

CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

MÉTODO DE IMPUTAÇÃO INCREMENTAL, SIMPLES E EFICIENTE PARA DADOS AUSENTES EM SISTEMAS EVOLUTIVOS

GIOVANNI AMORMINO DA SILVA JÚNIOR

Orientador: Prof. Alisson Marques da Silva
CEFET-MG

BELO HORIZONTE
NOVEMBRO DE 2020

GIOVANNI AMORMINO DA SILVA JÚNIOR

MÉTODO DE IMPUTAÇÃO INCREMENTAL, SIMPLES E EFICIENTE PARA DADOS AUSENTES EM SISTEMAS EVOLUTIVOS

Dissertação apresentada ao Programa de Pós-graduação em Modelagem Matemática e Computacional do Centro Federal de Educação Tecnológica de Minas Gerais, como requisito parcial para a obtenção do título de Mestre em Modelagem Matemática e Computacional.

Área de concentração: Modelagem Matemática e Computacional

Linha de pesquisa: Sistemas Inteligentes

Orientador: Prof. Alisson Marques da Silva
CEFET-MG

CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL
BELO HORIZONTE
NOVEMBRO DE 2020

Silva Júnior, Giovanni Amormino da
S586m Método de imputação incremental, simples e eficiente para dados ausentes em sistemas evolutivos / Giovanni Amormino da Silva Júnior. – 2020.
65 f.

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Modelagem Matemática e Computacional.
Orientador: Alisson Marques da Silva.
Dissertação (mestrado) – Centro Federal de Educação Tecnológica de Minas Gerais.

1. Estatísticas – Métodos – Teses. 2. Aleatória variáveis – Teses. 3. Análise de sistemas – Teses. 4. Sistemas difusos – Teses. 5. Processamento de sinais – Teses. I. Silva, Alisson Marques. II. Centro Federal de Educação Tecnológica de Minas Gerais. III. Título.

CDD 519.535

Elaboração da ficha catalográfica pela bibliotecária Jane Marangon Duarte, CRB 6º 1592 / Cefet/MG

SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
COORDENAÇÃO DO CURSO DE MESTRADO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

"MÉTODO DE IMPUTAÇÃO INCREMENTAL, SIMPLES E EFICIENTE PARA DADOS AUSENTES EM SISTEMAS EVOLUTIVOS"

Dissertação de Mestrado apresentada por **Giovanni Amormino da Silva Júnior**, em 30 de novembro de 2020, ao Programa de Pós-Graduação em Modelagem Matemática e Computacional do CEFET-MG, e aprovada pela banca examinadora constituída pelos professores:

Prof. Dr. Alisson Marques da Silva (Orientador)
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dr. Daniel Furtado Leite
Universidade Federal de Lavras

Prof. Dr. Eduardo Habib Bechelane Maia
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dr. Thiago de Souza Rodrigues
Centro Federal de Educação Tecnológica de Minas Gerais

Visto e permitida a impressão,

Prof. Dr. Thiago de Souza Rodrigues
Coordenador do Programa de Pós-Graduação em
Modelagem Matemática e Computacional

Dedico este trabalho a Deus e aos meus pais, por toda luta ao longo de suas vidas para que eu tivesse acesso a uma educação melhor.

Agradecimentos

Primeiramente a Deus e aos meus pais que sempre estiveram ao meu lado, sempre lutaram para me dar muito mais do que eles tiveram e sempre se dedicaram em fazer de mim uma versão melhor do que eles, doando parte de suas vidas em prol deste objetivo.

À minha avó, Ana, e minha tia, Adalva, que contribuiram muito com este objetivo dos meus pais e, assim como meus pais, foram fundamentais para minha formação moral e intelectual, participando ativamente da minha vida, nos ajudando nos momentos difíceis e me auxiliando no meu "primeiro emprego". Em nome de vocês agradeço a todos da família que também contribuíram.

À minha namorada, Gabriela, por todo companherismo, apoio e carinho todo esse tempo, não permitindo que eu desistisse e me mantivesse forte para alcançar os meus objetivos.

Ao meu orientador, Alisson, por ter me aceitado como seu orientando, sem nenhuma obrigação de fazê-lo e sem me conhecer, pela paciência, por ter me ajudado quando tive dificuldade, por me guiar pelo melhor caminho para que fizéssemos um bom trabalho e, principalmente, pela amizade. Aproveitando para agradecer também aos meus amigos do que fizeram parte desta jornada caminhando ao meu lado e tornando o trabalho mais leve.

Ao Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) e à Coordenação do Programa de Pós-Graduação em Modelagem Matemática e Computacional, por toda estrutura que envolve o curso, e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo apoio financeiro no início da pesquisa.

Aos meus antigos professores Moisés e Bráulio, que me deram suas recomendações sem as quais não poderia ter a oportunidade de iniciar este trabalho. Em nome deles também agradeço a todos os meus antigos professores que ao longo da minha jornada fizeram seu melhor para que hoje eu esteja neste ponto.

A todos que fazem parte da minha história, toda conquista minha sempre será também de vocês a quem sempre dedicarei o meu muito obrigado!

“Do you feel cold and lost in desperation
You build up hope but failure’s all you’ve known
Remember all the sadness and frustration
And let it go, let it go.”
(Iridescent - Linkin Park, 2010)

Resumo

Uma grande dificuldade enfrentada no desenvolvimento de aplicações que utilizam fluxos de dados para resolver problemas de previsão são os dados ausentes. Apesar de existirem técnicas para reduzir os impactos ocasionados por este problema, a maioria dos sistemas não são modelados de forma preventiva para possibilitar o tratamento adequado deste tipo de ocorrência. Neste contexto, este trabalho introduz uma nova abordagem denominada MDP (*Missing Data Procedure*) que possibilita que modelos evolutivos estimem as variáveis ausentes e, simultaneamente, realizem a previsão da saída utilizando todas as variáveis da amostra. A abordagem proposta foi implementada no eNFN (*evolving Neo-Fuzzy Neuron*) como um modelo de múltiplas entradas e múltiplas saídas e denominada eNFN-MDP. O MDP também foi implementado no ALMMo (*Autonomous Learning Multi-Model*), porém, como um modelo de múltiplas entradas e uma saída, chamado de ALMMo-MDP. Experimentos computacionais foram realizados em tarefas de previsão e identificação de sistemas não lineares em cenários de Ausência Completamente Aleatória (*Missing Completely at Random* – MCAR) e Ausência Aleatória (*Missing at Random* – MAR) para avaliar o desempenho do eNFN-MDP e do ALMMo-MDP. Os resultados obtidos foram comparados com métodos alternativos de imputação e mostraram que o eNFN-MDP e o ALMMo-MDP obtiveram um desempenho comparável ou superior ao das abordagens alternativas. Eles foram capazes de estimar as variáveis ausentes e realizar a previsão da saída com precisão. Portanto, os resultados experimentais sugerem a abordagem proposta como uma alternativa simples e eficiente como método de imputação para modelos evolutivos.

Palavras-chave: Dados Ausentes, Ausência Completamente Aleatória, Ausência Aleatória, Sistemas Evolutivos, Sistemas Fuzzy, Processamento *Online*.

Abstract

A great challenge in real-world applications driven use data streams to solve prediction problems is that of handling missing data. Although there are techniques to reduce the effects caused by this issue, most systems are not modeled in a preventive way to enable an adequate treatment of this type of occurrence. In this context, this work introduces a new approach called MDP (Missing Data Procedure). The proposed approach allow evolving systems to estimate missing values and, simultaneously, compute the output using all variables of the sample. The proposed approach was implemented in eNFN (evolving Neo-Fuzzy Neuron) as a model with multiple inputs and multiple outputs and called eNFN-MDP. MDP was implemented in ALMMo (Autonomous Learning Multi-Model) as a model of multiple inputs and single output, called ALMMo-MDP. Computational experiments were performed aiming at forecasting and nonlinear systems identification in scenarios of Missing Completely at Random (MCAR) and Missing at Random (MAR) to evaluate the performance of eNFN-MDP and ALMMo-MDP. The results obtained were compared with alternative imputation methods. We have noticed that eNFN-MDP and ALMMo-MDP have similar or better performance than alternative approaches. They can estimate the missing values and accurately predict the output. Therefore, the experimental results suggest the proposed approach as a simple and efficient alternative for data imputation in evolving modeling.

Keywords: Missing Data, Missing Completely at Random, Missing at Random, Evolving Systems, Fuzzy Systems, Online Processing.

Lista de Figuras

Figura 1 – Estrutura Genérica do *Missing Data Procedure* 19
Figura 2 – Abordagem MDP para modelos com múltiplas entradas e saídas. 20
Figura 3 – Abordagem MDP para modelos com múltiplas entradas e única saída. . 21
Figura 4 – Estrutura básica do eNFN-MDP. 22
Figura 5 – Estrutura básica de um gráfico *boxplot*. 34
Figura 6 – Boxplot do eNFN-MDP na previsão do nível da pressão sonora nos cenários MCAR. 37
Figura 7 – Boxplot do eNFN-MDP na previsão do nível da pressão sonora nos cenários MAR. 37
Figura 8 – Boxplot do ALMMo-MDP na previsão do nível da pressão sonora nos cenários MCAR. 39
Figura 9 – Boxplot do ALMMo-MDP na previsão do nível da pressão sonora nos cenários MAR. 39
Figura 10 – Boxplot dos MDPs na previsão do nível da pressão sonora nos cenários MCAR. 40
Figura 11 – Boxplot dos MDPs na previsão do nível da pressão sonora nos cenários MAR. 41
Figura 12 – Boxplot do eNFN-MDP na previsão da coordenada atômica calculada w' nos cenários MCAR. 43
Figura 13 – Boxplot do eNFN-MDP na previsão da coordenada atômica calculada w' nos cenários MAR. 44
Figura 14 – Boxplot do ALMMo-MDP na previsão da coordenada atômica calculada w' nos cenários MCAR. 45
Figura 15 – Boxplot do ALMMo-MDP na previsão da coordenada atômica calculada w' nos cenários MAR. 46
Figura 16 – Boxplot dos MDPs na previsão da coordenada atômica calculada w' nos cenários MCAR. 47
Figura 17 – Boxplot dos MDPs na previsão da coordenada atômica Calculada w' nos cenários MAR. 48
Figura 18 – Boxplot do eNFN-MDP na previsão da umidade absoluta nos cenários MCAR. 49
Figura 19 – Boxplot do eNFN-MDP na previsão da umidade absoluta nos cenários MAR. 50
Figura 20 – Boxplot do ALMMo-MDP na previsão da umidade absoluta nos cenários MCAR. 51

Figura 21 – Boxplot do ALMMo-MDP na previsão da umidade absoluta nos cenários MAR. 52
Figura 22 – Boxplot dos MDPs na previsão da umidade absoluta nos cenários MCAR. 53
Figura 23 – Boxplot dos MDPs na previsão da umidade absoluta nos cenários MAR. 54
Figura 24 – Boxplot do eNFN-MDP na identificação de processo não linear nos cenários MCAR. 55
Figura 25 – Boxplot do eNFN-MDP na identificação de processo não linear nos cenários MAR. 56
Figura 26 – Boxplot do ALMMo-MDP na identificação de processo não linear nos cenários MCAR. 57
Figura 27 – Boxplot do ALMMo-MDP na identificação de processo não linear nos cenários MAR. 58
Figura 28 – Boxplot dos MDPs na identificação de processo não linear nos cenários MCAR. 59
Figura 29 – Boxplot dos MDPs na identificação de processo não linear nos cenários MAR. 60

Lista de Tabelas

Tabela 1 – Desempenho RMSE do eNFN-MDP na previsão do nível da pressão sonora nos cenários MCAR. 36
Tabela 2 – Desempenho RMSE do eNFN-MDP na previsão do nível da pressão sonora nos cenários MAR. 36
Tabela 3 – Desempenho RMSE do ALMMo-MDP na previsão do nível da pressão sonora nos cenários MCAR. 38
Tabela 4 – Desempenho RMSE do ALMMo-MDP na previsão do nível da pressão sonora nos cenários MAR. 39
Tabela 5 – Desempenho RMSE dos MDPs na previsão do nível da pressão sonora nos cenários MCAR. 40
Tabela 6 – Desempenho RMSE dos MDPs na previsão do nível da pressão sonora nos cenários MAR. 41
Tabela 7 – Desempenho RMSE do eNFN-MDP na previsão da coordenada atômica calculada w' nos cenários MCAR. 42
Tabela 8 – Desempenho RMSE do eNFN-MDP na previsão da coordenada atômica calculada w' nos cenários MAR. 43
Tabela 9 – Desempenho RMSE do ALMMo-MDP na previsão da coordenada atômica calculada w' nos cenários MCAR. 44
Tabela 10 – Desempenho RMSE do ALMMo-MDP na previsão da coordenada atômica calculada w' nos cenários MAR. 45
Tabela 11 – Desempenho RMSE dos MDPs na previsão da coordenada atômica calculada w' nos cenários MCAR. 46
Tabela 12 – Desempenho RMSE dos MDPs na previsão da coordenada atômica calculada w' nos cenários MAR. 47
Tabela 13 – Desempenho RMSE do eNFN-MDP na previsão da umidade absoluta nos cenários MCAR. 48
Tabela 14 – Desempenho RMSE do eNFN-MDP na previsão da umidade absoluta nos cenários MAR. 49
Tabela 15 – Desempenho RMSE do ALMMo-MDP na previsão da umidade absoluta nos cenários MCAR. 50
Tabela 16 – Desempenho RMSE do ALMMo-MDP na previsão da umidade absoluta nos cenários MAR. 51
Tabela 17 – Desempenho RMSE dos MDPs na previsão da umidade absoluta nos cenários MCAR. 52
Tabela 18 – Desempenhos RMSE do MDPs na previsão da umidade absoluta nos cenários MAR. 53

Tabela 19 – Desempenho RMSE do eNFN-MDP na identificação de processo não linear nos cenários MCAR. 54
Tabela 20 – Desempenho RMSE do eNFN-MDP na identificação de processo não linear nos cenários MAR. 55
Tabela 21 – Desempenho RMSE do ALMMo-MDP na identificação de processo não linear nos cenários MCAR. 56
Tabela 22 – Desempenho RMSE do ALMMo-MDP na identificação de processo não linear nos cenários MAR. 57
Tabela 23 – Desempenho RMSE dos MDPs na identificação de processo não linear nos cenários MCAR. 58
Tabela 24 – Desempenho RMSE dos MDPs na identificação de processo não linear nos cenários MAR. 59

Lista de Algoritmos

Algoritmo 1 – Procedimento de Validação do *Missing Data Procedure* MIMO. 23

Algoritmo 2 – Procedimento para Criação de Funções de Pertinência 27

Algoritmo 3 – Procedimento para Exclusão de Funções de Pertinência 28

Algoritmo 4 – *eNFN with Missing Data Procedure*. 29

Algoritmo 5 – ALMMo *with Missing Data Procedure*. 31

Algoritmo 6 – Gerador de Conjuntos de Dados MCAR 33

Algoritmo 7 – Gerador de Conjuntos de Dados MAR 34

Sumário

1 – Introdução 1
1.1 Motivação e Relevância 1
1.2 Objetivos 2
1.2.1 Objetivo Geral 2
1.2.2 Objetivos Específicos 3
1.3 Contribuições do Trabalho 3
1.4 Publicações 3
1.5 Organização do Trabalho 4

2 – Fundamentação Teórica 5
2.1 Dados Ausentes (*Missing Data*) 5
2.2 Técnicas de Deleção 7
2.3 Técnicas de Imputação 7
2.3.1 *Expectation-Maximization* (EM) 8
2.3.2 *Regularized Expectation Maximization* (RegEM) 10
2.3.3 *K-Nearest Neighbors Input* (KNNI) 10
2.3.4 *Singular Value Decomposition Input* (SVDI) 11
2.3.5 *Bayesian Principal Component Analysis* (BPCA) 12
2.3.6 *Local Least Squares Imputation* (LLSimput) 13
2.3.7 *K-Mean* e *Fuzzy K-Mean Imputation* (KMI) 13
2.3.8 *Global Most Common Attribute* (GMC) 14
2.3.9 *Fuzzy-Rough Nearest Neighbor Imputation* (FRNNI) 14
2.3.10 Evolving Fuzzy Granular Predictor (eFGP) 16

3 – Modelagem Evolutiva para Estimação de Variáveis Ausentes 19
3.1 Missing Data Procedure (MDP) 19
3.2 eNFN *with Missing Data Procedure* (eNFN-MDP) 21
3.2.1 *Missing Data Procedure* - MIMO 23
3.2.2 Estimação das Variáveis e Cálculo da Saídas 23
3.2.3 Atualização dos Parâmetros 24
3.2.4 Criação de Funções de Pertinências 25
3.2.5 Exclusão de Funções de Pertinência 26
3.2.6 Hiperparâmetros e Algoritmo do eNFN-MDP 28
3.3 ALMMo *with Missing Data Procedure* (ALMMo-MDP) 28
3.3.1 *Missing Data Procedure* - MISO 30

4 – Experimentos Computacionais 32
4.1 Metodologia dos Experimentos 32
4.1.1 Construção de Conjuntos de Dados com Variáveis Ausentes 32
4.1.2 Medidas de Desempenho 33
4.2 Previsão do Nível de Pressão Sonora - Airfoil Self-Noise 35
4.2.1 eNFN-MDP na Previsão do Nível de Pressão Sonora 36
4.2.2 ALMMo-MDP na Previsão do Nível de Pressão Sonora 38
4.2.3 MDPs na Previsão do Nível de Pressão Sonora 40
4.3 Previsão da Coordenada Atômica Calculada w' - Carbon Nanotubes 41
4.3.1 eNFN-MDP na Previsão da Coordenada Atômica Calculada w' . . . 42
4.3.2 ALMMo-MDP na Previsão da Coordenada Atômica Calculada w' . . 43
4.3.3 MDPs na Previsão da Coordenada Atômica Calculada w' 45
4.4 Previsão da Umidade Absoluta - Air Quality 47
4.4.1 eNFN-MDP na Previsão da Umidade Absoluta 48
4.4.2 ALMMo-MDP na Previsão da Umidade Absoluta 50
4.4.3 MDPs na Previsão da Umidade Absoluta 51
4.5 Identificação de Processo Não Linear 53
4.5.1 eNFN-MDP na Identificação de Processo Não Linear 54
4.5.2 ALMMo-MDP na Identificação de Processo Não Linear 56
4.5.3 MDPs na Identificação de Processo Não Linear 57

5 – Considerações Finais 61

Referências 62

1 Introdução

Neste capítulo inicial apresenta-se, na próxima seção, uma descrição inicial do problema abordado neste trabalho, com os desafios e as motivações que norteiam esta pesquisa. Na Seção 1.2 são apresentados o objetivo geral e os objetivos específicos a serem alcançados. A Seção 1.3 destaca as contribuições e a Seção 1.4 as publicações, frutos do desenvolvimento deste trabalho. Por fim, na Seção 1.5, é descrito sucintamente o conteúdo de cada um dos capítulos.

1.1 Motivação e Relevância

Uma grande dificuldade enfrentada no desenvolvimento de aplicações do mundo real que utilizam dados para resolver problemas de previsão e identificação de sistemas são os dados ausentes. A ocorrência de dados ausentes pode se dar por vários motivos, como sensores não confiáveis, observações incompletas, oclusão parcial do sinal desejado, interferência na comunicação, restrição da banda de transmissão de dados, dentre outros (COOKE; MORRIS; GREEN, 1997; AMIRI; JENSEN, 2016). Independente da causa, a ocorrência de dados ausentes tem sérias implicações para a descoberta de conhecimento e para o uso de algoritmos em aplicações de previsão e identificação de sistemas (FARHANGFAR; KURGAN; PEDRYCZ, 2007; SANTOS et al., 2019).

Essencialmente, os dados ausentes podem ser tratados de duas maneiras: (i) remover as variáveis ausentes ou amostras com variáveis ausentes (GELMAN; HILL, 2006); (ii) imputar valores para as variáveis ausentes (KANG, 2013). Diversos métodos de imputação foram propostos, como a imputação de zero (AMIRI; JENSEN, 2016), da média (GARCIA; LEITE; SKRJANC, 2019), da mediana (GARCIA et al., 2019), dentre outros. Métodos estatísticos e técnicas de aprendizado de máquina também são utilizados para estimar o valor da variável ausente (GELMAN; HILL, 2006; ENDERS; BARALDI, 2018). Após a imputação, o conjunto de dados pode ser analisado normalmente, como uma base completa (AMIRI; JENSEN, 2016).

Os métodos de imputação que utilizam análises estatísticas e/ou técnicas de aprendizado de máquina, em sua maioria, foram propostos para aplicações no contexto de dados estáticos, nos quais os dados coletados em um intervalo de tempo específico são suficientes para representar o comportamento do processo no futuro (GARCIA; LEITE; SKRJANC, 2019). Embora estes métodos possam ser utilizados para solucionar o problema de dados ausentes, geralmente eles possuem algumas necessidades que os limitam, tais como:

- ter disponível previamente o conjunto de dados completo (OSMAN; ABU-MAHFOUZ;

PAGE, 2018);

- necessidade de um valor inicial para ser otimizado (AMIRI; JENSEN, 2016);
- processar as amostras várias vezes para determinar os melhores parâmetros do modelo de imputação (AMIRI; JENSEN, 2016; GARCIA et al., 2019);
- não lidar com situações de variações nas condições ambientais do sistema (mudança de conceito), típicas de ambientes *online* (GARCIA; LEITE; SKRJANC, 2019).

Em ambientes em que os dados chegam em fluxo e estão sujeitos a mudanças de conceito são necessários sistemas com processamento *online*, com uma única passagem nos dados, e que possuam mecanismos para lidar com as variações nos dados (LUGHOFER; ANGELOV, 2011). Os sistemas com essas características são chamados de sistemas evolutivos (ŠKRJANC et al., 2019).

Os sistemas evolutivos possuem a característica de desenvolver sua estrutura simultaneamente ao ajuste dos parâmetros à medida que novas amostras são apresentadas (SILVA et al., 2014). Estes sistemas possuem um mecanismo de aprendizado incremental e utilizam apenas a amostra de dados atual, ou seja, as amostras do fluxo de dados são processadas uma única vez e, então, são descartadas (SILVA, 2014). Destaca-se que a grande maioria dos sistemas evolutivos necessitam que os valores de todas as variáveis estejam disponíveis para funcionar corretamente (GARCIA et al., 2019).

Apesar da existência de técnicas para reduzir os impactos ocasionados por dados ausentes, a maioria dos sistemas não são modelados de forma preventiva para possibilitar o tratamento adequado deste tipo de problema (SCHAFER; GRAHAM, 2002; KANG, 2013). Neste contexto, este trabalho introduz uma abordagem simples e efetiva, chamada de MDP (*Missing Data Procedure*), que possibilita que modelos evolutivos realizem a previsão da saída e, simultaneamente, estimem o valor das variáveis ausentes. O MDP foi proposto para aplicações em ambientes não estacionários, envolvendo condições ambientais voláteis, ou seja, ambientes em que os dados chegam em forma de fluxo e estão sujeitos a variações e mudanças de conceito. No entanto, destaca-se que os modelos com o MDP também podem ser empregados em ambientes com dados estáticos.

1.2 Objetivos

1.2.1 Objetivo Geral

O objetivo geral deste trabalho é propor e implementar uma nova abordagem para tratar dados ausentes, em problemas de previsão e identificação de sistemas, que possa ser incorporada a modelos evolutivos.

1.2.2 Objetivos Específicos

Os objetivos específicos deste trabalho são:

- Realizar uma revisão da literatura relacionada ao problema de dados ausentes;
- Propor e implementar algoritmos que gerem bases com dados ausentes para os cenários de Ausência Completamente Aleatória (*Missing Completely at Random* – MCAR) e Ausência Aleatória (*Missing at Random* – MAR);
- Propor e implementar um método de imputação para modelos evolutivos com múltiplas entradas e múltiplas saídas;
- Propor e implementar um método de imputação para modelos evolutivos com múltiplas entradas e uma saída;
- Realizar experimentos computacionais para avaliar e comparar os métodos propostos com métodos do estado da arte em cenários MCAR e MAR em problemas de previsão e identificação de sistemas.

1.3 Contribuições do Trabalho

As principais contribuições deste trabalho são:

- Proposição de uma abordagem simples e efetiva para imputação de valores em ambientes com dados ausentes. A abordagem proposta pode ser implementada na maioria dos sistemas evolutivos, sejam eles com uma ou múltiplas saídas, e possibilita a imputação de valores para amostras com uma ou mais variáveis ausentes.
- Implementação da abordagem proposta no modelo *fuzzy* evolutivo com múltiplas entradas e múltiplas saídas eNFN (*evolving Neo-Fuzzy Neuron*) (SILVA et al., 2014; SILVA, 2014).
- Implementação da abordagem proposta no modelo *fuzzy* evolutivo com múltiplas entradas e uma única saída ALMMo (*Autonomous Learning Multi-Model*) (ANGELOV; GU; PRÍNCIPE, 2017).

1.4 Publicações

O presente trabalho gerou as seguintes publicações:

- ***Missing Data: Brief Review and Case Studies***. CILAMCE 2019 - *Proceedings of the XL Ibero-Latin-American Congress on Computational Methods in Engineering*, ABMEC. Natal/RN, Brasil, 2019 (Publicado).
- **Dados Ausentes em Problemas de Previsão: Uma Breve Revisão e Estudos de Casos**. CBIC 2019 - *Proceedings of the 14th Brazilian Computational Intelligence Meeting*. Belém/PA, Brasil, 2019 (Publicado).

- ***A Simple and Efficient Incremental Missing Data Imputation Method for Evolving Neo-Fuzzy Network***. *Journal Evolving Systems.* 2020 (Submetido).

O primeiro artigo apresenta uma revisão de literatura sobre dados ausentes e ilustra experimentos computacionais realizados para avaliar os modelos evolutivos eMG, eTS e xTS e os métodos de imputação RegEM e Zero em problemas de regressão. O segundo artigo trata de uma revisão da literatura sobre processamento *online* de dados ausentes e apresenta experimentos computacionais com os modelos evolutivos ALMMo, eMG e eTS e os métodos de imputação Zero, Média e Último Valor. Por fim, o terceiro artigo introduz a abordagem proposta (MDP) aplicada ao modelo evolutivo eNFN. Os experimentos realizados neste artigo comparam o eNFN-MDP aos modelos eMG e eNFN empregando as técnicas de imputação Zero, Última Saída, Último Valor, Média e ao modelo eFGP que possui um mecanismo para tratar dados ausentes.

1.5 Organização do Trabalho

O presente trabalho é divido em 5 capítulos:

- Este capítulo inicial discorre com as motivações e os objetivos que tangenciam esta pesquisa. Além disso, apresenta as principais contribuições e as publicações decorrentes do seu desenvolvimento.
- O Capítulo 2 trata da revisão da literatura sobre dados ausentes. Inicialmente, são apresentados os conceitos fundamentais acerca deste tema e, em seguida, algumas das principais técnicas utilizadas para minimizar os problemas causados pelos dados ausentes.
- O capítulo 3 introduz a abordagem proposta. Inicialmente, detalha-se a abordagem proposta e como ela pode ser utilizada em modelos com uma única saída e em modelos com múltiplas saídas. Posteriormente, apresenta-se a implementação da abordagem proposta nos modelos evolutivos eNFN (SILVA et al., 2014; SILVA, 2014) e ALMMo (ANGELOV; GU; PRÍNCIPE, 2017).
- O Capítulo 4 ilustra experimentos computacionais realizados para a avaliar a abordagem proposta implementada nos modelos evolutivos eNFN e ALMMo. Os resultados obtidos são avaliados e comparados com resultados de métodos alternativos de imputação em tarefas de previsão e identificação de sistemas. As abordagens propostas também são comparadas com um modelo evolutivo, chamado eFGP (*evolving Fuzzy Granular Predictor*) (GARCIA; LEITE; SKRJANC, 2019; GARCIA et al., 2019), que possui um mecanismo para tratar dados ausentes.
- Por fim, o Capítulo 5 traz as considerações finais do trabalho e sugere as propostas de continuidade desta pesquisa.

2 Fundamentação Teórica

Neste capítulo é realizada uma revisão de literatura relacionada a dados ausentes. A Seção 2.1 discorre sobre os tipos de dados ausentes e a classificação quanto a aleatoriedade de sua ocorrência. A Seção 2.2 apresenta uma breve revisão das técnicas de deleção, enquanto a Seção 2.3 revisa as técnicas de imputação.

2.1 Dados Ausentes (*Missing Data*)

Na literatura é possível encontrar definições e interpretações ligeiramente diferentes sobre os tipos de dados ausentes. Contudo, no geral, os pesquisadores trabalham basicamente com três definições, classificadas de acordo com a forma do desaparecimento do dado, são elas: Ausência Completamente Aleatória (*Missing Completely at Random* – MCAR), Ausência Aleatória (*Missing at Random* – MAR) e Ausência Não Aleatória (*Not Missing at Random* – NMAR)[1] (KRAUSE et al., 2018; OSMAN; ABU-MAHFOUZ; PAGE, 2018).

MCAR é definido quanto a probabilidade de uma variável estar ausente não possuir qualquer relação com seu valor, com outras variáveis ou amostras do conjunto de dados (OSMAN; ABU-MAHFOUZ; PAGE, 2018). Em outras palavras, as variáveis são independentes entre si e a probabilidade de uma variável estar ausente é completamente aleatória (LITTLE et al., 2016). Portanto, a probabilidade de ausência de uma variável é a mesma que a probabilidade das demais (PEDERSEN et al., 2017). Este tipo de ausência é comum quando ocorrem falhas na coleta, um equipamento não funciona corretamente ou fica temporariamente indisponível, por exemplo (OSMAN; ABU-MAHFOUZ; PAGE, 2018; GARCIA; LEITE; SKRJANC, 2019). Estar em um cenário com MCAR, apesar dos problemas ocasionados pela ausência, possui uma vantagem estatística que faz com que a análise permaneça imparcial, uma vez que não há relação entre os valores ausentes e outras variáveis ou amostras no conjunto de dados. Assim, mesmo que ainda ocorra perda de poder estatístico ou analítico, os parâmetros ou valores estimados não são influenciados pela ausência (KANG, 2013). Segundo Osman, Abu-Mahfouz e Page (2018), o MCAR pode ser representado estatisticamente por:

$$f(M \mid Y,\ \phi) = f(M \mid \phi)\ \forall\ Y,\ \phi, \tag{1}$$

onde Y e M denotam um vetor de valores de dados observados e um vetor de indicadores de ausências, respectivamente, ϕ é um parâmetro desconhecido, e a função f indica a

[1] Também conhecida como *Missing Not Random* - MNAR.

distribuição de probabilidade condicional.

Por outro lado, no MAR a probabilidade de ausência de uma variável continua aleatória e independente de seus valores, contudo essa ausência possui relação com outras variáveis (KRAUSE et al., 2018; GARCIA et al., 2019). Para Kang (2013), a probabilidade de que uma variável esteja ausente depende do conjunto de variáveis observadas, mas não está relacionada aos seus valores. Uma das consequências disso é fazer com que, em uma amostra, a probabilidade de uma variável estar ausente seja maior que a probabilidade das demais variáveis estarem ausentes (SCHAFER; GRAHAM, 2002). É importante ressaltar que a ausência é condicional a outras variáveis, mas a ocorrência deste tipo de ausência é comum quando existem variáveis que dependem de processamentos ou seleções anteriores para definirem seus valores (ENDERS; BARALDI, 2018). Segundo Osman, Abu-Mahfouz e Page (2018), o MAR pode ser representado estatisticamente por:

$$f(M \mid Y,\ \phi) = f(M \mid Y_{obs},\ \phi)\ \forall\ Y_{miss},\ \phi, \tag{2}$$

onde Y_{obs} e Y_{miss} são os componentes observados e ausentes da variável de destino Y. O parâmetro desconhecido ϕ pode ser estimado relacionando Y_{obs} com outras variáveis.

Por fim, para o NMAR, no qual as características da ausência não se enquadram nos descritos anteriormente, é possível encontrar na literatura algumas divergências com relação a sua definição. Para alguns autores, a probabilidade de uma variável estar ausente depende dos potenciais valores desta variável, ou seja, a ausência está diretamente relacionada ao próprio valor ausente ou a natureza da variável (PEDERSEN et al., 2017; ENDERS; BARALDI, 2018; KRAUSE et al., 2018; OSMAN; ABU-MAHFOUZ; PAGE, 2018). Para outros autores, as variáveis ausentes dependem de outras variáveis ausentes (AMIRI; JENSEN, 2016). Normalmente, a obtenção de estimativas precisas para este tipo de mecanismo tende a ser mais complexa, pois não se tem as informações necessárias para especificar corretamente modelos das ausências (ENDERS; BARALDI, 2018). Segundo Osman, Abu-Mahfouz e Page (2018), o NMAR pode ser representado estatisticamente por:

$$f(M,\ Y \mid \theta,\ \phi) = f(Y \mid \theta)\ f(M \mid Y,\ \phi), \tag{3}$$

onde θ é um parâmetro da distribuição de Y que é estimado a partir dos dados observados e ϕ é um parâmetro que caracteriza a distribuição do padrão de ausência.

Há várias maneiras de tratar dados ausentes e, de uma forma intuitiva, é possível citar as seguintes maneiras: exclusão do registro, onde toda a amostra ou parte dela é descartada ou omitida, ou a inserção do valor. A Seção 2.2 detalha as técnicas de deleção e a Seção 2.3 as de imputação.

2.2 Técnicas de Deleção

Uma das técnicas utilizadas no tratamento de dados ausentes é a deleção ou omissão de amostras com valores ausentes. Existem duas técnicas principais de deleção (GELMAN; HILL, 2006; KANG, 2013): (i) a deleção da amostra que contém variáveis com dados ausentes; (ii) a deleção da variável com dado ausente. A deleção da amostra simplesmente deleta ou ignora todas as variáveis da amostra que contém algum valor ausente (OSMAN; ABU-MAHFOUZ; PAGE, 2018). Por outro lado, a deleção da variável com valor ausente, apesar de também envolver deleção ou omissão de valores, se trata de uma técnica mais seletiva e menos radical quando comparada a técnica anterior (ZHANG, 2016). Neste caso, a amostra não é totalmente deletada e somente a variável com valor ausente é descartada do processamento, fazendo com que todas as demais possam ser devidamente computadas (GELMAN; HILL, 2006). Assim, esta técnica tenta minimizar a perda de dados e é eficaz quando o tamanho total do conjunto de dados é pequeno ou quando o número de amostras com dados ausentes é pequeno (SCHNEIDER, 2001; KANG, 2013).

No contexto onde todas as amostras do conjunto de dados estão disponíveis *a priori*, as técnicas que utilizam exclusão são recomendadas somente em casos nos quais a quantidade de amostras com dados ausentes é pequena, não exista uma relação entre as variáveis e a deleção não cause um grande impacto na matriz de covariância do sistema (KANG, 2013). Em contrapartida, nas aplicações em que as amostras chegam na forma de um fluxo de dados e os conjuntos de dados são potencialmente ilimitados (número de amostras) não há como prever o impacto da exclusão de uma amostra.

2.3 Técnicas de Imputação

Na literatura relacionada a dados ausentes é comum a utilização do termo imputação relacionado somente a métodos de estimativa e previsão de variáveis com valores ausentes. Contudo, neste trabalho, este termo também é considerado para os métodos tradicionais, como a imputação por zero ou pela média, por exemplo, uma vez que estes métodos também recorrem a imputação de algum valor em variáveis com valores ausentes. A imputação por zero é uma das técnicas de imputação mais simples e uma das mais utilizadas (AMIRI; JENSEN, 2016). Este método apresenta bons resultados quando o número de ausências é baixo com relação ao total de dados. Porém, seus resultados tendem a piorar consideravelmente à medida que a porcentagem de ausências aumenta (PRATAMA et al., 2016).

A imputação pela média, na qual o valor da variável ausente é substituído pelo valor médio da variável, também é uma técnica amplamente utilizada (GELMAN; HILL, 2006; KANG, 2013). Para Kang (2013), embora teoricamente a média seja uma estimativa razoável para imputar o valor de uma variável com valor ausente por motivos aleatórios,

na presença de uma grande quantidade de dados ausentes ou na falta de aleatoriedade destas ausências, a utilização da média pode acarretar resultados inconsistentes.

Outra técnica de imputação é a substituição do valor ausente de uma variável pelo seu último valor (GELMAN; HILL, 2006; KANG, 2013). Por assumir que o valor de uma variável não se altera ou pouco se altera entre duas amostras de dados, está técnica pode produzir uma estimativa tendenciosa (KANG, 2013). Uma técnica alternativa trata da utilização da última saída do algoritmo de previsão (GARCIA; LEITE; SKRJANC, 2019). Nesta técnica, ao invés de imputar um valor para a variável ausente e, após a imputação, realizar a previsão, a saída prevista para a última amostra sem dado ausente é replicada para a amostra que possui variáveis com valor ausente (GARCIA et al., 2019).

Algumas técnicas de imputação utilizam análises estatísticas e/ou técnicas de aprendizado de máquina para estimar o valor da variável ausente e, posteriormente, substituir o valor ausente pelo valor estimado (GARCIA et al., 2019). Geralmente, essa estimativa é realizada por meio da análise do relacionamento entre as variáveis de entrada e saída (KANG, 2013; ZHANG, 2016), ou entre uma amostra e suas variáveis associadas (OSMAN; ABU-MAHFOUZ; PAGE, 2018). Dentre os métodos estatísticos pode-se citar, métodos de regressão (KANG, 2013), *Expectation Maximization* (DEMPSTER; LAIRD; RUBIN, 1977), *Regularized Expectation Maximization* (RegEM) (SCHNEIDER, 2001), *Global Most Common Attribute* (GMC) (GRZYMALA-BUSSE et al., 2005), *Hot-Deck* (MYERS, 2011), *Cold-Deck* (KANG, 2013), Máxima Verossimilhança (DEMPSTER; LAIRD; RUBIN, 1977), por exemplo. Algumas técnicas de aprendizado de máquina aplicadas na estimativa de valores de variáveis com dados ausentes citadas na literatura são *K-Nearest-Neighbor Input* (KNNI) (TROYANSKAYA et al., 2001), *Singular Value Decomposition Input* (SVDI) (TROYANSKAYA et al., 2001), *Bayesian Principal Component Analysis* (BPCA) (OBA et al., 2003), *Local Least Squares Imputation* (LLSimput) (KIM; GOLUB; PARK, 2004), *Fuzzy K-Mean Imputation* (KMI) (LI et al., 2004), *Fuzzy-Rough Nearest Neighbor Imputation* (FRNNI) (AMIRI; JENSEN, 2016) e *Evolving Fuzzy Granular Predictor* (eFGP) (GARCIA; LEITE; SKRJANC, 2019; GARCIA et al., 2019). Nas próximas seções algumas dessas técnicas são detalhadas.

2.3.1 *Expectation-Maximization* (EM)

Proposto por Dempster, Laird e Rubin (1977), o EM (*Expectation-Maximization*) é uma abordagem geral para o cálculo iterativo de estimativas de máxima verossimilhança quando as amostras podem conter dados ausentes. Cada iteração do algoritmo consiste em uma etapa de expectativa seguida de uma etapa de maximização, justificando o nome dado ao algoritmo. De forma geral, os parâmetros de uma distribuição de probabilidade são estimados a partir de dados incompletos, maximizando iterativamente a probabilidade dos dados disponíveis serem vistos como uma função dos parâmetros (DEMPSTER; LAIRD; RUBIN, 1977; SCHNEIDER, 2001).

Seja $\mathbf{X} \in \Re^{n \times p}$ uma matriz de dados com n amostras e p variáveis, com o valor ausente para algumas variáveis em algumas amostras. Destes conjuntos de dados incompletos, a média $\mu \in \Re^{1 \times p}$ de amostras e a matriz de covariância $\sum \in \Re^{p \times p}$ das variáveis são estimadas. Para um determinado registro, $\mathbf{x} = \mathbf{X}_i$, com valor ausente, o vetor $x_a \in \Re^{1 \times p_a}$ que consiste de p_a variáveis em que, na amostra dada, os valores estão disponíveis, e o vetor $\mathbf{x}_m \in \Re^{1 \times p_m}$ que consiste nas p_m variáveis restantes onde, na amostra dada, os valores estão ausentes. Seja a média μ dividida em duas: $\mu_a \in \Re^{1 \times p_a}$, com a média dos valores das variáveis com valores disponíveis, e $\mu_m \in \Re^{1 \times p_m}$, com a média dos valores das variáveis com os valores ausentes. Para cada amostra com valor ausente, a relação entre as variáveis com valores ausentes e as variáveis com valores disponíveis é modelada pela regressão linear definida por:

$$\mathbf{X}_m = \mu_m + (\mathbf{x}_a - \mu_a)\mathbf{B} + \mathbf{e}, \tag{4}$$

onde $\mathbf{B} \in \Re^{p_a \times p_m}$ é uma matriz dos coeficientes de regressão, e $\mathbf{e} \in \Re^{1 \times p_m}$ é um vetor aleatório com média zero. Em cada iteração, o algoritmo estima a média μ e a matriz de covariância. As estimativas de máxima verossimilhança condicional da matriz dos coeficientes de regressão $\mathbf{B}$ e da matriz de covariância do resíduo são calculadas para cada amostra com valores ausentes. Com o modelo de regressão estimado para cada amostra, os valores ausentes são então preenchidos com valores estimados, e novas estimativas da média μ e da matriz de covariância são calculados a partir do conjunto de dados completo e da matriz de covariância residual.

Depois que, em todas as amostras, os valores ausentes são preenchidos com os valores estimados, a média da amostra do conjunto de dados concluído é uma nova estimativa da média das amostras. Uma nova estimativa da matriz de covariância segue a expectativa condicional dos produtos cruzados, conforme:

$$\hat{\sum}^{(t+1)} = \frac{1}{\tilde{n}} \sum_{i=1}^{n} \{\hat{\mathbf{S}}_i^t - [\hat{\mu}^{(t+1)}]^T \hat{\mu}^{(t+1)}\}, \tag{5}$$

onde, para cada amostra, a expectativa condicional $\hat{S}_i^t$ dos produtos cruzados é composta por três partes. A constante de normalização $\tilde{n}$ da estimativa da matriz de covariância é o número de graus de liberdade da matriz de covariância da amostra do conjunto de dados completo (SCHNEIDER, 2001).

A iteração seguinte do algoritmo EM é realizada com a atualização das estimativas $\hat{\mu}^{(t+1)}$ e $\hat{\sum}^{(t+1)}$ da média e da matriz de covariância. A iteração é interrompida quando o algoritmo converge, ou seja, quando as estimativas $\hat{\mu}^t$ e $\hat{\sum}^t$ e os valores inseridos $\hat{x}_m$ param de ter grandes mudanças (SCHNEIDER, 2001).

Segundo Schneider (2001), o algoritmo EM e os métodos derivados dele são aplicáveis apenas aos conjuntos de dados nos quais os valores ausentes estão ausentes aleatoriamente e o número de amostras deve ser bem maior que o número de variáveis, de modo que a matriz de covariância da amostra do conjunto de dados preenchida com valores imputados seja positiva.

2.3.2 *Regularized Expectation Maximization* (RegEM)

O RegEM[2], proposto por Schneider (2001), teve como ponto de partida o algoritmo EM (*Expectation Maximization*), citado anteriormente na Seção 2.3.1. Ele foi desenvolvido objetivando ser aplicado a bases de dados climáticas incompletas, nos quais o número de variáveis normalmente excede o número de registros.

Os parâmetros de regressão regularizados do RegEM são calculados com um método conhecido como *Ridge Regression*, por estatísticos, ou como *Tikhonov Regularization*, por matemáticos. Assim, o parâmetro de regularização que controla a filtragem é determinado pela validação cruzada generalizada, minimizando o erro médio quadrático esperado para os valores inseridos (SCHNEIDER, 2001). No geral, o modelo RegEM consiste nos mesmos passos do modelo EM com uma exceção na estimativa dada por:

$$\hat{B} = \hat{\sum}_{aa}^{-1} \hat{\sum}_{am}, \tag{6}$$

onde a estimativa $\hat{B}$ dos coeficientes de regressão é obtida pela submatriz $\hat{\sum}_{aa}$, que consiste das variâncias e covariâncias estimadas das variáveis para as quais, na amostra analisada, os valores estão disponíveis, e pela submatriz $\hat{\sum}_{am}$, que consiste das covariâncias cruzadas estimadas das variáveis para as quais os valores estão disponíveis com as variáveis para as quais os valores estão ausentes na amostra analisada. Para esta estimativa, em cada iteração e para cada amostra com valor ausente a matriz inversa $\hat{\sum}_{aa}^{-1}$ utilizada é obtida por:

$$\hat{\sum}_{aa}^{-1} \leftarrow (\hat{\sum}_{aa} + h^2 \hat{D})^{-1}, \tag{7}$$

onde $\hat{D} = Diag(\hat{\sum}_{aa})$ é a matriz diagonal constituída pelos elementos diagonais da matriz de covariância $\hat{\sum}_{aa}$ e o escalar h é um parâmetro de regularização (SCHNEIDER, 2001).

2.3.3 *K-Nearest Neighbors Input* (KNNI)

Introduzido inicialmente por Troyanskaya et al. (2001) em seu trabalho envolvendo dados ausentes em *microarrays* do DNA, o método baseado no KNN seleciona amostras com perfis de expressão semelhantes ao da amostra analisada para imputar os valores ausentes. Considerando, por exemplo, que a amostra A tem um valor ausente no experimento

[2] https://github.com/tapios/RegEM

1, este método encontrará K outras amostras, com o valor disponível no experimento 1 e com expressão mais próxima à de A nos $2 - N$ experimentos, onde N é o número total de experimentos. Então, uma média ponderada de valores no experimento 1 dos K mais próximos é utilizada como estimativa para o valor ausente na amostra. Com a utilização da média ponderada, a contribuição de cada amostra é ponderada pela similaridade de sua expressão com a do gene A (TROYANSKAYA et al., 2001).

O autor afirma ter considerado várias métricas (correlação de *Pearson*, minimização de variância) para verificação da similaridade, tendo escolhido a distância Euclidiana como sendo suficientemente precisa para calcular a similaridade genética (TROYANSKAYA et al., 2001).

2.3.4 *Singular Value Decomposition Input* (SVDI)

Assim como o KNNI, descrito na Seção 2.3.3, o SVDI também foi introduzido por Troyanskaya et al. (2001) em seu trabalho envolvendo dados ausentes em *microarrays* do DNA. Este método utiliza a decomposição de valor singular (8) para obter um conjunto de padrões de expressão mutuamente ortogonais que podem ser linearmente combinados para aproximar a expressão de todas as amostras no conjunto de dados.

$$A_{m \times n} = U_{m \times m} \sum_{m \times n} V^T_{n \times n} \tag{8}$$

A matriz V^T contém diferentes componentes principais, cuja contribuição para a expressão no espaço é quantificada por autovalores correspondentes na diagonal da matriz. Em seguida, os componentes principais mais significativos são identificados classificando os componentes de acordo com seu autovalor correspondente. Segundo o autor, a quantidade de componentes principais utilizados é definida empiricamente (TROYANSKAYA et al., 2001).

Uma vez que os K componentes principais mais significativos de V^T são selecionados, é estimado o valor ausente j no componente i, primeiro realizando uma regressão deste componente contra os K componentes principais e, então, utilizando os coeficientes da regressão para reconstruir j por meio de uma combinação linear dos K componentes principais. O j-ésimo valor do componente i e os valores de j dos K componentes principais não são utilizados para determinar esses coeficientes de regressão (TROYANSKAYA et al., 2001).

O autor ressalta que o SVD, base para o SVDI, somente pode ser executado em matrizes completas. Assim, a média de amostras foi substituída por todos os valores ausentes na matriz A, obtendo-se $A^{'}$. Em seguida, foi utilizado um método de maximização de expectativa para chegar à estimativa final. Cada valor ausente em $A^{'}$ é estimado utilizando

o algoritmo acima e, em seguida, o procedimento é repetido na matriz recém-obtida, até que a alteração total na matriz esteja abaixo de um limite determinado empiricamente, como $0{,}01$ (TROYANSKAYA et al., 2001).

É importante ressaltar que não foi constatada a existência de uma relação entre o método de maximização de expectativa utilizado por Troyanskaya et al. (2001) e o algoritmo EM introduzido por Dempster, Laird e Rubin (1977), apresentado na Seção 2.3.1.

2.3.5 *Bayesian Principal Component Analysis* (BPCA)

O método *Bayesian Principal Component Analysis* (BPCA) foi proposto por Oba et al. (2003) e aplicado na estimação de valores ausentes em *microarrays* de DNA. Segundo o autor, o método consiste em três etapas: regressão do principal componente (PC); estimativa *Bayesiana* e; maximização de estimativa.

O *Principal Component Analysis* (PCA) representa a variação D-dimensional de vetores de expressão genética y como uma combinação linear do eixo principal $w_l (1 \leq l \leq K)$, cujo número é relativamente pequeno ($K < D$), definido por:

$$y = \sum_{l=1}^{K} x_l w_l + \epsilon, \tag{9}$$

onde $x_l (1 \leq l \leq L)$ são os coeficientes lineares denominados contadores fatoriais (*factor score*) enquanto ϵ é o erro residual. Utilizando um número K especificamente determinado, o PCA obtém x_l e w_l de forma que a soma do erro quadrado $\|\epsilon\|^2$ sobre todo o conjunto de dados Y seja minimizada. Uma matriz de covariância S para os vetores de expressão $y_i (l \leq i \leq N)$ é dada por:

$$S = \frac{1}{N} \sum_{i=1}^{N} (y_i - \mu)(y_i - \mu)^T, \tag{10}$$

onde μ é o vetor médio de y: $\mu \overset{def}{=\!=} (\frac{1}{N}) \sum_{i=1}^{N} y_i$. Define-se $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_D$ e $u_1, u_2, ..., u_D$ como os autovalores e autovetores de S, respectivamente, e o l-ésimo vetor do eixo principal por $w_l = \sqrt{\lambda_l} u_l$, assim, a pontuação do l-ésimo fator para um vetor de expressão y é dado por $x_l = (\frac{w_l}{\lambda_l})^T y$. Até aqui considerou-se que não há valores ausentes e, a partir daqui, esta existência é assumida. Na regressão do PC, a parte ausente y^{miss} no vetor de expressão y é estimada a partir da parte observada y^{obs} utilizando o resultado do PCA. Define-se w_l^{obs} e w_l^{miss} como partes de cada eixo principal w_l, correspondente às partes observadas e ausentes, respectivamente, em y. De forma similar, define-se $W = (W^{obs}, W^{miss})$, onde W^{obs} ou W^{miss} denota uma matriz cujos vetores das colunas são

w_l^{obs},..., w_k^{obs} ou w_l^{miss},..., w_k^{miss}, respectivamente. A pontuação dos fatores $x = (x_l,..., x_k)$ para o vetor de expressão y são obtidos pela minimização do erro residual dada por:

$$err = \left\| y^{obs} - W^{obs} x \right\|^2 . \tag{11}$$

Como se trata de um problema de regressão, sua solução utilizando os mínimos quadrados é dada por:

$$x = (W^{obsT} W^{obs})^{-1} W^{obsT} y^{obs}. \tag{12}$$

Utilizando x, a parte ausente é estimada por:

$$y^{miss} = W^{miss} x. \tag{13}$$

2.3.6 *Local Least Squares Imputation* (LLSimput)

O método *Local Least Squares Imputation* (LLSimput) baseado em mínimos quadrados foi introduzido por Kim, Golub e Park (2004) e faz com que genes com valor ausente sejam representados como combinação linear de genes semelhantes. O método utiliza duas etapas na imputação de mínimos quadrados locais. O primeiro passo é selecionar os k genes pela norma $L2$ ou pelos coeficientes de correlação de Pearson. O segundo passo é a regressão e estimativa, independentemente de como os k genes são selecionados.

2.3.7 *K-Mean* e *Fuzzy K-Mean Imputation* (KMI)

Em seu trabalho, Li et al. (2004) faz uma comparação entre a utilização do método *K-Mean* e o método *Fuzzy K-Mean* para a imputação de dados ausentes. O objetivo do agrupamento é, dado um conjunto de objetos, dividir o conjunto de dados em grupos com base na similaridade destes objetos e minimizar a dissimilaridade intra-grupo. Essa dissimilaridade intra-grupo é medida pela soma das distâncias entre os objetos e o centroide do grupo ao qual estão atribuídos. Um centroide de um grupo representa o valor médio dos objetos no grupo. O algoritmo para imputação de dados ausentes com o método de agrupamento *K-Mean* pode ser dividido em três etapas: primeiro, selecione aleatoriamente K objetos de dados completos como K centroides; segundo, modifique iterativamente a partição para reduzir a soma das distâncias de cada objeto do centroide do grupo ao qual o objeto pertence; por fim, o processo termina quando o somatório das distâncias for menor que um limite especificado pelo usuário. A última etapa é o preenchimento de todos os atributos de não referência para cada objeto incompleto com base nas informações do grupo. Os objetos de dados que pertencem ao mesmo grupo são usados como vizinhos

mais próximos, e aplicamos um algoritmo de vizinho mais próximo para substituir os dados ausentes.

Por outro lado, na abordagem *fuzzy* o agrupamento fornece uma ferramenta de descrição melhor quando os grupos não são bem separados, como é o caso da ausência dos valores. Além disso, o agrupamento *K-mean* original pode ficar preso em um mínimo local se os pontos iniciais não forem selecionados corretamente. No entanto, os valores contínuos de associação no agrupamento *fuzzy* tornam os algoritmos resultantes menos suscetíveis a ficarem presos em mínimos locais. O algoritmo para a ausência de dados com o método de agrupamento *Fuzzy K-Mean* também possui três etapas: no processo de inicialização, são escolhidos os *K* centroides, distribuídos igualmente para evitar mínimos locais; na segunda etapa, são atualizadas iterativamente as funções de pertinência e os centroides até que a distância total atenda ao limite de distância especificado pelo usuário. Por fim, são atribuídos parâmetros de não referência para cada objeto com valor ausente. Em outro trabalho relacionado a dados ausentes, Liao et al. (2009) propõem a utilização do algoritmo *Fuzzy K-Mean* juntamente com a teoria de janela deslizante (*Sliding Window*) para limitar a quantidade de objetos utilizados na execução do algoritmo e demonstrar sua eficiência em comparação com a utilização completa.

2.3.8 *Global Most Common Attribute* (GMC)

A abordagem do atributo global mais frequente foi utilizada por Grzymala-Busse et al. (2005) para lidar com os valores de atributos ausentes nos conjuntos de dados que descrevem os nascimentos prematuros que possuíam uma grande quantidade de valores ausentes, além de alguns dados inconsistentes. Para atributos simbólicos, todo valor ausente foi substituído pelo valor do atributo mais comum. Para atributos numéricos, todos os valores ausentes devem ser substituídos pela média de todos os valores do atributo correspondente. Contudo, nos conjuntos de dados de nascimento prematuro, havia dois conceitos que descreviam o nascimento prematuro e completo. Dessa forma, para atributos simbólicos, todo valor ausente deve ser substituído pelo valor do atributo mais comum que ocorre para o mesmo conceito. Neste caso, para atributos numéricos, todos os valores ausentes devem ser substituídos pela média de todos os valores atribuídos, restritos ao mesmo conceito.

2.3.9 *Fuzzy-Rough Nearest Neighbor Imputation* (FRNNI)

A teoria dos conjuntos *Fuzzy-Rough* fornece uma excelente estrutura para lidar com a incerteza. Como não são essencialmente problemas de otimização, os métodos *Fuzzy-Rough* não iteram através de etapas de algoritmo. Em outras palavras, estes métodos não precisam de um critério de parada e não dependem de parâmetros especificados pelo usuário. Outro ponto positivo para utilização de técnicas *Fuzzy-Rough* é sua simplicidade e

compreensibilidade. As decisões são tomadas com base no cálculo das semelhanças *fuzzy* entre as instâncias. Além disso, os métodos trabalham com facilidade e eficácia na presença de ruído, podendo também lidar com dados impreciso e valores ausentes, inclusive, não necessitando de estimativas iniciais neste último caso (AMIRI; JENSEN, 2016).

Na teoria dos conjuntos de dados *Rough*, o conjunto de dados é representado como um sistema $(X, \mathcal{A})$ em que $X = x_1, x_2, ..., x_n$ é um conjunto finito de instâncias de dados e $\mathcal{A} = a_1, a_2, ..., a_m$ é um conjunto finito e não vazio de atributos. Para $B \subseteq C$ há uma relação *B-indiscernibility* dada por:

$$R_B = (x,y) \in X^2 | \forall a \in B \Rightarrow a(x) = a(y), \tag{14}$$

onde R_B é uma relação de equivalência e $[x]_{R_B}$ divide o conjunto X em classes equivalentes definidas pelos atributos presentes em B. Um sistema de decisão é representado por $(X, \mathcal{A} \cup d)$. Segundo o autor, este é um tipo especial de sistema de informações que possui um atributo adicional, chamado de atributo de decisão, que contém uma classe para cada instância de dados. Se $A \subseteq X$, então as aproximações superiores e inferiores do conjunto A são dadas por (AMIRI; JENSEN, 2016):

$$R_B \downarrow A = x \in X | [x]_{R_B} \subseteq A \tag{15}$$

$$R_B \uparrow A = x \in X | [x]_{R_B} \cap A \neq \emptyset. \tag{16}$$

Todas as instâncias que pertencem a $R_B \downarrow$ definitivamente pertencem ao conjunto A, enquanto as instâncias que pertencem a $R_B \uparrow$ provavelmente são membro de A (AMIRI; JENSEN, 2016).

Sendo R uma relação de tolerância *fuzzy*, $(X, \cup \{d\})$ um sistema de decisão e A um subconjunto de X, então aproximações *Fuzzy-Rough* superiores e inferiores de A são definidas por:

$$(R_B \downarrow A)(y) = inf_{x \in N} I(R_B(x,y), A(x)) \tag{17}$$

$$(R_B \uparrow A)(y) = sup_{x \in N} T(R_B(x,y), A(x)), \tag{18}$$

onde I é um implicador, T é uma T-norma e B é um subconjunto de atributos do conjunto $\mathcal{A}$ (AMIRI; JENSEN, 2016). Explicando de outra forma, nas aproximações superiores e inferiores, se um objeto estiver relacionado a todos dos elementos de um conjunto, ele pertence a aproximação inferior e, se um objeto estiver relacionado a pelo menos um dos

elementos de um conjunto, ele pertence a aproximação superior desse conjunto (AMIRI; JENSEN, 2016).

A partir desta teoria, Jensen e Cornelis (2011 apud AMIRI; JENSEN, 2016) apresentam o *Fuzzy-Rough Nearest Neighbor* (FRNN), um algoritmo para previsões baseado no algoritmo KNN que utilizam *Fuzzy-Rough* com aproximações superiores e inferiores. Neste algoritmo as aproximações superiores e inferiores são definidas como:

$$(R \downarrow R_d z)(y) = inf_{t \in N} I(R(y,t), R_d(t,z)) \tag{19}$$

$$(R \uparrow R_d z)(y) = sup_{t \in N} T(R(y,t), R_d(t,z)), \tag{20}$$

onde R_{d^z} é a relação de tolerância *fuzzy* que calcula a semelhança entre duas instâncias para a característica de decisão. $R_d z$, relativo ao conjunto *fuzzy*, significa a similaridade entre as instâncias z e t com a respectiva característica de decisão d e é calculada por $R_d(t,z)$, que é uma relação de tolerância *fuzzy*. Em geral, $R_a z(t)$ significa a similaridade entre as instâncias z e t em relação ao atributo a (AMIRI; JENSEN, 2016).

Com base nesta teoria, foi proposto por Amiri e Jensen (2016) uma extensão do FRNN denominado *Fuzzy-Rough Nearest Neighbor Imputation* (FRNNI). Para cada instância do conjunto de dados, y, contendo pelo menos um valor ausente para o atributo a, o algoritmo encontra os k vizinhos mais próximos e os coloca em um conjunto N. Então o algoritmo aproxima o valor ausente através dos vizinhos mais próximos de y. Após isso, são calculados as aproximações superiores e inferiores da instância z, utilizando a média deles para produzir a associação final M. O algoritmo realiza este processo para todas as instâncias que pertencem a N, retornando um valor baseado nestes cálculos, ou seja, o valor retornado é baseado no valor de todos os vizinhos. Amiri e Jensen (2016) também propuseram outros dois modelos que são variações do FRNNI. Um deles foi denominado *Vaguely Quantified Nearest Neighbor Imputation* (VQNNI) e o outro *Ordered Weighted Average Fuzzy-Rough Nearest Neighbors* (OWANNI).

2.3.10 Evolving Fuzzy Granular Predictor (eFGP)

Em Garcia, Leite e Skrjanc (2019) e Garcia et al. (2019) foi introduzido um preditor, chamado eFGP (*evolving Fuzzy Granular Predictor*), que utiliza uma abordagem híbrida baseada em deleção e imputação múltipla para tratar dados ausentes em ambientes dinâmicos e não estacionários nos quais os dados estão sujeitos a variações e mudanças de contexto. Trata-se de um modelo evolutivo baseado em regras *fuzzy* granulares modificado para aproximação de funções e previsão de séries temporais em um contexto *online* com a possibilidade de existirem dados ausentes. Basicamente, o eFGP atua de três maneiras: (i) quando não há variáveis ausentes, o algoritmo utiliza todas as variáveis para fazer a

previsão; (ii) se a amostra possui uma variável ausente, a previsão é realizada com termos reduzidos, ou seja, sem considerar a variável ausente; (iii) quando a amostra possui mais de uma variável ausente, o eFGP realiza a imputação múltipla considerando os pontos centrais de grânulos de informações fuzzy trapezoidais para as variáveis ausentes e realiza a previsão com a amostra completa.

A previsão do eFGP é realizada de acordo com a quantidade de variáveis ausentes na amostra. Quando a amostra não possui variáveis ausentes, o algoritmo utiliza funções completas p^i e o cálculo da saída $\hat{y}$ é obtido por (21):

$$\hat{y} = \frac{\sum_{i=1}^{c} \Psi_{com}^{i} p^i(x_1, ..., x_n)}{\sum_{i=1}^{c} \Psi_{com}^{i}}, \tag{21}$$

em que, Ψ_{com}^{i} é computado por (22) e $p^i(x)$ por (23).

$$\Psi_{com}^{i} = T(A_1^i(x_1), ..., A_n^i(x_n)), \tag{22}$$

$$p^i(x) = \alpha_0^i + \sum_{j=1}^{n} \alpha_j^i x_j, \tag{23}$$

onde p^i é uma função afim e $A_j^i = (\underline{\underline{a}}_j^i, \underline{a}_j^i, \overline{\overline{a}}_j^i, \overline{a}_j^i)$ são funções de pertinência trapezoidais relacionadas à i-ésima regra. Quando existe uma única variável ausente, as funções de termos reduzidos q_θ^i são empregadas e a saída $\hat{y}$ é calculada por (24):

$$\hat{y} = \frac{\sum_{i=1}^{c} \Psi_{inc}^{i} q_\theta^i(x_1, ..., x_{\theta-1}, x_{\theta+1}, ..., x_n)}{\sum_{i=1}^{c} \Psi_{inc}^{i}}, \tag{24}$$

em que, em que, Ψ_{com}^{i} é dado por (25) e $q_\theta^i(x)$ por (26).

$$\Psi_{inc}^{i} = T(A_1^i(x_1), ..., A_{\theta-1}^i(x_{\theta-1}), A_{\theta+1}^i(x_{\theta+1}), ..., A_n^i(x_n)), \tag{25}$$

$$q_\theta^i(x) = \beta_{0\theta}^i + \sum_{j=1, j\neq\theta}^{n} \beta_{j\theta}^i x_j \tag{26}$$

onde q_k^i é uma função afim e $B^i = (\underline{\underline{b}}_j^i, \underline{b}_j^i, \overline{\overline{b}}_j^i, \overline{b}_j^i)$ são funções de pertinência trapezoidais relacionadas à i-ésima regra. Por fim, quando existe mais de uma variável ausente, é escolhida a regra mais ativa para $x^{[h]}$ utilizando (27):

$$\Psi_{inc}^{i} = T(A_1^i(x_1), ..., A_{\theta_1-1}^i(x_{\theta_1-1}), A_{\theta_1+1}^i(x_{\theta_1+1}), ..., A_{\theta_2-1}^i(x_{\theta_2-1}), A_{\theta_2+1}^i(x_{\theta_2+1}), ..., A_n^i(x_n)). \tag{27}$$

Em seguida é feita uma imputação múltipla utilizando os pontos médios do granulo mais ativo para $x^{[h]}$, conforme (28) e (29).

$$x_{\theta_1}^{[h]} = mp(A_{\theta_1}^i) = \frac{(\underline{a}_{\theta_1}^i + \overline{a}_{\theta_1}^i)}{2}. \tag{28}$$

$$x_{\theta_2}^{[h]} = mp(A_{\theta_2}^i) = \frac{(\underline{a}_{\theta_2}^i + \overline{a}_{\theta_2}^i)}{2}. \tag{29}$$

Para finalizar, o algoritmo, então, calcula a previsão $\hat{y}$ utilizando funções completas p^i através de (21), (22) e (23).

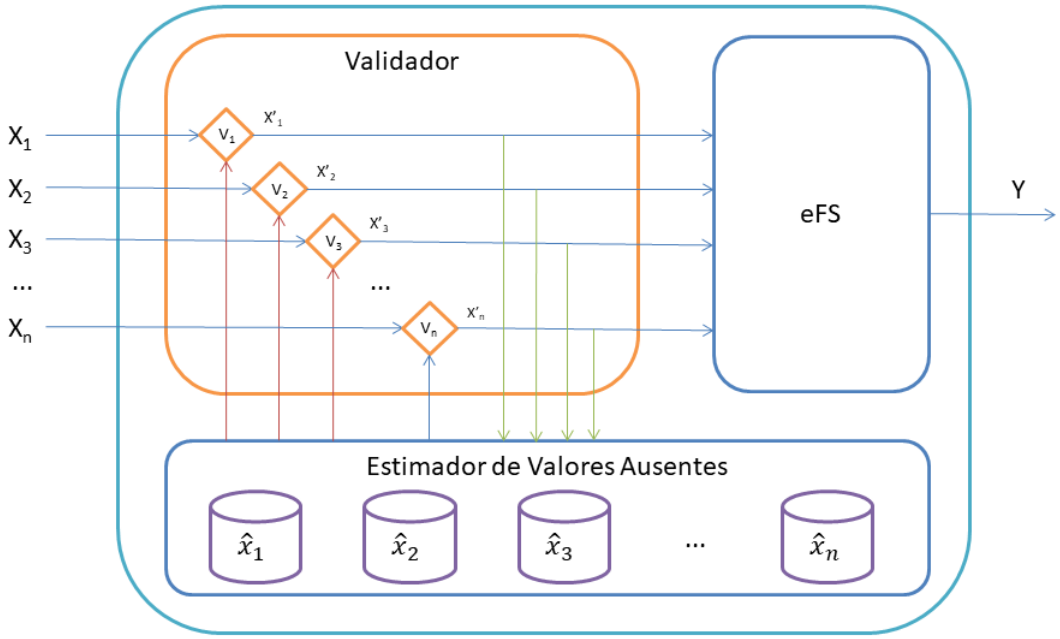
3 Modelagem Evolutiva para Estimação de Variáveis Ausentes

Este capítulo introduz a abordagem proposta para imputação de valores em variáveis ausentes e demonstra sua aplicação em dois modelos evolutivos. Inicialmente, na Seção 3.1, a abordagem proposta, chamada de MDP (*Missing Data Procedure*) é detalhada. Na sequência, a Seção 3.2, apresenta a abordagem proposta implementada no eNFN (*evolving Neo-Fuzzy Neuron*), denominada eNFN-MDP (*evolving Neo-Fuzzy Neuron with Missing Data Procedure*). Por fim, a Seção 3.3, ilustra a implementação do MDP no modelo evolutivo ALMMo (*Autonomous Learning Multi-Model*), denominada ALMMo-MDP (*Autonomous Learning Multi-Model with Missing Data Procedure*).

3.1 Missing Data Procedure (MDP)

O *Missing Data Procedure* (MDP) possibilita, a cada amostra do conjunto de dados, que os modelos evolutivos estimem os valores das variáveis ausentes, realizem a imputação desses valores quando necessário e realizem a previsão da saída. O MDP é ilustrado na Figura 1 e pode ser aplicado tanto para modelos com múltiplas entradas e múltiplas saídas quanto para modelos com múltiplas entradas e única saída. A abordagem proposta pode ser sumarizada em 3 etapas como se segue:

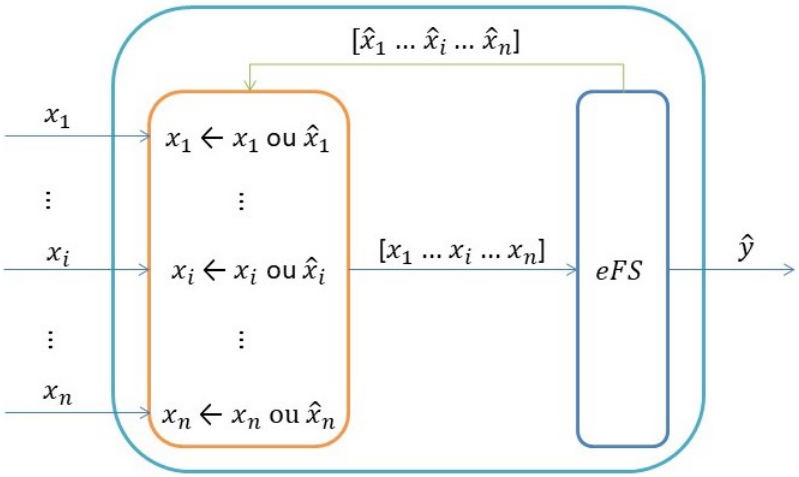
Figura 1 – Estrutura Genérica do *Missing Data Procedure*



- **Validação**: são instanciadas n verificações, onde n é o número de variáveis de entrada, que recebem a variável de entrada e, caso ela esteja ausente, imputam o valor previsto pelo Estimador. Dessa forma, independente se a variável possui ou não valor ausente, ela sempre será encaminhada com um valor válido para a etapa de previsão.
- **Estimativa**: o Estimador utiliza os valores de x_t como entrada, em que x é uma amostra do conjunto de dados com n variáveis e t é o passo atual, para estimar os valores de x_{t+1}, ou seja, estimam os valores de x para a próxima amostra. Os valores estimados para x_{t+1} são armazenados em $\hat{x}$.
- **Previsão**: nesta etapa o *evolving Fuzzy System* (eFS) calcula a saída utilizando todas as variáveis da amostra de dados.

Para os sistemas *fuzzy* evolutivos (eFS) com múltiplas entradas e saídas, o MDP realiza as etapas de estimativa e previsão simultaneamente, conforme pode ser verificado na Figura 2.

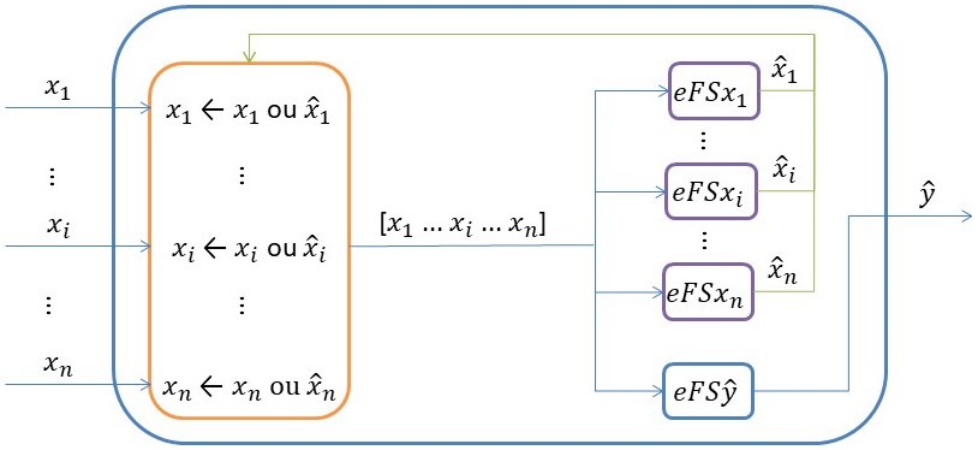
Figura 2 – Abordagem MDP para modelos com múltiplas entradas e saídas.



Para os eFS com múltiplas entradas e uma saída, o MDP cria uma instância do eFS para cada variável de entrada, para realizar a estimativa das variáveis ausentes, e uma instância para calcular a saída desejada, cada instância com múltiplas entradas e uma saída. Dessa forma, para o processamento de um conjunto de dados com n variáveis de entrada e uma saída devem ser instanciados $n + 1$ modelos, n responsáveis pela estimativa das variáveis de entrada e 1 para o cálculo da saída. Cada eFS recebe todas as variáveis de entrada e estima o valor da sua respectiva variável de entrada ou da saída. A estrutura do MDP proposta para modelos com múltiplas entradas e uma saída é sumarizada na Figura 3.

A próxima seção ilustra a aplicação do MDP em um eFS com múltiplas entradas e múltiplas saídas, denominado eNFN-MDP, e a Seção 3.3 demonstra o MDP implementado em um eFS com múltiplas entradas e uma saída, denominado ALMMo-MDP.

Figura 3 – Abordagem MDP para modelos com múltiplas entradas e única saída.



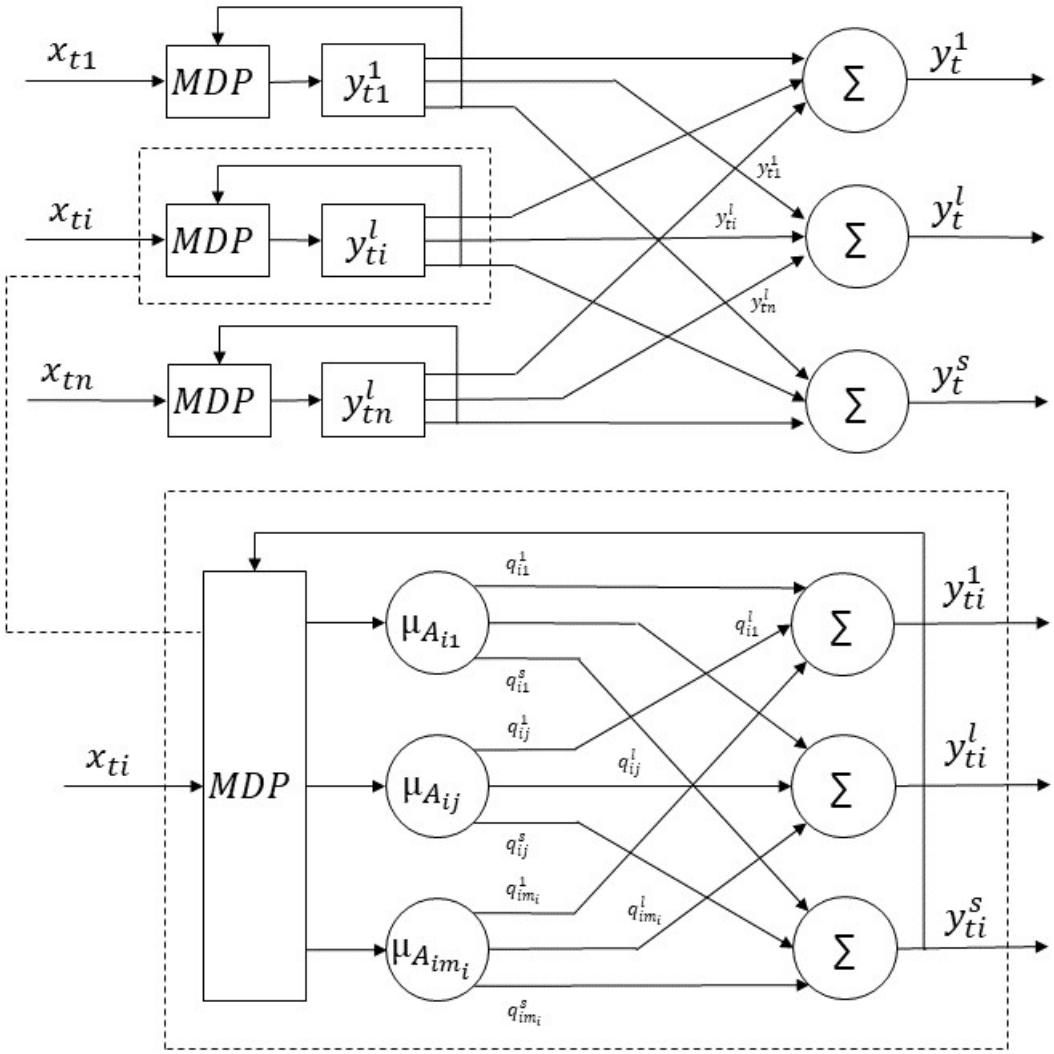
3.2 eNFN *with Missing Data Procedure* (eNFN-MDP)

O eNFN-MDP (*evolving Neo-Fuzzy Neuron with Missing Data Procedure*) é construído sob a estrutura da rede *neuro-fuzzy* evolutiva eNFN (*evolving Neo-Fuzzy Neuron*) proposta em Silva et al. (2014) e generalizada para um modelo com múltiplas entradas e múltiplas saídas (MIMO) em Silva (2014). O eNFN-MDP é um previsor que, simultaneamente, realiza a previsão e estima os valores das variáveis com dados ausentes, ou seja, a cada nova amostra ele realiza a previsão da saída e estima o valor de cada uma das variáveis de entrada para o próximo passo. Assim, caso exista uma ou mais variáveis ausentes em uma amostra, o eNFN-MDP imputa para as variáveis ausentes o valor estimado no passo anterior e realizada a previsão com a amostra completa.

A rede é composta por um conjunto de modelos de Takagi-Sugeno de Ordem $Zero$ com funções de pertinência triangulares e complementares. Ela inicia o processo de aprendizado do zero e sua estrutura flexível evolui, baseado no erro de modelagem, com a inclusão de funções de pertinência à medida que novas informações são fornecidas pelo fluxo de dados de entrada e com a exclusão de funções de pertinência quando informações recebidas anteriormente não representam o conhecimento atual. Os parâmetros da rede são atualizados por um algoritmo baseado no Gradiente com taxa de aprendizagem ótima. A Figura 4 ilustra a estrutura do eNFN-MDP e suas etapas são sumarizadas a seguir:

1. **Inicialização das Funções de Pertinência:** o domínio de cada variável de entrada é, inicialmente, granularizado por duas funções de pertinência triangulares e complementares. Esta etapa é realizada somente uma vez, utilizando os limites inferior (min_{x_i}) e superior (max_{x_i}) de cada variável de entrada.
2. ***Missing Data Procedure*:** verifica se a amostra de dados atual (x_t) possui alguma variável com valor ausente, caso positivo, imputa o valor estimado ($\hat{x}_i$). A Seção 3.3.1 detalhada este procedimento.
3. **Adaptação de Contexto:** visa ajustar os valores de min_{x_i} e max_{x_i} caso os valores

Figura 4 – Estrutura básica do eNFN-MDP.



da amostra atual x_{ti} excedam esses limiares. Portanto, $se\ x_{ti} < min_{x_i}\ então\ min_{x_i} = x_{ti}\ e\ b_{i1} = min_{x_i}$ ou $se\ x_{ti} > max_{x_i}\ então\ max_{x_i} = x_{ti}\ e\ b_{im_i} = max_{x_i}$, em que b_{i1} é o valor modal da primeira função de pertinência da variável i e b_{im_i} é o valor modal da última função de pertinência da variável i.

4. **Ativação das Funções de Pertinência:** para cada variável de entrada é calculado o grau de ativação das funções de pertinência $\mu_{A_{ij}}(x_{ti})$. Este procedimento é detalhado na Seção 3.2.2.
5. **Estimação das Variáveis e Cálculo das Saídas:** o eNFN-MDP possui s saídas, sendo $s = n + o$, em que n é o número de variáveis de entrada e o é a número de saídas. O eNFN-MDP realiza a previsão das o saídas e estima o valor das n variáveis de entrada concomitantemente. A Seção 3.2.2 ilustra o procedimento completo para estes cálculos.
6. **Atualização dos Parâmetros:** o aprendizado dos parâmetros do eNFN-MDP é realizado por um método baseado no Gradiente com taxa de aprendizado α ótima. Os procedimentos para cálculo da taxa de aprendizado α e ajuste dos parâmetros q_{ij}^l são

apresentados na Seção 3.2.3.

7. **Atualização do Valor Modal:** para cada variável de entrada encontra-se o índice da função de pertinência mais ativa (b_i^*) para a amostra x_{ti}. Se $b_i^* \neq 1$ e $b_i^* \neq m_i$, a b_i^*-ésima função de pertinência terá seu valor modal atualizado por

$$b_{b_i^*}^{novo} = b_{b_i^*}^{atual} + \beta(x_{ti} - b_{b_i^*}^{atual}), \tag{30}$$

em que β é a taxa de atualização.

8. **Criação de Funções de Pertinência:** este procedimento visa granularizar o domínio das variáveis de entrada para reduzir o erro de saída de maneira uniforme. No eNFN-MDP o refinamento é realizado com base no erro da função de pertinência mais ativa e é descrito na Seção 3.2.4.
9. **Exclusão de Funções de Pertinência:** o eNFN-MDP exclui funções de pertinência utilizando um conceito de idade da regra introduzido por (ANGELOV; FILEV, 2005). O conceito de idade da regra é empregado para determinar o tempo de inatividade de uma função de pertinência, isto é, o número de amostras que uma função de pertinência fica sem ser ativada (SILVA et al., 2014). A Seção 3.2.5 ilustra esse processo.

3.2.1 *Missing Data Procedure* - MIMO

Para cada a variável x_{ti} da amostra x_t, o MDP realiza a etapa de validação que verifica a ausência do valor da variável. Caso seja encontrado uma ou mais variáveis ausentes, o MDP imputa o valor estimado $\hat{x}_i$ para a variável com valor ausente. Este processo se repete para cada amostra recebida no fluxo de dados e o cálculo da saída sempre é realizado com todas as variáveis. Excepcionalmente, caso a ausência ocorra na primeira amostra, em que ainda não há estimativa disponível, o valor ausente é substituído por zero. O Algoritmo 1 sumariza a etapa de validação do *Missing Data Procedure* MIMO.

Algoritmo 1: Procedimento de Validação do *Missing Data Procedure* MIMO.

```
for i = 1 : n do
    if x_ti = missing_value then
        if t ≠ 1 then
            x_ti = x̂_i
        end
        else
            x_ti = 0;
        end
    end
end
```

3.2.2 Estimação das Variáveis e Cálculo da Saídas

O eNFN-MDP, simultaneamente, realizada a previsão das saídas e estima as variáveis ausentes. O domínio das variáveis de entrada x_{ti} é granularizado por m_i funções de

pertinência triangulares e complementares. A saída de cada um dos modelos de Takagi-Sugeno de Ordem $Zero$, um para cada variável de entrada i e para cada saída l é definida pelo conjunto dessas m_i^l regras:

$$\begin{aligned} R_i^{l1}\ Se\ x_{ti}\ \ A_{i1}\ Então\ y_{ti}^l &= \mu_{A_{i1}} q_{i1}^l \\ R_i^{lj}\ Se\ x_{ti}\ \ A_{ij}\ Então\ y_{ti}^l &= \mu A_{ij} q_{ij}^l \\ R_i^{lm_i}\ Se\ x_{ti}\ \ A_{im_i}\ Então\ y_{ti}^l &= \mu A_{im_i} q_{im_i}^l. \end{aligned} \tag{31}$$

A saída individual y_{ti}^l é obtida por:

$$y_{ti}^l = \frac{\sum_{j=1}^{m_i} \mu_{A_{ij}}(x_{ti}) q_{ij}^l}{\sum_{j=1}^{m_i} \mu_{A_{ij}}(x_{ti})}, \tag{32}$$

na qual $\mu_{A_{ij}}(x_{ti})$ é o grau de ativação da função de pertinência A_{ij} pela entrada x_{ti} e q_{ij}^l é o parâmetro da rede. Dado que as funções de pertinência são triangulares e complementares, independentemente do valor de x_{ti} apenas duas funções de pertinência terão seu grau de ativação maior que zero e soma do grau de ativação das funções ativas será sempre um, isto é:

$$\sum_{j=1}^{m_i} \mu_{A_{ij}}(x_{ti}) = \mu_{A_{ik_i}}(x_{ti}) + \mu_{A_{ik_i+1}}(x_{ti}) = 1. \tag{33}$$

Portanto, (32) pode ser reescrita como:

$$y_{ti}^l = \mu_{A_{ik_i}}(x_{ti}) q_{ik_i}^l + \mu_{A_{ik_i+1}}(x_{ti}) q_{ik_i+1}^l, \tag{34}$$

em que k_i e $k_i + 1$ são os índices das funções de pertinência ativadas para a variável i. Cada uma das l saídas no instante t (y_t^l) é obtida pela soma das saídas individuais y_{ti}^l conforme (35).

$$y_t^l = \sum_{i=1}^{n} y_{ti}^l. \tag{35}$$

3.2.3 Atualização dos Parâmetros

Os parâmetros q_{ij}^l são atualizados por um algoritmo baseado no Gradiente com taxa de aprendizagem ótima (CAMINHAS; GOMIDE, 2000). Durante a adaptação dos parâmetros, somente as funções de pertinência ativas, A_{ik_i} e A_{ik_i+1}, são utilizadas e, dessa

forma, somente os parâmetros destas funções são atualizados a cada passo. Portanto, a atualização dos pesos $q^l_{ik_i}$ e $q^l_{ik_i+1}$ pode ser definida por:

$$\begin{aligned} q^l_{ik_i} &= q^l_{ik_i} - \alpha(y_t - \hat{y}_t)\mu_{A_{ik_i}}(x_{ti}), \\ q^l_{ik_i+1} &= q^l_{ik_i+1} - \alpha(y_t - \hat{y}_t)\mu_{A_{ik_i+1}}(x_{ti}), \end{aligned} \tag{36}$$

e a taxa de aprendizagem α é obtida por (CAMINHAS; GOMIDE, 2000):

$$\alpha = \frac{1}{\sum_{i=1}^{n} \mu_{A_{ik_i}}(x_{ti})^2 + \mu_{A_{ik_i+1}}(x_{ti})^2}. \tag{37}$$

3.2.4 Criação de Funções de Pertinências

A criação de funções de pertinência refina a granularização do domínio das entradas visando reduzir o erro de saída de maneira uniforme (SILVA et al., 2015). O refinamento é realizado com base no erro e na variância do erro global e no erro local da função de pertinência mais ativa. O erro global ($\mu\hat{\mu}_{g_t}$) é calculado recursivamente para todas as saídas por:

$$\mu\hat{\mu}_{g_t} = \frac{\sum_{l=1}^{s} \hat{\mu}_{g^l_t}}{s} = \frac{\sum_{l=1}^{s} \hat{\mu}_{g^l_{t-1}} - \beta(\hat{\mu}_{g^l_{t-1}} - e_t)}{s}, \tag{38}$$

onde β é a taxa de aprendizado, s é o número de saídas, $\hat{\mu}^l_{gt}$ é o erro médio de cada uma das l saídas e $e_t = y^l_t - \hat{y}^l_t$. A variância média do erro global ($\sigma\hat{\sigma}^2_{g_t}$) é obtida recursivamente por:

$$\sigma\hat{\sigma}^2_{g_t} = \frac{\sum_{l=1}^{s} \hat{\sigma}^2_{g^l_t}}{s} = \frac{\sum_{l=1}^{s}(1-\beta)(\hat{\sigma}^2_{g^l_{t-1}} + \beta(\hat{\mu}_{g^l_t} - e_t)^2)}{s}, \tag{39}$$

em que $\hat{\sigma}^2_{g^l_t}$ a variância do erro de cada uma das l saídas. De forma similar, o valor médio $\hat{\mu}b^*_{ti}$ do erro local correspondente à função de pertinência mais ativa (b^*_i) é calculado recursivamente para cada entrada x_{ti} por:

$$\hat{\mu}_{b^*_{ti}} = \hat{\mu}_{b^*_{ti-1}} - \beta(\hat{\mu}_{b^*_{ti-1}} - e_t). \tag{40}$$

Uma nova função de pertinência é criada e inserida sempre que $\hat{\mu}_{b^*_{ti}} > \mu\hat{\mu}_{gt} + \sigma\hat{\sigma}^2_{gt}$ e $dist > \tau$, em que $dist$ é a distância entre o valor modal da função a ser criada e suas funções adjacentes. O limiar τ evita modelos complexos e *overfitting* e é calculado iterativamente por:

$$\tau = \frac{max_{x_i} - min_{x_i}}{\eta}, \tag{41}$$

onde η é um hiperparâmetro definido empiricamente.

A criação e inserção de funções de pertinência atualiza a granularização da i-ésima variável do domínio de entrada da seguinte forma:

- Se $b_i^* > 1$ e $b_i^* \neq m_i$, $dist$ é obtida por (42) e a função mais ativa é substituída por duas funções de pertinência com os valores modais calculados conforme (43) e (44).

$$dist = \frac{(b_{i,b_i^*+1} - b_{i,b_i^*-1})}{3}. \tag{42}$$

$$b_{i,novo_1} = b_{i,b_i^*-1} + dist. \tag{43}$$

$$b_{i,novo_2} = b_{i,b_i^*-1} + 2 * dist. \tag{44}$$

- Se a função mais ativa for a primeira ($b_i^* = 1$), $dist$ é calculada por (45) e uma nova função de pertinência é criada e inserida entre a primeira e a segunda com seu valor modal calculado de acordo com (46).

$$dist = \frac{(b_{i,b_i^*+1} - b_{i,b_i^*})}{2}. \tag{45}$$

$$b_{i,novo} = b_{i,b_i^*} + dist. \tag{46}$$

- Se a função mais ativa for a última ($b_i^* = m_i$), $dist$ é computada por (47) e uma nova função de pertinência é criada entre a penúltima e a última com seu valor modal obtido conforme (48).

$$dist = \frac{(b_{i,b_i^*} - b_{i,b_i^*-1})}{2}. \tag{47}$$

$$b_{i,novo} = b_{i,b_i^*} - dist. \tag{48}$$

As funções criadas recebem o valor do erro médio estimado para a função mais ativa e o número de funções para aquela variável é atualizado, assim como os limites inferior e/ou superior das funções de pertinência adjacentes são atualizados, porém sem modificação de seus valores modais. A alteração dos limites é necessária para garantir que as funções de pertinência se mantenham complementares (SILVA et al., 2015). O Algoritmo 2 detalha o procedimento para criação de funções de pertinência.

3.2.5 Exclusão de Funções de Pertinência

Para exclusão das funções de pertinência, é utilizado o conceito de idade da regra que determina o tempo de inatividade de uma função de pertinência, ou seja, o intervalo de

Algoritmo 2: Procedimento para Criação de Funções de Pertinência

```
Calcular μ̂μ_{g_t^l} - Eq. (38);
Calcular σ̂σ²_{g_t^l} - Eq. (39);
for i = 1 : n do
    Encontrar b_i^*;
    Calcular μ̂_{b_{t_i}^*} - Eq. (40);
    Calcular τ - Eq. (41);
    if (μ̂_{b_{t_i}^*} > μ̂μ_{g_t} + σ̂σ²_{g_t}) e (dist > τ) then
        if (b_i^* > 1) e (b_i^* ≠ m_i) then
            Calcular dist - Eq. (42);
            Remover b_i^*;
            Criar Nova Função de Pertinência - Eq. (43);
            Criar Nova Função de Pertinência - Eq. (44);
            Atualizar os parâmetros;
            m_i = m_i + 1;
        end
        if (b_i^* = 1) then
            Calcular dist - Eq. (45);
            Criar Nova Função de Pertinência - Eq. (46)
            Atualizar os parâmetros;
            m_i = m_i + 1;
        end
        if (b_i^* = m_i) then
            Calcular dist - Eq. (47);
            Criar Nova Função de Pertinência - Eq. (48);
            Atualizar os parâmetros;
            m_i = m_i + 1;
        end
    end
end
```

tempo que uma função de pertinência fica sem ser ativada (SILVA et al., 2015). Esta idade é dada por:

$$idade_j = t - ativa_j, \tag{49}$$

onde j é o índice da função de pertinência, t é a instância de tempo atual e $ativa$ é a instância de tempo que a j-ésima função de pertinência foi ativada. Para cada variável de entrada i, encontra-se o índice da função de pertinência com maior tempo de inatividade (b_i^-), que será excluída se:

$$idade_{b_i^-} > w \; e \; m_i > 2, \tag{50}$$

onde $idade_{b_i^-}$ é a instância de tempo em que a função indexada por b_i^- está inativa e w indica o limiar de tempo de inatividade das funções de pertinência. Após a exclusão de uma função de pertinência, é necessário atualizar a granularização do domínio da variável de entrada, que ocorre de acordo com as seguintes condições:

- Se $b_i^- > 1$ e $b_i^- \neq m_i$, a função indexada por b_i^- é excluída e as funções adjacentes terão seus limites superior e inferior modificados.
- Se $b_i^- = 1$, a função indexada por b_i^- é excluída e o valor modal da função adjacente é definido pelo valor de min_{x_i}.

- Se $b_i^- = m_i$, a função indexada por b_i^- é excluída e o valor modal da função adjacente é definido pelo valor de max_{x_i}.

O Algoritmo 3 ilustra o procedimento para exclusão de funções de pertinência.

Algoritmo 3: Procedimento para Exclusão de Funções de Pertinência

```
for i = 1 : n do
    Atualizar idade_i;
    Encontrar b_i^-;
    if (idade_{b_i^-} > ω) e (m_i > 2) then
        Excluir a função de pertinência indexada por b_i^-;
        Atualizar os parâmetros;
    end
end
```

3.2.6 Hiperparâmetros e Algoritmo do eNFN-MDP

O eNFN-MDP, assim como o eNFN, possui 3 hiperparâmetros definidos pelo usuário:

- β: a taxa de aprendizado empregada para atualizar os valores modais das funções de pertinência, para calcular o valor médio e a variância do erro, e para computar o erro médio da função mais ativa;
- η: parâmetro que auxilia no cálculo da menor distância permitida entre o valor modal da função de pertinência a ser criada e os valores modais de suas funções adjacentes;
- ω: limiar para exclusão de função de pertinência e está associado a capacidade do eNFN-MDP de esquecer o conhecimento antigo sempre que esse torna-se inútil.

Informações detalhadas sobre os hiperparâmetros, incluindo sugestões de valores, podem ser obtidas em Silva et al. (2014). O Algoritmo 4 ilustra o procedimento para lidar com dados ausentes, o processo de aprendizagem e o cálculo das saídas do eNFN-MDP generalizado para problemas com múltiplas entradas e múltiplas saídas.

3.3 ALMMo *with Missing Data Procedure* (ALMMo-MDP)

O MDP para modelos com múltiplas entradas e uma saída (MISO) foi implementado no ALMMo (*Autonomous Learning Multi-Model*) (ANGELOV; GU; PRÍNCIPE, 2017) e denominado ALMMo-MDP (*Autonomous Learning Multi-Model with Missing Data Procedure*). O ALMMo é um modelo *fuzzy* evolutivo que possui uma estrutura composta por regras *fuzzy* do tipo AnYa. Estas regras são inseridas em uma lista que também possui os pontos focais das nuvens de dados nas quais as regras são baseadas. As nuvens de dados, por sua vez, são os conjuntos de amostras de dados com propriedades comuns agrupadas em torno de seus pontos focais (ANGELOV, 2012; ANGELOV; YAGER, 2012).

No ALMMo, a identificação de pontos focais das nuvens de dados ocorre para estabelecer parâmetros locais de produção. Consequentemente, as amostras de dados são

Algoritmo 4: *eNFN with Missing Data Procedure*.

Entrada: x_t, y_t, β, η, w;
Saída: $\hat{y}_t$;
Inicializar funções de pertinência;
for $t = 1, 2, ...$ **do**
  Ler x_t, y_t;
  Missing Data Procedure MIMO - Alg. 1;
  Adaptação de contexto;
  for $i = 1 : n$ **do**
    Calcular $\mu_{A_{ik_i}}(x_{ti})$, $\mu_{A_{ik_i+1}}(x_{ti})$ - Eq. (33);
  end
  for $l = 1 : s$ **do**
    for $i = 1 : n$ **do**
      Calcular y_{ti}^l - Eq. (35);
    end
    Calcular $\hat{Y}_t^l$ - Eq. (35);
    Calcular α - Eq. (37);
    Atualizar $q_{ik_i}^l$, $q_{ik_i+1}^l$ - Eq. (36);
  end
  //Atualiza variáveis estimadas
  for $i = 1 : n$ **do**
    $\hat{x}_i = \hat{Y}_t^i$
  end
  //Saída prevista
  $i = 1$;
  for $l = n + 1 : s$ **do**
    $\hat{y}_{ti} = \hat{Y}_t^l$;
    $i = i + 1$;
  end
  Atualizar Valor Modal - Eq. (30);
  Procedimento - Criação de funções de pertinência - Alg. (2);
  Procedimento - Exclusão de funções de pertinência - Alg. (3);
end

atribuídas aos pontos focais mais próximos para formular a estrutura das nuvens de dados e a atribuição dos dados é como se segue:

$$SE(j^* = \underset{i=1,2,...,N}{argmin}(||x - \mu_i||))ENTÃO(\Xi_{j*} \leftarrow x), \tag{51}$$

em que, x = $[x_1, x_2, ..., x_M]^T$ são amostras de dados no espaço R^M e os índices referenciam o tempo em que as amostras foram observadas; $argmin$ é o método que busca a nuvem de dados mais próxima desse ponto; Ξ_i corresponde a i-ésima nuvem de dados e μ_i representa o centro daquela nuvem; N é o número de nuvens de dados de amostras locais observadas em R^M, e também corresponde ao número de regras.

Quando novas amostras são recebidas, a média global μ_k e a média escalar dos produtos x_k são atualizados para μ_{k+1} e x_{k+1}, de acordo com (52) e (53).

$$\mu_k = \frac{k-1}{k}\mu_{k-1} + \frac{1}{k}x_k;\ \mu_1 = x_1. \tag{52}$$

$$x_k = \frac{k-1}{k}x_{k-1} + \frac{1}{k}||x_k||^2; x_1 = ||x_1||^2. \tag{53}$$

Na sequência, são calculadas as densidades unimodais D da amostra de dados x_{k+i} e todos os pontos focais identificados da i-ésima classe, pertencente à nuvem de dados $\mu_{K,i}(i = 1, 2, ..., N_K)$, por meio de (54).

$$D_k(x_i) = \frac{1}{1 + \frac{||x_i - \mu_k||^2}{x_k - ||\mu_k||^2}}. \tag{54}$$

Por fim, é verificado se uma regra será criada da seguinte forma: caso a Condição 1, representada por (55), seja atendida, uma nova regra/nuvem de dados é adicionada em torno de x_{K+1} e os parâmetros atualizados; caso a Condição 1 não seja satisfeita, o algoritmo continua procurando a nuvem de dados mais próxima de x_{K+1}. A Condição 2, representada por (56), por sua vez, verifica se x_{K+1} está próximo da nuvem de dados ou não. Caso a Condição 2 seja satisfeita, os parâmetros das regras *fuzzy*/nuvem de dados próximos são atualizados. Por fim, se nenhuma das condições forem satisfeitas, uma nova regra/nuvem de dados é adicionada em torno de x_{K+1}.

$$\begin{aligned} \textit{Condição 1}: \quad & Se\ (D_{k+1}(x_{k+1}) > \max_{i=1,2,...,N_k}(D_{k+1}(\mu_{k,i}))) \\ & ou\ (D_{k+1}(x_{k+1}) < ,in \underset{i=1,2,...,N_k}{} (D_{k+1}(\mu_{k,i}))) \\ & Então\ (x_{k+1}\ é\ um\ novo\ ponto\ de\ foco). \end{aligned} \tag{55}$$

$$\begin{aligned} \textit{Condição 2}: \quad & Se\ (D_{k+1,i}(x_{k+1}) \geq \tfrac{1}{1+n^2}) \\ & Então\ (atualiza\ os\ dados\ próximos) \end{aligned} \tag{56}$$

3.3.1 *Missing Data Procedure* - MISO

Inicialmente, para cada variável x_{ti} da amostra x_t, o MDP cria uma instância ALMMox_i, responsável por realizar a estimativa dos valores daquela variável de entrada, e uma instância ALMMo$\hat{y}$, responsável pela previsão da saída $\hat{y}$. Caso exista uma ou mais variáveis ausentes na amostra, o MDP imputa o valor estimado $\hat{x}_i$ para a variável com valor ausente. Este processo se repete para cada amostra recebida no fluxo de dados. Caso a ausência ocorra na primeira amostra, assim como descrito para o eNFN-MDP, o valor ausente é substituído por zero. O Algoritmo 5 sumariza o ALMMo-MDP.

Algoritmo 5: ALMMo *with Missing Data Procedure*.

Entrada: x_t, y_t;
Saída: $\hat{y}_t$;
$n =$ número de variáveis de x_t;
Inicializar n instâncias ALMMox;
Inicializar instância ALMMo$\hat{y}$;
for $t = 1, 2, ...$ **do**
  Ler x_t, y_t;
  // Verifica ausência de valores nas variáveis
  for $i = 1 : n$ **do**
    if $x_{ti} = missing_value$ **then**
      if $t \neq 1$ **then**
        $x_{ti} = \hat{x}_i$
      end
      else
        $x_{ti} = 0$;
      end
    end
  end
  Calcula $\hat{y}_t$ executando ALMMo$\hat{y}$
  // Atualiza variáveis estimadas
  for $i = 1 : n$ **do**
    Calcula $\hat{x}_i$ executando ALMMox_i;
  end
end

4 Experimentos Computacionais

Este capítulo apresenta os experimentos computacionais realizados para avaliar o desempenho da abordagem proposta implementada no eNFN e no ALMMo. A Seção 4.1 apresenta a metodologia dos experimentos. Nesta seção são detalhados os algoritmos utilizados para construção de conjuntos de dados com variáveis ausentes e as medidas de desempenho utilizadas. A abordagem será avaliada na previsão do nível de pressão sonora na Seção 4.2, na previsão de coordenada atômica na Seção 4.3, na previsão da umidade absoluta na Seção 4.4 e, por fim, na identificação de sistema não linear na Seção 4.5.

4.1 Metodologia dos Experimentos

Esta seção apresenta como os experimentos computacionais foram conduzidos. A Seção 4.1.1 detalha o procedimento para criação dos conjuntos de dados com variáveis ausentes. Nesta seção são introduzidos os algoritmos utilizados na construção dos conjuntos com dados ausentes para os cenários MCAR e MAR e também a taxa de ausência utilizada em cada experimento. A Seção 4.1.2 destaca as medidas de desempenho utilizadas para avaliar os resultados e também apresenta como os parâmetros dos modelos foram inicializados.

4.1.1 Construção de Conjuntos de Dados com Variáveis Ausentes

Para os experimentos foram selecionados conjuntos de dados que não contém dados ausentes, por ser conveniente para o propósito das experiências e análises comparativas (GARCIA et al., 2019). Esse procedimento se justifica por possibilitar a definição de diferentes percentuais de ausência para os cenários MCAR e MAR, e também por permitir a comparação dos resultados experimentais das técnicas de imputação de dados ausentes com o conjunto completo. Para isso, foram desenvolvidos dois algoritmos que têm por objetivo, a partir de um conjunto de dados completo, criar novos conjuntos com variáveis ausentes de acordo com uma taxa de ausência pré-determinada.

Os conjuntos de dados para os cenários MCAR foram criados utilizando o Algoritmo 6. Este algoritmo recebe como entrada um conjunto de dados completo e a taxa de ausência para o novo conjunto. Com isso ele calcula o número de amostras que deve possuir dados ausentes a partir do número de amostras do conjunto original e da taxa de ausência. Então, o algoritmo sorteia as amostras que possuirão dados ausentes. Por fim, ele faz uma iteração em cada amostra sorteada para conter um dado ausente e sorteia a variável que ficará com o valor ausente. Para cada conjunto de dados foram criados 6 novos conjuntos com taxas

de ausência de 1%, 5%, 10%, 15%, 20% e 30%.

Algoritmo 6: Gerador de Conjuntos de Dados MCAR

Data: ConjuntoOriginal, taxaAusencia
Result: Conjunto de dados MCAR baseado no conjunto original
Obtém a quantidade de amostras e variáveis de entrada do conjunto original
Calcula a quantidade de amostras a receber dado ausente
Sorteia aleatoriamente as amostras que receberão dado ausente
Para *cada amostra sorteada*
  Sorteia uma variável aleatoriamente
  Faz o conjunto receber o dado ausente na amostra iterada e variável sorteada
Fim

O Algoritmo 7 foi utilizado para gerar os conjuntos de dados ausentes em cenários MAR. Este algoritmo recebe o conjunto de dados original, a taxa de ausência para a variável com maior propensão e a taxa de ausência para as demais variáveis. O algoritmo seleciona aleatoriamente uma variável para ser a variável com maior propensão a possuir valor ausente e calcula a probabilidade das demais variáveis estarem ausentes com base no número de variáveis e na taxa mínima de ausência estabelecida. Em seguida, o algoritmo faz uma iteração em cada variável para que cada uma ocupe uma quantidade de posições no vetor de sorteio relativa à sua porcentagem probabilística de ausência. Por fim, ao iterar em cada amostra do conjunto de dados original, o algoritmo seleciona aleatoriamente um índice do vetor de sorteio e, caso seja sorteado o índice de uma variável, a variável fica com o valor ausente. Este algoritmo foi utilizado com as taxas de ausência de 5x1%, 10x1%, 10x5%, 20x5%, 20x10%, 30x5% e 30x10%, onde o primeiro dígito se refere a taxa de ausência da variável mais propensa e o segundo se refere a taxa de ausência de cada uma das demais variáveis.

4.1.2 Medidas de Desempenho

Os experimentos computacionais foram executados de forma *online*, isto é, durante a estimativa dos parâmetros e avaliação do desempenho os modelos evoluem sua estrutura para todas as amostras do conjunto de dados. O desempenho é avaliado para todas as amostras pelo RMSE (*Root Mean Square Error*) calculado por:

$$RMSE = \frac{1}{N}\sum_{t=1}^{N}\sqrt{(y_t - \hat{y}_t)^2}, \tag{57}$$

onde N é o número de amostras, y_t é a saída estimada e $\hat{y}_t$ é a saída desejada.

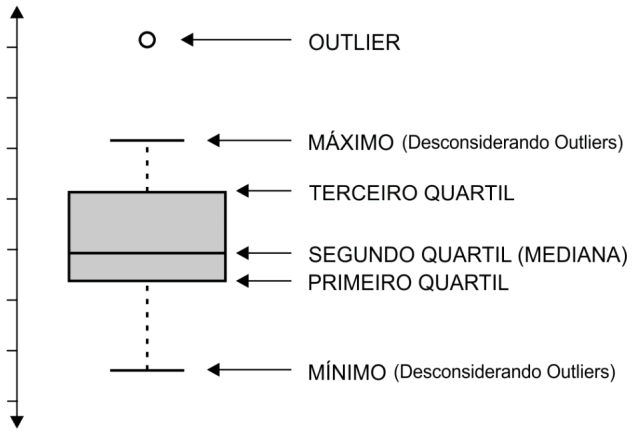
Para análise estatística dos resultados foi utilizado o *boxplot* (diagrama de caixa). O *boxplot* é um gráfico utilizado para avaliar a distribuição empírica dos resultados representando a variação destes dados por meio de quartis. Basicamente o gráfico é composto pelos

Algoritmo 7: Gerador de Conjuntos de Dados MAR

```
Data: ConjuntoOriginal, taxaAusenciaMax, taxaAusenciaMin
Result: Conjunto de dados MAR baseada no conjunto original
Obtém a quantidade de amostras e variáveis do conjunto original
Cria uma lista de sorteio de tamanho 100 e inicializa com zero
Sorteia uma variável para ter taxa de ausência máxima
Calcula a taxa de ausência das demais variáveis
Para cada variável do conjunto original
    if a variável for a sorteada para taxa máxima then
        Sorteia posições da lista de sorteio usando a taxaAusenciaMax para
         definir a quantidade
    else
        Sorteia posições da lista de sorteio usando a taxa de ausência calculada
         para definir a quantidade
    end
    Altera os valores das posições selecionadas da lista de sorteio com o valor
     da variável
Fim
Para cada amostra do conjunto original
    Sorteia aleatoriamente uma posição da lista de sorteio
    if a posição sorteada não for igual a zero then
        Obtém a variável pelo valor sorteado da lista de sorteio
        Altera o conjunto original na amostra iterada e variável obtida com o valor
         ausente
    end
Fim
```

limites inferior e superior, que formam a haste principal, o primeiro e terceiro quartis, que formam a caixa principal que contempla 50% dos dados e a mediana (segundo quartil) dos dados, e os *outliers*, que representam os resultados discrepantes. A Figura 5 representa a estrutura básica de um *boxplot*.

Figura 5 – Estrutura básica de um gráfico *boxplot*.



O *boxplot* mostra a distribuição dos resultados com base em cinco números (mínimo, primeiro quartil, mediana, terceiro quartil e máximo). Pode-se afirmar, sem perda de valor

estatístico que se duas caixas não se sobrepõem, há uma diferença estatística entre os dois modelos, sendo possível afirmar que a caixa com resultado inferior (menor erro) é estatisticamente superior. Dessa forma, o objetivo é analisar qual modelo é estatisticamente superior. Caso haja semelhança estatística, definida pela sobreposição entre as caixas do gráfico, é aplicado um teste T, unilateral, pareado e com um nível de significância de $0{,}05$, para avaliar se há superioridade estatística.

Os experimentos com a abordagem eFGP foram executados 3 vezes, sendo apresentado a média dos resultados. Como demais modelos são determinísticos os seus experimentos foram executados somente 1 vez. Os parâmetros do eNFN-MDP foram definidos com $\omega = 100$, $\eta = 10$ e $\beta = 0{,}01$. Para o ALMMo-MDP os parâmetros iniciais foram $\Omega = 10$, $\eta_0 = 0{,}1$ e $D_0 = 0{,}8$. Os parâmetros do eFPG foram definidos a partir de testes com as bases originais completas onde cada base foi executada utilizando cada conjunto de parâmetros $\rho = \sigma = [0{,}1\ 0{,}2\ 0{,}3\ 0{,}4\ 0{,}5\ 0{,}6]$ e $h_r = [48\ 100\ 250\ 500\ 750\ 1000]$ e o melhor par de parâmetros foi definido para cada base pela melhor média dos resultados de 3 execuções, sendo: $\rho = \sigma = 0{,}4$ e $h_r = 48$, para a previsão do nível de pressão sonora; $\rho = \sigma = 0{,}6$ e $h_r = 1000$, para a previsão da coordenada atômica Calculada w'; $\rho = \sigma = 0{,}6$ e $h_r = 1000$, para a previsão da umidade absoluta e; $\rho = \sigma = 0{,}4$ e $h_r = 500$, na identificação de processo não linear. Para geração dos conjuntos com dados ausentes, os conjuntos de dados foram normalizados entre 0 e 1 antes de passarem pelo processo de geração dos dados ausentes.

4.2 Previsão do Nível de Pressão Sonora - Airfoil Self-Noise

Esta seção apresenta os experimentos computacionais realizados para avaliar os modelos propostos na previsão do nível de pressão sonora utilizando o conjunto de dados *Airfoil Self-Noise*[1] (DUA; GRAFF, 2017). O conjunto obtido de testes aerodinâmicos e acústicos realizados pela NASA (*National Aeronautics and Space Administration*), possui 1503 amostras com 5 variáveis de entrada: frequência (*hertz*), ângulo (grau), comprimento da corda (metro), velocidade do fluxo livre (metros por segundo), espessura do deslocamento lateral da sucção (metro). Dada as entradas, o objetivo é prever o nível de pressão sonora (decibel). *Airfoil Self-Noise* é um *benchmark* clássico e já foi utilizado em Brooks, Pope e Marcolini (1989), Lopez, Balsa-Canto e Oñate (2008), Lau et al. (2009), Garcia, Leite e Skrjanc (2019), para citar somente alguns.

O eNFN-MDP e o ALMMo-MDP foram avaliados na previsão do nível de pressão sonora em comparação com as técnicas de imputação pela última saída, pela média dos últimos 5 valores e por zero nas Seções 4.2.1 e 4.2.2, respectivamente, e comparados com o modelo eFGP na Seção 4.2.3.

[1] http://archive.ics.uci.edu/ml/datasets/airfoil+self-noise

4.2.1 eNFN-MDP na Previsão do Nível de Pressão Sonora

Os resultados experimentais com o conjunto de dados *Airfoil Self-Noise* para os cenários MCAR são sintetizados na Tabela 1. O eNFN-MDP obteve o melhor desempenho em 5 dos 6 experimentos e também média dos experimentos. No cenário com taxa de ausência de 15%, a imputação pela média obteve o melhor resultado, seguido pelo eNFN-MDP. Em comparação ao resultado do conjunto original, o eNFN-MDP e a imputação pela média obtiveram médias melhores do que o resultado do conjunto original.

Tabela 1 – Desempenho RMSE do eNFN-MDP na previsão do nível da pressão sonora nos cenários MCAR.

Ausência	**Última Saída**	**Média**	**Zero**	**eNFN-MDP**
0%	0,09325			
1%	0,09445	0,09300	0,09382	0,09121
5%	0,10035	0,09144	0,09195	0,09078
10%	0,10541	0,09339	0,09486	0,09080
15%	0,11767	0,09329	0,09370	0,09417
20%	0,12378	0,09437	0,09836	0,09363
30%	0,14863	0,09364	0,09799	0,09356
Média	0,11505	0,09319	0,09511	0,09236

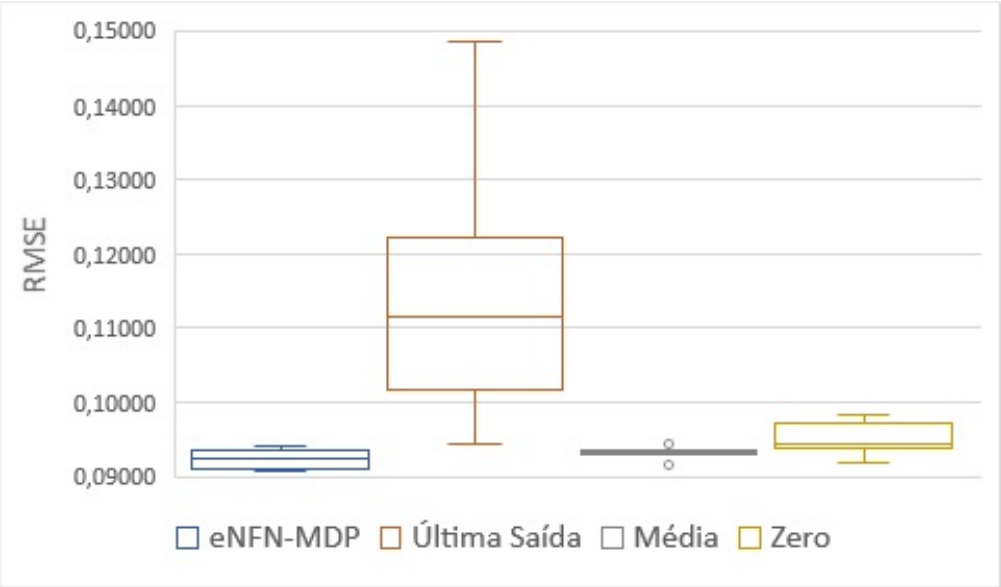
A Figura 6 ilustra o *boxplot* para o eNFN-MDP na previsão do nível da pressão sonora nos cenários MCAR. Os resultados sugerem que o eNFN-MDP é estatisticamente superior a última saída. Em relação aos demais, o teste T entre o eNFN-MDP e eles obteve o valor de $0{,}01$ em comparação com a imputação por zero, indicando que o eNFN-MDP é estatisticamente superior, e $0{,}08$ em relação a imputação pela média dos últimos 5 valores, indicando que não há diferença estatística entre eles. Considerando os outros índices analisados, o eNFN-MDP possui a menor mediana, enquanto a imputação pela média possui a menor diferença entre seus limites superior inferior.

A Tabela 2 ilustra os resultados experimentais obtidos para os cenários MAR. O eNFN-MDP obteve o melhor resultado em 6 dos 7 experimentos e na média dos resultados. No experimento com taxa de ausência de 30x5%, a média obteve o melhor resultado, seguido pelo eNFN-MDP. Em comparação com o conjunto original, somente o eNFN-MDP obteve uma média melhor que o resultado obtido com o conjunto original.

Tabela 2 – Desempenho RMSE do eNFN-MDP na previsão do nível da pressão sonora nos cenários MAR.

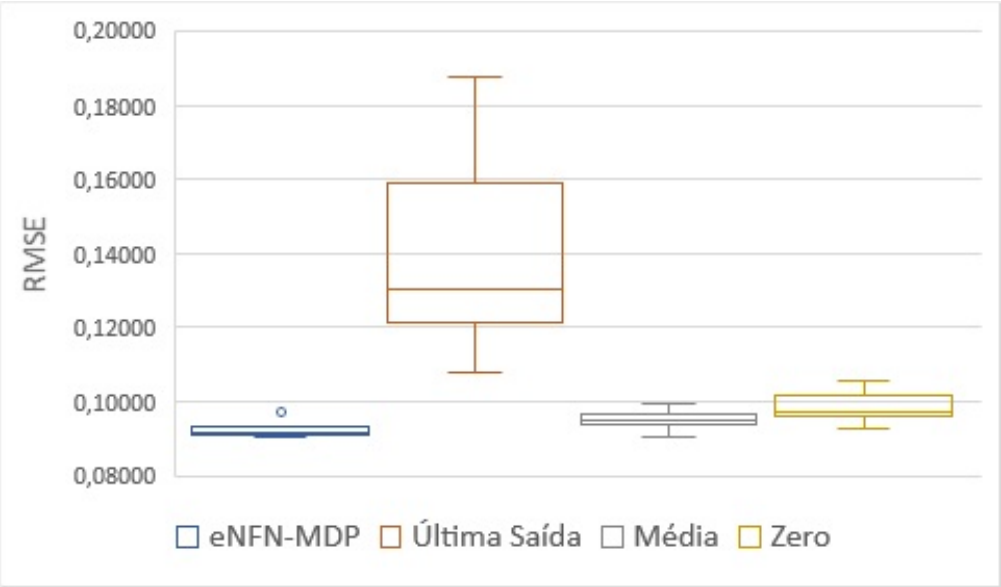
Ausência	**Última Saída**	**Média**	**Zero**	**eNFN-MDP**
0x0%	0,09325			
5x1%	0,10795	0,09351	0,09612	0,09061
10x1%	0,11295	0,09430	0,09702	0,09173
10x5%	0,13058	0,09489	0,09556	0,09139
20x5%	0,13010	0,09961	0,10535	0,09309
20x10%	0,15766	0,09868	0,10207	0,09723
30x5%	0,16051	0,09063	0,09272	0,09100
30x10%	0,18747	0,09486	0,10111	0,09333
Média	0,14103	0,09521	0,09857	0,09263

Figura 6 – Boxplot do eNFN-MDP na previsão do nível da pressão sonora nos cenários MCAR.



O *boxplot* ilustrado na Figura 7 apresenta os resultados obtidos pelo eNFN-MDP na previsão do nível da pressão sonora nos cenários MAR. Estes mostram que o eNFN-MDP é estatisticamente superior a imputação pela última saída e por zero. Comparando com os resultados da imputação pela média, onde houve sobreposição, foi obtido o valor de $0{,}01$ no teste T, indicando que o eNFN-MDP é estatisticamente superior.

Figura 7 – Boxplot do eNFN-MDP na previsão do nível da pressão sonora nos cenários MAR.



4.2.2 ALMMo-MDP na Previsão do Nível de Pressão Sonora

Os resultados do ALMMo no conjunto Airfoil Self-Noise para os cenários MCAR são apresentados na Tabela 3. O ALMMo-MDP obteve o menor RMSE em 4 dos 6 cenários e na média dos resultados. Para o cenário com 1% de ausência a imputação pela última saída obteve o melhor desempenho e no cenário com 5% foi a imputação por zero, ambos seguidos pelos ALMMo-MDP. Com relação aos resultados do conjunto original, a média dos resultados do ALMMo-MDP, da imputação pela média e por zero obtiveram proximidade com o RMSE do conjunto original em duas casas decimais.

Tabela 3 – Desempenho RMSE do ALMMo-MDP na previsão do nível da pressão sonora nos cenários MCAR.

Ausência	**Última Saída**	**Média**	**Zero**	**ALMMo-MDP**
0%	0,12644			
1%	0,12550	0,12769	0,12869	0,12762
5%	0,12786	0,12726	0,12690	0,12808
10%	0,13125	0,12943	0,12925	0,12634
15%	0,13396	0,12831	0,12981	0,12827
20%	0,14025	0,12775	0,13150	0,12683
30%	0,14323	0,13113	0,13341	0,12766
Média	0,13368	0,12860	0,12993	0,12747

A Figura 8 mostra o *boxplot* dos resultados experimentais do ALMMo-MDP na previsão do nível da pressão sonora nos cenários MCAR. Pode-se observar que o ALMMo-MDP é estatisticamente superior a imputação pela última saída e por zero. A execução do teste T comparando o ALMMo-MDP e a imputação pela média obteve o valor de $0{,}09$, indicando que não há superioridade estatística. Com relação aos demais índices analisados, o ALMMo-MDP obteve a menor mediana e a menor diferença entre os limites superior e inferior.

A Tabela 4 apresenta os resultados para os cenários MAR. O ALMMo-MDP obteve o melhor resultado em 5 dos 7 cenários e na média dos resultados. No cenário com taxa de ausência de 5x1% a imputação por zero obteve o melhor desempenho e, para a taxa de 10x5% este foi obtido pela imputação pela média dos últimos 5 valores, ambos seguidos pelo ALMMo-MDP. Em comparação com o resultado do conjunto original, a média dos resultados do ALMMo-MDP obteve uma proximidade de três casas decimais, enquanto a média da imputação pela média obteve uma proximidade de duas casas decimais.

A Figura 9 ilustra o *boxplot* para os cenários MAR na previsão do nível da pressão sonora pelo ALMMo-MDP. O *boxplot* sugere que o ALMMo-MDP é estatisticamente superior a imputação pela última saída e por zero. O teste T entre o ALMMo-MDP e a imputação pela média obteve o valor de $0{,}08$, indicando que não houve superioridade estatística. Considerando os demais índices, o ALMMo-MDP obteve a melhor mediana e a menor distância entre seus limites.

Figura 8 – Boxplot do ALMMo-MDP na previsão do nível da pressão sonora nos cenários MCAR.

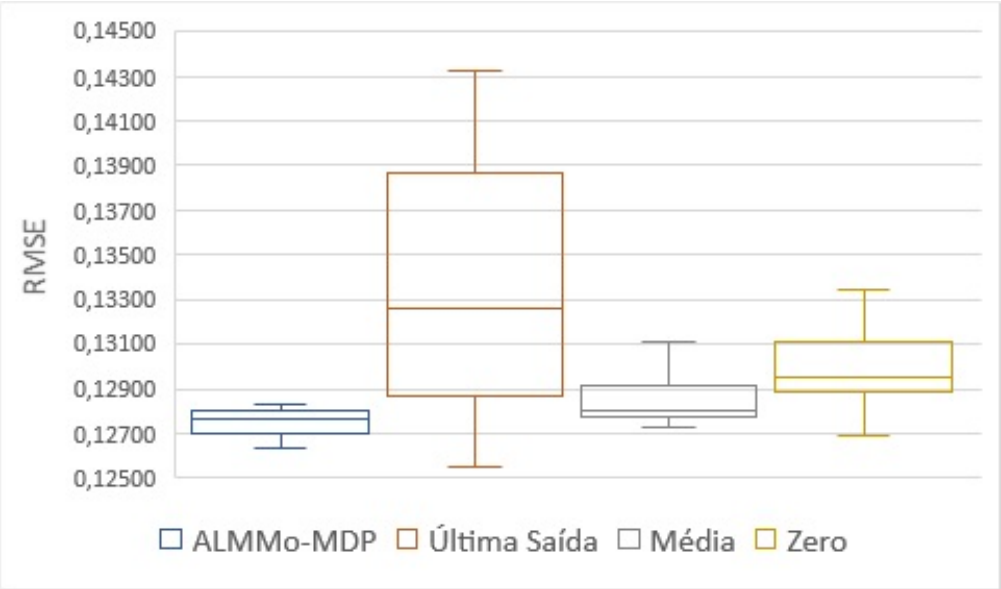
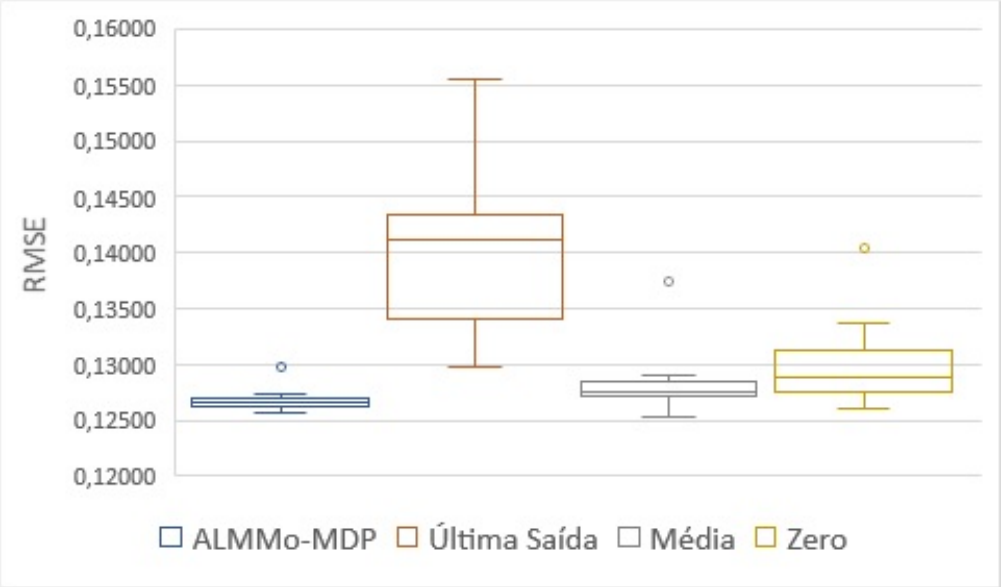


Tabela 4 – Desempenho RMSE do ALMMo-MDP na previsão do nível da pressão sonora nos cenários MAR.

Ausência	Última Saída	Média	Zero	ALMMo-MDP
0x0%		0,12644		
5x1%	0,12978	0,12703	0,12596	0,12643
10x1%	0,13487	0,12779	0,12766	0,12661
10x5%	0,13334	0,12523	0,12876	0,12742
20x5%	0,14110	0,12720	0,12752	0,12669
20x10%	0,14299	0,13737	0,14049	0,12971
30x5%	0,14395	0,12761	0,12901	0,12592
30x10%	0,15548	0,12905	0,13366	0,12573
Média	0,14021	0,12875	0,13044	0,12693

Figura 9 – Boxplot do ALMMo-MDP na previsão do nível da pressão sonora nos cenários MAR.



4.2.3 MDPs na Previsão do Nível de Pressão Sonora

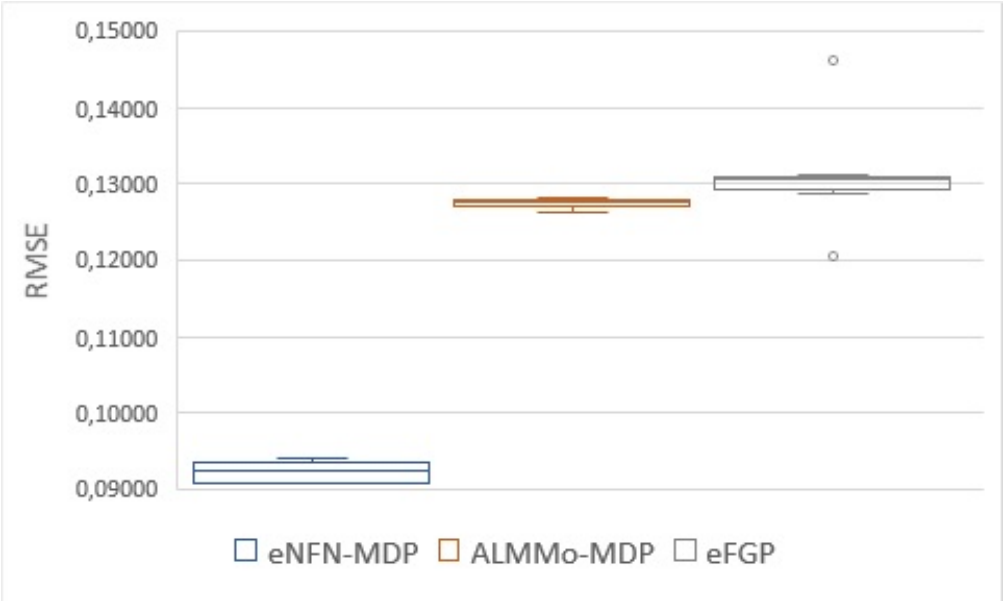
A Tabela 5 ilustra o RMSE do ALMMo-MDP, eNFN-MDP e eFGP para os cenários MCAR no conjunto *Airfoil Self-Noise*. Em todos os cenários o eNFN-MDP obteve o melhor desempenho, seguido pelo ALMMo-MDP. Na média dos resultados o eNFN-MDP obteve o menor RMSE seguido do ALMMo-MDP e do eFGP, respectivamente. Com relação ao resultado do conjunto original, o eNFN-MDP obteve uma média dos resultados melhor que o resultado do conjunto original enquanto o ALMMo-MDP e o eFGP obtiveram uma proximidade de duas casa decimais.

Tabela 5 – Desempenho RMSE dos MDPs na previsão do nível da pressão sonora nos cenários MCAR.

Ausência	eNFN-MDP	ALMMo-MDP	eFGP
0%	0,09325	0,12644	0,12258
1%	0,09121	0,12762	0,12050
5%	0,09078	0,12808	0,13053
10%	0,09080	0,12634	0,13082
15%	0,09417	0,12827	0,13110
20%	0,09363	0,12683	0,12886
30%	0,09356	0,12766	0,14617
Média	0,09236	0,12747	0,13133

O *boxplot* apresentado na Figura 10 mostra os resultados experimentais dos modelos MDPs comparando com os resultados do eFGP para os cenários MCAR na previsão do nível da pressão sonora. De acordo com os resultados, a análise do gráfico indica que o eNFN-MDP é estatisticamente superior aos demais, seguido pelo ALMMo-MDP.

Figura 10 – Boxplot dos MDPs na previsão do nível da pressão sonora nos cenários MCAR.



Os resultados do ALMMo-MDP e do eNFN-MDP são comparados com os resultados do eFGP para os cenários MAR na Tabela 6. Como é possível verificar, o eNFN-MDP obteve o melhor resultado em todos os cenários seguido pelo ALMMo-MDP, o que também ocorreu

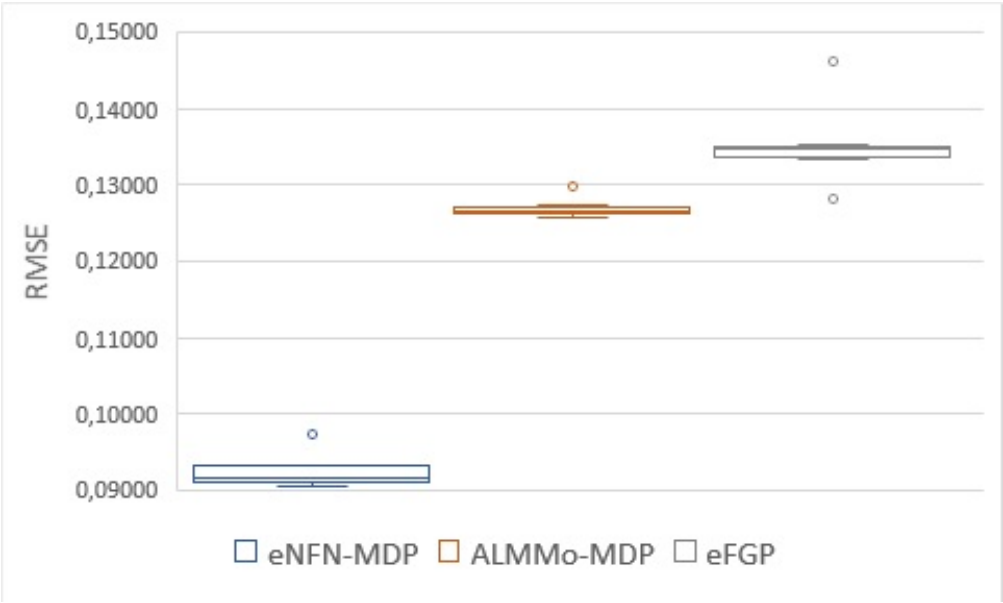
na média dos resultados, consequentemente. Em comparação com o conjunto original, o eNFN-MDP obteve uma média dos resultado melhor que o resultado do conjunto original, enquanto o ALMMo-MDP obteve a média dos resultados com uma proximidade de duas casas decimais e o eFGP com a média dos resultados com uma casa de proximidade ao resultado do conjunto original.

Tabela 6 – Desempenho RMSE dos MDPs na previsão do nível da pressão sonora nos cenários MAR.

Ausência	**eNFN-MDP**	**ALMMo-MDP**	**eFGP**
0x0%	0,09325	0,12644	0,12258
5x1%	0,09061	0,12643	0,12815
10x1%	0,09173	0,12661	0,13352
10x5%	0,09139	0,12742	0,13388
20x5%	0,09309	0,12669	0,13537
20x10%	0,09723	0,12971	0,14630
30x5%	0,09100	0,12592	0,13463
30x10%	0,09333	0,12573	0,13463
Média	0,09263	0,12693	0,13521

O *boxplot* apresentado na Figura 11 mostra os resultados experimentais dos modelos MDPs em comparação com os resultados do eFGP na previsão do nível da pressão sonora para os cenários MAR. Sua análise indica uma superioridade estatística do eNFN-MDP em relação aos demais modelos, seguido pelo ALMMo-MDP.

Figura 11 – Boxplot dos MDPs na previsão do nível da pressão sonora nos cenários MAR.



4.3 Previsão da Coordenada Atômica Calculada w' - Carbon Nanotubes

Esta seção apresenta os experimentos computacionais realizados para avaliar os modelos propostos na previsão da coordenada atômica calculada w' no conjunto de dados

Carbon Nanotubes[2] (DUA; GRAFF, 2017). O conjunto de dados possui 10.721 amostras com 7 variáveis de entrada, sendo elas: índice quiral n do vetor quiral selecionado, índice quiral m do vetor quiral selecionado, coordenada atômica inicial u, coordenada atômica inicial v, coordenada atômica inicial w, coordenada atômica calculada u' e coordenada atômica calculada v'. O objetivo é realizar a previsão da coordenada atômica calculada w'. Este *benchmark* foi utilizado nos trabalhos de Acı e Avcı (2016), Al. (2017), por exemplo.

O eNFN-MDP e o ALMMo-MDP foram avaliados na previsão da coordenada atômica calculada w' em comparação com as técnicas de imputação pela última saída, pela média dos últimos 5 valores e por zero nas Seções 4.3.1 e 4.3.2, respectivamente, e comparados com o modelo eFGP na Seção 4.3.3.

4.3.1 eNFN-MDP na Previsão da Coordenada Atômica Calculada w'

Os resultados obtidos para os cenários MCAR, presentes na Tabela 7, mostram que o eNFN-MDP obteve o melhor resultado na metade dos experimentos, seguido pela média dos últimos 5 valores. A imputação pela média obteve o melhor resultados nos cenários com 1%, 5% e 20% de ausência, seguido pelo eNFN-MDP. Na média dos resultados o eNFN-MDP obteve o melhor resultado. Com relação ao conjunto original, a média dos resultados do eNFN-MDP e da imputação pela média obtiveram uma casa decimal de proximidade do resultado do conjunto original.

Tabela 7 – Desempenho RMSE do eNFN-MDP na previsão da coordenada atômica calculada w' nos cenários MCAR.

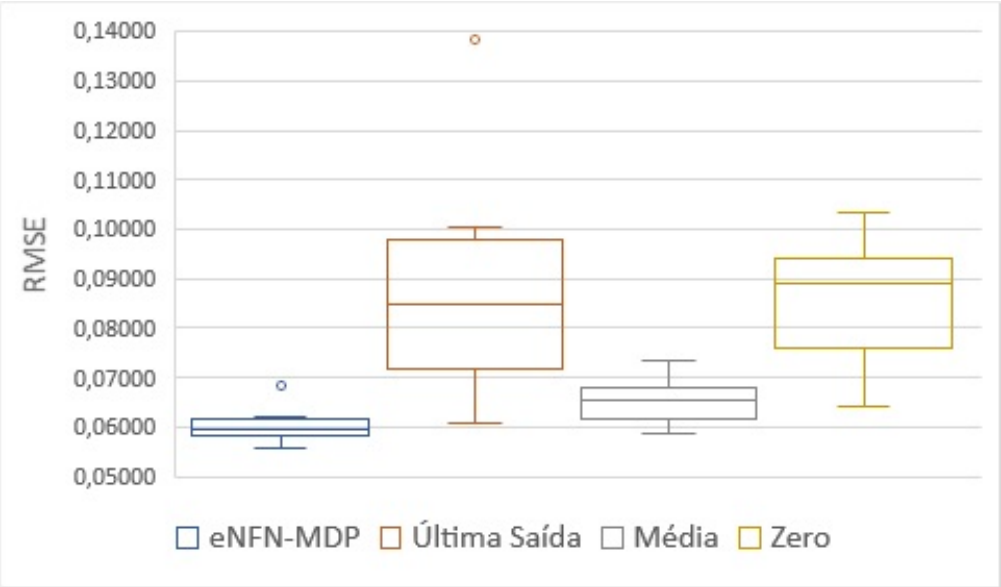
Ausência	Última Saída	Média	Zero	eNFN-MDP
0%	0,05764			
1%	0,06090	0,05888	0,06425	0,05924
5%	0,06933	0,06083	0,07235	0,06213
10%	0,07950	0,06596	0,08599	0,05770
15%	0,09030	0,06836	0,09234	0,05586
20%	0,10021	0,06475	0,09494	0,06856
30%	0,13832	0,07333	0,10356	0,06013
Média	0,08976	0,06535	0,08557	0,06060

A Figura 12 apresenta o *boxplot* dos resultados obtidos para os cenários MCAR na previsão da coordenada atômica calculada w' para o eNFN-MDP. O gráfico sinaliza que os resultados do eNFN-MDP são estatisticamente superiores aos resultados da imputação pela última saída e por zero, uma vez que não houve sobreposição. Comparando o eNFN-MDP e a imputação pela média, o teste T obteve $0{,}09$, indicando que não houve superioridade estatística. Para os demais índices analisados, o eNFN-MDP obteve a menor mediana e a menor diferença entre os limites superior e inferior.

Considerando os resultados para os cenários MAR, apresentados na Tabela 8, verifica-se que o eNFN-MDP obteve o melhor resultados em 6 dos 7 cenários e na média

[2] http://archive.ics.uci.edu/ml/datasets/Carbon+Nanotubes

Figura 12 – Boxplot do eNFN-MDP na previsão da coordenada atômica calculada w' nos cenários MCAR.



dos resultados. No cenário com 5x1% a imputação pela média dos últimos 5 valores obteve o melhor resultados seguido pelo eNFN-MDP. Comparando com o resultado do conjunto original, a média dos resultados do eNFN-MDP obteve um resultado melhor do que o conjunto original, enquanto a média dos resultados da imputação pela média obteve uma casa decimal de proximidade.

Tabela 8 – Desempenho RMSE do eNFN-MDP na previsão da coordenada atômica calculada w' nos cenários MAR.

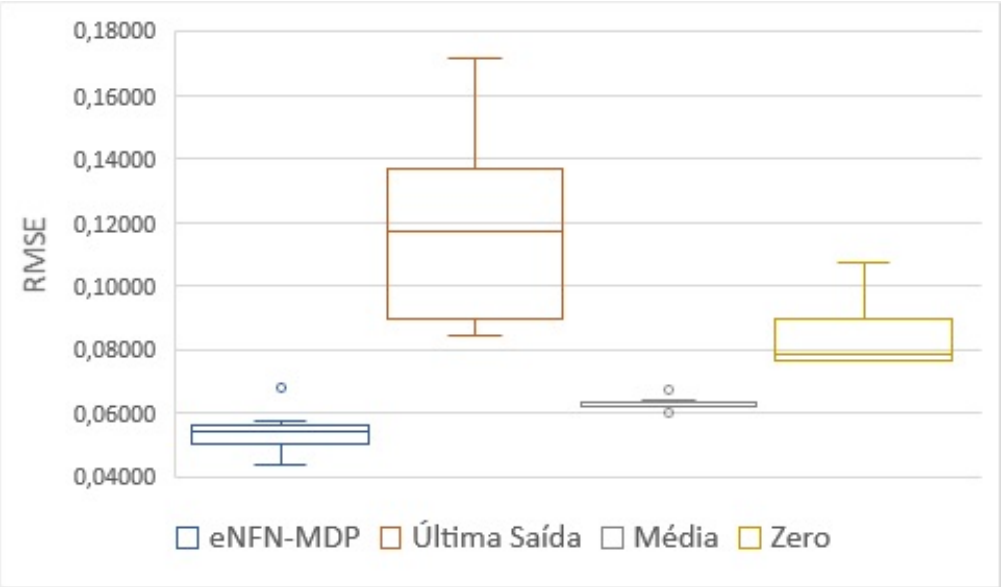
Ausência	Última Saída	Média	Zero	eNFN-MDP
0x0%		0,05764		
5x1%	0,08642	0,06745	0,10713	0,06779
10x1%	0,08423	0,06002	0,07645	0,05731
10x5%	0,09281	0,06328	0,07657	0,05440
20x5%	0,11737	0,06278	0,07823	0,04343
20x10%	0,13208	0,06334	0,08769	0,05419
30x5%	0,14105	0,06177	0,07661	0,04789
30x10%	0,17125	0,06383	0,09120	0,05221
Média	0,11789	0,06321	0,08484	0,05389

O *boxplot* dos resultados experimentais do eNFN-MDP para os cenários MAR na previsão da coordenada atômica calculada w' é apresentado na Figura 13. O gráfico indica que, como não houve sobreposição, os resultados do eNFN-MDP foi estatisticamente superior aos demais resultados.

4.3.2 ALMMo-MDP na Previsão da Coordenada Atômica Calculada w'

Conforme é possível verificar na Tabela 9, que apresenta os resultados obtidos nos experimentos dos cenários MCAR, o ALMMo-MDP obteve o melhor resultado em 5 dos 6 e na média dos resultados. No cenário com 15% de ausência o melhor resultado foi obtido

Figura 13 – Boxplot do eNFN-MDP na previsão da coordenada atômica calculada w' nos cenários MAR.



pela imputação pela média dos últimos 5 valores, seguido pelo ALMMo-MDP. Em relação o conjunto original, a média dos resultados do ALMMo-MDP e das técnicas obtiveram a proximidade de uma casa decimal com o resultado do conjunto original.

Tabela 9 – Desempenho RMSE do ALMMo-MDP na previsão da coordenada atômica calculada w' nos cenários MCAR.

Ausência	Última Saída	Média	Zero	ALMMo-MDP
0%	0,01550			
1%	0,02048	0,01737	0,03005	0,01731
5%	0,03672	0,02583	0,05095	0,02607
10%	0,04254	0,03197	0,07476	0,02739
15%	0,05706	0,03244	0,08637	0,02545
20%	0,06649	0,03576	0,09273	0,03315
30%	0,09493	0,03814	0,11481	0,03002
Média	0,05304	0,03025	0,07494	0,02656

A Figura 14 ilustra os resultados dos experimentos na previsão da coordenada atômica calculada w' nos cenários MCAR para o ALMMo-MDP. A Figura sugere que, como não houve sobreposição das caixas, o ALMMo-MDP é estatisticamente superior a substituição pela última saída e por zero. Em comparação entre o ALMMo-MDP e a média, o teste T obteve o valor de $0{,}02$, indicando que o ALMMo-MDP é estatiscamente superior.

De acordo com os resultados dos cenários MAR, presentes na Tabela 10, o ALMMo-MDP obteve o melhor resultado em todos os cenários e, consequentemente, na média deles, seguido pela imputação pela média dos últimos 5 valores. Análogo ao ocorrido para os cenários MCAR, a média dos resultados do ALMMo-MDP e das técnicas obtiveram uma casa decimal de proximidade em comparação com o resultado do conjunto original.

O *boxplot* dos resultados experimentais do ALMMo-MDP nos cenários MAR na

Figura 14 – Boxplot do ALMMo-MDP na previsão da coordenada atômica calculada w' nos cenários MCAR.

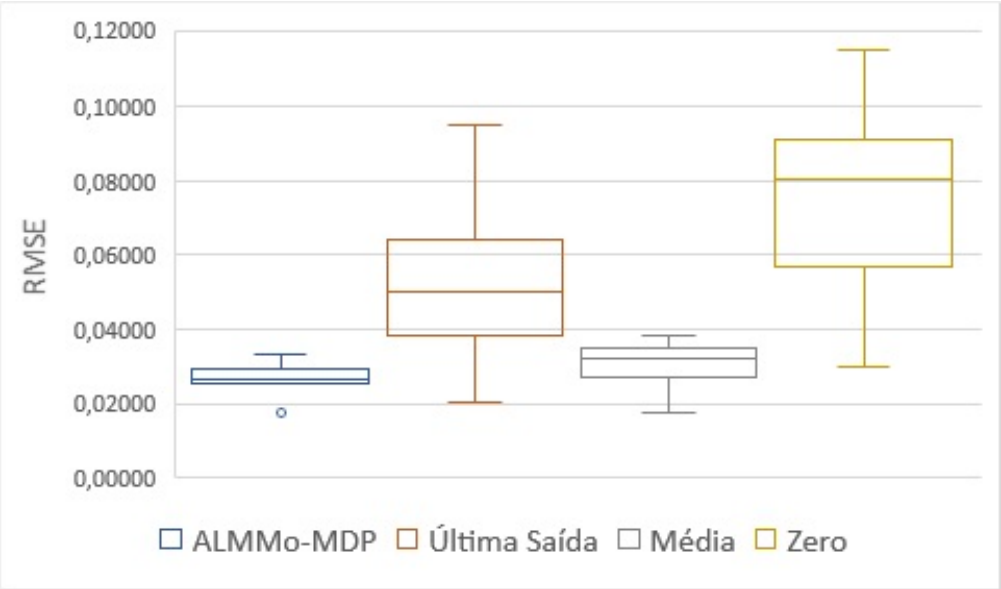


Tabela 10 – Desempenho RMSE do ALMMo-MDP na previsão da coordenada atômica calculada w' nos cenários MAR.

Ausência	Última Saída	Média	Zero	ALMMo-MDP
0x0%		0,01550		
5x1%	0,05107	0,04111	0,12276	0,03303
10x1%	0,05372	0,02105	0,06091	0,01704
10x5%	0,06050	0,02082	0,05720	0,01839
20x5%	0,08056	0,02198	0,05819	0,02052
20x10%	0,09398	0,02894	0,08021	0,02296
30x5%	0,09409	0,02320	0,05831	0,01993
30x10%	0,11265	0,02762	0,08365	0,02262
Média	0,07808	0,02639	0,07446	0,02207

previsão da coordenada atômica calculada w' são apresentados na Figura 15. O gráfico ressalta que o ALMMo-MDP é superior a imputação pela última saída e por zero. Em relação a imputação pela média, o teste T obteve o resultado de $0{,}00$, indicando que o ALMMo-MDP é estatiscamente superior.

4.3.3 MDPs na Previsão da Coordenada Atômica Calculada w'

Pela Tabela 11, onde estão representados os resultados do ALMMo-MDP comparados aos resultados do eNFN-MDP e do eFGP para os cenários MCAR do conjunto Carbon Nanotubes, verifica-se que, para todas as taxas de ausência e na média dos resultados o ALMMo-MDP obteve o melhor resultado, seguido pelo eNFN-MDP. Comparando com o resultado do conjunto original, a média de todos os modelos obtiveram uma casa decimal de proximidade com o resultado do conjunto original.

A Figura 16 compila em um gráfico *boxplot* os resultados dos experimentos dos cenários MCAR na previsão da coordenada atômica calculada w' para os modelos MDPs

Figura 15 – Boxplot do ALMMo-MDP na previsão da coordenada atômica calculada w' nos cenários MAR.

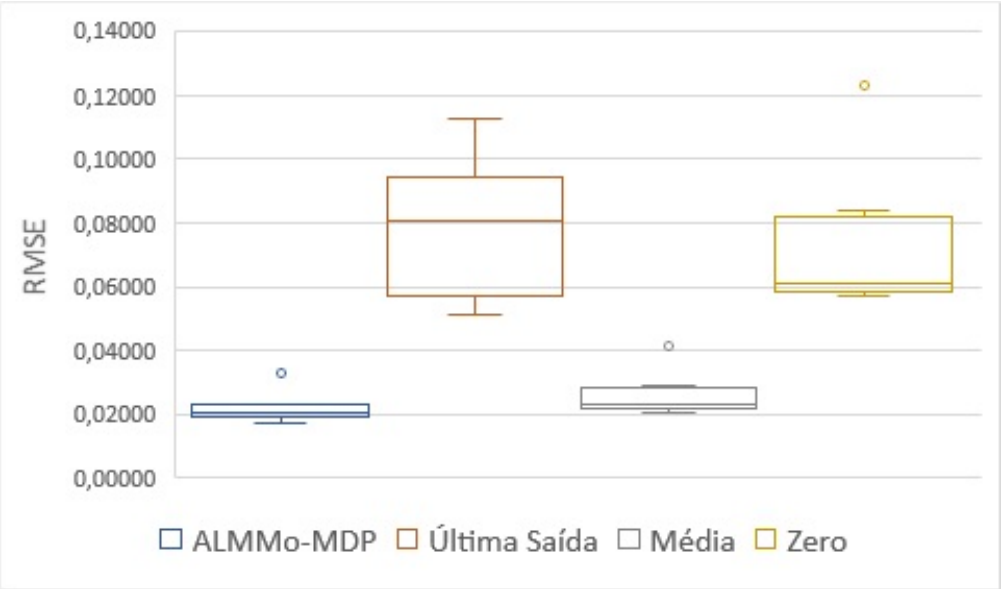


Tabela 11 – Desempenho RMSE dos MDPs na previsão da coordenada atômica calculada w' nos cenários MCAR.

Ausência	eNFN-MDP	ALMMo-MDP	eFGP
0%	0,05764	0,01550	0,05836
1%	0,05924	0,01731	0,05288
5%	0,06213	0,02607	0,07082
10%	0,05770	0,02739	0,06809
15%	0,05586	0,02545	0,07701
20%	0,06856	0,03315	0,08410
30%	0,06013	0,03002	0,09127
Média	0,06060	0,02656	0,07403

em comparação com o eFGP. Pelos resultados, é possível definir que o ALMMo-MDP foi estatisticamente superior aos demais modelos, seguido pelo eNFN-MDP.

Os resultados dos cenários MAR, apresentados na Tabela 12, mostram que o ALMMo-MDP obteve o melhor resultado em todos os cenários e na média dos resultados, seguido pelo eNFN-MDP. Com exceção do eNFN-MDP, cuja média dos resultados foi menor que o resultado do conjunto original, a média dos resultados dos demais modelos obtiveram uma casa decimal de proximidade com o resultado do conjunto original.

O *boxplot* ilustrado na Figura 17 apresenta os resultados na previsão da coordenada atômica calculada w' para os modelos MDPs comparando com o eFGP nos cenários MAR. Como não houve sobreposição, os resultados do ALMMo-MDP são estatisticamente superiores aos demais resultados, seguido pelo eNFN-MDP.

Figura 16 – Boxplot dos MDPs na previsão da coordenada atômica calculada w' nos cenários MCAR.

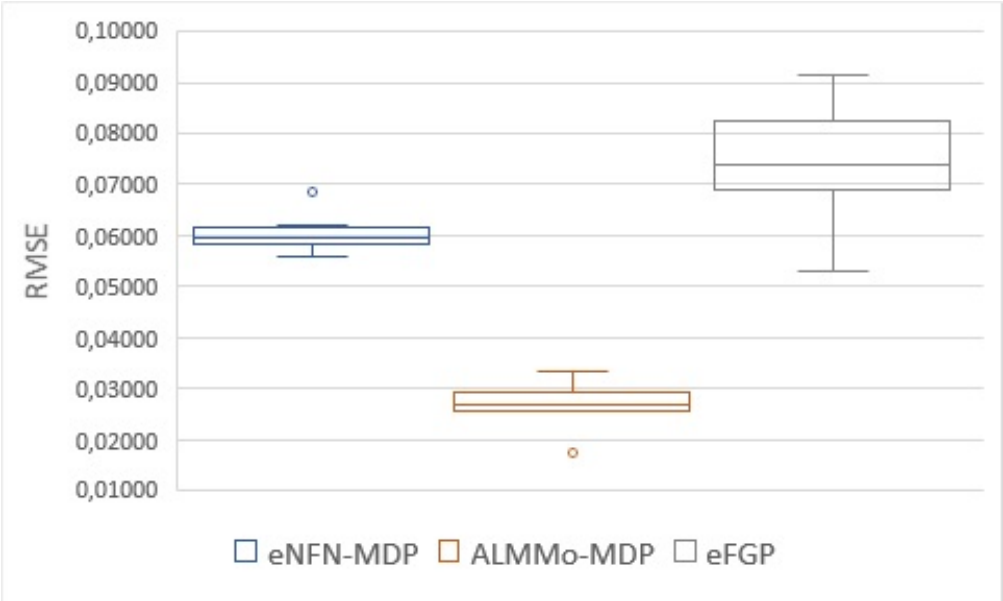


Tabela 12 – Desempenho RMSE dos MDPs na previsão da coordenada atômica calculada w' nos cenários MAR.

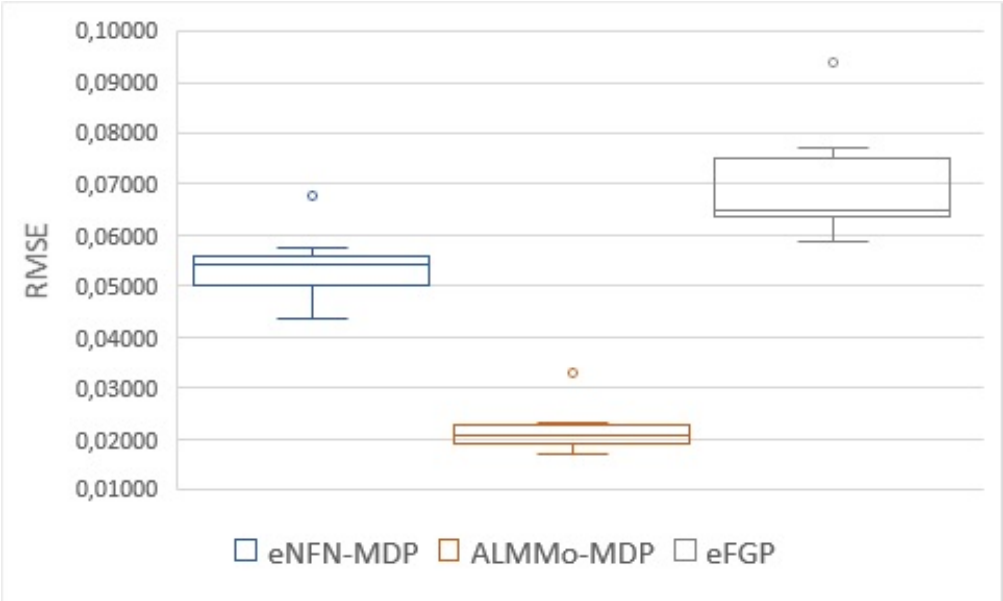
Ausência	eNFN-MDP	ALMMo-MDP	eFGP
0x0%	0,05764	0,01550	0,05836
5x1%	0,06779	0,03303	0,09401
10x1%	0,05731	0,01704	0,06312
10x5%	0,05440	0,01839	0,05862
20x5%	0,04343	0,02052	0,06389
20x10%	0,05419	0,02296	0,07288
30x5%	0,04789	0,01993	0,06486
30x10%	0,05221	0,02262	0,07730
Média	0,05389	0,02207	0,07067

4.4 Previsão da Umidade Absoluta - Air Quality

Nesta seção o desempenho dos modelos propostos é avaliado na previsão da umidade absoluta. Os experimentos foram conduzidos com o conjunto de dados Air Quality[3] (DUA; GRAFF, 2017). Os dados são provenientes de um dispositivo multissensor a gás implantado em uma cidade italiana e possui 9357 amostras com 12 variáveis de entrada: concentração média horária de CO (mg/m^3), resposta média do sensor relacionada ao óxido de estanho (PT08.S1), concentração média total de hidrocarbonetos não metânicos ($microg/m^3$), concentração média de benzeno ($microg/m^3$), resposta média do sensor relacionada a titânia (PT08.S2), concentração média de NOx (ppb), resposta média do sensor relacionada ao óxido de tungstênio NOx (PT08.S3), concentração média de NO^2 ($microg/m^3$), resposta média do sensor relacionada ao óxido de tungstênio NO^2 (PT08.S4), resposta média do sensor relacionada ao óxido de índio (PT08.S5), temperatura (graus), umidade relativa (%). O objetivo é realizar a previsão da umidade absoluta. Este *benchmark*

[3] http://archive.ics.uci.edu/ml/datasets/Air+Quality

Figura 17 – Boxplot dos MDPs na previsão da coordenada atômica Calculada w' nos cenários MAR.



foi utilizado, por exemplo, nos trabalhos de Vito et al. (2008), Vito et al. (2009) e Vito et al. (2012).

O eNFN-MDP e o ALMMo-MDP foram avaliados na previsão umidade absoluta em comparação com as técnicas de imputação pela última saída, pela média dos últimos 5 valores e por zero nas Seções 4.4.1 e 4.4.2, respectivamente, e comparados com o modelo eFGP na Seção 4.4.3.

4.4.1 eNFN-MDP na Previsão da Umidade Absoluta

Para os experimentais relativos aos cenários MCAR, presentes na Tabela 13, verifica-se que o eNFN-MDP obteve o melhor resultado em todos os cenários e na média dos resultados. Somente a média do eNFN-MDP foi inferior ao resultado obtido para o conjunto original enquanto a imputação pela média obteve uma proximidade de duas casas decimais.

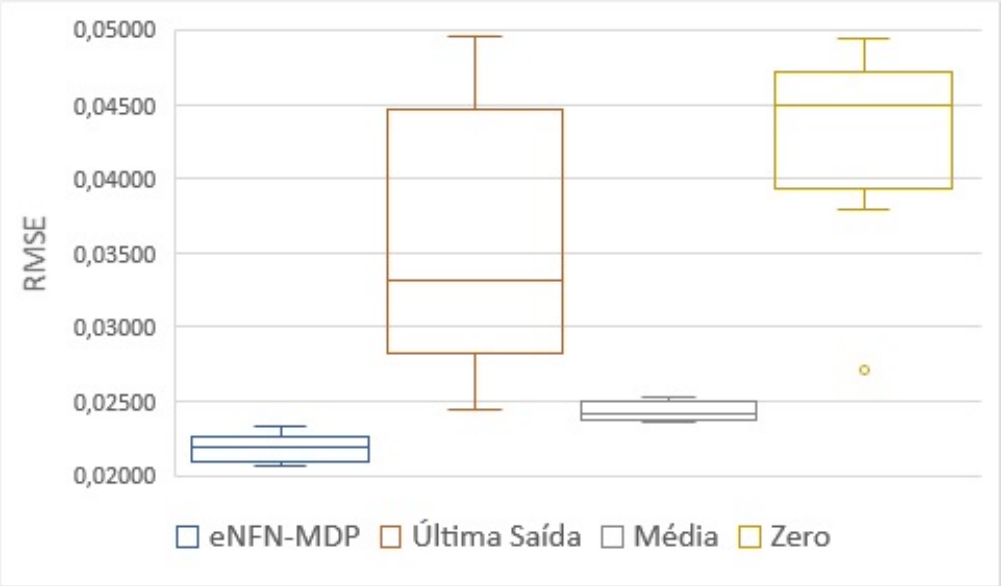
Tabela 13 – Desempenho RMSE do eNFN-MDP na previsão da umidade absoluta nos cenários MCAR.

Ausência	Última Saída	Média	Zero	eNFN-MDP
0%	0,02223			
1%	0,02445	0,02373	0,02709	0,02065
5%	0,02712	0,02402	0,03800	0,02061
10%	0,03147	0,02356	0,04324	0,02179
15%	0,03475	0,02422	0,04668	0,02210
20%	0,04795	0,02521	0,04741	0,02282
30%	0,04958	0,02532	0,04942	0,02327
Média	0,03589	0,02434	0,04197	0,02187

Na Figura 18 são apresentados em um *boxplot* os resultados experimentais dos cenários MCAR para o eNFN-MDP na previsão da umidade absoluta. Pelos resultados, é

possível definir que o eNFN-MDP é estatisticamente superiores aos resultados das demais técnicas analisadas.

Figura 18 – Boxplot do eNFN-MDP na previsão da umidade absoluta nos cenários MCAR.



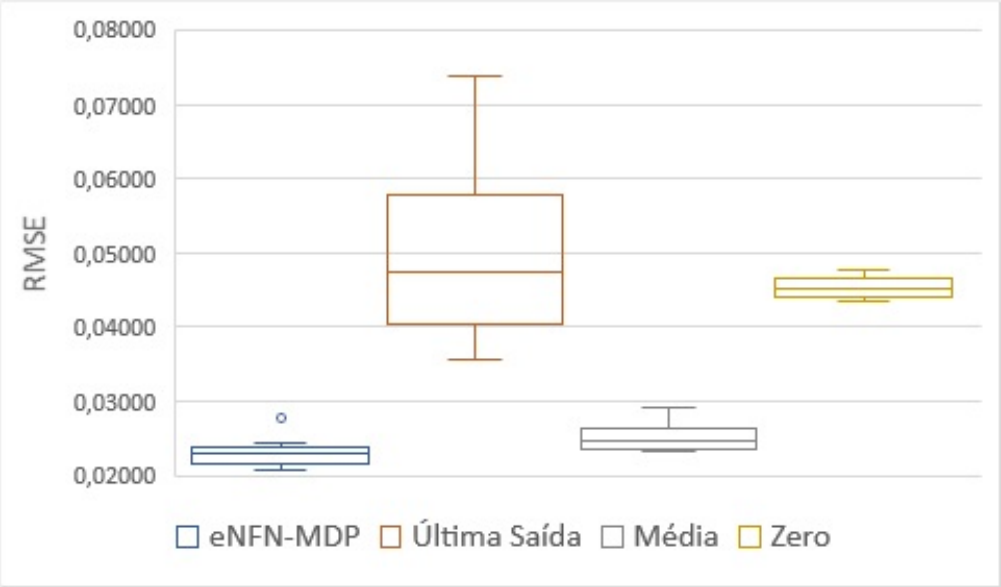
A Tabela 14, que resume os resultados para os cenários MAR, mostra que, análogo ao ocorrido nos cenários MCAR, o eNFN-MDP obteve o melhor resultado em todos os cenários e na média, seguido pela imputação pela média. Analisando o resultado do conjunto original, a média dos resultados do eNFN-MDP e da imputação pela média obtiveram duas casas decimais de proximidade em relação ao resultado do conjunto original.

Tabela 14 – Desempenho RMSE do eNFN-MDP na previsão da umidade absoluta nos cenários MAR.

Ausência	Última Saída	Média	Zero	eNFN-MDP
0x0%		0,02223		
5x1%	0,04196	0,02352	0,04523	0,02213
10x1%	0,03571	0,02477	0,04458	0,02295
10x5%	0,03887	0,02816	0,04747	0,02424
20x5%	0,04740	0,02462	0,04545	0,02313
20x10%	0,05421	0,02351	0,04359	0,02083
30x5%	0,07381	0,02910	0,04764	0,02768
30x10%	0,06162	0,02324	0,04363	0,02092
Média	0,05051	0,02527	0,04537	0,02312

O gráfico *boxplot* relativo aos resultados dos cenários MAR do eNFN-MDP na previsão da umidade absoluta é apresentado na Figura 19. O gráfico confirma a superioridade estatística do eNFN-MDP em relação aos resultados da imputação pela última saída e por zero. O teste T entre o eNFN-MDP e a imputação pela média obteve o valor de $0{,}00$, indicando que o eNFN-MDP é estatiscamente superior.

Figura 19 – Boxplot do eNFN-MDP na previsão da umidade absoluta nos cenários MAR.



4.4.2 ALMMo-MDP na Previsão da Umidade Absoluta

Os resultados apresentados na Tabela 15, referente aos cenários MCAR, mostram que o ALMMo-MDP obteve o melhor resultado em 5 dos 6 cenários e na média dos resultados. No cenário com 15% de ausência a imputação pela média obteve o melhor resultado, seguido pelo ALMMo-MDP. Comparando com o conjunto original, a média dos resultados do ALMMo-MDP obteve três casas de proximidade com o resultado do conjunto original, enquanto a média dos resultados da imputação pela média obteve duas casas de proximidade.

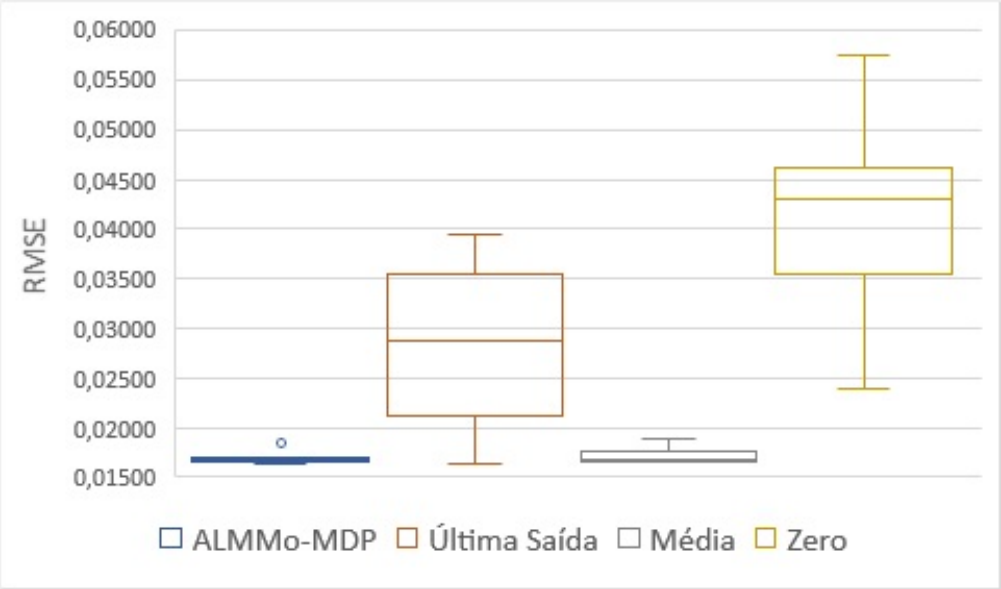
Tabela 15 – Desempenho RMSE do ALMMo-MDP na previsão da umidade absoluta nos cenários MCAR.

Ausência	Última Saída	Média	Zero	ALMMo-MDP
0%	0,01632			
1%	0,01643	0,01658	0,02391	0,01642
5%	0,01936	0,01647	0,03321	0,01636
10%	0,02627	0,01703	0,04194	0,01689
15%	0,03129	0,01663	0,04409	0,01671
20%	0,03939	0,01771	0,04678	0,01681
30%	0,03674	0,01891	0,05748	0,01845
Média	0,02825	0,01722	0,04123	0,01694

A Figura 20 ilustra o *boxplot* dos resultados experimentais para o ALMMo-MDP nos cenários MCAR na previsão da umidade absoluta. Verifica-se pelos resultados que há uma superioridade estatística entre os resultados do ALMMo-MDP e da imputação pela média em relação aos demais. Comparando o ALMMo-MDP e a imputação pela média, o teste T obteve o resultado de $0{,}05$, indicando que o ALMMo-MDP é estatisticamente superior.

Em relação aos resultados dos cenários MAR, apresentados na Tabela 16, o ALMMo-MDP obteve o melhor resultado para todas as taxas de ausência e na média dos resultados,

Figura 20 – Boxplot do ALMMo-MDP na previsão da umidade absoluta nos cenários MCAR.



consequentemente. Em relação aos resultados do conjunto original, a médias dos resultados do ALMMo-MDP e da imputação pela média obtiveram duas casa decimais de proximidade com o resultado do conjunto original.

Tabela 16 – Desempenho RMSE do ALMMo-MDP na previsão da umidade absoluta nos cenários MAR.

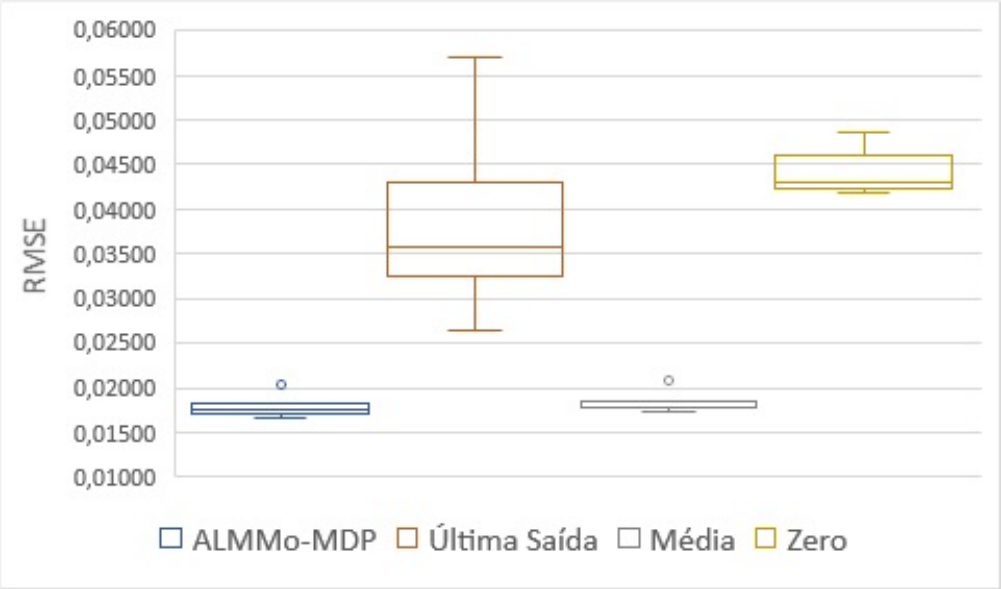
Ausência	Última Saída	Média	Zero	ALMMo-MDP
0x0%		0,01632		
5x1%	0,03567	0,01791	0,04189	0,01744
10x1%	0,02647	0,01847	0,04172	0,01826
10x5%	0,03018	0,01855	0,04794	0,01743
20x5%	0,03493	0,01856	0,04402	0,01815
20x10%	0,04492	0,01723	0,04305	0,01655
30x5%	0,05710	0,02090	0,04870	0,02045
30x10%	0,04109	0,01745	0,04251	0,01683
Média	0,03862	0,01844	0,04426	0,01787

O *boxplot* com os resultados experimentais na previsão da umidade nos cenários MAR para o ALMMo-MDP é apresentado na Figura 21. Verifica-se pelo gráfico que o ALMMo-MDP foi estatisticamente superior a imputação pela última saída e por zero. Em relação a comparação com a imputação pela média, onde houve sobreposição dos resultados, o teste T com o ALMMo-MDP obteve $0{,}00$, indicando que o ALMMo-MDP é estatisticamente superior.

4.4.3 MDPs na Previsão da Umidade Absoluta

A comparação dos resultados nos cenários MCAR do eNFN-MDP, ALMMo-MDP e do eFGP, apresentada na Tabela 17, mostra que para todos os cenários analisados e na média dos resultados o ALMMo-MDP obteve o melhor resultado, seguido do eNFN-MDP.

Figura 21 – Boxplot do ALMMo-MDP na previsão da umidade absoluta nos cenários MAR.



Comparando com o conjunto original, a média dos resultados do ALMMo-MDP obteve três casas decimais de proximidade com o resultado do conjunto original enquanto a média dos resultados do eNFN-MDP obteve duas casa e do eFGP obtive uma casa de proximidade.

Tabela 17 – Desempenho RMSE dos MDPs na previsão da umidade absoluta nos cenários MCAR.

Ausência	eNFN-MDP	ALMMo-MDP	eFGP
0%	0,02223	0,01632	0,03743
1%	0,02065	0,01642	0,03696
5%	0,02061	0,01636	0,03755
10%	0,02179	0,01689	0,03870
15%	0,02210	0,01671	0,03973
20%	0,02282	0,01681	0,04427
30%	0,02327	0,01845	0,04609
Média	0,02187	0,01694	0,04055

O gráfico *boxplot* presente na Figura 22 apresenta os resultados obtidos nos cenários MCAR para os modelos MDPs na previsão da umidade absoluta em comparação com o eFGP. Pelo gráfico, os resultados do ALMMo-MDP foram estatisticamente superiores aos demais modelos, seguido pelo eNFN-MDP.

Os resultados do eNFN-MDP, ALMMo-MDP e eFGP nos cenários MAR, dispostos na Tabela 18, mostram que, análogo ao ocorrido nos cenários MCAR, o modelo ALMMo-MDP obteve o melhor resultado em todos os cenário e na média dos resultados, seguido pelo eNFN-MDP. Na comparação com o conjunto original, a média dos resultados do ALMMo-MDP e do eNFN-MDP obtiveram duas casas de proximidade em comparação ao resultado do conjunto original enquanto a média dos resultados do eFGP obteve uma casa de proximidade.

O *boxplot* na Figura 23 é formado com os resultados obtidos nos cenários MAR na

Figura 22 – Boxplot dos MDPs na previsão da umidade absoluta nos cenários MCAR.

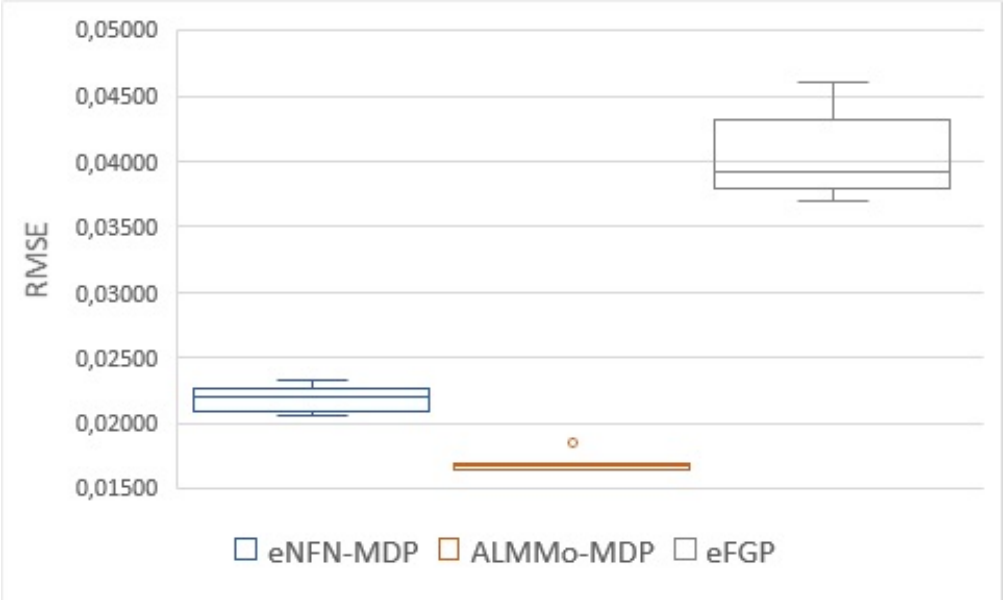


Tabela 18 – Desempenhos RMSE do MDPs na previsão da umidade absoluta nos cenários MAR.

Ausência	eNFN-MDP	ALMMo-MDP	eFGP
0x0%	0,02223	0,01632	0,03743
5x1%	0,02213	0,01744	0,03839
10x1%	0,02295	0,01826	0,03895
10x5%	0,02424	0,01743	0,04195
20x5%	0,02313	0,01815	0,03699
20x10%	0,02083	0,01655	0,04714
30x5%	0,02768	0,02045	0,04402
30x10%	0,02092	0,01683	0,04830
Média	0,02312	0,01787	0,03699

previsão da umidade absoluta para os modelos MDPs juntamente com os resultados do eFGP. A comparação entre os modelos mostra que há uma superioridade estatística dos resultados do ALMMo-MDP em relação aos demais modelos, seguido pelos resultados do eNFN-MDP.

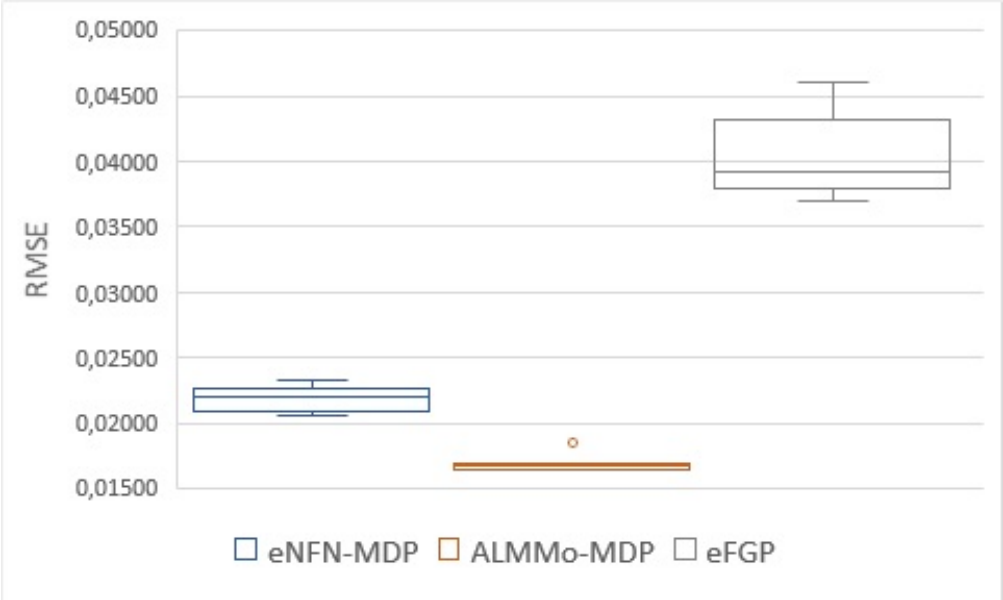
4.5 Identificação de Processo Não Linear

Nesta seção os algoritmos são avaliados na identificação de processo não linear. Trata-se de um *benchmark* clássico. Este processo não linear a ser identificado é definido por:

$$y_t = \frac{y_t y_{t-2}(y_{t-1} - 0{,}5)}{1 + (y_{t-1})^2 + (y_{n-2})^2} + u_{t-1}, \tag{58}$$

onde $u_t = \sin(\frac{2\pi t}{25}))$, e $y_0 = y_1 = 0$. O modelo possui 5200 amostras e é descrito por $\hat{y}_t = f(y_{t-1}, y_{t-2}, u_{t-1})$. O objetivo é prever a saída, $\hat{y}_t$ com base nas informações passadas

Figura 23 – Boxplot dos MDPs na previsão da umidade absoluta nos cenários MAR.



e nas saídas. Este conjunto de dados também foi utilizado por Silva (2014).

A avaliação do eNFN-MDP e o ALMMo-MDP na identificação de processo não linear em comparação foi feita juntamente com as técnicas de imputação pela última saída, pela média dos últimos 5 valores e por zero nas Seções 4.5.1 e 4.5.2, respectivamente, e comparados com o eFGP na Seção 4.5.3.

4.5.1 eNFN-MDP na Identificação de Processo Não Linear

Na Tabela 19, onde são apresentados os resultados obtidos para os cenários MCAR, mostra que o eNFN-MDP obteve o melhor resultado para todas as taxas de ausência e na média dos resultados. Em relação o conjunto original, a média dos resultados do eNFN-MDP e das técnicas obtiveram uma casa decimal de proximidade com o resultado do conjunto original.

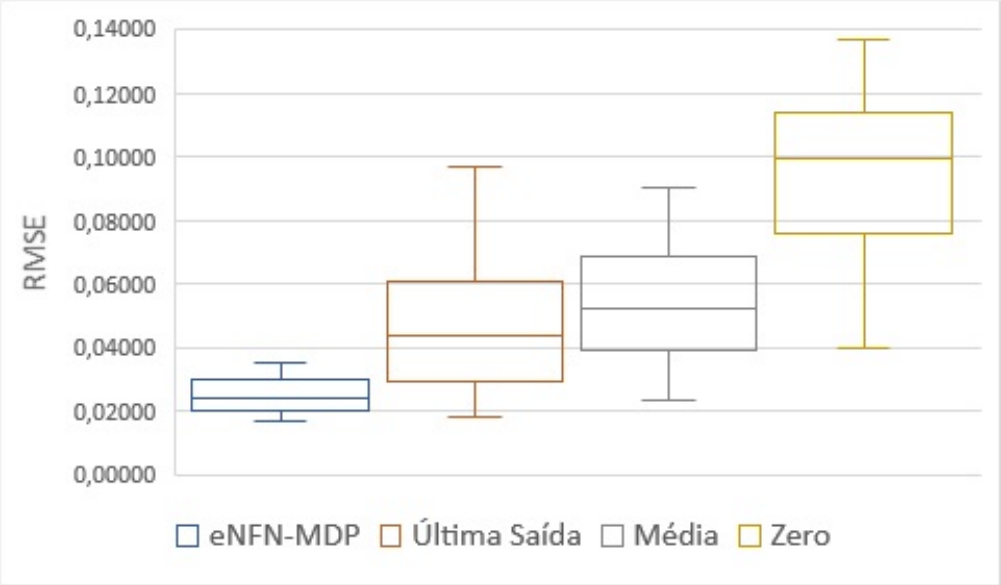
Tabela 19 – Desempenho RMSE do eNFN-MDP na identificação de processo não linear nos cenários MCAR.

Ausência	**Última Saída**	**Média**	**Zero**	**eNFN-MDP**
0%	0,01489			
1%	0,01783	0,02313	0,03998	0,01680
5%	0,02632	0,03640	0,06963	0,01904
10%	0,03709	0,04676	0,09515	0,02192
15%	0,05009	0,05807	0,10383	0,02577
20%	0,06448	0,07172	0,11757	0,03132
30%	0,09663	0,09034	0,13706	0,03500
Média	0,04874	0,05440	0,09387	0,02498

A Figura 24 ilustra o *boxplot* com os resultados experimentais para o eNFN-MDP na identificação do processo não linear nos cenários MCAR. O gráfico mostra a existência

de uma superioridade estatística dos resultados do eNFN-MDP em comparação com a imputação pela média e por zero. Em comparação entre o eNFN-MDP e a imputação pela última saída, o teste T obteve $0{,}02$, indicando que o eNFN-MDP é estatisticamente superior.

Figura 24 – Boxplot do eNFN-MDP na identificação de processo não linear nos cenários MCAR.



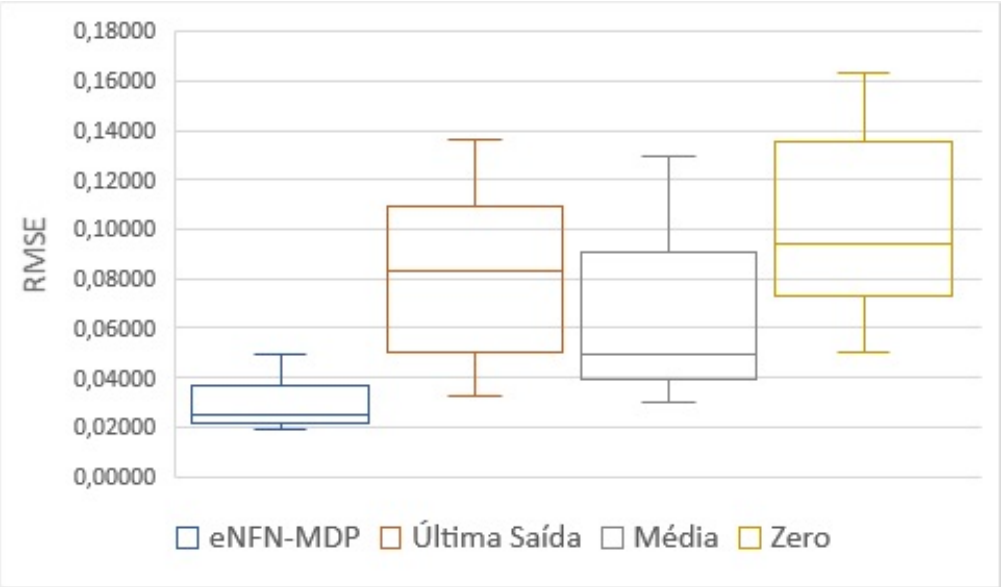
Para os resultados referentes aos cenários MAR, dispostos na Tabela 20, observa-se que, análogo aos cenários MCAR, o eNFN-MDP obteve o melhor resultado para todas as taxas de ausência e na média dos resultados. Em comparação com o conjunto original, a média dos resultados do eNFN-MDP obteve uma casa decimal de proximidade do resultado do conjunto original, assim como a imputação pela média e pela última saída.

Tabela 20 – Desempenho RMSE do eNFN-MDP na identificação de processo não linear nos cenários MAR.

Ausência	Última Saída	Média	Zero	eNFN-MDP
0x0%		0,01489		
5x1%	0,03228	0,02963	0,05060	0,01881
10x1%	0,04256	0,02997	0,05387	0,01904
10x5%	0,05724	0,07590	0,11511	0,03192
20x5%	0,08354	0,04908	0,09152	0,02379
20x10%	0,09138	0,10568	0,15498	0,04172
30x5%	0,12693	0,04892	0,09386	0,02498
30x10%	0,13585	0,12914	0,16324	0,04974
Média	0,08140	0,06691	0,10331	0,03000

O *boxplot* para o eNFN-MDP nos cenários MAR na identificação de processo não linear é apresentado na Figura 25. Neste cenário, como não houve sobreposição dos resultados, é comprovada a superioridade estatística do eNFN-MDP em relação a imputação pela última saída, pela média e por zero.

Figura 25 – Boxplot do eNFN-MDP na identificação de processo não linear nos cenários MAR.



4.5.2 ALMMo-MDP na Identificação de Processo Não Linear

Os resultados dispostos na Tabela 21, referentes aos cenários MCAR, mostram que o ALMMo-MDP obteve o melhor resultado para todas as taxas de ausência e na média dos resultados. Comparando com o resultado do conjunto original, a média do ALMMo-MDP obteve duas casas decimais de proximidade enquanto os demais obtiveram uma casa decimal de proximidade.

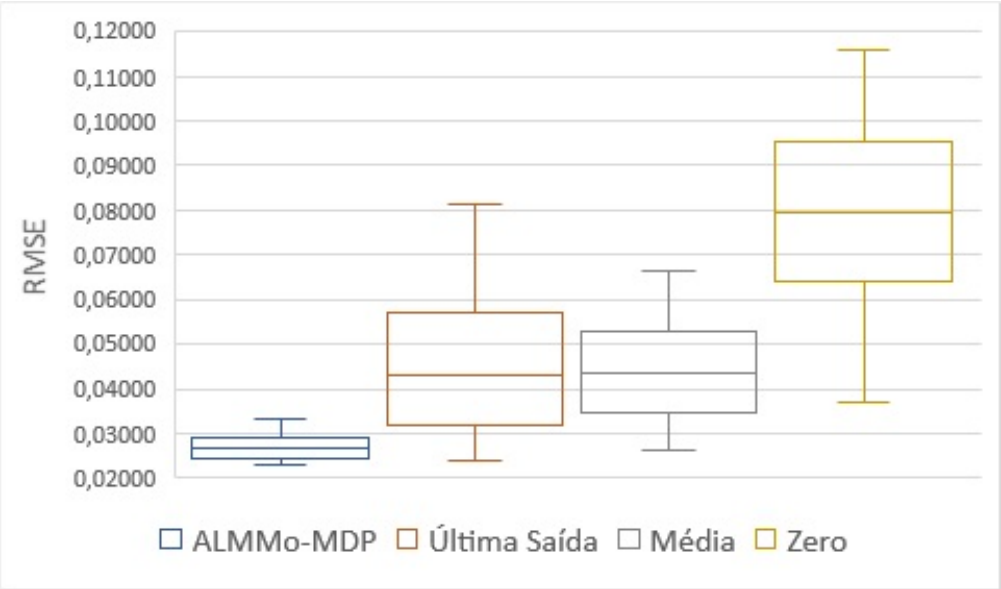
Tabela 21 – Desempenho RMSE do ALMMo-MDP na identificação de processo não linear nos cenários MCAR.

Ausência	Última Saída	Média	Zero	ALMMo-MDP
0%		0,02193		
1%	0,02393	0,02613	0,03701	0,02278
5%	0,03006	0,03335	0,06037	0,02403
10%	0,03811	0,03917	0,07549	0,02564
15%	0,04813	0,04761	0,08363	0,02805
20%	0,05976	0,05446	0,09892	0,02933
30%	0,08130	0,06634	0,11574	0,03347
Média	0,04688	0,04451	0,07853	0,02722

Na Figura 26 é apresentado o *boxplot* com os resultados experimentais para o ALMMo-MDP na identificação do processo não linear nos cenários MCAR. Como não houve sobreposição dos resultados, é possível afirmar que o ALMMo-MDP é estatisticamente superior em comparação as outras técnicas analisadas.

Analisando os resultados apresentados na Tabela 22, referentes aos cenários MAR, verifica-se que, análogo ao ocorrido para os cenários MCAR, o ALMMo-MDP obteve o melhor resultado para todas as taxas de ausência e na média dos resultados. Em relação ao conjunto original, a média dos resultados do ALMMo-MDP obteve uma casa decimal de

Figura 26 – Boxplot do ALMMo-MDP na identificação de processo não linear nos cenários MCAR.



proximidade do resultado dela enquanto a média dos demais obtiveram a proximidade de uma casa decimal.

Tabela 22 – Desempenho RMSE do ALMMo-MDP na identificação de processo não linear nos cenários MAR.

Ausência	Última Saída	Média	Zero	ALMMo-MDP
0x0%		0,02193		
5x1%	0,03687	0,02912	0,04413	0,02328
10x1%	0,04208	0,02895	0,04888	0,02361
10x5%	0,05284	0,05720	0,10085	0,02971
20x5%	0,07245	0,04025	0,07526	0,02629
20x10%	0,07814	0,07409	0,11953	0,03440
30x5%	0,10248	0,04112	0,07544	0,02669
30x10%	0,10494	0,08968	0,12563	0,04188
Média	0,06997	0,05149	0,08425	0,02941

A Figura 27 ilustra o *boxplot* referente aos resultados experimentais para o ALMMo-MDP na identificação de processo não linear nos cenários MAR. Os resultados mostram que o ALMMo-MDP é superior estatisticamente a imputação pela última saída, pela média e por zero.

4.5.3 MDPs na Identificação de Processo Não Linear

A Tabela 23, referente aos resultados dos cenários MCAR, mostra que o eNFN-MDP obteve o melhor resultado em 4 dos 6 cenários e na média dos resultados, enquanto o ALMMo-MDP obteve o melhor resultado nos outros dois. A média dos resultados do ALMMo-MDP obteve duas casas decimais de proximidade do resultado do conjunto original enquanto a média dos resultado do eNFN-MDP e do eFGP obtiveram a proximidade de uma casa decimal.

Figura 27 – Boxplot do ALMMo-MDP na identificação de processo não linear nos cenários MAR.

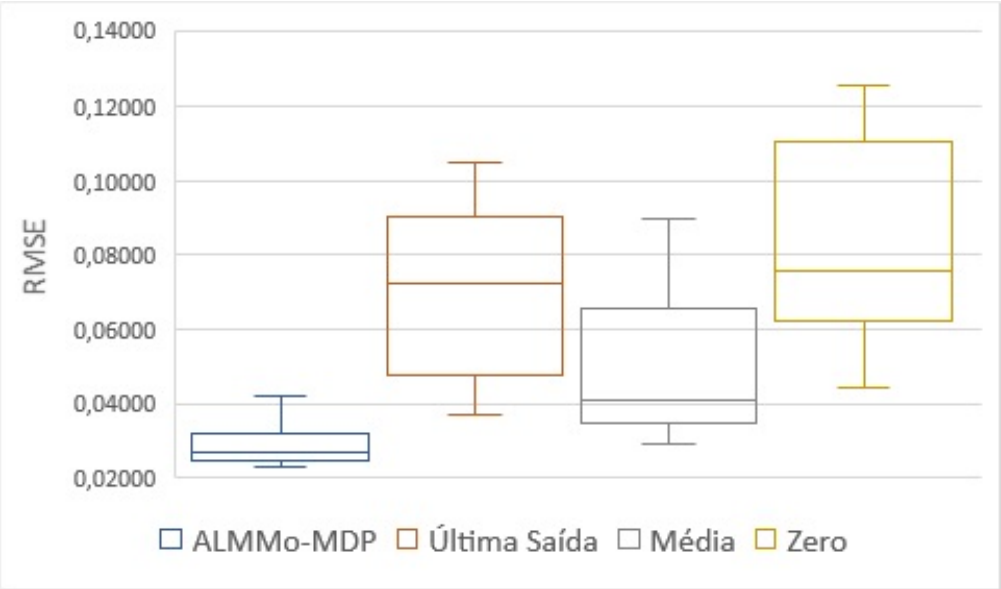


Tabela 23 – Desempenho RMSE dos MDPs na identificação de processo não linear nos cenários MCAR.

Ausência	eNFN-MDP	ALMMo-MDP	eFGP
0%	0,01489	0,02193	0,01999
1%	0,01680	0,02278	0,03192
5%	0,01904	0,02403	0,04203
10%	0,02192	0,02564	0,04758
15%	0,02577	0,02805	0,03792
20%	0,03132	0,02933	0,04894
30%	0,03500	0,03347	0,04525
Média	0,02498	0,02722	0,04227

O *boxplot* da Figura 28 mostra os resultados experimentais dos MDPs em comparação com os resultados do eFGP na identificação do processo não linear para os cenários MCAR. O gráfico indica que o eNFN-MDP e o ALMMo-MDP foram estatisticamente superiores ao eFGP. A sobreposição entre os resultados deles e o resultado do teste T, que obteve o valor de $0{,}08$, indicam não haver uma superioridade estatística entre eles. Com relação aos demais índices, o eNFN-MDP obteve a menor mediana e o ALMMo-MDP obteve a menor diferença entre os limites inferior e superior.

Para os resultados dos cenários MAR, apresentados na Tabela 24, o ALMMo-MDP obteve o melhor resultados em três cenários e na média, enquanto o eNFN-MDP obteve o melhor resultado em dois cenários e o eFGP em um cenário. Com relação ao conjunto original, o ALMMo-MDP obteve duas casas decimais de proximidade do resultado do conjunto original enquanto a média dos resultados do eNFN-MDP e do eFGP obtiveram uma casa decimal de proximidade.

O *boxplot* apresentado na Figura 29 apresenta de forma gráfica os resultados obtidos pelos MDPs juntamente com os resultados do eFGP na identificação de processos não

Figura 28 – Boxplot dos MDPs na identificação de processo não linear nos cenários MCAR.

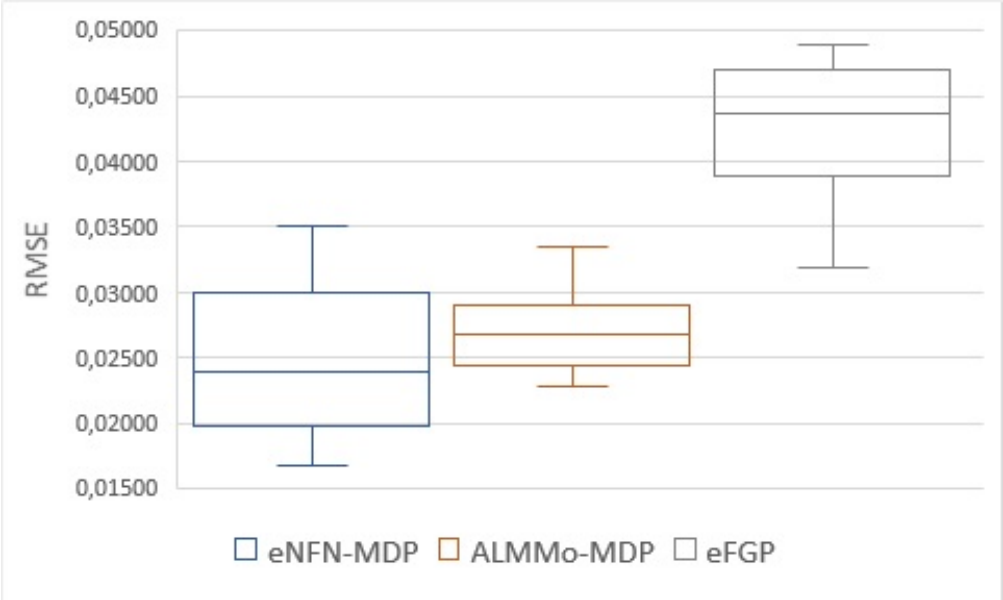
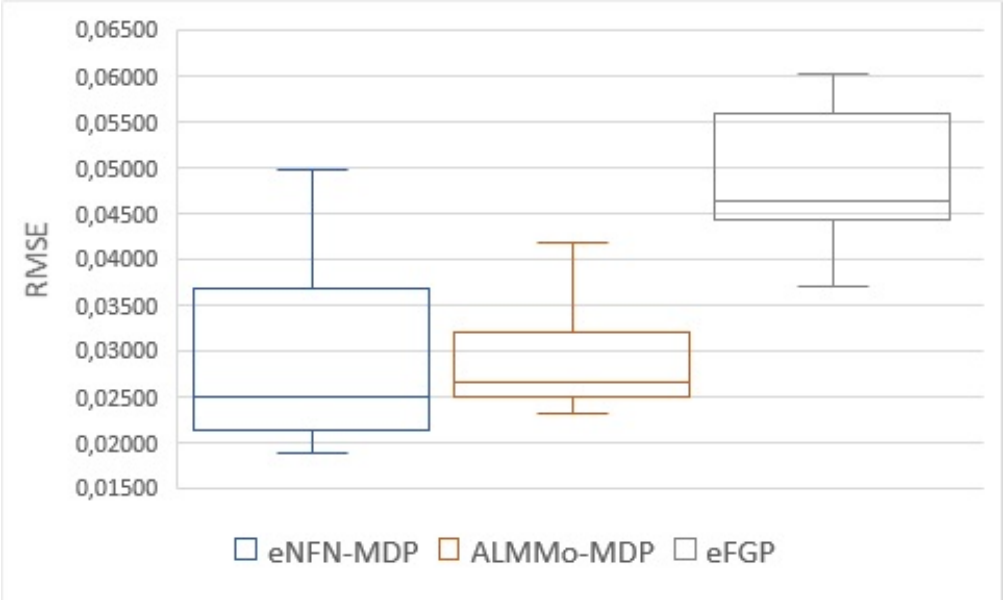


Tabela 24 – Desempenho RMSE dos MDPs na identificação de processo não linear nos cenários MAR.

Ausência	eNFN-MDP	ALMMo-MDP	eFGP
0x0%	0,01489	0,02193	0,01999
5x1%	0,01881	0,02328	0,03698
10x1%	0,01904	0,02361	0,04302
10x5%	0,03192	0,02971	0,04648
20x5%	0,02379	0,02629	0,04580
20x10%	0,04172	0,03440	0,05286
30x5%	0,02498	0,02669	0,06018
30x10%	0,04974	0,04188	0,05902
Média	0,03000	0,02941	0,03698

lineares nos cenários MAR. O gráfico mostra que o eNFN-MDP e o ALMMo-MDO foram estatisticamente superiores ao eFGP. O teste T entre eles gerou o valor $0{,}39$, indicando que não houve superioridade estatística entre eles. Considerando so demais índices, o eNFN-MDP obteve a menor mediana, enquanto o ALMMo-MDP obteve a menor diferença entre os limites superior e inferior.

Figura 29 – Boxplot dos MDPs na identificação de processo não linear nos cenários MAR.



5 Considerações Finais

Este trabalho introduziu uma nova abordagem denominada *Missing Data Procedure* (MDP). O MDP possibilita que modelos evolutivos realizem o cálculo das saídas e estimem os valores para as variáveis ausentes, simultaneamente. A abordagem foi implementada em dois modelos evolutivos, que foram chamados de eNFN-MDP e ALMMo-MDP. Apesar de ser desenvolvido para aplicações envolvendo condições ambientais voláteis os modelos também podem ser empregados em ambientes com dados estáticos.

Considerando o objetivo geral definido na Seção 1.2, tem-se a abordagem proposta descrita no Capítulo 3. O primeiro objetivo específico, referente a revisão da literatura, foi realizado ao longo do Capítulo 2. Para o segundo objetivo específico, referente a proposta e implementação de algoritmos para gerar bases com dados ausentes, foram propostos e implementados os algoritmos descritos na Seção 4.1.1. A implementação de um método de imputação para modelos evolutivos com múltiplas entradas e múltiplas saídas, terceiro objetivo específico, foi alcançado na Seção 3.2. O quarto objetivo específico referente a implementação de um método de imputação para modelos evolutivos com múltiplas entradas e saída única foi introduzido na Seção 3.3. Por fim, a validação dos modelos propostos em problemas de previsão e identificação de sistemas dinâmicos não lineares, último objetivo específico, foi abordada no Capítulo 4.

Um conjunto extensivo de experimentos computacionais foram realizados para avaliar o eNFN-MDP e o ALMMo-MDP em cenários com ausência completamente aleatória (MCAR) e em cenários com ausência aleatória (MAR). Inicialmente os resultados computacionais dos modelos propostos foram comparados com 3 técnicas alternativas de imputação: a imputação pela última saída; a imputação pela média dos últimos 5 valores e; a imputação por zero. Posteriormente os resultados do eNFN-MDP e do ALMMo-MDP foram comparados entre si e com um modelo específico para tratar dados ausentes chamado de eFGP. Os resultados experimentais mostram que a abordagem proposta é promissora. Destaca-se que os resultados obtidos pelos modelos propostos se mantiveram estáveis e pouco dispersos com relação ao percentual de dados ausentes e próximos aos resultados obtidos com a base original, sugerindo a sua eficiência.

Para trabalhos futuros, é interessante abordar um estudo do custo computacional dos modelos eNFN-MDP e do ALMMo-MDP, bem como suas aplicações em sistemas com dinâmica rápida. Destaca-se também a possibilidade de aplicar a abordagem MDP em outros modelos evolutivos. Outra abordagem a ser considerada é a implementação de um método que considere o valor modal da função mais ativa para a variável ausente como um método de imputação.

Referências

ACI, M.; AVCI, M. Artificial neural network approach for atomic coordinate prediction of carbon nanotubes. **Applied Physics A**, Springer, v. 122, n. 7, p. 631, 2016. Citado na página 42.

AL., M. A. E. Reducing simulation duration of carbon nanotube using support vector regression method. **JOURNAL OF THE FACULTY OF ENGINEERING AND ARCHITECTURE OF GAZI UNIVERSITY**, p. 901–907, 2017. Citado na página 42.

AMIRI, M.; JENSEN, R. Missing data imputation using fuzzy-rough methods. **Neurocomputing**, Elsevier, v. 205, p. 152–164, 2016. Citado 7 vezes nas páginas 1, 2, 6, 7, 8, 15 e 16.

ANGELOV, P. **Autonomous learning systems: from data streams to knowledge in real-time**. [S.l.]: John Wiley & Sons, 2012. Citado na página 28.

ANGELOV, P.; FILEV, D. Simpl_ets: A simplified method for learning evolving takagi-sugeno fuzzy models. In: IEEE. **The 14th IEEE International Conference on Fuzzy Systems, 2005. FUZZ'05.** [S.l.], 2005. p. 1068–1073. Citado na página 23.

ANGELOV, P.; YAGER, R. A new type of simplified fuzzy rule-based system. **International Journal of General Systems**, Taylor & Francis, v. 41, n. 2, p. 163–185, 2012. Citado na página 28.

ANGELOV, P. P.; GU, X.; PRÍNCIPE, J. C. Autonomous learning multimodel systems from data streams. **IEEE Transactions on Fuzzy Systems**, IEEE, v. 26, n. 4, p. 2213–2224, 2017. Citado 3 vezes nas páginas 3, 4 e 28.

BROOKS, T. F.; POPE, D. S.; MARCOLINI, M. A. Airfoil self-noise and prediction. **NASA Scientific and Technical Information (STI)**, NASA Technical Reports Server, p. 1–146, 1989. Citado na página 35.

CAMINHAS, W.; GOMIDE, F. A fast learning algorithm for neofuzzy networks. In: **Proc. Information Processing and Management of Uncertainty in Knowledge Based Systems**. [S.l.: s.n.], 2000. v. 1, n. 1, p. 1784–1790. Citado 2 vezes nas páginas 24 e 25.

COOKE, M.; MORRIS, A.; GREEN, P. Missing data techniques for robust speech recognition. In: IEEE. **1997 IEEE International Conference on Acoustics, Speech, and Signal Processing**. [S.l.], 1997. v. 2, p. 863–866. Citado na página 1.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 39, n. 1, p. 1–22, 1977. Citado 2 vezes nas páginas 8 e 12.

DUA, D.; GRAFF, C. **UCI Machine Learning Repository**. 2017. Disponível em: <http://archive.ics.uci.edu/ml>. Citado 3 vezes nas páginas 35, 42 e 47.

ENDERS, C. K.; BARALDI, A. N. Missing data handling methods. **The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development**, Wiley Online Library, p. 139–185, 2018. Citado 2 vezes nas páginas 1 e 6.

FARHANGFAR, A.; KURGAN, L. A.; PEDRYCZ, W. A novel framework for imputation of missing values in databases. **IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans**, IEEE, v. 37, n. 5, p. 692–709, 2007. Citado na página 1.

GARCIA, C. et al. Evolvable fuzzy systems from data streams with missing values: With application to temporal pattern recognition and cryptocurrency prediction. **Pattern Recognition Letters**, Elsevier, v. 128, p. 278–282, 2019. Citado 7 vezes nas páginas 1, 2, 4, 6, 8, 16 e 32.

GARCIA, C.; LEITE, D.; SKRJANC, I. Incremental missing-data imputation for evolving fuzzy granular prediction. **IEEE Transactions on Fuzzy Systems**, IEEE, 2019. Citado 7 vezes nas páginas 1, 2, 4, 5, 8, 16 e 35.

GELMAN, A.; HILL, J. **Data analysis using regression and multilevel/hierarchical models**. [S.l.]: Cambridge university press, 2006. 529–543 p. Citado 3 vezes nas páginas 1, 7 e 8.

GRZYMALA-BUSSE, J. W. et al. Handling missing attribute values in preterm birth data sets. In: SPRINGER. **International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing**. [S.l.], 2005. p. 342–351. Citado 2 vezes nas páginas 8 e 14.

JENSEN, R.; CORNELIS, C. Fuzzy-rough nearest neighbour classification. In: **Transactions on rough sets XIII**. [S.l.]: Springer, 2011. p. 56–72. Citado na página 16.

KANG, H. The prevention and handling of the missing data. **Korean journal of anesthesiology**, Korean Society of Anesthesiologists, v. 64, n. 5, p. 402, 2013. Citado 6 vezes nas páginas 1, 2, 5, 6, 7 e 8.

KIM, H.; GOLUB, G. H.; PARK, H. Missing value estimation for dna microarray gene expression data: local least squares imputation. **Bioinformatics**, Oxford University Press, v. 21, n. 2, p. 187–198, 2004. Citado 2 vezes nas páginas 8 e 13.

KRAUSE, R. W. et al. Missing network data a comparison of different imputation methods. In: IEEE. **2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)**. [S.l.], 2018. p. 159–163. Citado 2 vezes nas páginas 5 e 6.

LAU, K. et al. A neural networks approach for aerofoil noise prediction. **Scipedia**, International Centre for Numerical Methods in Engineering (CIMNE), 2009. Citado na página 35.

LI, D. et al. Towards missing data imputation: a study of fuzzy k-means clustering method. In: SPRINGER. **International Conference on Rough Sets and Current Trends in Computing**. [S.l.], 2004. p. 573–579. Citado 2 vezes nas páginas 8 e 13.

LIAO, Z. et al. Missing data imputation: a fuzzy k-means clustering algorithm over sliding window. In: IEEE. **2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery**. [S.l.], 2009. v. 3, p. 133–137. Citado na página 14.

LITTLE, T. D. et al. Missing data. **Developmental psychopathology**, Wiley Online Library, p. 1–37, 2016. Citado na página 5.

LOPEZ, R.; BALSA-CANTO, E.; OÑATE, E. Neural networks for variational problems in engineering. **International Journal for Numerical Methods in Engineering**, Wiley Online Library, v. 75, n. 11, p. 1341–1360, 2008. Citado na página 35.

LUGHOFER, E.; ANGELOV, P. Handling drifts and shifts in on-line data streams with evolving fuzzy systems. **Appl. Soft Comput.**, v. 11, p. 2057–2068, 03 2011. Citado na página 2.

MYERS, T. A. Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data. **Communication methods and measures**, Taylor & Francis, v. 5, n. 4, p. 297–310, 2011. Citado na página 8.

OBA, S. et al. A bayesian missing value estimation method for gene expression profile data. **Bioinformatics**, Oxford University Press, v. 19, n. 16, p. 2088–2096, 2003. Citado 2 vezes nas páginas 8 e 12.

OSMAN, M. S.; ABU-MAHFOUZ, A. M.; PAGE, P. R. A survey on data imputation techniques: water distribution system as a use case. **IEEE Access**, IEEE, v. 6, p. 63279–63291, 2018. Citado 5 vezes nas páginas 2, 5, 6, 7 e 8.

PEDERSEN, A. B. et al. Missing data and multiple imputation in clinical epidemiological research. **Clinical epidemiology**, Dove Press, v. 9, p. 157, 2017. Citado 2 vezes nas páginas 5 e 6.

PRATAMA, I. et al. A review of missing values handling methods on time-series data. In: IEEE. **2016 International Conference on Information Technology Systems and Innovation (ICITSI)**. [S.l.], 2016. p. 1–6. Citado na página 7.

SANTOS, M. S. et al. Generating synthetic missing data: A review by missing mechanism. **IEEE Access**, IEEE, v. 7, p. 11651–11667, 2019. Citado na página 1.

SCHAFER, J. L.; GRAHAM, J. W. Missing data: our view of the state of the art. **Psychological methods**, American Psychological Association, v. 7, n. 2, p. 147, 2002. Citado 2 vezes nas páginas 2 e 6.

SCHNEIDER, T. Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. **Journal of climate**, v. 14, n. 5, p. 853–871, 2001. Citado 4 vezes nas páginas 7, 8, 9 e 10.

SILVA, A. M. et al. A fast learning algorithm for evolving neo-fuzzy neuron. **Applied Soft Computing**, Elsevier, v. 14, p. 194–209, 2014. Citado 6 vezes nas páginas 2, 3, 4, 21, 23 e 28.

SILVA, A. M. et al. Adaptive input selection and evolving neural fuzzy networks modeling. **International Journal of Computational Intelligence Systems**, v. 8, p. 3–14, 2015. ISSN 1875-6883. Citado 3 vezes nas páginas 25, 26 e 27.

SILVA, A. M. da. **Sistemas Neuro-Fuzzy Evolutivos: Novos Algoritmos de Aprendizado e Aplicações.** 2014. Tese (Doutorado em Engenharia Elétrica), Sistemas de Computação e Telecomunicações, Inteligência Co Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil. Citado 5 vezes nas páginas 2, 3, 4, 21 e 54.

TROYANSKAYA, O. et al. Missing value estimation methods for dna microarrays. **Bioinformatics**, Oxford University Press, v. 17, n. 6, p. 520–525, 2001. Citado 4 vezes nas páginas 8, 10, 11 e 12.

VITO, S. D. et al. Semi-supervised learning techniques in artificial olfaction: A novel approach to classification problems and drift counteraction. **IEEE Sensors Journal**, IEEE, v. 12, n. 11, p. 3215–3224, 2012. Citado na página 48.

VITO, S. D. et al. On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. **Sensors and Actuators B: Chemical**, Elsevier, v. 129, n. 2, p. 750–757, 2008. Citado na página 48.

VITO, S. D. et al. Co, no2 and nox urban pollution monitoring with on-field calibrated electronic nose by automatic bayesian regularization. **Sensors and Actuators B: Chemical**, Elsevier, v. 143, n. 1, p. 182–191, 2009. Citado na página 48.

ZHANG, Z. Missing data imputation: focusing on single imputation. **Annals of translational medicine**, AME Publications, v. 4, n. 1, 2016. Citado 2 vezes nas páginas 7 e 8.

ŠKRJANC, I. et al. Evolving fuzzy and neuro-fuzzy approaches in clustering, regression, identification, and classification: A survey. **Information Sciences**, v. 490, 03 2019. Citado na página 2.