Análise Comparativa de Técnicas de Aprendizado para Imputação de Dados Faltantes em Séries Temporais de Dados Meteorológicos

Anderson Silvério Mendrot Filho
Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo - UNIFESP
São José dos Campos, Brasil
anderson.mendrot@unifesp.br

Mateus Alves de Oliveira
Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo - UNIFESP
São José dos Campos, Brasil
mateusalves.info@gmail.com

***Resumo*—Séries temporais meteorológicas costumam apresentar valores faltantes, devido a problemas nos instrumentos de leitura. Dessa forma, a imputação de dados se mostra de grande utilidade. O presente projeto visa o uso dos algoritmos de aprendizado de máquina de regressão linear e *multilayer* perceptron para imputação de dados de temperatura na base de dados coletados na estação de Mirante de Santana de 1981 a 2010. Foram obtidos resultados com valor de erro próximos a 2 nos melhores casos, e concluiu-se que, apesar dos resultados não terem sido os mais adequados, há uma margem considerável para melhora por estudos mais aprofundados.**

***Index Terms*—Imputação, temperatura, meteorologia, séries temporais**

I. Introdução e motivação

Uma série temporal é um conjunto de observações em que cada uma delas ocorre em um tempo específico [1]. Sua principal característica é a dependência entre as observações adjacentes, sendo de grande relevância o estudo de tais dependências [2]. Dentre as séries temporais, a climatologia se mostra de grande importância. Seus estudos, que englobam detecção de mudança climática e diversas análises estatísticas, necessitam da obtenção de dados sem perda de informações realizadas em diferentes estações por um longo período de tempo, de tal forma que seja possível a representação das características meteorológicas da região desejada [3].

Dessa forma, faz-se necessária a disponibilidade de uma série completa e homogênea [4], a partir da qual seja possível realizar inferências estatísticas a partir de observações com sequencialidade preservada. Porém, por ser um processo contínuo, a perda de informações em séries temporais deste tipo é inevitável pela impossibilidade de fazer medições completas [3]. Tal situação é recorrente, e pode ser resultado de amostragem insuficiente, erros nas medições ou falha na aquisição de dados, tornando necessário o uso de técnicas de imputação de dados para o combate ao problema [6].

O Brasil disponibiliza banco de dados meteorológicos para pesquisa. Um deles é o do INMET, denominado BDMEP [7], o qual abriga dados meteorológicos diários em forma digital de séries históricas das várias estações meteorológicas convencionais da rede de estações do INMET.

O presente trabalho propõe um estudo comparativo entre diferentes técnicas de aprendizado, analisando o desempenho destas no preenchimento de dados faltantes em séries temporais meteorológicas após diversas execuções de algoritmos de aprendizado de máquina. O estudo comtempla *multilayer perceptron* (MLP) e Regressão Linear.

II. Fundamentação Teórica

A. Multilayer Perceptron

Um Perceptron Multicamadas, ou *Multilayer Perceptron* (MLP) é uma técnica famosa em Inteligência Artificial. MLP é uma das possíveis implementações de redes neurais artificiais, técnica que se inspira em um cérebro humano, cujo objetivo é modelar uma rede para se ajustar a um conjunto de dados.

Para isso, são criados vértices, chamados neurônios. Cada neurônio possui um valor diferente, o qual é obtido através de uma função de ativação aplicada a seu conjunto de entrada. Após a aplicação da função de ativação, o valor obtido é transmitido para o próximo neurônio da rede. A rede mais simples, proposta por [39], é denominada *Perceptron*, e possui apenas uma camada de entrada de dados e uma de saída.

Um possível aperfeiçoamento da rede *Perceptron* é a adição de uma ou mais camadas intermediárias, chamadas camadas ocultas, o que torna possível a resolução de problemas que não são lineares. Essa derivação de técnica recebeu o nome de MLP, e está representada na Figura 7. O objetivo do uso de redes neurais é a busca pelos pesos W = [w1,...wn] das arestas que mais se adequem ao modelo desejado.

Tal algoritmo de aprendizado é composto por 4 passos:

- Inicialização: Atribuir valores aleatórios aos pesos e limites;
- Ativação: Calcular os valores dos neurônios das camadas ocultas e da camada de saída;
- Treinar Passos: Calcular os erros , a correção e atualizar os pesos das camadas ocultas e de saída;
- Iteração: Repetir o processo a partir do passo 2 até se chegar em um valor de erro aceitável de acordo com o critério.

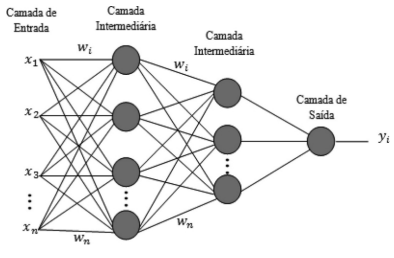

Figura 1. Multilayer Perceptron [40]

B. Regressão Linear

A Regressão Linear é um método de análise de regressão utilizado para investigar relações entre variáveis dependentes, denotadas por y, e independentes, denotadas por *x1*,*x2*,...,*xn*. Há diversos tipos de regressão linear, sendo uma delas a regressão linear simples, na qual há uma variável dependente e uma independente, além de coeficientes $\beta 0$, $\beta 1$ e uma taxa de erro ξ. A fórmula está representada na Figura 2 [38].

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Figura 2. Regressão linear simples [38]

Outro método é a regressão linear múltipla, que por sua vez utiliza mais de uma variável independente e uma variável dependente. Assim, assume-se que a variável dependente é uma função linear dos parâmetros do modelo. A forma da regressão linear está na Figura 3. Nela, há um vetor B = [$\beta 0$,$\beta 1$,...,βp] de coeficientes de regressão, X = [*x1*,*x2*,..,*xp*] de variáveis independentes, y como variável dependente. No aprendizado de máquina, a regressão linear é utilizada para previsão do valor da variável dependente em função dos valores das variável independentes [38].

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

Figura 3. Regressão linear [38]

C. Raiz do Erro Quadrático Médio (RMSE)

Medida de erro comumente usada para medir a qualidade do ajuste de um modelo a uma base, representa o desvio padrão amostral das diferenças entre os valores previstos *Pi* e os valores observados *Oi*, com n sendo o número de observações [37]. É um boa medida por mostrar, na maioria dos casos, explicitamente o valor que o algoritmo inteligente busca minimizar.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \left(P_i - O_i\right)^2}{n}}$$

Figura 4. Equação do RMSE

D. Validação Cruzada

Cross-Validation, ou validação cruzada é uma técnica estatística usada para avaliar modelos de *Machine Learning* dividindo a base de dados em subconjuntos, treinando o algoritmo em uma parte e avaliando no subconjunto complementar de dados. Uma das mais empregadas é a *k-folds*, na qual a base é dividida em k subconjuntos, sendo k-1 empregados no treino e 1 no teste. Em seguida, o subconjunto de teste é trocado k vezes, até todos os subconjuntos terem sidos usados para teste uma vez [31].

E. Teste de Wilcoxon Signed-Rank

Testes estatísticos são importantes para a análise de resultados de algoritmos inteligentes. O teste de *Wilcoxon Signed-Rank* utiliza o conceito de teste de hipóteses para verificação de correlação entre duas amostras pareadas da mesma população, de tal forma que há a hipótese nula *Ho* e a de rejeição *H1*: se *Ho* é verdadeiro, significa que as amostras possuem distribuição semelhante; caso contrário, a hipótese nula é rejeitada e a hipótese *H1*, de que as distribuições não são semelhantes,é aceita [36]. Para tanto, no projeto é utilizado um valor de confiança de 95% para as comparações, e é calculado o valor p de proximidade.

III. Trabalhos Relacionados

Há várias técnicas para lidar com o problema. A mais simples delas é a substituição de valores perdidos pela média dos dados válidos. Porém, seu uso pode dificultar a análise por conta das perturbações causadas na sua estrutura [5], pois pode reduzir a variância dos dados. Outro método comum é a interpolação linear, em que um dado faltante é preenchido pela média de seus valores vizinhos [4].

Outras técnicas, como a regressão linear, incluem as medidas de imputação única, em que são utilizados os valores de outras variáveis para auxílio na imputação de observações perdidas [7]. O principal problema da imputação única é a substituição do valor ausente por apenas um estimado, o que faz com que haja incerteza se aquele único valor representaria o valor real [3]. Por fim, há técnicas que consideram o erro gerado pelas estimativas, e entre elas se encontra a imputação múltipla, em que os processos de imputação são realizados diversas vezes, analisados por algum método estatístico e combinados para gerar o resultado final [7].

Dentro do escopo apresentado, métodos como redes neurais artificiais, *k-Nearest Neighbors* e *Support Vector Machine* se mostram algumas alternativas para o tratamento do problema. As redes neurais artificiais (RNA), por exemplo, são utilizadas no trabalho de [6]. Tal trabalho realiza um estudo do uso de diferentes algoritmos para preenchimento de dados faltantes de qualidade do ar, com a rede neural do tipo *multilayer perceptron* (MLP) obtendo desempenho semelhante aos mapas auto-organizáveis, e superior aos métodos kNN e baseado em regressão. O MLP também é utilizado em [8], pois foi considerado altamente flexível e útil para sistemas de imputação, tendo tido melhor desempenho comparado a imputação por média e baseado em regressão. Por fim, o algoritmo é utilizado

no trabalho de [3] para comparação entre diversos algoritmos na imputação de dados meteorológicos da Turquia.

O SVM, por sua vez, é outra possibilidade de algoritmo de aprendizagem de máquina, e é utilizado em [10] para comparação de seu desempenho com as RNAs e algoritmos tradicionais, obtendo desempenho semelhante às RNAs e melhor que os tradicionais.

Já o kNN é utilizado como base para formulação do algoritmo de [11], que descreve como vantagens do kNN a sua capacidade de servir tanto para modelos quantitativos quanto qualitativos, podendo ser útil para cálculo de média e contagem de valor mais frequente, respectivamente. O trabalho de [14] também criou um método para imputação usando como base o kNN por considerá-lo simples, porém eficiente.

A regressão linear aparece sendo bem utilizada em projetos. O projeto de [12], por exemplo, a utiliza para imputação em séries de dados de precipitação diária, assim como também faz uso de regressão linear, kNN e substituição pelo valor médio, por exemplo. Como resultado, foi obtido que a regressão linear múltipla supera os demais algoritmos testados. O trabalho de [20], por sua vez, utilizou fatores meteorológicos diversos, com ênfase a insolação. A regressão linear múltipla foi comparada com o método de interpolação inverso da potência das distâncias e com o algoritmo de imputação MICE (Multivariate Imputation by Chained Equations). Neste projeto, o método MICE se mostrou o mais vantajoso, seguido pela regressão múltipla linear. Já o trabalho de [25] também utilizou dados de precipitação para comparar o uso da imputação utilizando a regressão linear simples, o kNN e o inverso da potência das distâncias.

IV. Proposta

Para este projeto, foi utilizado de Regressão Linear e Perceptron Multicamadas para realizar a previsão dos valores faltantes. Para o primeiro experimento, foram utilizados *gaps* de 5, 10 e 15 entradas para o período de um ano, com taxas de aprendizado de 0,01 e 0,2 e 0,4 e com 2 e 3 camadas ocultas para o MLP. Após avaliar os resultados, optou-se por separar em estações (primavera, verão, outono e inverno) em *gaps* de 1 até 15 entradas, com as mesmas taxas de aprendizagem e camadas ocultas.

V. Metodologia Experimental

A. Base de dados

Para o estudo da imputação de dados foi escolhida a base de dados do INMET denominado BDMEP (Banco de Dados Meteorológicos para Ensino e Pesquisa). Esta base contém, de acordo com o INMET, "dados meteorológicos diários em forma digital, de séries históricas das várias estações meteorológicas convencionais da rede de estações do INMET com milhões de informações, referentes às medições diárias, de acordo com as normas técnicas internacionais da Organização Meteorológica Mundial" [16].

Dessa forma, foram escolhidos os dados referentes a estação meteorológica do Mirante de Santana por conta da sua relevância para o INMET, pois é a principal estação meteorológica da cidade de São Paulo e foi verificado que praticamente não contém valores faltantes para o intervalo de anos que se resolveu considerar.

A base de dados referente a Mirante de Santana abriga informações de 1961 até os dias atuais, porém foram escolhidos dados no intervalo de 01/01/1971 a 01/01/2011 (o que totaliza 21268 linhas). Esta redução no conjunto de dados a ser utilizado foi realizada pois, conforme recomendações da Organização Meteorológica Mundial (OMM), 30 anos são suficientes para representar os valores climáticos de determinada região. Além disso, o ano inicial de 1981 foi escolhido seguindo a recomendação da OMM para que tal contagem de 30 anos seja iniciada no primeiro ano de cada década. Tais especificações são denominadas normais meteorológicas [15].

Alguns valores presentes na base de dados apresentam dados faltantes. No caso das temperaturas, os intervalos de *gaps* que ultrapassam mais de 2 dias consecutivos estão apresentadas na Tabela II.

Os dados armazenados são obtidos diariamente, em dois períodos distintos. No primeiro caso, os dados são obtidos às 00:00, sendo que dentre os valores exibidos anteriormente, o único obtido neste caso é a temperatura mínima, a qual é medida e armazenada às 12:00. Tal situação, que pode ser observada na Figura 6, gera *gaps* entre os valores das linhas.

B. Escolha dos atributos

O atributo escolhido para a realização do trabalho foi a temperatura máxima diária. Para a seleção do atributo, foi feita uma pesquisa no *Google Scholar* com o uso da *string* de busca "*meteorological time series imputation*"com a posterior análise dos quinze primeiros artigos retornados pela busca pelo filtro de busca por relevância. Em seguida, os atributos utilizados em cada artigo foram enumerados, e estão representados na Tabela I.

Assim, é possível concluir também que há quatro atributos presentes nos trabalhos que se encontram na base utilizada, sendo tais atributos os quatro mais utilizados. Os valores de temperatura e precipitação estão presentes em 7 deles, enquanto que a velocidade do vento é utilizada em 5 projetos, e a umidade do ar, em 4 projetos.

Desta forma, eles se mostram atributos importantes a serem considerados para o estudo da imputação de valores faltantes. No caso do parâmetro de temperatura, a base contém disponíveis os valores de temperatura máxima, mínima e compensada média. Porém, foi verificado que este último não está presente nos trabalhos estudados, o que levou a decisão de considerar apenas as temperaturas máxima e mínima. Como opção de estudo, resolveu-se realizar o presente projeto apenas com o uso da temperatura máxima.

A velocidade do vento média e a umidade do ar, por sua vez, contém uma grande quantidade de valores faltantes na base, o que dificulta o estudo por conta de não haver valores com os quais comparar após a imputação realizada. Assim, o atributo escolhido para uso será a temperatura máxima diária.

Projeto	Media de avaliação	Medida de erro	Atributos utilizados	Algoritmos utilizados
Bauer; Deistler; Sherrer (2001) [25]	Acurácia	Raiz do erro quadrático médio	Precipitação diária e temperatura diária	Modelo autoregressivo
Campozano et al. (2014) [12]	Não especifica	Erro médio, erro médio absoluto, raiz do erro quadrático médio	Precipitação diária e temperatura diária	Inverso da potência das distâncias, kNN, regressão linear, regressão linear múltipla, valor médio
Ferrari; Ozaki (2014) [26]	Não especifica	Raiz do erro quadrático médio	Precipitação	Inverso da potência das distâncias, kNN, regressão linear
Junger; de Leon (2015) [23]	Acurácia, precisão	Raiz do erro quadrático médio, erro médio absoluto	Temperatura e umidade relativa	Maximização de expectativa
Junninen et al. (2004) [6]	Acurácia	Índice de concordância, coeficiente de determinação, raiz do erro quadrático médio, erro absoluto médio	Velocidade do vento, direção do vento, umidade, temperatura, qualidade do ar	Interpolação, kNN, Mapas de Kohonen, MLP, REGEM
Niska et al. (2004) [28]	Não especifica	Não especifica	Qualidade do ar	Algoritmo genético
Nkiaka; Nawas; Lovett (2016) [22]	Acurácia	Não especifica	Não especifica	Mapas de Kohonen
Park; Genton; Ghosh (2007) [18]	Eficiência	Erro quadrático médio	Precipitação e descarga fluvial	Método de imputação desenvolvido
Paatero et al. (2005) [19]	Não especifica	Não especifica	Temperatura máxima, temperatura mínima, temperatura média, ponto de orvalho, umidade relativa, pressão, precipitação, velocidade do vento, irradiação solar	Inverso da potência das distâncias, kNN, regressão linear
Roberts (2005) [27]	Não se refere a imputação	Não se refere a imputação	Não se refere a imputação	Não se refere a imputação
Shukur; Lee (2015) [29]	Acurácia	Raiz do Erro Quadrático Médio	Velocidade do vento	Redes neurais artificiais
Simolo et al. (2010) [20]	Acurácia	Erro médio, erro médio absoluto, raiz do erro quadrático médio	Precipitação diária	Método de imputação desenvolvido
Turrado et al. (2014) [21]	Não especifica	Raiz do Erro Quadrático Médio, erro absoluto médio	Temperatura, umidade, precipitação, irradiação, horas de sol	Inverso da potência das distâncias, MICE, Regressão linear múltipla
Yi et al. (2016) [24]	Acurácia	Erro quadrático médio, erro relativo médio	Qualidade do ar, velocidade do vento, umidade	Método de imputação desenvolvido
Yozgatligil et al.(2013) [3]	Acurácia	Raiz do erro quadrático médio	Precipitação mensal máxima, temperatura mensal média	Valor médio, Maximização de expectativa, MLP

Tabela I
PROJETOS E FATORES RELACIONADOS

Tabela II
INTERVALOS DE DADOS DE TEMPERATURA FALTANTES NA BASE DE DADOS

Dado	Intervalo faltante
Temperaturas máxima e mínima (ºC)	02/12/1980-31/01/1981, 02/04/1981-30/04/1981, 02/02/1982-31/03/1982, 02/01/1983,30/01/1983, 01/02/1983-31/12/1983

C. Técnicas e métodos

A seguir, foram escolhidas as técnicas a serem utilizadas. Para tanto, os mesmos artigos foram analisados, e os resultados das técnicas utilizadas se encontram na Tabela I. Pode-se perceber que os métodos mais utilizados são o kNN e o inverso da potência das distâncias, com 4 artigos utilizando cada um deles. A regressão linear também foi utilizada em 5 trabalhos, sendo que em 2 deles foi especificado o uso da regressão linear múltipla. As redes neurais artificiais foram utilizadas em 3 projetos, com 2 deles utilizando a classe *multilayer perceptron*.

Os métodos do inverso da potência das distâncias, de valor médio e de máximização das expectativas, apesar de serem bem citados, não serão estudados por serem métodos de interpolação ou estatísticos. Já algoritmos mais complexos como os mapas de Kohonen e os algoritmos genéticos não serão incluídos por terem sido pouco utilizados, o que também vale para outros algoritmos e métodos restantes.

O algoritmo da regressão linear se mostra ser um dos mais citados em relação a uso. Apesar de aparecer em trabalhos como os de [26] e [21] como tendo pior desempenho que outros algoritmos apresentados em tais artigos, foi escolhido para ser utilizado neste projeto, pela sua relevância, aparecendo em boa parte da pesquisa.

Portanto, por serem alguns dos mais citados na pesquisa, a Regressão Linear e as redes neurais artificiais foram escolhidas para implementação.

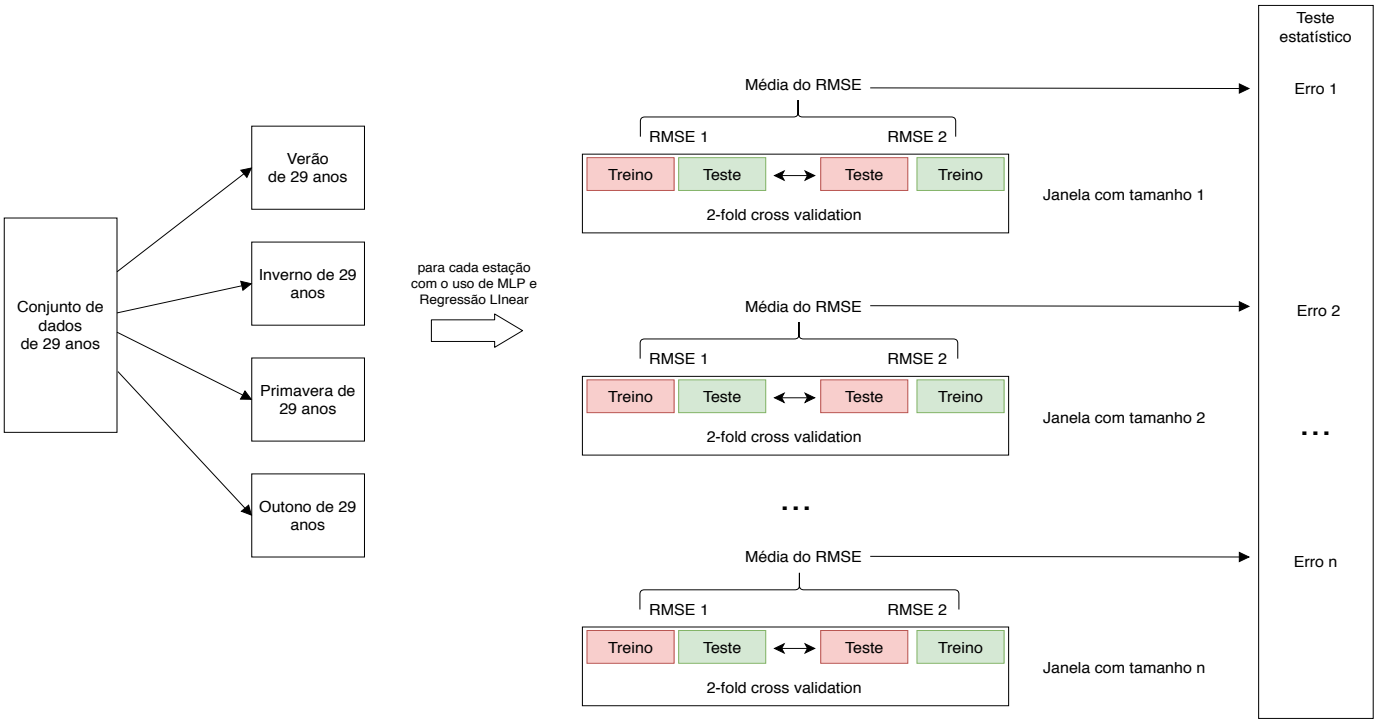

Figura 5. Protocolo do processo experimental

Estacao	Data	Hora	Precipitacao	TempMaxima	TempMinima	Insolacao	Evaporacao Piche	Temp Comp Media	Umidade Relativa Media	Velocidade do Vento Media
83781	04/07/1987	0		26		0	3.9	19.88	66	1.466667
83781	04/07/1987	1200	0		11.8					
83781	05/07/1987	0		21.4		0	4.5	15.38	76.5	3.8
83781	05/07/1987	1200	0		11.9					
83781	06/07/1987	0		22.6		0	2.2	17.6	79.5	2.333333
83781	06/07/1987	1200	0		13.6					
83781	07/07/1987	0		25.3		5.2	1.8	19.14	63	2.9
83781	07/07/1987	1200	0		13.4					
83781	08/07/1987	0		25.5		1.8	5.1	18.96	69.75	2.033333
83781	08/07/1987	1200	0		15					
83781	09/07/1987	0		20.4		2	4.4	16.1	84.5	1.433333
83781	09/07/1987	1200	7.4		14					
83781	10/07/1987	0		19.5		3	1.4	14.42	89.75	2.233333
83781	10/07/1987	1200	0		11.8					
83781	11/07/1987	0		26		6.2	1	16.82	75.75	2.233333
83781	11/07/1987	1200	0		9.9					
83781	12/07/1987	0		24.3		0.1	3.3	18.34	73	1.366667
83781	12/07/1987	1200	0		15					
83781	13/07/1987	0		28.2		6.5	3.1	22.26	50.75	3.3
83781	13/07/1987	1200	0		14.5					
83781	14/07/1987	0		29		5.2	8.2	22.16	64.25	1.266667
83781	14/07/1987	1200	0		16.6					
83781	15/07/1987	0		29.3		8	4.6	22.16	58.25	1.633333
83781	15/07/1987	1200	0		16.4					
83781	16/07/1987	0		28.6		6.9	8.2	23.48	46	5.066667
83781	16/07/1987	1200	0		17.2					
83781	17/07/1987	0		29		7.5	10.5	23.5	46	3.666667
83781	17/07/1987	1200	0		20					
83781	18/07/1987	0		29		6.3	8.7	22.62	50	1.733333
83781	18/07/1987	1200	0		16.5					

Figura 6. Estrutura dos dados

D. Janelas de dados geradas para experimentos

O estudo envolve a escolha de regiões do conjunto de dados para criação de janelas (*gaps*) visando a execução dos algoritmos de imputação, para posterior comparação por medidas de avaliação e testes estatísticos. Uma janela é, em uma série temporal, uma região dos dados em que não há informações.

Para a base utilizada, foram verificados intervalos relacionados entre si. Após tal análise, foi escolhido utilizar as estações como período. Ou seja, os dados foram divididos em oito: duas partições para verão, inverno, primavera e outono, com cada partição contendo 15 e 14 anos de dados, respectivamente). Assim, a janela de dados será de um ponto por vez (ou seja, um dia), visando a imputação de tais dados durante as execuções dos algoritmos.

E para se estimar esse valor, realizaram-se análises com período crescente de dias. Ou seja, o primeiro experimento utilizou um dia anterior ao dia com temperatura predita, o segundo utilizou dois dias anteriores, e assim por diante. Dessa forma, foi definido que os experimentos seriam finalizados quando os valores de erro se tornassem altamente crescentes entre eles.

Tal escolha de divisão de partições influencia diretamente no uso dos algoritmos escolhidos. Como houve o particionamento entre quatro estações diferentes, então a implementação de uma rede neural para cada estação se torna necessária. Além disso, a Regressão Linear realiza experimentos disjuntos, cada um deles utilizando os dados de uma estação.

E. Medidas de avaliação

É possível concluir também que a medida de avaliação mais usada é a acurácia, a qual está presente em 8 trabalhos. Além disso, a raiz do erro quadrático médio (RMSE) foi a mais utilizadas dentre as medidas do cálculo do erro. Portanto, tais medidas serão utilizadas no trabalho.

F. Protocolo de validação

Após a definição dos métodos, é necessário verificar a acurácia dos dados de estimativa [29] e dos artigos analisados. Uma forma popular para tal se chama validação cruzada [21] [29], a qual é uma técnica estatística utilizada para o particionamento de conjuntos de dados em treino e teste. Nesta técnica, os dados são divididos em N diferentes partições com tamanhos aproximadamente iguais, de tal forma que N-1 partições são utilizadas para treino, e o restante para teste em cada experimento. Este procedimento é repetido até que cada

partição tenha sido utilizada no máximo uma vez para testes [31].

O primeiro passo foi a escolha do número de partições. O primeiro experimento utilizou uma par

O primeiro passo foi a escolha do número de partições. Como visou-se utilizar o protocolo de validação cruzada do tipo 2-fold, então a forma para divisão do conjunto de dados em partições se deu de forma quase igualitária. Assim, para cada estação, foram criadas uma partição contendo dados de 15 anos, e outra contendo 14 anos, respectivamente.

Cada partição contém os dados de cada ano contínuos ao longo do tempo. Tal escolha de continuidade dos dados em cada ano se deu de acordo com [26], que afirma que séries temporais são geradas por processos que evoluem com o tempo. Desta forma, necessitam de uma continuidade dos dados, sendo preciso analisar dia-a-dia em sequência para uma predição mais efetiva dos dados seguintes.

Além disso, visando não tornar o modelo dependente da entrada em sequência dos anos, as partições contiveram os anos distribuídos de forma aleatória em relação a entrada. Ou seja, para uma partição de 1995 a 2010 de uma estação, o primeiro ano poderia ser 2005, o seguinte 1994, e assim por diante.

G. Testes estatisticos

Para a análise de qual teste estatístico utilizar para comparação entre os algoritmos a serem utilizados, primeiramente foram estudados os artigos obtidos pela pesquisa por relevância anteriormente revisados. Porém, somente o trabalho de [26] indicou que utilizou algum teste estatístico. Tal método, denominado teste de distribuição de Gumbel, não se aplica ao presente projeto por ser utilizado quando há muitos dados faltantes e *outliers*.

Portanto, uma segunda pesquisa na literatura demonstrou que o *Wilcoxon Signed-Rank Test* é utilizado por trabalhos como [32] e [33] para comparação entre os métodos de imputação criados pelos próprios com outros já existentes, tendo sido escolhido para uso no presente trabalho. A Figura 5 exibe todo o processo de protocolo experimental.

VI. Resultados e Discussões

Para os testes, a ideia inicial foi fazer a previsão de um dia com base nos dias anteriores, de acordo com os *gaps* e períodos estipulados na Seção III, para o ano todo. O primeiro teste foi realizado com uma rede neural MLP para regressão, e os dados obtidos estão apresentados na Tabela III.

Para MLP, foram feitos testes com 2, 3 e 4 camadas ocultas. Os resultados da rede com 4 camadas foram muito insatisfatórios, com uma taxa de erro muito alta. Logo, não foram considerados nessa análise.

Para os testes com duas e três camadas, foram realizados com diferentes taxas de aprendizado, sendo eles 0.01, 0.2 e 0.4 e para cada taxa, foram usadas 3 números de épocas diferentes, sendo elas 500, 1000 e 3000. Após avaliar os resultados, foi notado que eles indiferem com a variação de épocas, então também não foram considerados nessa análise.

Tabela III
Resultado da execução da rede neural MLP

Número de entradas	**Camadas ocultas**	**Taxa de aprendiza-gem**	**RMSE**
5	2	0.01	3.0158
5	2	0.2	3.6356
5	2	0.4	4.1541
5	3	0.01	3.0176
5	3	0.2	3.6377
5	3	0.4	4.2974
10	2	0.01	3.0153
10	2	0.2	3.0884
10	2	0.4	3.1166
10	3	0.01	3.0176
10	3	0.2	3.6377
10	3	0.4	4.2399
15	2	0.01	3.0148
15	2	0.2	3.8653
15	2	0.4	4.3887
15	3	0.01	3.0291
15	3	0.2	3.8046
15	3	0.4	4.174

Observando a tabela, nota-se que em quase todos os casos, o RMSE (*Root Mean Squared Error*) se mantém muito alto, variando entre 3 e 3,1 nos melhores casos. Portanto, o algoritmo não foi eficiente na previsão dos dados.

O segundo teste foi realizado através de regressão linear, utilizando 5, 10 e 15 entradas de temperatura de dias consecutivos para a previsão do próximo. Os resultados estão exibidos na Tabela IV.

Tabela IV
Resultado da Regressão Linear

Número de entradas	**RMSE**
5	3.0168
10	3.0137
15	2.9756

Pode-se ver que os resultados estão mais estáveis, com uma correlação baixa entre eles e o número de entradas (dias). Porém, os resultados continuam insatisfatórios, mostrando ainda um erro relativamente alto.

À primeira vista, os resultados obtidos não são satisfatórios, mas analisando os gráficos gerados a partir das temperaturas preditas, juntamente com as observadas, na Figura 7, podemos notar que a rede neural se aproxima dos valores reais. Pode-se dizer que os dados previstos se distanciam dos observados em picos de temperatura, sejam eles altos ou baixos, mas em pontos médios, eles se aproximam. De forma semelhante, o mesmo comportamento é apresentado no gráfico gerado pelos testes de regressão linear, na Figura 8.

Em ambos os casos, o erro médio fica em torno de 3. Ainda é um valor alto para o erro, o que mostra que os algoritmos ainda não estão otimizados. Mas por se tratarem de dados de uma série temporal, a margem de erro considerada é de 2 graus, para cima ou para baixo, um valor relativamente próximo do 2,9756 do melhor caso testado neste projeto.

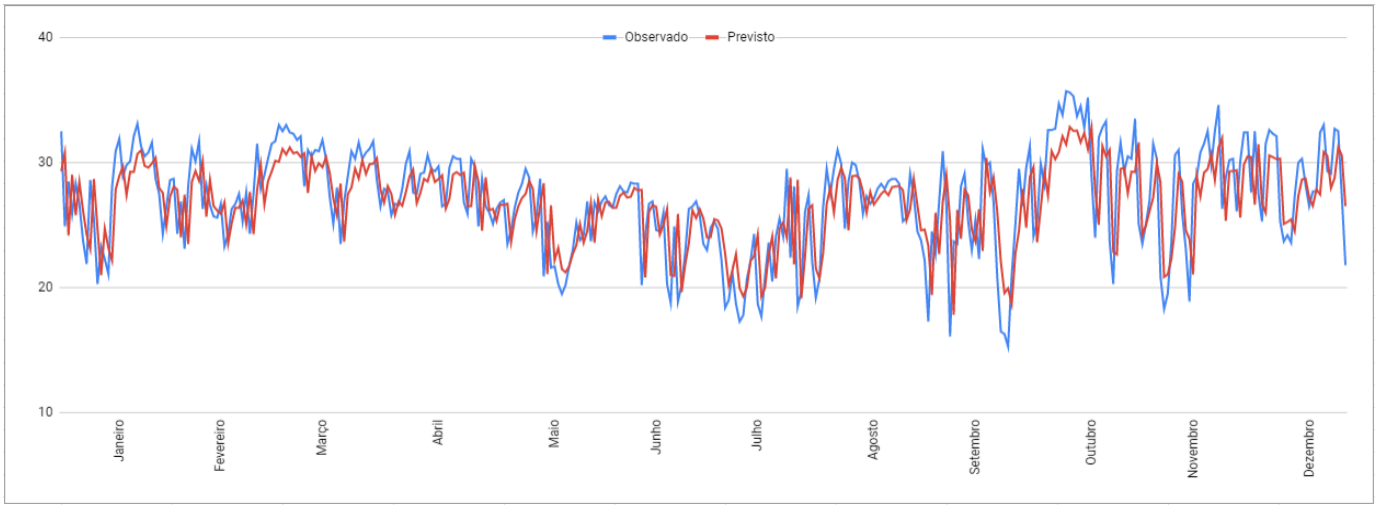

Figura 7. Gráfico de temperaturas preditas e observadas para o ano de 2002 pela rede neural MLP

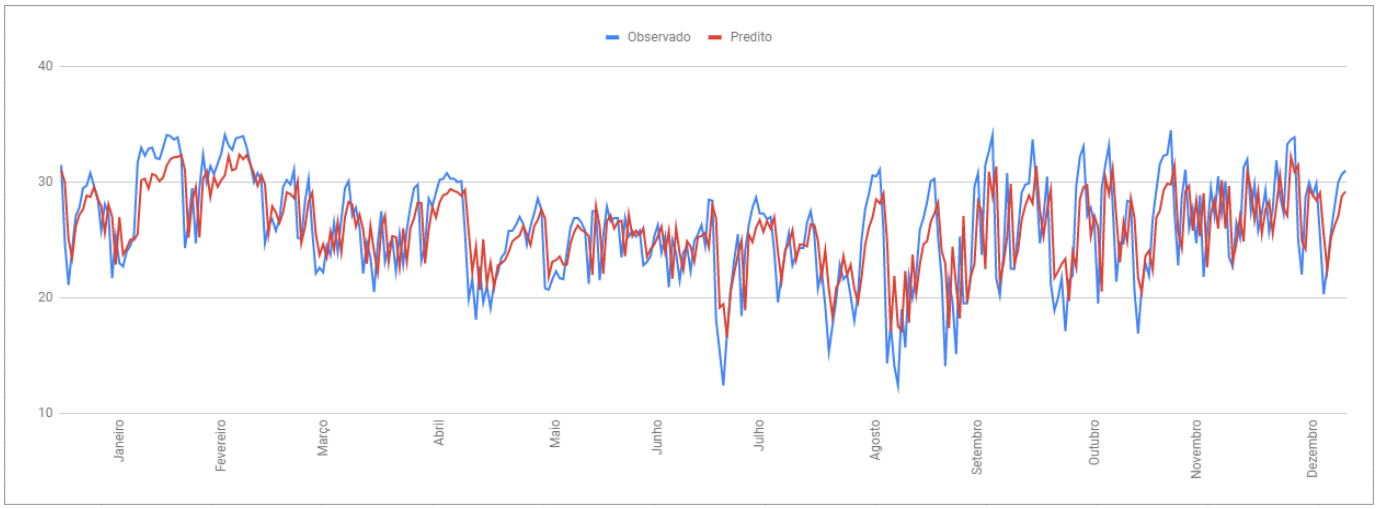

Figura 8. Gráfico de temperaturas preditas e observadas para o ano de 2003 por Regressão Linear

Em seguida, foi optado por trocar a abordagem anual dos dados para uma menor, definida por estações, como exibido na Figura 6. Nesse segundo experimento, foram gerados testes com Redes Neurais MLP e Regressão Linear em *gaps* de 1 à 15. Assim, para tal experimento, considerou-se a modificação de diferentes valores de parâmetros por testes em MLPs visando melhores comparações de desempenho. Tais parâmetros foram:

- Janelas (entradas na rede): de 1 a 15
- Taxa de aprendizado : 0.01; 0.2; 0.4
- Arquitetura de rede: 2 camadas ocultas (3 neurônios em cada), 2 camadas ocultas (4 neurônios em cada), 3 camadas ocultas (4 neurônios na primeira, 3 na segunda e 4 na terceira, partindo das entradas à saida)
- Épocas de treinamento: 500, 1000, 2000

Nessa abordagem, foi notado que os algumas estações possuem uma previsibilidade melhor em relação as outras. Estações como inverno e primavera mostraram um erro maior que a média do experimento passado, com o RMSE variando entre 3,5 e 3,6 para o inverno e 3,1 e 3,2 para a primavera, nos melhores casos de ambos. Já para o verão e outono, notou-se uma melhora significativa no erro, pois o verão, por exemplo, mostrou uma média de de 2,2 à 2,3. O outono, por sua vez, de 2,5 nos melhores casos.

Outro ponto notável foi que, para os MLPs, as redes com 3 camadas ocultas foram, em média, piores que as redes com 2. A taxa de aprendizado também se mostrou importante, pois os menores valores de erro se mostraram diretamente associadas aos valores de aprendizado, ou seja, 0,01 de taxa de aprendizado conseguiu uma média de resultados melhor que 0,2, que por sua vez conseguiu resultados melhores que 0,4. Ao se realizar uma analise geral nessas redes, verificou-se a seguinte configuração:

- Taxa de Aprendizado: 0,01;
- Épocas: 500;
- Arquitetura de Rede: duas camadas ocultas, cada uma

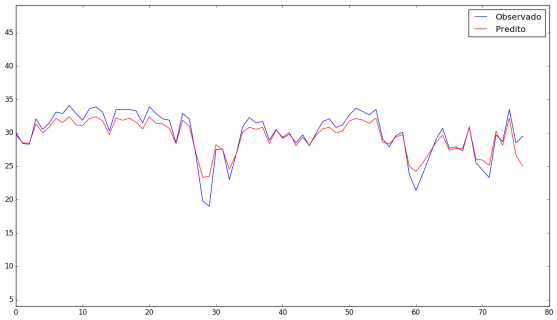

Figura 9. Gráfico de temperaturas preditas e observadas para o verão de 2000 por MLP

com 4 neurônios.

Entre os melhores resultados obtidos, essa configuração de rede foi a mais recorrente, além de ser absoluta nos melhores resultados com 10 ou mais *inputs* na rede. Na Figura 9, pode-se observar uma previsão utilizando essa configuração de rede para o MLP, onde os dados preditos se aproximaram bastante dos observados. A Tabela V exibe uma comparação entre as taxas de erro obtidas entre a melhor configuração da rede MLP e os valores médios da Regressão Linear.

Tabela V
COMPARAÇÃO DE TAXAS DE ERRO ENTRE REGRESSÃO LINEAR E MLP

Estação	**Regressão Linear**	**MLP**
Outono	2.51163333333	2.5309
Inverno	3.67033	3.73645
Primavera	3.23540333333	3.21255
Verão	2.25087333333	2.2208

Outra forma utilizada para a análise dos dados foi o teste estatístico de *Wilcoxon Signed-Rank* [36]. Para tal, analisou-se a média dos valores de erro para cada janela utilizada nos algoritmos MLP e Regressão Linear, os quais foram obtidos pelo uso do método de validação cruzada 2-fold. Assim, os valores foram comparados de forma pareada. O processo é exibido na Figura 6.

No caso da Regressão Linear, o cálculo para a obtenção dos valores de RMSE utilizados para o teste estatístico envolveu justamente todos os valores de temperatura calculados. Porém, tal situação não acontece no caso da MLP, o qual possui diferentes configurações de rede, de forma que são várias as possibilidades de amostras a serem escolhidas para teste.

Dessa forma, optou-se por verificar dentro dos valores das quatro redes neurais criadas para as estações qual continha a melhor configuração em relação a média do RMSE para cada janela. Ou seja, para cada rede, realizou-se um sistema de votação em relação às redes com 15 diferentes entradas visando encontrar tal configuração.

Dessa forma, a rede escolhida para as quatro estações foi a com taxa de aprendizado igual a 0.01, quantidade de épocas igual a 500 e duas camadas ocultas, com 4 neurônios em cada camada. A Tabela VI exibe os valores obtidos com a aplicação do teste de *Wilcoxon Signed-Rank*.

Tabela VI
RESULTADOS DO TESTE DE WILCOXON SIGNED-RANK POR ESTAÇÃO

Estação	**valor-p**
Outono	0.030908013652429556
Inverno	0.053474367154484302
Primavera	0.39424586721077726
Verão	0.023095732617466558

Os resultados do teste estatístico mostram que, para uma taxa de confiança de 95%, os modelos obtidos para outono e verão não contém uma distribuição semelhante. Por sua vez, os modelos de inverno e primavera se mostram relacionados em relação a isso, com o valor-p da primavera sendo bem discrepante em relação aos das outras estações.

Por fim, apesar de altos valores da RMSE no primeiro experimento e no segundo, nos casos da primavera e inverno, pode-se afirmar que o trabalho obteve resultados relativamente bons para o problema proposto.

A seguir, alguns problemas encontrados na obtenção de resultados melhores:

- A falta de correlação dos dados se mostrou uma dificuldade, pois as abordagens escolhidas (por ano e por estações) se mostraram ineficientes em alguns casos, com bons resultados apenas para verão e outono. Talvez uma abordagem diferente se mostre mais efetiva. Para isso, seria necessário um conhecimento mais aprofundado do campo da meteorologia para estabelecer correlações mais efetivas para a imputação dos dados;
- Por serem séries temporais, alguns fenômenos meteorológicos influenciam nas medições, como frentes frias, que alteram os padrões de vários dias consecutivos ou fenômenos maiores, e como *La niña*, que podem influenciar um ano inteiro;
- Para testes mais precisos, um estudo sobre escolhas de técnicas e seus parâmetros talvez poderia ajudar. A escolha destes não é algo trivial. Portanto, tanto um estudo mais profundo na literatura quanto uma maior quantidade de testes empíricos podem levar a melhores resultados.

VII. CONCLUSÃO E TRABALHOS FUTUROS

A imputação de dados faltantes para séries temporais ainda é um terreno com mais perguntas do que respostas. Ainda se vê a necessidade de estudos mais aprofundados na área, e de maneiras mais eficientes para a utilização de métodos de aprendizado de máquina na meteorologia. Foram encontradas dificuldades na busca de melhores modelos e configurações para eles, mas sua utilização para essa área da ciência mostra-se promissor. Um trabalho conjunto com técnicos de meteorologia poderia contribuir para um melhor desempenho.

Também foi observado uma ligeira superioridade em resultados satisfatórios por parte do MLP em relação a Regressão Linear, mostrando que métodos mais complexos, se bem trabalhados, podem apresentar melhores resultados.

Finalmente, neste projeto, concluiu-se que:

- Uma taxa de aprendizado pequena se saiu melhor para o modelo de MLP;
- Uma arquitetura de rede com poucas camadas trouxe taxas de erro menores;
- A menor taxa de erro foi obtida com o uso de uma baixa quantidade de épocas;
- É difícil de se prever os *outliers* das temperaturas, sejam elas muito baixas ou altas;
- Existem estações do ano que apresentam erros menores (verão e outono);
- Os algoritmos MLP e RL se mostraram muito semelhantes nos testes.

Para trabalhos futuros, é necessário reavaliar o modo em que os dados foram abordados. Dessa forma, eles podem ser analisados para permitir novas tentativas de encontrar melhores tamanhos de *gaps*, correlações mais relevantes entre os dados e estruturas melhores para as redes neurais. Outra possibilidade é o uso de uma base de dados maior com informações de outras estações meteorológicas, para maior variedade de dados.

REFERÊNCIAS

[1] Brockwell, Peter J., Richard A. Davis, and Matthew V. Calder. Introduction to time series and forecasting. Vol. 2. New York: springer, 2002.

[2] Ehlers, Ricardo S. "Análise de séries temporais."Laboratório de Estatística e Geoinformação. Universidade Federal do Paraná (2007).

[3] Yozgatligil, Ceylan, et al. "Comparison of missing value imputation methods in time series: the case of Turkish meteorological data."Theoretical and applied climatology 112.1-2 (2013): 143-167.

[4] Ramos-Calzado, P., et al. "A novel approach to precipitation series completion in climatological datasets: application to Andalusia."International Journal of Climatology: A Journal of the Royal Meteorological Society 28.11 (2008): 1525-1534.

[5] Norazian, Mohamed Noor, et al. "Estimation of missing values in air pollution data using single imputation techniques."ScienceAsia 34.3 (2008): 341-345.

[6] Junninen, Heikki, et al. "Methods for imputation of missing values in air quality data sets."Atmospheric Environment 38.18 (2004): 2895-2907.

[7] BDMEP - Banco de Dados Meteorológicos para Ensino e Pesquisa. Internet: http://www.inmet.gov.br/portal/index.php?r=bdmep/bdmep. 13 de abril, 2019.

[8] Silva-Ramírez, Esther-Lydia, et al. "Missing value imputation on missing completely at random data using multilayer perceptrons."Neural Networks 24.1 (2011): 121-129.

[9] Layanun, Vichaya, Supachai Suksamosorn, and Jitkomut Songsiri. "Missing-data imputation for solar irradiance forecasting in Thailand."2017 56th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE). IEEE, 2017.

[10] Richman, Michael B.; Trafalis, Theodore B.; Adrianto, Indra. Missing data imputation through machine learning algorithms. In: Artificial intelligence methods in the environmental sciences. Springer, Dordrecht, 2009. p. 153-169.

[11] Sridevi, S., et al. "Imputation for the analysis of missing values and prediction of time series data."2011 International Conference on Recent Trends in Information Technology (ICRTIT). IEEE, 2011.

[12] Campozano, Lenin, et al. "Evaluation of infilling methods for time series of daily precipitation and temperature: The case of the Ecuadorian Andes."Maskana 5.1 (2014): 99-115.

[13] Wu, Shin-Fu, Chia-Yung Chang, and Shie-Jue Lee. "Time series forecasting with missing values."2015 1st International Conference on Industrial Networks and Intelligent Systems (INISCom). IEEE, 2015.

[14] Amiri, Mehran, and Richard Jensen. "Missing data imputation using fuzzy-rough methods."Neurocomputing 205 (2016): 152-164.

[15] INMET. Metodologia. Disponível em: http://www.inmet.gov.br/webcdp/climatologia/normais/imagens/normais/textos/metodologia.pdf. Acessado em 11 de maio de 2019.

[16] INMET. BDMEP - Banco de Dados Meteorológicos para Ensino e Pesquisa. Disponível em: http://www.inmet.gov.br/portal/index.php?r=bdmep/bdmep. Acessado em 11 de maio de 2019.

[17] Coulibaly, P.; Evora, N. D. Comparison of neural network methods for infilling missing daily weather records. Journal of hydrology, v. 341, n. 1-2, p. 27-41, 2007.

[18] Park, Jung Wook; GENTON, Marc G.; GHOSH, Sujit K. Censored time series analysis with autoregressive moving average models. Canadian Journal of Statistics, v. 35, n. 1, p. 151-168, 2007.

[19] Paatero, Pentti et al. Estimating time series of aerosol particle number concentrations in the five HEAPSS cities on the basis of measured air pollution and meteorological variables. Atmospheric Environment, v. 39, n. 12, p. 2261-2273, 2005.

[20] Simolo, C. et al. Improving estimation of missing values in daily precipitation series by a probability density function-preserving approach. International Journal of Climatology, v. 30, n. 10, p. 1564-1576, 2010.

[21] Turrado, Concepción et al. Missing data imputation of solar radiation data under different atmospheric conditions. Sensors, v. 14, n. 11, p. 20382-20399, 2014.

[22] Nkiaka, E.; Nawaz, N. R.; Lovett, J. C. Using self-organizing maps to infill missing data in hydro-meteorological time series from the Logone catchment, Lake Chad basin. Environmental monitoring and assessment, v. 188, n. 7, p. 400, 2016.

[23] Junger, W. L.; De Leon, A. Ponce. Imputation of missing data in time series for air pollutants. Atmospheric Environment, v. 102, p. 96-104, 2015.

[24] Yi, Xiuwen et al. ST-MVL: filling missing values in geo-sensory time series data. 2016.

[25] Bauer, Gerd; DEISTLER, Manfred; SCHERRER, Wolfgang. Time series models for short term forecasting of ozone in the eastern part of Austria. Environmetrics: The official journal of the International Environmetrics Society, v. 12, n. 2, p. 117-130, 2001.

[26] Ferrari, Gláucia Tatiana; Ozaki, Vitor. Missing data imputation of climate datasets: Implications to modeling extreme drought events. Revista Brasileira de Meteorologia, v. 29, n. 1, p. 21-28, 2014.

[27] Roberts, Steven. An investigation of distributed lag models in the context of air pollution and mortality time series analysis. Journal of the Air Waste Management Association, v. 55, n. 3, p. 273-282, 2005.

[28] Niska, Harri et al. Evolving the neural network model for forecasting air pollution time series. Engineering Applications of Artificial Intelligence, v. 17, n. 2, p. 159-167, 2004.

[29] Shukur, Osamah Basheer; Lee, Muhammad Hisyam. Imputation of missing values in daily wind speed data using hybrid AR-ANN method. Modern Applied Science, v. 9, n. 11, p. 1, 2015.

[30] Gutierrez-Corea, F.-V., Manso-Callejo, M.-A., Moreno-Regidor, M.-P., Velasco-Gómez, J. (2014). Spatial Estimation of Sub-Hour Global Horizontal Irradiance Based on Official Observations and Remote Sensors. Sensors, 14(4), 6758–6787. doi:10.3390/s140406758

[31] Refaeilzadeh, Payam; Tang, Lei; Liu, Huan (2009). Cross-validation. Encyclopedia of database systems, p. 532-538.

[32] Aydilek, Ibrahim Berkan; Arslan, Ahmet. A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. Information Sciences, v. 233, p. 25-35, 2013.

[33] Ravi, Vadlamani; Krishna, Mannepalli. A new online data imputation method based on general regression auto associative neural network. Neurocomputing, v. 138, p. 106-113, 2014.

[34] Bergmeir, C., Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. Information Sciences, 191, 192–213. doi:10.1016/j.ins.2011.12.028

[35] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

[36] WILCOXON, Frank. Individual comparisons by ranking methods. Biometrics bulletin, v. 1, n. 6, p. 80-83, 1945.

[37] ADETILOYE, Taiwo; AWASTHI, Anjali. Predicting Short-Term Congested Traffic Flow on Urban Motorway Networks. In: Handbook of Neural Computation. Academic Press, 2017. p. 145-165.

[38] YAN, Xin; SU, Xiaogang. Linear regression analysis: theory and computing. World Scientific, 2009.

[39] ROSENBLATT, Frank. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review, v. 65, n. 6, p. 386, 1958.

[40] COUTINHO, E. R.; SILVA, R. M.; DELGADO, A. R. S. Utilização de Técnicas de Inteligência Computacional na Predição de Dados Meteorológicos. Revista Brasileira de Meteorologia, v. 31, n. 1, p. 24-36, 2016.