dados, sendo utilizados quatro *datasets* de tamanhos variados e sob uma suposição de dados ausentes completamente aleatória (MCAR). As técnicas comparadas foram: média, K-vizinhos mais próximos (KNN), Fuzzy K-vizinhos mais próximos (FKM), decomposição em valores singulares (SVD), análise de componentes principais bayesiana (bPCA) e método de imputação múltipla por equações encadeadas (MICE). Foram utilizados quatro critérios para avaliação do desempenho de cada uma das técnicas, a saber: erro médio quadrático (RMSE), erro de classificação não supervisionado (UCE), erro de classificação supervisionada (SCE) e tempo de execução. Os resultados obtidos pelo autor, sugerem que os métodos de imputação mais populares (média, KNN, SVD e MICE), considerando o ano de publicação do trabalho, não são necessariamente os mais eficientes. As técnicas bPCA e FKM apresentaram um melhor desempenho nos dados utilizados, sendo que o uso de FKM foi melhor observado em pequenos conjuntos de dados.

A pesquisa de Ahmad et al. [12] aborda o problema de dados ausentes em conjuntos de dados médicos. O autor reforça que simplesmente remover as partes faltantes dos conjuntos de dados originais pode trazer mais problemas do que soluções, para isso é realizada a comparação entre métodos de aprendizagem de máquina para imputação em um *dataset* de informações cardiovasculares. O trabalho indica que deve haver uma atenção maior na fase de pré-processamento dos dados e conclui que técnicas de aprendizado de máquina podem ser a melhor abordagem para imputar valores ausentes no conjunto de dados observado.

Em seu trabalho, Mani [13] demonstra o uso de RNRs para o desafio de processar frases com exatamente uma palavra faltando e determinar onde deve ocorrer a predição e também descobrir a palavra correta para inserir no local previsto. Ainda com o uso de RNRs, o trabalho de Ghazi et al. [14] aborda o uso desse método para prever valores ausentes em séries temporais de dados clínicos para modelagem de progressão de doença. Kim et al. [15] em sua pesquisa utilizaram rede neural recorrente do tipo LSTM para previsão de dados ausentes em exames médicos e realiza a comparação do uso desse método com a técnica de regressão linear.

Os trabalhos sobre imputação de dados em SHM, aqui apresentados, têm seu foco aplicado em redes de sensores sem fio, e abordam em sua maioria apenas recuperação dos dados na transmissão entre os sensores. Não é realizada uma comparação entre técnicas de imputação de dados ausentes, característica encontrada em trabalhos de outras áreas do conhecimento. Assim, este trabalho se diferencia das outras propostas, pois realiza a comparação de técnicas de imputação de dados ausentes, utilizando aprendizado de máquina, no contexto de SHM e faz uma avaliação entre os métodos propostos.

III. Metodologia

O trabalho consiste em utilizar uma base de dados de SHM, realizar amputação de valores em posições aleatórias e depois realizar a imputação, nas mesmas posições, de valores preditos através do uso de RNRs. Os algoritmos escolhidos para realizar a imputação foram: *Gated Recurrent Units* (GRU) e *Long Short-Term Memory* (LSTM).

Em um contexto de monitoramento e avaliação da integridade estrutural, mais do que prever um valor ausente em um dado conjunto de dados, esse valor deve representar um significado. Neste caso, após o processo de imputação, esse conjunto de dados é submetido a um teste de detecção de dano a fim de comparar o desempenho da detecção de danos entre a base original (*baseline*) e as bases preditas. Os algoritmos utilizados para a detecção de dano na base original e nas bases preditas foram: *Mahalanobis Square Distance* (MSD) e *kernel principal component analysis* (kPCA). O fluxo da metodologia utilizada pode ser observado na Figura 1.

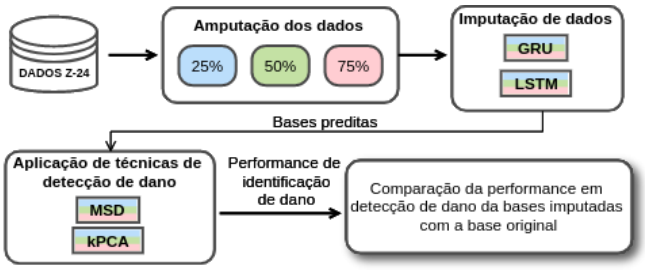

Fig. 1. Fluxo da metodologia utilizada.

Para realizar a imputação de dados ausentes no contexto de SHM, foi utilizado o conjunto de dados da Ponte Z-24 na Suíça. A Ponte Z-24, construída em 1961, possuía a intenção de ligar as aldeias de Utzenstorf e Koppigen na Suíça. Essa ponte é uma estrutura pós-tensionada de vigas de caixa de concreto e possui um vão principal no meio de 30 metros e dois vãos laterais de 14 metros cada.

A ponte precisou ser demolida para dar espaço a uma ponte maior, mas antes da demolição foram feitos testes para o projeto *System Identification to Monitor Civil Engineering* (SIMCES), para extrair dados que pudessem ser usados para testes de detecção de dano [16]. Foi efetuado um programa de monitorização a longo prazo, de 11 de Novembro de 1997 a 10 de Setembro de 1998, para quantificar a variabilidade operacional e ambiental presente na ponte e detectar dano introduzido artificialmente, de forma controlada, no último mês de operação. A cada hora, durante 11 min, oito acelerômetros capturavam as vibrações da ponte e uma série de sensores mediam parâmetros ambientais, como temperatura em diversos locais [17]. Dessa maneira, a base de dados tem um longo período de medições sem dano e um período com dano, possibilitando realização de testes e validações com os dados gerados. As primeiras quatro frequências naturais estimadas (características), de hora em hora, de 11 de novembro de 1997 a 10 de setembro de 1998, com um total de 3.932 observações, estão representadas na Figura 2.

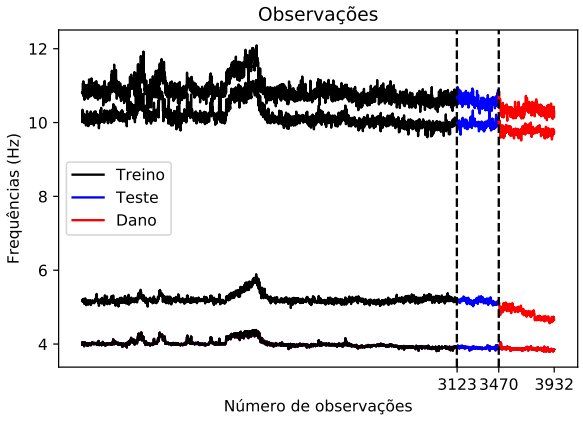

Fig. 2. Quatro primeiras frequências naturais ou características da Ponte Z-24.

A. Método de amputação dos dados

Devido o conjunto de dados da Z-24 se apresentar completo, o que facilita a posterior tarefa de comparação de resultados preditos, foi necessário realizar o processo de amputação dos dados a fim de simular a falha de algum sensor, interrupção de sinal, ou alguma situação em que culminasse com a perda de dados no conjunto. O principal sistema de classificação de mecanismos de dados ausentes foi proposto por Rubin [18]. Para melhor compreendê-lo, assumiremos que a variável Y representa o conjunto de dados completo, então: $Y = \{Y_o, Y_a\}$. Temos que Y_o é um valor observado no conjunto Y enquanto que Y_a representa um valor ausente no mesmo conjunto. O tratamento dos dados ausentes de uma forma aceitável depende da natureza de sua ausência.

Existem atualmente quatro mecanismos de dados ausentes na literatura, que são [19]:

- MCAR - *Missing Completely at Random*
- MAR - *Missing at Random*
- MNAR - *Missing Not at Random*
- MBND - *Missing By Natural Design*

Este trabalho assume que os dados ausentes são do tipo MCAR que ocorre quando a probabilidade de que as respostas que estão faltando não está relacionada aos valores específicos que, em princípio, deveriam ter sido obtidos ou ao conjunto de respostas observadas, significa que a causa que levou aos dados faltantes é um evento aleatório. Little e Rubin [20] expressam essa relação matematicamente como:

$$P(M|Y_o, Y_a) = P(M), \tag{1}$$

onde $M \in \{0, 1\}$ representa a indicação de um valor ausente. $M = 1$ se Y é um valor observado e $M = 0$ se Y for um valor ausente. Y_o representa os valores observados em Y enquanto Y_a representa os valores ausentes de Y. A partir da Equação 1, a probabilidade de uma entrada perdida em uma variável não está relacionada a Y_o ou Y_a. O conjunto de dados foi amputado em 25%, 50% e 75% para posterior imputação conforme observado na Figura 1.

B. Técnicas utilizadas para imputação

O uso de redes neurais recorrentes aplicadas as séries temporais tem seu uso justificado em um dos seus benefícios que é o uso de "memória", que funciona como um *buffer* que permite análises mais profundas de informações sensíveis ao contexto. As RNRs contam ainda com forte desempenho de previsão, bem como a capacidade de capturar dependências temporais de longo prazo e observações de comprimento variável [21].

1) LSTM: As redes LSTM são baseadas nas RNRs e foram desenvolvidas para sanar o problema do desaparecimento de gradiente. LSTMs estendem uma RNR com células de memória, em vez de unidades recorrentes, para armazenar e produzir informações, facilitando o aprendizado de relações temporais em escalas de tempo longas. Uma célula de memória LSTM é composta de três portas: porta de entrada *input gate* (i_t), porta de esquecimento *forget gate* (f_t) e porta de saída *output gate* (o_t). A porta de entrada controla o impacto do valor de entrada no estado da célula de memória. A porta de saída controla o impacto do estado da célula de memória na saída na etapa de tempo atual. A porta de esquecimento determina quanto do valor da memória anterior deve ser passado para a próxima etapa de tempo [22].

2) GRU: Do mesmo modo que a rede LSTM, a GRU segue o papel de manter uma espécie de memória de curto prazo. Da mesma forma que a unidade LSTM, a GRU possui unidades de portas que modulam o fluxo de informações dentro da unidade, mas sem possuir células de memória separadas. Na GRU são utilizadas apenas duas portas de controle, uma porta de atualização (z_t) (*update gate*) que faz o papel da porta de esquecimento e de entrada, e uma porta de redefinição (r_t) (*reset gate*) [21].

3) Avaliação dos resultados de imputação: Para avaliação dos resultados obtidos após a aplicação das RNRs, é adotado a métrica *Mean Absolute Percentage Error* (MAPE). O MAPE considera valores reais alimentados em modelos e valores ajustados a partir do modelo, considerando a diferença absoluta entre os dois como porcentagem do valor real [23]. O MAPE é definido por:

$$MAPE = \sum_{i=1}^{N} \frac{\frac{Y_{(t)} - \widehat{Y}_{(t)}}{Y_{(t)}} \times 100}{N} \tag{2}$$

onde $Y_{(t)}$ é o valor da série temporal no período t, $\widehat{Y}_{(t)}$ é o valor da previsão para o período t e N é o total de observações.

4) Entrada e arquitetura das RNRs utilizadas: Os dados de entrada para rede foram modelados com 1 classe de entrada e 1 sinal de saída para um valor discreto. Na Tabela I temos temos a divisão das entradas, atributos, dados para treinamento, dados para teste, e passos-a-frente como parâmetros utilizados para ambas as redes. O critério para seleção dos dados de entrada para treinamento e teste foi dividido em 75% e 25% respectivamente. A opção de passos a frente seleciona 50 valores de observações passadas para prever o próximo valor, com isso a rede consegue observar uma grande quantidade de dados durante o estágio de treinamento. A Tabela II mostra a arquitetura utilizada nas duas redes, o critério de seleção do número de neurônios e camada *Dropout* foram escolhidos após testes de performance executados em cada configuração.

TABELA I
DADOS DE ENTRADA DAS RNRS

Observações	Atributos	Treino	Teste	Passos a frente
3932	1	2949	983	50

TABELA II
ARQUITETURA DAS RNRS UTILIZADAS NO TRABALHO

Camada	Neurônios	Ativação	Função Perda	Regularização
1	50			Dropout(0,1)
2	25			Dropout(0,1)
3	25			Dropout(0,1)
4	25			Dropout(0,1)
5	1	Linear	Huber	

C. Técnicas utilizadas para detecção de dano

Essas técnicas lidam com as características como entrada de dados e geram indicadores de dano (*damage indicator* - DI) como saídas do seu processamento. Sob certas condições, um limiar pode ser estabelecido para classificar os DIs em sem dano ou com dano, considerando um determinado grau de confiança nos dados de treinamento. Neste estudo, o grau de confiança é igual a 95%.

1) MSD: MSD é uma medida de distância para detecção estatística multivariada de desvios [24]. Ao considerar uma matriz de treinamento, $\mathbf{X}$, com um vetor de médias multivariado, μ, e uma matriz de covariância, Σ, a MSD (ou DI no contexto de SHM) entre um vetor de características a partir

de $\mathbf{X}$ e qualquer novo vetor de características (ou observação) a partir da matriz de teste, $\mathbf{Z}$, é calculada tal que

$$\mathrm{DI}(\mathbf{z}) = (\mathbf{z} - \mu)\,\Sigma^{-1}\,(\mathbf{z} - \mu)^{\top}. \tag{3}$$

O pressuposto é que se uma nova observação for obtida a partir de dados coletados da condição danificada, que podem incluir fontes de variabilidade operacional e ambiental, a observação está mais longe da média da condição normal. Por outro lado, se uma observação é obtida de um sistema dentro de sua condição não danificada, mesmo com variabilidade operacional e ambiental, esse vetor de características está mais próximo da média da condição normal.

2) kPCA: O algoritmo kPCA é a melhoria não linear do PCA [25]. Seja $\mathcal{X} \in \mathbb{R}^d$ o espaço de entrada tal que as observações $\mathbf{x}_i \in \mathcal{X}$, $i = 1, \ldots, n$. Cada observação $\mathbf{x}$ é então mapeada para um espaço de características $\mathcal{H}$ de d_ϕ dimensões, aplicando as funções de mapeamento ϕ_m, $m = 1, \ldots, d_\phi$, onde

$$\phi(\mathbf{x}) = \begin{bmatrix}\phi_1(\mathbf{x}) & \phi_2(\mathbf{x}) & \ldots & \phi_{d_\phi}(\mathbf{x})\end{bmatrix}^{\top}. \tag{4}$$

Ao empregar o "truque" do kernel [26], $\mathcal{K} : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ é definida como uma função de kernel escalar semi-definida positiva que satisfaz para todos $\mathbf{x}_i$, $\mathbf{x}_j \in \mathcal{X}$:

$$\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^{\top}\,\phi(\mathbf{x}_j). \tag{5}$$

$\mathcal{K}(\cdot)$ define um produto interno que permite mapear as observações implicitamente para um espaço de altas dimensões. Com

$$\mathbf{\Phi} = [\phi(\mathbf{x}_1) \quad \phi(\mathbf{x}_2) \quad \ldots \quad \phi(\mathbf{x}_n)] \tag{6}$$

sendo a matriz $d_\phi \times n$ de observações mapeadas e $\mathbf{K} = \mathbf{\Phi}^{\top}\mathbf{\Phi}$ sendo a matriz de kernel $n \times n$. O "truque" do kernel consiste em especificar o kernel $\mathcal{K}(\cdot)$ em vez do mapeamento ϕ. Neste trabalho, o kernel Gaussiano é aplicado:

$$\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\parallel \mathbf{x}_i - \mathbf{x}_j \parallel^2}{2\sigma^2}\right), \tag{7}$$

onde o kernel não linear implicitamente define um espaço de características de alta dimensão com um parâmetro de largura de banda σ^2.

Os autovalores Σ e os correspondentes autovetores $\mathbf{U}$ podem ser calculados usando a decomposição de valores singulares (SVD) para resolver o problema generalizado de autovalores [25],

$$\mathbf{K}\mathbf{U} = \mathbf{U}\Sigma. \tag{8}$$

Em seguida, Σ_1 e $\mathbf{U}_1$ devem ser definidos como segue,

$$\Sigma = [\Sigma_1 \quad \Sigma_2], \Sigma_1 \in \mathbb{R}^{r \times r}, \tag{9}$$

$$\mathbf{U} = [\mathbf{U}_1 \quad \mathbf{U}_2], \mathbf{U}_1 \in \mathbb{R}^{n \times r}, \tag{10}$$

onde Σ_1 compreende os r maiores autovalores e $\mathbf{U}_1$ os correspondentes autovetores.

Existem vários métodos para otimizar o parâmetro σ^2 do kernel Gaussiano [27], requerendo apenas que $n \geq d$. A maximização da entropia da informação a partir da matriz de kernel, $\mathbf{K}$, é o método mais indicado no contexto de detecção de dano em SHM.

Similarmente, muitos critérios foram propostos para determinar o número de componentes principais r selecionados no espaço de características de alta dimensão [28, 29]. Neste estudo, r é calculado de forma a compreender aproximadamente toda a variabilidade normal encontrada nos dados de treinamento, i.e., 99% da variabilidade é retida.

Considerando que um modelo de dados sem dano foi estabelecido na fase de treinamento, na fase de teste o DI é gerado para cada nova observação $\mathbf{z}_i \in \mathbb{R}^d$, $i = 1, \ldots, l$. Primeiramente, a nova observação deve ser mapeada no espaço de características de alta dimensão na forma de $\mathbf{\Phi}(\mathbf{z}_i)^{\top}\mathbf{\Phi}$ (ou $\mathbf{\Phi}^{\top}\mathbf{\Phi}(\mathbf{z}_i)$), usando $\mathbf{X}$ e $\mathbf{z}_i$ in 5.

Em seguida, os autovetores $\mathbf{U}_1$ devem ser normalizados,

$$\mathbf{u}_m \rightarrow \frac{\mathbf{u}_m}{\sqrt{\Sigma_{m,m}}}, m = 1, \ldots, r. \tag{11}$$

Finalmente, um DI é gerado para uma nova observação l tal que,

$$\mathrm{DI}(\mathbf{z}_l) = \mathbf{\Phi}(\mathbf{z}_l)^{\top}\,\mathbf{\Phi}\mathbf{U}_1\mathbf{U}_1^{\top}\mathbf{\Phi}^{\top}\mathbf{\Phi}(\mathbf{z}_l). \tag{12}$$

IV. Resultados

A. Resultados da etapa de imputação

Para o treinamento dos modelos preditivos foi realizado a divisão do conjunto de dados entre treino e teste representando 75% e 25%, respectivamente. A base de dados da Z-24 é representada por uma matriz $D^{3932 \times 4}$ onde cada coluna representa uma frequência natural. Na etapa de treino, teste e previsão dos valores a serem imputados, as frequências foram divididas e os modelos aplicados individualmente em cada uma delas.

Para alcançar os resultados pretendidos foram realizadas vinte rodadas de treinamento em cada uma das quatro frequências e para cada rede neural, GRU e LSTM, visando lidar com a variabilidade estocástica presente nas redes neurais. Com as redes treinadas, vinte modelos para cada rede, foi realizada a etapa de geração das bases amputadas e predição, conforme descrição do algoritmo 1:

Como métrica para avaliação das bases imputadas foi utilizada a técnica MAPE e os resultados consolidados são apresentados nas Tabelas de 1 a 4.

Para cada frequência natural foi realizado o cálculo com as variações na taxa de amputação em 25%, 50% e 75%, o que conforme esperado mostra que quanto maior a taxa de dados ausentes, maior a variação do erro em comparação ao conjunto de testes. Conforme observado nas Tabelas de III a VI, os melhores resultados foram alcançados com o uso do modelo LSTM para imputação. Nas Figuras de 3 a 5 é demonstrada a performance da previsão com LSTM, que foi a técnica de

Algoritmo 1: Geração das bases consolidadas (Amputadas + Preditas)

Número de Bases Amputadas = 10;
Carregar Base de dados da Z24;
for $i \leftarrow 0$ **to** 10 **do**
Gerar amputação nos dados nas bases da Z24;
Criar matriz invertida das posições amputadas;
Criar vetor para armazenar os resultados;
for $i \leftarrow 0$ **to** 20 **do**
Carregar modelos de predição;
Realizar predição sobre os dados de teste;
Armazenar resultado predito no vetor;
end
Realizar soma dos resultados preditos;
Tirar média dos valores e armazenar;
Unir os vetores de base amputada + média dos valores preditos;
end
Result: Base consolidada (base amputada + predita)

TABELA III
MAPE FREQUÊNCIA 1.

Tx de amputação	GRU	LSTM
25%	0,0805	0,0797
50%	0,1694	0,1674
75%	0,2494	0,2468

TABELA IV
MAPE FREQUÊNCIA 2.

Tx de amputação	GRU	LSTM
25%	0,2369	0,1750
50%	0,4834	0,3530
75%	0,7264	0,5341

TABELA V
MAPE FREQUÊNCIA 3.

Tx de amputação	GRU	LSTM
25%	0,1331	0,1322
50%	0,2751	0,2735
75%	0,4006	0,3984

TABELA VI
MAPE FREQUÊNCIA 4.

Tx de amputação	GRU	LSTM
25%	0,1905	0,1868
50%	0,3876	0,3805
75%	0,5849	0,5738

imputação que obteve os melhores resultados, para as quatro frequências de acordo com o percentual de amputação nos dados.

B. Resultados da etapa de detecção de danos

Após a etapa de imputação de dados, as bases preditas para cada frequência foram unidas e serviram de entrada para a etapa de detecção de dano conforme demonstrado da Figura 1. Conforme observado em Figueiredo e Santos [17], os dados usados para treino dos algoritmos de detecção de dano representam 90% de todo o conjunto de dados sem dano, observações 1 a 3122. Na Figura 2 é ilustrada a divisão adotada para aplicação desta etapa.

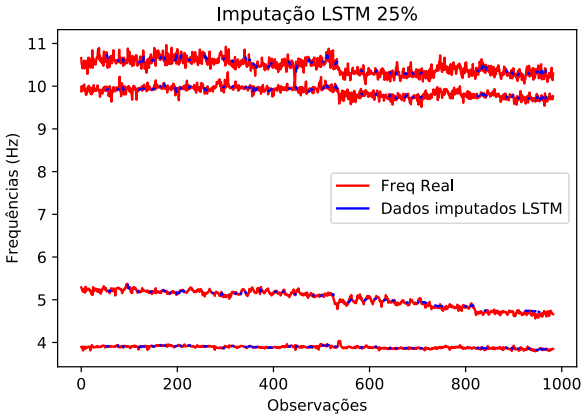

Fig. 3. Previsão LSTM com 25% de dados amputados nas quatro frequências.

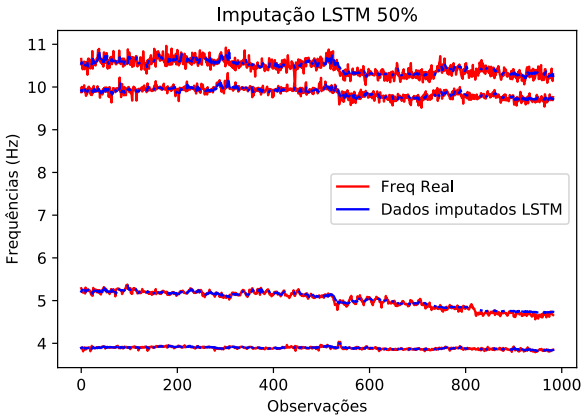

Fig. 4. Previsão LSTM com 50% de dados amputados nas quatro frequências.

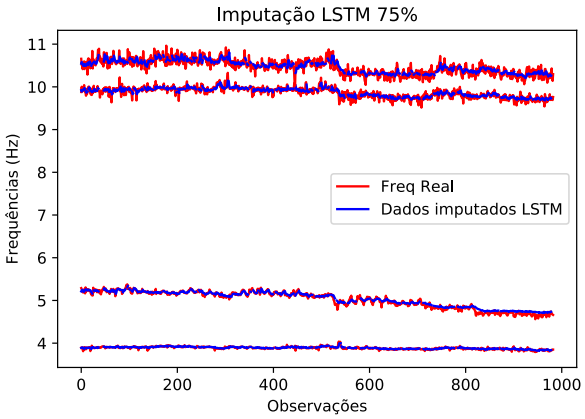

Fig. 5. Previsão LSTM com 75% de dados amputados nas quatro frequências

Os resultados dos testes da detecção de dano podem classificar corretamente uma amostra dos dados tanto como contendo dano, quanto sendo sem dano. Ao ser classificada com dano uma amostra que de fato apresenta dano, este é um verdadeiro positivo, caso seja classificado como sem dano uma amostra sem dano é um verdadeiro negativo. Existem casos ainda, que as amostras são classificadas como sendo falsos positivos e falsos negativos, respectivamente erros do Tipo 1 e 2 (T1 e T2) [30].

De modo a demonstrar a condição *baseline* da estrutura representada nos dados, os modelos de detecção de dano gerados foram aplicados no conjunto total dos dados originais: 3932 observações. Dessa maneira temos uma visão da performance dos modelos nos dados originais para posterior comparação com as bases imputadas. Na Tabela VII são apresentados os valores de detecção de T1 e T2 na condição *baseline*.

TABELA VII

DETECÇÃO DE ERROS T1 E T2, BASELINE.

Algoritmo	Erros T1	Erros T2	% T1	% T2
MSD	162	190	4,6685	41,1255
kPCA	171	4	4,9279	0,8658

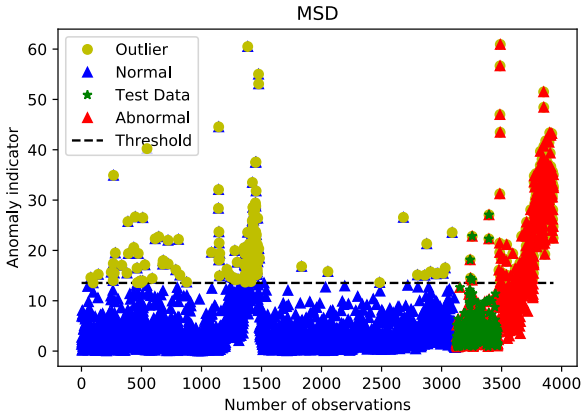

Fig. 6. Aplicação do algoritmo MSD demonstrando a condição *baseline* da estrutura.

As Tabelas de VIII a X demonstram os resultados médios, considerando as dez bases imputadas por LSTM, de erros T1 e T2, as Tabelas de XI a XIII demonstram os resultados médios, considerando as dez bases imputadas por GRU, variando de acordo com a taxa de imputação.

TABELA VIII

DETECÇÃO DE ERROS T1 E T2 EM BASE IMPUTADA EM 25% POR LSTM.

Algoritmo	Erros T1	Erros T2	% T1	% T2
MSD	158 ± 1,74	201 ± 3,95	4,6685	43,5064
kPCA	166 ± 2,28	4 ± 1,16	4,7838	0,8658

V. CONCLUSÕES

A aplicação dos modelos preditivos utilizando as RNRs possibilitou a imputação das séries temporais com diferentes

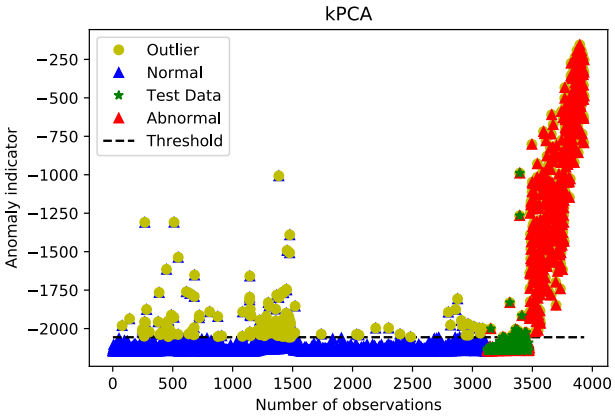

Fig. 7. Aplicação do algoritmo kPCA demonstrando a condição *baseline* da estrutura.

TABELA IX

DETECÇÃO DE ERROS T1 E T2 EM BASE IMPUTADA EM 50% POR LSTM.

Algoritmo	Erros T1	Erros T2	% T1	% T2
MSD	153 ± 2,18	215 ± 2,92	4,4092	46,5367
kPCA	156 ± 2,27	4 ± 1,04	4,4956	0,8658

TABELA X

DETECÇÃO DE ERROS T1 E T2 EM BASE IMPUTADA EM 75% POR LSTM.

Algoritmo	Erros T1	Erros T2	% T1	% T2
MSD	152 ± 1,51	234 ± 2,57	4,3804	50,6493
kPCA	154 ± 1,73	7 ± 1,00	4,4380	1,5151

TABELA XI

DETECÇÃO DE ERROS T1 E T2 EM BASE IMPUTADA EM 25% POR GRU.

Algoritmo	Erros T1	Erros T2	% T1	% T2
MSD	158 ± 1,74	210 ± 5,03	4,5533	45,4545
kPCA	166 ± 2,19	4 ± 0,89	4,7838	0,8658

TABELA XII

DETECÇÃO DE ERROS T1 E T2 EM BASE IMPUTADA EM 50% POR GRU.

Algoritmo	Erros T1	Erros T2	% T1	% T2
MSD	153 ± 2,18	233 ± 4,4	4,4092	50,4329
kPCA	156 ± 2,2	4 ± 0,87	4,4956	0,8658

TABELA XIII

DETECÇÃO DE ERROS T1 E T2 EM BASE IMPUTADA EM 75% POR GRU.

Algoritmo	Erros T1	Erros T2	% T1	% T2
MSD	152 ± 1,51	267 ± 5,70	4,3804	57,7922
kPCA	154 ± 1,34	7 ± 0,94	4,4380	1,5151

taxas de amputação na base de dados de SHM utilizada no trabalho. Os resultados encontrados neste trabalho indicam que com a mesma arquitetura de rede, quantidade de neurônios e configuração de camadas, a LSTM e GRU apresentam excelente resultado para imputação de dados, porém os resultados da LSTM são ligeiramente melhores. Este fato pode ser justificado pela capacidade em utilizar portas que permitem ajustes de peso, e assim, modificar as informações no tempo

de maneira que possa melhor prever os estados futuros. Essa característica torna as redes LSTM ideais para o processamento de dados sequenciais no tempo, característica da base de dados SHM utilizada neste trabalho.

O bom desempenho da imputação por LSTM é refletido também na etapa de detecção de dano, onde é possível observar a melhor performance em comparação ao *baseline* se levarmos em conta a imputação por GRU. Em termos de custo computacional a GRU apresenta melhor desempenho, menor tempo de treinamento, justificado pela ausência de um portão se comparado com a LSTM, porém, em termos de acurácia de previsão não se saiu tão bem quanto a LSTM.

Os modelo treinados apresentaram resultados satisfatórios mostrando que as redes neurais recorrentes constituem uma boa alternativa para reduzir os efeitos adversos causados pelos dados ausentes em um contexto de SHM.

REFERÊNCIAS

[1] H. N. Li, T. H. Yi, and P. Qiao, "Special issue on health monitoring technologies for civil infrastructure," *Journal of Aerospace Engineering*, vol. 30, no. 2, pp. 2–4, 2017.

[2] C. R. Farrar and K. Worden, *Structural Health Monitoring: A Machine Learning Perspective*, 2012.

[3] J. P. Lynch and P. T. R. S. A, "An overview of wireless structural health monitoring for civil structures An overview of wireless structural health monitoring for civil structures," pp. 345–372, 2007.

[4] Y. Yang and S. Nagarajaiah, "Harnessing data structure for recovery of randomly missing structural vibration responses time history: Sparse representation versus low-rank structure," *Mechanical Systems and Signal Processing*, vol. 74, pp. 165–182, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.ymssp.2015.11.009

[5] Y. Bao, H. Li, X. Sun, Y. Yu, and J. Ou, "Compressive sampling-based data loss recovery for wireless sensor networks used in civil structural health monitoring," *Structural Health Monitoring*, vol. 12, no. 1, pp. 78–95, 2013.

[6] M. G. Peter Schmitt e Jonas Mandel, "A Comparison of Six Methods for Missing Data Imputation," *Journal of Biometrics & Biostatistics*, vol. 06, no. 01, pp. 1–6, 2015.

[7] Z. Zou, Y. Bao, H. Li, B. F. Spencer, and J. Ou, "Embedding compressive sensing-based data loss recovery algorithm into wireless smart sensors for structural health monitoring," *IEEE Sensors Journal*, vol. 15, no. 2, pp. 797–808, 2015.

[8] Y. F. Luo, Z. W. Ye, X. N. Guo, X. H. Qiang, and X. M. Chen, "Data missing mechanism and missing data real-time processing methods in the construction monitoring of steel structures," *Advances in Structural Engineering*, vol. 18, no. 4, pp. 585–601, 2015.

[9] Z. Chen, Y. Bao, H. Li, and B. F. Spencer, "A novel distribution regression approach for data loss compensation in structural health monitoring," *Structural Health Monitoring*, vol. 17, no. 6, pp. 1473–1490, 2018.

[10] ——, "LQD-RKHS-based distribution-to-distribution regression methodology for restoring the probability distributions of missing SHM data," *Mechanical Systems and Signal Processing*, vol. 121, no. 451, pp. 655–674, 2019.

[11] G. Fan, J. Li, and H. Hao, "Lost data recovery for structural health monitoring based on convolutional neural networks," *Structural Control and Health Monitoring*, vol. 26, no. 10, pp. 1–21, 2019.

[12] N. Ahmad, N. A. Ghani, and A. A. Kamil, "Machine Learning-Based Missing Value Imputation Method for Clinical Datasets Machine Learning-Based Missing Value Imputation Method for Clinical Datasets," no. August, pp. 701–711, 2013.

[13] A. Mani, "Solving Text Imputation Using Recurrent Neural Networks," pp. 1–7, 2015.

[14] M. M. Ghazi, M. Nielsen, A. Pai, M. J. Cardoso, M. Modat, S. Ourselin, and L. Sørensen, "Robust training of recurrent neural networks to handle missing data for disease progression modeling," no. Midl, pp. 1–9, 2018. [Online]. Available: http://arxiv.org/abs/1808.05500

[15] H. G. Kim, G. J. Jang, H. J. Choi, M. Lim, and J. Choi, "Medical examination data prediction with missing information imputation based on recurrent neural networks," *International Journal of Data Mining and Bioinformatics*, vol. 19, no. 3, pp. 202–220, 2017.

[16] B. Peeters and G. De Roeck, "One-year monitoring of the Z24-bridge: Environmental effectsversus damage events," *Earthquake Engineering and Structural Dynamics*, vol. 30, no. 2, pp. 149–171, 2001.

[17] Figueiredo e Santos, "Vibration-Based Techniques for Damage Detection and Localization in Engineering Structures, pp. 1-39 (2018)," *An Automated Irrigation System Using Arduino Microcontroller*, vol. 1908, no. January, pp. 2–6, 2018.

[18] D. B. Rubin, "Inference and Missing Data," *Biometrika*, vol. 63, no. 3, pp. 581 – 592, 1976.

[19] C. A. Leke and T. Marwala, *Deep Learning and Missing Data in Engineering Systems*. Springer, 2019, vol. 48.

[20] Little e Rubin, *Statistical Analysis with Missing Data*. John Wiley & Sons, 2020.

[21] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent Neural Networks for Multivariate Time Series with Missing Values," *Scientific Reports*, vol. 8, no. 1, pp. 1–12, 2018. [Online]. Available: http://dx.doi.org/10.1038/s41598-018-24271-9

[22] R. T. Gonzalez and D. A. Couto Barone, "Using deep learning and evolutionary algorithms for time series forecasting," *ESANN 2019 - Proceedings, 27th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pp. 631–636, 2019.

[23] M. F. BARROS, "Análise e Previsão de Séries Temporais Utilizando Amortecimento Exponencial com Múltiplos Ciclos e Técnicas de Simulação na Produção de Energia Eólica," *Dissertação de Mestrado*, pp. 1–74, 2015.

[24] K. Worden, G. Manson, and N. R. J. Fieller, "Damage detection using outlier analysis," *Journal of Sound and Vibration*, vol. 229, no. 3, pp. 647–667, 2000.

[25] B. Schölkopf, A. Smola, and K.-R. Muller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[26] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. Pittsburgh PA, United States: ACM, 1992, pp. 144–152.

[27] D. Widjaja, C. Varon, A. Dorado, J. A. K. Suykens, and S. V. Huffel, "Application of kernel principal component analysis for single-lead-ECG-derived respiration," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 4, pp. 1169–1176, 2012.

[28] I. Jolliffe, *Principal Component Analysis*, 2nd ed. New York NY, United States: Springer-Verlag, 2002.

[29] P. R. Peres-Neto, D. A. Jackson, and K. M. Somers, "How many principal components? stopping rules for determining the number of non-trivial axes revisited," *Computational Statistics & Data Analysis*, vol. 49, no. 4, pp. 974–997, 2005.

[30] M. A. Pereira de Lima, C. Sales, A. Santos, R. Santos, M. Silva, J. Costa, M. Carvalho, and M. Cruz, "A framework for data compression and damage detection in structural health monitoring applied on a laboratory three-story structure," *Revista Brasileira de Computação Aplicada*, vol. 8, no. 2, p. 129, 2016.