Revista Brasileira de Climatologia
Brazilian Journal of Climatology
OPEN ACCESS

SISTEMA WEB PARA PRÉ-PROCESSAMENTO E ANÁLISE DE DADOS METEOROLÓGICOS

Web System for pre-processing and analysis of Weather data

Sistema web de preprocesamiento y análisis de datos meteorológicos

Walingson da Silva da Costa
Universidade Estadual do Mato grosso
walingson.costa@unemat.br

Rivanildo Dallacort
Universidade Estadual do Mato grosso
rivanildo@unemat.br

Marcos Antônio Camillo de Carvalho
Universidade Estadual do Mato grosso
marcocarvalho@unemat.br

Silmara Bispo dos Santos
Universidade Federal do Mato grosso
silmara@ufr.edu.br

Resumo: O entendimento do tempo e do clima é indispensável para a tomada de decisões assertivas em diversos campos da atuação humana, necessitando, portando de dados consistentes e confiáveis. Deste modo, o objetivo com este trabalho foi descrever as funcionalidades de um sistema (web) desenvolvido com intuito de identificar erros e imputar dados ausentes em séries históricas de dados meteorológicos, descrevendo as características e erros da base de dados do INMET (Instituto Nacional de Meteorologia) nos municípios de Matupá MT e Sinop MT. O sistema foi construído com a linguagem de programação Python, as bibliotecas Scikit-learn, SciPy, Pandas, Plotly e o Framework Streamlit. Para validação do sistema foi utilizado série histórica de dados meteorológicos fornecidos pelo INMET, tratados suas falhas e os valores ausentes foram imputados com o algoritmo KNNImputer. A assertividade da imputação de valores ausentes foi verificada através das métricas de Acurácia, Precisão, Recall, F1-score e Erro Quadrático Médio (QMS). Tais métricas são oriundas de comparação de valores previstos e valores originais por matriz de confusão. O sistema foi eficiente na identificação de outliers e na imputação de valores ausentes, identificando 100% dos valores discrepantes das variáveis analisadas.

Palavras-chave: Data Wrangling. Mineração de dados. Computação aplicada.

Abstract: Understanding the weather and climate is essential for making assertive decisions in various fields of human activity, requiring, therefore, consistent and reliable data. Thus, the objective of this work was to describe the functionalities of a system (web) developed to identify errors and impute missing data in historical series of meteorological data, describing the characteristics and errors of the INMET (National Institute of Meteorology) database from the municipalities of Matupá MT and Sinop MT. The system was built with the Python programming language, the Scikit-learn, SciPy, Pandas, Plotly libraries and the Streamlit Framework. For the validation of the system, a historical series of meteorological data provided by INMET was used, its failures were treated and the missing values were imputed with the KNNImputer algorithm. The assertiveness of imputation of missing values was verified through the metrics of Accuracy, Precision, Recall, F1-score and Mean Square Error (QMS). These metrics are derived from the comparison of predicted values and original values by confusion matrix. The system was efficient in identifying outliers and imputing missing values, identifying 100% of the outliers of the variables analyzed.

Keywords: Data Wrangling. Data mining. Applied computing.

Resumen: La comprensión del tiempo y el clima es esencial para la toma de decisiones asertivas en diversos campos de la acción humana, requiriendo, por tanto, datos consistentes y fiables. De esta forma, el objetivo de este trabajo fue describir las funcionalidades de un sistema (web) desarrollado para identificar errores e imputar datos faltantes en series históricas de datos meteorológicos, describiendo las características y errores en la base de datos del INMET (Instituto Nacional de Meteorología) en los municipios de Matupá MT y Sinop MT. El sistema fue construido con el lenguaje de programación Python, las librerías Scikit-learn, SciPy, Pandas, Plotly y el Framework Streamlit. Para validar el sistema, se utilizaron series de datos meteorológicos históricos proporcionados por el INMET, se trataron sus fallos y se imputaron los valores perdidos con el algoritmo KNNImputer. La asertividad de la imputación de los valores perdidos se verificó a través de las métricas de Exactitud, Precisión, Recall, F1-score y Error Cuadrático Medio (QMS). Estas métricas se derivan de la comparación de los valores predichos y los valores originales mediante la matriz de confusión. El sistema fue eficiente en la identificación de valores atípicos y en la imputación de valores perdidos, identificando el 100% de los valores discrepantes de las variables analizadas.

Palabras clave: Gestión de datos. Minería de datos. Informática aplicada.

Submetido em: 20/08/2021
Aceito para publicação em: 25/03/2022
Publicado em: 16/04/2022

1. INTRODUÇÃO

O Estado de Mato Grosso é rico em biodiversidade abrangendo 3 biomas (Amazônia, Cerrado e Pantanal), cujo clima não é homogêneo, requerendo grandes quantidades de estações de monitoramento meteorológico. O Estado possui apenas 39 estações administradas pelo INMET, com deficiência ou inexistência de monitoramento meteorológico de superfície em algumas regiões.

Quanto aos dados disponibilizados pelo INMET, algumas estações e períodos apresentam inconsistências e falhas. Requerendo assim, metodologias e ferramentas que facilitem o tratamento de dados brutos, favorecendo a geração de informações e conhecimentos, sendo que, para o processo de tomada de decisões, o conhecimento está no topo da hierarquia, tendo como base os dados que serão adequados, rearranjados e transformados em informações (GUIMARÃES; BEZERRA, 2019). Contudo, as decisões em um empreendimento estão suscetíveis a erros, haja vista que, caso os dados não forem cuidadosamente tratados a decisão tomada pode ser catastrófica.

A escassez e a confiabilidade questionável dos dados contribuem para decisões pouco assertivas (TARAPANOFF, 2006). Portanto, a informação de qualidade deve estar disponível no momento certo e no lugar correto para quem queira fazer uso (CHAFFEY; WHITE, 2011).

Quando há muitos dados a serem analisados ou a base de dados apresenta problemas e erros no processo de coleta, a dificuldade de mineração aumenta. Por esse motivo, séries históricas de dados meteorológicos devidamente preparadas para análise, são escassas. Quanto maior for a série histórica, maior será o poder computacional requerido para a avaliação e processamento (BILALLI et al., 2018). Dificultando a geração de informações e conhecimentos para usuários que não domina técnicas de manipulação dos dados e que não possua uma infraestrutura computacional adequada.

Softwares específicos para análise de dados meteorológicos também são escassos, sendo assim necessário o desenvolvimento de sistemas aplicados à preparação de dados desta natureza para análises exploratórias explícitas e implícitas levando-se em conta a disponibilidade e acessibilidade para realizar estes procedimentos.

Portanto, neste trabalho foi desenvolvido um sistema para pré-processamento de dados meteorológicos com base em Data Wrangling, denominado PAP Meteor (Preparação, Análise e Previsão de dados meteorológicos), cuja finalidade é o tratamento de dados para

aplicação na agropecuária e áreas afins, e facilitar a localização e manipulação destes dados em bases públicas. Deste modo, tem como objetivo demostrar métodos e funcionalidades do PAP Meteor, sistema desenvolvido neste trabalho, para preparação de dados meteorológicos visando geração de conhecimento e apoio a decisões estratégicas, além de descrever e corrigir os erros e falhas da base de dados do INMET da estação de superfície no município de Matupá MT e de Sinop MT.

2. METODOLOGIA

2.1. Ferramentas de desenvolvimento e validação do sistema

O sistema PAP Meteor foi desenvolvido utilizando a linguagem de programação Python, com o Framework Streamlit 0.58.2 e as bibliotecas Pandas, Numpy, Scipy, Scikit-learn e Plotly. Apresentando uma "interface" simples para interação com o usuário, contendo orientações de como deve estar configurada a base de dados. Possui também um formulário para a entrada e processamento dos dados.

Para testar o desempenho do sistema desenvolvido, precisamente os módulos incumbidos de identificar, corrigir erros e imputar informações ausentes, foram utilizados dados de estação meteorológica de superfície, sendo uma estação convencional (Matupá MT) e uma automática (Sinop MT).

Dos dados adquiridos, foram selecionadas as variáveis, umidade relativa (mínima, média e máxima), precipitação e temperatura (mínima e máxima). Para a variável de umidade relativa, foi utilizado apenas os valores de umidade relativa média. Os dados oriundos de estação convencional são diários, com dois registros no dia, sendo das 00:00 às 12:00 e das 12:00 às 18:00. Já dados de estações automáticas são horários das 00:00 às 23:00, no qual, foram enviados em sua forma bruta no formato .CSV, contendo erros e ausência de registros.

Para preparar os dados, o sistema importa um conjunto de dados com extensão .CSV[1] com encode UTF-8[2], com separador definido pelo usuário. Tal arquivo não deve haver

[1] Extensão .CSV: são valores separados por vírgula que podem ser criados ou editados por editores de planilhas eletrônicas.

[2] Encode: é um tipo de codificação binária de comprimento variável, que pode representar qualquer caractere universal (ex: ç,~,´, etc.).

cabeçalho, apenas os rótulos das colunas, cujos separadores decimais devem ser representados com ponto (.). Após importados, são transformados em um quadro de dados (Data Frame), que é como uma matriz, porém, suas colunas são nomeadas e suportam diferentes classes de dados, facilitando a identificação das variáveis no processo de codificação do software.

São utilizadas várias classes para preparação dos dados. Primeiro o sistema percorre a base e retorna as informações básicas, tais como, quantidade de linhas e nomes das colunas, além do tipo e quantidade de dados não nulos. Na segunda etapa, o sistema faz um resumo estatístico dos dados brutos apresentando a contagem de registro por variáveis meteorológicas, a média aritmética de cada variável, o desvio padrão, os mínimos e máximos registros e os quartis (25%, 50% e 75%) dos dados. A terceira etapa corresponde em identificar valores presentes e ausentes por colunas, retornando em uma tabela, o nome das variáveis e a somatória desses registros. Por fim, é necessário identificar possíveis erros nos dados, sendo indispensável a identificação de ruídos nos dados (outliers).

Para identificação de outlier, o sistema inspeciona os dados identificando pontos atípicos. Para tal, possui barras deslizantes (Figura 1), no qual o usuário configura os padrões climáticos para sua região. Vale lembrar que, o resumo estatístico fornecido nos módulos anteriores, servem de base para configuração dos parâmetros considerados erros de dados.

Com exceção das variáveis de temperaturas, considera-se dados errados, aqueles que estiverem abaixo de zero. Deste modo o programa, apresenta os erros em forma de tabelas. No qual, os registros considerados errados são excluídos e identificados como Nan (valores ausentes ou nulos, acrônimo em inglês para Not a Number) no Dataframe, que posteriormente serão preenchidos juntos com os registros faltantes.

Conforme a Figura 1 o usuário filtra temperatura mínima abaixo de 10 °C e acima de 28 °C. Tal função retorna uma tabela com as datas e as demais variáveis nos parâmetros especificados e a contagem da quantidade dos registros na condição configurada.

Utilizar as barras de configurações do sistema, requer do usuário a compreensão das condições climáticas da região. O fato do usuário não possuir estas informações, pode ser um agente contribuidor para a persistência de outliers. Por isso, o sistema faz uma varredura na base de dados e analisa as validações básicas propostas na Tabela 1.

Figura 1 - Sistema de filtros do PAP Meteor.

Fonte: Elaborado pelos autores (2021).

A Tabela 1 possui caráter generalista, ou seja, contempla as condições ambientais de todo território brasileiro. No entanto, para o município de Matupá-MT e Sinop MT, foram estabelecidos os critérios descritos na Tabela 2 para caracterização de dados errados ou suspeitos, embasados nos dados fornecidos pelo INMET (2020) para esta região. Lembrando que o usuário pode especificar estes critérios no sistema.

Tabela 1 - Parâmetros de identificação básica de erros em dados meteorológicos.

Tipo de verificação	Parâmetro de validação de dados
Validação de lógica	Temp. Mínima < Temp. Máxima
Validação de limites	Temperatura -8 °C a 45 °C
	Precipitação ≧ 0 mm < 500 mm
	Umidade Relativa >0% e <100%

Fonte: Elaborado pelos autores (2021).

Tabela 2 - Condições para os dados brutos da estação do INMET em Matupá MT e Sinop MT a serem considerados errados ou suspeitos

Variável	Parâmetro	Unidade de Medida
Temperatura Mínima (Matupá)	< 05 ou > 28	°C
Temperatura Máxima	< 15 ou > 45	°C
Precipitação	< 0 ou > 220	mm/dia
Umidade Relativa	<0 ou >100	%

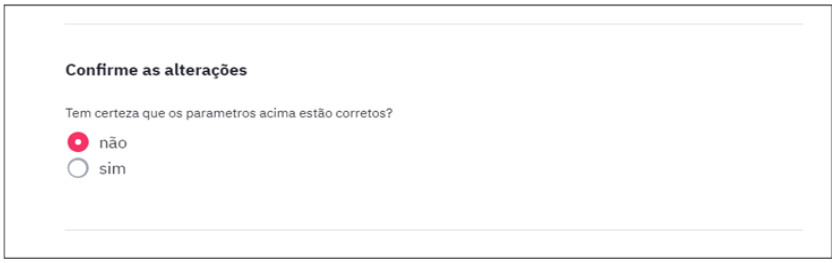
Fonte: Elaborado pelos autores (2021).

Para Sinop, devido à série histórica ser horária, somente foram atribuídos os limites mínimos e máximos para precipitação e umidade relativa.

Com a identificação de erros e dados ausentes devidamente contabilizados a ação seguinte do usuário é a correção dos dados. Para tal o usuário deverá confirmar se os parâmetros para identificação de ruído estão devidamente configurados, através de dois

botões (sim ou não). Caso o usuário esteja consciente dos parâmetros, sua ação é marcar o botão sim (Figura 2). Desse modo os valores considerados anormais serão automaticamente substituídos por nulos (Nan), que posteriormente será preenchido com algoritmo imputador. Por fim resta preencher os dados ausentes.

Figura 2 - Confirmação de parâmetros para identificar dados anormais. Se pressionado o botão sim, os dados na condição configurada será considera nulo (Nan).



Fonte: Elaborado pelos autores (2021).

Para preenchimento dos dados ausentes foram utilizados o método K-Vizinhos Mais Próximos (KNNImputer). Este algoritmo, por padrão utiliza a métrica de distância euclidiana, suportando registros ausentes e utilizada para encontrar vizinhos mais próximo (TROYANSKAYA et al., 2001). Cada registro ausente é imputado utilizando o valor do vizinho mais próximo. O valor atribuído ao vizinho ausente é mediado uniformemente ou ponderado pela distância de cada vizinho. A configuração utilizada foi:

- N_neighbors (Número de amostras vizinhas a serem usadas para imputação) = 30;
- Peso (Função de peso usada na previsão) = Distância (pontos de peso pelo inverso de sua distância. Neste caso, vizinhos mais próximos de um ponto de consulta terão uma influência maior do que vizinhos mais distantes);
- Metric (Métrica de distância para pesquisar vizinhos) = 'nan_euclidean' (Distância euclidiana dos valores ausentes).

Para certificar a validade e eficiência do método KNNImputer na correção de dados meteorológicos, é necessário a comparação de registros tratados (ausentes imputados) com bases consistentes e originais. No entanto, devido à dificuldade de encontrar em base de dados pública, série histórica meteorológica devidamente consistente e confiável, fez-se necessário, tratar os dados brutos, identificar, corrigir as outliers e imputar os valores ausentes. Haja vista que os mecanismos de coleta de dados são passíveis de erros, além de falharem na coleta de alguns dados, gerando lacunas que podem afetar a geração de

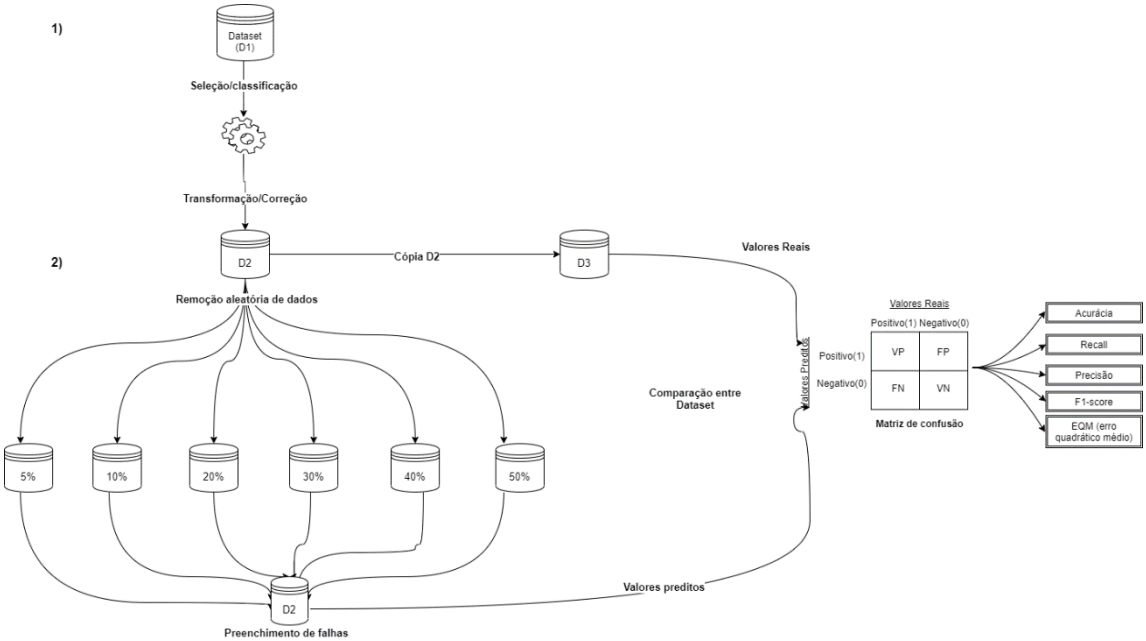
conhecimento. Deste modo, a partir da série corrigida, foi adotado a metodologia descrita na Figura 3. Conforme mostrado na Figura 3:

1) A partir de uma série histórica com dados brutos (D1), os dados foram tratados, corrigidos os erros e preenchido os valores ausentes com o método KNNImputer, formando o Dataset D2;

2) O conjunto D2 foi duplicado, formando o dataset D3 (controle). O dataset D2 original foi removido aleatoriamente por etapas de validação, 5, 10, 20, 30, 40 e 50% dos dados. A cada remoção aleatória foi feito uma validação com matriz de confusão comparando com Dataset D3. O modelo foi validado com as métricas de Acurácia, Recall, Precisão, F1-score e EQM (Erro Quadrático Médio).

Além da validação através da remoção aleatória de porcentuais de dados, também se verificou o quanto cada ano adicionado na série história afeta o modelo KNNImputer. Foi então, imputado dados em série histórica com 1 a 6 anos de registros e analisados através de matriz de confusão a acurácia, precisão e o erro quadrático médio em cada situação.

Figura 3 - Metodologia aplicada para avaliação do desempenho do algoritmo imputador KNNImputer.



Fonte: Elaborado pelos autores (2021).

A Acurácia corresponde ao cálculo da precisão, da fração ou a contagem das previsões

corretas. Onde $\hat{y}_i$ é o valor previsto do i a amostra e y_i é o valor verdadeiro correspondente, deste modo a fração das previsões corretas sobre é definido como:

$$(1)\ Acuracia(y, \hat{y}) = \frac{1}{n_{\text{amostras}}} \sum_{i=0}^{n_{\text{amostras}}-1} 1(\hat{y}_i = y_i)$$

O Recall avalia a proporção entre acertos e o total de segmentos avaliados. Esta métrica indica o quão bom o modelo foi para a identificação dos pontos corretos. Onde os valores vp (verdadeiro positivo) são divididos pelos valores $vp + fn$ (falso negativo).

$$(2)\text{Recall} = \frac{vp}{vp + fn},$$

A Precisão corresponde à capacidade de evitar falsos positivos (fp), cuja fórmula consiste na divisão dos verdadeiros positivos (vp)pela soma de verdadeiros (vp) positivos e falsos positivos (fn).

$$(3)\text{Precisão} = \frac{vp}{vp + fp},$$

O F1_score é a média ponderada da precisão e do recall, tal métrica define a qualidade geral do modelo.

$$(4)\ F_\beta = (1 + \beta^2)\frac{\text{precisão} \times \text{recall}}{\beta^2 \text{precisão} + \text{recall}}$$

Por fim, o erro quadrático médio (EQM) tem por função comparar estimadores, de modo que o estimador mais eficaz é aquele com menor variância.

$$(5)EQM(y, \hat{y}) = \frac{1}{n_{\text{amostra}}} \sum_{i=0}^{n_{\text{amostra}}-1} (y_i - \hat{y}_i)^2$$

A partir das métricas da matriz de confusão foi possível avaliar a qualidade da imputação do algoritmo K-vizinhos mais próximos (KNNImputer)

3. RESULTADOS E DISCUSSÕES

Os resultados serão apresentados de acordo com a sequência dos módulos exibidos pelo sistema. Sendo o primeiro módulo responsável por exibir as informações básicas dos dados. O segundo módulo verifica e elimina outliers e por fim no terceiro módulo imputa registros ausentes.

3.1. Identificações básicas da base de dados

Os dados utilizados do Município de Matupá MT possuem seis (06) colunas com 25.567 registros no intervalo de 01/01/1987 a 31/12/2020. No município de Sinop MT os dados apresentam oito (8) colunas com 105.187 registros no intervalo de 01/01/2009 a 31/01/2020. Ambos datasets apresentam em suas colunas, dados de temperatura, umidade relativa e precipitação.

Os registros válidos (não nulos) somam 54,47% dos dados para Matupá MT (estação convencional) e 77,2% para Sinop MT (estação automática). Para a estação convencional, a quantidade de registros válidos são semelhantes, exceto para umidade relativa média com maior quantidade de dados registrados. A menor consistência foi atribuída a variável precipitação com 41,52% dos lançamentos, com uma média anual de 322 inscrições por ano (considerando 2 registros diários). O mesmo ocorre na estação automática de Sinop MT, cujos registros de precipitação também sofreram maior quantidade de falhas em relação as outras variáveis, totalizando 28% de dados faltantes.

A razão para que haja menor quantidade de dados válidos para precipitação, pode ser explicada devido à estrutura dos instrumentos. A precipitação em estações convencionais, geralmente é medida com Pluviômetro, que está sujeito a ação do vento, topografia, ser obstruído por sujidades (folhas, objetos, etc.), além de problemas inerentes a erros humanos, por exemplo a danificação acidental de dados e equipamentos (WMO, 2008) e (SEIBERT; MORÉN, 1999). As estações automáticas também não estão imunes a problemas nos registros, haja vista, que estão sujeitas a danos físicos, afetando a qualidade dos registros ou até sua interrupção (STRASSBURGER, et al., 2011).

Em Matupá MT houve grandes oscilações nos registros no período de 1987 a 2000 para todas as variáveis, com ênfase para Precipitação e Umidade Relativa (Figura 4A). Já entre 2000 a 2020 houve uma padronização na consistência dos registros. Tais fatos mencionados sugerem, algumas hipóteses:

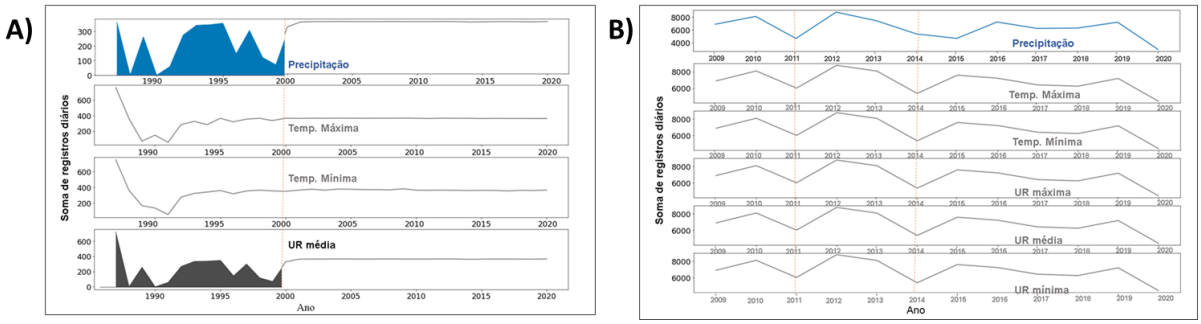
a) No período de 1990 a 2000, houve problemas sérios nos equipamentos, ou ficou sem operador;

b) Os instrumentos no período de 2000 em diante foram substituídos por equipamentos mais confiáveis e assertivos;

c) Houve incremento no quadro de pessoal a partir de 2003, ou teve programas de

calibração de instrumentos e capacitação de operadores.

Figura 4 – A) Quantidade média de registros válidos (não nulos) no período de 1985 a 2020 em Matupá MT. B) Quantidade média de registros válidos (não nulos) no período de 2009 a 2020 em Sinop MT.



Fonte: Elaborado pelos autores (2021).

No ano de 1994 através do Ofício 269/INMET de 11/11/94, encaminhado ao ministro da agricultura, o relatório de atividades do INMET, apresentando em caráter de urgência a necessidade de recompor o quadro de pessoal do Instituto, apontando para vários distritos do INMET com problemas financeiros e necessidades de treinamento operacional e calibração de instrumentos (INMET, 2000).

No relatório anual da (9°) DISME (Cuiabá) do INMET no ano de 2000, relatou-se sobre o treinamento no programa “Qualidade 2000”, incluindo calibração instrumental. Ainda neste período, houve várias reformas e substituições de equipamentos, reparos em abrigos e melhorias no sistema de transmissão de dados em estações no Mato Grosso. Em 2001 houve a recuperação da base física das estações de Gleba Celeste, Pe. Ricardo Remetter, Matupá, Merure e Cáceres.

Deste modo, de 2000 a 2019 houve estabilização e melhora nos registros coletados, principalmente em estações automáticas, como exemplo de Sinop com poucas oscilações abruptas de falhas de registros (Figura 4B). Contudo, ainda há sérios problemas a serem sanados, principalmente relacionado a mão de obra, manutenção de instalações, equipamentos e a falta de reposição de peças INMET (2017).

Devido a estes problemas aludidos, nos dados observados neste trabalho, ocorreram discrepâncias e registros duvidosos. A partir do resumo estatístico dos dados brutos foi possível evidenciar erros nas variáveis temperatura máxima e temperatura mínima. Nos registros de temperatura máxima, o mínimo e máximo valor registrado é de 9,60 °C a 40,20

°C em Matupá e de 10 °C a 40,00 °C em Sinop. No entanto, dadas as características climatológicas e de relevo das regiões, a probabilidade de temperatura máxima inferior a 10 °C é baixíssima. Não obstante, a temperatura mínima também apresenta erros, com registros mínimos de 1,40 °C na estação convencional e 4,60 °C na estação automática.

De acordo com Matupá (2020), houve registros mínimos de 4 °C no município de Matupá MT. No entanto, segundo a base de dados do INMET, houve 5 (cinco) registros inferiores a 6 °C. Contudo, tais registros são equivocados, já que, datam 13 e 27 de janeiro e 01 de fevereiro (registro de 1,4 ºC, 1,7 °C e 5,4 °C), época extremamente úmida e quente na região. Neste caso, o município está localizado na região norte de Mato Grosso, fazendo limite com a nordeste, considerada a mais quente do estado, e muito distante da região mais fria que é a sudeste (RAMOS et al., 2017). Falha em temperatura também ocorreu nos dados de Sinop, registrando 4,6 °C as 12:00 do dia 07/02/2020, no que todos os outros registros estavam em média de 24,0° C. Tais falhas podem ocorrer por vários motivos, dentre eles, falhas nos sensores, na transmissão de dados e até problemas de calibração de instrumentos (BABA; VAZ; COSTA, 2014).

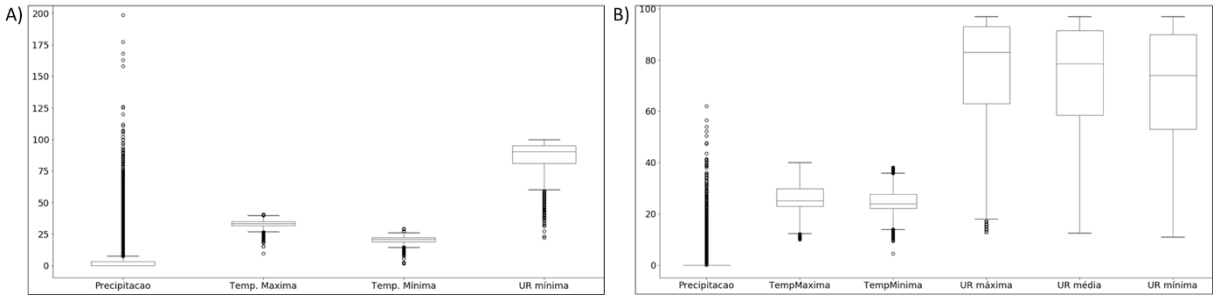
As verificações lógicas e de limites propostos na Tabela 1, resultaram com 100% dos dados brutos dentro dos parâmetros estabelecidos para ambas as estações meteorológicas. Isso demonstra que embora haja vários registros ausentes, a integridade dos dados possui boa qualidade.

Os erros dos dados brutos foram pequenos. Na estação de Matupá a variável Temperatura Mínima houve 7 registros menores que 6 °C e maior que 28 °C, sendo 5 registros inferiores a 6 °C e 2 registros superiores a 28 °C. Para a temperatura máxima, foi identificado apenas um registro com temperatura inferior a 15 °C. Tal registro ocorreu dia 15/04/2002, época com bastante chuva e calor predominante na região.

Os erros no conjunto de dados também podem ser visualizados através de gráfico do tipo Boxplot (Figura 5). Há uma grande dispersão nos dados de Precipitação em Matupá (esperado para este tipo de variável), fato que também pode ser observado no resumo estatístico disponibilizado pelo sistema, com desvio padrão 12,84 mm ao dia em Matupá. No entanto, em Sinop a mesma variável apresenta desvio padrão de 1,56 mm ao dia (Figura 5B). A diferença entre os dois datasets é justificada devido à organização dos dados (diário e horário), de modo que, em estações com dados horários há menor amplitude entre os

registros, afetando as médias diárias das variáveis (ENSOR; ROBESON, 2008). Já para as variáveis de temperatura e umidade relativa é bem visível as outliers com valores incomuns bem nítidos.

Figura 5: A) Identificação de outliers através de boxplot na estação de Matupá MT e B) em Sinop MT



Fonte: Elaborado pelos autores (2021).

3.2. Imputação de dados ausentes.

A série histórica de Matupá MT, possui 45,53% dos ausentes, com destaque para variável Precipitação, sendo seus períodos mais críticos de coleta nos anos 1998 a 2000. De modo que, anualmente em média 226 dias deixavam de serem registrados, o equivalente 8 dias no mês. Para a temperatura, a média de ausentes por ano foram de 210 dias, ou seja, em média a perda de 7 dias/mês. Já para a variável umidade relativa foram perdidos em média apenas 2 registros por mês. Quanto as falhas nos registros, na estação de Sinop, a variável precipitação também apresenta o maior número de falhas.

Foi imputado na base de dados de Matupá MT 54.755 e 487.305 registros na base de Sinop MT através do método k-vizinhos mais próximos (KNNImputer). O método apresentou melhores resultados em conjunto de dados com 10% ausentes (Tabela 3) e série histórica com 3 anos de registros para estação automática com série horária. No entanto, para estações com série histórica diária, a acurácia e precisão diminuem à medida que o número de falhas aumenta, consequentemente o erro quadrático médio (QMS) também aumenta (Tabela 4).

Em estação automática com série histórica horária, à medida que aumenta a quantidade de falhas, o KNNImputer diminui sua acurácia e precisão na proporção de 10:15 e 10:7, ou seja, a cada 10% de falhas adicionadas diminui em média 15% da acurácia e 7% da precisão (Figura 6). Já para estação convencional com série histórica diária, a proporção de

acurácia é de 10:7 e 10:4 na métrica de precisão.

Tabela 3: Pontuação por variáveis nas previsões imputadas em dados ausentes pelo algoritmo KNNImputer em conjunto de dados com 5, 10, 20, 30, 40 e 50% de valores ausentes (Nan) em série histórica com dados horários.

Métrica	Percentual	UR. Máxima	UR. Mínima	Temp. Máxima	Temp. Mínima	Precipitação
Acurácia	5%	0,53	0,52	0,57	0,56	0,98
Acurácia	10%	0,92	0,91	0,93	0,92	1,00
Acurácia	20%	0,82	0,82	0,83	0,83	0,99
Acurácia	30%	0,71	0,72	0,74	0,74	0,99
Acurácia	40%	0,62	0,63	0,66	0,65	0,98
Acurácia	50%	0,53	0,52	0,57	0,56	0,98
F1 score	5%	0,58	0,57	0,58	0,63	0,63
F1 score	10%	0,90	0,91	0,92	0,93	0,95
F1 score	20%	0,84	0,83	0,86	0,86	0,82
F1 score	30%	0,76	0,74	0,80	0,75	0,72
F1 score	40%	0,66	0,67	0,70	0,71	0,68
F1 score	50%	0,58	0,55	0,63	0,64	0,56
Precisão	5%	0,71	0,69	0,83	0,82	0,94
Precisão	10%	0,92	0,92	0,93	0,96	1,00
Precisão	20%	0,86	0,85	0,92	0,91	0,97
Precisão	30%	0,84	0,81	0,89	0,88	0,97
Precisão	40%	0,77	0,77	0,87	0,86	0,95
Precisão	50%	0,72	0,68	0,83	0,83	0,95
QMS	5%	9,94	11,09	2,61	2,75	0,84
QMS	10%	2,60	3,08	1,40	0,93	0,37
QMS	20%	6,07	7,12	1,70	1,82	0,64
QMS	30%	8,38	9,11	2,12	2,28	0,69
QMS	40%	9,12	9,91	2,39	2,51	0,79
QMS	50%	9,94	10,92	2,62	2,76	0,85
Recall	5%	0,53	0,52	0,49	0,55	0,56
Recall	10%	0,90	0,91	0,93	0,91	0,92
Recall	20%	0,82	0,82	0,82	0,83	0,78
Recall	30%	0,71	0,71	0,74	0,69	0,64
Recall	40%	0,61	0,63	0,62	0,63	0,61
Recall	50%	0,53	0,51	0,53	0,56	0,49

Fonte: Elaborado pelos autores (2021).

Para precipitação, a acurácia foi de 99,99% em série histórica de 3 anos com dados horários e 10% de falhas. A acurácia diminui em média 1% à medida que adiciona 10% de falhas (Figura 6A) e o erro quadrático médio também aumenta. Outro sim, agrupamentos de 3 anos com 10% de falhas apresentam melhores resultados de precisão, sendo que, em séries

com 5 anos, a precisão diminui 3% (Figura 6B). Já a acurácia não apresenta uma diminuição significativa, porém, o QMS aumenta gradualmente.

O preenchimento de falhas para precipitação ainda é considerado um problema de difícil solução. Em 2005, nos trabalhos de Chibana et al., (2005), não haviam métodos para imputar com eficiência os dados diários e horários de precipitação, no qual recomendou-se que os preenchimentos ocorressem em dados mensais ou anuais. Desde então, foram aplicadas e desenvolvidas várias metodologias para o preenchimento de registros ausentes, no entanto, a maioria dos trabalhos limita-se a séries com dados mensais e anuais.

Dentre os métodos de preenchimento o que apresenta maior destaque é a ponderação regional, com resultados satisfatórios para dados com poucas lacunas e para dados mensais, (NASCIMENTO et al., 2010), (SOARES; SILVA, 2017) e (DIAZ; PEREIRA; NOBREGA, 2018). Porém, o preenchimento de falhas em dados diários e horários ainda é pouco discutido na literatura.

Métodos de regressão linear múltipla (RLM) e ponderação regional (PR) tiveram bons desempenhos nos trabalhos de Bier; Ferraz (2017), Ventura et al.(2016), na imputação de dados de temperatura. A utilização de redes neurais artificiais resultou em 97% a 99% de precisão no preenchimento de falhas em dados mensais de temperatura e umidade relativa nos trabalhos de Coutinho et al.(2018). No entanto, os mesmos resultados não se aplicam para falhas de precipitação, com a pesquisa de Brubacher, Oliveira e Guasselli (2020) relatando vantagens das RLMs em relação as RNAs possivelmente em função da forte correlação linear entre os dados de precipitação de cada local em relação a sua vizinhança.

As RNAs também foram utilizadas nos trabalhos de Depiné et.al (2014) para dados horários de precipitação, porém, teve dificuldade de reproduzir chuvas de verão, com as RNAs variando conforme as condições locais de cada região (tamanho da bacia, distribuição espacial de chuva na bacia), qualidade dos dados de entrada e da escolha e da configuração da rede a ser utilizada.

Com o PAP Meteor foi possível uma imputação satisfatória para dados diários e horários de temperatura, umidade e precipitação (inclusive em meses secos), fato que dificilmente é possível imputar falhas com qualidade usando regressão linear múltipla (COSTA et al., 2012). No entanto, nos meses secos em série diária com dois registros, a imputação não é satisfatória em conjuntos de dados com falhas superiores a 30% (Figura 7). Neste caso, recomenda-se transformar a série para apenas um registro ao dia.

Tabela 4: Pontuação por variáveis nas previsões imputadas em serie com 10% de registros ausentes de 1 a 6 anos de dados em série histórica com dados horários.

Métrica	Ano	UR. Máxima	UR. Mínima	Temp. Máxima	Temp. Mínima	Precipitação
Acurácia	1	0,94	0,93	0,95	0,94	1,00
Acurácia	2	0,93	0,93	0,94	0,94	1,00
Acurácia	3	0,94	0,94	0,95	0,95	1,00
Acurácia	4	0,93	0,93	0,94	0,94	1,00
Acurácia	5	0,93	0,93	0,94	0,94	1,00
Acurácia	6	0,93	0,94	0,94	0,94	1,00
F1 score	1	0,91	0,92	0,96	0,94	0,90
F1 score	2	0,92	0,93	0,94	0,93	0,94
F1 score	3	0,92	0,93	0,93	0,93	0,93
F1 score	4	0,93	0,93	0,94	0,94	0,95
F1 score	5	0,92	0,92	0,94	0,94	0,93
F1 score	6	0,93	0,93	0,94	0,95	0,92
Precisão	1	0,93	0,93	0,97	0,96	0,99
Precisão	2	0,93	0,93	0,95	0,96	1,00
Precisão	3	0,93	0,93	0,97	0,96	1,00
Precisão	4	0,93	0,93	0,96	0,95	0,98
Precisão	5	0,93	0,93	0,96	0,96	0,97
Precisão	6	0,94	0,93	0,96	0,96	0,98
QMS	1	1,68	2,45	0,55	0,61	0,36
QMS	2	3,13	2,04	0,68	0,68	0,32
QMS	3	3,14	2,60	0,63	0,63	0,32
QMS	4	3,30	2,91	0,73	0,77	0,39
QMS	5	3,15	3,00	0,63	0,83	0,49
QMS	6	3,12	3,05	0,60	0,63	0,46
Recall	1	0,91	0,92	0,95	0,92	0,89
Recall	2	0,92	0,92	0,92	0,93	0,91
Recall	3	0,92	0,92	0,91	0,93	0,90
Recall	4	0,92	0,93	0,92	0,93	0,93
Recall	5	0,92	0,92	0,92	0,93	0,91
Recall	6	0,92	0,93	0,93	0,93	0,90

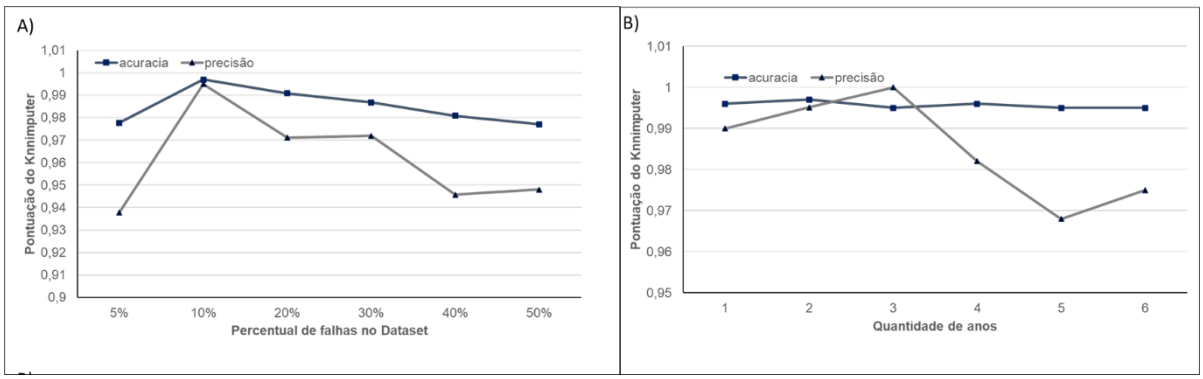
Fonte: Elaborado pelos autores (2021).

Para as variáveis de temperatura (máxima e mínima), o KNNImputer também foi bastante preciso, tanto em série com dados diários, como para dados horários (Figura 8). Na série diária (Matupá MT) a acurácia diminui em média 7% à medida que aumenta 10% de falhas, consequentemente o QMS também aumenta. O mesmo ocorre para a precisão, quando se adiciona 10% de falhas a precisão diminui em média 4,7%. Em série horária (Sinop

MT), as melhores acurácias ficaram em conjuntos com 10% de falhas (Figura 8 C e D).

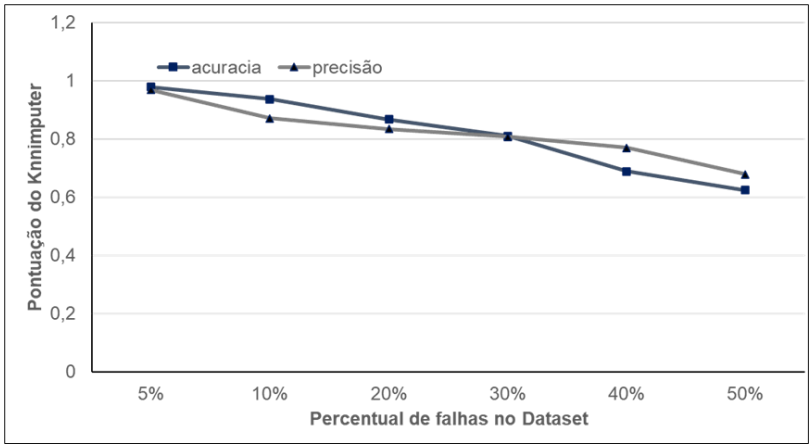
As variáveis de temperatura e umidade relativa também são preenchidas satisfatoriamente com ponderação regional e regressão linear múltipla (BIER; FERRAZ, 2017) e (YAGUCHI et al., 2016). No entanto, não é adequado para preenchimento de lacunas em estações no Mato Grosso, haja vista que, a distância entre estações é relativamente longa e com grande variabilidade de relevo. Tais fatores implicam em predições errôneas, considerando que tais variações no relevo podem ocorrer microclimas e ventos acentuados, afetando o desempenho das estimativas. (FENSTERSEIFER, 2013).

Figura 61: Pontuação do preenchimento de falhas do KNNImputer em dados de Precipitação da estação automática de Sinop MT de 2009 a 2011. A) Série com 5 a 50% de falhas em dados horários. B) Séries com 10% de falhas de 1 a 6 anos de registros.



Fonte: Elaborado pelos autores (2021).

Figura 7: Pontuação do preenchimento de falhas do KNNImputer em dados de Precipitação da estação convencional de Matupá MT (série com dados diários) de 2005 a 2008.

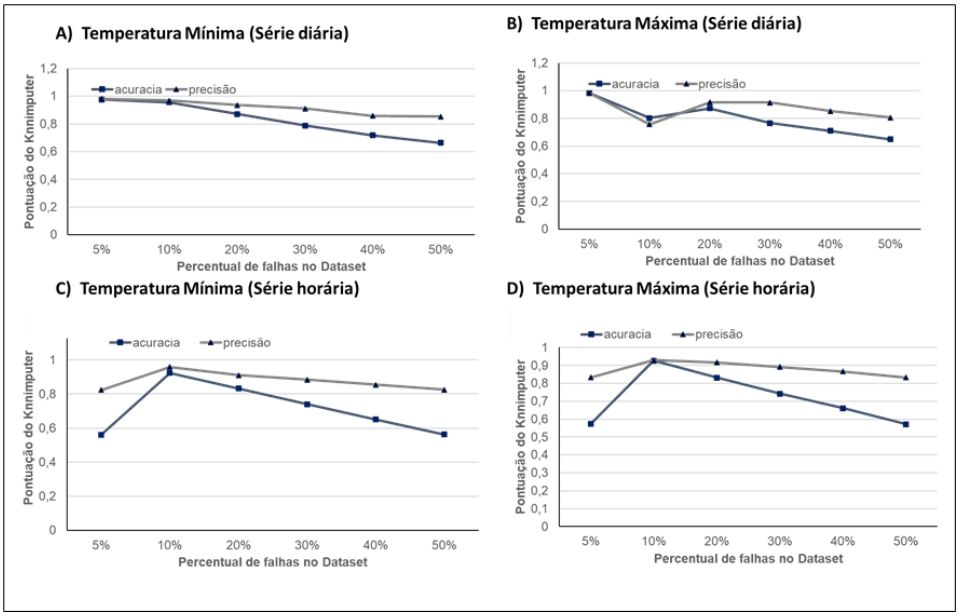


Fonte: Elaborado pelos autores (2021).

Todos os métodos citados para preenchimento de falhas, consomem elevado custo

computacional. Desde modo, limita-se este tipo de procedimento a usuários específicos, porém, o PAP Meteor está hospedado em um servidor na nuvem (Google Cloud Plataform) incumbido de todo processamento, tornando este procedimento acessível para qualquer usuário. Silva et al.(2015) utilizou recursos de computação paralela em placas gráficas de propósito geral (GPGPU) para melhorar a eficiência no processamento de dados desta natureza. Como o PAP Meteor utiliza uma plataforma Web, para melhorar a capacidade de processamento é necessário em futuras versões implementar processamento distribuído, dispensando o usuário da aquisição de hardware para processar os dados.

Figura 8: Pontuação do preenchimento de falhas do KNNImputer em dados de Temperatura mínima e máxima de Matupá MT (A e B) e Sinop MT de 2005 a 2008(C e D).



Fonte: Elaborado pelos autores (2021).

4. CONSIDERAÇÕES FINAIS

O PAP Meteor demonstrou ser eficiente em preencher falhas em dados meteorológicos em séries históricas diárias e horárias. O sistema tem maior precisão em série histórica com dados horários com 10% de falha nos registros e com duração de 3 (três) anos.

A acurácia do modelo diminui 16% à medida que é adicionado 10% de falhas em dados horários e 7% em dados diários. Já a precisão diminui em média 7% em série histórica com dados horários e 4,3% em dados diários. O QMS aumenta 20% a cada 10% de falhas adicionada em série diária e horária.

O PAP Meteor é uma alternativa acessível para correção e imputação de registros ausentes em séries com dados diários e horários, considerando que está disponível em uma plataforma web.

REFERÊNCIAS

BABA, Ricardo Kazuo; VAZ, Maria Salete Marcon Gomes; COSTA, Jéssica da. Correção de dados agrometeorológicos utilizando métodos estatísticos. **Revista Brasileira de Meteorologia**, [S.L.], v. 29, n. 4, p. 515-526, dez. 2014. DOI: http://dx.doi.org/10.1590/0102-778620130611.

BIER, Anderson Augusto; FERRAZ, Simone Erotildes Teleginski. Comparação de Metodologias de Preenchimento de Falhas em Dados Meteorológicos para Estações no Sul do Brasil. **Revista Brasileira de Meteorologia**, v. 32, n. 2, p. 215–226, 2017. DOI 10.1590/0102-77863220008.

BILALLI, Besim *et al*. Intelligent assistance for data pre-processing. **Computer Standards & Interfaces**, [S.L.], v. 57, p. 101-109, mar. 2018. DOI: http://dx.doi.org/10.1016/j.csi.2017.05.004.

BRUBACHER, João Paulo; OLIVEIRA, Guilherme Garcia de; GUASSELLI, Laurindo Antonio. Preenchimento de Falhas em Séries Temporais de Precipitação Diária no Rio Grande do Sul. **Revista Brasileira de Meteorologia**, [S.L.], v. 35, n. 2, p. 335-344, jun. 2020. DOI: http://dx.doi.org/10.1590/0102-7786352035.

CHAFFEY, Dave; WHITE, Gareth. **Business Information Management: Improving Performance Using Information Systems**. 2. Ed. [s.l.]:Financial Times/Prentice Hall, 2011. ISBN 1784483648, 9781784483647.620 p.

CHIBANA, Eduardo Yasuji *et al.* Preenchimento de Falha de Dados Climáticos. In: XIV Congresso Brasileiro de Agrometeorologia, 2005, Campinas - SP. In: CONGRESSO BRASILEIRO DE AGROMETEOROLOGIA., 2005. **Anais** [...]. Santa Maria - RS, 2005. 8 p. Disponível em: < http://www.sbiagro.org.br/pdf/v_congresso/Trabalho41.pdf> Acessado em: 12/04/2019.

COSTA, Rafaela Lisboa *et al*. Imputação Multivariada de Dados Diários de Precipitação e Análise de Índices de Extremos Climáticos (Imputation Multivariate of Precipitation Daily Data and Analysis of Climate Extremes Index). **Revista Brasileira de Geografia Física**, [S.L.], v. 5, n. 3, p. 661-675, 5 nov. 2012. DOI: http://dx.doi.org/10.26848/rbgf.v5i3.232861.

COUTINHO, Eluã Ramos *et al*. Application of Artificial Neural Networks (ANNs) in the Gap Filling of Meteorological Time Series. **Revista Brasileira de Meteorologia**, [S.L.], v. 33, n. 2, p. 317-328, jun. 2018. DOI: http://dx.doi.org/10.1590/0102-7786332013.

DEPINÉ, Haline *et al*. Preenchimento de Falhas de Dados Horários de Precipitação Utilizando Redes Neurais Artificiais. **Revista Brasileira de Recursos Hídricos**, [S.L.], v. 19, n. 1, p. 51-63, 2014. http://dx.doi.org/10.21168/rbrh.v19n1.p51-63

DIAZ, Caio César Farias; PEREIRA, João Antonio dos Santos; NOBREGA, Ranyere Silva. Comparação de dados estimados pelo método da ponderação regional (PR) e dados estimados pelo TRMM para o preenchimento de falhas de precipitação na bacia hidrográfica do Rio

Pajeú. **Revista Brasileira de Climatologia**, [S.L.], v. 22, p. 324-339, 7 maio 2018. Universidade Federal do Paraná. DOI: http://dx.doi.org/10.5380/abclima.v22i0.46911.

ENSOR, Leslie A.; ROBESON, Scott M. Statistical Characteristics of Daily Precipitation: comparisons of gridded and point datasets. **Journal Of Applied Meteorology And Climatology**, [S.L.], v. 47, n. 9, p. 2468-2476, 1 set. 2008. American Meteorological Society. DOI: http://dx.doi.org/10.1175/2008jamc1757.1.

FENSTERSEIFER, Cesar Augusto**. Qualidade das estimativas de precipitações derivadas de satélites na bacia do Alto Jacuí –RS**. 2013. Dissertação (Metrado em Engenharia Civil e Ambiental) – Universidade Federal de Santa Maria, UFSM RS. Santa Maria –RS, 2013. 126p.

GUIMARÃES, André José Ribeiro; BEZERRA, Cicero Aparecido. Gestão de dados: uma abordagem bibliométrica. **Perspectivas em Ciência da Informação**, v. 4, p. 171–186, 2019. DOI 10.1590/1981-5344/4192.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Cidades do Mato Grosso IBGE**. 2019. Disponível em: https://cidades.ibge.gov.br/brasil/mt/historico. Acesso em: 18 abr. 2020.

INSTITUTO NACIONAL DE METEOROLOGIA. **BDMEP Matupá MT INMET 2020**. Disponível em: https://bdmep.inmet.gov.br/. Acesso em: 20 jan. 2021.

INSTITUTO NACIONAL DE METEOROLOGIA. **Relatório do gestor INMET exercício de 2000**. Brasilia: [s.n.], 2000. Disponível em: http://www.inmet.gov.br/html/informacoes/relatorio_gestor/pdf/SEDE_REL_GESTOR_2000.pdf. Acesso em: 8 maio 2020.

INSTITUTO NACIONAL DE METEOROLOGIA. **Relatório de gestão INMET exercício 2017**. Brasília: [s.n.], 2017. Disponível em: http://www.inmet.gov.br/html/informacoes/relatorio_gestor/pdf/RelatorioGestao2017.pdf. Acesso em: 8 maio 2020.

MATUPÁ, Prefeitura. **Geografia de Matupá Mato Grosso**. set. 2020. Disponível em: https://www.matupa.mt.gov.br/Nossa-Cidade/Geografia/. Acesso em: 20 mai. 2020.

NASCIMENTO, Telma Santos do *et al.* Preenchimento de falhas em banco de dados pluviométricos com base em dados do CPC (CLIMATE PREDICTION CENTER): estudo de caso do rio solimões-amazonas. **Revista Brasileira de Climatologia**, [S.L.], v. 7, p. 143-158, 30 set. 2010. DOI: http://dx.doi.org/10.5380/abclima.v7i0.25643.

RAMOS, Henrique da Cruz *et al.* Precipitação e temperatura do ar para o estado de mato grosso utilizando krigagem ordinária. **Revista Brasileira de Climatologia**, v. 13, n. 0, p. 2237–8642, 2017. DOI: https://doi.org/10.5380/abclima.v20i0.43762