WALINGSON DA SILVA DA COSTA

SISTEMA WEB PARA PRÉ-PROCESSAMENTO E ANÁLISE DE DADOS METEOROLÓGICOS

Dissertação de Mestrado

ALTA FLORESTA-MT

2021

	WALINGSON DA SLVA DA COSTA	Diss. MESTRADO	PPGBioAgro 2021

UNIVERSIDADE DO ESTADO DE MATO GROSSO
FACULDADE DE CIÊNCIAS BIOLÓGICAS E AGRÁRIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIODIVERSIDADE E AGROECOSSISTEMAS AMAZÔNICOS

WALINGSON DA SILVA DA COSTA

SISTEMA WEB PARA PRÉ-PROCESSAMENTO E ANÁLISE DE DADOS METEOROLÓGICOS

Dissertação apresentada à Universidade do Estado de Mato Grosso, como parte das exigências do Programa de Pós-Graduação em Biodiversidade e Agroecossistemas Amazônicos, para a obtenção do título de Mestre em Biodiversidade e Agroecossistemas Amazônicos.Orientador: Prof. Dr. Rivanildo Dallacort

ALTA FLORESTA-MT

2021

Catalogação na publicação

Faculdade de Ciências Biológicas e Agrárias

Walter Clayton de Oliveira CRB 1/2049

C837s

COSTA, Walingson da Silva da Costa.

Sistema Web para Pré-Processamento e Análise de Dados Meteorológicos / Walingson da Silva da Costa Costa – Alta Floresta, 2021.

64 f.; 30 cm. (ilustrações) Il. color. (sim)

Trabalho de Conclusão de Curso (Dissertação/Mestrado) – Curso de Pós-graduação Stricto Sensu (Mestrado Acadêmico) Biodiversidade e Agroecossistemas Amazônicos, Faculdade de Ciências Biológicas e Agrárias, Câmpus de Alta Floresta, Universidade do Estado de Mato Grosso, 2021.

Orientador: Rivanildo Dallacort

1. Data Wrangling. 2. Mineração de Dados. 3. Computação Aplicada. 4. Imputação de Dados. 5. Amazônia. I. Walingson da Silva da Costa Costa. II. Sistema Web para Pré-Processamento e Análise de Dados Meteorológicos: .

CDU 556.5

SISTEMA WEB PARA PRÉ PROCESSAMENTO E ANÁLISE DE DADOS METEOROLÓGICOS

Walingson da Silva da Costa

Dissertação apresentada à Universidade do Estado de Mato Grosso, como parte das exigências do Programa de Pós-Graduação em Biodiversidade e Agroecossistemas Amazônicos, para a obtenção do título de Mestre em Biodiversidade e Agroecossistemas Amazônicos.

Aprovada em:

_____

Prof. Dr. Rivanildo Dallacort Orientador –

UNEMAT/ PPGBioAgro

_____

Prof. Dr. Marco Antonio Camillo De Carvalho

UNEMAT/ PPGBioAgro

_____

Prof. Dra. Silmara Bispo dos Santos

UFR – Universidade Federal de Rondonópolis

Dedico

A meus pais Irenilda e Leonidio, a minha filha Cecília e meus irmãos Welington e Elizaine.

AGRADECIMENTOS

A Deus, por me dar forças e ânimo em todos os momentos.

Ao Prof. Dr. Rivanildo Dallacort, por ter aceitado o desafio de me orientar e pela paciência na transmissão do conhecimento.

Ao corpo docente e aos técnicos da UNEMAT e do Programa de Pós-Graduação em Biodiversidade e Agroecossistemas Amazônicos por contribuírem para a qualidade na Universidade Pública e na minha formação.

Aos membros da banca de qualificação.

A meus pais e irmãos, pelo apoio nas adversidades.

Aos colegas de curso pelo apoio e trocas de experiencias.

“... melhor é a sabedoria do que os rubins; e de tudo o que se deseja nada se pode comparar com ela. Eu, a Sabedoria, habito com a prudência e acho a ciência dos conselhos!”

Provérbios 8: 11 - 12

SUMÁRIO

LISTA DE TABELAS

ARTIGO 1:

ARTIGO 2:

LISTA DE FIGURAS

ARTIGO 1:

ARTIGO 2:

LISTA DE SIGLAS E ABREVIATURAS

PAP Meteor Preparação, Análise e Previsão de dados Meteorológicos

INMET Instituto Nacional de Meteorologia

QMS Erro Quadrático Médio

OMM Organização Mundial de Meteorologia

GPGPU Unidade de Processamento Gráfico de Propósito Geral

RESUMO

O conhecimento das condições meteorológicas é fundamental para tomada de decisões em agroecossistemas, sendo necessárias informações precisas do clima e das condições atmosféricas. Para tanto, algumas etapas são requeridas para ser construído. A coleta de dados é a etapa inicial para este processo, porém, está sujeita a erros que fatalmente prejudicará as análises subsequentes e a geração do conhecimento, implicando em decisões errôneas. As informações necessárias para construção do conhecimento nem sempre estão disponíveis ou são confiáveis. Desde modo, mecanismos para tratamento, análise e previsões são imprescindíveis no gerenciamento de agroecossistemas garantindo eficiência e assertividade nas decisões. O objetivo deste trabalho é descrever o funcionamento do PAP Meteor (Preparação, Análise e Previsão de dados Meteorológicos), aplicando dados fornecidos pelo Instituto Nacional de Meteorologia (INMET) em estação meteorológica de superfície nos municípios de Matupá MT e de Sinop MT. O PAP Meteor é um sistema (WEB) desenvolvido com a linguagem de programação Python, subdividido em 3 módulos. O módulo de pré-processamento é responsável por fazer a leitura da base de dados e retornar suas informações principais, além de identificar anomalias e imputar registros ausentes. O módulo de análise exploratória realiza um resumo estatístico dos dados, análise de correlação além de explorar os dados com tabelas e gráficos dinâmicos. Nos dados meteorológicos de Matupá foram identificadas inconsistências em temperatura e precipitação, além de 55,1 % de falha nos registros, em Sinop as falhas somam 28%. O sistema foi eficiente em imputar dados faltantes de temperatura, umidade relativa e precipitação. Em Matupá as temperaturas variam entre 10 e 40 °C com médias anuais de 33 °C com tendência positiva de aumento. A precipitação é predominante nos meses de janeiro a abril e de outubro a dezembro. O PAP Meteor pode contribuir no processo de geração de conhecimento, contribuindo para maior sustentabilidade e racionalização de recursos em agroecossistemas.

Palavras Chave: Data Wrangling, mineração de dados, computação aplicada, Amazônia

ABSTRACT

Knowledge of meteorological conditions is fundamental for decision making in agroecosystems, requiring accurate information of climate and weather conditions. For this, some steps are necessary to be built. Data collection is the initial step in this process, but is subject to errors that will fatally affect subsequent analysis and the generation of knowledge, leading to erroneous decisions. The information needed for knowledge construction is not always available or reliable. Thus, mechanisms for treatment, analysis, and forecasting are essential in the management of agro ecosystems, ensuring efficiency and assertiveness in decisions. The objective of this work is to describe the operation of PAP Meteor (Preparation, Analysis and Forecast of Meteorological Data), applying data provided by the National Institute of Meteorology (INMET) in surface meteorological station in the municipality of Matupá MT and Sinop MT. The PAP Meteor is a system (WEB) developed with Python programming language, subdivided into 3 modules. The pre-processing module is responsible for reading the database and returning its main information, besides identifying anomalies and imputing missing records. The exploratory analysis module performs a statistical summary of the data, correlation analysis, and explores the data with dynamic tables and graphs. In the meteorological data of Matupá, inconsistencies in temperature and precipitation were identified, besides 55.1% of missing records. The system was efficient in imputing missing data for temperature, relative humidity and precipitation. In Matupá the temperatures vary between 10 and 40 °C with annual averages of 33 °C with a positive tendency to increase. Precipitation is predominant in the months of January through April and October through December. The months of May to August have the highest rates of insolation. PAP Meteor can contribute to the process of knowledge generation, contributing to greater sustainability and rationalization of resources in agro ecosystems.
Keywords: Data Wrangling, data mining, applied computing, Amazon

1. INTRODUÇÃO GERAL

A dinâmica entre dados e informações contribuem para a formação dos maiores ativos de uma organização, Sendo considerados a informação e o conhecimento como bens de grande valor (CEREZO-NARVÁEZ; et al., 2021). Portanto, se necessita abranger informações internas e externas, sistematicamente coletadas, analisadas e disseminadas, de modo a transformar informações em conhecimento estratégico (RODRIGUES; BLATTMANN; RODRIGUES, 2014). Para tal, são empregados mecanismos para coleta e armazenamento de dados. Contudo, os processos de coleta de dados são passíveis de erros, contendo dados imprecisos, incorretos e lacunas. Tais problemas, implicam em informações imprecisas e tomadas de decisões errôneas, fatalmente prejudicando as atividades e ações de uma instituição.

Sob este ponto de vista, para tomar decisões pautadas em dados são necessários cuidados e técnicas desde a coleta dos dados até a geração de conhecimento. Deste modo, para as decisões estratégicas são empregados diversos métodos e tecnologias, iniciando pelo tratamento dos dados.

Erros e falhas em série histórica de dados meteorológicos são recorrentes, (MACHADO; ASSIS, 2020). Além de registros errados, o maior problema das séries históricas são a ausência de registros (OLIVEIRA et al., 2010), sendo necessário a implementação de modelos e ferramentas para identificação e correção de erros e falhas.

Há vários métodos para limpeza e correção de dados meteorológicos. Tais como os métodos de Krigagem, inverso da distância, regressão linear e ponderação regional (BABA; VAZ; COSTA, 2014). Contudo, há carência de ferramentas que facilitam ou torna acessível a pesquisadores e a cidadãos comuns a limpeza e a correção de dados desta natureza. Outro sim, a eficiência nos métodos é relativa, dependendo do nível de qualidade dos dados, de como estão organizados e dos períodos de coleta (anual, mensal, diário, decendial, etc.). Para preenchimento de falhas, por exemplo, a metodologia de regressão linear múltipla e ponderação regional foram eficientes nos trabalhos de Oliveira et al. (2010). Já para elementos climáticos anuais a Interpolação polinomial foi considerada um bom método (SLUITER,

2009). Porém, tais métodos não são eficientes para todas as variáveis, para precipitação por exemplo, os modelos citados não tiveram uma acurácia significativa. Portanto a escolha da metodologia é relativa, cabendo o pesquisador avaliar a estrutura e integridade dos seus dados.

À medida que os dados meteorológicos são “arrumados”, estão aptos a serem transformados em informações e com isso ocorre a geração de conhecimentos auxiliando a tomada de decisão. A análise exploratória de dados são meios úteis para a geração de “insight” e a base para tomada de decisões estratégicas.

Portanto, neste trabalho investigamos mecanismos para descoberta automática de erros e “limpeza” de dados, bem como a análise exploratória a partir de estatística descritiva, gráficos e técnicas de aprendizagem de máquina. Conferindo a gestores e pesquisadores o poder de acessar e processar seus dados em dispositivos multiplataforma sem grande esforço computacional e conhecimento de programação. Contribuindo, portanto, para decisões mais assertivas, implicando todas as esferas da sociedade que faz uso dos conhecimentos climáticos, em especial aos agroecossistemas, permitindo o uso racional de recursos e uma produtividade mais sustentável.

Diante disso, a dissertação será dividida em dois capítulos. No primeiro capítulo propõe-se o desenvolvimento de um sistema (web) no qual, o usuário submete uma série histórica de dados no formato .CSV. Após isso, a ferramenta deverá retornar à apresentação dos dados, informações básicas do conjunto, além de, identificar valores discrepantes, duplicados e ausentes através de módulos de configuração de parâmetro da base. Também, a partir dos erros devidamente identificados, deverá haver a função de corrigir os dados e imputar os registros ausentes, preparando o caminho para análises posteriores.

O segundo capítulo, o usuário poderá utilizar a série histórica preparada pelo sistema descrito anteriormente ou utilizar dados brutos (originais). Desta maneira poderá se verificar a estatística descritiva básica, análise exploratória e a decomposição sazonal.

2. REFERÊNCIAS BIBLIOGRÁFICAS

BABA, Ricardo Kazuo; VAZ, Maria Salete Marcon Gomes; COSTA, Jéssica da. Correção de dados agrometeorológicos utilizando métodos estatísticos. **Revista Brasileira de Meteorologia**, v. 29, n. 4, p. 515–526, 1 out. 2014. DOI 10.1590/0102-778620130611. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-77862014000400005. Acesso em: 6 abr. 2020.

CEREZO-NARVÁEZ, Alberto; PASTOR-FERNÁNDEZ, Andrés; OTERO-MATEO, Manuel; BALLESTEROS-PÉREZ, Pablo; RODRÍGUEZ-PECCI, Francisco. Knowledge as an Organizational Asset for Managing Complex Projects: the case of naval platforms. **Sustainability**, [S.L.], v. 13, n. 2, p. 885, 17 jan. 2021. MDPI AG. http://dx.doi.org/10.3390/su13020885.

MACHADO, Lilian Aline; ASSIS, Wellington Lopes. Comparação entre métodos de preenchimento de falhas em séries de dados meteorológicos da bacia hidrográfica do Rio das Velhas (MG). **Revista Geografias**, v. 26, n. 1, p. 73–90, 5 fev. 2020. Disponível em: https://periodicos.ufmg.br/index.php/geografias/article/view/19216. Acesso em: 19 jan. 2021.

MARTINS, Claudia; OLIVEIRA, Henrique; OLIVEIRA, Allan Gonçalves de. Uma abordagem computacional para preenchimento de falhas em dados micrometeorológicos. **Revista Brasileira de Ciências Ambientais**, p. 61–70, 2013. Disponível em: https://www.academia.edu/17220119/Uma_abordagem_computacional_para_preenchimento_de_falhas_em_dados_micrometeorológicos. Acesso em: 26 nov. 2020.

RODRIGUES, Charles; BLATTMANN, Ursula; RODRIGUES, Charles. Gestão da informação e a importância do uso de fontes de informação para geração de conhecimento. **Perspectivas em Ciência da Informação**, v. 19, p. 4–29, 2014. DOI 10.1590/1981-5344/1515. Disponível em: http://dx. Acesso em: 12 mar. 2020.

SLUITER, R. **Intern rapport**. . De Bilt: [s.n.], 2009. Disponível em: https://www.snap.uaf.edu/attachments/Interpolation_methods_for_climate_data.pdf. Acesso em: 6 abr. 2020.

3. CAPÍTULOS

3.1. SISTEMA PARA ANÁLISE E CORREÇÃO DE DADOS METEOROLÓGICOS

Resumo – O entendimento do tempo e do clima é indispensável para decisões assertivas em diversos campos da atuação humana. Necessitando, portando de dados consistentes e confiáveis para inferências e tomadas de decisão. Deste modo, o objetivo deste trabalho é 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 imputados os valores ausentes 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 – The understanding of time and climate is indispensable for assertive decisions in different fields of human activity. In need, carrying consistent and reliable data for inferences and decision making. Thus, the objective of this work is to describe the functionalities of a (web) system developed with the intention of identifying errors and imputing missing data in historical series of meteorological data, describing the characteristics and errors of the INMET database (National Institute of Meteorology) in the municipality 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. To validate the system, a historical series of meteorological data provided by INMET was used, its faults were treated and the missing values were imputed with the KNNImputer algorithm. The assertiveness of the imputation of missing values was verified through the metrics of Accuracy, Precision, Recall, F1-score and Mean Square Error (QMS). Such metrics are derived from the comparison of predicted values and original values per confusion matrix. The system was efficient in identifying outliers and imputing missing values, identifying 100% of the outliers of the analyzed variables.

Keyword: Data Wrangling, data mining, applied computing

Introdução

O Estado de Mato Grosso é rico em biodiversidade abrangendo 3 biomas (Amazônia, Cerrado e Pantanal). Cujo clima não é homogêneo, requerendo grande quantidade de estações de monitoramento meteorológico. No Estado possui apenas 39 estações administradas pelo INMET, sendo, deficiente ou inexistente o 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 facilite o tratamento de dados bruto, favorecendo a geração de informações e conhecimentos. Considerando que, para o processo de tomada de decisão, o conhecimento está no topo da hierarquia, tendo como base os dados, que serão adequados, rearranjados e transformados em informação (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; WOOD, 2005). Deste modo, não basta conhecer a importância e o valor da informação e conhecimento, mas deve haver estratégias que viabilize a qualidade destas. A mineração de dados (data mining) é considerada uma alternativa para extrair conhecimentos de grande volume de dados, possibilitando a descoberta de padrões e relações além da geração de regras para predizer e correlacionar dados (AVELAR; ROCHA; CRUZ, 2017). Tal tarefa consome cerca de 50-80% do tempo para as análises (BILALLI *et al.*, 2018).A metodologia de Data Wrangling é uma alternativa para a preparação dos dados.

Data Wrangling significa transformação e preparação de dados. É um processo de mapeamento e transformação de dados brutos, com intuito de deixá-los mais apropriado para análises posteriores. Em suma, este método diz respeito ao ato de coletar, limpar, combinar, normalizar, estruturar e

organizar dados (RATTENBURY et al., 2017). Tal processo não é tão simples de ser realizado.

A dificuldade de preparação aumenta quando há muitos dados a serem analisados ou quando a base apresenta problemas e erros no processo de coleta. Tanto que, em série histórica de dados meteorológicos devidamente preparadas para análise, são escassas. Outro sim, quanto maior a quantidade de dados a ser avaliado e processado, requer maior poder computacional para tal tarefa (BILALLI *et al.*, 2018). De modo que, alguns processos são inviáveis para pessoas que não possuem infraestrutura e domínio de técnicas de limpeza de dados (Data Cleansing).

Sistemas específicos para análise de dados meteorológicos também são escassos. De modo que, se faz necessário o desenvolvimento de sistemas aplicados a preparação de dados desta natureza para análises exploratórias explícitas e implícitas. Considerando ainda, a disponibilidade e acessibilidade para realizar os procedimentos supracitados.

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 pauta é a importância da geração do conhecimento a partir de dados meteorológicos na agropecuária e áreas afins, e da dificuldade em encontrar e/ou manipula-los 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.

Material e Métodos

Ferramentas de desenvolvimento

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.

Dados utilizados e validação do sistema

Para testar o desempenho do sistema desenvolvido, precisamente os módulos incumbidos de identificar, corrigir erros, e imputar informações ausentes foram necessários dados de estação meteorológica de superfície. Deste modo, foi solicitado ao Instituto Nacional de Meteorologia séries históricas de dados de estações convencionais e automáticas.

Dos dados adquiridos, foram selecionadas as variáveis umidade relativa (mínima, média e máxima), precipitação e temperatura (máxima e mínima) (Tabela 1), lembrando que para a variável de umidade relativa, a estação convencional de Matupá apresenta apenas umidade relativa média. Os dados oriundos de estação convencional são diários com dois registros no dia, sendo as 00:00 as 12:00 e das 12:00 as 18:00. Já dados de estações automáticas são horários das 00:00 as 23:00, no qual, foram enviados em sua forma bruta no formato .CSV, contendo erros e ausência de registros.

Tabela 1: Unidade de medidas dos dados utilizados para teste do sistema.

Variável	**Unidade de Medida**	**Categorias de dados**
Data\hora	Data\hora	Date
Precipitação	mm\dia	Float
UR[1] Máxima	%	Float
UR Mínima	%	Float
UR média	%	Float
Temp. Máxima	°C	Float
Temp. Mínima	°C	Float

Fonte: Elaboração própria

Os dados correspondem a duas estações meteorológicas de superfície. Uma estação convencional situada em Matupá MT, localizada na Latitude -10.1916° e longitude -54.9461° com altitude de 272 metros, cujo código da

[1] Umidade Relativa

OMN é o 83214. A segunda estação é automática e está no município de Sinop MT na Latitude -11,98, longitude -55,57 e altitude de 366,57 metros. A partir da disponibilidade dos dados, foi possível a identificação dos principais problemas a serem solucionados pelo sistema. Sendo um dos maiores desafios a definição de metodologias para tratamento dos dados. A seguir será apresentado o fluxo de funcionamento do sistema.

Métodos de Tratamento dos dados

Para preparar os dados, o sistema importa um conjunto de dados .CSV com encode UTF-8, cujo separador é definido pelo usuário. Tal arquivo não deve haver 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. Internamente, o sistema classifica a base de dados em ordem crescente (mais antigo para mais recente) por hora, mês, dia e data.

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, 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. Na quarta etapa ocorre a verificação de dados duplicados, retornado à somatória e as linhas duplicadas se houverem. Até esta etapa o sistema não realiza nenhuma modificação nos dados, apenas retorna informações básicas da base. No entanto, é necessário identificar

possíveis erros nos registros, sendo indispensável a identificação de ruídos nos dados (outliers).

Identificação de outliers

Prevalece em Mato Grosso o clima tropical super úmido de monção (IBGE, 2019). Este, com temperaturas altas na maior parte do ano, dificilmente abaixo de 10 °C. Considerando isso, as etapas subsequentes do sistema correspondem a identificação de outliers e correção de dados.

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.

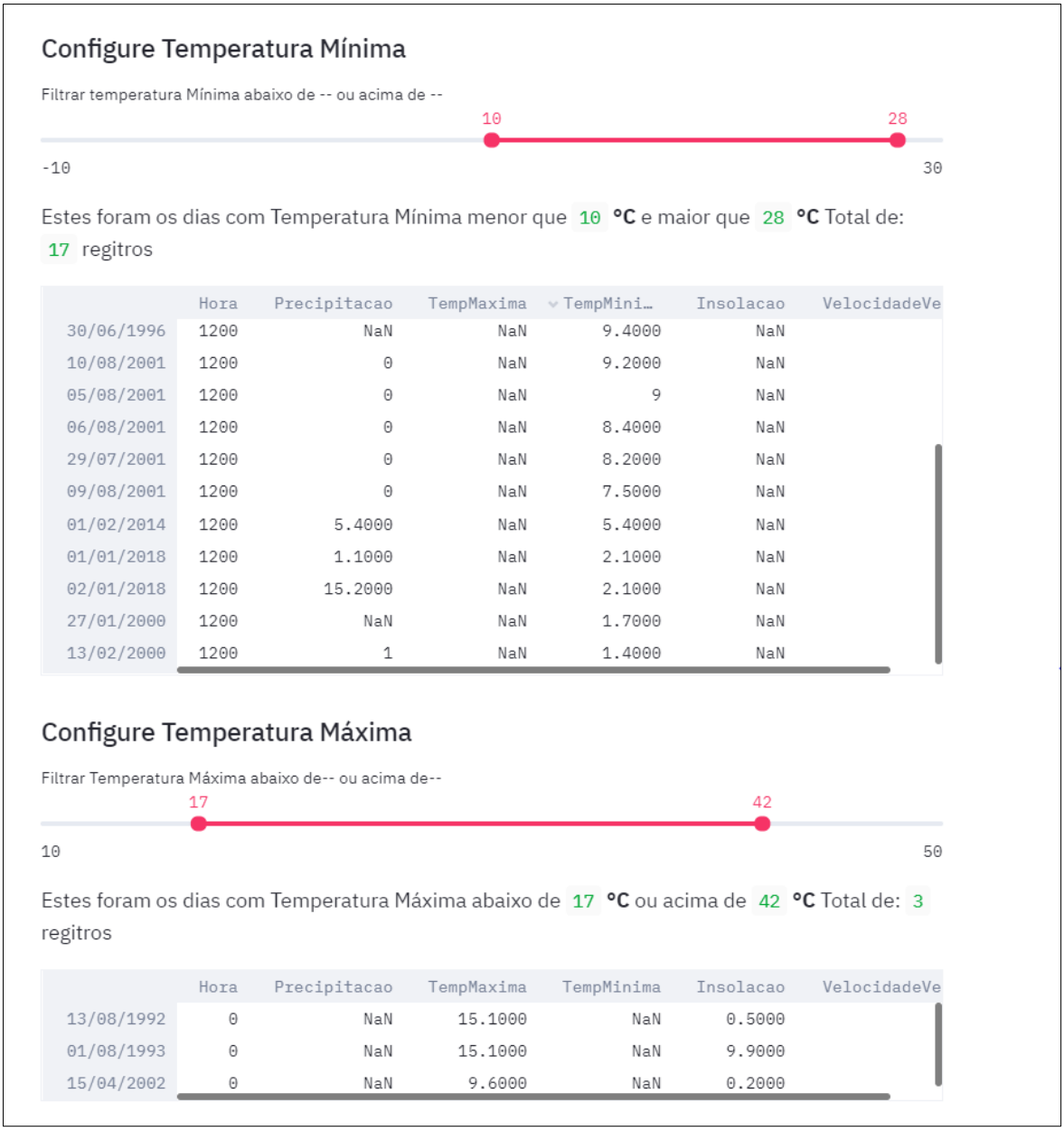


Figura 1 - Sistema de filtros do PAP Meteor

Conforme a Figura 1 o usuário filtra temperatura mínima abaixo de 10 °C e acima de 28 °C e temperatura máxima abaixo de 17 ºC ou acima de 42 °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 configuração 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 2.

Tabela 2: 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: Elaboração própria

A Tabela 2 possui caráter generalista, ou seja, contempla as condições ambientais de todo território brasileiro. No entanto, para o município de Matupá-MT e Sinop MT, foi estabelecido os critérios descritos na Tabela 3 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 3: 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: Elaboração própria

Para Sinop, devido à série histórica ser horária, somente foi atribuído limites mínimos e máximos para precipitação e umidade relativa. Para temperaturas, foi considerado somente os registros mínimos conforme a Tabela 3.

Com a identificação de erros e dados ausentes devidamente contabilizada 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.

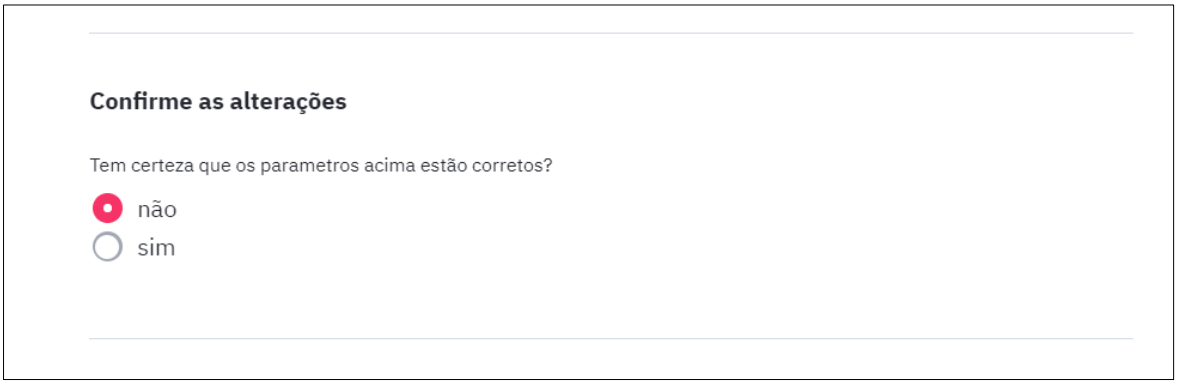


Figura 2: Confirmação de parâmetros para identificar dados anormais. Se pressionado o botão sim, os dados na condição configurada será considera nulo (Nan)

Correção de falhas (dados ausentes)

Para preenchimento dos dados ausentes foram utilizados o método K-Vizinhos Mais Próximos (KNNImputer).

O método k-vizinhos mais Próximo, 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’ (Distancia euclidiana dos valores ausentes).

Validação do algoritmo imputador (KNNImputer)

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 e identificar e 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 a Figura 3:

1) A partir de uma série histórica com dados brutos (D1), cujos 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. Validando o modelo com as métricas de Acurácia, Recall, Precisão, F1-score e EQM (Erro Quadrático Médio).

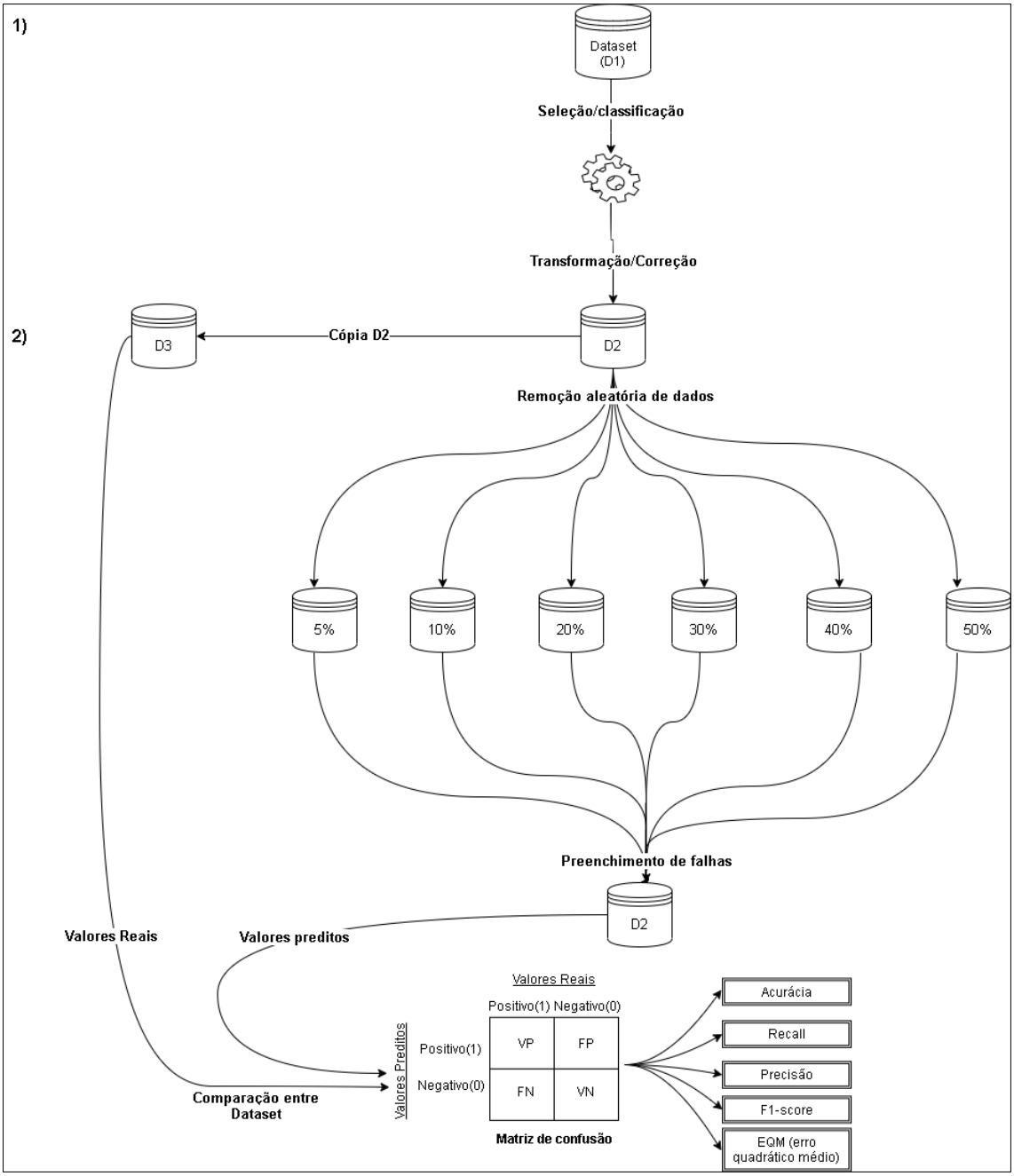


Figura 2: Metodologia aplicada para avaliação do desempenho do algoritmo imputador KNNImputer.

Além da validação através da remoção aleatória de porcentuais de dados, também foi verificado o quanto cada ano adicionado na série história afeta o modelo KnnImputer. De modo que, foi imputado dados em série histórica de 1 a 6 anos de dados e analisado através de matriz de confusão a acurácia, precisão e o erro quadrático médio em cada situação.

Métricas de avaliação do modelo de imputação de dados (KNNImputer)

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 n_{amostras} 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) é dividido pelos valores $vp + fn$ (falso negativo).

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

A Precisão corresponde à capacidade de evitar falso positivos (fp), cuja fórmula consiste na divisão dos verdadeiros positivos (vp) pela soma de verdadeiro (vp) positivo 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)

Resultados e Discussão

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.

Identificações básicas da base de dados

As informações básicas extraídas pelo PAP Meteor da estação meteorológica de superfície do INMET no Município de Matupá MT são demonstradas na Tabela 4, apresentando, seis (06) colunas com 25.567 registros no intervalo de 01/01/1987 a 31/12/2020 contendo 54,47% de dados válidos.

Tabela 4: Síntese de conjunto de dados da estação meteorológica de superfície do INMET no município de Matupá MT

Colunas	**Registros válidos**		**Tipo de Dados**
	Total	**Percentual**	
Data/Hora	25 567	100,00%	Int64
Precipitação	10 617	41,52%	Float64
Temp. Máxima	11 695	45,7%	Float64
Temp. Mínima	11 878	46,5%	Float64
UR média	21 518	84,1%	Float64

Fonte: Elaboração própria

Os dados do município de Sinop MT apresentam oito (8) colunas (Tabela 5) com 105.187 registros no intervalo de 01/01/2009 a 31/01/2020.

Tabela 5: Síntese de conjunto de dados da estação meteorológica de superfície do INMET no município de Sinop MT

Colunas	Registros válidos		Tipo de Dados
	Total	Percentual	
Data/Hora	105 187	100%	Int64
Precipitação	75 856	72%	Float64
Temp. Máxima	82 289	78%	Float64
Temp. Mínima	82 289	78%	Float64
UR Mínima	82 289	78%	Float64
UR Média	82 289	78%	Float64
UR Máxima	82 289	78%	Float64

Fonte: Elaboração própria

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 aferida com Pluviômetro, que está sujeito a ação do vento, topografia, ser obstruído por sujidades (folhas, objetos, etc.), além de problemas inerentes a erros humanos (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 4). Já entre 2000 a 2020 houve uma padronização na consistência dos registros. Tais fatos mencionados sugerem, algumas hipóteses:

a) O 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.

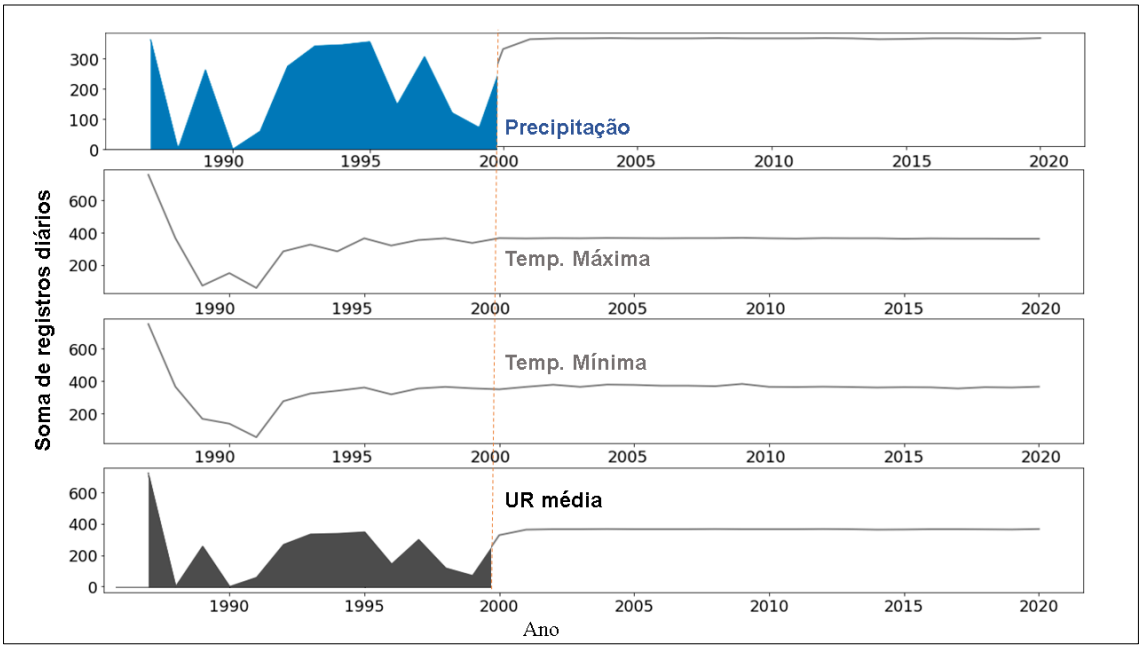


Figura 4: Quantidade média de registros válidos (não nulos) no período de 1985 a 2020 em Matupá MT.

No período de 1990 a 2000 houve o término da digitalização das normais climatológicas, com aplicativo específico, podendo calcular o balanço hídrico em todas as estações meteorológicas, contidas nos Normais Climatológicos. Sendo que, no processo de digitalização pode ter havido erros e perda de dados.

Outro fato histórico importante aconteceu em 1994. No qual através do Ofício 269/INMET de 11/11/94 foi encaminhado ao ministro da agricultura o relatório de atividades, 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, relata treinamento no programa “Qualidade 2000”, incluindo calibração instrumental. Ainda neste período houveram várias reformas e substituição 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 5). No entanto, estações automáticas embora possua menor número de falhas, não estão imunes a falha, estando sujeitas a vários problemas físicos, afetando a qualidade geral dos dados e até a interrupção dos registros (STRASSBURGER et al., 2011). 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).

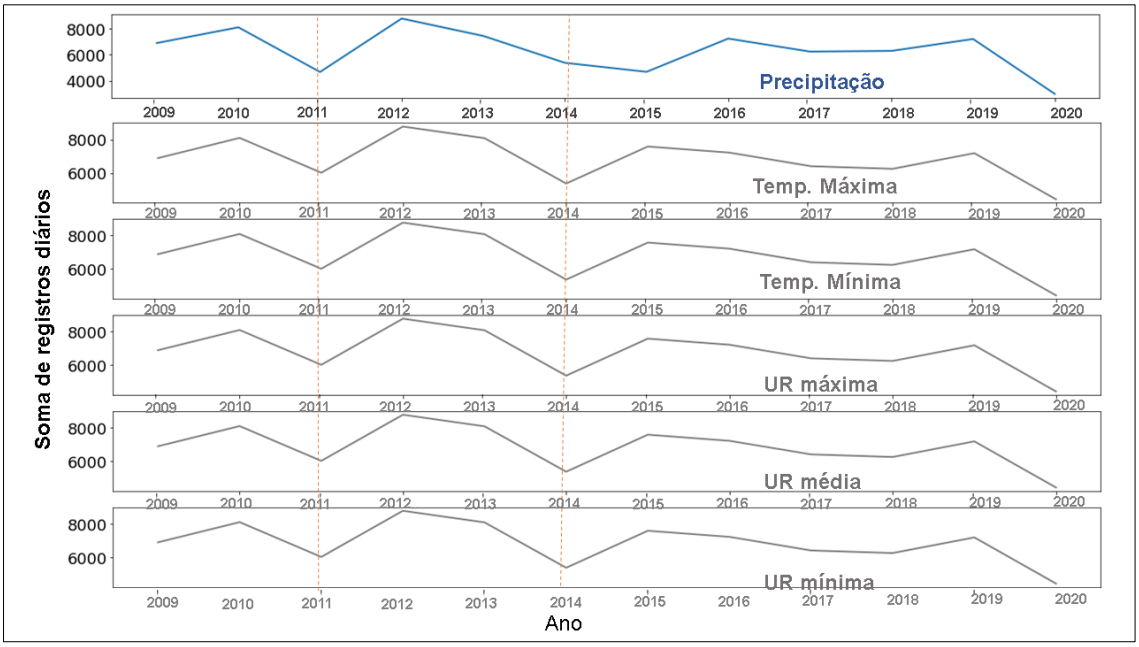


Figura 3: Quantidade média de registros válidos (não nulos) período de 2009 a 2020 em Sinop MT.

Devido a estes problemas aludidos, há na série histórica disponibilizada pelo INMET nas estações verificadas, valores discrepantes e duvidosos.

Verificação de outliers

A partir do resumo estatístico dos dados brutos foi possível evidenciar erros nas variáveis temperatura máxima e temperatura mínima (Tabela 5 e 6).

Tabela 5: Resumo estatístico dos dados brutos da Estação Meteorológica de Matupá MT no período de 1987 a 2020

	Precipitação (mm)	Temp. Máxima (°C)	Temp. Mínima (°C)	UR (%)
Contagem	10617	11695	11878	21518
Média	5,09	32,82	20,05	86,12
Desvio Padrão	12,81	2,70	2,40	11,78
Min	0,00	9,60	1,40	22,00
25%	0,00	31,40	18,80	81,00
50%	0,00	33,00	20,60	90,00
75%	3,00	34,50	21,80	95,00
Max	198,40	40,20	29,00	100,00

Fonte: Elaboração própria

Tabela 6: Resumo estatístico dos dados brutos da Estação Meteorológica de Sinop MT no período de 2009 a 2020

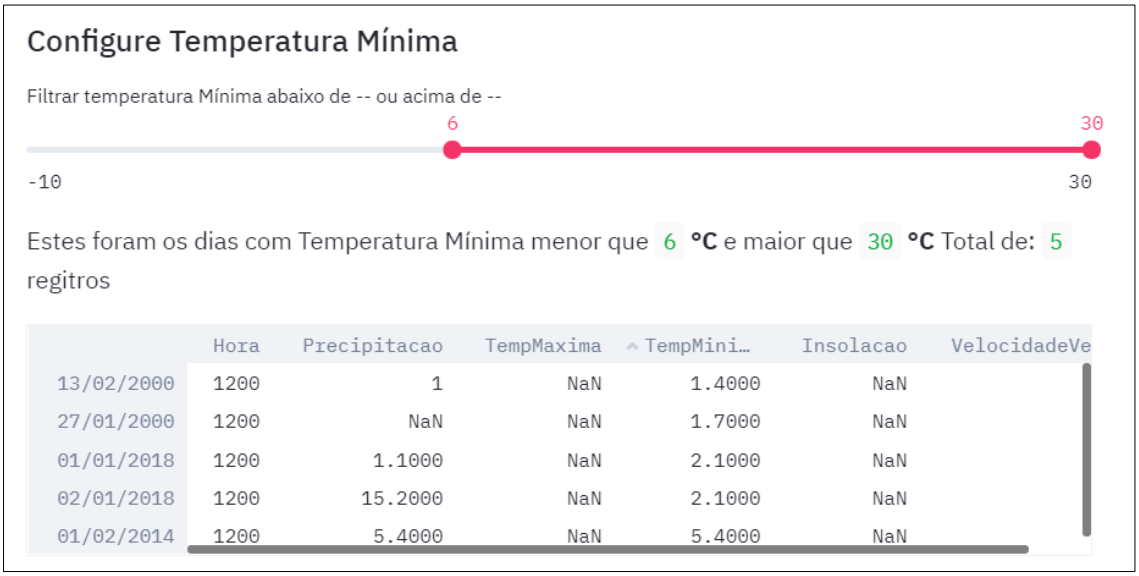
	Precipitação (mm)	Temp. Máxima (°C)	Temp. Mínima (°C)	UR máxima(%)	UR media(%)	UR mínima(%)
Contagem	75,856	82,289	82,290	82,290	82,290	82,290
Média	0,200108	26,37	24,92	75,51	72,48	69,45
Desvio Padrão	1,56	4,50	4,08	20,60	21,30	22,29
Min	0,00	10,10	4,60	13,00	12,50	11,00
25%	0,00	22,90	22,20	63,00	58,50	53, 00
50%	0,00	25,20	23,90	83,00	78,50	74,00
75%	0,00	29,90	27,70	93,00	91,50	90,00
Max	62,00	40,00	38,10	97,00	97,00	97,00

Fonte: Elaboração própria

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, dada 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 conforme a Figura 7. 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. Considerando ainda que 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).



	Hora	Precipitacao	TempMaxima	TempMini…	Insolacao	VelocidadeVe
13/02/2000	1200	1	NaN	1.4000	NaN	
27/01/2000	1200	NaN	NaN	1.7000	NaN	
01/01/2018	1200	1.1000	NaN	2.1000	NaN	
02/01/2018	1200	15.2000	NaN	2.1000	NaN	
01/02/2014	1200	5.4000	NaN	5.4000	NaN	

Figura 4: Filtro de Temperatura Mínima do sistema de limpeza de dados Meteorológicos. Temperaturas Mínimas abaixo de 6ºC ou maior que 30 °C, totaliza 5 registros em Matupá MT.

Nas verificações lógicas e de limites propostos na Tabela 2, resultaram com 100% dos dados brutos dentro dos parâmetros estabelecidos para ambas estações meteorológicas. Isso demonstra que embora haja bastante 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 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 (Tabela 7).

Tabela 7: Dados 'brutos' diários considerados errados ou suspeitos da estação do INMET em Matupá MT no período de 1987 a 2020.

Variável	Total de erros	Percentual
Temperatura Mínima $< 5\,°C$ ou $> 28\,°C$	7	0,027%
Temperatura Máxima $< 15\,°C$ ou $> 45\,°C$	1	0,004%
Precipitação $< 0mm$ ou $> 220mm$	0	0,000%
Umidade Relativa média $< 0\,\%$ ou $> 100\%$	0	0.000%

Fonte: Elaboração própria

Os erros no conjunto de dados também podem ser visualizados através de gráfico do tipo Boxplot (Figura 7 e 8). 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 (Tabela 6 e 7) 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 8). 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.

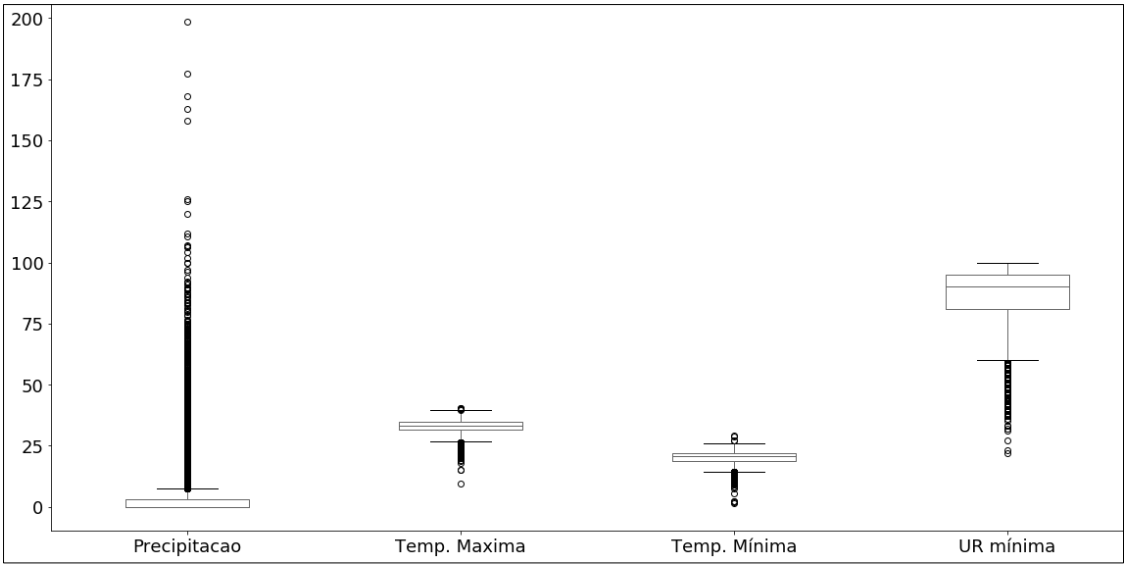


Figura 5: Identificação de outliers através de boxplot na estação de Matupá MT

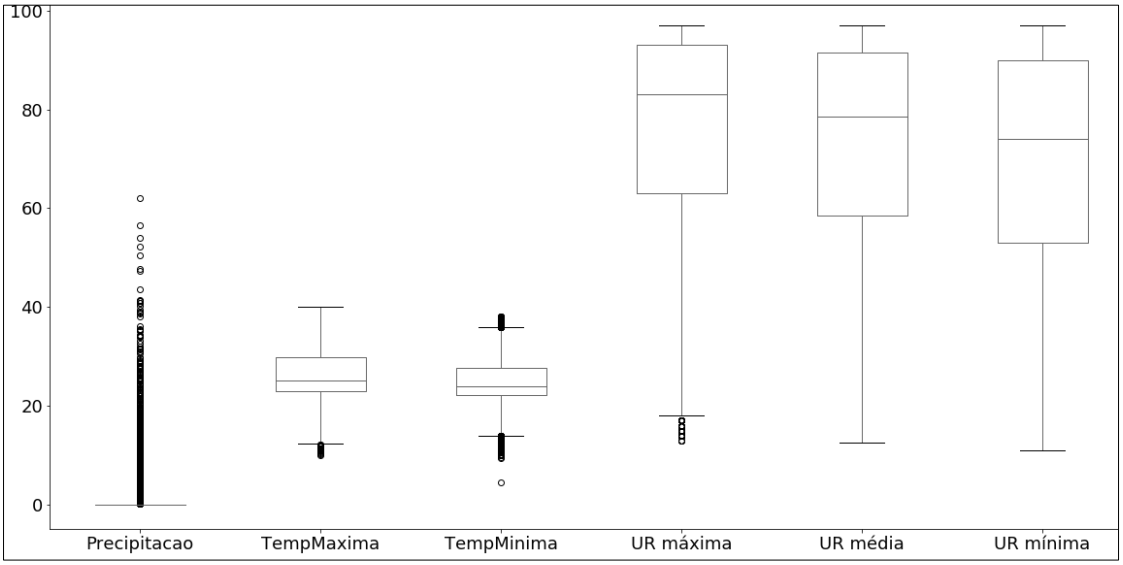


Figura 6: Identificação de outliers através de boxplot na estação de Sinop MT

Imputação de dados ausentes.

A série histórica de Matupá MT possui 45,53% dos ausentes, com destaque para variável Precipitação, (Figura 9) sendo seus períodos mais críticos de coleta nos anos 1998 a 2000. De modo que, anualmente em média 226 dias deixava de ser registrado, o equivalente 8 dias no mês. Para 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 foi perdido em média apenas 2 registros por mês.

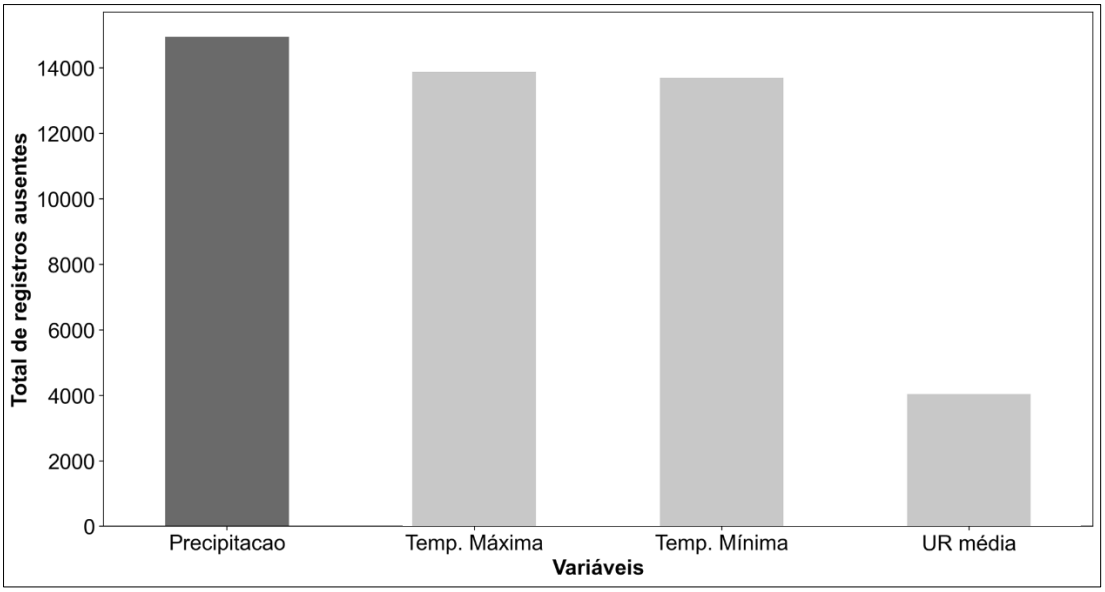


Figura 7: Dados ausentes separados por variáveis em estação do INMET no Município de Matupá MT (1986-2020).

Quanto as falhas nos registros, na estação de Sinop, a variável precipitação também apresenta o maior número de falhas (Figura 10).

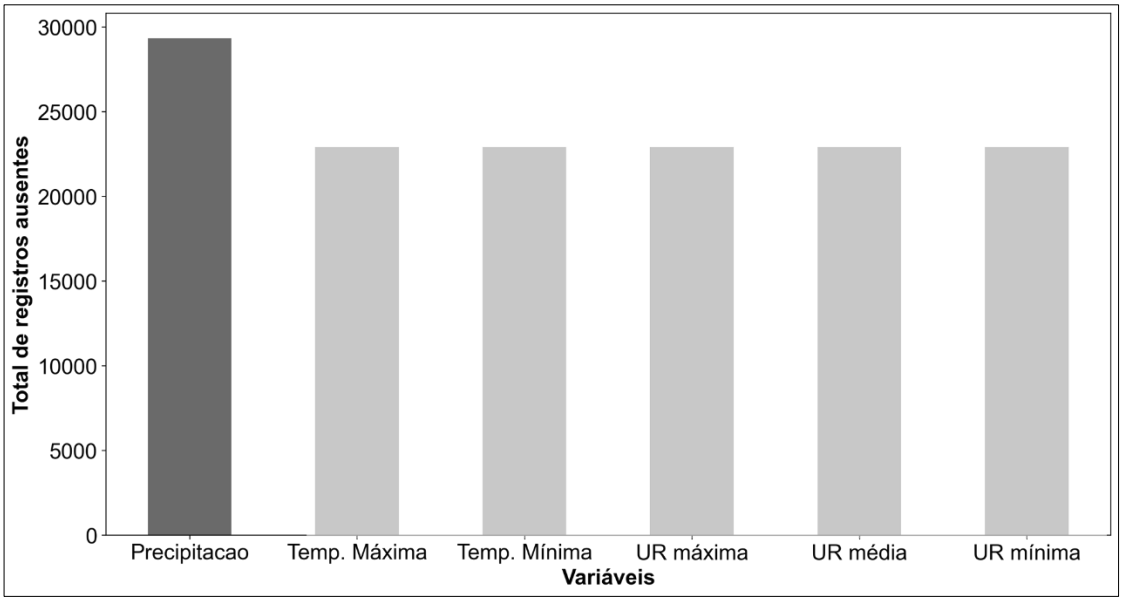


Figura 8: Dados ausentes separados por variáveis em estação do INMET no Município de Sinop MT (2009-2020)

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% de dados ausentes (Tabela 7) e com series histórica com 3 anos de registros (Tabela 8) 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 falha aumenta, consequentemente o erro quadrático médio (QMS) também aumenta.

Tabela 6: 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: Elaboração própria

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 (Figura 11) e 7% da precisão. Já para estação convencional com série histórica diária a proporção de acurácia é de 10:7 e 10:4 na métrica de precisão.

Tabela 7: 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: Elaboração própria

Para precipitação, a acurácia foi de 100% 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 11A) e o erro quadrático médio também aumenta. Outro sim, agrupamentos de 3 anos com 10% de falhas apresenta melhores resultados de precisão, sendo que, séries com 5 anos a precisão diminui 3% (Figura 11B), a acurácia não apresenta uma diminuição significativa, porém, o QMS aumenta gradualmente.

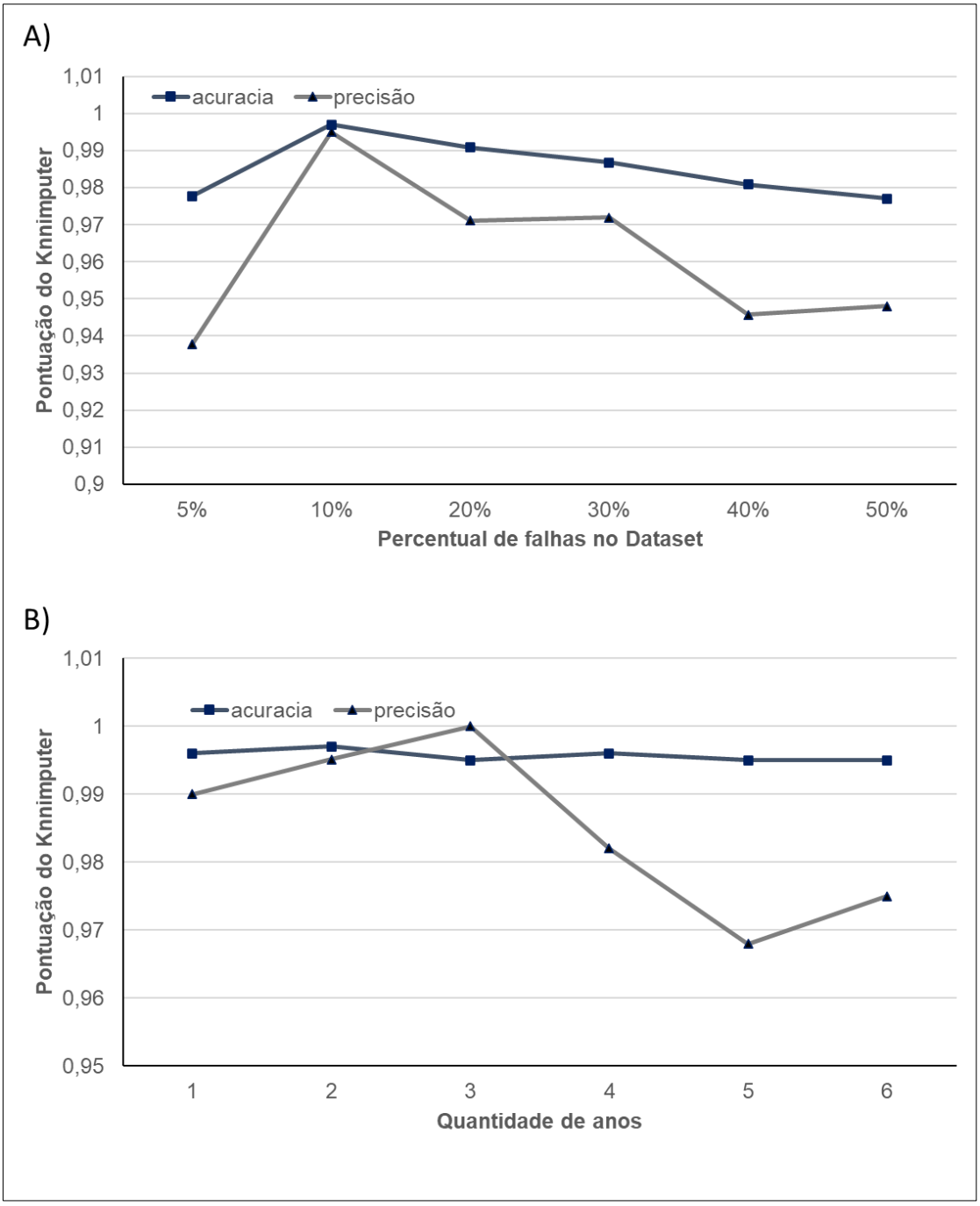


Figura 9: Pontuação do preenchimento de falhas do KNNImputer em dados de Precipitação da estação automática de Sinop MT de 2009 a 2011. A) Série com

5 a 50% de falhas em dados horários. B) Séries com 10% de falhas de 1 a 6 anos de registros.

O preenchimento de falhas para precipitação ainda é considerado um problema de difícil solução. Em 2005 nos trabalhos de Chibana et al., (2005) não havia métodos para imputar com eficiência dados diários e horários de precipitação, no qual recomendou que os preenchimentos ocorressem em dados mensais ou anuais. Desde então, foi aplicado e desenvolvido várias metodologias para o preenchimento de registros ausentes, no entanto, a maioria dos trabalhos limita-se a série com dados mensais e anuais. Dentre os métodos de preenchimento o que apresenta maior destaque é a ponderação regional, apresentando resultados satisfatórios em dados com poucas lacunas e com dados mensais, (NASCIMENTO et al., 2010), (SOARES; SILVA, 2017) e (DIAZ; PEREIRA; NOBREGA, 2018). Contudo, o preenchimento de falhas em dados diários e horários ainda é pouco discutido.

Métodos de regressão linear múltipla e ponderação regional tiveram bom desempenho nos trabalhos de Bier; Ferraz (2017) e de Ventura *et al.*(2016) na imputação de dados de temperatura, porém não encontraram nenhum método com o mesmo desempenho para falhas de precipitação. Contudo, com o PAP Meteor foi possível uma imputação satisfatória para dados horários, 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 12). Neste caso, recomenda-se transformar a série para apenas um registro ao dia.

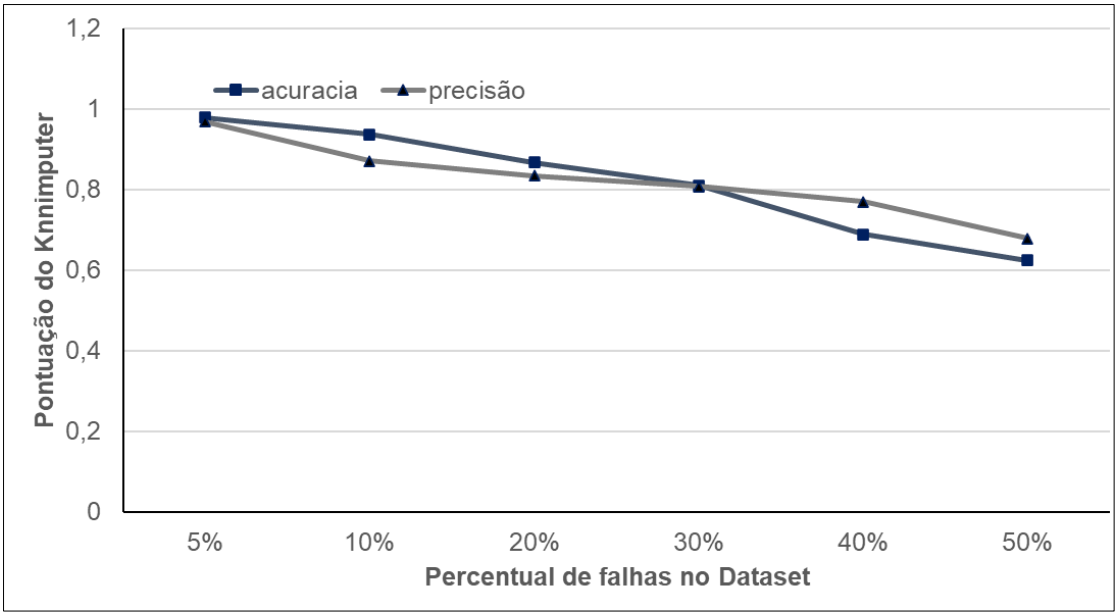


Figura 10: Pontuação do preenchimento de falhas do KNNImputer em dados de Precipitação da estação convencional de Matupá MT (série com dados diários) de 2005 a 2008.

Para as variáveis de temperatura (máxima e mínima) o KNNImputer também foi bastante preciso tanto em série com dados diários e horários (Figura 13). 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 13 C e D).

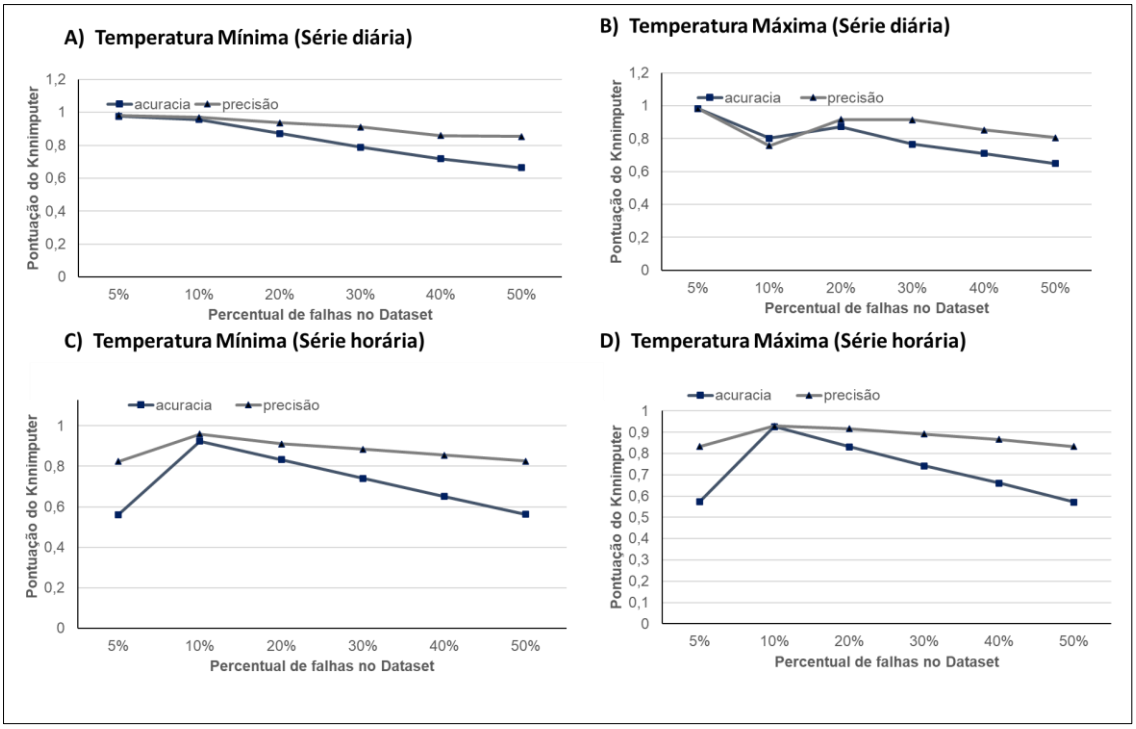


Figura 11: Pontuação do preenchimento de falhas do KNNImputer em dados de Temperatura mínima e máxima de Matupá MT e Sinop MT de 2005 a 2008. Em A) e B) representa a pontuação para 5 a 50% de falhas de temperatura na estação de Matupá MT. Em C) e D) representa a pontuação para 5 a 50% de falhas de temperatura na estação de Sinop MT.

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 grande variabilidade de relevo. Tais fatores implica em predição errônea, considerando que tais variações no relevo, pode ocorrer micro climas e ventos acentuados, afetando o desempenho das estimativas. (FENSTERSEIFER, 2013).

Todos os métodos citados para preenchimento de falhas consomem elevado custo computacional. Desde modo, limita 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.

Conclusões

O PAP Meteor demonstrou ser eficiente em preencher falhas em dados meteorológico, em séries históricas diárias e horária. 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 Bibliográficas

AVELAR, Cátia Fabíola Parreira de; ROCHA, Thiago Augusto Hernandes; CRUZ, Flávia Juliesse Soares. MINERAÇÃO DE DADOS. **Revista Vianna Sapiens**, v. 8, n. 2, p. 25, 12 dez. 2017. DOI 10.31994/rvs.v8i2.232.

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. FapUNIFESP (SciELO). 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. Disponível em: http://www.scielo.br/scielo.php?pid=S0102-77862017000200215&script=sci_abstract&tlng=pt. Acesso em: 6 abr. 2020.

BILALLI, Besim *et al.* Intelligent assistance for data pre-processing. **Computer Standards and Interfaces**, v. 57, p. 101–109, 1 mar. 2018. DOI 10.1016/j.csi.2017.05.004.

CHAFFEY, Dave; WOOD, Steve. Business information management : improving performance using information systems. [s.l.]: **Prentice Hall/Financial Times**, 2005.

CHIBANA, E. Y.; FLUMIGNAN, D.; MOTA, R. G.; VIEIRA, A. de S.; FARIA, R. T. Estimativa de falhas em dados meteorológicos. Anais online. In: V CONGRESSO BRASILEIRO DE AGROINFORMÁTICA, SBI-AGRO. Londrina/PR: Anais online, 2005, 8 p. Disponível em: < http://www.sbiagro.org.br/pdf/v_congresso/Trabalho41.pdf> Acessado em: 12/04/2019.

COSTA, Rafaela Lisboa; SILVA, Fabrício Daniel dos Santos; SARMANHO, Gabriel Fonseca; LUCIO, Paulo Sérgio. 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. Revista Brasileira de Geografia Fisica. http://dx.doi.org/10.26848/rbgf.v5i3.232861.

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 Parana. 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. http://dx.doi.org/10.1175/2008jamc1757.1.

FENSTERSEIFER, C.A.; Qualidade das estimativas de precipitações derivadas de satélites na bacia do Alto Jacuí –RS. Dissertação de mestrado. Programa de Pós-Graduação em Engenharia Civil e Ambiental. Santa aria –RS, 2013. 126p.

IBGE. IBGE Mato Grosso. 2019. **Instituto Brasileiro de Geografia e Estatística**. Disponível em: https://cidades.ibge.gov.br/brasil/mt/historico. Acesso em: 18 abr. 2020.

INMET, Instituto Nacional de Meteorologia. INMET :: BDMEP Matupá MT. 2020. **Banco Dados INMET**. Disponível em: https://bdmep.inmet.gov.br/. Acesso em: 20 jan. 2021.

INMET, Instituto Nacional de Meteorologia. **Relatório do gestor 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.

INMET, Instituto Nacional de Meteorologia. **Relatório de gestão 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.

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. Disponível em: http://dx.doi.org/10.1590/1981-5344/4192. Acesso em: 8 abr. 2020.

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

NASCIMENTO, Telma S. do; SARAIVA, Jaci Maria B.; SENNA, Renato; AGUIAR, Francisco Evandro O.. 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. Universidade Federal do Parana. 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. Disponível em: https://revistas.ufpr.br/revistaabclima/article/view/43762. Acesso em: 22 out. 2020.

RATTENBURY, Tye *et al.* Principles of Data Wrangling: Practical Techniques for Data Preparation. **O'Reilly Media**. 1. ed. Sebastopol CA: [s.n.], 2017. Disponível em: https://books.google.com.br/books?id=SEUqDwAAQBAJ&pg=PA30&lpg=PA30&dq=data+wrangling+artigos&source=bl&ots=ny77mqG5J7&sig=ACfU3U2QQCXbRZt3cQmk3BNRa-HCp5aW3A&hl=pt-BR&sa=X&ved=2ahUKEwiMxrLF7d7oAhWLKLkGHQztAgUQ6AEwA3oECAwQLg#v=onepage&q=data wrangling artigos&f=false. Acesso em: 10 abr. 2020.

SEIBERT, Jan; MORÉN, Ann-Sofie. Reducing systematic errors in rainfall measurements using a new type of gauge. **Agricultural And Forest Meteorology**, [S.L.], v. 98-99, p. 341-348, dez. 1999. Elsevier BV. http://dx.doi.org/10.1016/s0168-1923(99)00107-0.

SILVA, Fábio Cardozo da *et al.* Tratamento de Grandes Volumes de Dados Meteorológicos Através de Workflows Científicos Paralelos em ambientes GPGPU-CUDA. 2015. **Anais** [...]. Rio de Janeiro: [s.n.], 2015.

STRASSBURGER, André Samuel; MENEZES, Antônio José Elias Amorim de; PERLEBERG, Tângela Denise; EICHOLZ, Eberson Diedrich; MENDEZ, Marta Elena Gonzalez; SCHÖFFEL, Edgar Ricardo. Comparação da temperatura do ar obtida por estação meteorológica convencional e automática. **Revista Brasileira de Meteorologia**, [S.L.], v. 26, n. 2, p. 273-278, jun. 2011. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/s0102-77862011000200011.

TARAPANOFF, Kira. Inteligência, informação e conhecimento**.** 1. ed. Brasília: **IBICT, UNESCO**, 2006. v. 1.

TROYANSKAYA, Olga *et al.* Missing value estimation methods for DNA microarrays. **BIOINFORMATICS**, v. 17, n. 6, p. 520–525, fev. 2001. Disponível em: https://academic.oup.com/bioinformatics/article-abstract/17/6/520/272365. Acesso em: 6 maio 2020.

VENTURA, Thiago Meirelles *et al.* Análise da aplicabilidade de métodos estatísticos para preenchimento de falhas em dados meteorológicos (analysis methods of application for statistical data in meteorology). **Revista Brasileira de Climatologia**, v. 19, n. 0, p. 2237–8642, 17 out. 2016. DOI 10.5380/abclima.v19i0.44989. Disponível em: https://revistas.ufpr.br/revistaabclima/article/view/44989. Acesso em: 26 nov. 2020.

YAGUCHI, Silvia Manami; MASSIGNAM, Angelo Mendes; RICCE, Wilian da Silva; PANDOLFO, Cristina. PREENCHIMENTO DE FALHAS DOS DADOS DIÁRIOS DE TEMPERATURA MÁXIMA E MÍNIMA DO AR. **Ciência e Natura**, [S.L.], v. 38, n. 3, p. 1419-1425, 28 set. 2016. Universidad Federal de Santa Maria. http://dx.doi.org/10.5902/2179460x19502.

WMO. **Guide to Meteorological Instruments and Methods of Observation**. . Geneva: [s.n.], 2008. Disponível em: www.wmo.int. Acesso em: 7 maio 2020

3.2. SISTEMA PARA ESTATÍSTICA DESCRITIVA, ANÁLISE EXPLORATÓRIA DE DADOS METEOROLÓGICOS

Resumo – A sistematização e organização de dados climáticos é fundamental para inferir e tomar decisões assertivas e eficientes. O objetivo deste trabalho foi experimentar os módulos de análise exploratória e estatística descritiva do sistema PAP Meter (Preparação, análise e previsão meteorológica) aplicando duas séries histórica com diferentes percentuais de falhas nos registros (45,53% e 22,8%), além de descrever a distribuição temporal de Precipitação e temperatura dos municípios de Matupá MT e Sinop MT. Foi analisado através do sistema mencionado, uma série temporal disponibilizada pelo Instituto Nacional de Meteorologia (INMET) em um intervalo de 11 a 33 anos. O sistema possibilita a sumarização dos dados além de uma visão global dos dados de forma dinâmica e interativa. O módulo de análise exploratória possibilita uma visão holística dos dados além de decompor os dados com agrupamentos anuais e mensais em forma de tabelas e gráficos dinâmicos. No módulo de estatística básica é possível sumarizar os dados exibindo em forma de tabela as medidas de tendência central, dispersão, correlação e quartis. Os municípios de Matupá MT e Sinop MT apresentam temperaturas variando em média entre 10 e 40 °C, com duas estações bem definidas, uma quente e outra chuvosa. As novas funcionalidades do PAP Meteor facilitam a compreensão das variáveis meteorológicas através de recursos visuais e sem o dispêndio de infraestrutura computacional e de programação.

Palavras-chave: Análise de dados, mineração de dados, Amazônia.

Abstract – The systematization and organization of climatic data is fundamental to infer and make assertive and efficient decisions. The objective of this work was to try the exploratory analysis and descriptive statistics modules of the PAP Meter system (Preparation, analysis and meteorological forecast) applying two historical series with different percentages of failures in the records (45.53% and 22.8%), in addition to to describe the temporal distribution of Precipitation and temperature in the municipalities of Matupá MT and Sinop MT. It was analyzed through the mentioned system, a time series made available by the National Meteorological Institute (INMET) in an interval of 11 to 33 years. The system makes it possible to summarize the data in addition to a global view of the data in a dynamic and interactive way. The exploratory analysis module provides a holistic view of the data in addition to decomposing the data with annual and monthly groupings in the form of dynamic tables and graphs. In the basic statistics module, it is possible to summarize the data by displaying the measures of central tendency, dispersion, correlation and quartiles in a table form. The municipalities of Matupá MT and Sinop MT have temperatures varying on average between 10 and 40 ° C, with two well-defined seasons, one hot and the other rainy. The new features of PAP Meteor facilitate the understanding of meteorological variables through visual resources and without the expense of computational and programming infrastructure.

Keywords: Data analysis, data mining, Amazon.

Introdução

Para tomar decisões em agroecossistemas é necessário considerar as condições climatológicas. Haja vista que, há nestes grande dependência dos fatores climáticos (FIORIN; ROSS, 2015). Para tanto, é necessário conhecimento, requerendo boas fontes de informação e ferramentas para mineração de dados.

Quando se trata de ferramentas para mineração de dados, está incluído desde a infraestrutura computacional até técnicas específicas para tal atividade. De maneira que, nem sempre existem pessoas capacitadas ou condições adequadas para que possam ser desenvolvidas essas atividades. No que tange a construção de conhecimento, em relação às condições meteorológicas e climáticas, se faz necessário construir informações e desenvolver etapas de mineração de dados requeridas com eficiência (VIANNA *et al.*, 2017).

A sistematização, organização e sintetização dos dados climáticos inicia-se pelo monitoramento das condições atmosféricas, oriundas de instrumentos instalados em estações meteorológicas. Sendo, portanto, necessário a aplicação de métodos para descrever de forma quantitativa e qualitativa as características dos dados coletados.

Métodos estatísticos são fundamentais para a compreensão e inferências a respeito das condições do tempo atmosférico. A estatística descritiva é essencial para compreender as características básicas de um determinado conjunto de dados. Deste modo, as medidas de tendência central (média, mediana e moda), de dispersão (desvio padrão, coeficiente de variação, quartis, etc.) e de correlação, são elementos importantes para compreensão e interpretação dos dados climáticos (AKAMINE; YAMAMOTO, 2013). Tais conhecimentos contribuem para várias atividades humanas, como exemplo, na agricultura, na produção de energia, no transporte e no desenvolvimento de políticas públicas, etc. (FIORIN; ROSS, 2015).

Possibilitando aos gestores públicos e privados regular ou mitigar os impactos climáticos na economia (SILVA; JARDIM, 2019).

O resultado da sistematização por estatística descritiva das variáveis climáticas, podem ser potencializados com a adoção de análise exploratória de dados (SILVA, JARDIM, 2019). Haja vista, que tal metodologia facilita a interpretação dos resultados obtidos nas análises estatísticas cuja aplicação em séries históricas de dados meteorológicos promovem a interatividade intermediada por representações visuais, como gráficos de linhas, de área, histogramas, diagramas de caixa (box plot), entre outras categorias de esboços. Assim sendo, em séries históricas é fundamental a introdução de mecanismos visuais para identificação do comportamento dos dados ao longo do tempo.

Desta maneira, a utilização de módulos descritivos e sistemas web contribui para o conhecimento das características climáticas, como exemplo a região de Matupá – MT e Sinop - MT, que se destacam na produção agropecuária, com a produção de bovinos e milho (IBGE, 2020). A utilização de sistemas web possibilitam a potencialização da assertividade nas tomadas de decisões em diversas áreas de atuação humana, uma vez que facilita a compreensão das características do tempo e clima de uma determinada região, necessitando apenas de um conjunto confiável de dados cronológicos de estações meteorológicas.

Mediante o exposto, este artigo tem objetivo de descrever novas funcionalidades de um sistema (web) o PAP Meteor. Cujas funcionalidades limitam em identificar e corrigir erros e falhas em série histórica de dados meteorológicos. Este capítulo demonstra o funcionamento das novas incrementações do sistema supracitado, apresentando os módulos desenvolvidos para estatística descritiva e análise exploratória (nesta ordem) além de explorar a distribuição temporal de temperatura e precipitação na base de dados do INMET de Matupá MT e Sinop MT.

Material e Métodos

Características das novas implementações do sistema (Web) PAP Meteor

O sistema continua com uma “interface” simples, (Figura 1) com um guia de orientação ao usuário e um campo para entrada de dados.

Figura 1: ‘Interface’ do sistema

As ações do usuário para acesso as novas implementações consistem em selecionar a base de dados e escolher as ações do sistema:

a) Estatística descritiva: neste módulo o usuário escolhe as variáveis que deseja aplicar estatística descritiva (medidas de tendência central, dispersão e correlação de Pearson). Podendo também optar por sumarização de todo conjunto de dados, retornando a contagem, a média, o desvio padrão, o mínimo, o máximo e quartis (25,50 e 75%).

b) Explorar dados: nesta etapa o usuário pode escolher duas opções:

I. Explorar base: o usuário escolhe a variável que deseja verificar com o sistema, retornando gráficos de valores médios, mínimos e máximos agrupados por ano e meses. Além de tabela com os respectivos dados plotados nos gráficos. Os gráficos são interativos, possibilitando o usuário filtrar períodos específicos através de botões e mecanismo de arrastar e soltar (“drag drop”),

além de poder fazer transferência dos dados e das figuras plotadas.

O funcionamento dos módulos depende da inserção de dados no formato especificado nas orientações disponibilizados pelo sistema.

Cada módulo foi testado aplicando dados do Município de Matupá MT e Sinop MT, de modo que, nos resultados e discussões apresentam tabelas e gráficos gerados pelo sistema.

Coleta de dados

Para teste e validação do sistema foi utilizado dados históricos da estação meteorológica convencional de superfície no Município de Matupá MT e Sinop MT, pertencente ao Instituto Nacional de Meteorologia (INMET). A estação de Matupá MT está registrada com código 8314 da Organização Mundial de Meteorologia (OMM), localizada nas coordenadas geográficas de latitude -10.1916°, longitude -54.9461° e altitude de 272 metros cujos dados são diários, com 2 registros ao dia 0h a 12h e 12h a 18h. A estação de Sinop MT está localizada na Latitude -11,98, longitude -55,57 e altitude de 366,57 metros, com código da OMN A917, cujos dados são horários das 0h às 23h.

Foram utilizados dados históricos de temperatura (máxima e mínima) e precipitação.

Os dados originais disponibilizados pelo INMET, apresenta registros entre os anos de 1987 a 2020, iniciando no dia 01/01/1987 até 31/12/2020 para o munício de Matupá MT, já para Sinop MT os dados iniciam em 01/01/2009 a 31/12/2020. No entanto, apesar de disponibilizar uma série relativamente longa, os dados possuem falhas (ausência de registros) além de dados discrepantes. Tais anormalidades foram possíveis de serem observadas utilizando os módulos de estatística descritiva e análise exploratória do sistema apresentado neste trabalho.

De modo a avaliar o comportamento do sistema supracitado e as características climáticas dos referidos municípios, os dados brutos foram submetidos aos módulos de análise descritiva e análise exploratória. Posteriormente, os dados brutos foram corrigidos e imputado registros ausentes com o método de K-vizinhos mais próximos (KNNIMPUTER) de Troyanskaya *et al.*, (2001) implementado na 1° versão do PAP Meteor. A partir

dos dados devidamente corrigidos, também foi submetido as novas funcionalidades do sistema, assim como aos dados brutos. Deste modo, é possível comparar a distribuição temporal de ambos os conjuntos de dados (original e corrigido).

Resultados e Discussão

O fluxo dos resultados e discussões acompanha os módulos desenvolvidos do sistema. A ordem de apresentação são os módulos de estatística descritiva e análise exploratória.

Estatística descritiva

A Tabela 1 apresenta um resumo estatístico descritivo das variáveis meteorológicas de Matupá MT, comparando os resultados dos dados brutos com corrigidos. No Município de Matupá, a média de temperatura mínima diária é de 20,16 °C e a de temperatura máxima 32,08 °C. A variável temperatura mínima apresentava registro mínimo de 1,4 °C e a versão corrigida apresenta 7,5 °C. Tal correção foi aplicada analisando as datas com registros inferiores a 7,5 °C considerado incomum para a região, no qual se verificou que esses registros ocorreram em meses com incidência de altas temperaturas. Esta, discrepância pode ser verificada no módulo de análise exploratória do sistema com a plotagem de histograma e diagrama de caixa (boxplot). Além de acompanhamento do comportamento de temperatura através de gráfico de linha da série temporal agrupado por meses. A mesma anomalia aconteceu no município de Sinop, com registro de 4,6 °C as 12h, considerando que os registros antecedentes e posteriores estavam em média 24 °C. Tais, discrepâncias são comuns em séries históricas, devido falha nos instrumentos, ações antrópicas e outras categorias de interferências (BIER; FERRAZ, 2017).

Tabela 1: Estatística básica das variáveis climáticas para o município de Matupá - MT, com base em registros entre 1987 a 2020.

	Temp. Máxima (°C)		Temp. Mínima (°C)		Precipitação (mm)		Insolação (horas)	
#	Originais	Corrigidos	Originais	Corrigidos	Originais	Corrigidos	Originais	Corrigidos
Média	32,79	32,08	20,06	20,17	5,08	5,93	6,06	5,04
Mediana	33	33,4	20,06	20,57	0	4,77	6,4	3,69
Moda	33,2	34,4	21,8	20	0	0	0	0
Mínimo	9,6	15,10	1,4	7,5	0	0	0	0
Máximo	40,2	40,2	29,0	27,1	198,4	198,4	11,7	11,7
Desvio Padrão	2,68	2,62	2,41	1,76	12,79	8,38	3,33	2,90

Fonte: INMET (2020)

Análise exploratória

No período de 1987 a 1989 houveram lacunas nos registros de precipitação no município de Matupá - MT (Figura 2A). Devido à grande quantidade de dados ausentes neste período a imputação de dados pode não ser confiável ou afetar valores médios das variáveis (Figura 2B). O maior número de falhas está nos registros das 0h às 12h (80%). No entanto, as imputações podem chegar a 100% de acurácia, se os dados forem transformados em apenas um registro ao dia atribuindo uma média diária para as variáveis (Figura 2C).

As falhas nos dados afetam a média das variáveis, principalmente precipitação nos meses secos (JARDIM; SILVA, 2019). No entanto, o preenchimento das falhas pelo PAP Meteor é eficiente, haja vista que, utiliza a distância euclidiana entre os vizinhos mais próximos (algoritmo KNNImputer). Considerando que o sistema agrupa os anos com maior integridade nos registros e classifica os dados em horas, mês e dia, contribuindo para uma menor distância entre os vizinhos. Consequentemente, as estimativas são mais fidedignas, tanto para séries horárias (0h as 23:00) (Figura 3B) como para série diária. Porém deve se ter atenção ao utilizar o

PAP Meteor em séries com dois ou três registros ao dia, já que a quantidade exagerada de falhas nestes registros afeta as médias, diárias, decendiais, mensais e anuais das variáveis, recomendado nestes casos, transformar os dados em um único registro diário (Figura 2B).

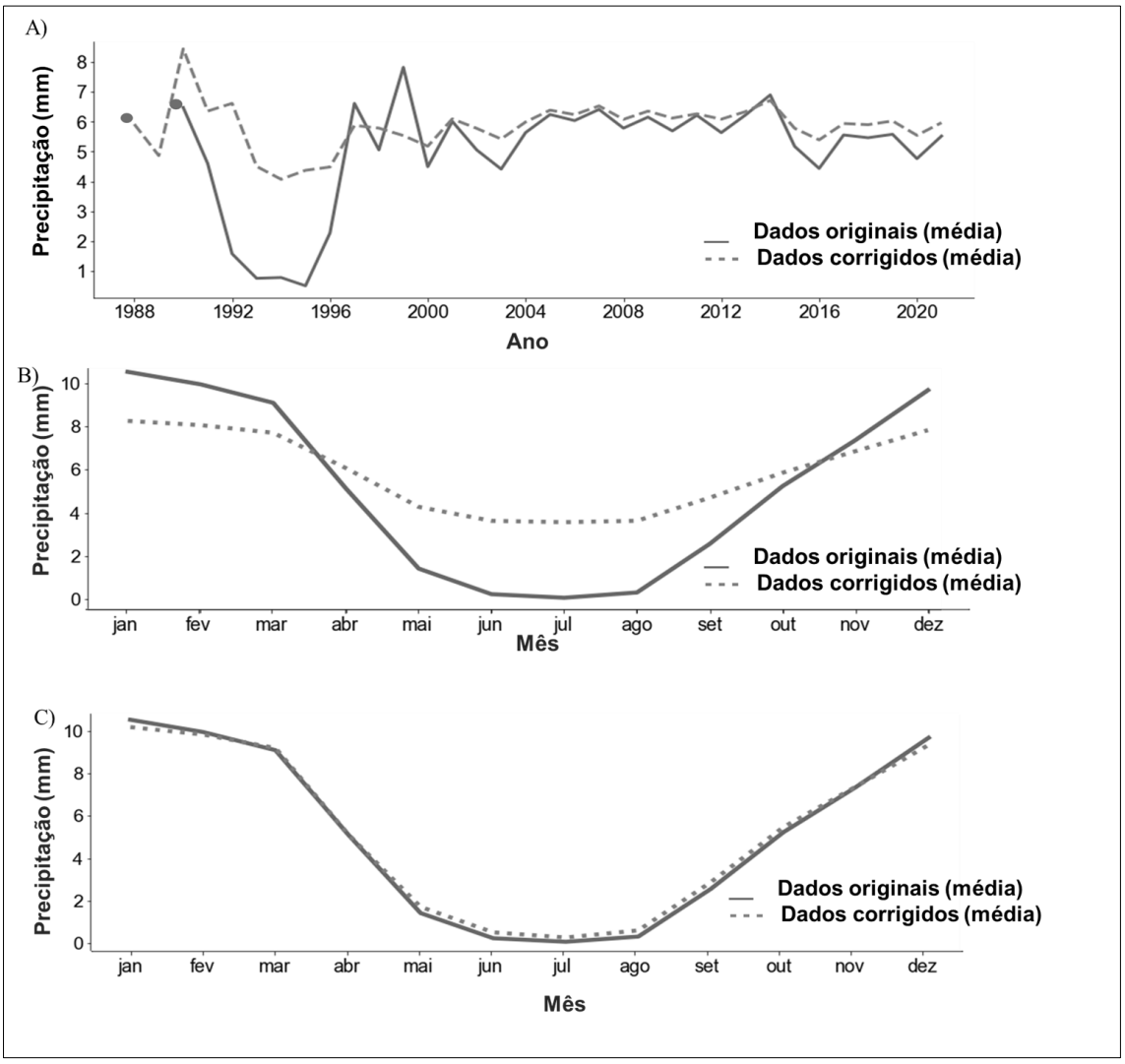


Figura 2: Registros de precipitação média diária no município de matupá – MT, no período de 1987 a 2020. Em a) representa as lacunas nos registros de precipitação ao longo dos anos, em b) Média diária de precipitação no decorrer dos meses com dois (2) registros diários, em c) Média diária de precipitação no decorrer dos meses com um (1) registros diários.

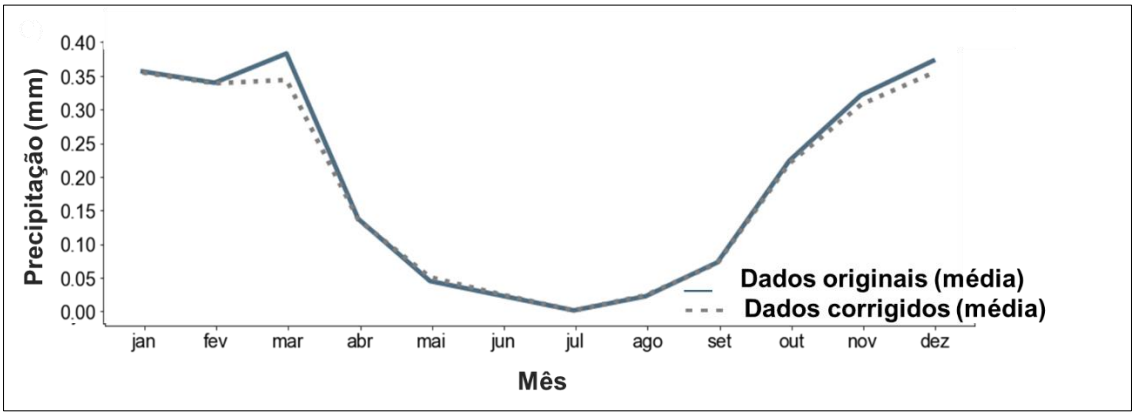


Figura 3: Média horária de precipitação no município de Sinop – MT, no período de 2009 a 2020

Os municípios de Matupá e Sinop apresentam uma estação seca (março a setembro) e chuvosa (outubro a abril) (Figura 2B e 3), com Sinop apresentando uma média anual de 1.622 mm ao ano. Tais médias também foram observadas nos trabalhos de Ramos et. al (2017) e de Marcuzzo; Melo; Rocha (2011), apontando médias anuais de até 2.200 mm no norte do estado.

Para temperatura mínima na base de Matupá, a série histórica apresenta alguns valores atípicos nos anos 1997, 2000, 2014 e 2019 (linha seccionada Figura 4A), chegando a registrar 1,4 °C. Tal anomalia foi tratada (linhas contínuas Figura 4A). Deste modo, os meses com menores registros é junho, julho e agosto, sendo o mês de julho o mais frio, com temperaturas médias de 18,6 °C. As temperaturas mais baixas concentram no sudeste do Mato Grosso com os registros mínimos variando entre 16 a 18 °C (RAMOS et al., 2017).

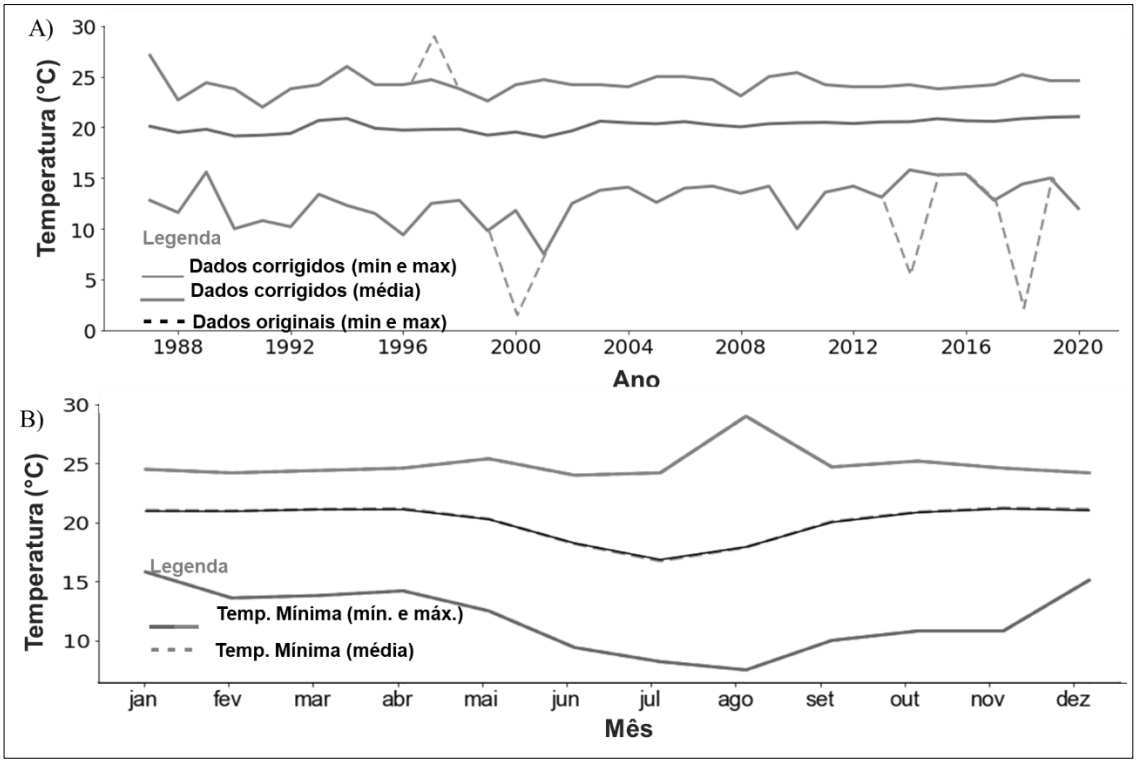


Figura 4: Distribuição de temperatura mínima no município de Matupá – MT, em A) Erros nos registros, em B) Distribuição mensal de temperatura mínima no período de 1986 a 2020.

Em Matupá, nos anos de 2011 e 2019 registraram as maiores temperaturas com 40,2 °C (Figura 5A). A média anual de temperatura máxima é de 32,7° C, já os meses mais quentes é julho, agosto e setembro. Sendo, portanto, o mês de agosto o mais quente, com temperatura média de 34,8 °C (Figura 5B).

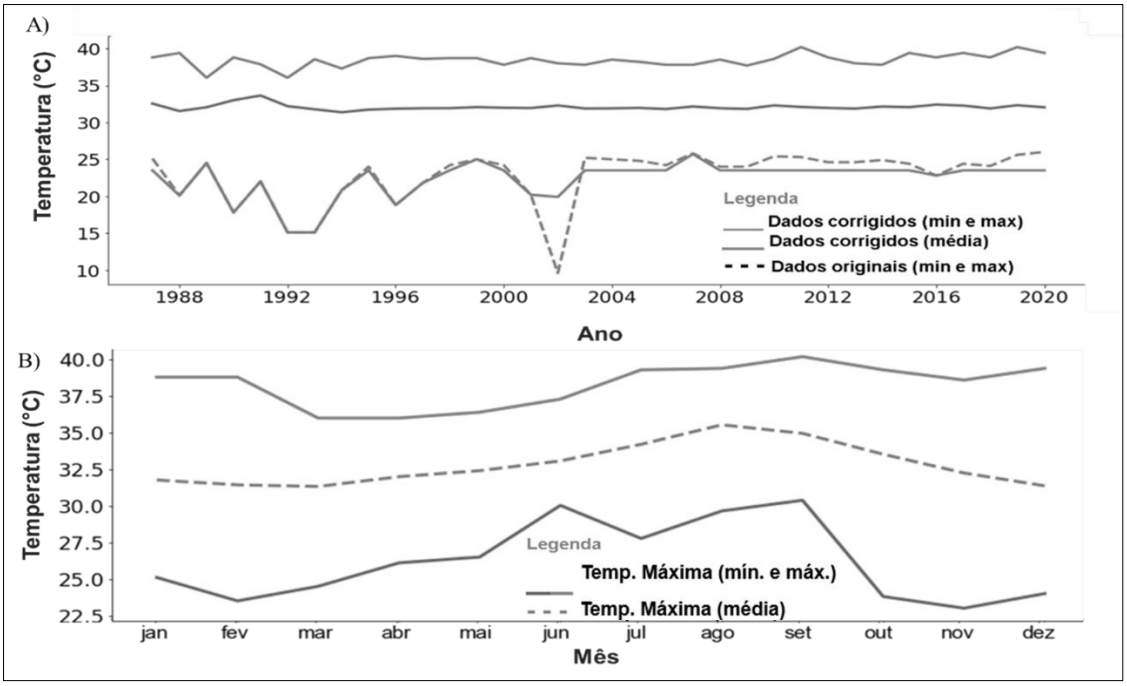


Figura 5: Registros de temperatura máxima no município de Matupá – MT, no período de 1987 a 2020

Em Sinop de 2009 a 2020, os anos mais quentes foram 2014, 2015 e 2019 com média anual próxima aos 27,0 °C, alcançando picos de até 40,0 °C (Figura 6A). Os meses mais quentes são agosto, setembro e outubro, sendo setembro o mês mais quente com média de 28,5 °C, atingindo até 40,0 °C (Figura 6B).

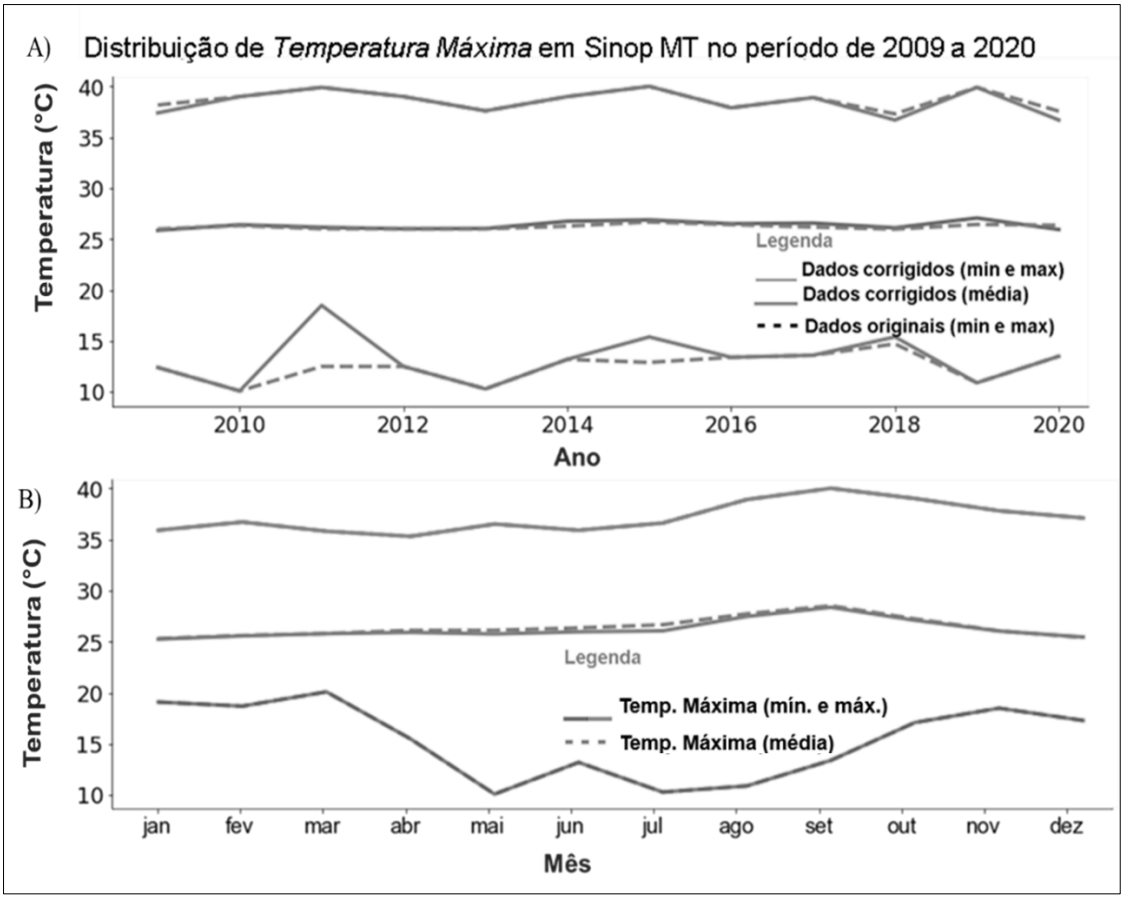


Figura 6: Registros de temperatura máxima no município de Sinop – MT, no período de 2009 a 2020,

Embora Matupá e Sinop localizam na região norte do Mato grosso, a dinâmica da temperatura se comporta de formas diferentes, considerando que as condições de relevo, vegetação, uso e ocupação do solo são distintas nos municípios, além de estarem separados com ~204 km um do outro e Sinop com ~84 metros a mais de altitude (GOMES; SANTOS, 2001). Tais fatos contribuem para diferenças nas temperaturas, considerando ainda que Sinop está mais ao sul da região norte, onde a temperatura e precipitação vai diminuindo gradativamente a medida que se avança ao sul do Mato Grosso (RAMOS *et al.*, 2017).

Conclusão

Os módulos do PAP Meteor, demonstram as características básicas do tempo em uma série histórica, de modo a facilitar e identificar padrões nas variáveis meteorológicas.

O PAP Meteor por ser um sistema web, facilita a sumarização e interpretação de dados meteorológicos. Apresentando resultados satisfatórios, independente da organização dos dados (horário e diário).

Quanto a dados ausentes, séries com dois a três registros diários, se apresentarem muitas falhas em algum de seus horários, afeta a média das variáveis. Sendo recomendado neste caso, atribuir uma média diária ou usar dados horários (apresenta melhor assertividade). Erros nos dados também é um fator que afeta as médias diárias ou horárias.

As temperaturas do município de Matupá MT oscilaram entre 10 °C a 40 °C com médias anuais de 33 °C. Sendo agosto o mês com maior temperatura e julho o mês com menor temperatura. Quanto a chuvas, é predominante nos meses de janeiro a abril e de outubro a dezembro.

Em Sinop MT as temperaturas também variam de 10 °C a 40°C, sendo o mês de julho o mais frio e setembro o mais quente. A precipitação também é predominante nos meses de janeiro a abril e de setembro a dezembro.

Nos conjuntos de dados originais e corrigidos não houve grandes diferenças nas médias das variáveis.

O sistema é eficiente e cumpre com seu escopo, sumarizando e facilitando a interpretação de dados meteorológicos.

Referências Bibliográficas

AKAMINE, Carlos Takeo; YAMAMOTO, Roberto Katsuhiro. **Estudo Dirigido de Estatística Descritiva**. 3° edição ed. São Paulo SP: [s.n.], 2013. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9788536517780/cfi/27!/4/2@100:0.00. Acesso em: 19 jun. 2020.

BELINDA, Pereira Da Cunha; AUGUSTIN, Sérgio. **Sustentabilidade ambiental: estudos jurídicos e sociais**. I ed. Caxias d0 Sul: EDUCS – Editora da Universidade de Caxias do Sul, 2014. v. I

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. Disponível em: http://www.scielo.br/scielo.php?pid=S0102-77862017000200215&script=sci_abstract&tlng=pt. Acesso em: 6 abr. 2020.

IBGE, Instituto Brasileiro de Geografia e Estatística. **Matupá MT**: economia cidades. Economia cidades. 2020. Disponível em: https://www.ibge.gov.br/cidades-e-estados/mt/matupa.html. Acesso em: 23 out. 2020.

JARDIM, C. H.; AION ANGELU FERRAZ SILVA. Aplicação de técnicas de preenchimento de falhas de dados de pluviosidade mensal e anual para o noroeste do estado de Minas Gerais - Brasil. **Revista Geografias**, *[S. l.]*, v. 25, n. 2, p. 83–106, 2019. Disponível em: https://periodicos.ufmg.br/index.php/geografias/article/view/16058. Acesso em: 22 mar. 2021.

VIANNA, Luiz Fernando De Novaes *et al.* Bancos de Dados Meteorológicos: Análise dos Metadados das Estações Meteorológicas no Estado de Santa Catarina, Brasil. **Revista Brasileira de Meteorologia**, p. 53–64, 2017. DOI 10.1590/0102-778632120150119. Disponível em: http://dx.doi.org/10.1590/0102-778632120150119. Acesso em: 30 nov. 2020.

FIORIN, Tatiana Tasquetto; ROSS, Meridiana Dal. **Climatologia Agrícola** . Santa Maria-RS: [s.n.], 2015. Disponível em: http://estudio01.proj.ufsm.br/cadernos_fruticultura/terceira_etapa/arte_climatologia_agricola.pdf. Acesso em: 8 abr. 2020.

GOMES, Marco Antonio Villarinho; SANTOS, Mário Vital dos. **Zoneamento sócio-econômico-ecológico: Diagnóstico socioeconômico do Estado de mato Grosso e assistência técnica na formulação da 2°:** aspectos das formações vegetais/ uso e ocupação do solo - folha mir-320 ⁝ sinop ⁝ memória técnica. Cuiabá: Cnec, 2001. 49 p.

MARCUZZO, Francisco F N; MELO, Denise C R; ROCHA, Hudson M. Distribuição Espaço-Temporal e Sazonalidade das Chuvas no Estado do Mato Grosso. **Revista Brasileira de Recursos Hídricos**, v. 16, p. 157–167, 2011. Disponível em: http://rigeo.cprm.gov.br/jspui/handle/doc/18875. Acesso em: 17 nov. 2020.

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. 20, n. 0, p. 2237–8642, 1 ago. 2017. DOI 10.5380/abclima.v20i0.43762. Disponível em: https://revistas.ufpr.br/revistaabclima/article/view/43762. Acesso em: 17 nov. 2020.

SILVA, Aion Angelu Ferraz; JARDIM, Carlos Henrique. Técnicas de estatística descritiva e análise exploratória na caracterização têmporo-espacial da pluviosidade da região de Unaí-MG. jun. 2019. **Anais** [...]. Fortaleza CE: [s.n.], jun. 2019. Disponível em: http://www.editora.ufc.br/images/imagens/pdf/geografia-fisica-e-as-mudancas-globais/983.pdf. Acesso em: 19 jun. 2020.

TROYANSKAYA, Olga *et al.* Missing value estimation methods for DNA microarrays. **Bioinformatics**, v. 17, n. 6, p. 520–525, fev. 2001. Disponível em: https://academic.oup.com/bioinformatics/article-abstract/17/6/520/272365. Acesso em: 6 maio 2020.