XL CILAMCE
IBERO-LATIN AMERICAN CONGRESS ON COMPUTATIONAL METHODS IN ENGINEERING
NOVEMBER 11-14, 2019
Praiamar Natal Hotel & Convention
Natal, RN-BRAZIL

MISSING DATA: BRIEF REVIEW AND CASE STUDIES

MISSING DATA: BREVE REVISÃO E ESTUDOS DE CASOS

Giovanni Amormino da Silva Júnior
Alisson M. Silva
Paulo E. M. Almeida
gigio_jr@hotmail.com
alisson@cefetmg.br
pema@cefetmg.br
Programa de Pós-Graduação em Modelagem Matemática e Computacional
CEFET-MG - Centro Federal de Educação Tecnológica de Minas Gerais

Abstract. A major difficulty faced in developing real applications that use data streams to solve prediction and/or classification problems are the missing data. Although there are techniques to reduce the impacts caused by this problem, most systems are not preventively modeled to allow the adequate treatment of this type of occurrence. Unreliable or damaged sensors, partial occlusion, interference in communication, restrictions in the data transmission band, are some of the reasons that can cause this problem. Some authors classify this lack of data regarding their randomness and show that in some cases this absence may also be related to the lack of other attributes. Basically there are two ways of dealing with missing data: i) exclusion, where all or part of the sample is removed or ignored; and ii) imputation, where the missing value is replaced by zero, by the average of the variable up to the sample with the problem or by an estimated value, where the problematic variable has its estimated value by some model that, in some cases, can lead to variables and/or previous values. In this context, this article presents a review of the literature addressing the main methodologies used to address the problem of missing data. In addition, case studies are presented for comparing results and defining situations where each type of approach is best employed.

Keywords: Missing Data, Missing Completely at Random (MCAR), Missing at Random (MAR)

Resumo. Uma grande dificuldade enfrentada no desenvolvimento de aplicações reais que utilizam fluxos de dados para resolver problemas de previsão e/ou classificação são os dados ausentes. Apesar de existirem técnicas para reduzir os impactos ocasionados por este problema, a maioria dos sistemas não são modelados de forma preventiva para possibilitar o tratamento adequado deste tipo de ocorrência. Sensores não confiáveis ou danificados, oclusão parcial, interferência na comunicação, restrições na banda de transmissão de dados, são alguns dos motivos que podem ocasionar este problema. Alguns autores classificam esta ausência dos dados com relação a sua aleatoriedade e mostram que, em alguns casos, esta ausência pode também estar relacionada a falta de outros atributos. Basicamente existem duas formas de lidar com os dados ausentes: i) exclusão, onde toda ou parte da amostra é removida ou ignorada; e ii) imputação, onde o valor ausente é substituído por zero, pela média da variável até a amostra com o problema ou por um valor estimado, onde a variável problemática tem seu valor estimado por algum modelo que, em alguns casos, pode levar em consideração outras variáveis e/ou valores anteriores. Neste contexto este artigo apresenta uma revisão da literatura abordando as principais metodologias utilizadas para tratar o problema de dados ausentes. Além disso, estudos de casos são apresentados para comparação de resultados e definição de situações onde cada tipo de abordagem por ser melhor empregada.

Palavras-Chave: Dados Ausentes, Ausência Completamente Aleatória, Ausência Aleatória.

1 Introdução

Uma grande dificuldade enfrentada no desenvolvimento de aplicações reais que utilizam fluxos de dados para resolver problemas de previsão e/ou classificação são os dados ausentes. Sensores não confiáveis, observações incompletas, oclusão parcial do sinal desejado, interferência na comunicação, restrição da banda de transmissão de dados, informações incompletas obtidas de especialistas ou de pesquisas públicas, são alguns dos motivos que podem levar ao problema em questão [1, 2]. Apesar de existirem técnicas para reduzir os impactos ocasionados por este problema, a maioria dos sistemas não são modelados de forma preventiva para possibilitar o tratamento adequado deste tipo de ocorrência. Na falta de arcabouços teóricos ou recursos, os responsáveis pela modelagem e desenvolvimento dos sistemas recorrem a edição dos dados ausentes para manter uma aparência de integridade. Contudo, tanto estas edições quanto a falta de uma abordagem adequada podem levar os sistemas a produzir respostas tendenciosas, ineficientes e pouco confiáveis, prejudicando qualquer conclusão feita a partir da análise dos dados [3, 4].

Os dados ausentes são parte de um conceito geral e mais amplo que também está relacionado a falta de precisão dos dados incluindo também números arredondados, agrupados, agregados, truncados ou censurados, por exemplo [3]. Além da perda de eficiência e da dificuldade da análise, o viés resultante das diferenças entre dados ausentes e completos são algumas consequências geralmente associadas a este problema e que complicam a análise dos dados [2]. Outro problema é a possibilidade de perda de poder estatístico, considerando que os dados ausentes podem ocasionar tendências na estimativa dos parâmetros, que pode haver redução da representatividade das amostras e que pode complicar a análise dos estudos, sendo importante ressaltar que cada um destes possíveis problemas pode ameaçar a validade dos testes e resultar em conclusões inválidas [4]. Segundo Schneider [5], a ausência dos dados é uma circunstância que complica a análise multivariada de dados climáticos, mas se trata de algo comum uma vez que a disponibilidade das medições climáticas varia espacial e temporalmente, fazendo com que os dados climáticos nem sempre estejam disponíveis. Este problema com as ausências também é comum em tarefas de reconhecimento de padrões, fazendo com que os dados a serem avaliados não estejam disponíveis, dificultando sua análise [4].

Nas comunidades de pesquisa, é possível encontrar definições e interpretações ligeiramente diferentes sobre os diferentes tipos de dados ausentes. Contudo, no geral, os pesquisadores trabalham basicamente com três definições, classificadas de acordo com a forma do desaparecimento, sendo elas: Ausência Completamente Aleatória (*Missing Completely at Random* – MCAR), Ausência Aleatória (*Missing at Random* – MAR) e Ausência Não Aleatória (*Not Missing at Random* – NMAR ou *Missing Not Random* - MNAR) [3, 4, 6–9].

O mecanismo MCAR pode ser definido quanto a probabilidade de uma variável estar ausente não possui qualquer relação com outras variáveis ou amostras da base de dados. Em outras palavras, a probabilidade de um valor estar ausente é completamente aleatória. Dessa forma, a probabilidade da ausência de um valor é a mesma que a probabilidade de ausência dos demais valores. Este tipo de ausência é comum quando um equipamento não funciona direito ou fica temporariamente indisponível, por exemplo [3, 4, 6, 7]. Apesar do problema que a ausência causa, este tipo de ausência possui uma vantagem estatística que faz com que a análise permaneça imparcial, uma vez que não há relação entre os valores ausentes e outras variáveis ou amostras na base de dados. Assim, mesmo que ainda haja perda de poder estatístico ou analítico, os parâmetros estimados não são influenciados pela ausência [4]. Uma constatação importante é que, se os dados estão ausente completamente ao acaso, ou seja, se trata-se de um mecanismo MCAR, excluir amostras com valores ausentes não deve alterar suas inferências [10].

Para o mecanismo MAR, por outro lado, a probabilidade de ausência de uma variável continua aleatória e independente dos valores, contudo essa ausência possui relação com outras variáveis da amostra [7–9]. A probabilidade de uma variável estar ausente depende apenas das informações disponíveis [10]. Para Little [6], os valores dos dados perdidos podem ser considerados como um efeito aleatório que pode ser previsto por outras variáveis no conjunto de dados. Para Kang [4], a probabilidade de que uma variável esteja ausente depende do conjunto de variáveis observadas, mas não está relacionada aos

valores específicos ausentes que se espera que sejam obtidos. Uma das consequências disso é fazer com que, em uma amostra, a probabilidade de uma variável estar ausente seja maior que a probabilidade das demais variáveis estarem ausentes [3]. É importante ressaltar que a ausência é condicional a outras variáveis, mas a ocorrência deste tipo de ausência é comum quando existem variáveis que dependem de processamentos ou seleções anteriores para definirem seus valores, como ocorre em um questionário onde algumas questões dependem de respostas anteriores, por exemplo [8].

Por fim, para o mecanismo NMAR ou MNAR, onde as características da ausência não se enquadram nos anteriores, a probabilidade de uma variável estar ausente depende dos potenciais valores desta variável, ou seja, a ausência está diretamente relacionada ao próprio valor ausente [3, 4, 7–9]. Little [6] também ressalta que, nesta situação, os dados estão ausentes por algum motivo sistemático que indisponível ao pesquisador. Este tipo de ausência pode ocorrer quando, em um questionário, uma pergunta é desconsiderada por não se aplicar ou por falta de conhecimento da pessoa que está respondendo, por exemplo [8]. Neste caso, a única maneira de se obter uma estimativa imparcial dos parâmetros é estimar os valores ausentes [4]. Normalmente, a obtenção de estimativas precisas para este tipo de mecanismo tende a ser mais complexos, pois não se tem as informações necessárias para especificar corretamente modelos das ausências [8].

De uma forma prática, estes mecanismos são importantes por ajudarem a descrever as condições necessárias para se obter parâmetros consistentes. A partir deles foi possível verificar que excluir casos incompletos de uma análise requer um mecanismo MCAR, além de gerar estimativas tendenciosas em mecanismos MAR e NMAR. Também que a estimativa de máxima verossimilhança com todos os dados observados disponíveis produz estimativas consistentes para mecanismos MCAR ou MAR, o mesmo podendo ser aplicado a imputação múltipla. Finalmente, que os mecanismos NMAR exigem que o especifique corretamente um modelo explícito para a distribuição das probabilidades de ausência [8].

O presente artigo tem por objetivo trazer uma revisão de literatura relacionada aos dados ausentes e comparar os resultados obtidos com duas das técnicas, mais especificamente a substituição por zero e o algoritmo RegEM, em bases MCAR e MAR geradas. Para isso, na seção 2 são abordadas algumas das principais técnicas para tratamento dos dados ausentes, na seção 3 são abordados os materiais e métodos utilizados nos experimentos, na seção 4 são abordados os experimentos e os resultados obtidos e, por fim, na seção 5 é feita a conclusão do estudo e dos experimentos.

2 Técnicas de Manipulação dos Dados Ausentes

Há várias maneiras de tratar dados ausentes e, de uma forma intuitiva, é possível citar as seguintes maneiras: exclusão do registro, onde toda a amostra ou parte dela é descartada ou omitida, ou a inserção do valor. A exclusão é recomendada somente em casos onde a quantidade de amostras com dados ausentes é pequena, não haja uma relação entre as variáveis e a deleção não cause um grande impacto no sistema. É importante ressaltar que a remoção de amostras com dados ausentes pode ocasionar perda substancial de informações. Além disso, excluir variáveis da análise devido a ausência de alguns dados significaria utilizar as informações de forma ineficiente, ressaltando a importância de se utilizar métodos confiáveis para se estimar os valores que estão faltando [5]. Assim como aplicações que dependem da base completa, aplicações que trabalham com *stream* de dados, dados online ou que realizam processamento em tempo real para realizar alguma previsão, por exemplo, dependem destes dados para realizar seus procedimentos [2]. Contudo, nestes casos, a remoção das amostras não pode ser considerada. Por outro lado, a inserção pode ser feita de formas simples, substituindo o valor ausente por zero ou por uma média, por exemplo, ou estimando os valores ausentes utilizando modelos que possam identificar relações entre as variáveis para estimar novos valores, sendo este um exemplo de uma das formas mais complexas [2, 11].

Para Kang [4] o melhor método possível é ter um bom planejamento do estudo e coletar cuidadosamente os dados para evitar o problema. Contudo, o autor também considera não ser incomum existir uma quantidade considerável de dados ausentes em um estudo. É exatamente a partir dessa existência que se torna necessário o estudo dos métodos e técnicas descritos nesta seção.

2.1 Técnicas de Deleção

Uma das técnicas mais comuns para este problema é a deleção ou omissão de amostras com valores ausentes, fazendo com que se trabalhe apenas com os dados disponíveis, sendo este um método padrão em muitos pacotes de regressão [4, 10, 12]. Muitos pacotes de softwares estatísticos a utilizam como solução padrão, apesar de alguns pesquisadores dizerem que esta técnica pode produzir resultados tendencioso [4, 10]. Apesar disso, sabe-se que também é possível obter resultados conservadores, caso a base tratada seja de um cenário MCAR, conforme visto anteriormente [4]. Por isso, esta técnica é aconselhável somente quando a deleção das amostras ausentes não representar um problema considerável pela quantidade de registros e o mecanismo a ser tratado for comprovadamente MCAR [4].

A deleção em pares, também observada como análise de dados disponíveis ou análise de variáveis completas, apesar de também envolver deleção ou omissão de valores, se trata de uma técnica menos radical quando comparada a técnica anterior. Neste caso, a amostra não é totalmente deletada e somente a variável com valor ausente é descartada do processamento, fazendo com que todas as demais possam ser devidamente computadas e com que esta técnica seja menos tendenciosa em cenários MCAR ou MAR [4, 10].

2.2 Técnicas de Imputação

Uma variedade de técnicas, das mais simples as mais complexas, trabalham com imputação de dados buscando substituir as variáveis com valores ausentes por valores estimados para que se possa utilizar todas as informações disponíveis. Dessa forma, todos os casos são preservados e, após o preenchimento de todas as variáveis ausentes, a base de dados pode ser analisada normalmente como uma base de dados completa, preservando o tamanho original da base [4, 10]. Segundo Gelman [10], existe uma tendência de se obter erros padrão mais baixos quando se utiliza apenas uma estratégia de imputação.

A substituição pela média, uma das formas simples de imputação, faz com que seja utilizado o valor médio de uma variável no lugar dos dados ausentes desta variável [4, 10]. Para Kang [4], embora teoricamente a média seja uma estimativa razoável para preencher o valor de uma variável ausente por motivos aleatórios, na presença de uma grande quantidade de dados ausentes ou na falta de aleatoriedade destas ausências, a utilização da média pode acarretar em resultados inconsistentes. O autor também ressalta que esta técnica não adiciona novas informações, apenas aumenta o tamanho da amostra e leva a uma subestimação dos erros. Para Gelman [10], por outro lado, este método pode distorcer severamente a distribuição dos dados para esta variável, levando a complicações com medidas incluindo, notadamente, subestimações do desvio padrão e distorcendo relações entre as variáveis, tendendo para zero as estimativas da correlação.

Outra técnica de imputação substitui o valor de uma variável ausente pelo último valor existente para a mesma variável [4, 10]. Apesar de se tratar de uma técnica simples, como esta técnica assume que os valores para uma variável não se alteraram para então poder substituir pelo último valor, a técnica produz uma estimativa tendenciosa, fazendo com que seja aconselhável não utilizá-la como abordagem primária [4].

Na imputação por regressão as variáveis existentes são utilizadas para que se possa realizar uma previsão que será utilizada para substituir o valor ausente, como se o valor previsto fosse o valor real da variável [4]. Primeiramente, é necessário ajustar um modelo de regressão definindo a variável de interesse como variável de resposta e outra variável relevante como covariáveis. Os coeficientes são estimados e, em seguida, os valores omissos podem ser previstos pelo modelo ajustado [12]. Kang [4] ressalta que, embora esta técnica possua algumas vantagens sobre as técnicas de deleção descritos anteriormente por evitar alterar a distribuição dos dados e o desvio padrão, assim como a técnica de substituição pela média, nenhuma informação nova é adicionada. Por outro lado, Zhang [12] ressalta que a imputação com regressão em outra ou mais variáveis pode produzir valores mais inteligentes.

Há também técnicas que utilizam a máxima verossimilhança para tratar os dados ausentes. A estimativa de máxima verossimilhança identifica os valores dos parâmetros da população com a maior probabilidade de produzir os dados da amostra [8]. Após a estimativa dos parâmetros utilizando os

dados disponíveis, os dados ausentes são estimados com base nos parâmetros que acabaram de ser estimados. De uma forma geral, esta técnica estima os valores das variáveis ausentes usando a distribuição condicional das outras variáveis [4]. O algoritmo EM (*Expectation Maximization*), por exemplo, foi desenvolvido por [13] e pode ser utilizado para criar um novo conjunto de dados, no qual todos os valores omissos são imputados com valores estimados pelos métodos de máxima verossimilhança. Inicialmente, o algoritmo faz a estimativa dos parâmetros. Essas estimativas são utilizadas para criar uma equação de regressão capaz de prever os dados ausentes. A etapa seguinte de maximização utiliza essas equações para imputar os dados ausentes. O algoritmo então volta ao passo inicial para estimar novamente os parâmetros e gerar novas equações, repetindo as etapas até que o sistema se estabilize, quando a matriz de covariância para a iteração subsequente é virtualmente a mesma que a da iteração anterior [4, 5].

Outra técnica interessante para tratar dados ausentes é a imputação múltipla, considerada uma alternativa á estimativa de máxima verossimilhança e que tem sido amplamente utilizada [4, 8]. Nesta técnica, em vez de substituir um único valor por outro, os valores ausentes são substituídos por um conjunto de valores plausíveis que contêm a variabilidade natural e a incerteza dos valores corretos [4]. Esta técnica começa com uma previsão dos dados ausentes utilizando os dados existentes das demais variáveis. Os valores ausentes são, então, substituídos pelos valores previstos e um conjunto de dados completo é criado e denominado conjunto de dados imputado. Como este processo é repetido dentro das iterações, são criados vários conjuntos de dados imputados, justificando assim o nome da técnica. Cada conjunto de dados múltiplo imputado produzido é então analisado utilizando os procedimentos de análise estatística padrão para dados completos, fornecendo vários resultados analíticos. Posteriormente, combinando esses resultados analíticos, é produzido um único resultado geral analítico. Além de restaurar a variabilidade natural dos valores ausentes, esta técnica incorpora a incerteza devido aos dados ausentes, o que resulta em uma inferência estatística válida [4]. Diferentemente da técnica de máxima verossimilhança, a imputação múltipla separa o tratamento de dados ausentes da análise estatística e, como a imputação normalmente emprega um modelo muito geral, um único conjunto de imputações geralmente pode suportar uma variedade de análises estatísticas [8].

2.3 RegEM

O RegEM[1] foi proposto por Schneider [5], teve como ponto de partida o algoritmo EM (*Expectation Maximization*), citado anteriormente neste artigo, e foi desenvolvido objetivando ser aplicado a bases de dados climáticas incompletas, nos quais o número de variáveis normalmente excede o número de registros.

A partir dos dados incompletos, o algoritmo EM calcula as estimativas de máxima verossimilhança dos parâmetros de qualquer distribuição probabilística. Para dados Gaussianos, cuja distribuição probabilística pode ser parametrizada pela média e pela matriz de covariância, o EM inicia com estas duas e depois percorre as etapas alternadas de atribuição de valores ausentes e reestimativa da média e da matriz de covariância a partir conjunto de dados completo e de uma estimativa da matriz de covariância do erro de inserção do dado ausente. Na etapa de inserção, os valores ausentes são inseridos por meio da expectativa condicional obtida através dos valores disponíveis e a matriz de covariância do erro dos valores inseridos é estimada. Na etapa de estimação, a média e a matriz de covariância são reestimadas, considerando o erro de inserção condicional para a matriz de covariância. As etapas de inserção e estimativa são repetidas até que os valores inseridos, a média estimada e a matriz de covariância parem de se alterar [5].

Segundo Schneider [5], para dados Gaussianos, o algoritmo EM é baseado em análises de regressão linear iteradas e precisou ser regularizado para resolver o problema abordado pelo autor envolvendo conjuntos de dados climáticos. Para o algoritmo RegEM, versão regularizada do EM, os parâmetros de regressão regularizados são calculados com um método conhecido como Ridge Regression, por estatísticos, ou como Tikhonov Regularization, por matemáticos. Assim, o parâmetro de regularização que controla a filtragem é determinado pela validação cruzada generalizada, minimizando o erro médio

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

quadrático esperado para os valores inseridos [5]. No geral, o modelo RegEM consiste nos mesmos passos do modelo EM com uma exceção na estimativa dada por

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

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

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

onde $\hat{D} = Diag(\hat{\sum}_{aa})$ é a matriz diagonal constituída pelos elementos diagonais da matriz de covariância $\hat{\sum}_{aa}$ e o escalar h é um parâmetro de regularização [5].

3 Materiais e Métodos

Esta seção detalha os conjuntos de dados e os modelos utilizados nos experimentos. Os métodos de avaliação dos resultados também são descritos nessa seção.

3.1 Bases de Dados Originais

Inicialmente, foram selecionadas duas bases de dados completas associadas a problemas de regressão (previsão). As bases de dados foram obtidas no Repositório da UCI [14] e são descritas a seguir:

- **DataSet 1**[2]: base de dados relacionada a testes aerodinâmicos e acústicos realizados pela NASA (*National Aeronautics and Space Administration*) compreendendo aerofólios de tamanhos diferentes NACA 0012 em várias velocidades de túnel de vento e ângulos de ataque. A base possui 1503 registros com 5 variáveis de entrada, sendo elas: frequência, ângulo, comprimento da corda, velocidade do fluxo livre, espessura do deslocamento lateral da sucção. O objetivo é prever o nível de pressão sonora em escala. Este conjunto de dados foi utilizado nos trabalhos de [15–17].
- **DataSet 2**[3]: base de dados contendo a contagem horária e diária de bicicletas de aluguel entre os anos de 2011 e 2012 no sistema Capital Bikeshare. A base possui 731 registros com 13 variáveis de entrada, são elas: estação do ano, ano, mês, se o dia era feriado ou não, dia da semana, se o dia era um dia de trabalho (exceto finais de semana e feriados), condição do tempo, temperatura, sensação térmica, humidade, velocidade do vento, contador de usuários casuais, contador de usuários registrados. O objetivo é realizar a previsão do número de bicicletas alugadas. Este conjunto de dados foi utilizado no trabalho de Fanaee [18].

Destaca-se que para os experimentos os dados foram normalizados com valores entre 0 e 1.

3.2 Bases de Dados com Dados Ausentes

A partir das bases de dados completas descritas na Seção 3.1 foram criadas bases com dados ausentes para os cenários de MCAR e MAR.

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

[3]http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset

Bases de Dados MCAR

Para criação da base MCAR de dados ausentes foi utilizado o Algoritmo 1. Este algoritmo recebe como entrada a base de dados completa e a taxa de ausência que a nova base deve ter. Com isso ele calcula o número de amostras que devem possuir dados ausentes a partir do número de amostras da base original e da taxa de ausência estipulada. Então, o algoritmo sorteia as amostras que possuirão dados ausentes. Por fim, ele faz uma iteração em cada amostra sorteada para conter um dado ausente e sorteia a variável que ficará com o valor ausente.

Para cada base de dados descrita na Seção 3.1 foram criadas 6 novas bases de dados com dados ausentes com taxas de ausência de 1%, 5%, 10%, 15%, 20% e 30%. A partir de cada base de dados com dados ausentes foram criadas duas novas bases, uma substituindo os valores ausentes por zero e a outra utilizando o RegEM para imputar os dados ausentes. Portanto, foram criadas 12 bases de dados para os experimentos com MCAR. Os dados das bases de dados foram normalizados para valores entre 0 e 1.

Data: baseOriginal, taxaAusencia
Result: uma base de dados MCAR baseada na base original
Obtém a quantidade de amostras e variáveis de entrada da base original
Calcula a quantidade de amostras a receber dado ausente
Sorteia aleatoriamente as amostras que receberão dado ausente
Para *cada amostra sorteada*
  Sorteia uma variável aleatoriamente
  Faz a base receber o dado ausente na amostra iterada e variável sorteada
Fim

Algorithm 1: Gerador de bases MCAR

Bases de Dados MAR

O Algoritmo 2 foi utilizado para gerar as bases de dados ausentes MAR. Este algoritmo recebe a base de dados original, a taxa de ausência para a variável com maior propensão e a taxa de ausência para as demais variáveis. O algoritmo seleciona uma variável da base para ser a variável com maior propensão a possuir valor ausente e calcula a probabilidade das demais variáveis estarem ausentes com base no número de variáveis e na taxa mínima de ausência estabelecida. Depois, são iniciados dois vetores para controle dos sorteios, um zerado e um com o valor de cada índice. Com isso, o algoritmo faz uma iteração em cada variável para que cada uma ocupe uma quantidade de posições no vetor de sorteio relativa à sua porcentagem probabilística de ausência. Por fim, ao iterar em cada amostra da base de dados original, o algoritmo seleciona aleatoriamente um índice do vetor de sorteio e, caso seja sorteado o índice de uma variável, a variável fica com o valor ausente.

O Algoritmo 2 foi utilizado com taxas de ausência de 5x1, 10x1, 10x5, 20x5, 20x10, 30x5 e 30x10 para criação de 7 bases de dados com dados ausentes MAR, onde o primeiro dígito se refere a taxa de ausência máxima e o segundo se refere a taxa de ausência mínima. Após este procedimento, a partir das 2 bases originais, foram criadas 14 bases MAR das quais 7 foram geradas por substituição por zero e as outras 7 utilizando o RegEM.

3.3 Algoritmos de Validação

Para avaliar as técnicas de imputação de dados descritas foram utilizados três algoritmos fuzzy evolutivos: eTS [19], xTS [20]) e eMG [21]. A seguir cada um desses modelos é descrito.

eTS

O eTS (*Evolving Takagi-Sugeno*) foi proposto em [19]. Trata-se de um modelo funcional do tipo Takagi-Sugeno que atualiza continua e recursivamente a estrutura da base de regras, inserindo novas

Data: baseOriginal, taxaAusenciaMax, taxaAusenciaMin
Result: uma base de dados MCAR baseada na base original
Obtém a quantidade de amostras e variáveis da base original
Cria uma lista de sorteio de tamanho 100 e todos os valores iguais a zero
Sorteia uma variável para ter taxa de ausência máxima
Calcula a taxa de ausência das demais variáveis
Para *cada variável da base original*
  if *a variável for a sorteada para taxa máxima* **then**
    Sorteia posições da lista de sorteio usando a taxaAusenciaMax para definir a quantidade
  else
    Sorteia posições da lista de sorteio usando a taxa de ausência calculada para definir a quantidade
  end
  Altera os valores das posições selecionadas da lista de sorteio com o valor da variável
Fim
Para *cada amostra da base original*
  Sorteia aleatoriamente uma posição da lista de sorteio
  if *a posição sorteada não for igual a zero* **then**
    Obtém a variável pelo valor sorteado da lista de sorteio
    Altera a base original na amostra iterada e variável obtida com o valor ausente
  end
Fim

Algorithm 2: Gerador de bases MAR

regras e alterando as existentes bem como seus parâmetros. O tipo das regras do eTS podem ser verificada na Equação (3) e sua saída é determinada pela média ponderada normalizada pela ativação das funções de pertinência.

$$\Re_i : SE\, x_1 \,\acute{e}\, A_{i1} \ldots e\, x_j \,\acute{e}\, A_{ij} \ldots e\, x_n \,\acute{e}\, A_{in}\, ENT\tilde{A}O\, y_i = q_{i0} + q_{i1}x_1 + \ldots + q_{ij}x_j + \ldots + q_{in}x_n \quad (3)$$

A cada nova amostra, este modelo utiliza um algoritmo de agrupamento recursivo não supervisionado para atualizar sua base de regras. A estrutura dos grupos é atualizada a cada iteração e cada grupo define o antecedente de uma regra. Para cada amostra processada existe a possibilidade da criação de um novo grupo ou da atualização dos valores de algum grupo existente [19]. Além de utilizar funções Gaussianas, o aprendizado do antecedente das regras do eTS é realizado empregando uma versão incremental do algoritmo eClustering onde a capacidade de um ponto ser escolhido como centro é definida pela função potencial representada por

$$P_i = \frac{1}{N}\sum_{j=1}^{N} e^{-\frac{4}{r^2}\|s_i - s_j\|^2}, \quad (4)$$

onde P_i é o potencial da *i*-ésima amostra, N é o número de amostras e r é o raio de influência do grupo [19]. Os parâmetros do consequente das regras são atualizados por mínimos quadrados recursivos.

xTS

xTS (*eXtended Takagi-Sugeno*) proposto por Angelov & Zhou [20] é uma versão extendida do eTS. Diferente do eTS, para este modelo o raio de abrangência de cada grupo é estimado recursivamente. Além disso, foi adicionado ao xTS um índice de idade utilizado para medir a qualidade dos grupos. Outro parâmetro utilizado para medir a qualidade dos grupos é a população, porém, neste caso, a população de um grupo é denominada como seu suporte e obtida por

$$S_l \leftarrow S_l + 1; l = \underset{l}{\overset{R}{\operatorname{argmin}}} \|S_t - S_l^*\|^2, l = 1...R, \tag{5}$$

onde S_l é o suporte do *l*-ésimo *cluster*. Este suporte descreve a quantidade de amostras que estão na zona de influência, indicando também o poder de generalização da regra. O suporte, ou o suporte divido pelo número de amostras, também é utilizado para avaliar a qualidade de uma regra, indicando se a regra pode ser ignorada (excluída) [20].

Assim como o eTS os parâmetros do consequente são atualizados por mínimos quadrados recursivos. Contudo, enquanto no eTS os raios de abrangência dos grupos são esféricos, no xTS eles são elipsoidais com forma adaptada recursivamente pelas informações espaciais obtidas pelo processamento de novas amostras [20].

eMG

O eMG (*Multivariable Gaussian Evolving Fuzzy*), um sistema *fuzzy* evolutivo com uma nova abordagem do aprendizado participativo foi proposto em Lemos, Caminhas & Gomide [21]. O modelo utiliza funções de pertinência Gaussianas multivariável na representação dos grupos e sua estrutura de agrupamento é atualizada recursivamente a cada passo do algoritmo, com seus limiares sendo definidos automaticamente.

Utilizando um algoritmo de agrupamento participativo evolutivo Gaussiano, o modelo obtém o antecedente das regras, enquanto os parâmetros do consequente são atualizados por um algoritmo recursivo de mínimos quadrados ponderados. O modelo trabalha tanto com a ideia de inclusão de novas regras, quanto com a modificação de parâmetros de regras existentes e união de grupos redundantes. Utilizando uma medida de compatibilidade e um índice de alerta calculados, a estrutura de grupo do modelo é atualizada a cada amostra processada [21]. A medida de compatibilidade, representada por

$$p_i^k = exp[-\frac{1}{2}M(x^k, c_i^k)], \tag{6}$$

utiliza o quadrado de uma medida de distância normalizada entre a amostra processada e os centros dos grupos, que produz grupos elipsoidais cujos eixos não são necessariamente paralelos aos eixos das variáveis de entrada. A distância normalizada (M) pode ser calculada por

$$M(x^k, c_i^k) = (x^k - c_i^k)(\sum_i^k)^{-1}(x^k - c_i^k)^T, \tag{7}$$

onde $\sum_i^k$ é a matriz de dispersão que representa a forma dos grupos, sendo estimada a cada iteração por

$$\sum_i^{k+1} = (1 - \alpha(p_i^k)^{1-\lambda_i^k})(\sum_i^k -\alpha(p_i^k)^{1-\lambda_i^k})(x^k - c_i^k)(x^k - c_i^k)^T). \tag{8}$$

As funções de pertinência que representam os grupos são caracterizadas por um vetor central e uma matriz de dispersão, representando a dispersão de cada variável de entrada. O modelo também utiliza um mecanismo de ajuste automático do valor de limiar de distância que se baseada na dimensão da entrada. Seu limiar de compatibilidade se dá por

$$T^p = exp[-\frac{1}{2}X^2_{m,\lambda}], \tag{9}$$

onde $X^2_{m,\lambda}$ é o λ o intervalo de confiança unilateral superior de uma distribuição *Qui-Quadrado* com *m* graus de liberdade, onde *m* é o número de entradas [21].

Para verificar a possibilidade de união de grupos semelhantes, a compatibilidade entre os grupos é frequentemente avaliada e, caso essa compatibilidade entre dois grupos seja maior que o limiar, isto é $p^k_i(c^k_j, c^k_i) > T_p$, ocorre a união dos grupos [21].

3.4 Medidas de Desempenho

Para avaliar o desempenho dos algoritmos e das metodologias imputação de dados foi utilizado como medida de erro o RMSE (*Root Mean Square Error*), descrito na Eq. 10, calculado utilizando o quadrado da diferença entre a saída obtida e a real.

$$RMSE = \frac{1}{H}\sum_{h=1}^{H}\sqrt{(\hat{y}^{[h]} - y^{[h]})^2} \tag{10}$$

4 Experimentos e Resultados

Esta seção detalha os experimentos realizados e apresenta os resultados obtidos. Inicialmente são apresentados os experimentos realizados empregando as 12 bases de dados MCAR seguido dos experimentos com as 14 bases MCAR. Os experimentos foram realizados com os algoritmos eTS, xTS e eMG. Para o eTS foram utilizados os valores: 750 para o Omega; modelos não fixos; objetivo de otimização global. Para o xTS foram utilizados os valores: 0,5 para o raio de influência; 750 para o Omega; modelos não fixos. Para eMG foram utilizados os valores: 0,01 para o α básico; 0,05 para o λ; 40 para o tamanho da janela; para a quantidade de pontos de treinamento foi utilizado a quantidade de registros de cada base; e para o σ inicial foi utilizada uma matriz quadrada de tamanho $n-1 \times n-1$, onde n é a quantidade de variáveis da base, com os valores 0.1 na diagonal principal e zero nas demais posições. As primeiras bases a serem executadas pelos algoritmos definidos na sessão anterior foram as originais para se obter os valores que serão a base para toda a análise dos demais resultados, seguida das bases MCAR e, por fim, as bases MAR.

4.1 Experimentos MCAR

A Tabela 1 ilustra os resultados obtidos para as bases MCAR geradas a partir da base de dados 1 com taxas de ausência de 1%, 5%, 10%, 15%, 20% e 30% e para a base original. Conforme pode ser verificado, para os cenários com 1% e 5% de ausência, a substituição por 0 conseguiu obter resultados melhores do que os da base original, contudo, o resultado foi piorando à medida que a taxa de ausência aumentava e, em comparação com a base original, sua proximidade máxima foi de 1 casa decimal. Por outro lado, as bases que utilizaram o RegEM como solução, no geral, obtiveram alguns resultados ligeiramente melhores que a base original em todos os algoritmos a partir dos cenários com taxa de ausência acima de 15% e mantiveram seus resultados com uma proximidade máxima de três casas decimais em alguns casos.

Analogamente, a Tabela 2 mostra os resultados obtidos para as bases MCAR geradas a partir da base de dados 2. Neste cenário, todos os resultados obtidos pela substituição por 0 foram piores do que os obtidos pela base original, mas, assim como ocorreu no cenário anterior, os resultados foram piorando

Table 1. Resultado para o cenário MCAR nas bases geradas da base original 1

Bases	**eTS**		**xTS**		**eMG**	
Original 1	0,12716		0,12764		0,12941	
Ausência	**Subst. 0**	**RegEM**	**Subst. 0**	**RegEM**	**Subst. 0**	**RegEM**
1%	0,05607	0,12805	0,05928	0,12457	0,05597	0,12807
5%	0,10679	0,13500	0,10463	0,12711	0,10979	0,13215
10%	0,15084	0,13092	0,15165	0,12170	0,13662	0,12820
15%	0,14772	0,12074	0,15091	0,12228	0,14565	0,12902
20%	0,19161	0,12180	0,18817	0,11836	0,16605	0,12869
30%	0,22208	0,11874	0,22440	0,12349	0,21250	0,12851

à medida que a taxa de ausência aumentava. Os resultados obtidos pela utilização do RegEM, assim como ocorreu para a base original 1, se mantiveram com uma proximidade de até três casas decimais, contudo, apesar de também ter obtido alguns resultados melhores do que a base original, neste cenário, os resultados melhores não ocorreram conforme a taxa de ausência aumentava, como verificado nos resultado anteriores. É importante ressaltar que para as bases com 10% e 15% de ausência que utilizaram como solução o RegEM, por motivo até então desconhecido, não obtiveram resultado no algoritmo xTS.

Table 2. Resultado para o cenário MCAR nas bases geradas da base original 2

Bases	**eTS**		**xTS**		**eMG**	
Original 2	0,02410		0,01478		0,04166	
Ausência	**Subst. 0**	**RegEM**	**Subst. 0**	**RegEM**	**Subst. 0**	**RegEM**
1%	0,03505	0,02432	0,03066	0,01483	0,05422	0,04186
5%	0,06717	0,02097	0,04871	0,01194	0,04632	0,02830
10%	0,08107	0,02525	0,10951	NaN	0,07062	0,02859
15%	0,12377	0,02488	0,15800	NaN	0,08665	0,04299
20%	0,12995	0,03454	0,09788	0,01389	0,10546	0,04662
30%	0,17204	0,02642	0,16160	0,01896	0,16396	0,02999

4.2 Experimentos MAR

A Tabela 3 ilustra os resultados obtidos para as bases MAR geradas a partir da base de dados 1 com taxas de ausência de 5x1, 10x1, 10x5, 20x5, 20x10, 30x5 e 30x10. Neste cenário, a substituição por 0 continuou obtendo resultados com proximidade de uma casa decimal dos resultados da base original, contudo é importante ressaltar que seu pior caso no cenário de ausência 30x5 foi muito superior a todos os outros obtidos até então, causando uma diferença máxima de 0.35288 no algoritmo eTS. Para as bases que utilizaram o RegEM, entretanto, os resultados se mantiveram próximos, com alguns resultados melhores que os da base original, inclusive, mas a proximidade máxima foi reduzida para duas casas decimais.

Table 3. Resultado para o cenário MAR nas bases geradas da base original 1

Bases	**eTS**		**xTS**		**eMG**	
Original 1	0,12716		0,12764		0,12941	
Ausência	**Subst. 0**	**RegEM**	**Subst. 0**	**RegEM**	**Subst. 0**	**RegEM**
5x1	0,13225	0,13659	0,12650	0,13074	0,15289	0,13276
10x1	0,13371	0,12567	0,12885	0,13818	0,13579	0,12096
10x5	0,11644	0,12609	0,10914	0,12197	0,09186	0,12088
20x5	0,10162	0,12909	0,09691	0,12217	0,10960	0,13026
20x10	0,15343	0,11284	0,16630	0,11979	0,14983	0,12082
30x5	0,48004	0,10725	0,45391	0,11104	0,44672	0,11057
30x10	0,14826	0,11604	0,13716	0,12120	0,12562	0,11971

Analogamente, a Tabela 4 mostra os resultados obtidos para as bases MAR geradas a partir da base de dados 2. Diferentemente dos resultados mostrados na Tabela 3, os resultados obtidos pela substituição por 0 neste cenário não houve valores extremamente altos em relação aos resultados da base original, obtendo alguns resultados melhores e um resultado equivalente ao da base original no cenário 5x1 no algoritmo xTS. Para as bases que utilizaram o RegEM, a maioria dos resultados obtidos foi melhor que os resultados da base original e a proximidade máxima observada foi de três casas decimais. É importante ressaltar que, assim como ocorreu para dois cenários utilizando o RegEM, conforme mostrado na Tabela 2, para as bases com ausência 10x5 e 30x10, que utilizaram como solução a substituição por 0, não se obtive resultado no algoritmo xTS, também por motivo até então desconhecido.

Table 4. Resultado para o cenário MAR nas bases geradas da base original 2

Bases	**eTS**		**xTS**		**eMG**	
Original 2	0,02410		0,01478		0,04166	
Ausência	**Subst. 0**	**RegEM**	**Subst. 0**	**RegEM**	**Subst. 0**	**RegEM**
5x1	0,02432	0,02375	0,01478	0,01442	0,04444	0,04426
10x1	0,02635	0,02359	0,01299	0,01295	0,03037	0,03538
10x5	0,02460	0,02261	NaN	0,01451	0,04887	0,04849
20x5	0,02529	0,02176	0,01539	0,01245	0,04595	0,04531
20x10	0,10783	0,02343	0,10007	0,01429	0,12747	0,04859
30x5	0,01837	0,01790	0,01047	0,01099	0,04863	0,04117
30x10	0,15033	0,02258	NaN	0,01285	0,16747	0,02837

5 Conclusão

De acordo com a revisão feita por este artigo é possível verificar a complexidade do tema abordado. Cada mecanismo de dados ausente merece uma atenção e um tratamento especial para que seu impacto nos resultados sejam os menores possíveis. Além disso, é importante lembrar que não há uma solução única que resolva totalmente o problema ou que seja aconselhável em todos os casos, o que ressalta a importância da escolha das técnicas para tratar cada mecanismos da melhor forma o problema, seja por meio de uma ou mais técnicas em conjunto.

Para as duas técnicas analisadas, ocorreram resultados piores, muito próximos e melhores em relação aos resultados obtidos com as bases originais. No geral, as bases que utilizaram o RegEM obtiveram uma maior proximidade dos resultados das bases originais em mais testes, demonstrando uma estabilidade interessante. Por outro lado, as bases que utilizaram a substituição por 0, apesar de obter um resultado equivalente ao resultado da base original, conforme mostrado na Tabela 4 no cenário de ausência 5x1, no geral, obteve resultados dispersos, crescente a medida que a ausência aumentava nos cenários MCAR e sem padrão um padrão verificado para os cenários MAR. Isso mostra que, à medida que a quantidade de dados ausente aumenta, a técnica de substituição por 0 perde eficiência e tende a ser cada vez menos confiável.

Um dos motivos especulados para a obtenção de resultados melhores das bases que utilizaram as técnicas em comparação com as bases originais é a proximidade dos valores reais das variáveis com zero, contudo essa hipótese não foi confirmada até o fechamento deste artigo. Outro ponto interessante a ser verificado é o motivo de 4 bases não obterem um resultado utilizando o algoritmo xTS, contudo, esta é uma análise mais relevante para o algoritmo, saindo completamente do escopo deste trabalho.

Como trabalhos futuros estão tanto o estudo de outras possíveis formas de geração de bases com dados ausentes, quanto a análise de outras técnicas e algoritmos que solucionem este problema, com o objetivo de se gerar um modelo próprio.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Os autores também são gratos ao CEFET-MG pelo apoio.

References

[1] Cooke, M., Morris, A., & Green, P., 1997. Missing data techniques for robust speech recognition. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pp. 863–866. IEEE.

[2] Farhangfar, A., Kurgan, L. A., & Pedrycz, W., 2007. A novel framework for imputation of missing values in databases. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 37, n. 5, pp. 692–709.

[3] Schafer, J. L. & Graham, J. W., 2002. Missing data: our view of the state of the art. *Psychological methods*, vol. 7, n. 2, pp. 147.

[4] Kang, H., 2013. The prevention and handling of the missing data. *Korean journal of anesthesiology*, vol. 64, n. 5, pp. 402.

[5] Schneider, T., 2001. Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of climate*, vol. 14, n. 5, pp. 853–871.

[6] Little, T. D., Lang, K. M., Wu, W., & Rhemtulla, M., 2016. Missing data. *Developmental psychopathology*, pp. 1–37.

[7] Pedersen, A. B., Mikkelsen, E. M., Cronin-Fenton, D., Kristensen, N. R., Pham, T. M., Pedersen, L., & Petersen, I., 2017. Missing data and multiple imputation in clinical epidemiological research. *Clinical epidemiology*, vol. 9, pp. 157.

[8] Enders, C. K. & Baraldi, A. N., 2018. Missing data handling methods. *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development*, pp. 139–185.

[9] Krause, R. W., Huisman, M., Steglich, C., & Sniiders, T. A., 2018. Missing network data a comparison of different imputation methods. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 159–163. IEEE.

[10] Gelman, A. & Hill, J., 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

[11] Amiri, M. & Jensen, R., 2016. Missing data imputation using fuzzy-rough methods. *Neurocomputing*, vol. 205, pp. 152–164.

[12] Zhang, Z., 2016. Missing data imputation: focusing on single imputation. *Annals of translational medicine*, vol. 4, n. 1.

[13] Dempster, A. P., Laird, N. M., & Rubin, D. B., 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, n. 1, pp. 1–22.

[14] Dua, D. & Graff, C., 2017. UCI machine learning repository.

[15] Brooks, T. F., Pope, D. S., & Marcolini, M. A., 1989. Airfoil self-noise and prediction.

[16] Lau, K., López, R., Oñate, E., Ortega, E., Flores, R., Mier-Torrecilla, M., Idelsohn, S., Sacco, C., & González, E., 2006. A neural networks approach for aerofoil noise prediction. *Master thesis*.

[17] Lopez, R., Balsa-Canto, E., & Oñate, E., 2008. Neural networks for variational problems in engineering. *International Journal for Numerical Methods in Engineering*, vol. 75, n. 11, pp. 1341–1360.

[18] Fanaee-T, H. & Gama, J., 2013. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, pp. 1–15.

[19] Angelov, P. P. & Filev, D. P., 2004. An approach to online identification of takagi-sugeno fuzzy models. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 34, n. 1, pp. 484–498.

[20] Angelov, P. & Zhou, X., 2006. Evolving fuzzy systems from data streams in real-time. In *2006 International symposium on evolving fuzzy systems*, pp. 29–35. IEEE.

[21] Lemos, A., Caminhas, W., & Gomide, F., 2010. Multivariable gaussian evolving fuzzy modeling system. *IEEE Transactions on Fuzzy Systems*, vol. 19, n. 1, pp. 91–104.