UNIVERSIDADE FEDERAL DA BAHIA

DISSERTAÇÃO DE MESTRADO

PSGF (Phase Space Gap Filling): Um novo método para substituição de valores ausentes em séries temporais caóticas

Marcos Ricardo Santos Oliveira

Programa de Pós-Graduação em Ciência da Computação

Salvador
06 de Julho de 2022

MARCOS RICARDO SANTOS OLIVEIRA

PSGF (PHASE SPACE GAP FILLING): UM NOVO MÉTODO PARA SUBSTITUIÇÃO DE VALORES AUSENTES EM SÉRIES TEMPORAIS CAÓTICAS

Esta Dissertação de Mestrado foi apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal da Bahia, como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Orientador: Prof. Dr. RICARDO ARAÚJO RIOS

Salvador
06 de Julho de 2022

Ficha catalográfica elaborada pela Biblioteca Universitária de Ciências e Tecnologias Prof. Omar Catunda, SIBI – UFBA.

O48 Oliveira, Marcos Ricardo Santos
PSGF (Phase Space Gap Filling): um novo método para substituição de valores ausentes em séries temporais caóticas / Marcos Ricardo Santos Oliveira. – Salvador, 2022.
83 f.

Orientador: Prof. Dr. Ricardo Araújo Rios

1. Substituição de dados ausentes. 2. Séries temporais caóticas. 3. Aprendizado de máquina. 4. Teoria do caos. I. Rios, Ricardo Araújo. II. Universidade Federal da Bahia. III. Título.

CDU: 004.8

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DA BAHIA
INSTITUTO DE COMPUTAÇÃO
PGCOMP - Programa de Pós-Graduação em Ciência da Computação
http://pgcomp.ufba.br

"*PSGF (Phase Space Gap Filling): Um novo método para substituição de valores ausentes em séries temporais caóticas*"

Marcos Ricardo Santos Oliveira

Dissertação apresentada ao Colegiado do Programa de Pós-Graduação em Ciência da Computação na Universidade Federal da Bahia, como requisito parcial para obtenção do Título de Mestre em Ciência da Computação.

Banca Examinadora

Prof. Dr.Ricardo Araújo Rios (Orientador PGCOMP)

Prof. Dr. Renato Porfirio Ishii (UFMS)

Prof. Dr. Ewaldo Eder Carvalho Santana (UFMA)

RESUMO

O tratamento de informações ausentes ou inválidas em bases de dados representa um grande desafio na área de Aprendizado de Máquina (AM) que, se não for realizado da maneira adequada, pode afetar a qualidade do modelo produzido ou, até mesmo, impedir a sua utilização. Em geral, esse problema ocorre por diversas razões como, por exemplo, erro no dispositivo utilizado para coleta de informações, problemas na transmissão entre os dispositivos de coleta e de armazenamento, e a ausência real de informação no sistema monitorado. Quando os dados são coletados de maneira independente e identicamente distribuída, os próprios modelos tradicionais de AM podem ser utilizados para tratar esse problema. Entretanto, quando há dependência temporal entre as observações coletadas, e.g. quando os dados são organizados como séries temporais, tais modelos não são adequados por não considerar o relacionamento existente entre os instantes de tempo das coletas. Para o tratamento desse tipo de dado, há diversas técnicas como métodos de interpolação (e.g. Lagrange, Newton e *Splines*) e *Singular Spectrum Analysis* (SSA). Contudo, experimentos realizados durante este projeto de mestrado demonstraram que as técnicas existentes apresentam resultados insatisfatórios quando as séries temporais possuem comportamento caótico, uma vez que informações sobre seus atratores no espaço de coordenadas de atraso (espaço fase) não são levados em consideração. Neste sentido, este projeto de mestrado apresenta um novo método que utiliza ferramenta de Sistemas Dinâmicos e Teoria do Caos para desdobrar séries do domínio temporal para o espaço fase, viabilizando, assim, a aplicação de técnicas de Aprendizado de Máquina na substituição de valores ausentes. Resultados obtidos enfatizam a importância desse novo paradigma de substituição de valores ausentes, apresentando uma superioridade do método proposto com relação às técnicas conhecidas no estado da arte.

Palavras-chave: Substituição de dados ausentes, Séries Temporais Caóticas, Aprendizado de Máquina

ABSTRACT

The preprocessing step performed to deal with missing or invalid information in datasets is a relevant task in Machine Learning (ML) applications to avoid producing wrong models and make feasible the usage of specific algorithms that do not work in such a condition. In general, missing values occours for different reasons as, for instance, problems in the device used to monitor a system, network issues between monitoring and storage services, and the authentic absence of data. By collecting data in an i.i.d (independent and identically distributed) manner, traditional ML models are able to replace missing values. However, when there are temporal dependencies between collected observations, e.g., time series, such models are unsuitable for not considering the existing relationship in time instants. The treatment of missing data in time series is performed by several techniques such as interpolation methods (e.g. Lagrange, Newton, and Splines) and Singular Spectrum Analysis (SSA). Experiments during this project highlighted that these methods provided poor results when the time series present a chaotic behavior once their attractors in the phase space are not taken into account. Therefore, this work presents a new method that considers Dynamical System and Chaos Theory tools to unfold series from the temporal domain into phase space, making it possible the adoption of ML models to replace missing values. Our results emphasize the importance of this new paradigm to deal with missing values, outperforming the state-of-the-art.

Keywords: Missing Value Imputation, Chaotic Time Series, Machine Learning

SUMÁRIO

Capítulo 1—Introdução 1

1.1 Contexto . . . 1
1.2 Motivação . . . 5
1.3 Hipótese . . . 6
1.4 Objetivos . . . 6
1.5 Organização . . . 7

Capítulo 2—Referencial Teórico 9

2.1 Considerações Iniciais . . . 9
2.2 Decomposição de Séries Temporais . . . 9
2.2.1 Singular Spectrum Analysis - SSA . . . 10
2.2.2 Empirical Mode Decomposition - EMD . . . 11
2.3 Substituição de valores ausentes – Gap Filling . . . 13
2.3.1 Remoção dos Valores Ausentes . . . 15
2.3.2 Interpolação . . . 15
2.3.3 Imputação . . . 15
2.4 Sistemas Dinâmicos e Teoria do Caos . . . 15
2.4.1 Espaço Fase . . . 16
2.4.2 Reconstrução no Espaço Fase . . . 16
2.5 Considerações Finais . . . 17

Capítulo 3—Revisão Sistemática da Literatura 19

3.1 Considerações Iniciais . . . 19
3.2 Fase I - Busca e Coleta de Artigos . . . 19
3.3 Fase II - Análise dos Artigos Selecionados . . . 21
3.4 Fase III - Conclusão . . . 30
3.4.1 Redes Neurais Artificiais (RNA) . . . 31
3.4.2 Imputação . . . 31
3.4.3 Análise Espectral Singular . . . 32

Capítulo 4—Novos métodos para substituição de valores ausentes 33

4.1 Considerações Iniciais . . . 33
4.2 Bidirecional Mean Distance Estimation (BMDE) . . . 33
4.3 PSGF - Phase Space Gap Filling . . . 37

4.3.1 Processo de substituição 39
4.3.2 Limitação do método 41
4.4 Considerações Finais 42

Capítulo 5—Configuração Experimental 43

5.1 Considerações Iniciais 43
5.2 Séries no Espaço Tempo 43
5.3 Séries Caóticas (Espaço Fase) 44
5.3.1 Atrator de Lorenz 44
5.3.2 Atrator de Rössler 44
5.3.3 Atrator de Hénon 45
5.3.4 Mapa Logístico 47
5.4 Organização dos experimentos 48
5.5 Avaliação 49

Capítulo 6—Resultados 53

6.1 Considerações Iniciais 53
6.2 Resultados Espaço Tempo 53
6.3 Resultados Espaço Fase 54
6.3.1 Variação de posição de janela 55
6.3.1.1 KNN 56
6.3.1.2 DWNN 57
6.3.1.3 Random Forest 58
6.3.1.4 SVR 59
6.3.1.5 Resumo do experimento - Variação de janela 60
6.3.2 Multilacuna 60
6.3.3 Variação completa 62

Capítulo 7—Conclusão 67

Apêndice A—Resultados complementares obtidos com o método proposto 69

LISTA DE FIGURAS

1.1 (a) Série Temporal com valores ausentes. (b) Aplicação de Spline Cúbicas para substituição de valores ausentes. *f(x)* em preto representa a função geradora e *f'(x)* em vermelho tracejado representa a função aproximada utilizando Splines Cúbicas. 3
1.2 Substituição em série caótica utilizando Splines Cúbicas. Em vermelho estimação com Splines Cúbicas, em azul os dados originais. 5

2.1 Série exemplo original antes de aplicada a decomposição. 12
2.2 Extração de IMFs: Conceitos de envelope superior e inferior 13
2.3 Resultado da decompsição EMD sobre a série exemplo 2.1. 14
2.4 Série produzida pelas equações de Lorentz. 18
2.5 Série de Lorenz - Desdobramento. 18

4.1 Execução do BMDE para substituição de valores ausentes. a) série temporal com dados ausentes; b) componente estocástico estimado com BMDE; c) componente determinístico estimado com Splines Cúbicas; d) Série temporal resultante do método BMDE. 35
4.2 Pontos extremos e *zero-crossings* dentro das janelas do subconjunto à esquerda e à direita da lacuna de valores ausentes. 36
4.3 Resultado da IMF após estimar os extremos para substituir os valores ausentes. 37
4.4 Representação do desdobramento de uma série com *m=3* e *d=5*. Utilizando a coluna D3 como rótulo para um algoritmo de aprendizado de máquina, o valor ausente (NA) pode ser calculado e replicado nas demais colunas. 40
4.5 Limitação para dimensão de separação. Nesta série de exemplo a dimensão de separação é limitada ao valor 3 devido a posição da lacuna de valores ausentes. 42

5.1 Série de Lorenz. 45
5.2 Série de Lorenz - Desdobramento. 45
5.3 Série de Rössler 46
5.4 Série de Rössler - Desdobramento 46
5.5 Série de Hénon. 46
5.6 Série de Hénon - Desdobramento. 46
5.7 Diagrama de bifurcação do modelo logístico com a variação de r. 47

5.8 Organização do experimento de variação de janela - Uma técnica é aplicada em todas as séries e para cada série será gerada uma série com uma lacuna de valores ausentes em uma janela diferente. 50
5.9 Organização do experimento multilacuna - Uma técnica é aplicada em todas as séries e para cada série será gerada uma série com 8 lacunas de tamanhos variados. 51
5.10 Organização do experimento de variação completa - A técnica KNN é aplicada em todas as séries e para cada série será gerada uma ou mais lacunas de valores ausentes com tamanhos variados. Cada experimento será repetido 40 vezes para cada porcentagem de perda de dados. 52

6.1 Um conjunto de 6 experimentos usando BMDE (linhas azuis), SSA (linhas vermelhas), Splines Cúbicas (linhas roxas) e os resultados esperados (linhas pretas). 55
6.2 Plotagem de violino comparando BMDE, SSA e Splines Cúbicas em um grande *dataset*. 56
6.3 Substituição de valores ausentes utilizando o PSGF e o SSA. Pode-se perceber que com parametrização e métodos simples, a substituição pode gerar resultados promissores em relação à técnicas que consideram apenas o espaço tempo. 57
6.4 Distribuição dos erros do experimento de variação completa sobre a série de Lorenz 63
6.5 Distribuição dos erros do experimento de variação completa sobre a série de Henon. 64
6.6 Distribuição dos erros do experimento de variação completa sobre a série do Mapa Logístico. 65
6.7 Distribuição dos erros do experimento de variação completa sobre a série de Rössler. 66

A.1 Distribuição dos erros do PSGF + KNN para cada série e cada dimensão de separação. O método apresenta mais estabilidade sobre as séries do Mapa Logístico e a série de Rössler 70
A.2 Distribuição dos erros do PSGF + DWNN para cada série e cada dimensão de separação. A série de Rössler foi a única que apresentou uma melhoria sobre a técnica KNN 71
A.3 Distribuição dos erros do PSGF + RF para cada série e cada dimensão de separação. A técnica RF não apresentou nenhuma melhoria sobre as outras técnicas. 72
A.4 Distribuição dos erros do PSGF + SVR para cada série e cada dimensão des separação. A técnica SVR apresentou altos índices de erros. 73
A.5 Resultados dos experimentos de variação de janela. Cruzamento das dimensões de separação de menores erros com a posição das lacunas. Nota-se que a posição da lacuna afeta as séries de maneiras diferentes entre si. . . 74

A.6 Resultados dos experimentos de variação de janela. Cruzamento das dimensões de separação de menores erros com a posição das lacunas. Nesses experimentos as séries de Lorenz, Rossler e o Mapa Logístico apresentam maiores erros quando a posição da lacuna está localizada nas primeiras janelas. 75
A.7 Resultados dos experimentos de variação de janela. Cruzamento das dimensões de separação de menores erros com a posição das lacunas. Novamente as séries não apresentam similaridade entre seus resultados e suas posições de janela. 76
A.8 Resultados dos experimentos de variação de janela. Cruzamento das dimensões de separação de menores erros com a posição das lacunas. É possível perceber que séries de Henon, Rossler e o Mapa Logístico apresentam picos semelhantes na posição 3 da janela de valores ausentes. . . . 77

LISTA DE TABELAS

1.1 Exemplo de um conjunto de dados com valor ausente. Neste caso, um conjunto de dados com valor ausente foi simulado removendo o valor de um atributo da base de dados Iris. 2

3.1 Número de artigos após cada critério. 21
3.2 Número de publicação por país de autores. 22
3.3 Número de publicações por ano. 23
3.4 Relação de número de publicações e tipo. 23
3.5 Autores, ano e maiores pontuações. 23
3.6 Sumário dos artigos selecionados 24
3.7 Tabela de referência do sumário 29

6.1 Dimensões de separação para cada série - KNN. 58
6.2 Dimensões de separação para cada série - DWNN. 58
6.3 Dimensões de separação para cada série - RF. 59
6.4 Dimensões de separação para cada série - SVR. 60
6.5 Resumo do experimento multilacuna. As tabelas resumem os valores de RMSE para cada técnica, série e tamanho de lacuna. Os valores foram mapeados em diferentes cores: melhores resultados em tons de verde e piores resultados em tons de vermelho. 61

Capítulo 1

INTRODUÇÃO

1.1 CONTEXTO

O monitoramento e o armazenamento de dados coletados de sistemas têm crescido de maneira significativa nos últimos anos. Em 2015, cerca de $2,5$ bilhões de GB de dados foram coletados no mundo a partir de diferentes dispositivos e sistemas como, por exemplo, sensores remotos, sinais de GPS e redes sociais (ZIKOPOULOS et al., 2015a, 2015b). De acordo com Özköse, Ari e Gencer (2015), estima-se que 90% do volume de dados coletados até aquele ano fora produzido nos 2 anos anteriores.

Esse aumento significativo na quantidade de dados tem dificultado a tarefa de especialistas na análise e extração de novas informações. Buscando superar essas dificuldades, técnicas de Aprendizado de Máquina (AM) têm sido propostas visando induzir hipóteses que sejam capazes de descrever relações entre os dados analisados (Mitchell et al., 1997; Faceli et al., 2015).

Entretanto, como as técnicas de AM são diretamente ajustadas sobre os dados coletados (*data-driven models*), a presença de valores inválidos ou ausentes pode levar a inferências e conclusões erradas ou, até mesmo, inviabilizar sua aplicação. Esse problema pode ser ocasionado por diferentes razões como: i) monitoramento de uma nova variável (logo, dados históricos terão valores ausentes para essa variável); ii) problemas nos dispositivos de monitoramento (e.g. falha elétrica, defeito técnico, obstrução entre coletor e a variável observada); e iii) problemas na transmissão entre a coleta e o armazenamento. De acordo com os trabalhos publicados por Little e Rubin (2019) e por Sentas e Angelis (2006), valores inválidos ou ausentes podem seguir 3 padrões de comportamento:

- *Missing Completely at Random* (MCAR): O valor ausente em determinada variável não tem qualquer relação com outras variáveis observadas (ausente ou não);
- *Non-Ignorable Missingness* (NIM): A probabilidade de existir valores ausentes em uma variável depende da própria variável.

- *Missing at Random* (MAR): A probabilidade de existirem dados ausentes em uma variável depende de outras variáveis observadas.

Em geral, o tratamento desses valores inválidos ocorre de acordo com o tipo de dado coletado. Por exemplo, considerando que os dados coletados são independentes e identicamente distribuídos (iid), pode-se aplicar as próprias técnicas de AM para substituição de valores ausentes. Para exemplificar essa aplicação, considere a Tabela 1.1 que apresenta parte do conjunto de dados Iris. Esse conjunto de dados é bem conhecido na literatura de AM e contém 50 instâncias de cada espécie (*Setosa*, *Versicolor* e *Virginica*) da flor do tipo Iris. Cada instância foi registrada com 5 atributos: Comprimento e Largura da Sépala; Comprimento e Largura da Pétala; e Espécie. Neste exemplo, o atributo Espécie, que indica o rótulo da instância, foi desconsiderado. Esse conjunto de dados originalmente não possui nenhum valor ausente. Entretanto, para fins de ilustração, substituiu-se o valor real da Largura da Pétala da instância de número 10, que era igual a $0,2$, por "NA".

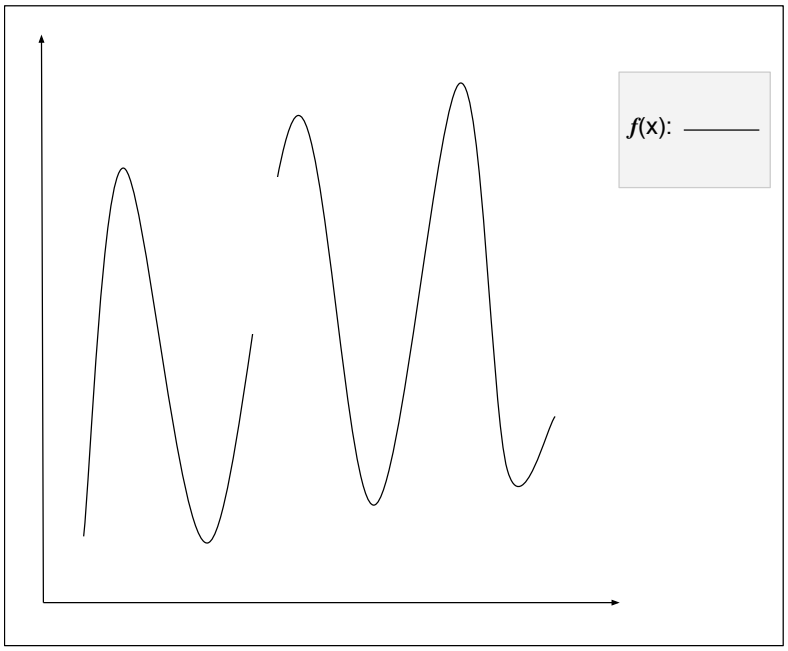
Tabela 1.1 Exemplo de um conjunto de dados com valor ausente. Neste caso, um conjunto de dados com valor ausente foi simulado removendo o valor de um atributo da base de dados Iris.

	Comprimento da Sépala	Largura da Sépala	Comprimento da Pétala	Largura da Pétala
8	5,00	3,40	1,50	0,2
9	4,40	2,90	1,40	0,2
10	4,90	3,10	1,50	**NA**
11	5,40	3,70	1,50	0,2
12	4,80	3,40	1,60	0,2
13	4,80	3,00	1,40	0,1

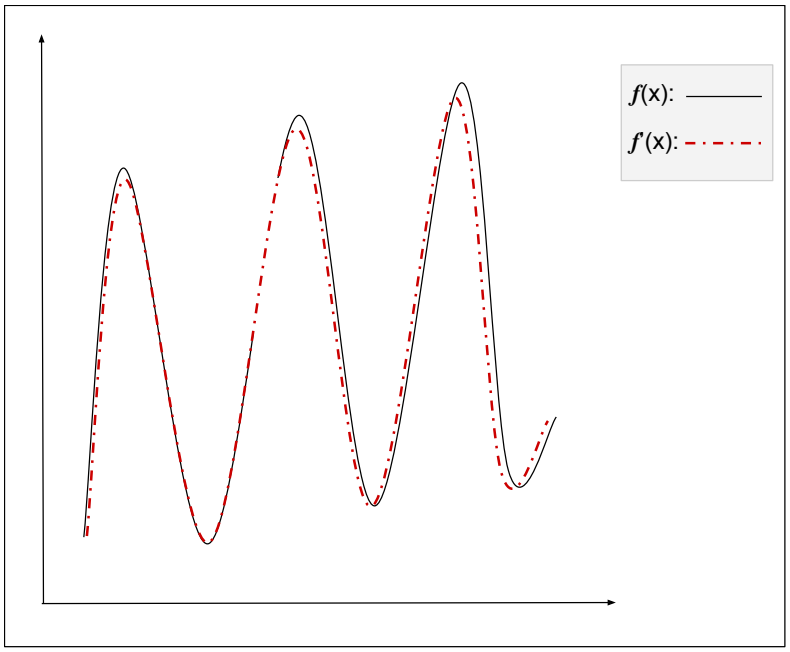
Esse problema pode ser facilmente resolvido utilizando a técnica de Aprendizado de Máquina Supervisionado KNN (*K-Nearest Neighbors*) (Mitchell et al., 1997; HASTIE et al., 1999). De maneira resumida, essa técnica visa encontrar um número k de vizinhos mais próximos de uma determinada instância. Nesse sentido, separou-se a Largura da Pétala de todas as instâncias em um vetor ν. Em seguida, verificou-se quais eram as k instâncias mais semelhantes à instância 10 mostrada na Tabela 1.1. Por fim, calculou-se a média da Largura da Pétala considerando os k vizinhos mais próximos (semelhantes) da instância 10. O valor da média foi, então, usado para substituir o valor ausente[1]. Para este exemplo, ao escolher um valor $k = 6$, obteve-se um valor substituto igual a $0,2$.

Contudo, a aplicação desse método não é adequada quando existe dependência entre dados. Por exemplo, se os dados apresentam uma dependência temporal entre suas observações, como em Séries Temporais (BOX; JENKINS; REINSEL, 1994), deve-se utilizar técnicas específicas para tratar esse problema (e.g., métodos de interpolação). Visando ilustrar essa situação, considere a Figura 1.1(a), a qual representa uma série temporal sintética criada a partir da combinação de uma função senoidal e uma tendência. Neste exemplo, parte da série temporal foi removida e em seu lugar foram adicionados valores "NA".

[1]Na literatura, essa versão do algoritmo é comumente utilizada em problemas de regressão.



(a)



(b)

Figura 1.1 (a) Série Temporal com valores ausentes. (b) Aplicação de Spline Cúbicas para substituição de valores ausentes. *f(x)* em preto representa a função geradora e *f'(x)* em vermelho tracejado representa a função aproximada utilizando Splines Cúbicas.

Para substituir esses valores ausentes, pode-se, por exemplo, utilizar o método de interpolação Spline Cúbica. De maneira resumida, essa técnica analisa um conjunto

de observações (chamados de nós de interpolação) $X = \{x_0, x_1, \ldots, x_n\}$, os quais foram produzidos por uma regra geradora $f(\cdot)$ que define o comportamento de um sistema. Uma vez que $f(\cdot)$ é desconhecida, o objetivo, então, dos métodos de interpolação é estimar uma função aproximada $f'(\cdot)$ tal que $f(x_i) = f'(x_i), \forall x_i \in X$ (RUGGIERO; LOPES, 1996). Essa função é utilizada para estimar os valores ausentes nos instantes de tempo t, simplesmente, calculando a função $f'(x_t)$. A Figura 1.1(b) ilustra o resultado final desta interpolação.

Apesar de apresentar um excelente resultado para essa série temporal sintética, as funções de interpolação possuem fortes limitações quando as séries possuem um outro tipo de comportamento não muito incomum na natureza: o comportamento caótico.

Sistemas caóticos apresentam um conceito fundamental chamado sensibilidade às condições iniciais. Este conceito indica que pequenas alterações na condições iniciais geram grandes alterações a longos prazos, tornando previsões de modelos caóticos pouco previsíveis (ALLIGOOD; SAUER; YORKE, 1997b). Séries temporais caóticas podem parecer estocásticas por apresentarem um caráter de instabilidade. No entanto, se as condições iniciais são exatamente conhecidas um sistema caótico pode ser replicado e previsto. Devido a complexidade dos sistemas caóticos, encontrar as condições iniciais com precisão é uma tarefa complicada e qualquer erro pode levar a previsões completamente diferentes do esperado.

Para ilustrar essa limitação, considere a Figura 1.2 cujos valores de observação foram produzidos a partir de um Sistema de Lorenz, que descreve a circulação de fluidos uniformemente aquecidos a partir da parte inferior e resfriado em sua parte superior. Definindo seus parâmetros com valores iguais a $\sigma = 10$, $\rho = 28$ e $\beta = 8/3$, obtém-se uma série temporal com comportamento caótico.

Assim como experimentado na série anterior, essa série de Lorenz teve parte de suas observações substituídas por "NA". Entretanto, após a interpolação com Splines Cúbicas, nota-se que os valores obtidos (linha vermelha) diferem consideravelmente dos valores esperados (linha azul). Esse problema também acontece com outros métodos de substituição de valores ausentes para séries temporais como, por exemplo, o método SSA (*Singular Spectrum Analysis*) que é o principal método de substituição de valores ausentes utilizado na literatura.

No entanto, é importante destacar que esse problema ocorre porque tais técnicas foram desenvolvidas para analisar o comportamento de séries considerando apenas suas relações no domínio temporal. A partir da Revisão Sistemática da Literatura (Capítulo 3), pode-se observar que a ausência de técnicas que façam essa substituição modelando dependências entre observações em um número maior de dimensões, como visto em séries caóticas, ainda é um problema em aberto na área de Análise de Séries Temporais.

Este trabalho investigou a substituição de valores ausentes em séries temporais a partir de duas perspectivas: (i) utilizando de técnicas de decomposição e realizando a substituição nos componentes resultantes; e (ii) empregando ferramentas de Sistemas Dinâmicos e Teoria do Caos para substituir valores ausentes no espaço fase.

1.2 MOTIVAÇÃO

Dados inválidos ou ausentes[2] em bases de dados é um tipo de problema comumente enfrentado na modelagem de sistemas reais conforme discutido em Junger e Leon (2015). Nesse trabalho, os autores apresentam as vantagens do uso de técnicas de substituição de valores ausentes no monitoramento de poluentes no ar. Segundo os autores, a substituição de dados inválidos não é uma tarefa trivial. Além disso, dependendo do contexto, a aplicação de técnicas que exigem uma estimativa elevada de parâmetros pode não ser viável e, por outro lado, métodos simples tendem enviesar a modelagem obtida.

É importante destacar, ainda, que tais técnicas são importantes principalmente em cenários onde não se pode repetir o processo de coleta de dados. Neste contexto, Brás e Menezes (2007) discutem que há uma probabilidade de 85% de existir pelo menos um valor ausente em microarranjos de DNA. Por razões econômicas ou relacionadas à disponibilidade de mais amostras biológicas, repetir os experimentos para obter uma matriz completa de expressão genética é geralmente inviável.

O problema da substituição de valores ausentes (*"gap filling"*) tem sido abordado por diferentes perspectivas. Conforme discutido anteriormente, há pesquisadores que dedicam esforços para desenvolver novos métodos adequados para dados independentes e identica-

[2]No contexto deste trabalho de mestrado, valores inválidos e ausentes são tratados da mesma maneira, uma vez que, em ambas as situações, valores substitutos deverão ser estimados antes da aplicação de técnicas de modelagem.

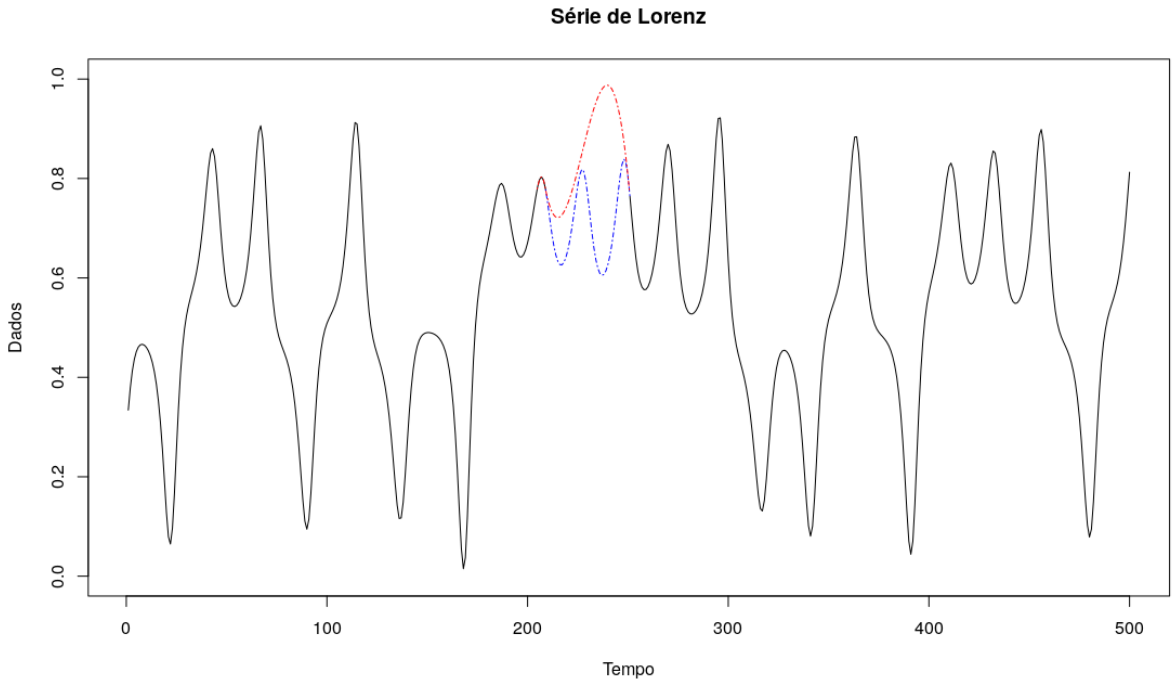


Figura 1.2 Substituição em série caótica utilizando Splines Cúbicas. Em vermelho estimação com Splines Cúbicas, em azul os dados originais.

mente distribuídos. Por outro lado, existem pesquisas que propõem novas técnicas para substituição de valores ausentes em séries temporais. No caso de séries, observou-se que há necessidade de desenvolvimento de um novo método para um subconjunto específico de dados: séries temporais caóticas, o qual tem sido pouco explorado na literatura.

Após a execução de uma Revisão Sistemática da Literatura (RSL), notou-se que apenas dois trabalhos, dos 21 selecionados, lidavam com séries temporais caóticas. O trabalho publicado por Facchini e Mocenni (2011) abordou substituição de valores ausentes em séries com comportamento caótico. O artigo apresenta uma técnica chamada *twin surrogates* aplicada sobre dados de concentração de oxigênio dissolvido coletados do lago de Ortobello na Itália e sobre séries temporais obtidas do mapa logístico (comportamento caótico). Prasomphan, Lursinsap e Chiewchanwattana (2009) também abordaram séries temporais caóticas com um novo algoritmo de *Bootstrap*[3] que utiliza processos de imputação. No entanto, os dois trabalhos não realizam o processo de imputação considerando a série desdobrada no espaço fase, o que permitiria compreender melhor as relações entre suas observações (conforme descrito no Capítulo 2.4). Com base nessas observações, definiu-se a hipótese para este projeto, a qual é detalhada a seguir.

1.3 HIPÓTESE

A ausência de trabalhos desenvolvidos para substituição de valores ausentes em séries temporais caóticas, conforme observado na Revisão Sistemática da Literatura desenvolvida neste projeto, e, de acordo com estudos preliminares, o baixo desempenho obtido com a aplicação de técnicas tradicionais levaram ao desenvolvimento da seguinte hipótese para este trabalho:

A substituição de valores ausentes em séries temporais caóticas é realizada com melhor desempenho se modelos forem aplicados sobre suas observações transformadas no espaço de coordenadas de atraso.

1.4 OBJETIVOS

Esta dissertação de mestrado tem como objetivo principal investigar a hipótese apresentada anteriormente. Para isso, foi desenvolvida uma nova técnica de substituição de valores ausentes que utiliza ferramentas da área de Sistemas Dinâmicos e Teoria do Caos (ALLIGOOD; SAUER; YORKE, 1997a). Essas ferramentas foram projetadas para transformar uma série temporal caótica, reconstruindo suas observações do domínio temporal ($\mathbb{R}$) para o espaço fase – também referenciado como espaço de coordenada de atraso – ($\mathbb{R}^m$, sendo m o número de dimensões após a reconstrução). Cada etapa planejada e desenvolvida para alcançar esse objetivo está brevemente discutida nos próximos capítulos.

[3]Método para estimar a distribuição de um estimador ou estatística de teste por reamostragem de dados ou um modelo estimado a partir dos dados (HÄRDLE; HOROWITZ; KREISS, 2003).

1.5 ORGANIZAÇÃO

Esta dissertação de mestrado está organizada da seguinte forma: o Capítulo 2 apresenta um referencial teórico, discutindo de maneira geral os principais conceitos abordados neste projeto; no Capítulo 3, uma Revisão Sistemática da Literatura sobre substituição de dados ausentes em séries temporais é apresentada; no Capítulo 4, a proposta e a metodologia apresentada para o desenvolvimento deste projeto de mestrado são discutidos em detalhes; o Capítulo 5 apresenta e explica os experimentos desenvolvidos tanto para o espaço tempo quanto para o espaço fase; o Capítulo 6 apresenta um conjunto de resultados preliminares sobre o espaço tempo que foram publicados na principal conferência nacional de Inteligência Artificial (8th Brazilian Conference on Intelligent Systems – BRACIS) e também os resultados obtidos sobre o espaço fase; por fim, o Capítulo 7 apresenta as conclusões e trabalhos futuros a serem desenvolvidos acerca dos temas apresentados.

Capítulo

2

REFERENCIAL TEÓRICO

2.1 CONSIDERAÇÕES INICIAIS

Neste capítulo, serão apresentados os conceitos fundamentais para elaboração desta dissertação de mestrado. Alguns desses conceitos também são detalhados e referenciados no Capítulo 3, que apresenta a Revisão Sistemática da Literatura. Inicialmente, serão introduzidos conceitos básicos de decomposição de Séries Temporais. Em seguida, serão abordados métodos de substituição de valores ausentes tradicionalmente considerados na literatura. E, por fim, serão apresentados os principais conceitos de Sistemas Dinâmicos e Teoria do Caos.

2.2 DECOMPOSIÇÃO DE SÉRIES TEMPORAIS

Uma série temporal pode ser definida como uma sequência de observações coletadas ao longo do tempo (MORETTIN; TOLOI, 2006), cujos valores são, normalmente, influenciados por três componentes principais: tendência, sazonalidade e estocasticidade. O componente de tendência representa o crescimento ou decrescimento da série ao longo do tempo. O componente sazonal representa a recorrência ou repetição de estados do sistema analisado. Por fim, o componente estocástico representa influências que, de maneira aleatória, afetam os valores produzidos pelos sistemas.

Em análise de séries temporais, a decomposição de séries é utilizada para identificar e, individualmente, modelar esses componentes. A Equação 2.1 representa os componentes que são considerados nesse trabalho, onde $x(t)$ representa a série temporal, $T(t)$ o componente de tendência, $S(t)$ o componente de sazonalidade e $\varepsilon(t)$ representa a estocasticidade(MORETTIN; TOLOI, 2006; RIOS; MELLO, 2013a). Neste trabalho, será considerado, ainda, que $T(t)$ e $S(t)$ somadas correspondem a um componente determinístico.

$$x(t) = T(t) + S(t) + \varepsilon(t) \tag{2.1}$$

Decompor uma série tem por objetivo separar informação embutida de forma que seja mais simples identificar padrões como, por exemplo, sazonalidade, tendência e ruídos. Nesta seção, serão discutidas duas técnicas de decomposição: Análise Espectral Singular (Singular Spectrum Analysis - SSA) e Decomposição de Modo Empírica (Empirical Mode Decomposition - EMD). Ambas separam, usando diferentes abordagens, séries em componentes aditivos. É importante destacar que, além de permitir decompor séries temporais, SSA é a principal ferramenta utilizada na literatura para substituição de valores ausentes.

2.2.1 Singular Spectrum Analysis - SSA

A técnica é executada em duas etapas: i) decomposição das observações; e ii) reconstrução substituindo valores inválidos considerando importantes informações implícitas presentes nos componentes decompostos como tendências, sazonalidades, ciclos, periodicidades com amplitudes variadas (HASSANI, 2007).

Na etapa de decomposição, um conjunto de observações $X = \{x_1, ..., x_n\}$, com $|X| = n$, é transformado em uma série com L dimensões, produzindo uma matriz de trajetórias conforme apresentada na Equação 2.2, tal que $K = n - L + 1$.

$$\mathbf{X} = [X_1 : \ldots : X_K] = (x_{ij})_{i,j=1}^{L,K} = \begin{pmatrix} x_1 & x_2 & x_3 & \ldots & x_K \\ x_2 & x_3 & x_4 & \ldots & x_{K+1} \\ x_3 & x_4 & x_5 & \ldots & x_{K+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & x_{L+2} & \ldots & x_N \end{pmatrix} \tag{2.2}$$

Em seguida, o produto entre a matriz trajetória e sua transposta , $\mathbf{S} = \mathbf{X}\mathbf{X}^T$, é decomposto pelo método SVD (*Singular Value Decomposition*) produzindo seus autovalores e autovalores ortogonais e normalizados (HASSANI, 2007; GOLYANDINA; OSIPOV, 2007b; GOLYANDINA; NEKRUTKIN; ZHIGLJAVSKY, 2001). O SVD da matriz trajetória pode descrito conforme a Equação 2.3.

$$\mathbf{X} = \mathbf{E_1} + ... + \mathbf{E_d} \tag{2.3}$$

Na etapa de reconstrução, é realizado um agrupamento das matrizes elementares $\mathbf{E_i}$ produzidas nos passos anteriores. Para tanto, o SVD apresentado na Equação 2.3 é particionado em subconjuntos disjuntos: $X = E_{I1} + ... + E_{Im}$.

Logo, a saída desse passo é um conjunto de somas de matrizes resultantes da matriz trajetória. O objetivo do agrupamento é diminuir o número de componentes (matrizes elementares) na SVD da matriz trajetória (GOLYANDINA; OSIPOV, 2007b). A decisão de escolha desses conjuntos é fundamental para a aplicação do método SSA e é baseada na propriedade de separabilidade entre conjuntos. A separabilidade é mensurada por correlação ponderada, *i.e.*, sejam dois conjuntos de observações $X_t^{(1)}$ e $X_t^{(2)}$, a correlação ponderada entre eles é representada por:

$$\rho_{12}^{(\omega)} = \frac{< X_t^{(1)}, X_t^{(2)} >_\omega}{\parallel X_t^{(1)} \parallel_\omega \parallel X_t^{(2)} \parallel_\omega}$$

A norma da i-ésima subsérie é dada por $\parallel X_t^{(i)} \parallel_\omega = \sqrt{< X_t^{(i)}, X_t^{(i)} >_\omega}$, sendo que o produto interno é definido por $< X_t^{(i)}, X_t^{(j)} >= \sum_{c=1}^{N} \omega_c X_c^{(i)} X_c^{(j)}$ com os pesos $\omega_c = min\{c, L, n-c\}$ e $L \leq n/2$.

Assim, com base na análise da correlação ponderada, pode-se estimar a separabilidade dos componentes (HASSANI, 2007). Nesse sentido, Golyandina e Osipov (2007b) criaram uma técnica de substituição de dados inválidos, usando SSA, que extrai componentes aditivos de sinais, como tendência e sazonalidade, e simultaneamente substitui valores inválidos presentes no conjunto de dados. De maneira geral, quando valores inválidos estão localizados no final do conjunto de dados (lado direito da série de observações), essa técnica substitui tais valores com base nos componentes reconstruídos usando SSA. Para os demais valores, na etapa de reconstrução, após escolher-se um subespaço e uma projeção de vetores no espaço de coordenadas de atraso, obtém-se um valor para substituição do primeiro dado inválido a partir da combinação linear de valores anteriores e coeficientes dos componentes principais extraídos com o método SSA. Em seguida, esse passo é repetido até que todos os valores inválidos sejam substituídos (GOLYANDINA; OSIPOV, 2007b).

2.2.2 Empirical Mode Decomposition - EMD

Apesar de não ter sido citado na RSL (maiores detalhes na Seção 3), o método EMD, desenvolvido com base na Transformada de Hilbert (HUANG et al., 1998), é utilizado para decomposição, processamento e análise de sinais. No contexto de substituição de valores ausentes, foi encontrado um artigo que utiliza o EMD de maneira semelhante ao método proposto em (MOGHTADERI; BORGNAT; FLANDRIN, 2012). De maneira resumida, esse trabalho apresenta uma modificação do algoritmo EMD para que o mesmo possa decompor séries temporais contendo valores ausentes, realizando as respectivas substituições em cada um dos monocomponentes extraídos, os quais são somados ao final para encontrar os valores resultantes.

O EMD, assim como o SSA, é um método de decomposição de sinais em componentes aditivos. Cada um desses componentes é chamado de IMF (Intrinsic Mode Function) que devem satisfazer dois critérios:

- O número de extremos locais e o número de *zero-crossings*[1] de cada IMF deve variar em no máximo 1;
- A média dos envelopes superior e inferior deve ser igual ou próximo a zero.

[1]Neste trabalho optou-se por cognominar *zero-crossings*, assim como na literatura, os pontos do eixo Y que tem valor 0 e tocam o eixo X nas IMFs

As IMFs são calculadas por um processo chamado de peneiragem (Sifting Process) onde cada componente resultante da média entre o envelope superior e o envelope inferior é subtraída da série atual e então submetida aos critérios de uma IMF. Se o componente extraído não passar pelos critérios de uma IMF, o processo se repete até que uma IMF seja encontrada. Um envelope é calculado através dos valores máximos e mínimos da série, onde o envelope superior corresponde a uma Spline Cúbica que interpola os valores máximos da série referenciada. Analogamente, o envelope inferior interpola os valores mínimos.

A Figura 2.1 apresenta uma série não ruidosa com componentes de sazonalidade e tendência. Na Figura 2.2, a série começa a ser processada com a identificação de seus valores máximos e mínimos. Em seguida, uma função de Spline Cúbica é utilizada para interpolar esses valores em envelopes superior e inferior, respectivamente. Por fim, calcula-se a média entre esses envelopes. A Figura 2.2 ilustra o envelope superior em vermelho, o envelope inferior em azul e a média entre os dois é representada na linha pontilhada.

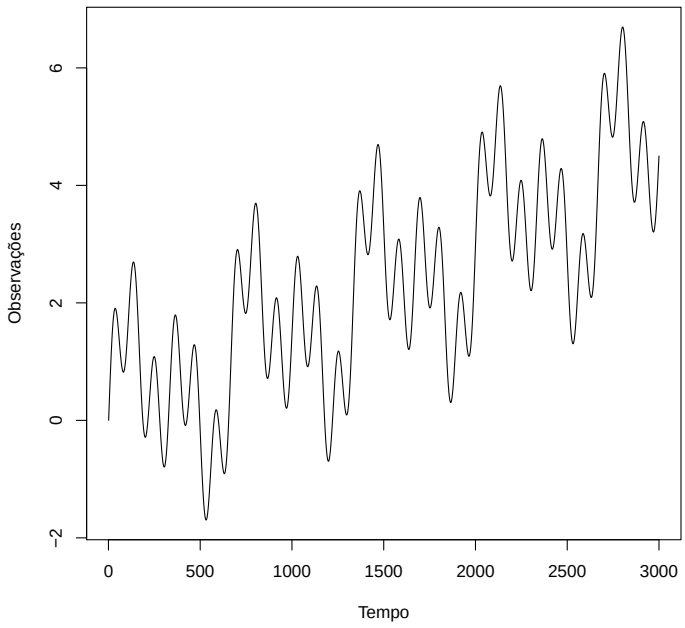


Figura 2.1 Série exemplo original antes de aplicada a decomposição.

Todos os passos para a execução do EMD estão descritos no Algoritmo 1. Por fim, a Figura 2.3 demonstra o resultado da decomposição de uma série com 3000 pontos, sendo que: a) apresenta a série original, b-e) ilustram as IMFs extraídas e f) o resíduo.

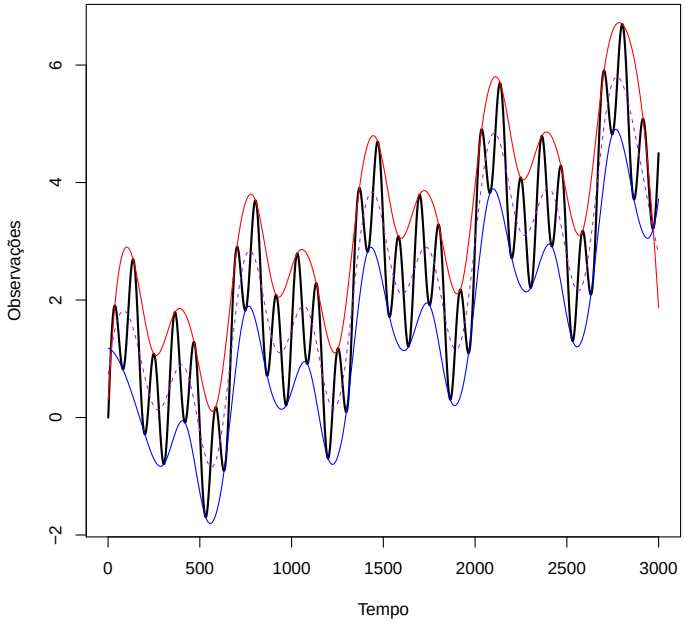


Figura 2.2 Extração de IMFs: Conceitos de envelope superior e inferior. Em vermelho o envelope superior, em azul o envelope inferior e a linha pontilhada representam a média dos envelopes.

Algorithm 1 Algoritmo EMD.

procedure DECOMP
  $Residuo \leftarrow X(t)$
  $I_1(x) \leftarrow X(t)$
  $i \leftarrow 1$, $k \leftarrow 1$
  while *Media Resíduo* $\neq 0$ **or** *Monótono* **do**
    while $I_i(t)$ *Não é uma IMF* **do**
      $EnvSup(t) \leftarrow$ *spline de máximos locais*
      $EnvInf(t) \leftarrow$ *spline de mínimos locais*
      $MediaEnvs(t) \leftarrow (EnvSup(t) + EnvInf(t))/2$
      $I_i(t) \leftarrow I_i(t)$ - $MediaEnvs(t)$
      $i \leftarrow i + 1$
    $IMF_k(t) \leftarrow I_i(t)$
    $Residuo \leftarrow Residuo$ - $IMF_k(t)$
    $k \leftarrow k + 1$

2.3 SUBSTITUIÇÃO DE VALORES AUSENTES – GAP FILLING

A complexidade ao lidar com dados ausentes depende do sistema monitorado, da frequência e do tamanho (quantidade de valores ausentes em sequência) com que aparecem, além

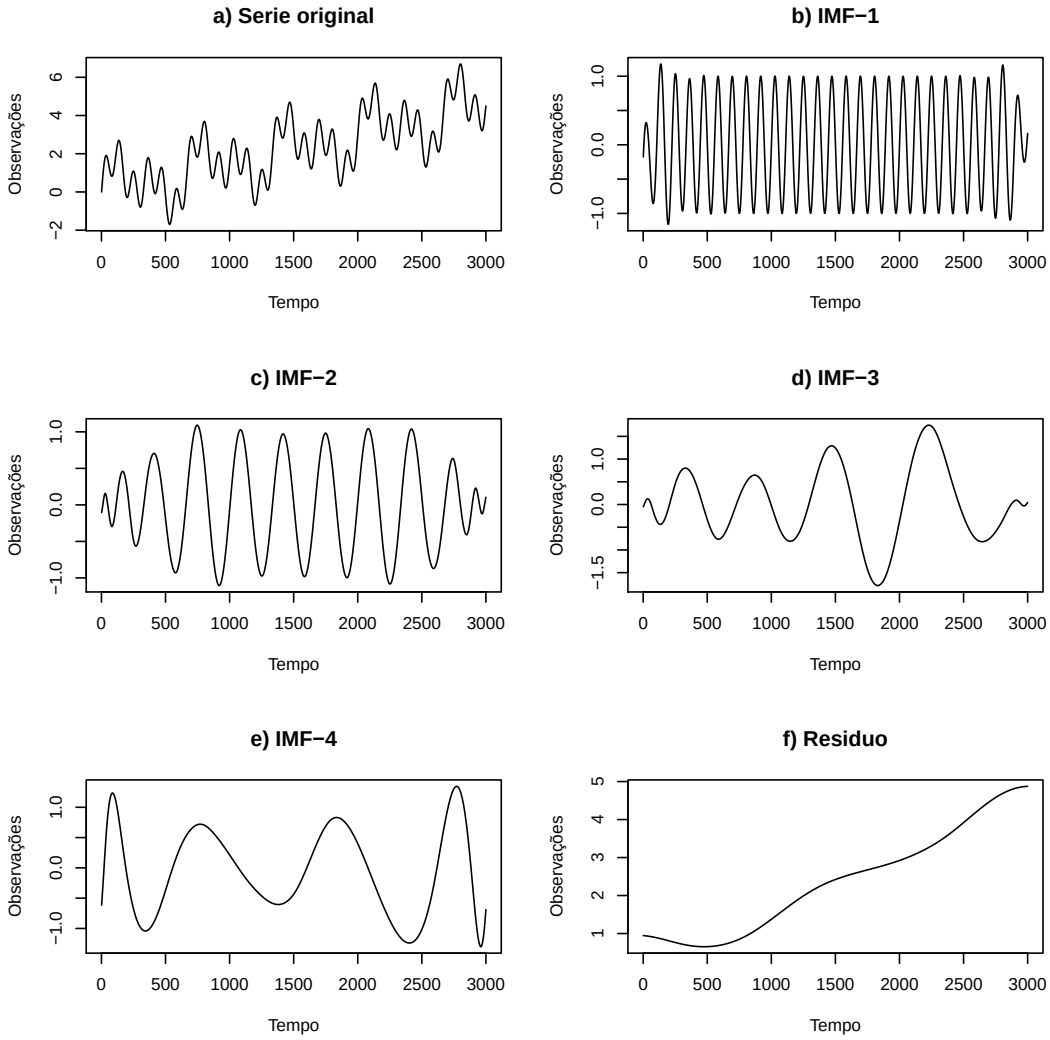


Figura 2.3 Resultado da decompsição EMD sobre a série exemplo 2.1.

da posição (e.g. início, meio ou final) na série analisada. Supondo que os dados não apresentem ruído e sazonalidade, uma forma simples de realizar substituição de dados ausentes é por meio da aplicação de uma regressão linear. Se os dados apresentam uma sazonalidade bem definida (e.g. uma função senoidal) com pouco ruído e sem tendência, pode-se utilizar um método de interpolação como a Spline Cúbica. Por outro lado, há cenários em que a simples remoção do valor ausente é suficiente para modelar um dado sistema. De maneira resumida, conforme discutido nesta seção, diferentes técnicas de substituição de valores ausentes podem ser adotadas de acordo com as especificidades de cada sistema.

2.3.1 Remoção dos Valores Ausentes

As técnicas mais conhecidas neste caso são *listwise deletion* e *pairwise deletion*, as quais simplesmente removem dados ausentes. Recomenda-se utilizar esta técnica quando os valores ausentes são do tipo MCAR. Quando a evidência do MCAR não é suportada, é necessário buscar uma outra abordagem para tratar os valores ausentes. A técnica *pairwise deletion* é um pouco mais sofisticada pois leva em conta o modelo avaliado. Enquanto *listwise* remove todas as instâncias com dados ausentes, *pairwise* remove apenas as instâncias que compõe um modelo ou análise. (OLINSKY; CHEN; HARLOW, 2003)

2.3.2 Interpolação

Os valores ausentes são interpolados por uma função estimada $f'(\cdot)$ que aproxima os pontos produzidos por uma função original $f(\cdot)$. Esta função irá interpolar todos os pontos disponíveis e, então, servir de função geradora para os dados ausentes. A função interpoladora deve ser escolhida de acordo com a base de dados. Se a base apresenta apenas uma tendência, uma função interpolante de grau 1 é suficiente. A função interpoladora mais utilizada é a Spline Cúbica, na qual um conjunto de funções do terceiro grau interpolam todos os valores disponíveis.

2.3.3 Imputação

Uma imputação ocorre quando o valor é substituído por dados que fazem sentido dentro do contexto. Em alguns casos, valores ausentes podem ser resolvidos pela simples repetição de um valor anterior que seja suficientemente válido para substituir o dado ausente. Outras alternativas estão relacionadas à aplicação da moda ou da média dos valores mais próximos. Existem ainda modos mais complexos de imputação como a imputação múltipla, imputação em classes, maximização de expectativa etc. Métodos de regressão também entram neste grupo bem como algumas técnicas de Aprendizado de Máquina como Árvores de Regressão e Redes Neurais Artificiais. Outros meios de imputação são discutidos no Capítulo 3

2.4 SISTEMAS DINÂMICOS E TEORIA DO CAOS

O estudo dos sistemas dinâmicos começa com as primeiras tentativas de descrever movimentos de sistemas físicos através de equações matemáticas. Isaac Newton conseguiu explicar a órbita dos planetas afirmando que existe atração gravitacional entre os corpos, provando, assim, que a força de atração entre dois corpos é proporcional ao produto da massa entre os corpos e inversamente proporcional ao quadrado da distância entre eles. Muitos outros estudos foram derivados deste estudo e melhorados utilizando equações diferencias para descrever comportamentos mais complexos (ALLIGOOD; SAUER; YORKE, 1997b).

Com o estudo das equações diferenciais, pode-se perceber que soluções mantidas em um espaço fechado teriam dois tipos de estados: i) estado estático, alcançado com a perda de energia dentro do sistema ou ii) estado periódico ou quase periódico. Quase

periodicidade é a propriedade de um sistema que demonstra uma periodicidade irregular, diferentemente da periodicidade que mostra comportamento regular após uma certa quantidade de tempo (ALLIGOOD; SAUER; YORKE, 1997b).

Cientistas sabiam que sistemas compostos por grande quantidade de partículas seriam extremamente complexos. Então, descobriu-se um terceiro tipo de movimento que foi chamado de iii) caótico e que, inclusive, pode ser observado em sistemas mais simples. Em sistemas com esse comportamento, pode-ser dizer que pequenas mudanças no estado inicial resulta em significativas perturbações em seu comportamento futuro (ALLIGOOD; SAUER; YORKE, 1997b).

No cenário atual, o comportamento caótico pode ser observado em várias áreas da ciência e são frequentemente usados na biologia e na física. O estudo desse comportamento é realizado pelo desdobramento do sistema no espaço fase.

2.4.1 Espaço Fase

Séries temporais são usualmente avaliadas no domínio temporal. Entretanto, existem situações nas quais algumas características podem não ser corretamente modeladas ao analisar a série em um espaço unidimensional. A estratégia, então, é reconstruir a série em espaços multidimensionais para, assim, entender melhor órbitas, atratores e repulsores. Este estudo foi originalmente desenvolvido por (WHITNEY, 1936) aplicando variáveis diferenciáveis para desdobrar funções em espaços multidimensionais. Esta análise ficou conhecida como **espaço fase**.

Ao analisar séries temporais no espaço fase busca-se encontrar pontos fixos e órbitas que explicam como a série se comporta ao longo do tempo. Define-se então como ponto fixo:

Definição 1 (Ponto Fixo)**.** *Seja f um mapa em $\mathbb{R}$ e ρ um número tal que $f(\rho) = \rho$. Se todos os pontos proximos a ρ, considerando uma vizinhança V, são atraídos a ρ, então ρ é chamado de ponto fixo de atração. Por outro lado, se todos os pontos são repelidos de ρ, então ele é chamado de ponto fixo repulsor.*

Analiticamente, órbitas são regiões onde os pontos são atraídos ou afastados de um ponto fixo. Para reconstruir uma série temporal no espaço fase, pode-se utilizar o Teorema de Imersão proposto por (TAKENS, 1981), o qual define que uma série temporal $\{x_0, x_1, \ldots, x_{n+1}\}$ pode ser reconstruída em um espaço multidimensional $x_n(m, \tau) = \{x_n, x_{n+\tau}, \ldots, x_{n+(m-1)\tau}\}$, também chamado de coordenadas de atraso, onde m é a dimensão embutida e τ é o atraso no tempo (*time delay*) também chamado de dimensão de separação. A saída desta técnica representa um conjunto de pontos em um espaço de m dimensões (normalmente Euclidiano).

2.4.2 Reconstrução no Espaço Fase

Como visto anteriormente, dois parâmetros são necessários para representar a série no espaço fase: a dimensão embutida e a de separação (*time delay*). Essas dimensões podem ser estimadas por duas técnicas bem conhecidas na literatura de Sistemas Dinâmicos e Teoria do Caos: Falsos Vizinhos Mais Próximos (False Nearest Neighbors - FNN) para

estimar a dimensão embutida e Autoinformação Mútua (Auto Mutual Information - AMI) para estimar a distância de separação.

A FNN proposta por Kennel, Brown e Abarbanel (1992) calcula os elementos mais próximos de cada ponto no espaço fase começando com uma única dimensão. Após calculadas todas as distâncias insere-se uma nova dimensão e recalcula-se as distâncias. Se esta distância aumentar, considera-se os pontos como falsos vizinhos e isso mostra a necessidade de mais uma dimensão.

Para estimar a distância de separação, Fraser e Swinney (1986) apresentam um estudo e afirmam que a AMI apresenta os melhores resultados. De maneira geral, a técnica funciona de forma incremental, onde inicia-se com uma distância de separação de valor 1 e incrementa-se o valor até que se encontre o primeiro mínimo formado pela função dos deslocamentos. A informação mútua média é definida pela Equação 2.4:

$$I(X,Y) = \int P_{XY}(x,y) log_2 \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} dxdy \tag{2.4}$$

onde X e Y seguem, respectivamente, as funções de distribuição de probabilidade P_X e P_Y, e X e Y ocorrem em pares com distribuição conjunta P_{XY} (KENNEL; BROWN; ABARBANEL, 1992).

Para efeito de visualização, considere uma série temporal produzida pelos atratores de Lorentz definidos pela Equação 2.5:

$$\begin{aligned} \frac{\partial x}{\partial t} &= \sigma(y - x) \\ \frac{\partial y}{\partial t} &= x(\rho - z) - y \\ \frac{\partial z}{\partial t} &= xy - \beta z \end{aligned} \tag{2.5}$$

Os parâmetros foram ajustados visando a produção de uma série caótica da seguinte maneira: $\sigma = 10$, $\rho = 28$ e $\beta = 8/3$

A Figura 2.4 mostra a série com comportamento caótico produzida pelas equações de Lorentz. Pode-se perceber que trata-se de uma série de alta complexidade de modelagem e previsão de seus resultados futuros.

Os resultados dos cálculos utilizando FNN e AMI foram, respectivamente, $m = 3$ e $\tau = 5$. A Figura 2.5 mostra a reconstrução dessa série no espaço fase.

2.5 CONSIDERAÇÕES FINAIS

Este capítulo apresentou um conjunto de conceitos fundamentais para compreensão das soluções propostas neste trabalho, cujo objetivo principal foi comprovar a hipótese apresentada na introdução. A seguir, serão apresentados os resultados obtidos com a Revisão Sistemática da Literatura, a qual foi conduzida visando identificar os principais trabalhos publicados na literatura da área de substituição de valores ausentes em séries temporais.

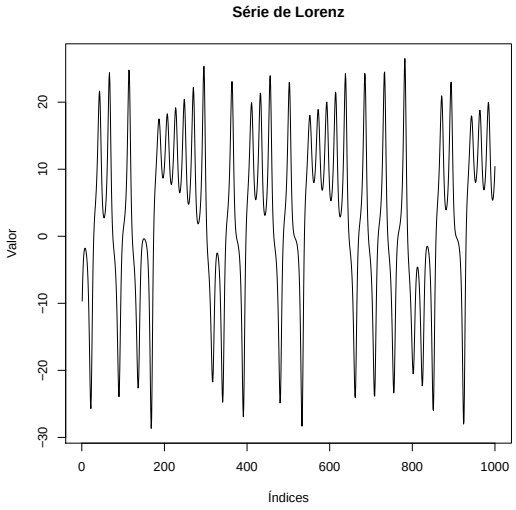


Figura 2.4 Série produzida pelas equações de Lorentz.

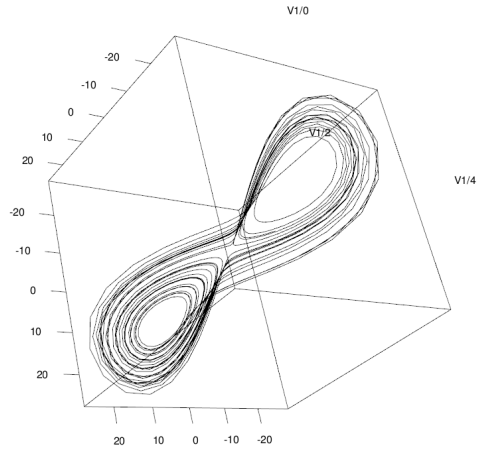


Figura 2.5 Série de Lorenz - Desdobramento.

Capítulo 3

REVISÃO SISTEMÁTICA DA LITERATURA

3.1 CONSIDERAÇÕES INICIAIS

Visando encontrar métodos que abordam estimação de valores ausentes em séries temporais, executou-se uma Revisão Sistemática da Literatura (Systematic Literature Review – SLR), a qual define um conjunto de regras para buscar em repositórios digitais artigos publicados sobre um determinado tema em estudo (KITCHENHAM et al., 2009).

A revisão sistemática é realizada definindo, inicialmente, as perguntas principal e secundárias. Em seguida, um conjunto de palavras-chave é escolhido para auxiliar na busca por artigos relacionados ao tema abordado. Após essa fase de busca, inicia-se uma fase de leitura e análise dos artigos coletados visando responder a pergunta principal e as secundárias.

O produto final de uma SLR permite realizar diversas análises sobre o estado da arte de um determinado tema de pesquisa como, por exemplo, o número de publicações anuais e principais autores.

É importante destacar que, neste trabalho, a SLR foi realizada com apoio da ferramenta StArt desenvolvida pelo Laboratório de Engenharia de Software da Universidade Federal de São Carlos[1] (FABBRI et al., 2016). Nas seções a seguir, todas as fases da Revisão Sistemática da Literatura realizada neste projeto são apresentadas em detalhes.

3.2 FASE I - BUSCA E COLETA DE ARTIGOS

Na primeira fase, é definida todas as características da busca a ser realizada na SLR: repositórios de busca, palavras-chave, string de busca, idioma padrão, perguntas principal e secundárias, critérios de inclusão e exclusão e os passos de execução.

Logo, para esta SLR foi definida a seguinte pergunta principal da pesquisa:

Quais são as técnicas comumente adotadas para resolver o problema de dados ausentes em séries temporais?

[1]Disponível em ⟨http://lapes.dc.ufscar.br/tools/start_tool⟩. Último acesso: 26 de setembro de 2019.

Após definida a pergunta principal, um conjunto de perguntas secundárias diretamente associadas à pesquisa também foram definidas com o intuito de validar a pergunta principal. O objetivo desse conjunto é avaliar de modo mais direto e detalhado todos os trabalhos encontrados dentro do tema analisado.

PS.1 - Quais tipos de aplicações práticas podem tirar vantagem ao usar tais técnicas?
PS.2 - Como tais técnicas são avaliadas?
PS.3 - Por que os autores usam tais técnicas?
PS.4 - Quais são as limitações das técnicas?
PS.5 - Qual a frequência de artigos publicados por ano?
PS.6 - Quem é o pesquisador principal na área?
PS.7 - Existe alguma evidência que o assunto estudado foi limitado por falta de estudos primários?

Em seguida, definiu-se um repositório online que é comumente utilizados pela comunidade científica:

- Scopus (https://www.scopus.com/home.url)

Decidiu-se também pelo uso do Inglês como idioma padrão, logo, os artigos escritos em outros idiomas foram descartados. No próximo passo, definiu-se as palavras-chave considerando a pergunta principal e o objetivo do estudo:

- *time series*
- *gap filling*
- *missing data*

Baseando-se nas palavras-chave, criou-se uma *string* de busca que ao ser inserida no campo de busca do repositório permite recuperar todos os artigos que podem estar relacionados ao tema estudado. A string de busca definida foi:

```
(‘‘time series’’) AND (‘‘gap filling’’ OR ‘‘missing data’’)
```

Mesmo após a escolha das palavras-chave e string de busca, muitos artigos recuperados podem não estar relacionados à pesquisa. Então, para o próximo passo, visando atingir o maior número possível de documentos que são relevantes para a análise, definiu-se critérios de inclusão e exclusão para selecionar tais documentos. Foram incluídos na análise apenas os documentos que perfizeram as seguintes condições:

- O trabalho lida com ausência de dados em séries temporais?
- É um estudo primário?

Após aplicados os critérios de inclusão, os seguinte critérios de exclusão foram definidos:

- O trabalho apresenta uma técnica que é muito restrita a um problema específico
- O trabalho não apresenta um modelo analítico bem definido
- A avaliação da técnica não é satisfatória
- O trabalho não tem uma revisão literária satisfatória

Assim, o artigo que se encaixou em um ou mais critérios de exclusão foi descartado.

Se uma referência de algum artigo for julgada interessante para revisão, ela também pode ser adicionada trazendo mais robustez aos estudos.

3.3 FASE II - ANÁLISE DOS ARTIGOS SELECIONADOS

Após aplicar a string de busca no repositório escolhido, 615 artigos foram encontrados onde alguns deles foram classificados como duplicados e, assim, retirados da fase de análise. Nesta fase, lê-se título e resumo de todos os artigos individualmente e aplica-se os critérios de inclusão para separar os que não estavam dentro da abordagem desejada. Como consequência, 156 foram selecionados para serem lidos e avaliados aplicando-se os critérios de exclusão. Finalmente, 21 documentos foram selecionados como mais relacionados com o tema avaliado. A Tabela 3.1 apresenta a classificação final após a utilização dos critérios de inclusão e exclusão.

Tabela 3.1 Número de artigos após cada critério.

Critérios	Número de papers
Total de papers	615
Remoção de duplicados	609
Inclusão	156
Exclusão	21

Após a leitura de cada documento os mesmos foram organizados em um apêndice que destaca o Autor, Contexto, Técnicas, Métodos, Objetivo e Avaliação. Além dessas informações, ainda apresenta-se na Tabela 3.2 a distribuição dos artigos por país de seus autores. Em relação à pergunta secundária PS.6, a SLR não pôde apontar tal informação, pois nenhum dos autores apareceu em mais de uma publicação dos artigos selecionados.

A Tabela 3.3 apresenta o número de artigos publicados por ano até a data em que a revisão sistemática foi realizada, mostrando que 2010 foi o ano onde publicou-se mais novas técnicas de substituição de dados ausentes. Esse tipo de informação enfatiza o quanto o tema abordado tem sido estudado ao longo dos anos, respondendo à pergunta secundária PS.5.

Com o uso da ferramenta de apoio - StarT - para a organização dos documentos, outra métrica para a análise se fez interessante: a pontuação que cada documento recebe de acordo com a presença de palavras-chave ao longo do paper. Foi estabelecido palavras-chave encontradas no título somam 7 pontos, as encontradas no resumo somam 3 e nas

Tabela 3.2 Número de publicação por país de autores.

País	Número de papers
Itália	4
Estados Unidos	3
China	3
França	2
Alemanha	2
Portugal	1
Brasil	1
Tailâdia	1
Coréia do Sul	1
Turquia	1
Russia	1
Espanha	1

palavras-chave do documento somam 2. A Tabela 3.5 apresenta os artigos com maior pontuação. A Tabela 3.4 apresenta o tipo de publicação dos artigos selecionados.

Alta pontuação não indica que o artigo é mais relevante na revisão, apenas que alguns artigos podem estar mais relacionados ao tema. Do mesmo modo que baixa pontuação também não indica que o artigo está fora do tema. Por exemplo, dois artigos com pontuação 0 passaram pelos critérios de seleção.

Além dessa pontuação, os trabalhos selecionados receberam uma qualificação de prioridade de leitura sendo essas: Muito baixa, baixa, alta e muito alta. Conforme apresentado na última coluna da Tabela 3.6. A Tabela 3.7 referencia seu identificador com o ID da Tabela 3.6.

Tabela 3.3 Número de publicações por ano.

Ano	Frequência
2015	1
2014	1
2013	2
2012	2
2011	1
2010	4
2009	2
2008	2
2007	3
2006	2
2003	1

Tabela 3.4 Relação de número de publicações e tipo.

Tipo	Frequência
Journal	14
Conferência	7

Tabela 3.5 Autores, ano e maiores pontuações.

Autor	Ano	Pontuação
Verger et al. (2013)	2013	38
Facchini e Mocenni (2011)	2011	23
Huo et al. (2008)	2008	22
Golyandina e Osipov (2007b)	2007	19
Junger e Leon (2015)	2015	18
Cano e Andreu (2010)	2010	18
Kim e Pachepsky (2010)	2010	17

Tabela 3.6: Sumário dos artigos selecionados

ID	Contexto	Técnica	Método	Objetivo	Avaliação	P	R
1	Método baseado em regressão linear múltipla e é utilizado para reconstrução de dados de temperatura gravadas diariamente em múltiplas estações de uma região	Regressão Linear Múltipla e Imputação	identifica-se dois grupos de estações vizinhas que podem ser usadas para a reconstrução, uma para o período que precede o gap e outro para o que sucede o gap, seleciona-se o período a ser considerado, identifica-se qual sub-conjunto de estações tem a melhor correlação com a estação do gap alvo e identifica-se o menor tamanho da amostragem que minimiza a reconstrução do erro	Identificar qual o melhor conjunto de estações e períodos para minimizar o erro de estimação da reconstrução	Comparação de resultados com os resultados de uma de suas referências e com inspeção visual em erros de reconstrução	16	MA
2	Propõe-se um novo método baseado em Fuzzy Similarity para reconstrução de missing data.	Fuzzy Similarity	Aplicado em um modelo de monitoramento de condições de componentes industriais online por modelos empíricos orientados a dados. A técnica calcula uma medida de similaridade difusa de um segmento de uma série temporal que contém gap e uma série temporal usada como referencia sem gaps, associa-se um peso a cada segmento de referencia e então reconstroem-se os valores como uma média ponderada do segmentos de referencia.	O objetivo do método é preencher lacunas geradas por falhas de sensores de monitoramento em tempo real para que os dados registrados já sejam guardados sem erros e não gerem informações incorretas sobre as condições dos componentes.	O método é avaliado de acordo com sua acurácia na reconstrução dos pontos, usando a métrica Mean Square Error (MSE) entre as reconstruções e os valores reais.	8	A
3	Observação de oscilações semelhantes às solares para sondar o interior de estrelas.	Recurrent Neural Network e Wavelet Decomposition	A técnica baseia-se em duas redes neurais (Recurrent neural network - RNN) com mesma topologia em seu número de neurônios mas treinadas separadamente, onde a primeira é treinada para prever pontos a frente enquanto a segunda é treinada para prever pontos atrás obtendo assim uma reconstrução do tipo *forward and backward*, inicialmente a técnica não apresentava tanta precisão até que adicionou-se como entrada uma decomposições wavelet dos sinais. Essa técnicas é capaz de reconstruir sinais em coeficientes wavelet e também de prever esses coeficientes para a predição dos sinais.	O trabalho tem como objetivo a reconstrução de observações fotométricas para identificar as reais oscilações das frequências de estrelas.	A avaliação do método é feita sobre uma simulação usando uma série do telescópio Kepler sobre uma estrela observada por um período de um mês.	0	A
4	Propõe a aplicação de uma nova metodologia para reconstrução de séries temporais de dados ecológicos através de gap filling.	Twin surrogates	Para a reconstrução: É necessário um preprocessamento para normalizar a série, deve-se também testar se as duas séries temporais que demonstram dinâmicas suficientemente iguais antes do gap; Na série onde existe o gap, identifica-se dois conjuntos de regiões onde um corresponde ao gap e o outro a um segmento posicionado antes do gap para se dar continuidade a fase de gap filling.	Substituição de valores ausentes em dados de oxigênio dissolvido na lagoa de Ortobello. O autor faz testes com séries temporais reais e séries temporais caóticas	MSE e inspeção visual	23	MA
5	Aborda uma novo método de gap filling utilizando Radial Basis Function Neural Networks.	Radial Basis Function Neural Networks	Para esse método, um sistema complexo de redes modulares de duas camadas foi desenvolvido para preencher os dados ausentes. Esse procedimento consiste em duas partes principais, construção de uma topologia de rede perceptron multicamadas (MLP - Multi-layer Perceptron) e uma série temporal química meteorológica foi usada como entrada para treinamento e otimização de duas redes RBF para gap filling de vapor de água turbulenta e dados de fluxo de CO2.	Esse trabalho tem como objetivo a substituição de valores inválidos em séries de fluxo de vapor de água e avaliação da qualidade do CO2 em áreas urbanas.	A avaliação do método é feita por inspeção visual de tabelas, correlações e outras informações diretas citadas pelo autor.	13	MA

ID	Contexto	Técnica	Método	Objetivo	Avaliação	P	R
6	Propõe um novo método para gap filling utilizando de processos autorregressivos (AR) e estimação de parâmetros através de Expectation Maximization (EM)	Expectation Maximization.	O método baseia-se em modelar séries temporais de oscilações solares por processos estocásticos autorregressivos (AR) e em seguida estima-se os parâmetros com EM que é baseado em 2 passos: Passo de expectativa (E-step), onde as variáveis ocultas são estimadas assumindo que os parâmetros já são conhecidos, e o Passo de maximização, onde as variáveis ocultas são literalmente usadas e os parâmetros são corrigidos. O gap filling é realizado nessa fase (M) e é feito de forma iterativa, então a série resultante da primeira estimação deve ser usada para uma segunda estimação que deve melhorar os resultados até se atingir um resultado bem avaliado.	O trabalho tem como objetivo reconstruir séries temporais Heliosismológicas para a análise de ondas de pressão no Sol.	A avaliação do método foi baseado em dados simulados usando oscilações de frequência adquiridas pela rede BISON como entrada pra simular séries temporais com e sem gaps.	6	A
7	Apresenta uma modificação do algoritmo de imputação de missing values baseado no reuso de dados estimados: weighted K-NN imputation (k-nearest neighbours).	Weighted K-NN	O método, baseado em iterações, logo foi nomeado de IKNNimputation, primeiramente substitui todos os MV (missing values) por valores obtidos atráves de médias e obtém-se uma primeira matriz completa. No próximo passo, constrói-se uma matriz de genes-candidatos, computa-se a distância euclidiana entre todos os candidatos, imputa-se o MV através de média ponderada dos níveis de expressão dos K genes mais próximos. Depois da imputação de todos os MV tem-se uma matriz completa, então soma-se as diferenças ao quadrado de matriz construída e das iterações anteriores, o algoritmo para quando atingir o critério de convergência.	O objetivo do trabalho é substituir observações inválidas de genes considerados inválidos em microarranjos de DNA ou mRNA. Monitoramento simultâneo de genes.	A avaliação do método foi feita por vários métodos tais como, NRMSE, coeficientes de correlação entre valores reais e estimados e o método também foi comparado com outros métodos de estimativa baseada em cluster.	13	MA
8	Propõe um método para imputação de missing data utilizando o algoritmo EM (expectation maximization).	Expectation Maximization	O método baseia-se em estimar o vetor médio e uma matiz de covariância de uma distribuição normal multivariada com missing data. Ele se resume em três passos: substituir os MV através de estimativas, estimar os parâmetros de média e covariância, estimar o nível para cada série temporal uni-variada, re-estimar os MV usando as estimativas atualizadas dos parâmetros e dos níveis da série temporal. Esses passo se repetem até que um critério de convergência seja atingido.	O objetivo do trabalho é ser aplicado para reconhecer os efeitos de poluição do ar na saúde, conceito epidemiológico.	Inspeção visual e comparação com outros métodos e modelos, tal como NN e ARIMA.	18	MA
9	Propõe uma nova técnica para imputação de MV utilizando um método de Bootstrap e uma imputação proposta no artigo.	Imputation with Bootstrap	Autor considera 4 tipos diferentes de situações em que os gaps podem ser encontrados e para cada uma delas um conjunto de dados de Bootstrap é gerado e usado para a imputação. Essas quatro situações remetem ao crescimento ou decrescimento da curva em que gaps são encontrado e também ao tamanho do gap. Após uma série de definições de localidade dos gaps ao longo da série e dos dados sem gap o processo realiza a checagem do crescimento ou decrescimento da curva através do cálculo de tangentes. Após algumas verificações, decide-se qual algoritmo será utilizado para realizar a imputação.	O objetivo do trabalho era descrever essa nova técnica e compara-las com outras já existentes.	Para a avaliação do método o autor usa de uma abordagem MCAR com gaps que vão de 10% a 70% variando em 10% e compara-se com outros 3 métodos: Splines Cúbicas, MI interpolation e Varies Window Similarity Measure (VWSM).	15	A

ID	Contexto	Técnica	Método	Objetivo	Avaliação	P	R
10	Apresenta um novo modo de preencher MV através de métodos inovativos (TES e TESWN). Esse método baseia-se na técnica TES (Two-directional Exponecial Smoothing) e TESWN (TES with White Noite), o método TES estima MV baseando-se em autocorrelações para contar com fato que os MV podem ocorrer de forma não aleatória.	TES and TESWN	Para a reconstrução, primeiro gera-se um conjunto de dados usando um método de média, depois uma ES para frente para preencher os MV e depois um ES para trás e, assim, os valores recolocados serão a média dos valores preditos pelas duas iterações anteriores (Forward and Backward). O método TESWN diferencia-se de modo que ele adiciona um ruído branco para contar com efeitos estocásticos.	Gap filling utilizado no controle da qualidade da água.	O método de avaliação desse método é feito com uma base de dados com observações quase completas em um período de três dias, então remove-se intencionalmente pontos para testar a qualidade do método.	22	MA
11	Apresenta o método Integrative Missing Value Estimation (iMISS) que utiliza múltiplos datasets de referencia para melhorar estimação de MV aplicado em tecnologias de microcroarrays com dados genéticos	Integrative Missing Value Estimation, k-NN, LLS imputation	O algoritmo do iMISS procura pelo melhor dataset de referência que pode ser usado dado microarray. Para isso ele propõe um método de imputação de submatriz. A performance do algoritmo pode ser influenciada pelo número de datasets utilizados como referencia, pela tolerância ao nível de ruído e pelo número de amostras.	Apresentar novo método de gap filling que apresenta melhoria sobre o LLS (Local Least Square)	Para avaliar o método, utiliza-se da métrica RMSE que considera os valores reais e os valores imputados.	0	A
12	Apresenta um novo método baseado em PPCA (Probabilistic Principal Component Analysis - Análise de componentes principais probabilísticos) para imputação de MV em dados de fluxo de volume de tráfego.	PPCA and MLE	A técnica PPCA contém duas ferramentas principais: Análise de componentes principais e Maximum Likelihood Estimation (MLE). O PPCA tende a separar a parte significante do fluxo do tráfego onde o MLE é aplicado para estimar o MV e construir, assim, um modelo regressivo latente. O autor cita que esse método pode ser considerado como uma reformulação do algoritmo de maximum-likelihood e ele também cita o uso do algoritmo EM.	Provar que o PCCA apresenta um melhora sobre os métodos convencionais reduzindo o RMSE em no mínimo 25%	O método é avaliado com a métrica do RMSE e é comparado com outros outros métodos: Historical Imputation e Splines Cúbicas.	5	MA
13	Introduz um novo método onde uma RBFNN foi desenvolvida para ser o estimador da reconstrução de missing data.	Radial Basis Function Neural Network	Utiliza-se de uma rede neural de 3 camadas, uma camada de entrada, uma segunda (hidden layer) onde se encontram as funções de ativação de cada unidade, e uma terceira camada que é uma única unidade linear que está completamente conectada a camada oculta. Utiliza uma série de muitas observações (30-anos Zooplankton study) e foi necessário o particionamento em grupos dessas observações.	Apresentar o potencial do uso de neural networks para reconstrução de missing values	Para avaliação do trabalho o autor utilizado do MSE e compara o resultados com outros dois método : Back-propagation based neural networks e Splines Cúbicas. O autor afirma que seu método tem melhores resultados que os citados.	7	A
14	Mostra uma nova técnica para reconstruir missing data em precipitações (chuva) diárias. Os dados utilizados foram coletados em diferentes estações de observações, ou seja, cada estação faz uma observação diferente ao longo do tempo.	Bootstrap Resampling Method, Artificial Neural Netwoks e Regression Trees	Faz-se uso de um algoritmo de Bootstrap (Boostrap Resampling Method) para selecionar qual estação vizinha deve ser usada para a reconstrução e para avaliação estatística da precisão da reconstrução de todos os 3 métodos (RT, ANN e ANN + RT).	O objetivo desse trabalho é desenvolver uma técnica para reconstrução dos MV combinando duas outras técnicas : RT (Regression Trees) e ANN (Artificil neural network).	A técnica foi avaliada através de coeficiente de correlação e RMSE	17	MA
15	Propõe um novo algoritmo para prever MV baseado em RBFNN (Radial Basis Function). A RBF funciona de acordo com o padrão, sendo 3 camadas, uma de entrada, uma camada oculta e uma de saída.	Radial Basis Function Neural Networks	O método se resume em dividir a série em segmentos onde cada um desses segmentos contém alguns MV, usar os dados de cada segmentos para gerar uma equação artificial da série temporal (encontrar os pesos na aproximação da RBF), calcular o erro em cada segmento, calcular a soma dos erros quadráticos de cada segmento e substituir os valores de acordo com o valor mínimo da somas dos quadrados.	Apresentar novo método de gap filling com RBFNN	Inspeção visual com valores originais e estimados	12	MA

ID	Contexto	Técnica	Método	Objetivo	Avaliação	P	R
16	Mostra uma modelagem para reconstrução de MV utilizando de MSFD (Multiple Sine Function Decomposition) em médias de temperatura mensal.	Multiple Sine Function Decomposition	O método é descrito da seguinte forma: Primeiramente, utiliza-se uma função seno para aproximar os valores dos dados existentes e, então, três parâmetros da função seno são obtidos. Subtrai-se os valores aproximados dos valores alvos e armazena-se os resultados dos residuais como os novos alvos para a próxima função de decomposição senoidal, esse processo é repetido de acordo com um parâmetro. Ao final desses processo, o modelo resultante deve ser usado para reconstruir diretamente os MV.	Apresentar um abordagem iterativa de aprendizado de máquina para construção de um modelo de gap filling	Para a avaliação do modelo, utiliza-se de três métricas: SMAPE (Symmetric Mean Absolute Percentage Error), RMSE (Root Mean Squared Error) e a MRE (Mean Relative Error).	0	A
17	Propõe um novo método de imputação única e o compara a outros métodos de imputação única e multipla já conhecidos.	Site Dependent Effect Method e Imputation	Utiliza-se de observações de intervalo de duas horas da concentração de PM10 medidas em oito estações diferentes. O método proposto, chamado SDEM (Site Dependent Effect Method), utiliza 3 médias diferentes baseadas nas observações dos efeitos semanais, diários e por hora, estimando o MV de acordo com as médias das oito estações. O autor também utiliza um modelo baseado em imputação múltipla pra incluir uma incerteza e dar naturalidade à variação dos dados.	Apresentação de um novo método de gap filling baseado em observações de múltiplas estações.	Para a avaliação de performance do método foram usados: coeficiente de correlação, um índice de aceitação, RMSD (Root Mean Square Deviation) e MAD (Mean Absolute Deviation).	2	B
18	Apresenta uma abordagem de reconstrução de MV baseada em suportes de regressão vetorial aplicada em microarrays de expressão genética. Utilizando Framelet Based Kernels para aproximar funções com estrutura multiescalar para suavização de dados ruidosos.	Regressions Arrays e Framelet Based Kernels	O autor cita que o algoritmo trabalha com aprendizado através de exemplos, utilizando medidas válidas para fazer a reconstrução, ainda mostra que resolve problemas de overfitting usando um princípio minimização de erros. O kernel usado foi o WMFK (Weighted Multiscale Framelet Kernel).	O objetivo do trabalho é mostrar a melhoria na performance da reconstrução usando as técnicas citadas.	Para a avaliação, mediu-se a precisão do método comparando os dados recuperados com os originais através da métrica NRMSE (Normalized Root Mean Square Error) e comparando com outras técnicas: Interpolação Linear e Splines Cúbicas.	5	A
19	Propõe uma nova técnica para suavizar e preencher gaps em séries temporais LAI (Leaf Area Index) derivadas de observações espaciais.	Neural Networks	Esse método é baseado em padrões sazonais dos pixels. Utiliza de modelos fenológicos e usa essas informações para realizar o gap filling aproveitando também da climatologia do ambiente observado. Utiliza-se uma série de 20 anos de observações para testar a abordagem sob várias condições. Redes neurais também são utilizadas nas observação, mas não diretamente na parte de reconstrução.	Apresentar método de gap filling em séries de informações fenológicas (Índice de Área de Folhas)	Para avaliar o método utiliza-se de RMSE. A performance foi avaliada contra os métodos AG(Asymmetric Gaussian) e SG(Savitzky-Golay).	38	MA
20	Aplica SSA (Singualar Spectrum Analisys) em séries temporais que apresentam MV e propõe um método para preencher dados inválidos.	Singular Spectrum Analisys	O algoritmo proposto extrai componentes de uma séries temporal e realiza o gap filling simultaneamente. O autor faz o uso do Cartepillar-SSA e o modifica para tratamento de missing data. O resultado da aplicação do Cartepillar-SSA é uma soma de componentes de uma série temporal, mas para tratar missing data ela terá dois estágios: decomposição e reconstrução. No estágio de reconstrução realiza-se a escolha de um subespaço, projeção de *lagged vectors* incompletos, projeção de *lagged vectors* incompletos e a medida de uma média diagonal.	Descrição formal completa do método e modificações para vários tratamentos	Para exemplificar e testar a técnica utilizou-se de uma série real e implantou-se gaps artificiais. Aparentemente não usou nenhuma métrica ou comparação com nenhum outro método/técnica.	19	MA

ID	Contexto	Técnica	Método	Objetivo	Avaliação	P	R
21	Propõe uma nova técnica para lidar com MV baseada em Multiple Imputation (MI). O trabalho aborda um novo ponto de vista para MI em séries temporais que lida com o efeito da distancia e ainda evita ruídos nas simulações.	Multiple Imputation	O trabalho segue os princípios de Markov (Markov chain) e os algoritmos MCMC (Markov Chain Monte Carlo). O processo de MI é composto de 3 estágios: Imputaçao, onde o número de imputações é definido e a distribuição probabilística é aproximada de acordo com o número de MV e de valores observados usando os algoritmos MCMC; Análise, todos os dados simulados são analisados usando método padrões; Tanque (Pool), nesse ponto vários resultados já foram gerados, então eles são combinados com regras de inferência.	Descrever toda a teoria do método assim como pontos fortes e implementação.	A avaliação é feita por meio de 5 testes em simulações utilizando a linguagem R.	18	MA

Tabela 3.7 Tabela de referência do sumário

Identificador	Referência
1	(TARDIVO; BERTI, 2012)
2	(BARALDI et al., 2013)
3	(CAPIZZI; NAPOLI; PATERNò, 2012)
4	(FACCHINI; MOCENNI, 2011)
5	(SCHMIDT; WRZESINSKY; KLEMM, 2008)
6	(ROTH; ZHUGZHDA, 2010)
7	(BRáS; MENEZES, 2007)
8	(JUNGER; LEON, 2015)
9	(PRASOMPHAN; LURSINSAP; CHIEWCHANWATTANA, 2009)
10	(HUO et al., 2008)
11	(HU et al., 2006)
12	(QU et al., 2009)
13	(HONG; CHEN, 2003)
14	(KIM; PACHEPSKY, 2010)
15	(UYSAL, 2007)
16	(ZHANG et al., 2014)
17	(PLAIA; BONDI, 2006)
18	(ZHANG; LIU; YAN, 2010)
19	(VERGER et al., 2013)
20	(GOLYANDINA; OSIPOV, 2007a)
21	(CANO; ANDREU, 2010)

Toda área que exija armazenamento e análise de uma série temporal está sujeita ao impacto negativo causado por perda de dados. Em resposta a PS.1, Tardivo e Berti (2012) apresentam uma aplicação que utiliza de uma abordagem baseada em regressão para reconstrução automática de dados de temperatura utilizando diversas estações de coleta de dados. Já Roth e Zhugzhda (2010) utilizam, substituição de valores ausentes baseado na modelagem de dados de oscilações solares que são constantemente afetados por erros levando assim a uma má interpretação espectral. Brás e Menezes (2007), usam de técnicas de substituição de dados ausentes em microvetores de DNA que guardam informações sobre níveis de RNA mensageiros de milhares de genes de diferentes células e tecidos. Nota-se que técnicas de substituição de dados ausentes são constantemente utilizadas em muita áreas de pesquisa em que opta-se pela análise de series temporais.

Em resposta a PS.2, a avaliação das técnicas de substituição de dados ausentes é feita da seguinte forma: i) escolhe-se uma série temporal; ii) retira-se intervalos de valores da série original, que são guardados para serem comparados com os novos valores estimados; iii) aplica-se a técnica proposta sobre a série escolhida; iv) compara-se os valores gerados pela técnica com os valores originais. Todos os artigos encontrados nessa revisão utilizam desse método para avaliar as técnicas descritas e, também, utilizam de métricas de avaliação de erro tais como Mean Square Error (MSE) e Root Mean Square Error (RMSE).

Além da comparação com valores originais, é comum verificar os resultados e compará-los com outras técnicas conhecidas na literatura. Por exemplo, Qu et al. (2009) utiliza do RMSE e confronta sua técnica com Splines Cúbicas e com a Imputação Histórica. Existem alguns autores que optam apenas pela inspeção visual da série para comparar os resultados estimados com os originais.

Um dos motivos que leva profissionais e pesquisadores a lidar com valores ausentes e não simplesmente ignorar-los, é o fato de que existem técnicas utilizadas sobre séries temporais que não podem ser executadas se existirem valores ausentes (NA) no corpo da série. Valores ausentes também geram inconsistência em banco de dados. Por exemplo, como assumir que determinado mês foi o mais chuvoso durante todo o ano se 20% das observações foram perdidas e estão registradas como NA? Então, respondendo a PS.3, existem casos que a estimação dos valores ausentes se faz necessária. Baraldi et al. (2013) apresentam uma abordagem onde monitora-se as condições e o estado de componentes industriais através de modelos empíricos aplicados a usinas nucleares. O artigo ainda exemplifica: "Se um sensor falhar ao prover um valor de entrada a série, o modelo pode não ser capaz de inferir o estado do componente. Sendo assim, é importante que as informações de entrada estejam completas para o treinamento do modelo e para seu uso.". Outro contexto onde não pode existir valores ausentes é na decomposição de séries temporais, muito bem apresentado por Golyandina e Osipov (2007a) que utiliza da Decomposição de Valor Singular (SVD) para decompor uma série em componentes aditivos.

Contudo, existem dois problemas que causam menor acurácia nos valores estimados, ou seja, limitam a utilização das técnicas: o tamanho dos intervalos de valores ausentes e a frequência em que aparecem. Por outro lado, Cano e Andreu (2010) mostram que a qualidade dos dados disponíveis afeta sua técnica baseada em Imputação Múltipla, ou seja, as limitações variam entres as técnicas e afetam também o peso que cada um delas tem no resultado da estimação dos valores. Cano e Andreu (2010) mostram ainda que o erro cresce conforme cresce o tamanho do intervalo de valores ausentes e que a taxa do aumento do erro pode ser representada por uma função exponencial. Sendo assim, quanto mais intervalos, quanto maior o tamanho dos intervalos e dependendo da posição dos intervalos os erros serão altos. Respondendo assim a PS.4, o tamanho dos intervalos de valores ausentes, a frequência em que aparecem e a posição onde aparecem é o que mais limita as técnicas. Comumente, para avaliar um novo método, os testes são preparados com variações das limitações citadas.

Em resposta a PS.7, os artigos encontrados permitiram realizar um levantamento bibliográfico sobre técnicas e métodos de substituição de valores ausentes, detalhando o estado da arte atual e as principais técnicas utilizadas.

3.4 FASE III - CONCLUSÃO

Esta seção discute brevemente as técnicas mais utilizadas dentre os artigos que passaram pelos critérios de inclusão e exclusão.

3.4.1 Redes Neurais Artificiais (RNA)

Redes Neurais Artificiais ganharam muita popularidade nos últimos anos em pesquisas de matemática teórica e também em ciências aplicadas, elas podem ser usadas para examinar relações desconhecidas entre variáveis ou resolver problemas onde não existe solução analítica (RIPLEY, 2006; SCHMIDT; WRZESINSKY; KLEMM, 2008). A principal vantagem do uso das Redes Neurais vem do seu poder de reconhecimento de padrões que, aplicado a um conjunto de dados, a rede pode ser treinada para reduzir erros na estimação dos dados. Schmidt, Wrzesinsky e Klemm (2008) utilizam Redes Neurais Artificiais para minimizar o desvio entre a saída da rede e os valores reais. Redes Neurais Artificiais são compostas por três camadas principais, uma camada de entrada (input layer) uma camada escondida (hidden layer) e uma camada de saída (output layer). A camada oculta pode ser composta por várias outras camadas. Todas as camadas de uma rede neural são compostas por neurônios (*neurons*) que podem ou não ser ativados de acordo com o conjunto de entrada e com a etapa de treinamento de uma rede neural.

Dentre os artigos estudados na Revisão Sistemática, encontrou-se métodos que utilizavam, por exemplo, de Funções de Base Radial (Radial Basis Functions - RBF) com Redes Neurais (SCHMIDT; WRZESINSKY; KLEMM, 2008; UYSAL, 2007; HONG; CHEN, 2003). Ao utilizar essas redes, cada neurônio da camada de saída passa a ser uma combinação linear de n funções de base radial, ou seja, cada neurônio da camada intermediária passa a ter uma RBF como função de ativação. Além de RBF ainda foram encontrados métodos com Redes Neurais Recorrentes (CAPIZZI; NAPOLI; PATERNò, 2012), e com Árvores de Regressão (AR + RNA) (KIM; PACHEPSKY, 2010).

3.4.2 Imputação

Imputação é um técnica que preenche valores ausentes por valores substitutos. Diferente das Redes Neurais, uma imputação não usa de funções complexas ou interpolações para substituir esses valores ausentes. Existem situações em que valores ausentes são fáceis de ser reconstruídos por trazer apenas um comportamento linear. Nesses casos, apenas uma média dos valores vizinhos pode resolver o problema, ou até repetir o último valor encontrado. A imputação também pode utilizar da correlação entre variáveis de tempo e espaço para estimar os valores.

Plaia e Bondi (2006) abordam imputação em um contexto onde existem várias zonas de observação ambientais e os valores ausentes são tratados de forma que, dada uma série temporal sendo armazenada por uma determinada estação, os valores ausentes serão calculados com médias semanais, diárias e horárias entre as estações mais próximas. Este método foi chamado de *Site-Dependent Effect Method*. Métodos de imputação são extremamente ligados ao contexto da aplicação. Outros artigos encontrados pela revisão também utilizam de técnicas de imputação e outras técnicas para adaptarem seus métodos ao problema estudado. Além dessas técnicas, Prasomphan, Lursinsap e Chiewchanwattana (2009) utilizam de imputação com Bootstrap. Junger e Leon (2015) utilizam uma modificação iterativa do algoritmo Expectation Maximization para estimar o vetor médio e uma matriz de covariância de uma distribuição normal multivariada.

3.4.3 Análise Espectral Singular

Mesmo sendo abordada apenas uma vez dentre os artigos analisados, a Análise Espectral Singular é considerada o estado da arte na substituição de valores ausentes em séries temporais. A Análise Espectral Singular (*Singular Spectrum Analisys - SSA*) é um método de decomposição de séries temporais em componentes aditivos. Essa técnica decompõe uma série temporal em vários componentes podendo eles ser sazonal, estocástico, periódico etc (GOLYANDINA; OSIPOV, 2007a). O processo de substituição de valores ausentes com SSA é formado por dois estágios: i) Descomposição; ii) Reconstrução. No primeiro estágio, a série é transformada em uma matriz de Hankel, então aplica-se uma Decomposição de Valor Singular (*Singular Value Decomposition - SVD*) sobre essa matriz.

O SVD tem como saída várias matrizes elementares. Após a aplicação do SVD, inicia-se a fase de reconstrução onde as matrizes elementares são agrupadas com base na teoria da separabilidade entre conjuntos. Nesta fase acontece a substituição dos valores ausentes, utilizando-se da combinação linear entre valores anteriores e coeficientes dos componentes extraídos pelo SSA, que acontece na forma de predição. Assim, depois de realizada a substituição, soma-se os componentes para reconstruir a série e finalizar o método.

Capítulo

4

NOVOS MÉTODOS PARA SUBSTITUIÇÃO DE VALORES AUSENTES

4.1 CONSIDERAÇÕES INICIAIS

Este capítulo apresenta os dois métodos de substituição de valores ausentes em séries temporais que foram desenvolvidos durante a realização desta dissertação. O primeiro método, apresentado na Seção 4.2, foi desenvolvido visando substituir valores ausentes no domínio temporal. O segundo método, detalhado na Seção 4.3, utiliza uma combinação de ferramentas da área de Sistemas Dinâmicos e Teoria do Caos com modelos propostos na área de Aprendizado de Máquina.

4.2 BIDIRECIONAL MEAN DISTANCE ESTIMATION (BMDE)

O método BMDE utiliza o método EMD para decompor a série temporal em componentes estocásticos e determinísticos e realizar a substituição de dados ausentes em cada componente. Após a decomposição da série temporal analisada, é necessário combinar as IMFs extraídas para reconstruir comportamentos importantes. Em estudos anteriores Rios e Mello (2016) comprovaram que a frequência máxima da próxima IMF extraída será menor que a anterior, ou seja, as IMFs são obtidas de altas para baixas frequências, confirmando a possibilidade de separar IMFs em componentes de acordo com frequência do sinal. De acordo com a área de processamento de sinais, os componentes de alta frequência correspondem a influências estocásticas, enquanto os de baixa frequência representam o determinismo. Portanto, pode-se estimar um ponto de corte entre IMFs para separar os componentes estocásticos e determinísticos (RIOS; MELLO, 2016, 2013b).

Ao considerar essa suposição, pode-se escolher um ponto de corte (k) de modo que a soma das primeiras k IMFs represente o componente estocástico, $\sum_{i=1}^{k} h_i(t)$, e a soma das demais como componente determinístico, $\sum_{i=k+1}^{J} h_i(t)$, onde J é o número de IMFs

extraídas. O resíduo resume a tendência, que pode ser analisada individualmente ou considerada como parte do comportamento determinístico.

No entanto, essa fase de decomposição não pode ser realizada quando a série temporal apresenta dados ausentes. O processo de *sifting* falha ao procurar extremos nesta situação. Para superar esse problema, criamos um método que aplica EMD separadamente em observações localizadas antes (série temporal à esquerda) e depois (série temporal à direita) da lacuna com valores ausentes. Uma vez que o EMD é aplicado individualmente nas séries temporais à esquerda e à direita, o número de IMFs extraídas pode ser diferente entre si, impedindo a aplicação de imputação e interpolação diretamente nas IMFs. Essa questão não afeta o método BMDE porque as IMFs são agrupadas de acordo com suas influências estocásticas e determinísticas.

Com base na reconstrução desses componentes, BMDE executa paralelamente dois métodos de preenchimento de lacunas. Para o componente determinístico, utiliza-se o método de interpolação Splines Cúbicas. Esse método se encaixa perfeitamente no componente determinístico porque seu comportamento aproxima-se de uma função senoidal.

Por outro lado, o componente estocástico apresenta resultados ruins usando Splines Cúbicas para estimar os valores ausentes. Esta situação motivou o desenvolvimento de um novo método. Para compreender melhor este novo método, considere a Figura 4.1 que ilustra uma série temporal com ruído aditivo. O componente determinístico foi criado somando 3 saídas produzidas por cada função seno apresentada na Equação 4.1.

$$\begin{aligned} x'_t &= \sin(\pi t) + \sin(2\pi t) + \sin(6\pi t) \\ x''_t &= \sin(\pi t) + \sin(6\pi t) \\ x'''_t &= \sin(\pi t) + \sin(6\pi t) + \sin(12\pi t) \end{aligned} \tag{4.1}$$

O componente estocástico, por sua vez, foi criado por um processo puramente aleatório seguindo uma distribuição de probabilidade gaussiana com média igual a $\mu = 0$ e desvio padrão definido pela relação sinal-ruído igual a $3,0$. Após adicionar esses componentes, remove-se todas as observações dentro do intervalo $[480, 500]$, simulando um conjunto de valores ausentes (lacuna – *gap*), conforme mostrado na Figura 4.1(a).

Na próxima etapa, utilizou-se EMD para decompor a série temporal em dois componentes: estocástico e determinístico. Da série temporal à esquerda, antes da lacuna, foram extraídas 7 IMFs, nos quais os primeiros 3 foram combinados para criar os componentes estocásticos e os restantes foram usados como determinísticos. Em relação à série temporal à direita, depois da lacuna, foram extraídos 6 IMFs e estimou-se o ponto de corte igual a 2, ou seja, os primeiros 2 IMFs eram estocásticos e os $3 - 6$ IMFs eram determinísticos.

Usando Cubic Spline, os valores ausentes no componente determinístico foram estimados como mostrado na Figura 4.1(c). A execução do método de preenchimento de lacunas proposto no componente estocástico é exemplificado na Figura 4.2. Como se pode notar nessa figura, para uma IMF estocástica, inicialmente definiu-se duas janelas de subconjuntos, contendo uma porcentagem ω de valores que devem ser utilizados nas séries temporais esquerda e direita. Neste exemplo, definimos $\omega = 10\%$ como destacado pelos dois quadrados.

Em seguida, para cada IMF estocástica, estima-se os pontos extremos nas janelas de subconjunto. Nessa figura, máximos e mínimos são plotados usando pontos vermelhos e azuis, respectivamente. Além disso, também estima-se cruzamentos no eixo x em zero (*zero-crossing*), como mostrado pelos pontos pretos nesta figura.

Na próxima etapa, estima-se o número de *zero-crossings* dentro da lacuna. Para isso, calcula-se a distância média entre *zero-crossings* em cada janelas dos subconjuntos à esquerda e à direita. Usando tal distância, estima-se a quantidade de *zero-crossings* dentro da lacuna.

Em seguida, divide-se a lacuna em duas metades. A primeira metade da lacuna usa a janela do subconjunto à esquerda para estimar os pontos até o meio da lacuna e a segunda metade usa a janela do subconjunto à direita da lacuna para estimar novos pontos na segunda metade da lacuna.

Para a primeira metade, verifica-se inicialmente o último ponto antes da lacuna. Isso irá definir qual será o próximo tipo de ponto a ser estimado (máximo ou mínimo). Por exemplo, se este último ponto for um mínimo, o próximo deverá ser um *zero-crossing* seguido de um máximo. O valor do novo mínimo ou máximo é estimado selecionando

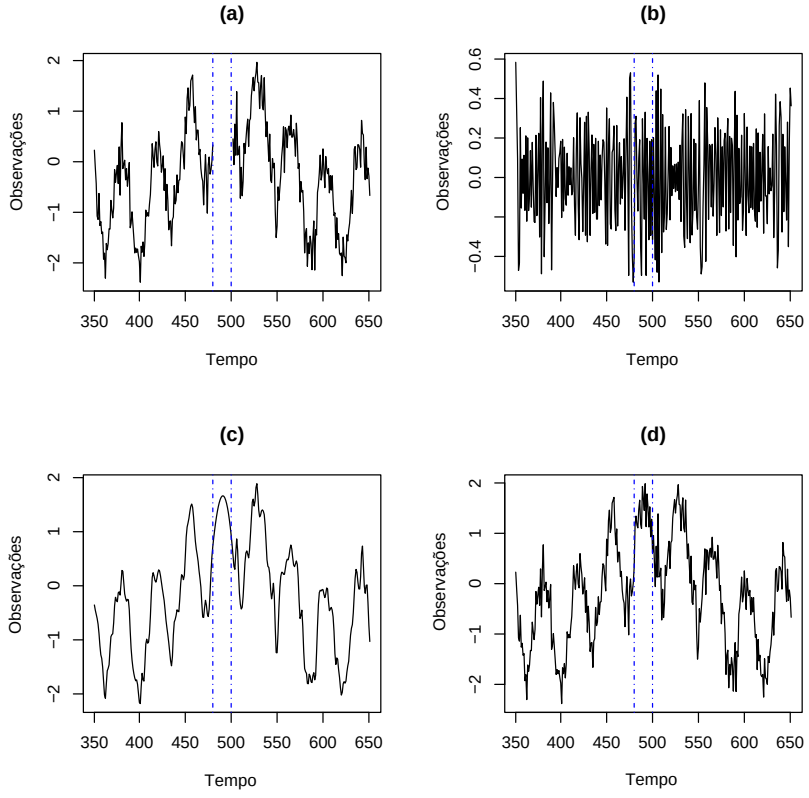


Figura 4.1 Execução do BMDE para substituição de valores ausentes. a) série temporal com dados ausentes; b) componente estocástico estimado com BMDE; c) componente determinístico estimado com Splines Cúbicas; d) Série temporal resultante do método BMDE.

aleatoriamente um valor mínimo ou máximo na janela do subconjunto à esquerda. Ou seja, se o próximo ponto é um mínimo, seleciona-se aleatoriamente um mínimo na janela à esquerda e o novo ponto terá o mesmo valor desse mínimo selecionado. O procedimento continua até a primeira metade da lacuna. Para a segunda metade da lacuna, o processo se repete utilizando o subconjunto da janela à direita. A Figura 4.3 apresenta os extremos estimados para cada janela de cada subconjunto (à direita e esquerda). Esses novos extremos substituem os valores ausentes na primeira IMF.

As etapas são executadas paralelamente em cada IMF. Em seguida, as IMFs, incluindo os novos valores, são somadas para compor o componente estocástico sem nenhum dado ausente, conforme mostrado na Figura 4.1(b).

Ao final do método, o algoritmo reconstrói a série original somando os componentes estocástico e determinístico, incluindo os valores estimados. O pseudocódigo apresentado no Algoritmo 2 mostra uma implementação simplificada do BMDE. A Figura 4.1(d) mostra o resultado final do método. Os resultados obtidos com a execução desse método foram publicados no 8th BRACIS (Brazilian Conference on Intelligent Systems).

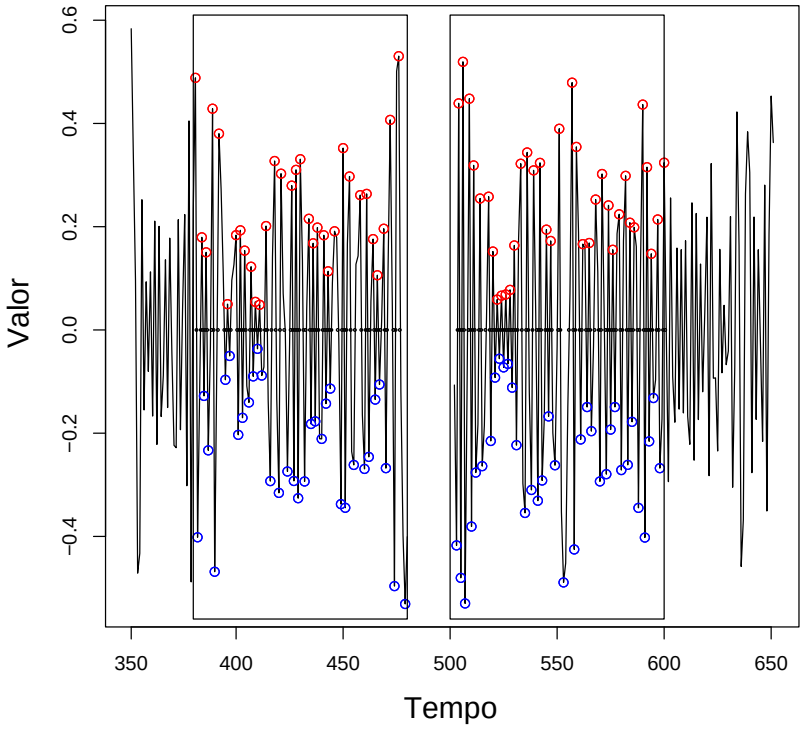


Figura 4.2 Pontos extremos e *zero-crossings* dentro das janelas do subconjunto à esquerda e à direita da lacuna de valores ausentes.

4.3 PSGF - PHASE SPACE GAP FILLING

O segundo método proposto visa realizar a substituição de valores ausentes em séries temporais caóticas considerando informações sobre os atratores e repulsores no espaço fase. De maneira geral, a execução desse método deve ser realizada em duas etapas. Primeiro, a série temporal é transformada do domínio temporal para o espaço fase. Em seguida, uma vez que a dependência temporal é removida das observações, utiliza-se técnicas tradicionais de AM para realizar a estimativa dos valores ausentes. Com isso, espera-se que a técnica de AM possa aprender o comportamento dos pontos e seja capaz de replicar tal comportamento nos valores ausentes identificados. É importante destacar que essas técnicas apresentam garantias de aprendizado apenas para dados independentes e identicamente distribuídos (MELLO; PONTI, 2018). No contexto deste trabalho, essa restrição é satisfeita, pois a série reconstruída desconsidera a dependência temporal (PAGLIOSA; MELLO, 2018).

Ao longo do desenvolvimento deste trabalho, foram considerados dois cenários que podem ser observados de acordo com o comportamento dos dados. O primeiro cenário é encontrado quando as dimensões de separação e dimensão embutida são conhecidas. Por outro lado, simulou-se situações na qual tais dimensões são desconhecidas. Estes dois

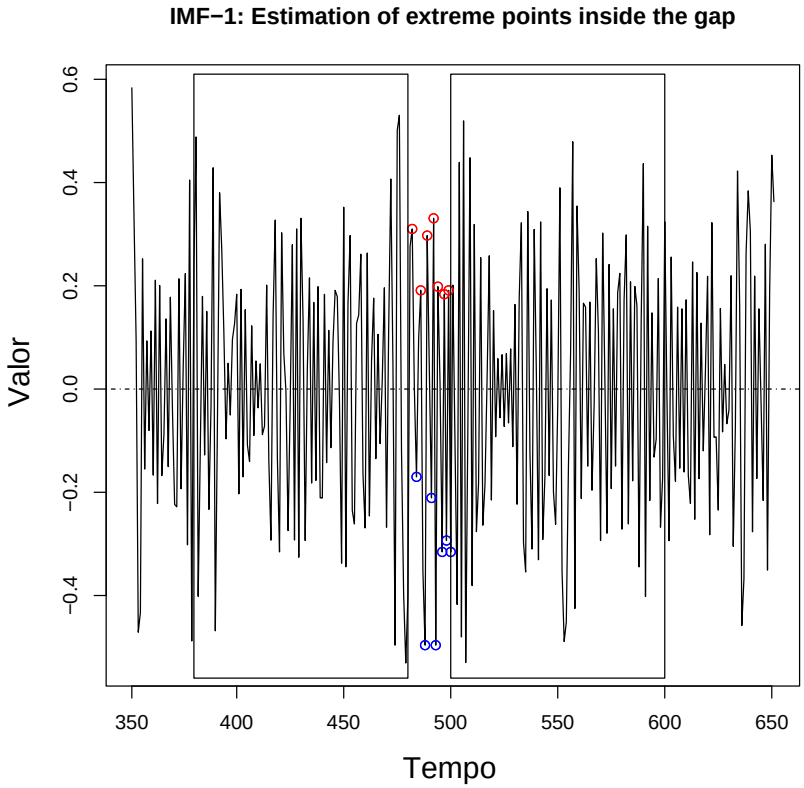


Figura 4.3 Resultado da IMF após estimar os extremos para substituir os valores ausentes.

Algorithm 2 Algoritmo BMDE.

```
1: procedure BMDE
2:     DecompEsq ← EMD(serieEsq)
3:     DecompDir ← EMD(serieDir)
4:     CompDetermEsq ← soma IMFs a partir do cutoff + residuo esquerdo
5:     CompDetermDir ← soma IMFs a partir do cutoff + residuo direito
6:     CompDeterm ← Concatena(CompDetermEsq, lacuna, CompDetermDir)
7:     FuncaoSpline ← SplineCubica(CompDeterm)
8:     for i = inicioLacuna to fimLacuna do
9:         CompDeterm[i] ← FuncaoSpline(CompDeterm[i])
10:
11:    CompEstoc ← inicia componente zerado do tamanho da série
12:    for i = 1 to cutoff do
13:        imfAtual ← Concatena(serieEsq[i,], lacuna, serieDir[i,])
14:        mediaZeroCross ← calcula distancia média entre zero crossings nas janelas
15:        ExtremosEsq ← encontra quantidade de extremos dentro janela esquerda
16:        ExtremosDir ← encontra quantidade de extremos dentro janela direita
17:        NumMedioExtremos ← Media(ExtremosEsq, ExtremosDir)
18:        DistanciaMinimaExtremos ← calcula distância mínima entre os extremos
19:
20:        for j = inicioLacuna to meioLacuna do
21:            imfAtual[j] ← Estima(j, janelaEsquerda, mediaZeroCross, NumMedioExtremos, DistanciaMinimaExtremos)
22:        for j = meioLacuna + 1 to fimLacuna do
23:            imfAtual[j] ← Estima(j, janelaDireita, mediaZeroCross, NumMedioExtremos, DistanciaMinimaExtremos)
24:
25:        CompEstoc ← Soma(CompEstoc, imfAtual)
26:
27:    if ainda existem NA em CompEstoc then
28:        FuncaoSpline ← SplineCubica(CompEstoc)
29:        for i = inicioLacuna to fimLacuna do
30:            CompEstoc[i] ← FuncaoSpline(CompEstoc[i])
31:
32:    SerieFinal ← Soma(CompEstoc + CompDeterm)
33:    return SerieFinal
```

parâmetros são os mais importantes para um bom desempenho do método desenvolvido e isso foi destacado pelos resultados experimentais abordados no Capítulo 6.

Propõe-se duas estratégias diferentes para cada cenário. Para o primeiro cenário, não é necessário estimar as dimensões, tornando assim o processo mais simples. Nesse sentido, sugere-se o seguinte processo: i) desdobramento da série para o espaço fase; ii)

treinamento da técnica de AM; iii) estratégia PSGF + AM para substituir os valores ausentes; e iv) reconstrução dos dados no espaço tempo. Ao final desse processo, espera-se obter uma série temporal sem dados ausentes. A estratégia de substituição dos dados com AM será detalhada na próxima seção.

O segundo cenário insere uma camada de complexidade: a estimação dos parâmetros para o desdobramento da série. A complexidade vem do fato que as técnicas encontradas para estimar os parâmetros não podem ser executadas com dados ausentes, logo, esta etapa deve ser inserida no processo. O escopo deste trabalho não lida com tal cenário mas, para ilustrar, existem algumas alternativas como: executar as técnicas FNN e AMI utilizado partes menores da série onde não existem valores ausentes; remover todos os valores ausentes e executar as técnicas; ou aplicar técnicas de substituição de valores ausentes no espaço tempo antes de aplicar essas técnicas.

Sugere-se o seguinte processo para o segundo cenário: i) substituir valores ausentes no espaço tempo; ii) desdobrar a série para o espaço fase; iii) remover os valores estimados no espaço tempo; iv) aplicar estratégia PSGF + AM para substituir os valores ausentes; e v) reconstruir os dados no espaço tempo.

O primeiro passo pode ser substituído por uma etapa de estimativa dos parâmetros para as dimensões e, assim, dispensando a necessidade do passo iii) de remoção dos valores substitutos estimados.

4.3.1 Processo de substituição

Esta seção descreve o principal passo do método, chamado *Propagação de Equivalência Embutida*, na qual a técnica de AM é treinada e o resultado é propagado ao longo das dimensões embutidas da série. Para ilustrar essa etapa, considere uma série temporal desdobrada de acordo com suas coordenadas de atraso com dimensão embutida $m = 3$ e dimensão de separação $d = 5$, conforme apresentado na Figura 4.4. Pode-se perceber na figura que os pontos se repetem com um intervalo d entre suas linhas e as colunas vizinhas. A ideia principal da substituição é utilizar essa informação para predizer um valor ausente (NA) em outras colunas.

Para treinamento das abordagens de AM supervisionado, considerado neste trabalho, considera-se a coluna D3 na Figura 4.4 como rótulo e remove-se todas as linhas que possuem dados ausentes da base de dados. A base sem dados ausentes é usada na fase de treinamento da técnica de AM escolhida. Neste trabalho optou-se por técnicas tradicionais de AM conhecidas na literatura: KNN (*K-Nearest Neighbors*) (SONG et al., 2017), DWNN (*Distance-Weighted Nearest Neighbors*) (YIGIT, 2013), SVR (*Support Vector Regression*) (AWAD; KHANNA, 2015) e Random Forests (BREIMAN, 2001). Tais técnicas foram encontradas, exceto a SVR, foram utilizadas em trabalhos encontrados a partir da RSL.

O próximo passo é calcular os valores ausentes ($N.A.$) na coluna D3 usando predições realizadas pelos modelos de AM. Com esse novo valor calculado, o mesmo será propagado nas colunas D2 e D1 na posição correta de acordo com a dimensão de separação definida ($d = 5$). A Figura 4.4 ilustra bem essa interação. Esse processo continua por todos os valores ausentes encontrados na coluna de rótulos. Após estimados todos os valores

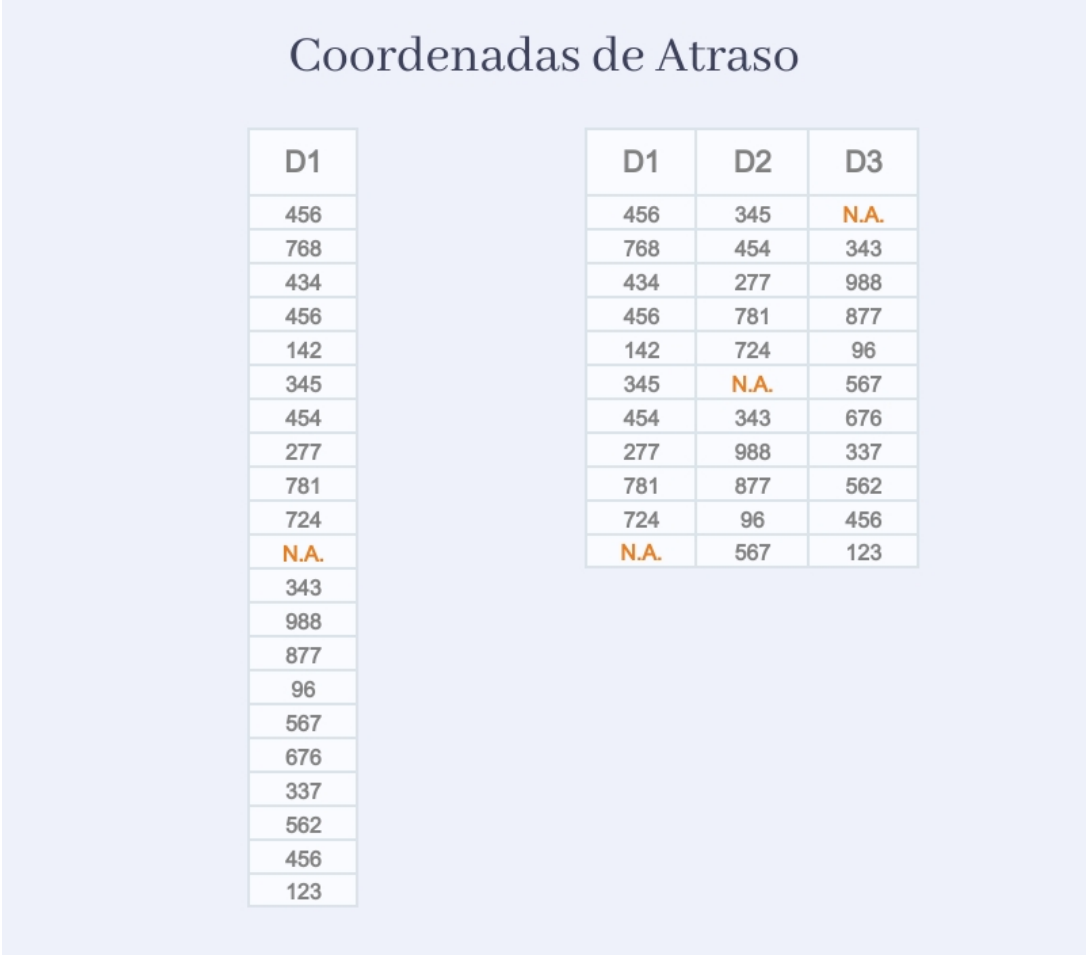


Figura 4.4 Representação do desdobramento de uma série com *m=3* e *d=5*. Utilizando a coluna D3 como rótulo para um algoritmo de aprendizado de máquina, o valor ausente (NA) pode ser calculado e replicado nas demais colunas.

ausentes a série temporal é reconstruída para o domínio temporal.

Algumas variações dessa ideia foram testadas para maiores lacunas de valores ausentes. Pensou-se, inicialmente, em um processo no qual a lacuna era dividida até sua metade e a parte superior seguia o processo acima descrito e a parte inferior da lacuna passava pelo processo de forma inversa, considerando coluna D1 como rótulo e estimando os pontos para cima (abordagem bidirecional). Outra variação é preencher toda a lacuna de valores ausentes de baixo pra cima usando a coluna D1 como rótulo (abordagem reversa). O pseudocódigo apresentado no Algoritmo 3 mostra uma implementação-exemplo do PSGF com a técnica KNN. A implementação pode variar um pouco se a técnica de AM escolhida gerar modelo, *i.e.*, Redes Neurais, fazendo com que a estimativa do valor ausente ocorra antes da linha 20.

Experimentou-se uma abordagem com reposição, porém, o processo torna-se pouco viável para técnicas que geram modelo por necessitar rebalanceamento a cada nova in-

teração. Nenhuma dessas variações apresentou melhorias visivelmente significantes e, por isso, não foram abordadas nos experimentos.

Algorithm 3 Algoritmo PSGF-KNN.

```
1:  procedure PSGFKNN
2:      DimSep ← inicializa com dimensão de separação
3:      DimEmbutida ← inicializa com dimensão embutida
4:      SerieDesdobrada ← Desdobramento(Serie, DimSep, DimEmbutida)
5:      ConjuntoTreino ← recebe os valores sem NA da SerieDesdobrada
6:      ColunaRotulo ← última coluna em SerieDesdobrada
7:
8:      for i = 1 to numero de linhas em SerieDesdobrada do
9:          if SerieDesdobrada[i, ColunaRotulo] é um valor ausente then
10:             NovoValor ← EstimaComKNN(SerieDesdobrada[i,],ConjuntoTreino)
11:             LinhaAtual ← i
12:             for ColunaAtual = 1 to DimSep do
13:                 SerieDesdobrada[LinhaAtual, ColunaAtual] ← NovoValor
14:                 LinhaAtual ← LinhaAtual + DimSep
15:     SerieReconstuida ← Redobramento(SerieDesdobrada, DimSep, DimEmbutida)
16:     return SerieReconstuida
```

4.3.2 Limitação do método

Encontrou-se uma limitação para casos onde uma lacuna de valores ausentes, ao aparecer no final da série, não terá todos os seus valores substituídos. A Figura 4.5 mostra os últimos 19 valores de uma série exemplo desdobrados no espaço fase. O processo estimará o valor na posição *(5, D2)* da representação 1 e esse valor será propagado para a posição *(11, D2)*. O mesmo processo se repete até a posição *(7, D2)* e o processo termina pois não há mais valor a ser estimado, não havendo equivalência entre os pontos *(8, D2)* e *(14, D2)*, uma vez que o ponto *(14, D2)* não existe. Sendo assim, o processo termina e existem 3 valores ausentes na série temporal como mostrado na matriz 2 da Figura 4.5.

No entanto, ao reduzir o valor da dimensão de separação o processo pode ser completado normalmente, assim como ilustrado na Figura 4.5 nas matrizes 3 e 4. Ou seja, para essa série, o método proposto acaba sendo limitado para uma dimensão de separação de valor menor ou igual a 3. Outra forma simples de superar essa limitação é, em casos como esse, não propagar os valor estimado.

Este tipo de interação acontece apenas em lacunas de valores ausentes que situam-se ao final da série temporal e parte dela é separada para a próxima dimensão embutida. Ao deparar-se com este tipo de cenário propõe-se avaliar se a redução da dimensão de separação afeta negativamente os resultados obtidos.

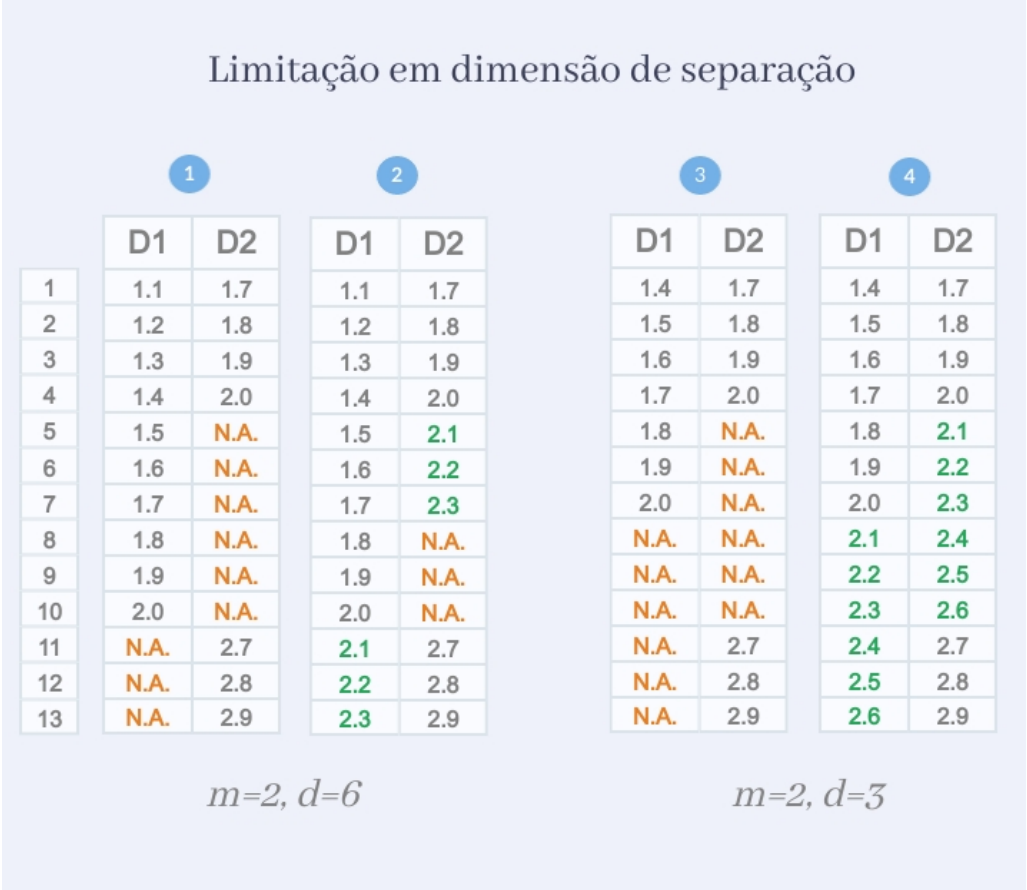


Figura 4.5 Limitação para dimensão de separação. Nesta série de exemplo a dimensão de separação é limitada ao valor 3 devido a posição da lacuna de valores ausentes.

4.4 CONSIDERAÇÕES FINAIS

Após detalhar os métodos desenvolvidos nesta dissertação, os próximos capítulos apresentam a construção do ambiente experimental considerado para avaliar os métodos e os resultados obtidos.

Capítulo

5

CONFIGURAÇÃO EXPERIMENTAL

5.1 CONSIDERAÇÕES INICIAIS

Este capítulo discorre sobre a descrição e organização dos experimentos desenvolvidos, tanto no espaço tempo quanto no espaço fase, para comprovar a hipótese apresentada na introdução. A primeira subseção visa explicar brevemente os experimentos realizados sobre o espaço tempo que foram publicados na principal conferência nacional de Inteligência Artificial (*8th Brazilian Conference on Intelligent Systems*). A segunda subseção descreve todos os detalhes dos experimentos e das formas de avaliação sobre a substituição de valores ausentes no espaço fase. Os resultados de ambos os experimentos são discutidos no Capítulo 6.

5.2 SÉRIES NO ESPAÇO TEMPO

Com o objetivo de validar a proposta, criou-se um conjunto de dados combinando diferentes níveis de determinismo e estocasticidade. Primeiramente, foram geradas 100 séries temporais combinando de 0 a 3 funções senoidais apresentadas na Equação 5.1. Em seguida, 100 séries temporais estocásticas foram criadas usando variáveis aleatórias normalmente distribuídas com média gaussiana igual a zero e desvio padrão limitado a $3,0$. Este parâmetro evita que o componente estocástico suprima o determinístico. Finalmente, as séries temporais estocásticas foram adicionadas às determinísticas, resultando em um conjunto de 100 séries temporais ruidosas.

$$\begin{aligned} x'_t &= \sin(\pi t) + \sin(2\pi t) + \sin(6\pi t) \\ x''_t &= \sin(\pi t) + \sin(6\pi t) \\ x'''_t &= \sin(\pi t) + \sin(6\pi t) + \sin(12\pi t) \end{aligned} \tag{5.1}$$

O tamanho das séries temporais variam de 1.000 até 2.000 observações. Para cada série temporal, removemos aleatoriamente 10% das observações, para simular uma lacuna, e as

usamos como *ground truth*. Os métodos usados para comparar nossa abordagem foram Cubic Spline e SSA. Esses métodos foram selecionados porque são amplamente adotados para substituir dados ausentes em séries temporais, assim como pode ser observado na RSL. Para avaliar o método proposto, considerou-se 3 medidas de distância amplamente utilizadas na área de análise de séries temporais: Dynamic Time Warping (DTW), Root Mean Square Error (RMSE) e Mean Absolute Error (MAE).

5.3 SÉRIES CAÓTICAS (ESPAÇO FASE)

Esta seção aborda as seguintes séries utilizadas para os experimentos no espaço fase: Lorenz, Rössler, Hénon e Mapa Logístico. Todas as técnicas utilizadas nos experimentos, assim como o método proposto, realizam substituição de valores ausentes no espaço fase afim de utilizar as informação das órbitas e atratores em cada uma das séries citadas. Após a substituição dos valores ausentes o erro é medido no espaço tempo. As próximas subseções apresentam detalhes sobre o processo de criação dessas séries.

5.3.1 Atrator de Lorenz

Edward Lorenz foi um meteorologista que utilizava um sistema de 12 equações diferenciais para modelar uma miniatura de atmosfera, visando entender e gerar um modelo para previsão do tempo (ALLIGOOD; SAUER; YORKE, 1997b). Em um dos seus experimentos, ele decidiu imprimir os resultados apenas com os 3 dígitos mais significantes. Ao tentar replicá-los, Lorenz identificou uma dependência sensitiva às condições inicias, percebendo, assim, que seria necessário reduzir o modelo para identificar se o comportamento seria mantido com um número menor de equações (ALLIGOOD; SAUER; YORKE, 1997b). Tal observação levou Lorenz a desenvolver um sistema capaz de descrever um mapa dinâmico usando as seguintes equações diferenciais:

$$\begin{aligned} x' &= \sigma(y - x) \\ y' &= x(\rho - z) - y \\ z' &= xy - \beta z, \end{aligned} \tag{5.2}$$

Lorenz descobriu que com os seguintes parâmetros: $\sigma = 10$ e $\beta = 8/3$ e ρ excedendo o valor 24.74 o sistema apresenta comportamento caótico. As Figuras 5.1 e 5.2 demonstram, respectivamente, o comportamento da série de Lorenz em forma de série temporal e desdobradas no espaço fase com os seguintes parâmetros: dimensão embutida $m = 3$ e de separação $d = 5$.

5.3.2 Atrator de Rössler

O atrator de Lorenz foi um dos primeiros amplamente conhecidos a partir de equações diferenciais simples. Após ele, muitos outros sistemas caóticos foram identificados (ALLIGOOD; SAUER; YORKE, 1997a). Um deles foi o atrator de Rössler, que foi desenvolvido a partir das seguintes equações:

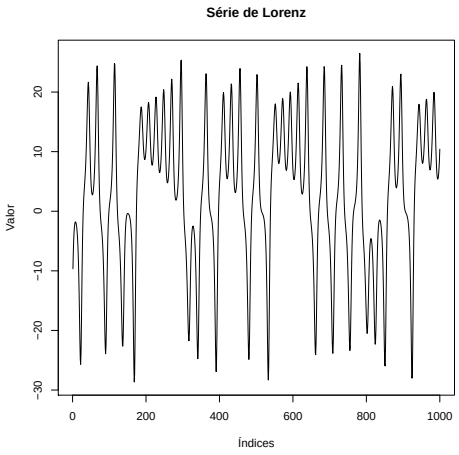


Figura 5.1 Série de Lorenz.

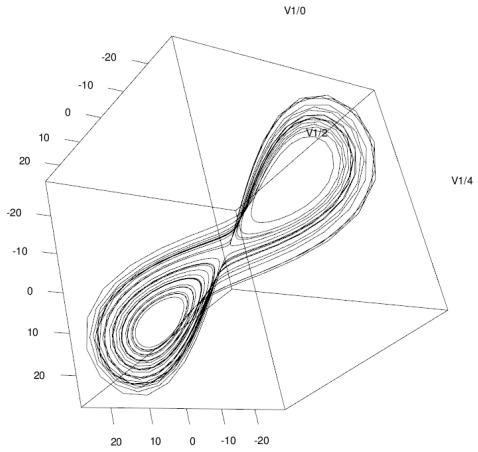


Figura 5.2 Série de Lorenz - Desdobramento.

$$\begin{aligned} x' &= -y - z \\ y' &= x + y\sigma \\ z' &= \rho + (x - \beta)z, \end{aligned} \tag{5·3}$$

Assim como o atrator de Lorenz, nem toda combinação de parâmetros apresenta comportamento caótico para atrator de Rössler. Enquanto os parâmetros $\sigma = \rho = 0.1$ e $\beta = 4$ geram um sistema periódico, ao alterar-se o valor do parâmetro $\beta = 9$ o sistema de Rössler apresenta comportamento caótico.

Inicialmente, o sistema de Rössler foi desenvolvido para ser um modelo simples de estudo do caos, mas o modelo também se mostrou útil no equilíbrio de reações químicas, ajudando a entender o caos químico em colisões e reações de partículas (SCOTT, 1993). As Figuras 5.3 e 5.4 demonstram o comportamento em forma de série temporal e o desdobramento sobre seu atrator no espaço fase usando os seguinte parâmetros: dimensão embutida $m = 3$ e de separação $d = 2$.

5.3.3 Atrator de Hénon

O atrator de Hénon surgiu com o mesmo objetivo do atrator de Rössler, encontrar um modelo simplificado com as mesmas propriedades do modelo de Lorenz. Esse modelo tinha como objetivo fazer a exploração numérica mais rápida e precisa para que as soluções pudessem ser acompanhadas por um período maior de tempo, e assim, fornecer um modelo matemático mais simples de ser analisado (HÉNON, 1976). Um exemplo de aplicação do mapa de Hénon é sua utilidade em algoritmos para criptografia de imagens de alta definição (SOLEYMANI; NORDIN; SUNDARARAJAN, 2014).

As Figuras 5.5 e 5.6 ilustram o comportamento da série temporal do mapa de Hénon e seu desdobramento no espaço fase. Assim como os atratores discutidos anteriormente,

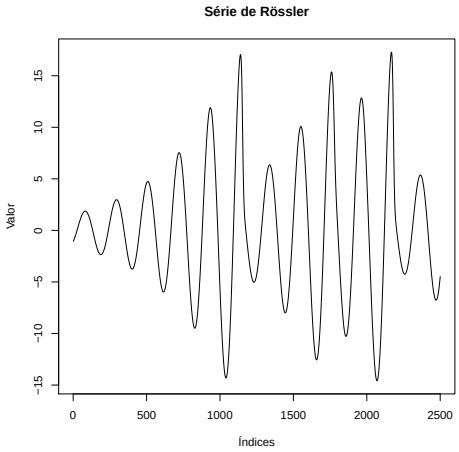


Figura 5.3 Série de Rössler

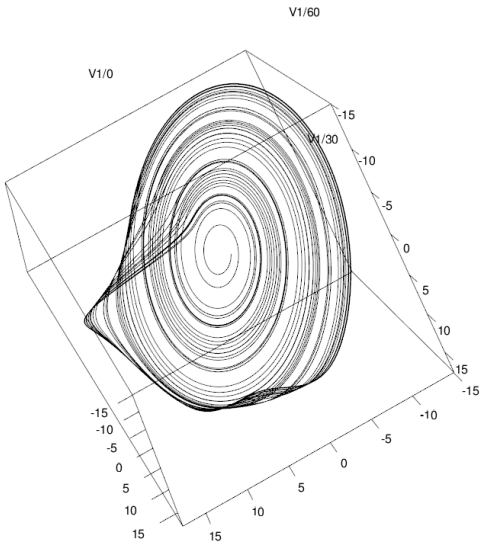


Figura 5.4 Série de Rössler - Desdobramento

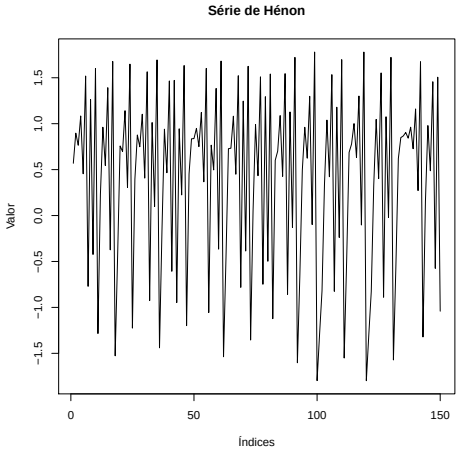


Figura 5.5 Série de Hénon.

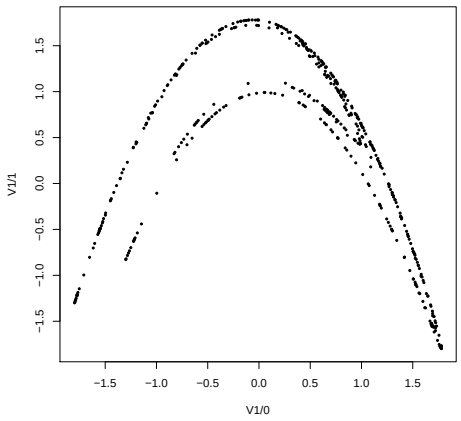


Figura 5.6 Série de Hénon - Desdobramento.

o mapa de Hénon não define sempre um sistema caótico usando os seguinte parâmetros: dimensão embutida $m = 2$ e de separação $d = 3$. O mapa de Hénon classico foi definido com os parâmetros $a = 1.4$ e $b = 0.3$.

$$\begin{aligned} x_{n+1} &= 1 - ax_n^2 + y_n \\ y_{n+1} &= bx_n \end{aligned} \tag{5.4}$$

5.3.4 Mapa Logístico

A família dos mapas logísticos aparece frequentemente em estudos de sistemas dinâmicos e em matemática aplicada (KRAFT, 1999). Um exemplo é sua utilidade no estudo das dinâmicas de populações, crescimento de tumores cancerígenos e modelos epidemiológicos. O mapa logístico apresenta dinâmicas caótica a partir da seguinte equação:

$$x_{n+1} = rx_n(1 - x_n) \tag{5·5}$$

Esse modelo representa a família dos mapas logísticos e, dependendo de suas condições iniciais e do valor de r, esta simples equação apresenta um comportamento caótico. A Figura 5.7 mostra o comportamento do modelo da família de mapas logísticos ao variar-se o valor de r a partir de 2.5, variando-se em 0.003 e ponto inicial $x = 0.1$. O mapa inicia com órbita de ponto fixo até r atingir o valor 3, então inicia-se um processo de bifurcação com o aumento de r até um momento em que a bifurcação atinge um comportamento caótico e imprevisível.

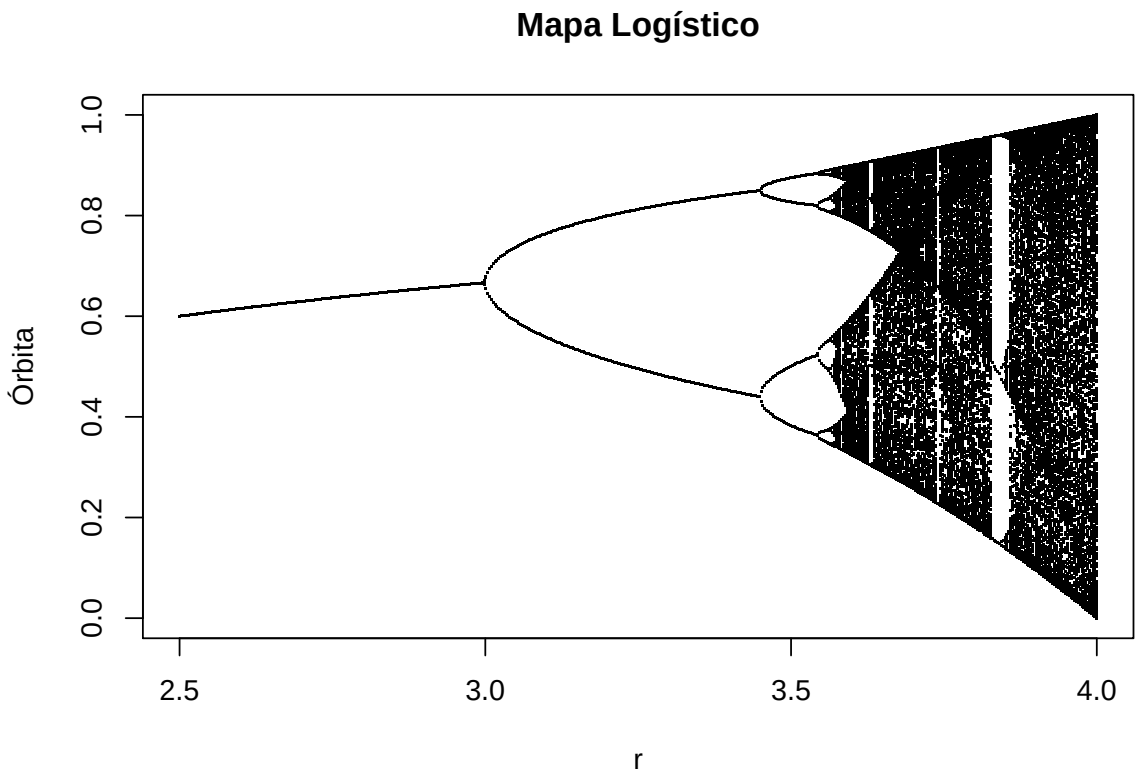


Figura 5.7 Diagrama de bifurcação do modelo logístico com a variação de r.

5.4 ORGANIZAÇÃO DOS EXPERIMENTOS

Conforme discutido no Capítulo 4.3, o método proposto possui uma etapa de predição de valores ausentes que permite utilizar diferentes métodos de Aprendizado de Máquina. Nos experimentos realizados nessa dissertação, as seguintes técnicas foram implementadas:

- K-Nearest Neighbors (KNN)
- Distance Weighted K-Nearest Neighbors (DWNN)
- Support Vector Regression (SVR)
- Random Forests (RF)

É importante destacar que, embora essas técnicas tenham sido escolhidas nesse ambiente experimental, o leitor poderá alterar-las por qualquer outro método de AM que melhor modele seus dados. Vale ressaltar que, para melhor desempenho do método, essa alteração poderá necessitar de algumas alterações no algoritmo do método. Técnicas que geram modelos e técnicas que não geram modelos devem ser inseridos de formas diferentes no método.

Para cada experimento executado com as técnicas KNN e DWNN, diversos valores de k foram testados variando de 1 a 15, sendo escolhido o valor de k que minimize a métrica de erro utilizada.

As técnicas RF e SVR necessitam de ajuste de outros parâmetros antes de serem utilizadas com a proposta deste trabalho. Para esse objetivo, utilizou o seguinte método de busca de melhor parâmetro: *Random Search*. Esse método define um espaço de busca para, através do processo de treinamento e validação, otimizar a combinação de parâmetros. O processo de treinamento e validação foi realizado com a abordagem *K-fold Cross Validation* com $k = 10$.

Assim como estudado e apontado na RSL, os erros de substituição de valores ausentes variam de acordo com os dados, posição, tamanho e frequência em que as falhas acontecem. Os experimentos deste trabalho também foram guiados com esse mesmo conceito, sendo divididos em 3 macroexperimentos: i) variação de posição de janela onde cada experimento apresenta apenas uma lacuna e varia-se sua posição; ii) multi-lacuna onde cada experimento contém pequenas lacunas igualmente espaçadas ao longo da série; iii) variação completa onde varia-se tamanho, posição e quantidade de lacunas ao longo da série de forma randômica.

Cada macroexperimento é executado sobre todas as séries e, para cada uma delas, também será avaliado o comportamento do erro em diversas dimensões de separação e sobre cada técnica citada anteriormente (KNN, DWNN, RT e SVR). Foi definido que a dimensão de separação será avaliada entre os valores 1 e 15. Esse intervalo de valores foi limitado a um valor factível que não aumentasse muito a complexidade dos experimentos. As Figuras 5.8, 5.9 e 5.10 demonstram visualmente como cada experimento foi executado.

Como mencionado anteriormente no Capítulo 4, os experimentos foram realizados considerando um cenário onde as dimensões embutidas são conhecidas. As dimensões

embutidas para cada série foram: $m = 3$ para Lorenz e Rössler, e $m = 2$ para Henon e Mapa Logístico (ALLIGOOD; SAUER; YORKE, 1997a; RIBEIRO; RIOS, 2021).

A Figura 5.8 apresenta passo a passo como o experimento de variação de janela foi executado. O fluxograma pode ser lido como a execução de um *loop* a cada camada descendente. De maneira geral, executou-se as seguintes etapas: i) seleciona-se um método de AM para ser experimentado; ii) define-se um atraso (dimensão de separação) para ser avaliado; iii) seleciona-se uma série para ser desdobrada com esse atraso; iv) busca-se uma janela de valores ausentes e, então, aplica-se o método proposto. Todo o processo é repetido até que todos os métodos de AM, valores de atraso, séries temporais e janelas tenham sido experimentados.

Os experimentos de multilacuna e variação completa seguem um processo semelhante, como pode ser visto nas Figuras 5.9 e 5.10, respectivamente. A diferença principal é a fase de inserção dos valores ausentes (última fase). No experimento de multilacunas apresentado pela Figura 5.9, insere-se múltiplas lacunas igualmente espaçadas e varia-se o tamanho dessas lacunas, enquanto que no experimento de variação completa, visto na Figura 5.10, a inserção de valores ausentes é feita de forma aleatória em tamanho, posição e quantidade de valores ausentes inseridos.

As séries temporais nos experimentos anteriores (variação de janela e multilacuna) possuíam 10.000 observações. No entanto, o tamanho das séries foi reduzido para 5.000 Observações no experimento de variação completa devido ao elevado custo computacional. Entretanto, essa redução não produz impacto sobre o resultado analisado, visto que o comportamento geral do sistema independe do número de observações.

5.5 AVALIAÇÃO

A forma de avaliação utilizada por estudos encontrados através da RSL consistem, normalmente, em dois tipos de avaliações: quantitativa e qualitativa. A avaliação quantitativa é feita por medidas de erros e a avaliação qualitativa por inspeção visual, onde a série é visualmente apresentada e, sobre ela, os valores ausentes substituídos com a técnica avaliada.

A avaliação qualitativa é tão importante quanto a quantitativa porque, em alguns casos, o comportamento geral dos valores substituídos é mais importante do que a exatidão da medida quantificadora.

Nos experimentos com séries caóticas, optou-se por utilizar apenas RMSE para quantificar o erro, além das avaliações qualitativas através de inspeção visual. A escolha do RMSE foi justificada porque as outras medidas (DTW e MAE) não apresentaram diferenças significativas no processo de avaliação temporal.

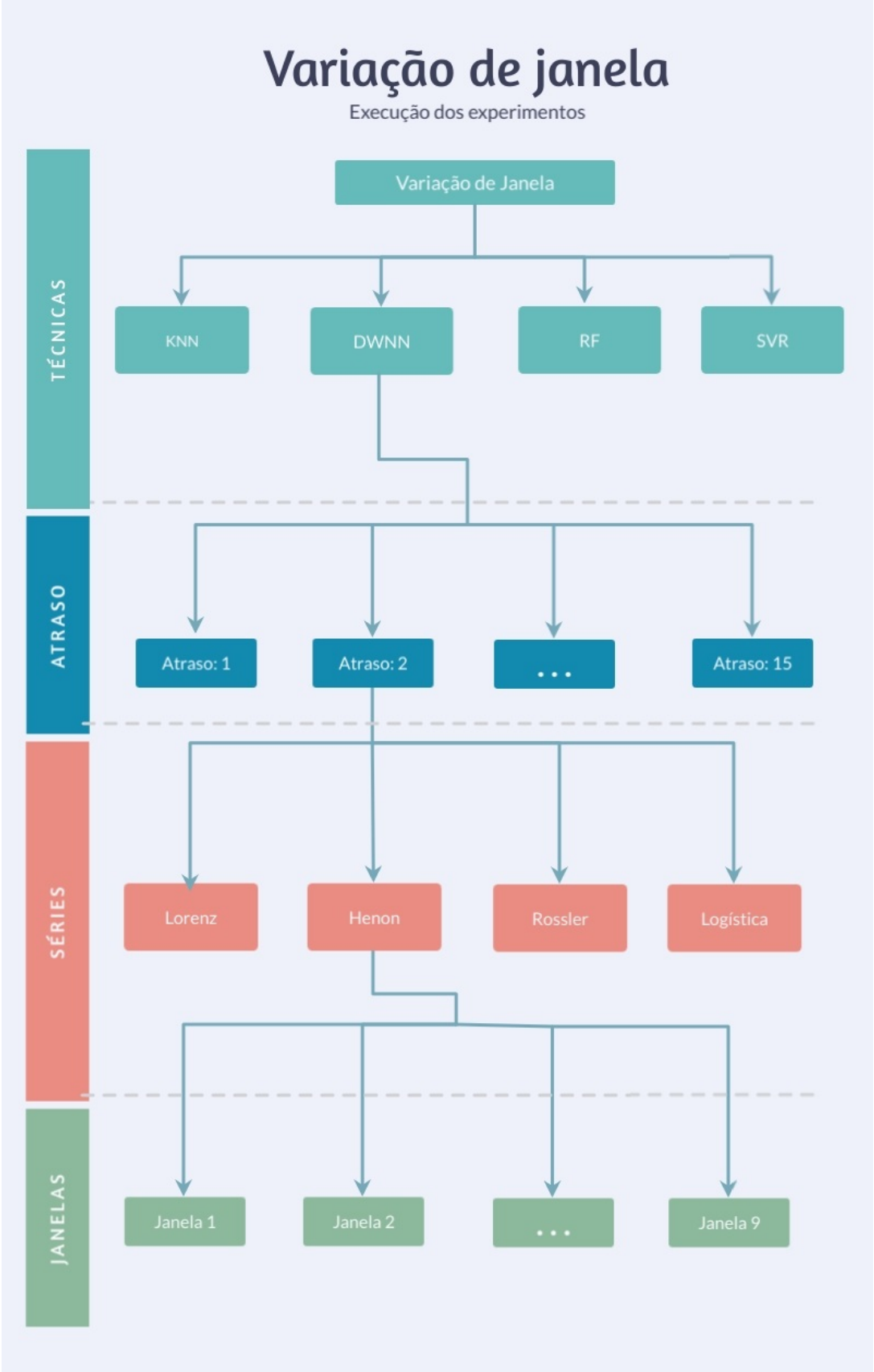


Figura 5.8 Organização do experimento de variação de janela - Uma técnica é aplicada em todas as séries e para cada série será gerada uma série com uma lacuna de valores ausentes em uma janela diferente.

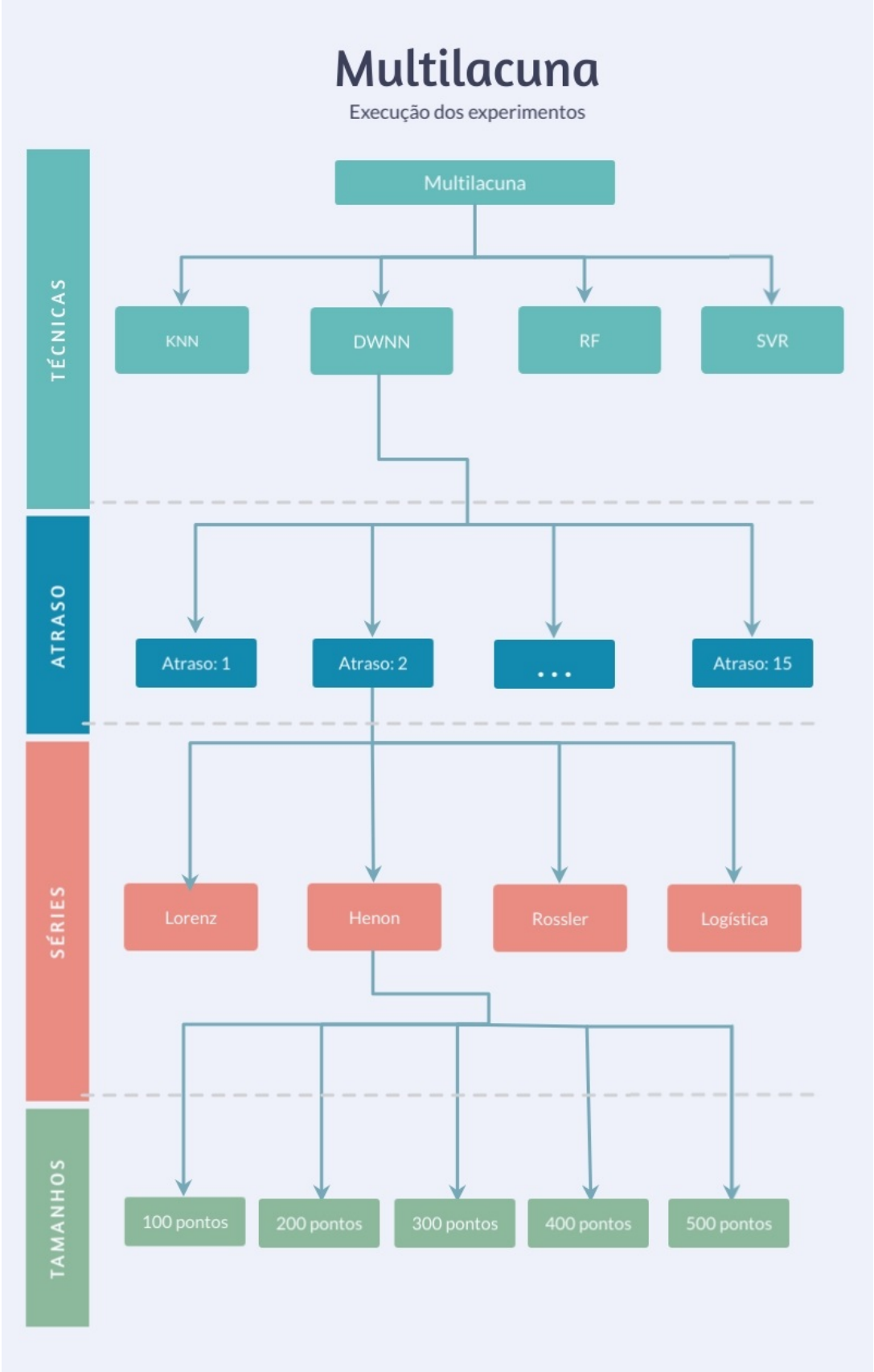


Figura 5.9 Organização do experimento multilacuna - Uma técnica é aplicada em todas as séries e para cada série será gerada uma série com 8 lacunas de tamanhos variados.

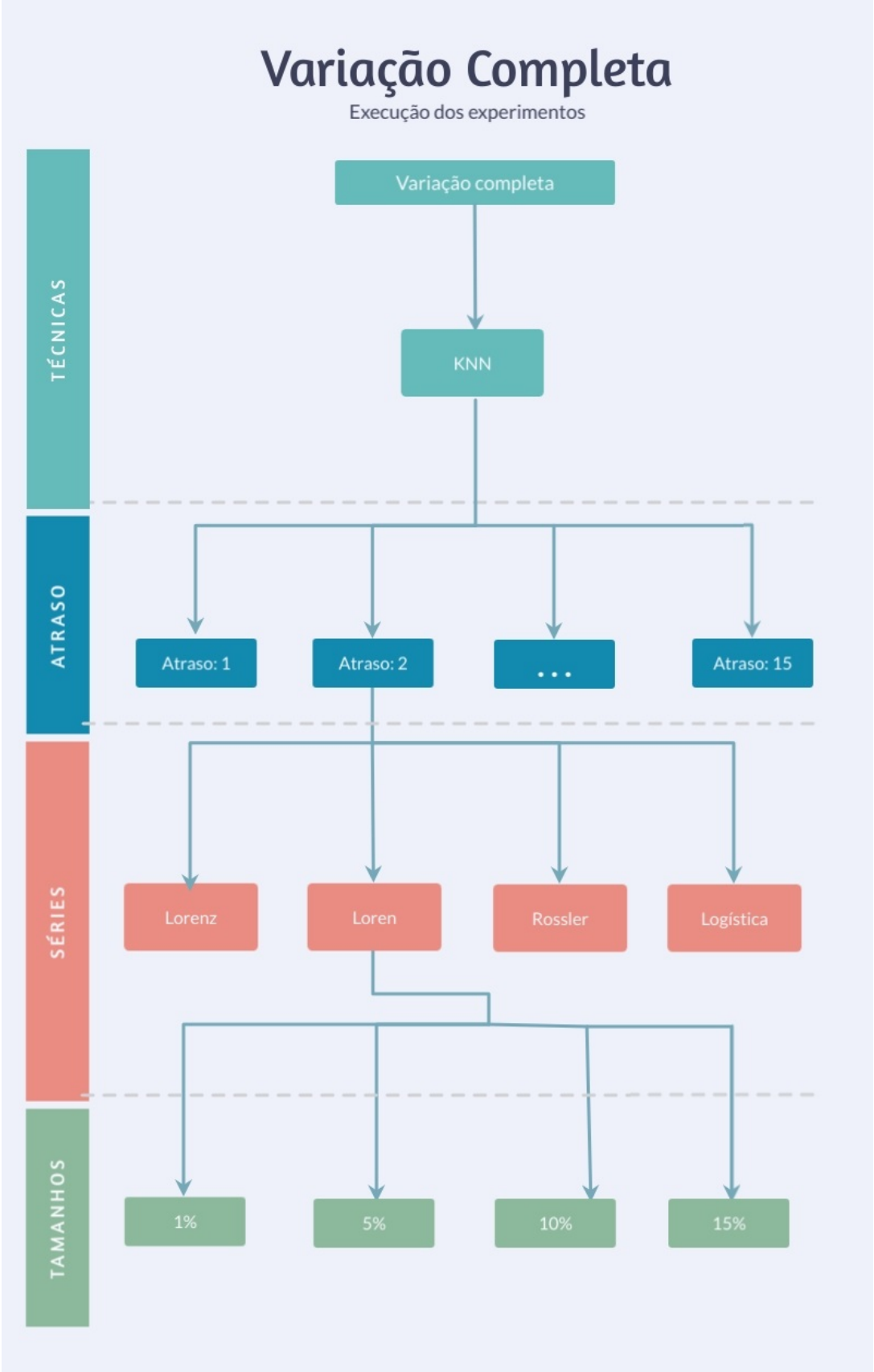


Figura 5.10 Organização do experimento de variação completa - A técnica KNN é aplicada em todas as séries e para cada série será gerada uma ou mais lacunas de valores ausentes com tamanhos variados. Cada experimento será repetido 40 vezes para cada porcentagem de perda de dados.

Capítulo

6

RESULTADOS

6.1 CONSIDERAÇÕES INICIAIS

Este capítulo apresenta e discute os resultados obtidos com os experimentos descritos no Capítulo 5. As duas próximas seções discutem separadamente os resultados no espaço tempo e no espaço fase. Diante do grande número de macroexperimentos realizados com as diferentes configurações e visando realizar uma discussão de maneira mais concisa, optou-se por apresentar nesse capítulo apenas a parte mais relevante dos resultados. No entanto, uma versão completa dos resultados é apresentada no Apêndice A.

6.2 RESULTADOS ESPAÇO TEMPO

Os resultados apresentados nesta seção foram divididos em duas partes. Na primeira, foi realizada uma inspeção visual com o objetivo de avaliar se a abordagem proposta fornece novos dados próximos aos dados originais. Na segunda parte, avaliou-se empiricamente a abordagem em todo o *dataset* e os resultados foram analisados usando medidas de distância.

A inspeção visual foi realizada em 6 séries temporais (Figura 6.2) com comportamento similar ao ambiente experimental discutido na Seção 5. Para todas as séries temporais, estimou-se uma substituição dos dados usando BMDE (linhas azuis), SSA (linhas vermelhas), e Splines Cúbias (linhas roxas).

Comparando as saídas produzidas por estes métodos com os dados originais na Figura 6.1(a) é possível notar que o SSA e o BMDE apresentaram resultados similares, principalmente comparando um comportamento mais generalizado. Um comportamento similar também pode ser encontrado na Figura 6.1(b) mesmo com um maior número de dados ausentes. Apesar de apresentar alguma variação nos valores individuais, nota-se que SSA e BMDE mantiveram o comportamento geral da série.

Na Figura 6.1(c), foi analisada outra série temporal com uma lacuna maior de dados ausentes. Neste caso, o BMDE apresentou melhores resultados se comparados com os valores esperados.

Os resultados apresentados na Figura 6.1(d) e 6.1(e) avaliam a substituição dos valores ausentes em casos onde os dados mudam de comportamento ao longo do tempo de maneira não-estacionária. Nas duas figuras, os três métodos foram capazes de replicar o comportamento geral dos dados, embora a Spline Cúbica mostre maior amplitude se comparada com os dados originais e os demais métodos.

O último experimento avaliado por inspeção visual, a Figura 6.1(f), simula uma série temporal sem comportamento determinístico, usando apenas um processo aleatório. Como esperado, a Spline Cúbica apresenta o pior resultado, por ter sido desenvolvida para modelar comportamentos determinísticos.

Embora estes experimentos tenham sido úteis para entender, detalhadamente, como os métodos de substituição de valores ausentes funcionam, os resultados obtidos foram também analisados com o apoio de medidas de distâncias. Resumidamente, aplicou-se as medidas de distância para calcular as diferenças entre os resultados estimados e os esperados. A Figura 6.1(a) enfatiza como o BMDE e o SSA apresentam comportamento similar ao estimar os dados ausentes. Apesar do BMDE apresentar alguns picos acima do SSA, a memória utilizada é consideravelmente menor pois a matriz de Hankel, que transforma a série temporal com N observações em uma matriz $N \times N$, não se faz necessária. Os resultados da Spline Cúbica foram apresentados em uma plotagem diferente (Figura 6.2) devido aos problemas de escala, amplificados pela influência estocástica.

O BMDE apresenta um novo método para substituição de valores ausentes em séries temporais. De acordo com os experimentos, foi possível notar excelentes resultados, modelando séries temporais independentemente de suas características estacionárias, estocásticas e lineares. Além disso, o BMDE é uma alternativa importante quando o tamanho da série temporal restringe seu uso em cenários de memória limitada.

6.3 RESULTADOS ESPAÇO FASE

Esta seção apresenta os resultados obtidos usando séries caóticas. Os experimentos foram divididos em 3 macroexperimentos. O primeiro macroexperimento (variação de janela) avalia como a posição da lacuna e a dimensão de separação afetam a imputação dos dados, aumentando ou reduzindo o erro. O segundo macroexperimento (multilacuna) avalia a variação do erro em casos com múltiplas lacunas variando-se, também, o tamanho da lacuna. O terceiro macroexperimento (variação completa) é uma combinação dos dois anteriores, onde avalia-se o tamanho, posição e a frequência das lacunas e como estas influenciam os resultados da substituição dos valores ausentes no espaço fase.

Para todos os experimentos as medidas de erro apresentaram comportamento muito similar, portanto, os resultados serão avaliados apenas com a medida RMSE. Antes de apresentar os resultados completos, é importante destacar que a aplicação dos métodos Spline e SSA produziram erros maiores que o método proposto por não conseguir modelar adequadamente as séries caóticas. Esses erros elevados acontecem porque tais métodos não foram originalmente construídas para analisar séries no espaço fase. Enquanto SSA realiza decomposição da série no espaço tempo, utilizando informações extraídas do agrupamento de pequenos componentes ortogonais (autovalores e autovetores), PSGF transforma a série para seu espaço de coordenadas de atraso e utiliza informações de suas

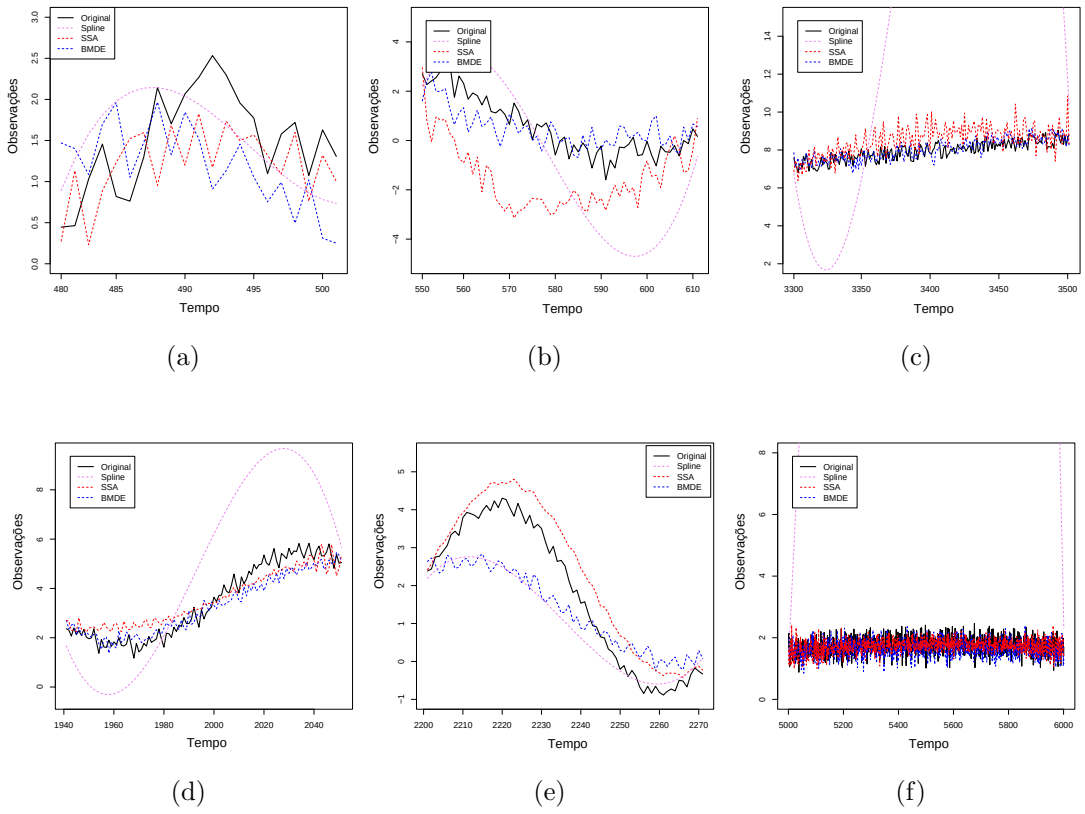


Figura 6.1 Um conjunto de 6 experimentos usando BMDE (linhas azuis), SSA (linhas vermelhas), Splines Cúbicas (linhas roxas) e os resultados esperados (linhas pretas).

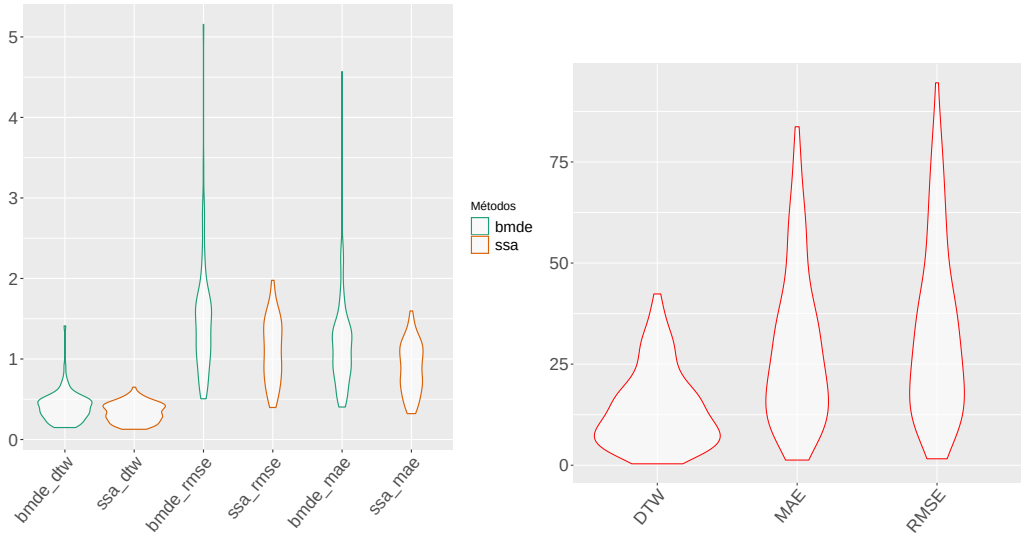
órbitas e atratores. Sendo assim, uma análise comparando resultados entre métodos sobre o espaço fase e métodos sobre de espaço tempo é, provavelmente, inadequada.

Para ilustrar esse cenário, observe a Figura 6.3. Após realizar uma análise da série para encontrar uma combinação ótima de parâmetros de dimensão embutida e de separação, a aplicação do método PSGF + KNN sobre o espaço fase apresenta resultados que condizem com o comportamento geral de série resultante. No entanto, ao utilizar a Análise Espectral Singular (SSA) o resultado não se aproxima do comportamento esperado.

Por essa razão, visando realizar uma avaliação mais justa, optou-se por remover esses métodos das análises e manter o foco no desempenho do método proposto em diferentes situações de valores ausentes.

6.3.1 Variação de posição de janela

Este experimento investiga o comportamento do erro em relação a posição em que a ausência de dados foi inserida na série temporal. O tamanho da lacuna para este macroexperimento foi definido como 1% do tamanho da série temporal (100 pontos) e, esta lacuna, é inserida em pontos fixos para todas as séries. Estes pontos foram escolhidos de maneira janelada, dividindo a série em 10 janelas e evitando inserir as lacunas de valores



(a) (b)

Figura 6.2 Plotagem de violino comparando BMDE, SSA e Splines Cúbicas em um grande *dataset*.

ausentes no início ou no fim da série, evitando assim, discussões sobre previsão de séries temporais. Os pontos escolhidos para a inserção dos valores ausentes foram os seguintes pontos: [900, 1900, 2900, 3900, 4900, 5900, 6900, 7900, 8900]. O fluxograma da Figura 5.8 exibe a organização deste experimento. Os resultados serão apresentados para cada técnica separadamente nas próximas subseções.

6.3.1.1 KNN A Figura A.1 apresenta a distribuição dos erros para cada dimensão de separação. Pode-se concluir a partir da análise dos erros que a dimensão de separação impacta nos resultados da substituição dos valores ausentes nas séries. Os resultados também mostram comportamentos bem diferentes entre as série, na qual a série de Rössler não apresenta grandes alterações em seus resultados ao aumentar a dimensão de separação em seus dados.

Nota-se ainda que não existe padrão entre os erros das séries porque, ao analisar a série de Lorenz, percebe-se um comportamento quase inverso aos erros da série de Hénon. A Tabela 6.1 resume as dimensões de separação com os melhores resultados para cada série temporal.

Cruzando as informações da Tabela 6.1 com a distribuição dos erros visto na Figura A.1 foi possível criar a Figura A.5 que apresenta um resumo dos experimentos com os melhores resultados para a execução do PSGF com o KNN. Este resumo apresenta um cruzamento da quantificação do erro (RMSE) com a posição em que a lacuna de valores ausentes aparece na série. Analisando a Figura A.5, pode-se concluir que a série de Lorenz é a mais afetada pela variação da posição da lacuna de valores ausentes, enquanto que as demais séries (Hénon, Logística e Rössler) demonstram comportamento estável. É possível notar, ainda, que o Mapa Logístico apresenta os melhores resultados para este

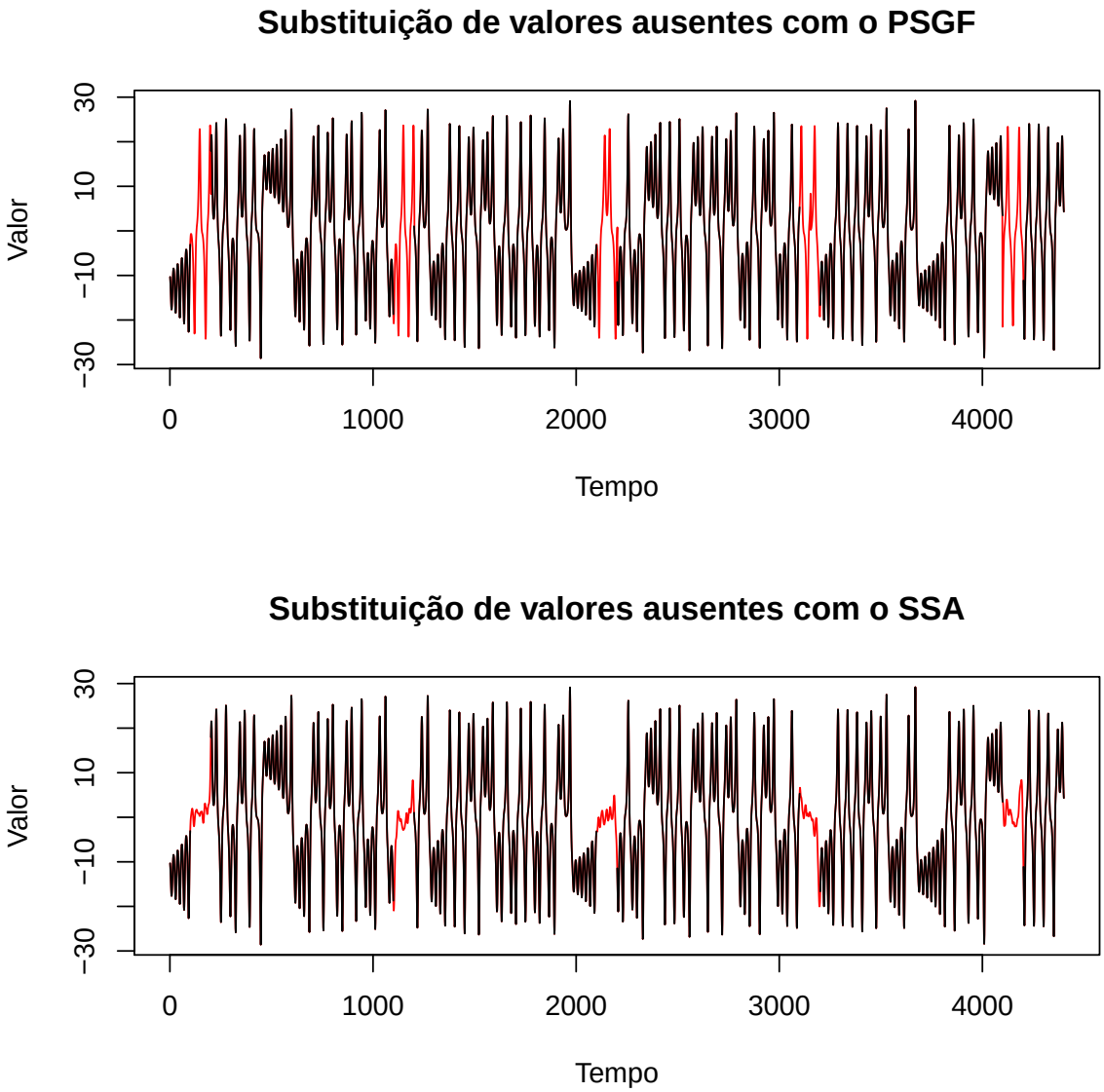


Figura 6.3 Substituição de valores ausentes utilizando o PSGF e o SSA. Pode-se perceber que com parametrização e métodos simples, a substituição pode gerar resultados promissores em relação à técnicas que consideram apenas o espaço tempo.

experimento com RMSE variando entre 2.24 e 2.74.

6.3.1.2 DWNN Os resultados das técnicas DWNN não apresentam muitas mudanças em relação a KNN. Pode-se notar na Figura A.2 que a distribuição dos erros em relação a dimensão de separação tem comportamento semelhante. No experimento com a série de Lorenz, pode-se verificar mais *outliers* no experimento com DWNN se comparado

Tabela 6.1 Dimensões de separação para cada série - KNN.

Série	Dimensão de Separação
Lorenz	1
Hénon	12
Rössler	9
Mapa logístico	14

ao KNN. Além disso, ao aumentar a dimensão de separação, o comportamento é bem semelhante e apresenta estabilidade nos resultados a partir da dimensão de separação com valor 10. Essa estabilidade parece estar mais presente com a técnica DWNN para a série de Lorenz.

Para a série Hénon, DWNN apresenta maior estabilidade nos resultados. No entanto, a técnica KNN mostra erros reduzidos após a dimensão 10, onde seus resultados se estabilizam. Para a série Rössler, os resultados são ainda mais semelhantes. Ambas as técnicas são bem estáveis para esta série e mostram pouca diferença na distribuição dos erros. Os experimentos com DWNN e KNN para série do Mapa Logístico apresentam baixa variação na magnitude dos resultados. Ambos variam entre 2.2 e 3. Neste caso, a escolha da dimensão de separação não apresenta diferença relevante. Mesmo a técnica KNN tendo melhores resultados com a dimensão de separação 14 e a DWNN com 6, a diferença entre os erros é mínima.

A Tabela 6.2 apresenta as melhores dimensões de separação para cada série no experimento com DWNN. A Figura A.6 mostra o erro ao longo da série variando-se a janela onde os valores ausentes aparecem. Percebe-se que, assim como na Figura A.5, a posição onde aparecem os valores ausentes afeta muito os resultados na série de Lorenz. Os experimentos com DWNN apresentam ainda menos estabilidade nos resultados mas, em algumas janelas, os erros são baixos, entregando assim uma erro médio menor para determinados experimentos. Por fim, nota-se que experimentos com a série Rössler apresentam comportamentos semelhantes entre as duas técnicas (KNN e DWNN), com janelas de pico e crescimento bem similares.

Tabela 6.2 Dimensões de separação para cada série - DWNN.

Série	Dimensão de Separação
Lorenz	1
Hénon	11
Rössler	11
Mapa logístico	6

6.3.1.3 Random Forest A técnica Random Forest, mesmo sendo uma abordagem diferente das anteriores, apresenta números semelhantes na distribuição dos erros em relação a dimensão de separação. Apenas a série de Rössler apresenta grandes mudanças

na distribuição dos erros, mostrando uma variação bem maior em relação às técnicas KNN e DWNN. A Figura A.3 mostra que para a dimensão de separação de valor 1, os erros variam de 20 a 70, mas ao analisar a dimensão de separação de valor 4, os valores voltam a se assemelhar com as técnicas KNN e DWNN. Este experimento mostra que uma boa escolha para dimensão de separação pode entregar resultados tão bons quanto outras técnicas. Isso pode ser observado na série de Hénon na Figura A.3, no qual os resultados com RF apresentaram comportamento bem semelhante à técnica KNN a partir da dimensão de separação igual a 7, tanto na distribuição dos erros quanto nos valores das dimensões de separação. O Mapa Logístico também apresenta bons resultados e sofre pouca variação na distribuição dos erros com a mudança da dimensão de separação.

Assim como abordado nas técnicas anteriores, também foi feito um cruzamento dos dados da Tabela 6.3 com a Figura A.3, produzindo a Figura A.7 que agrega os melhores resultados para este experimento. Pode-se perceber que a posição da janela afeta as séries de maneira diferente. Analisando este experimento e comparando-o com os anteriores, é possível notar que a série de Mapa Logístico é menos impactada com a variação da posição em que os valores ausentes aparecem na série. Para a série de Rössler os erros da RF aumentaram consideravelmente em relação as técnicas de vizinhos mais próximos, com picos que variam de 1 até 24.

Tabela 6.3 Dimensões de separação para cada série - RF.

Série	Dimensão de Separação
Lorenz	3
Hénon	13
Rössler	4
Mapa logístico	15

6.3.1.4 SVR A substituição dos valores ausentes utilizando a técnica SVR apresenta resultados diferentes das técnicas anteriores. Analisando a distribuição dos erros na Figura A.4, é possível identificar facilmente picos de variação de erros muito maiores. No entanto, após encontrada a dimensão de separação de melhor ajuste, os resultados voltam a se assemelhar aos resultados das outras técnicas. Para a série de Lorenz, os resultados começam a estabilizar após dimensão de separação com valor 10. Para Hénon, a partir do valor 7 e para o Mapa Logístico a partir do valor 6. Este comportamento diferencia-se apenas para a série de Rössler que apresenta variações maiores.

A visualização da Figura A.4 é afetada para alguns resultados por mostrar grandes diferenças entre a distribuição dos erros de cada dimensão de separação. Assim como nos experimentos anteriores, cruzar dados da Tabela 6.4 com os dados da Figura A.4 fornecerá uma melhor visualização para entender os melhores resultados desta técnica. As informações presentes na Figura A.8 mostram a técnica SVR com resultados muito similares à técnica KNN, apresentando melhor resultado que as demais para a série de Mapa Logístico e Hénon. O mesmo não pode ser observado para a série de Lorenz e Rössler, onde a técnica SVR apresentou erros maiores que as técnicas anteriores.

Tabela 6.4 Dimensões de separação para cada série - SVR.

Série	Dimensão de Separação
Lorenz	13
Hénon	13
Rössler	14
Mapa logístico	9

6.3.1.5 Resumo do experimento - Variação de janela O objetivo deste experimento é entender como a substituição dos valores ausentes foi afetada pela posição em que cada lacuna aparece e pelas diferentes dimensões de separação. Foi possível observar que os erros podem variar bastante com a posição da lacuna em alguns casos, assim como em séries temporais não-caóticas. Pode-se citar como exemplo os resultados exibidos na Figura A.8 para a série de Rössler, onde a posição 3 e a posição 5 apresentam resultados bem distantes.

Este experimento também mostra o quão é importante escolher corretamente a dimensão de separação ao realizar-se substituição de valores ausentes no espaço fase. Outro objetivo deste experimento é investigar se alguma dimensão de separação (dentre os melhores resultados) se repete entre as técnicas. Apenas para a série de Lorenz e Hénon foi possível observar-se uma certa repetição. Para a série de Lorenz, a melhor dimensão de separação se concentra entre os valores menores (exceto para a técnica SVR). Para a série de Hénon, concentra-se entre os valores maiores de dimensão de separação. Esta informação pode ser valiosa ao se realizar em grandes *benchmarks*, evitando-se assim valores altos para a série de Lorenz e valores baixos para a série de Hénon.

Por fim, este experimentou mostrou melhor desempenho com a combinação do método PSGF com a técnica KNN, mesmo com resultados de outras técnicas sendo bem próximos. Os próximos experimentos serão fundamentais para entender se esta interação se repete.

6.3.2 Multilacuna

O objetivo deste experimento é avaliar o comportamento do erro ao se inserir várias lacunas de valores ausentes ao longo da série temporal. Com esse experimento, é possível explorar o crescimento do erro em casos onde há perda de dados em várias regiões da série e com diversos tamanhos diferentes. É importante destacar que os experimentos apresentados nesta seção foram conduzidos considerando os valores de dimensão de separação estimados com os menores erros apresentados na seção anterior.

A Tabela 6.5 resume os resultados obtidos com todas as técnica de AM, séries e tamanhos da lacunas de valores ausentes, mostrando em uma escala de cores onde estão os melhores resultados para cada situação. Com esta tabela é possível concluir que, assim como esperando, o erro aumenta com maiores lacunas de valores ausentes. Além disso, é possível verificar que o tamanho da lacuna de valores ausentes afeta cada série de diferentes maneiras.

Em todas as técnicas de aprendizado de máquina avaliadas, a série de Lorenz apresenta alta sensibilidade ao tamanho do lacuna. Isto pode ser notado em todos os experimen-

	KNN				
Tamanho	100	200	300	400	500
Lorenz	2,5291	189,1345	225,7624	262,9993	290,8490
Henon	10,1780	14,4206	17,6865	20,1042	22,5455
Mapa Logístico	2,4530	3,3942	4,2684	4,9174	5,4776
Rossler	5,1130	8,0684	23,6726	26,0103	44,5500

	DWNN				
Tamanho	100	200	300	400	500
Lorenz	2,5267	190,2744	239,9160	277,0808	314,9308
Henon	10,6882	16,2129	20,3348	23,3331	25,6594
Mapa Logístico	2,8534	3,9209	4,8665	5,6079	6,3794
Rossler	5,1130	8,0684	14,9124	16,2743	27,1270

	Random Forest				
Tamanho	100	200	300	400	500
Lorenz	133,6320	271,2134	262,9723	317,1198	355,2621
Henon	10,7469	16,7399	20,6770	23,2474	26,7045
Mapa Logístico	3,1043	4,4361	5,5428	6,5059	7,2310
Rossler	8,6056	25,4153	46,7019	56,4698	67,5671

	Support Vector Regression				
Tamanho	100	200	300	400	500
Lorenz	124,6166	176,6741	219,4734	253,9623	283,1152
Henon	10,4158	14,5218	17,7421	20,1930	22,4299
Mapa Logístico	2,5243	3,4975	4,2369	4,8830	4,8830
Rossler	8,0505	15,8555	39,5903	48,8029	69,3401

Tabela 6.5 Resumo do experimento multilacuna. As tabelas resumem os valores de RMSE para cada técnica, série e tamanho de lacuna. Os valores foram mapeados em diferentes cores: melhores resultados em tons de verde e piores resultados em tons de vermelho.

tos para a séries Lorenz e, principalmente, se comparado o experimento com a técnica KNN para os tamanhos 100 e 200, cujos erros variam de $2,51$ até $189,13$. Este tipo de comportamento também pode ser observado com a técnica RF para a série de Rössler entre os tamanhos 100 e 200. No entanto, a série do Mapa Logístico não apresenta uma sensibilidade tão grande ao tamanho das lacunas, sendo a série que tem menor variação nos resultados para todas as técnicas. A tabela com os resultados mostra, ainda, que a substituição de valores ausentes em séries temporais caóticas é sensível ao comportamento e distribuição dos valores ausentes, assim como apontado no Capítulo 3 para séries temporais não caóticas.

Pode-se verificar que, aparentemente, a técnica RF teve um desempenho pior em relação as outras 3 técnicas. No entanto, esse comportamento não indica que ela não

deve ser utilizada no método proposto. Isso implica que a técnica RF pode exigir mais esforços para encontrar um modelo que entregue bons resultados. Vale ressaltar que o modelo gerado pela técnica RF é um modelo que depende muito de ajustes de seus hiperparâmetros e da seleção de atributos. Sendo assim, nos experimentos conduzidos neste trabalho, esta técnica pode tender a apresentar resultados não satisfatórios por dispor de poucos atributos para o treinamento do modelo.

O modelo gerado pela técnica SVR apresenta resultados similares a técnica KNN e com resultados ainda melhores para a série de Mapa Logístico no geral. Entretanto, para as séries de Lorenz e Rössler a discrepância entre seus resultados é mais alta. Uma outra vantagem das técnicas KNN e DWNN sobre a técnica SVR é o tempo de execução, afinal o treinamento do modelo e seus hiperparâmetros é muito mais complexo que o cálculo dos vizinhos mais próximos. Logo, ao escolher a técnica a ser utilizada para a estimativa do valor ausente, deve-se analisar seu custo computacional e espacial para garantir, por exemplo, possibilidade de execução em ambientes limitados e com restrições temporais.

Por fim, pode-se concluir que, para os experimentos realizados, as técnicas KNN e DWNN apresentaram resultados mais consistentes. Considerando, ainda, que as mesmas têm implementações e aplicações simples, elas têm potencial para ser um modo padrão (*default*) de execução de preenchimento de valores ausentes em grandes lotes de séries temporais caóticas. Entretanto, destaca-se que a principal vantagem desses experimentos é enfatizar a possibilidade de uso de qualquer algoritmo de AM na substituição de valores ausentes, de acordo com a preferência e necessidade do usuário.

6.3.3 Variação completa

A última fase de experimentação consistiu em inserir lacunas de valores ausentes com variação de tamanho, frequência e posição na série temporal avaliada. Para este experimento foi utilizada apenas a técnica KNN por apresentar melhor desempenho nos experimentos anteriores. No entanto, assim como foi destacado anteriormente, qualquer técnica de AM pode ser aplicada nesse cenário.

As lacunas sintéticas foram definidas pela quantidade de informação retirada da série temporal original. O experimento foi executado com perda de 1%, 5%, 10% e 15%. Escolheu-se limitar a perda até um máximo de 15% porque valores maiores descaracterizam as séries, impedindo uma avaliação adequada do método proposto. Entretanto, destaca-se que o método pode ser aplicado em situações reais com maiores taxas de valores ausentes.

Analisando a substituição na série de Lorenz, Figura 6.4, pode-se identificar comportamento similar aos resultados do experimento multilacuna. Ao aumentar a perda de dados, o erro também cresce drasticamente, como esperado. É possível verificar ainda grande variação entre cada experimento na série de Lorenz. Mesmo com perda de apenas 1%, a série já apresenta alta discrepância entre os erros. Dentre as séries apresentadas nesta dissertação, a série de Lorenz é a mais complexa e isso é refletido nos experimentos.

Os experimentos de variação completa com a série de Hénon, Figura 6.5, também apresentam comportamento similar aos anteriores, com baixa variação entre os experimentos mesmo aumentando a perda de informações e independente da técnica de AM

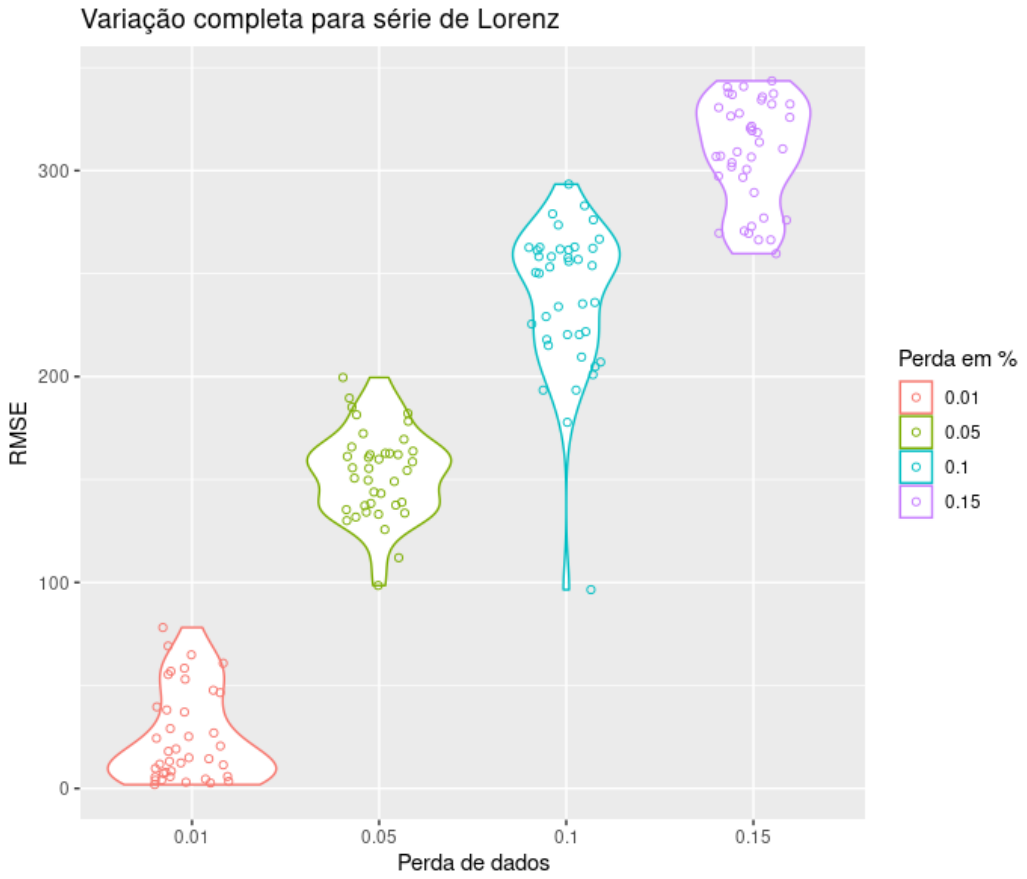


Figura 6.4 Distribuição dos erros do experimento de variação completa sobre a série de Lorenz

utilizada.

Assim como Hénon, a série do Mapa Logístico apresenta resultados similares aos anteriores, com baixa variação entre os erros no experimentos e foi a série que apresentou melhores resultados.

A série de Rössler, por outro lado, apresentou um comportamento diferente. Experimentos com variação na posição da janela mostraram menor variação nos erros, como pode ser conferido na Figura A.5. No entanto, os resultados dos experimentos multilacuna (Figura 6.5) e de variação completa (Figura 6.7) mostram uma variação mais alta com o aumento da perda dos dados. Ou seja, a série de Rössler apresenta mais sensibilidade à perda de dados e menos em relação à posição em que a lacuna aparece.

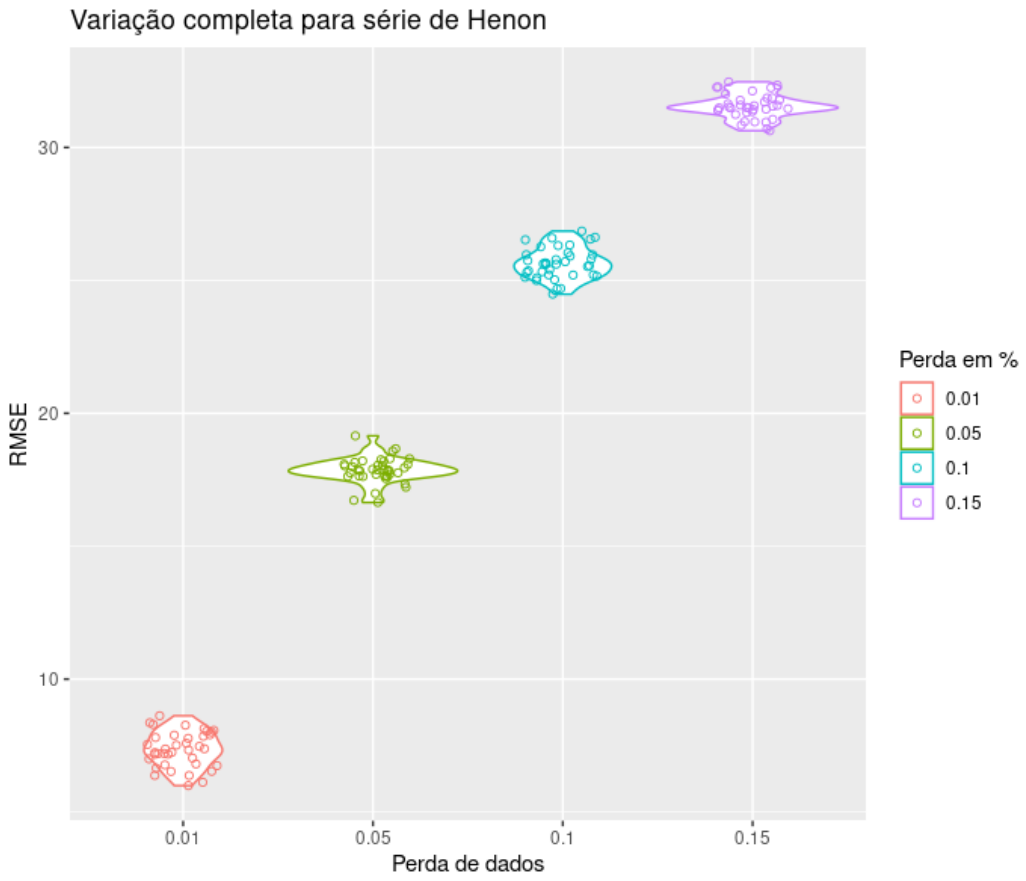


Figura 6.5 Distribuição dos erros do experimento de variação completa sobre a série de Henon.

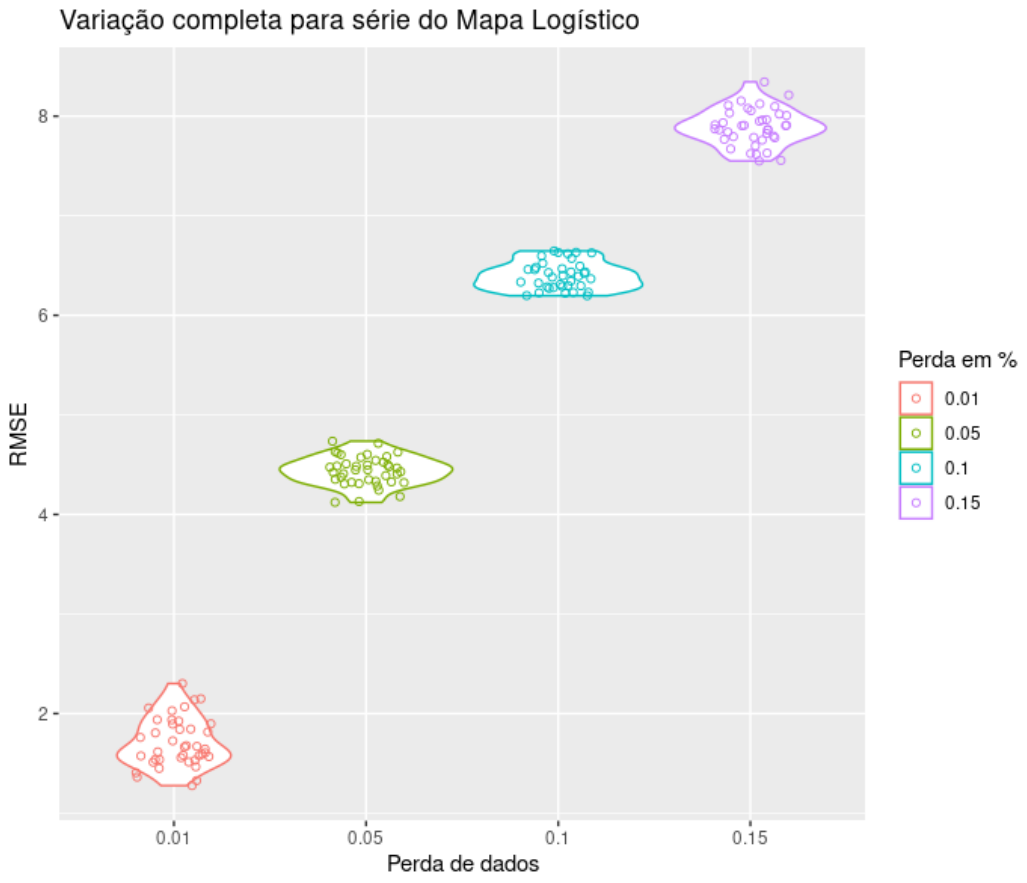


Figura 6.6 Distribuição dos erros do experimento de variação completa sobre a série do Mapa Logístico.

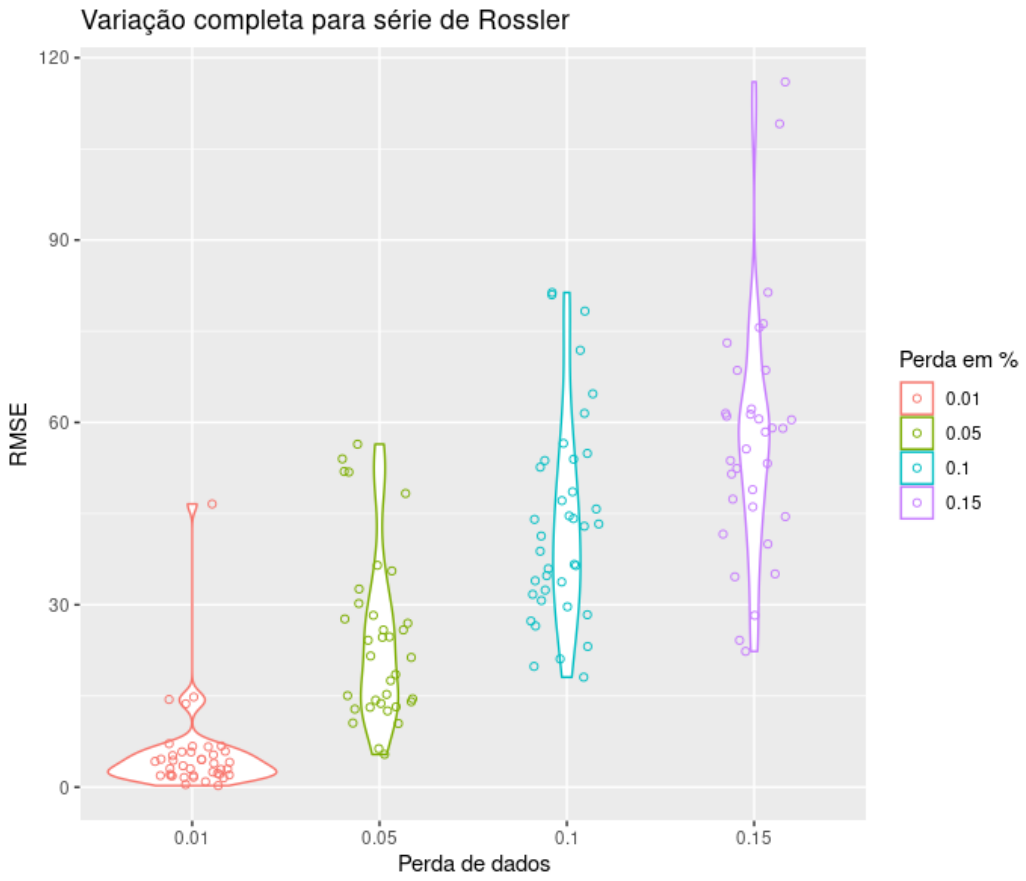


Figura 6.7 Distribuição dos erros do experimento de variação completa sobre a série de Rössler.

Capítulo 7

CONCLUSÃO

Esta dissertação discorreu sobre o problema de perda de dados em séries temporais, o qual afeta a modelagem de vários problemas reais em diferentes áreas do conhecimento. A Revisão Sistemática da Literatura (RSL), conduzida neste trabalho, constatou que o estudo de ferramentas para substituição de valores ausentes em séries temporais é um trabalho de alta relevância. Além disso, observou-se que o foco principal dos trabalhos identificados estavam focados no domínio temporal. A ausência de trabalhos realizados no espaço de coordenadas de atraso, comumente utilizado para modelar séries caóticas, motivou a realização desta pesquisa. Em resumo, dentre as diferentes definições de séries temporais, optou-se por investigar o problema em um contexto específico: a substituição de valores ausentes em séries temporais caóticas.

De maneira resumida, o principal objetivo desta pesquisa foi comprovar a seguinte hipótese: *A substituição de valores ausentes em séries temporais caóticas é realizada com maior acurácia quando modelos são aplicados sobre observações transformadas no espaço de coordenadas de atraso.* Visando demonstrar a viabilidade desta hipótese, resultados foram obtidos a partir de séries caracterizadas com diferentes padrões de valores ausentes como, por exemplo, o número de observações com problemas e a quantidade de janelas com tais observações.

Como consequência, dois novos métodos para substituição desses valores ausentes foram propostos. O primeiro método, chamado de BMDE, é focado na análise ao longo do tempo. Esse método apresentou resultados comparáveis às principais abordagens do estado da arte, com complexidade consideravelmente inferior, e foi publicado na principal conferência de Inteligência Artificial organizada pela Sociedade Brasileira de Computação: **8th Brazilian Conference on Intelligent Systems (BRACIS)** (DOI ⟨10.1109/BRACIS.2019.00138⟩). O segundo método, chamado PSGF, que até o momento da elaboração dessa dissertação encontra-se em fase de submissão, tem como principal objetivo modelar séries temporais caóticas com valores ausentes no espaço fase. Os resultados obtidos com esse método enfatizaram a importância de um estudo conduzido especificamente para modelagem de séries caóticas. Uma importante característica do método

proposto é a combinação de ferramentas de modelagem de séries temporais caóticas e algoritmos conhecidos da área de Aprendizado de Máquina (AM). Essa combinação fornece uma flexibilidade para que séries caóticas sejam modeladas usando diferentes vieses de aprendizado, dependendo do tipo de sistema usado para a produção das séries analisadas.

É importante destacar, ainda, que os métodos propostos foram avaliados em diferentes cenários, desde simples janelas de dados ausentes até casos extremos com diferentes janelas com tamanhos variados. Os resultados obtidos enfatizam a importância da proposta para o estado da arte de modelagem de séries caóticas. De acordo com a RSL e o conhecimento atual dos pesquisadores envolvidos, esse é o primeiro trabalho desenvolvido para substituição de valores ausentes, que combina ferramentas de Sistemas Dinâmicos e Teoria do Caos com Aprendizado de Máquina.

Apesar dos resultados promissores apresentados nesta dissertação, há limitações no método proposto que podem ser exploradas em trabalhos futuros para aumentar sua capacidade de aprendizado e generalização, possibilitando um desempenho ainda maior na substituição de valores ausentes. Uma dessas limitações está relacionada à presença de lacunas de valores ausentes próximas aos extremos da série. Nesse cenário, pode-se optar por uma regressão tradicional, uma vez que não há observações após as lacunas para auxiliar na modelagem do espaço fase.

Uma segunda lacuna importante é a avaliação das predições utilizadas para substituir valores ausentes, a qual ocorre no espaço tempo, i.e., após a substituição dos valores ausentes, o método proposto reconstrói a série do espaço fase para o domínio temporal e, em seguida, aplica métricas comumente utilizadas na predição de séries temporais. Como trabalho futuro, espera-se avaliar os erros diretamente no espaço fase utilizando técnicas bem fundamentadas na literatura como, por exemplo, as métricas derivadas do método *Recurrence Plots* (ECKMANN et al., 1995): *Cross Recurrence Plots* e *Joint Recurrence Plots*.

Ainda como trabalho futuro, espera-se disponibilizar um pacote em R com todas as funcionalidades para que a comunidade de séries temporais possam reproduzir os resultados obtidos nesta dissertação e utilizá-lo em outras aplicações. Por fim, espera-se modelar um problema real usando a proposta dessa dissertação como, por exemplo, no estudo de dados que monitoram o clima. É sabido que mudanças climáticas podem ser modeladas por atratores, como Lorenz. Assim, o método PSGF pode ser útil para preencher possíveis valores ausentes. Uma outra aplicação relacionada é a utilização do método proposto em dados de monitoramento de rios para suportar o processo de tomada de decisão em situações críticas como as enchentes que causam diversos problemas no Brasil, vitimando centenas de pessoas todos os anos.

APÊNDICE A

RESULTADOS COMPLEMENTARES OBTIDOS COM O MÉTODO PROPOSTO

Esse apêndice apresenta um conjunto de resultados que foram obtidos com o método proposto. Visando manter o texto da dissertação conciso, esses resultados foram retirados do texto principal. No entanto, considerando sua importância para avaliar o método proposto em diferentes cenários, optou-se por apresentá-los nesta seção complementar.

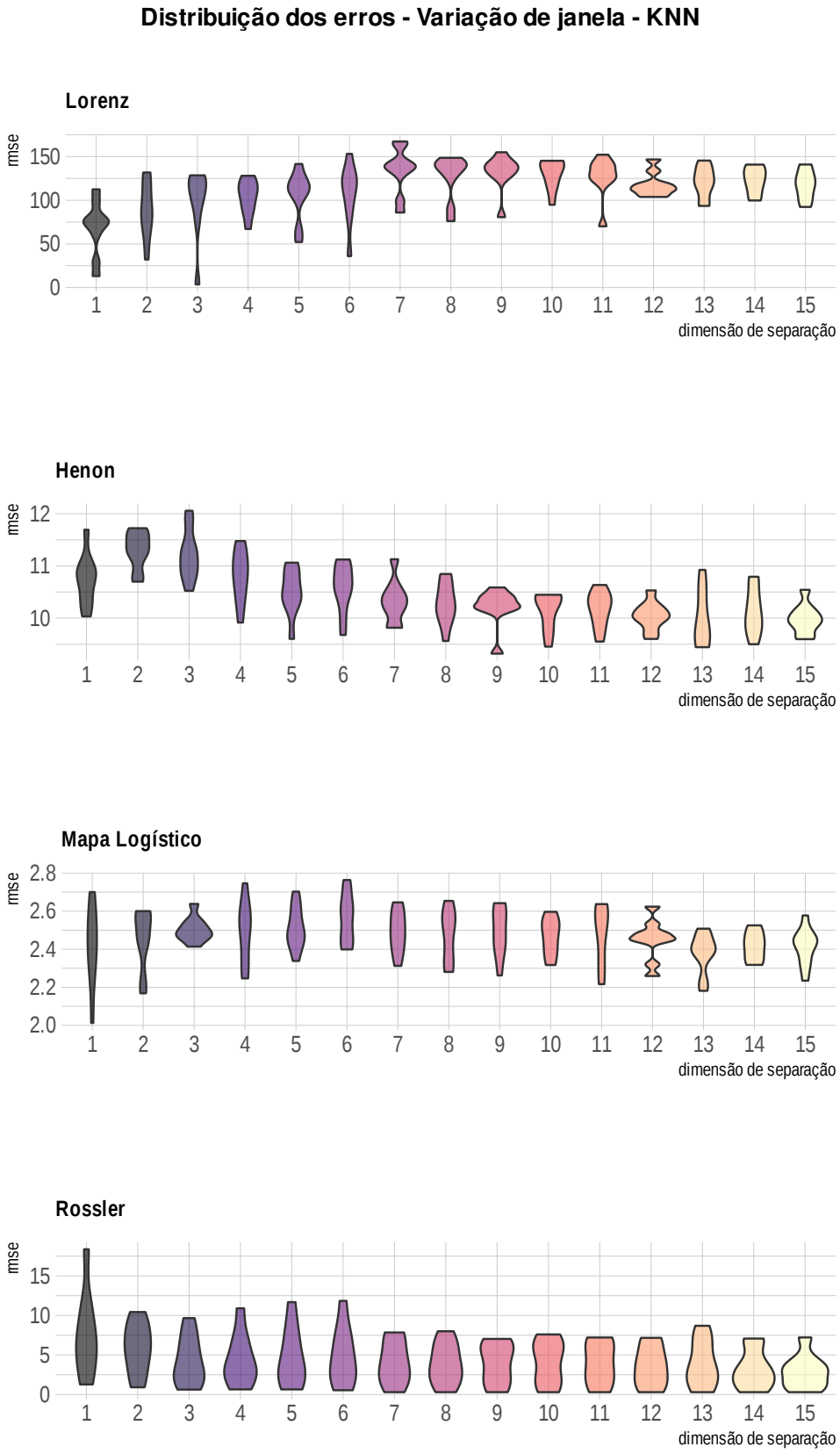


Figura A.1 Distribuição dos erros do PSGF + KNN para cada série e cada dimensão de separação. O método apresenta mais estabilidade sobre as séries do Mapa Logístico e a série de Rössler

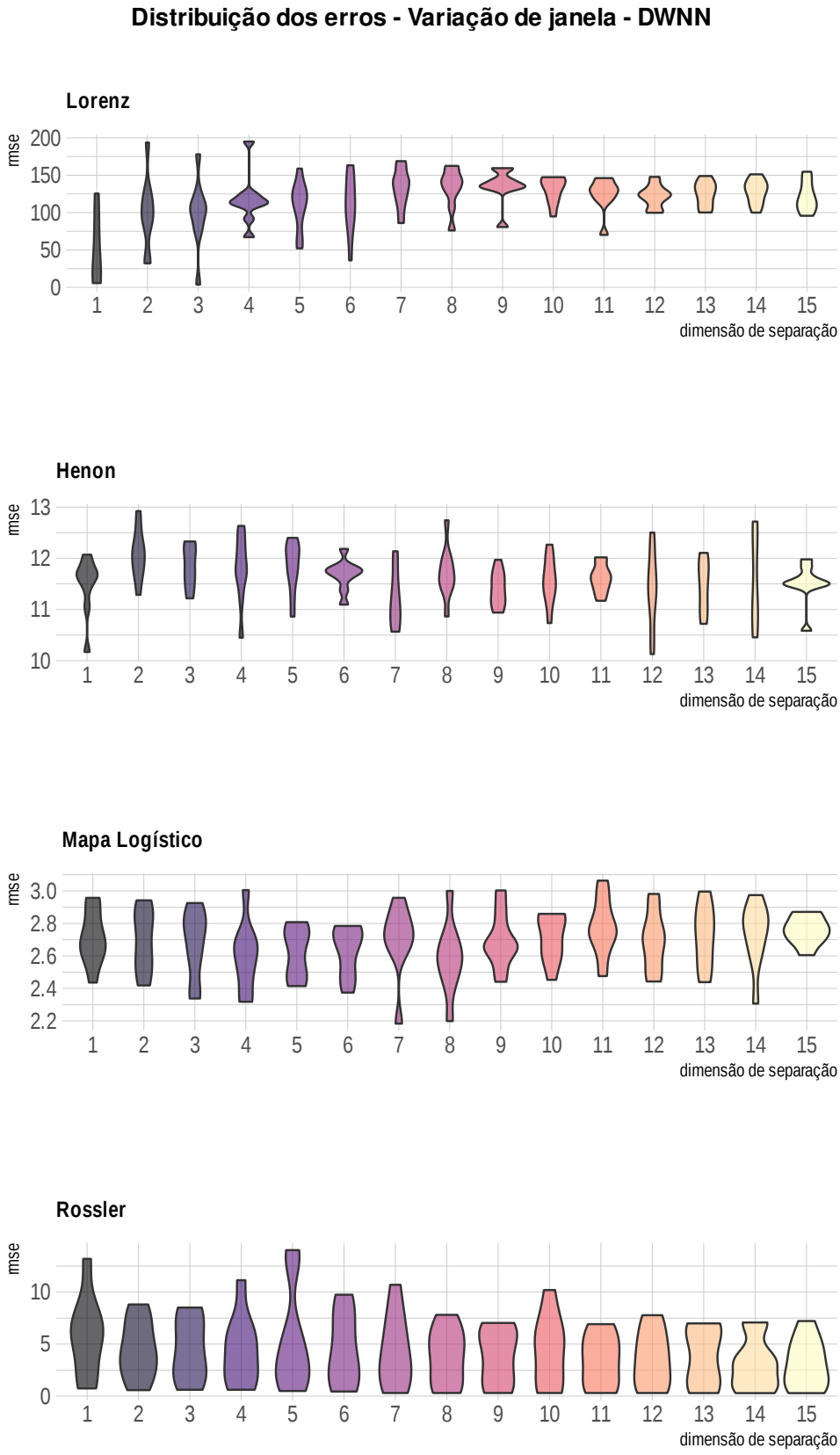


Figura A.2 Distribuição dos erros do PSGF + DWNN para cada série e cada dimensão de separação. A série de Rössler foi a única que apresentou uma melhoria sobre a técnica KNN

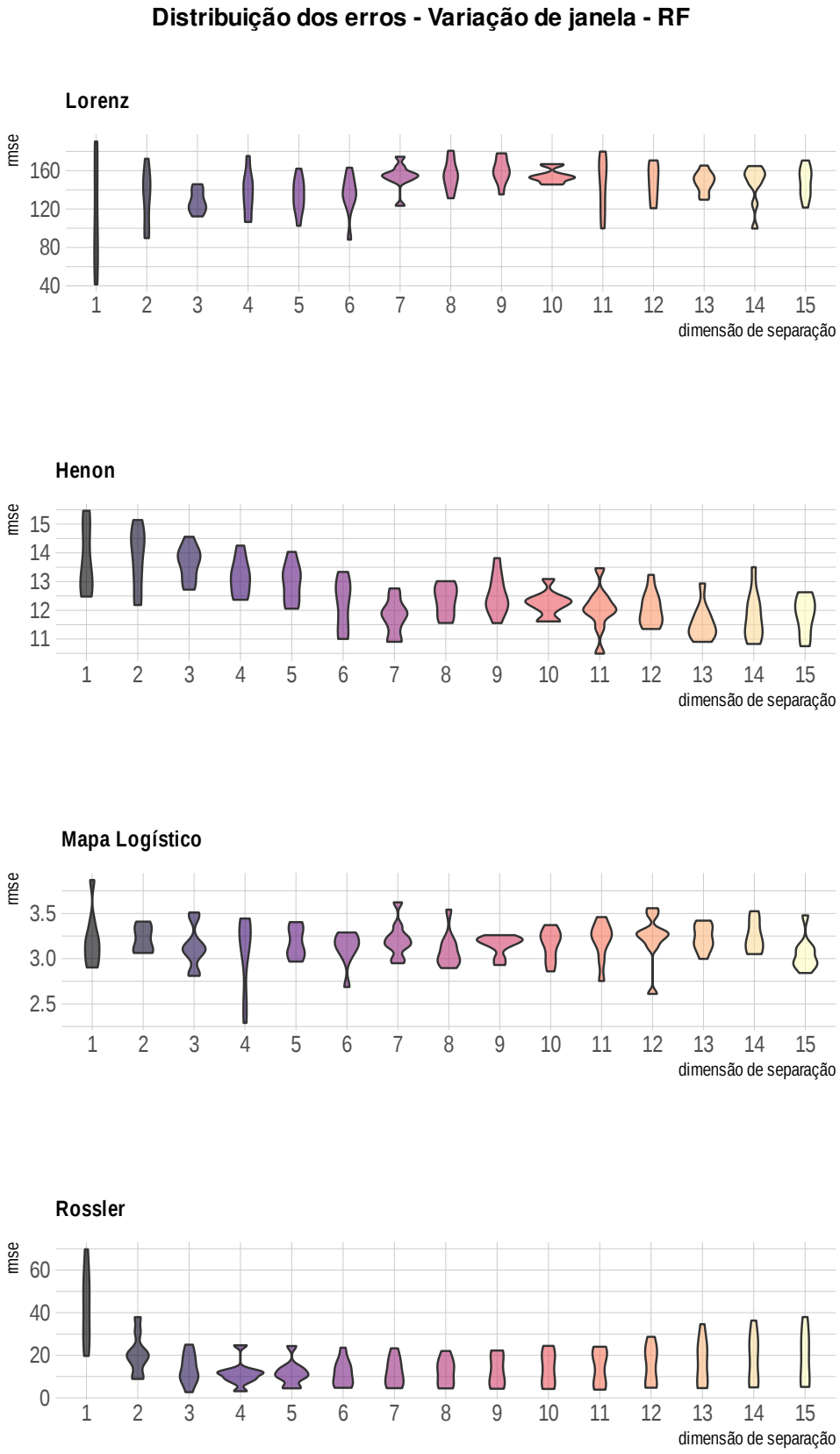


Figura A.3 Distribuição dos erros do PSGF + RF para cada série e cada dimensão de separação. A técnica RF não apresentou nenhuma melhoria sobre as outras técnicas.

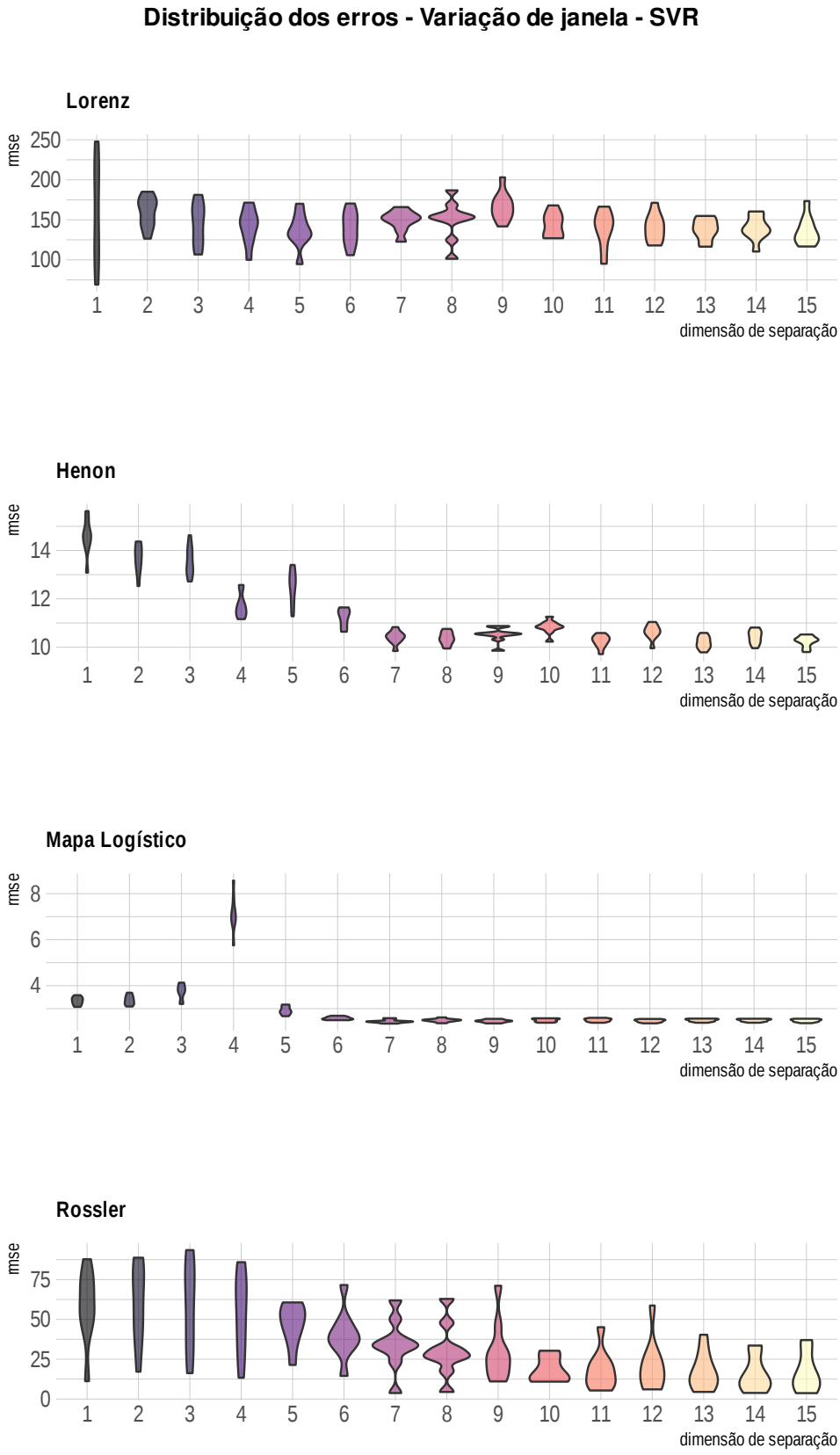


Figura A.4 Distribuição dos erros do PSGF + SVR para cada série e cada dimensão des separação. A técnica SVR apresentou altos índices de erros.

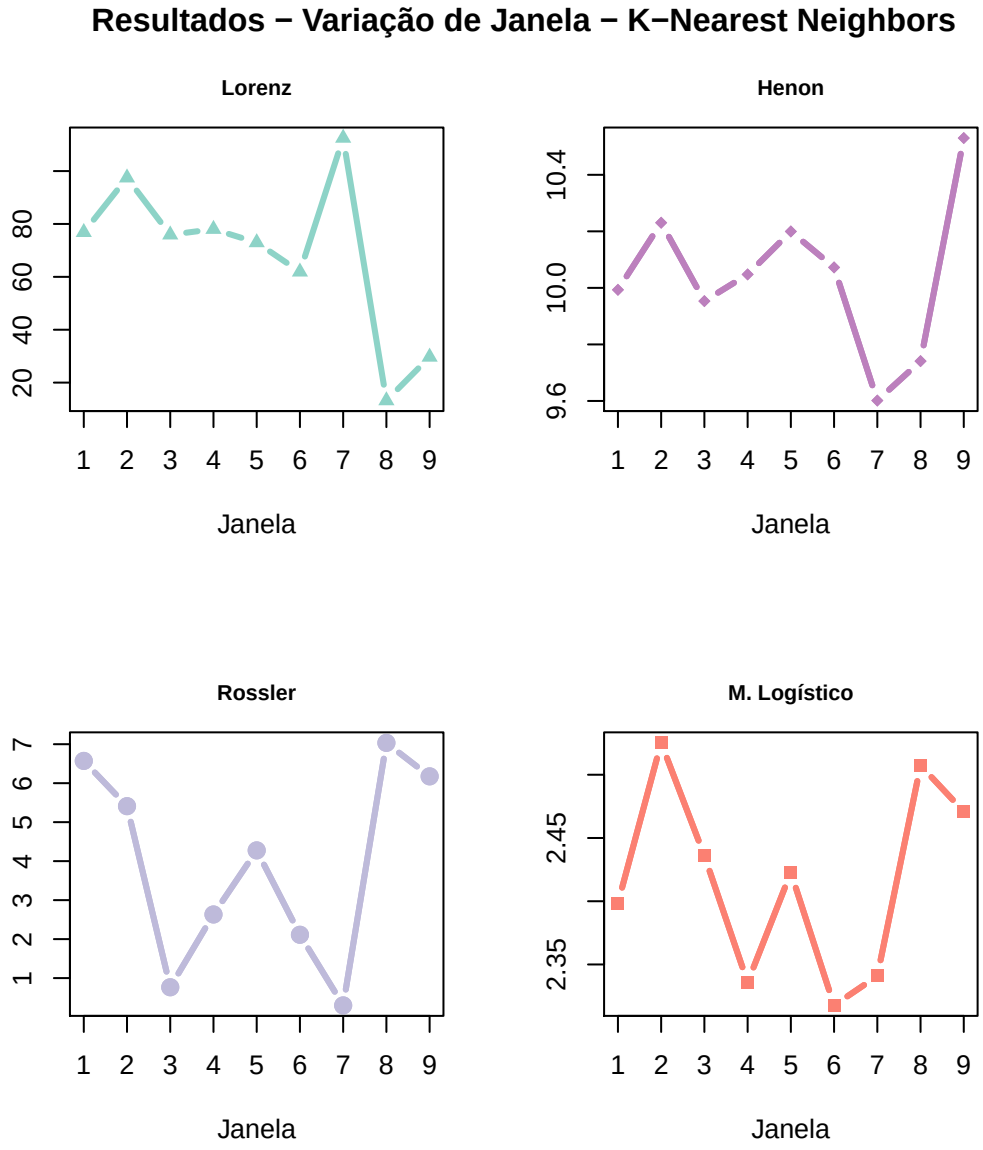


Figura A.5 Resultados dos experimentos de variação de janela. Cruzamento das dimensões de separação de menores erros com a posição das lacunas. Nota-se que a posição da lacuna afeta as séries de maneiras diferentes entre si.

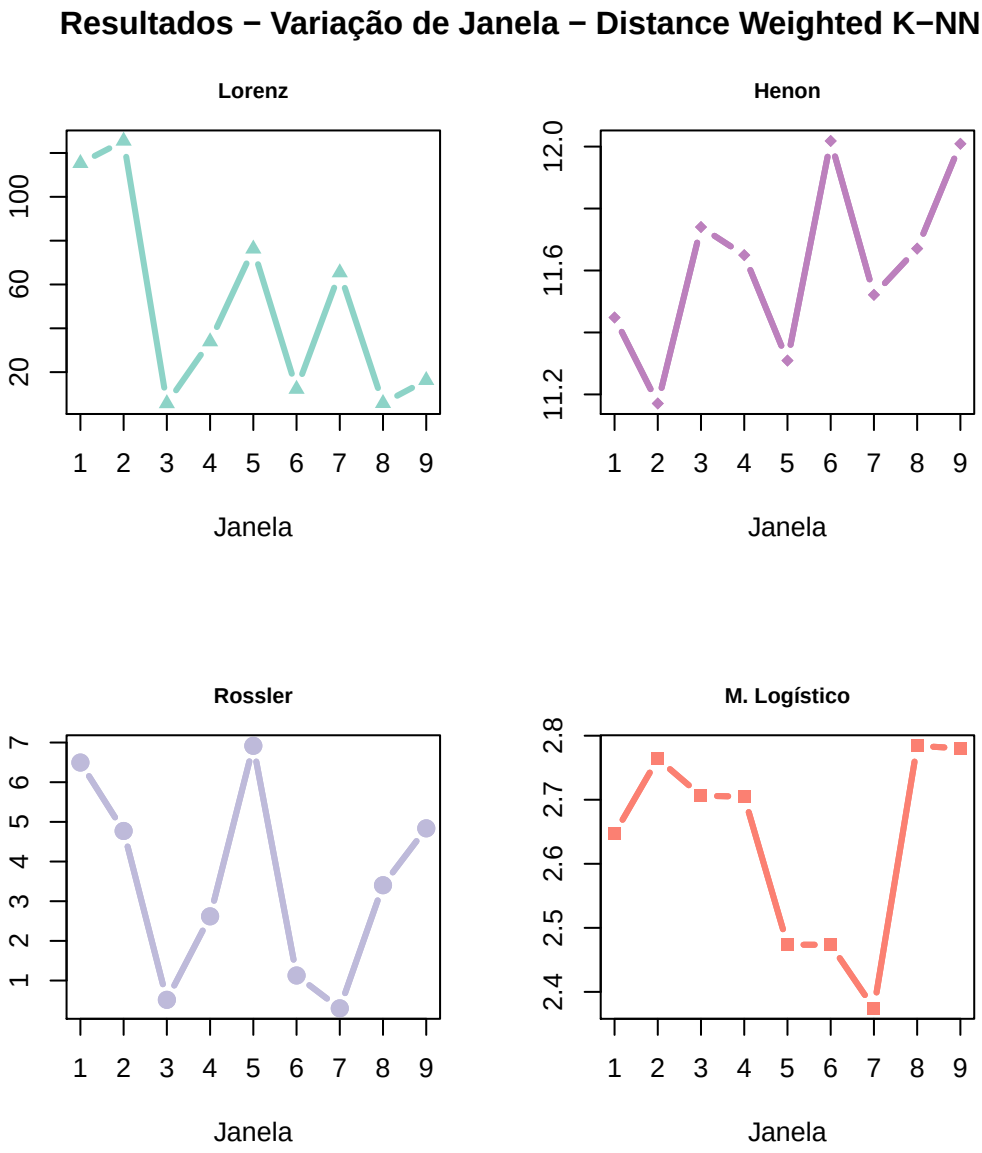


Figura A.6 Resultados dos experimentos de variação de janela. Cruzamento das dimensões de separação de menores erros com a posição das lacunas. Nesses experimentos as séries de Lorenz, Rossler e o Mapa Logístico apresentam maiores erros quando a posição da lacuna está localizada nas primeiras janelas.

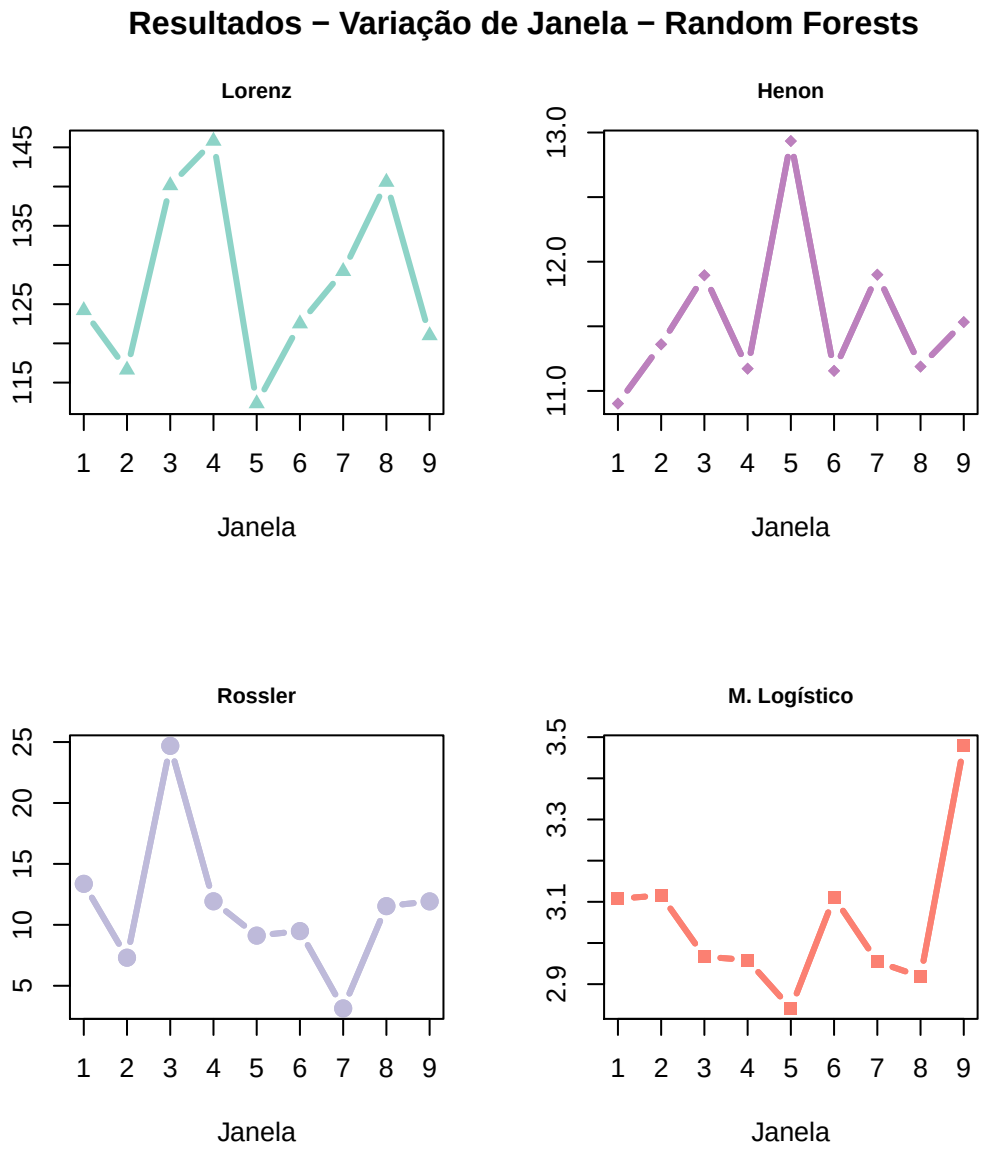


Figura A.7 Resultados dos experimentos de variação de janela. Cruzamento das dimensões de separação de menores erros com a posição das lacunas. Novamente as séries não apresentam similaridade entre seus resultados e suas posições de janela.

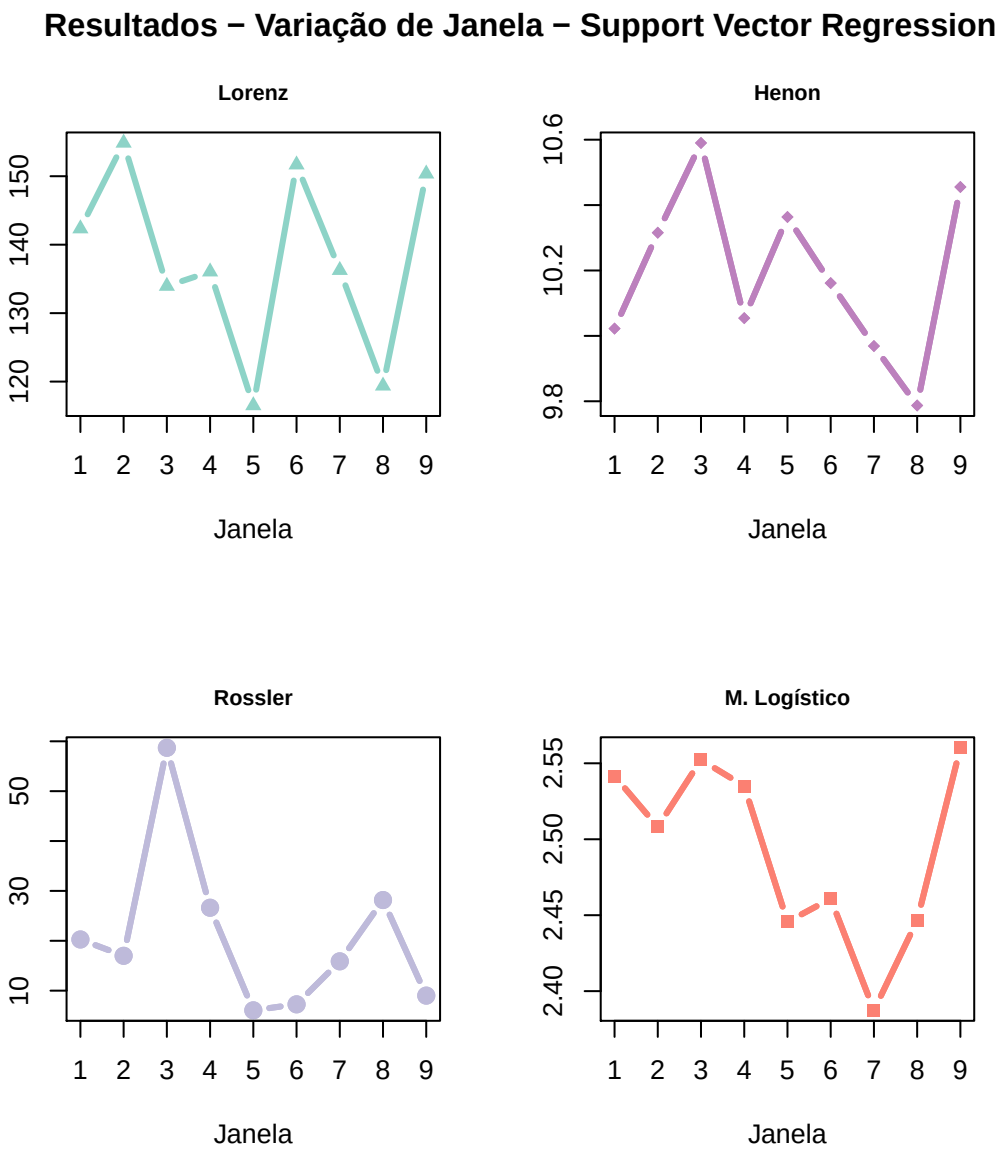


Figura A.8 Resultados dos experimentos de variação de janela. Cruzamento das dimensões de separação de menores erros com a posição das lacunas. É possível perceber que séries de Henon, Rossler e o Mapa Logístico apresentam picos semelhantes na posição 3 da janela de valores ausentes.

REFERÊNCIAS BIBLIOGRÁFICAS

ALLIGOOD, K.; SAUER, T.; YORKE, J. *Chaos: An Introduction to Dynamical Systems*. [S.l.]: Springer New York, 1997. (Textbooks in Mathematical Sciences).

ALLIGOOD, K. T.; SAUER, T. D.; YORKE, J. A. *Chaos: An Introduction to Dynamical Systems. 1996*. [S.l.]: Springer-Verlag, 1997.

AWAD, M.; KHANNA, R. Support vector regression. In: _____. *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. Berkeley, CA: Apress, 2015. p. 67–80. ISBN 978-1-4302-5990-9. Disponível em: ⟨https://doi.org/10.1007/978-1-4302-5990-9_4⟩.

BARALDI, P. et al. A fuzzy similarity based method for signal reconstruction during plant transients. *Chemical Engineering Transactions*, v. 33, p. 889–894, 2013.

BOX, G.; JENKINS, G. M.; REINSEL, G. *Time Series Analysis: Forecasting & Control*. 3ª. ed. [S.l.]: Prentice Hall, 1994. Hardcover. ISBN 0130607746.

BRáS, L.; MENEZES, J. Improving cluster-based missing value estimation of dna microarray data. *Biomolecular Engineering*, v. 24, p. 273–282, 2007.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001.

CANO, S.; ANDREU, J. Using multiple imputation to simulate time series. In: . [S.l.: s.n.], 2010. p. 117–122.

CAPIZZI, G.; NAPOLI, C.; PATERNò, L. An innovative hybrid neuro-wavelet method for reconstruction of missing data in astronomical photometric surveys. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, v. 7267 LNAI, p. 21–29, 2012.

ECKMANN, J.-P. et al. Recurrence plots of dynamical systems. *World Scientific Series on Nonlinear Science Series A*, WORLD SCIENTIFIC PUBLISHING, v. 16, p. 441–446, 1995.

FABBRI, S. et al. Improvements in the start tool to better support the systematic review process. In: *Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering*. New York, NY, USA: ACM, 2016. (EASE '16), p. 21:1–21:5. ISBN 978-1-4503-3691-8. Disponível em: ⟨http://doi.acm.org/10.1145/2915970.2916013⟩.

FACCHINI, A.; MOCENNI, C. Filling gaps in ecological time series by means of twin surrogates. *International Journal of Bifurcation and Chaos*, v. 21, p. 1085–1097, 2011.

Faceli, K. et al. *Inteligência Artificial – Uma Abordagem de Aprendizado de Máquina*. 1st. ed. [S.l.]: LTC, 2015.

FRASER, A. M.; SWINNEY, H. L. Independent coordinates for strange attractors from mutual information. *Physical review A*, APS, v. 33, n. 2, p. 1134, 1986.

GOLYANDINA, N.; NEKRUTKIN, V.; ZHIGLJAVSKY, A. A. *Analysis of time series structure: SSA and related techniques*. [S.l.]: CRC press, 2001.

GOLYANDINA, N.; OSIPOV, E. The "caterpillar-ssa method for analysis of time series with missing values. *Journal of Statistical Planning and Inference*, v. 137, p. 2642–2653, 2007.

GOLYANDINA, N.; OSIPOV, E. The "caterpillar"-ssa method for analysis of time series with missing values. *Journal of Statistical planning and Inference*, Elsevier, v. 137, n. 8, p. 2642–2653, 2007.

HÄRDLE, W.; HOROWITZ, J.; KREISS, J.-P. Bootstrap methods for time series. *International Statistical Review*, Wiley Online Library, v. 71, n. 2, p. 435–459, 2003.

HASSANI, H. Singular spectrum analysis: methodology and comparison. *Journal of Data Science*, Cardiff University and Central Bank of the Islamic Republic of Iran, v. 5, n. 2, p. 239–257, 2007.

HASTIE, T. et al. Imputing missing data for gene expression arrays. Stanford University Statistics Department Technical report, 1999.

HÉNON, M. A two-dimensional mapping with a strange attractor. In: *The theory of chaotic attractors*. [S.l.]: Springer, 1976. p. 94–102.

HONG, B.; CHEN, C. Radial basis function neural network-based nonparametric estimation approach for missing data reconstruction of non-stationary series. In: . [S.l.: s.n.], 2003. v. 1, p. 75–78.

HU, J. et al. Integrative missing value estimation for microarray data. *BMC Bioinformatics*, v. 7, 2006.

HUANG, N. E. et al. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. In: THE ROYAL SOCIETY. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*. [S.l.], 1998. v. 454, n. 1971, p. 903–995.

HUO, J. et al. Innovative missing data replacement methods using time series models. In: . [S.l.: s.n.], 2008. v. 316.

JUNGER, W.; LEON, A. Ponce de. Imputation of missing data in time series for air pollutants. *Atmospheric Environment*, v. 102, p. 96–104, 2015.

KENNEL, M. B.; BROWN, R.; ABARBANEL, H. D. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical review A*, APS, v. 45, n. 6, p. 3403, 1992.

KIM, J.-W.; PACHEPSKY, Y. Reconstructing missing daily precipitation data using regression trees and artificial neural networks for swat streamflow simulation. *Journal of Hydrology*, v. 394, p. 305–314, 2010.

KITCHENHAM, B. et al. Systematic literature reviews in software engineering - a systematic literature review. *Information and Software Technology*, v. 51, n. 1, p. 7 – 15, 2009. ISSN 0950-5849. Special Section - Most Cited Articles in 2002 and Regular Research Papers. Disponível em: ⟨http://www.sciencedirect.com/science/article/B6V0B-4TX182T-2/2/d714d8469c560c40f3cdb6bce5534036⟩.

KRAFT, R. L. Chaos, cantor sets, and hyperbolicity for the logistic maps. *The american mathematical monthly*, Taylor & Francis, v. 106, n. 5, p. 400–408, 1999.

LITTLE, R. J.; RUBIN, D. B. *Statistical analysis with missing data*. [S.l.]: John Wiley & Sons, 2019.

MELLO, R. F.; PONTI, M. A. *Machine Learning - A Practical Approach on the Statistical Learning Theory*. [S.l.]: Springer, 2018. ISBN 978-3-319-94988-8.

Mitchell, T. M. et al. *Machine learning. WCB*. [S.l.]: McGraw-Hill Boston, MA:, 1997.

MOGHTADERI, A.; BORGNAT, P.; FLANDRIN, P. Gap-filling by the empirical mode decomposition. In: IEEE. *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. [S.l.], 2012. p. 3821–3824.

MORETTIN, P. A.; TOLOI, C. Análise de séries temporais. In: *Análise de séries temporais*. [S.l.]: Blucher, 2006.

OLINSKY, A.; CHEN, S.; HARLOW, L. The comparative efficacy of imputation methods for missing data in structural equation modeling. *European Journal of Operational Research*, Elsevier, v. 151, n. 1, p. 53–79, 2003.

ÖZKÖSE, H.; ARI, E. S.; GENCER, C. Yesterday, today and tomorrow of big data. *Procedia - Social and Behavioral Sciences*, v. 195, p. 1042–1050, 2015. ISSN 1877-0428.

PAGLIOSA, L. de C.; MELLO, R. F. de. Semi-supervised time series classification on positive and unlabeled problems using cross-recurrence quantification analysis. *Pattern Recognition*, v. 80, p. 53 – 63, 2018. ISSN 0031-3203. Disponível em: ⟨http://www.sciencedirect.com/science/article/pii/S0031320318300815⟩.

PLAIA, A.; BONDI, A. Single imputation method of missing values in environmental pollution data sets. *Atmospheric Environment*, v. 40, p. 7316–7330, 2006.

PRASOMPHAN, S.; LURSINSAP, C.; CHIEWCHANWATTANA, S. Imputing time series data by regional-gradient-guided bootstrapping algorithm. In: . [S.l.: s.n.], 2009. p. 163–168.

QU, L. et al. Ppca-based missing data imputation for traffic flow volume: A systematical approach. *IEEE Transactions on Intelligent Transportation Systems*, v. 10, p. 512–522, 2009.

RIBEIRO, R. G.; RIOS, R. Temporal gap statistic: A new internal index to validate time series clustering. *Chaos, Solitons & Fractals*, Elsevier, v. 142, p. 110326, 2021.

RIOS, R. A.; MELLO, R. F. D. Improving time series modeling by decomposing and analyzing stochastic and deterministic influences. *Signal Processing*, Elsevier, v. 93, n. 11, p. 3001–3013, 2013.

RIOS, R. A.; MELLO, R. F. de. Improving time series modeling by decomposing and analyzing stochastic and deterministic influences. *Signal Processing*, v. 93, n. 11, p. 3001 – 3013, 2013. ISSN 0165-1684.

RIOS, R. A.; MELLO, R. F. de. Applying empirical mode decomposition and mutual information to separate stochastic and deterministic influences embedded in signals. *Signal Processing*, v. 118, p. 159 – 176, 2016. ISSN 0165-1684.

RIPLEY, B. D. *Pattern recognition and neural networks*. [S.l.]: Cambridge university press, 2006.

ROTH, M. c.; ZHUGZHDA, Y. b. Gapfilling interrupted helioseismic data with the em algorithm. *Astronomy Letters*, v. 36, p. 64–73, 2010.

RUGGIERO, M.; LOPES, V. da R. *Cálculo numérico: aspectos teóricos e computacionais*. [S.l.]: Makron Books do Brasil, 1996. ISBN 9788534602044.

SCHMIDT, A.; WRZESINSKY, T.; KLEMM, O. Gap filling and quality assessment of co2 and water vapour fluxes above an urban area with radial basis function neural networks. *Boundary-Layer Meteorology*, v. 126, p. 389–413, 2008.

SCOTT, S. K. *Chemical chaos*. [S.l.]: Oxford University Press, 1993.

SENTAS, P.; ANGELIS, L. Categorical missing data imputation for software cost estimation by multinomial logistic regression. *Journal of Systems and Software*, v. 79, p. 404–414, 03 2006.

SOLEYMANI, A.; NORDIN, M. J.; SUNDARARAJAN, E. A chaotic cryptosystem for images based on henon and arnold cat map. *The Scientific World Journal*, Hindawi, v. 2014, 2014.

SONG, Y. et al. An efficient instance selection algorithm for k nearest neighbor regression. *Neurocomputing*, Elsevier, v. 251, p. 26–34, 2017.

TAKENS, F. Detecting strange attractors in turbulence. In: *Dynamical systems and turbulence, Warwick 1980*. [S.l.]: Springer, 1981. p. 366–381.

TARDIVO, G.; BERTI, A. A dynamic method for gap filling in daily temperature datasets. *Journal of Applied Meteorology and Climatology*, v. 51, p. 1079–1086, 2012.

UYSAL, M. Reconstruction of time series data with missing values. *Journal of Applied Sciences*, v. 7, p. 922–925, 2007.

VERGER, A. b. et al. The cacao method for smoothing, gap filling, and characterizing seasonal anomalies in satellite time series. *IEEE Transactions on Geoscience and Remote Sensing*, v. 51, n. 4, p. 1963–1972, 2013.

WHITNEY, H. Differentiable manifolds. *Annals of Mathematics*, JSTOR, p. 645–680, 1936.

YIGIT, H. A weighting approach for knn classifier. In: IEEE. *2013 international conference on electronics, computer and computation (ICECCO)*. [S.l.], 2013. p. 228–231.

ZHANG, W.-F.; LIU, C.-C.; YAN, H. Temporal gene expression profiles reconstruction by support vector regression and framelet kernel. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, v. 6064 LNCS, p. 68–74, 2010.

ZHANG, Y. b. c. et al. Restoration of missing time-series data via multiple sine functions decomposition with guangzhou-temperature application. In: . [S.l.: s.n.], 2014. p. 459–464.

ZIKOPOULOS, P. et al. *Big Data Beyond the Hype – A Guide to Conversations for Today's Data Center*. [S.l.]: McGraw-Hill Education, 2015. ISBN: 978-0-07-184466-6.

ZIKOPOULOS, P. C. et al. *Harness the Power of Big Data*. [S.l.]: McGraw-Hill Education, 2015. ISBN: 978-0-07-184466-6.