

UNIVERSIDADE FEDERAL DE MATO GROSSO
INSTITUTO DE FÍSICA
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA AMBIENTAL

**GUIA PARA QUALIDADE: DO TRATAMENTO AO
ARMAZENAMENTO DOS DADOS AMBIENTAIS**

RAPHAEL PIRES FERREIRA

**ORIENTADOR: PROF. DR. PAULO HENRIQUE ZANELLA DE
ARRUDA**

Cuiabá, MT
Maio – 2021

UNIVERSIDADE FEDERAL DE MATO GROSSO
INSTITUTO DE FÍSICA
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA AMBIENTAL

**GUIA PARA QUALIDADE: DO TRATAMENTO AO
ARMAZENAMENTO DOS DADOS AMBIENTAIS**

RAPHAEL PIRES FERREIRA

*Tese apresentada ao Programa de Pós-graduação
em Física Ambiental da Universidade Federal de
Mato Grosso, como parte dos requisitos para
obtenção do título de Doutor em Física Ambiental.*

Orientador: Prof. Dr. Paulo Henrique Zanella de Arruda

Cuiabá, MT
Maio – 2021

Dados Internacionais de Catalogação na Fonte.

P667g Pires Ferreira, Raphael.
GUIA PARA QUALIDADE: DO TRATAMENTO AO ARMAZENAMENTO
DOS DADOS AMBIENTAIS / Raphael Pires Ferreira. -- 2021
75 f. ; 30 cm.

Orientador: Paulo Henrique Zanella de Arruda.
Tese (doutorado) - Universidade Federal de Mato Grosso, Instituto de Física,
Programa de Pós-Graduação em Física Ambiental, Cuiabá, 2021.
Inclui bibliografia.

1. Qualidade de dados. 2. Variáveis ambientais. 3. Armazém de dados. I. Título.

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Permitida a reprodução parcial ou total, desde que citada a fonte.



MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE MATO GROSSO
PRÓ-REITORIA DE ENSINO DE PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA AMBIENTAL

FOLHA DE APROVAÇÃO

TÍTULO: GUIA PARA QUALIDADE: DO AO ARMAZENAMENTO DOS DADOS AMBIENTAIS

AUTOR: DOUTORANDO RAPHAEL PIRES FERREIRA

Tese defendida e aprovada em **12 de maio de 2021**.

COMPOSIÇÃO DA BANCA EXAMINADORA

1. Prof. Dr. Paulo Henrique Zanella de Arruda – Orientador – Instituto de Física/UFMT
2. Prof. Dr. Denilton Carlos Gaio – Examinador Interno – Instituto de Física/UFMT
3. Prof. Dr. Raphael de Souza Rosa Gomes - Examinador Interno - Instituto de Computação - UFMT
4. Prof. Dr. João Paulo Ignácio Ferreira Ribas - Examinador Externo - Instituto de Computação – UFMT
5. Prof. Dr. Prof. Dr. Michael Jacques Lathuilliere - Examinador Externo - Stockholm Environment Institute (SEI), Stockholm, Sweden

Cuiabá-MT, 12/05/2021.



Documento assinado eletronicamente por **SERGIO ROBERTO DE PAULO, Coordenador(a) de Programas de Pós-Graduação em Física Ambiental - IF/UFMT**, em 12/05/2021, às 19:58, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **RAPHAEL DE SOUZA ROSA GOMES, Docente da Universidade Federal de Mato Grosso**, em 12/05/2021, às 20:44, conforme horário oficial de Brasília, com fundamento

no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **JOAO PAULO IGNACIO FERREIRA RIBAS, Docente da Universidade Federal de Mato Grosso**, em 12/05/2021, às 20:52, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **DENILTON CARLOS GAIO, Docente da Universidade Federal de Mato Grosso**, em 12/05/2021, às 22:29, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Michael Jacques Lathuillière, Usuário Externo**, em 13/05/2021, às 05:22, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **PAULO HENRIQUE ZANELLA DE ARRUDA, Docente da Universidade Federal de Mato Grosso**, em 13/05/2021, às 16:40, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



A autenticidade deste documento pode ser conferida no site http://sei.ufmt.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3492048** e o código CRC **4E37B885**.

DEDICATÓRIA

À Deus, à minha esposa Maria
Daniele e à minha família.

AGRADECIMENTOS

- À Deus, por proporcionar mais essa oportunidade em minha vida;
- À minha esposa Maria Daniele, que me incentiva e acredita em meu potencial e me inspira a sempre continuar crescendo;
- Ao meu orientador, professor Dr. Paulo Henrique Zanella de Arruda, pela paciência e incentivo durante o doutorado;
- Ao professor Dr. José de Souza Nogueira (Paraná), pelo trabalho que realizou no programa de Pós-Graduação em Física Ambiental e pela parceria durante esses anos de estudo;
- Ao professor Dr. Sérgio Roberto de Paulo (Serginho), pelo trabalho que realiza no programa de Pós-Graduação em Física Ambiental;
- Aos professores que ministraram as disciplinas no Programa de Pós-graduação em Física Ambiental com imensa dedicação;
- Ao Prof. Dr. Raphael de Souza Rosa Gomes pela disponibilidade e auxílio nas sugestões de melhoria da tese;
- Ao Dr. Michael Lathuillière pelas contribuições na melhoria da tese;
- Ao Prof. Dr. Denilton Carlos Gaio, pela dedicação, parceria e auxílio na melhoria da tese;
- Ao Prof. Dr. João Paulo Ignácio Ferreira Ribas, pela sua parceria nos anos de estudo e contribuições de melhoria da tese;
- Aos colegas que colaboraram durante todo o curso e contribuíram de forma direta e indireta a esse trabalho;
- À Soilce e ao Cesário que estão sempre dispostos a colaborar e me auxiliaram durante esses anos com os assuntos administrativos do programa;

SUMÁRIO

LISTA DE FIGURAS.....	VII
LISTA DE TABELAS.....	VIII
LISTA DE ABREVIACÕES E SIGLAS.....	IX
RESUMO.....	X
ABSTRACT.....	XI
1. INTRODUÇÃO.....	1
2. REVISÃO BIBLIOGRÁFICA.....	5
2.1. PRINCIPAIS REQUISITOS PARA DADOS CLIMATOLÓGICOS.....	5
2.2. DIMENSÕES DE QUALIDADE.....	6
2.3. DISPONIBILIDADE DE ACESSO AOS DADOS.....	8
2.4. INTEGRIDADE DOS DADOS.....	8
2.5. BANCO DE DADOS RELACIONAL.....	9
2.5.1. ARMAZÉM DE DADOS.....	10
2.6. Visualização de dados.....	11
3. MATERIAL E MÉTODOS.....	13
3.1. ÁREAS DE ESTUDO.....	13
3.1.1. PANTANAL.....	13
3.1.2. CERRADO.....	15
3.2. INDICADORES DE QUALIDADE.....	16
3.3. MÉTODOS DE VALIDAÇÃO DA QUALIDADE DE VARIÁVEIS AMBIENTAIS.....	19
3.3.1. VALORES FISICAMENTE IMPOSSÍVEIS OU CLIMATOLOGICAMENTE INCONSISTENTES.....	20
3.3.2. VERIFICAÇÃO DE <i>OUTLIERS</i>	20
3.3.3. VALORES ZERADOS OU IGUAIS CONSECUTIVOS.....	22
3.3.4. IDENTIFICAÇÃO DOS PERÍODOS QUE FALTAM DADOS.....	23
3.4. ESTRUTURA PARA ARMAZENAMENTO E DISPONIBILIZAÇÃO DOS DADOS.....	24
4. RESULTADOS.....	26
4.1. VALIDAÇÕES DE QUALIDADE DE VARIÁVEIS AMBIENTAIS.....	26
4.1.1. CASO DE ESTUDO COM DADOS DO PANTANAL.....	26
4.1.2. CASO DE ESTUDO COM DADOS HISTÓRICOS DO INMET.....	28
4.2. APLICAÇÃO DOS INDICADORES DE QUALIDADE.....	38

4.3. ESTRUTURA DE ARMAZENAMENTO E DISPONIBILIZAÇÃO DOS DADOS	
39	
5. CONCLUSÃO.....	46
REFERÊNCIAS BIBLIOGRÁFICAS.....	49
APÊNDICE.....	56

LISTA DE FIGURAS

FIGURA 1 – Fontes de dados climatológicos	10
FIGURA 2 - Localização da torre microclimática do SESC Pantanal	14
FIGURA 3 - Classificação de Koppen-Geiger para o Brasil	15
FIGURA 4 – As três variáveis com mais indicações de problemas para valores zerados.....	33
FIGURA 5 – As três variáveis com mais indicações de problemas para <i>outliers</i>	33
FIGURA 6 – As três variáveis com mais indicações de problemas para falhas de leitura.....	34
FIGURA 7 - Outliers por região e ano.....	35
FIGURA 8 - Valores zerados por região e ano	36
FIGURA 9 - Falhas de leitura por região e ano	37
FIGURA 10 - Modelo estrela para dados climatológicos.....	40
FIGURA 11 - Rotina de carga no DW	41
FIGURA 12 – Diagrama estrutural da PADC	42
FIGURA 13 – Média diária da temperatura do ar em °C.....	43
FIGURA 14 – Média diária das temperaturas do ar e do solo em °C	43
FIGURA 15 – Média horária da temperatura do ar em °C	44
FIGURA 16 - Exemplo de painel construído com as variáveis climatológicas.....	45

LISTA DE TABELAS

TABELA 1 - Dimensões de qualidade.....	6
TABELA 2 - Indicadores de qualidade (DQI)	17
TABELA 3 - Classificação dos DQI.....	18
TABELA 4 - Resultado de qualidade dos conjuntos analisados.....	26
TABELA 5 - Quantidade de registros da verificação de <i>outliers</i>	27
TABELA 6 - Resultado das análises de qualidade para os dados do INMET	28
TABELA 7 - Resultado das validações de qualidade por região e variável ambiental	29
TABELA 8 – Indicadores de Qualidade para o Pantanal e Cerrado	38
TABELA 9 - Indicadores de Qualidade para as regiões do Brasil.....	39

LISTA DE ABREVIACÕES E SIGLAS

DQI – Indicadores de Qualidade

PADC – Plataforma de Acesso a Dados Climatológicos

DW – Armazém de dados

PPGFA – Programa de Pós-graduação em Física Ambiental

GEE – Gases do efeito estufa

IPCC – Painel Intergovernamental sobre Mudanças Climáticas

SQL – Linguagem de consulta estruturada

INMET – Instituto Nacional de Meteorologia

LCA – Avaliação de ciclo de vida

GHCN – Rede Climatológica Histórica Global

ETL – Extração, transformação e carga

OSD – Base de dados de código aberto

NoSQL – Base de dados não convencionais

RESUMO

FERREIRA, R. P. Guia para qualidade: do tratamento ao armazenamento dos dados ambientais. Cuiabá, 2021, 73f. Tese (Doutorado em Física Ambiental) - Instituto de Física, Universidade Federal de Mato Grosso.

A obtenção de dados climatológicos através de sensores automáticos tem se tornado cada vez mais trivial, seja pela simplicidade de acessar portais que fornecem tais informações, ou pela compra e instalação de tais equipamentos. Porém, a massificação desses processos apresenta alguns desafios, a necessidade de se verificar e garantir a qualidade dessa grande quantidade de dados, bem como utilizar técnicas que garantam a disponibilidade desses dados para que as pesquisas sejam realizadas com maior precisão. Nesse sentido, diante da necessidade de bases confiáveis para a obtenção dos dados, esta pesquisa tem como objetivo sistematizar métodos que possibilitem avaliar a qualidade de dados ou conjunto de dados de variáveis ambientais, como a temperatura do ar, a temperatura do solo, dentre outras. Com a aplicação de forma automatizada das rotinas de qualidade, nota-se um ganho notável de eficiência para o pesquisador, por meio das validações, assim como, com a criação das técnicas em ambiente R, conforme apresentado nos resultados. A utilização desses métodos mostrou que melhorias em sensores do Instituto Nacional de Meteorologia (INMET) devem ser feitas, pois, vários dados apresentaram problemas de qualidade. Destaca-se também a criação da plataforma de código aberto, denominada Plataforma de Acesso a Dados Climatológicos (PADC). Esta plataforma visa garantir os conceitos de estruturação de bases de dados em formato de armazém de dados (DW), apresentando as garantias de um ambiente dedicado exclusivamente à entrega de dados, onde a inserção de novas informações é rigorosamente controlada. Por meio dessa estrutura obtém-se as tradicionais garantias de bases relacionais, como o conceito de facilidade de estruturas de tabela de dados e técnicas de backups tradicionais podem ser utilizadas. Garantindo assim, mais tempo para que o produto da pesquisa seja realizado de forma mais eficaz. Ademais, o painel dinâmico de dados abertos possibilita maior visibilidade, transparência e personalização para análises mais rápidas e aprimoradas. Dessa forma, a qualidade dos dados é preservada, não só em técnicas que funcionam no contexto dos dados, mas também na forma de seu armazenamento de longo prazo, garantindo que essas informações não serão alteradas. Salienta-se vários benefícios quando da utilização da plataforma, dentre eles ganho de tempo, quando permite dedicação à metodologia do trabalho e não necessariamente na preparação de dados. Vislumbra-se também resultados com menor viés e, portanto, mais confiáveis.

Palavras-chave: Qualidade de dados, Variáveis ambientais, Armazém de dados

ABSTRACT

FERREIRA, R. P. Guide to quality: from treatment to storage of environmental data. Cuiabá, 2021, 73f. Thesis (PhD in Environmental Physics) - Institute of Physics, Federal University of Mato Grosso.

Obtaining climatological data through automatic sensors has become increasingly trivial, whether due to the simplicity of accessing portals that provide such information, or by purchasing and installing such equipment. However, the massification of these processes presents some challenges, the need to verify and guarantee the quality of this large amount of data, as well using techniques that guarantee the availability of this data so that the research can be carried out with greater precision. In this sense, given the need for reliable bases for obtaining the data, this research aims to systematize methods that make it possible to assess the quality of data or data set of environmental variables, such as air temperature, soil temperature, among others. A notable gain in efficiency with these validations for the researcher is perceived with the automated application of quality routines, as well as with the creation of techniques in an R environment, as shown in the results. The use of these methods showed that improvements in sensors of the National Institute of Meteorology (INMET) should be made, because several data presented quality problems. Also noteworthy is the creation of the open source platform, called “Plataforma de Acesso a Dados Climatológicos” (PADC). This platform aims to guarantee the concepts of structuring databases in a data warehouse (DW) format, presenting the guarantees of an environment dedicated exclusively to data delivery, where the insertion of new information is strictly controlled. Through this structure, traditional guarantees of relational bases are obtained, as the concept of ease of data table structures and traditional backup techniques can be used. Thus, ensuring more time for the research product to be carried out more effectively. In addition, the dynamic panel of open data enables greater visibility, transparency and customization for faster and improved analysis. In this way, the quality of the data is preserved, not only in techniques that work in the context of the data, but also in the form of its long-term storage, ensuring that this information will not be changed. Several benefits are highlighted when using the platform, including time savings, when it allows dedication to the work methodology and not necessarily in the preparation of data. It is also possible to see results with less bias and therefore more reliable.

Keywords: Data quality, Environmental variables, Data warehouse

1. INTRODUÇÃO

Dados climáticos acessíveis e de alta qualidade têm aplicação generalizada, como monitorar a variabilidade e as mudanças climáticas, apoiar decisões em torno de riscos naturais, gerenciamento de riscos, redução de desastres e subsidiar previsões e projeções climáticas (AUSTRALIAN BUREAU OF METEOROLOGY; CSIRO, 2011), auxiliando em políticas públicas. Os dados assumem o papel de uma "matéria-prima" ilimitada e reutilizável para posterior processamento, criando aplicações que geram valor agregado, lucro e novos empregos (MINISTRY OF INTERIOR, 2015).

No Brasil, existem várias estações que medem e disponibilizam dados meteorológicos como precipitação, velocidade e direção do vento, umidade relativa do ar, pressão, etc. Só o Instituto Nacional de Meteorologia (INMET) administra mais de 400 estações, sendo que possui 10 Distritos Regionais que recebem, processam e enviam estes dados para a sede, localizada em Brasília-DF. A sede, por sua vez, processa estes dados e os envia por satélite para todo o mundo (INMET, 2021).

Com a facilidade na obtenção de dados ambientais, através da utilização de sensores e equipamentos para medições, é possível obter de forma rápida acesso a grandes quantidades de dados (SCIUTO et al., 2013). Estes dados estão disponíveis em diversos portais de informações sobre pesquisas ambientais ou portais governamentais da área ambiental e são coletados por meio de sensores instalados em diversas localidades, pode-se citar aviões de alta altitude, satélites, torres micrometeorológicas, entre outros.

Diante da grande quantidade de dados gerados para possibilitar realizar análises climatológicas, é fundamental que se tenham padrões e formas de verificação de sua qualidade. Com isso em mente, o controle de qualidade deve ser feito antes de qualquer tipo de análise, procurando eliminar quaisquer erros de leitura ou vieses não climatológicos no conjunto submetido a tal procedimento (STEPANEK et al., 2009). Porém, validações sobre a qualidade dos dados e seus metadados raramente estão disponíveis para pesquisadores (HARYOKO, 2012).

Pesquisas ambientais, como na área de mudanças climáticas, dependem não só dessa grande quantidade de dados, mas também que esses dados tenham qualidade, ou seja, a validação de qualidade dos dados é crucial para pesquisadores de toda parte (BOULANGER et al., 2010, ALEXANDERSSON; MOBERG 1997, CAUSSINUS; MESTRE 2004, RUSTICUCCI; BARRUCAND, 2004). O impacto de erros dessa natureza nas análises climatológicas ainda é amplamente desconhecido (HUNZIKER et al., 2018). No contexto das mudanças climáticas, os dados climáticos históricos definem o leque da variabilidade climática experimentada, além de fornecer contexto para a interpretação das mudanças climáticas projetadas para o futuro (JONES et al., 2013).

O problema com qualidade de dados não é uma exclusividade do mundo corporativo. Sadiq (2013) apresenta um caso do meio acadêmico/científico onde o sistema estabelecido em Nova Orleans, nos Estados Unidos da América, para avisar sobre furacões, falhou porque estava incompleto e inadequado, pois foi construído de forma desconexa durante muitas décadas utilizando dados de elevação desatualizados. Ou seja, não adianta apenas possuir muitos dados em uma enorme base sem nenhum tipo de trabalho qualitativo. Quando não se tem esse controle, o dado errado trabalhado torna o trabalho inútil, pois o resultado será incorreto.

Ao trabalhar com dados, o pesquisador tem a necessidade de garantir a qualidade de um conjunto de dados, mas é necessário que saiba como essa qualidade está sendo auferida e se algum evento importante não foi suprimido pelos procedimentos de qualidade (DURRE et al., 2008 e SCIUTO et al., 2013).

A validação da qualidade dos dados de temperatura é de extrema importância para a maioria das pesquisas ambientais, visto que essa variável é utilizada em diversos estudos sobre as alterações climáticas (FREE et al., 2005 e THORNE et al. 2005). O levantamento de dados precisos, de longos períodos, sobre a temperatura do ar ainda é um desafio para as pesquisas atmosféricas (MEIER et al., 2016).

A avaliação da qualidade de dados também é necessária com a popularização de sensores ou estações automáticas de leituras, que cada vez mais disponibilizam esses dados de forma automática na rede mundial de computadores, porém, sem nenhum

tipo de tratamento de qualidade. A utilização desses dados levanta questionamentos sobre sua qualidade (BELL et al., 2015), bem como quais os tipos de validação são possíveis de se realizar (MEIER et al., 2016).

Nos resultados das análises automáticas algumas inconsistências podem ser identificadas, como falsos positivos, o que necessitaria de um profissional com mais experiência na área para analisar e validar essas informações, porém, esse esforço já seria menor (SCIUTO et al., 2009), uma vez que o conjunto completo passou pela avaliação automática antes.

As análises e pesquisas realizadas nos últimos anos mostram que será cada vez mais necessário cruzar dados coletados de diferentes fontes (MARTIN et al., 2015), a fim de demonstrar algum aspecto ou tendência que não teria sido possível antes, devido à dificuldade em se obter tais informações. Como, por exemplo, a dificuldade em se entender uma possível evasão escolar em uma região em determinadas épocas do ano devido à alta temperatura e as salas que não possuem aparelhos de ar-condicionado. Assim, é possível entender a necessidade de possuir bases de dados sólidas e confiáveis para a realização de pesquisas.

Especificamente, com respeito ao Programa de Pós-Graduação em Física Ambiental (PPGFA), que possui equipamentos instalados em vários pontos do estado de Mato Grosso (MT), como no pantanal mato-grossense (Torre Baía das Pedras) e no cerrado (Torre da Fazenda Miranda), os dados são armazenados sem uma estratégia clara de padronização pré-estabelecida, o que pode dificultar sua disponibilidade e uso, bem como, não apresenta um plano para aferição da qualidade dos dados gerados pelos sensores.

O problema da falta de verificação da qualidade dos dados produzidos pelo PPGFA pode ser percebido em uma análise dos trabalhos disponíveis no site do programa. Verificando vinte e um trabalhos, que possuíam em seu título alguma informação sobre variáveis ambientais, como radiação solar, temperatura do ar, temperatura do solo, ou o termo micro meteorológico. Dessa forma, foi encontrado em apenas sete trabalhos (33%) a informação de algum controle de qualidade aplicado aos dados da pesquisa.

Diante do exposto, este estudo tem como objetivo geral apresentar um guia com detalhamento da sistematização de métodos e uma plataforma para que os dados de variáveis ambientais possam ser tratados para garantir a sua qualidade, bem como estabelecer princípios básicos para o seu armazenamento e disponibilização de forma aberta.

Para isso, objetivos específicos foram traçados:

1. Introduzir uma avaliação de qualidade aplicável a variáveis ambientais;
2. Apresentar a implementação de métodos para realizar a avaliação de qualidade dos dados de temperatura. Essa implementação será baseada na linguagem R, que consiste em um software livre para estatística computacional e gráficos, podendo ser executado em diversos ambientes Linux, Windows ou MacOS (R CORE TEAM, 2018);
3. Utilizar os métodos para verificar a qualidade dos dados da base histórica do INMET e de conjuntos disponibilizados pela equipe técnica do PPGFA;
e
4. Desenvolver e apresentar uma plataforma, utilizando apenas softwares de código aberto, desde o banco de dados para o armazenamento até uma ferramenta para fornecer acesso aos dados para pesquisadores interessados em realizar pesquisas, com a possibilidade de estender isso a outras fontes de dados disponíveis na internet, fornecendo um ambiente de dados abertos, para que qualquer pessoa possa acessar.

2. REVISÃO BIBLIOGRÁFICA

2.1. PRINCIPAIS REQUISITOS PARA DADOS CLIMATOLÓGICOS.

A climatologia evoluiu ao longo do tempo; não é apenas uma área que armazena dados, mas também uma que oferece serviços e respostas rápidas (SLATYER; BONNER, 1996). Dessa forma, melhores práticas para o gerenciamento de dados climáticos devem ser o ponto de partida para análises básicas e complexas (MARTIN et al., 2015).

Muitos locais, como instituições de pesquisa ou universidades, que são produtores de dados não se preocupam com a verificação de qualidade desses dados. Os dados e os metadados devem passar por um processo de qualidade, e isso é um fator importantíssimo (DÜSTERHUS; HENSE, 2012). Informações a respeito dos dados, como a sua confiabilidade, incertezas, integridade, idade e validade, são justamente informações de qualidade que precisam ser verificadas (JAYAWARDENE, SADIQ, INDULSKA, 2013 e WEIDEMA; WESNAES, 1996). Uma grande quantidade de dados com dúvidas sobre a sua qualidade gera uma desconfiança sobre o real valor desse conjunto de dados (SADIQ; INDULSKA, 2017).

Problemas com a qualidade de dados de estudos são questionados há algum tempo, conforme Lim e Boileau (1999) já levantam esse problema. Os inventários de gases do efeito estufa (GEE) que são feitos passam por esse questionamento e essa preocupação foi levada a especialistas em reuniões do Painel Intergovernamental sobre as Mudanças Climáticas (IPCC), e uma das primeiras propostas para resolução disso foi de se realizar uma avaliação sobre a qualidade dos dados gerados e agrupados para construção dos inventários (LIM; BOILEAU, 1999).

No Brasil, uma grande dificuldade é percebida mesmo nos momentos iniciais da coleta de dados, onde, muitas vezes, alguns dos requisitos básicos para a obtenção de uma rede de informação estruturada e confiável não são respeitados, como a falta de estações posicionadas a distâncias pré-estabelecidas para obter algumas validações de

qualidade em referências entre si. Muito se deve à falta de financiamento nacional, em muitos países, em que a ajuda externa, como o financiamento estrangeiro, é a única maneira de manter uma rede com dados meteorológicos viáveis (MARTIN et al., 2015), como certos projetos precisam.

Com essas dificuldades, o que pode ser feito é pensar em mecanismos práticos e tangíveis que qualquer tipo de projeto climatológico de armazenamento de dados possa aproveitar e garantir uma qualidade aceitável.

2.2. DIMENSÕES DE QUALIDADE

Sob o olhar de aspectos centrais na área de qualidade de dados, estão as dimensões de qualidade, com estudos em desenvolvimento há muito tempo, como Garvin (1987) que apresentou oito dimensões de qualidade para a avaliação de um produto; outro exemplo seria o trabalho de Russel e Taylor (2003), onde apresentaram dimensões de qualidade para tratar dados de um serviço de atendimento. Porém é importante ressaltar que os estudos da qualidade de dados é área transversal em qualquer esfera, seja ela acadêmica ou empresarial, uma vez que estipular métricas de qualidade e manter essas medições são imprescindíveis para qualquer pesquisa. Com isso, o domínio de qualidade de dados tem como fator crucial e fundamental as suas dimensões (BATINI et al., 2009; JAYAWARDENE et al., 2013). A Tabela 1 apresenta as dimensões de qualidade.

TABELA 1- Dimensões de qualidade

Dimensões	Definições
Acessibilidade	disponibilidade e facilidade de conseguir os dados
Quantidade de Dados Apropriados	a quantidade de dados é apropriada para as atividades
Credibilidade	veracidade dos dados

Integridade	sem falta de dados, atende a granularidade proposta
Representação Concisa	dados são objetivos
Representação Consistente	dados seguem sempre o mesmo formato
Facilidade de Manipulação	os dados são fáceis de manipular e aplicar
Livre de Erros	dados corretos e confiáveis
Interpretabilidade	definições estão claras, unidades, símbolos são apropriados
Objetividade	dados sem viés e imparciais
Relevância	dados são aplicáveis
Reputação	modo de geração tem aceitação
Segurança	acesso aos dados é seguro
Temporalidade	os dados não são antigos para sua aplicação
Entendimento	os dados são compreensíveis
Valor agregado	Dados são benéficos e apresentam vantagens para o seu uso

FONTE: Adaptado de Jayawardene et al. (2013)

É fundamental estabelecer medidas e adaptações de técnicas de qualidade de dados para a área ambiental, visto que, a qualidade dos dados acaba sendo preterida, ou não abordada. Na literatura, é possível encontrar diversas técnicas para se avaliar essa qualidade (BATINI et al., 2009), como as dimensões de qualidade propostas por JAYAWARDENE et al. (2013), ou técnicas baseadas em regras de negócio.

Carmo et al. (2016) utilizam técnicas de *visual analytics* (VA) para verificar a qualidade de dados ambientais gerados para o Sistema Integrado de Monitoramento Ambiental (SIMA). Nota-se então que a visualização dos dados pode ser uma ferramenta para acompanhar e auxiliar a qualidade de dados.

2.3. DISPONIBILIDADE DE ACESSO AOS DADOS

O acesso a dados abertos tornou-se uma questão importante em muitos níveis, pode-se alcançar transparência, conscientização pública, etc. (JAROLÍMEK; MARTINEC, 2016). Esse acesso é um ponto chave para novas descobertas e a construção de novos conhecimentos (ATENAS et al, 2015).

Alcançar um alto grau na disponibilidade de dados significa que qualquer pessoa possa realizar suas pesquisas sem qualquer barreira (LEE, 2006), sem se preocupar sobre como acessar determinada base de dados, ou como configurar um conjunto confiável de sensores em campo. Ao se garantir alto grau de disponibilidade de dados, é possível estimular a criação de conhecimento e impulsionar os trabalhos que buscam responder a grandes problemas globais.

2.4. INTEGRIDADE DOS DADOS

O termo integridade é autoexplicativo, ou seja, pode-se inferir que a integridade é garantida quando algo é o que deveria ser (GILBERT et. al., 2004), no entanto, vários autores lidam com a integridade dos dados, cada um com sua visão e definição, referenciados sobre qualidade, segurança, mudanças e fluxo de informações entre diferentes entidades. Com o foco estabelecido na qualidade e nas definições mais gerais, a integridade seria um parâmetro importante para trabalhar na qualidade dos dados (COURTNEY; WARE, 1989).

A integridade é fundamental para qualquer tipo de banco de dados. Uma vez armazenado, é necessário garantir que esse conjunto não sofra qualquer tipo de alteração, não armazene dados indesejados ou receba dados inesperados (IMRAN, et. al., 2017 e MCDOWALL, 2019), o que prejudica a apresentação e o uso corretos para análise. Portanto, é necessário estruturar o banco de dados para ser resiliente contra falhas, usando espelhamento, backup, verificação de falhas, etc. Tudo isso com uma boa tipagem de dados, ou seja, o tipo do dado deve ser definido para que não ocorra

problemas no seu armazenamento e as ferramentas de consulta a esses dados possam retornar o que eles realmente são, um número inteiro, ou número decimal com a precisão de casas definidas pela equipe de negócio, ou mesmo um texto ou sigla.

2.5. BANCO DE DADOS RELACIONAL

O banco de dados relacional ainda é a principal escolha quando se trata de armazenar dados para sistemas de informação. A prática na Agência Australiana de Meteorologia mostrou que a entrada e o gerenciamento de dados são mais eficientes quando os dados meteorológicos observados são armazenados em estruturas de tabelas relacionais (MARTIN et al., 2015).

A estrutura relacional apresenta ganhos em relação a outros modelos, como a fácil compreensão, tanto pelo programador quanto pelo usuário final, de que os valores são armazenados em tabelas e são recuperados pelas chamadas desses objetos (CODD, 1982), a linguagem de consulta estruturada (SQL) em si é simples de entender. Possui em sua base de operações álgebra relacional, que fornece todos os tipos de operações possíveis com SQL. A Figura 01 apresenta uma lista de fontes de dados climatológicos disponíveis.

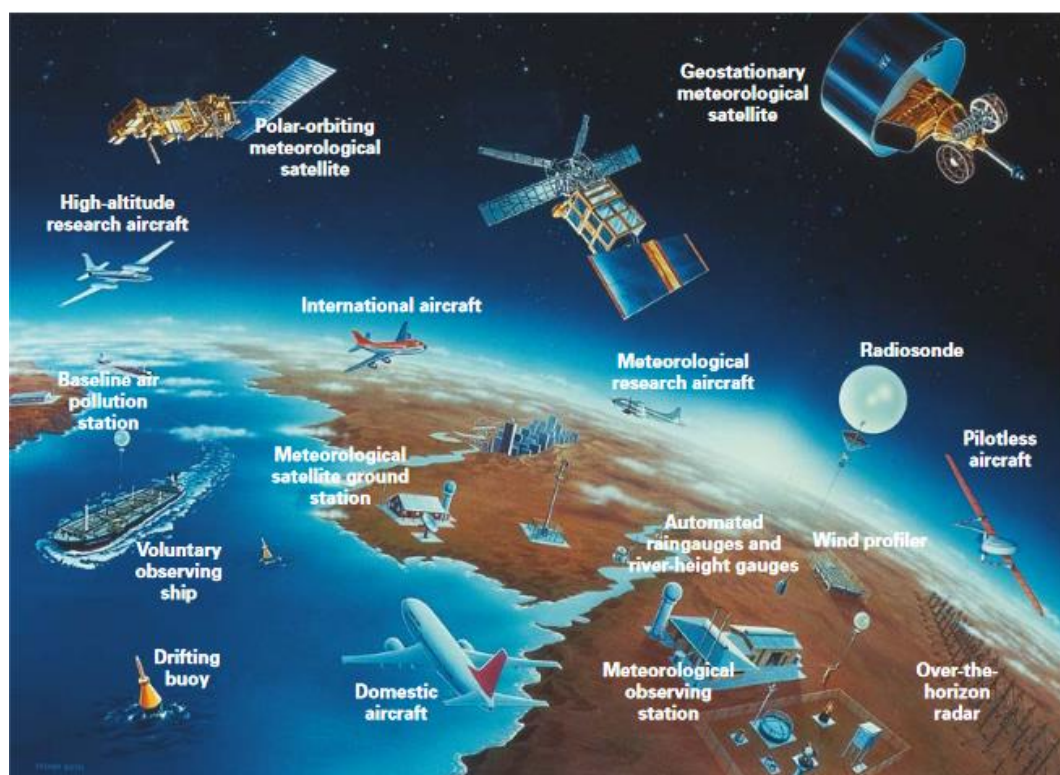


FIGURA 1 – Fontes de dados climatológicos
 Fonte: WMO (2014)

2.5.1. ARMAZÉM DE DADOS

Ao implementar um projeto de banco de dados, temos algumas alternativas na estruturação básica, e uma delas é a construção de um armazém de dados (DW). Nele, os diferentes dados gerados em fontes heterogêneas podem ser armazenados e usados em conjunto.

Uma evolução na forma de armazenamento de dados é necessária. O simples fato de armazená-los pode não atender às demandas futuras que a pesquisa precisa, como a capacidade de cruzar dados de diferentes fontes geradoras (INMON, 2002).

Conforme KIMBALL (2002), as características fundamentais para se ter um armazém de dados são:

1. O conteúdo de um DW deve ser compreensível, com os dados sendo intuitivamente óbvios para o usuário e o processo de obtê-los pode ser feito a partir de várias fontes diferentes;
2. As informações devem ser apresentadas de forma consistente, ou seja, os dados devem ser cuidadosamente assegurados de diversas fontes e a sua qualidade verificada;
3. O DW deve ser adaptativo e resiliente a mudanças, qualquer mudança não pode afetar os dados já presentes do armazém;
4. O armazém de dados deve garantir a segurança das informações contidas;
5. Deve apresentar os dados certos para a tomada de decisão, ou seja, os dados de interesse para organização devem estar presentes nele;
6. A comunidade deve realmente utilizar a solução, de nada adianta apenas a solução implantada, mas que não seja utilizada.

Sendo assim, o design para um armazém de dados pode seguir, principalmente, duas abordagens, uma está associada com o design Inmon e a segunda com o design Kimball (DEDIC; STAINER, 2017).

Kimball (2002) detalha que a abordagem associada a Inmon, existe a necessidade de normalizar todas as tabelas na preparação dos dados na terceira forma normal, para somente após isso finalizar o processo de extrair, transformar e carregar (ETL), que seria carregar os dados para a área de apresentação do DW, tornando assim o processo de ETL muito mais dispendioso e demorado para apresentar os dados finais e comumente e este é o maior ponto de fracasso de equipes que desenvolvem armazém de dados.

2.6. VISUALIZAÇÃO DE DADOS

A importância de tentar facilitar a visualização e o entendimento em diversas matérias, que envolvem a representação de dados, é remetida a antiguidade. A primeira representação gráfica conhecida da informação quantitativa é um gráfico de cerca de 950ac, são séries múltiplas anônimas mostrando a mudança de posição dos sete corpos celestiais mais proeminentes ao longo do espaço e do tempo (FRIENDLY, 2009).

Porém, foi apenas a partir do século 17 que Rene Descartes criou a representação gráfica de dados quantitativos em relação ao plano de coordenada bidimensional (AZZAM et al., 2013).

Conforme Azzam et al. (2013), a visualização de dados está pautada em três princípios:

1. É baseado em um conjunto de dados qualitativos ou quantitativos;
2. Resulta em uma imagem que representa um conjunto de dados;
3. Deve ser interpretável pelos espectadores, permitindo a sua análise.

A matéria de visualização de dados encontra na era moderna, principalmente com a massificação de computadores, um grande espaço para inovações e para auxiliar no entendimento de dados. Essa ferramenta auxilia também na questão de transparência, tanto no âmbito público como no privado, onde, a sociedade passa a cobrar uma maior facilidade na visualização e entendimento sobre dados publicados.

Existem várias ferramentas que já possibilitam explorar e visualizar os dados de forma gráfica, como Microsoft Power BI[®], Apache[®] Superset, IBM Cognos[®], entre outros.

A importância de possuir os dados disponíveis já em visualização gráfica podem ser constatados atualmente por meio dos vários sites na internet¹ que monitoram os dados sobre o avanço da pandemia de COVID-19 e mantém toda a população informada de uma maneira prática e de rápido entendimento.

¹ <https://www.bing.com/covid/local/brazil?cc=br> e <https://covid.saude.gov.br/>

3. MATERIAL E MÉTODOS

3.1. ÁREAS DE ESTUDO

3.1.1. PANTANAL

A região pantaneira é um local único para realização de estudos e pesquisas na área ambiental. A região é um bioma diferenciado no qual possui a maior concentração de fauna das Américas com características de outros biomas como: o Cerrado, o Chaco (ou Bosque Chiquitano), a Amazônia e a Mata Atlântica e se liga a duas bacias hidrográficas de importância transfronteira, a Amazônia e a do Prata que contribui para a ampliação das várias espécies da fauna e da flora (Almanaque Brasil, 2008; SANTOS, 2018).

O Pantanal é uma imensa área que faz parte da bacia do alto do rio Paraguai, cobrindo uma área de aproximadamente 140,000 km² (TONDATO, MATEUS; ZIOBER, 2010), sendo uma das maiores planícies sedimentares do mundo, com cotas altimétricas entre 80 e 150 metros (JUNK et al., 2005). A alternância entre estações chuvosas e secas define o clima como sazonal, apresentando variabilidade plurianual, *i. e.*, ciclos alternados de anos muito úmidos e muito secos (HAMILTON et al., 1996).

Um dos conjuntos de dados utilizado para validação dos métodos de qualidade foi coletado por sensores instalados na torre de análise microclimática na Reserva Particular do Patrimônio - RPPN SESC - Pantanal, no município de Poconé (Torre Sesc) (BARBOSA, 2016), conforme é apresentado na Figura 2.



FIGURA 2 - Localização da torre microclimática do SESC Pantanal

Fonte: BARBOSA (2016)

O clima da região, segundo a classificação Climática de Koppen-Geiger, é do tipo Aw, como é possível confirmar na Figura 4, que corresponde a invernos secos e verões chuvosos (CURADO, 2013; ALVARES et al, 2014). A temperatura média anual apresenta variações entre 22°C e 32°C (HASLAM et al, 2003).

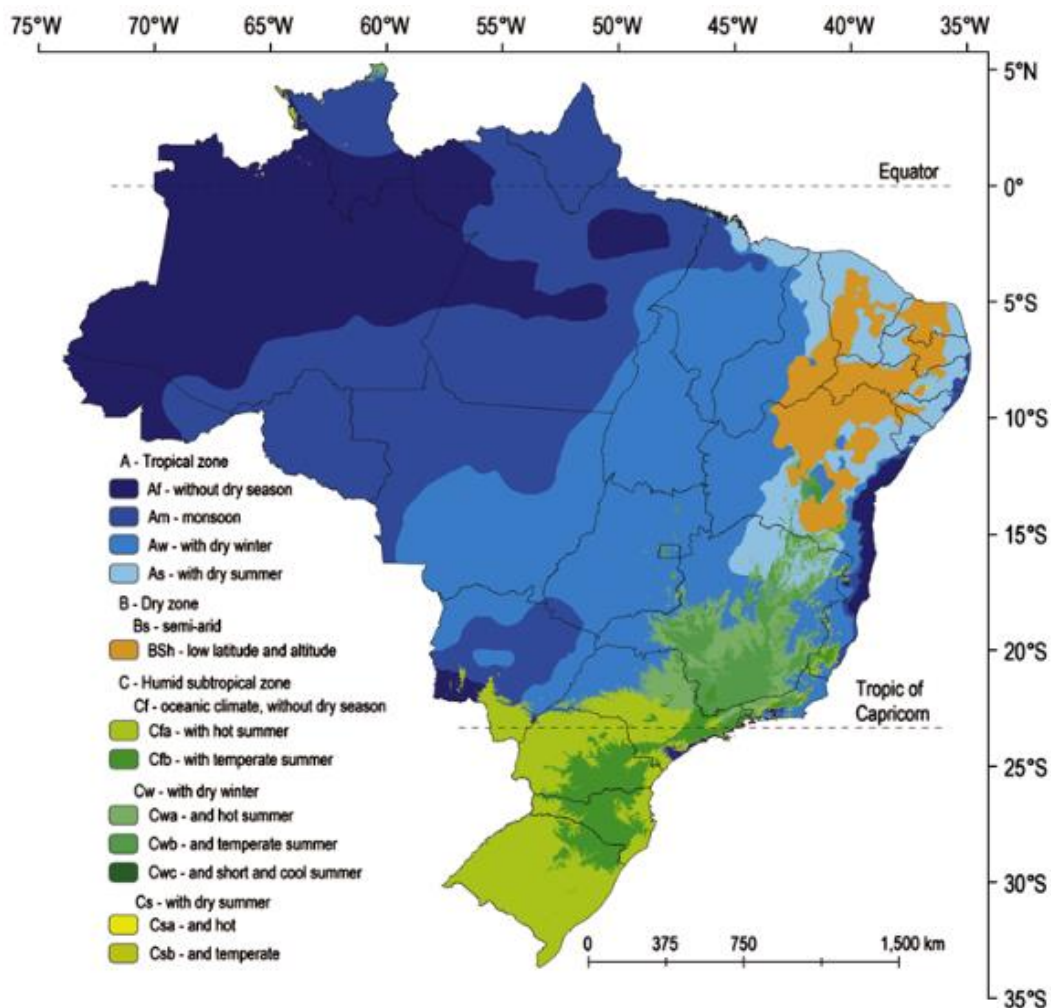


FIGURA 3 - Classificação de Köppen-Geiger para o Brasil

FONTE: ALVARES et al, 2014

3.1.2. CERRADO

O Cerrado é atualmente o bioma brasileiro com as maiores taxas de conversão de habitats naturais em terras agrícolas (LAPOLA et al. 2013). Conforme SILVA (2018), a combinação dessa escassez de dados e altas taxas de conversão de habitat representa um grande desafio para a conservação dos serviços ecossistêmicos dessa região.

Com a alta taxa de exploração pelo homem do bioma do Cerrado, estudos precisam ser implementados para avaliar a extensão dessas intervenções nos ciclos naturais, bem como para entender o impacto dessas ações.

Apenas 2,96% da vegetação natural do Cerrado é encontrada dentro de unidades de conservação integral (FRANÇOSO et al. 2015), uma proporção muito menor do que a das florestas brasileiras, com 17,5% totalmente protegidos (MMA 2009; GIBBS et al. 2015; OVERBECK et al. 2015).

Dessa forma, é possível entender a importância desse bioma para a região mato-grossense e brasileira, visto a alta demanda humana. E maiores quantidades de dados com qualidade são guias básicos para construção de novos conhecimentos sobre o bioma.

Seguindo a tendência da região em que se localiza, a Classificação de Köppen-Geiger é dada como Aw, ou seja, possui inverno seco e verão chuvoso (ALVARES et al, 2014), como é possível verificar na Figura 4.

3.2. INDICADORES DE QUALIDADE

Uma das técnicas de tratamento de dados seria a aplicação de indicadores de qualidade de dados (DQI), que é recomendada nos casos de avaliação de ciclo de vida (LCA), (BICALHO, SAUER, RAMBAUD; ALTUKHOVA, 2017; FINNVEDEN, 2000; JOHNSON, 2003; WEIDEMA, WESNAES, 1996). Essa avaliação irá dizer sobre os aspectos dos dados que mais podem influenciar os estudos de LCA (BICALHO et al., 2017) de modo que se pode traçar um paralelo com os dados de variáveis ambientais, utilizando-se dos DQI. Como nas LCA, os dados de variáveis ambientais servem como base para a criação de estudos e decisões de políticas públicas, como mitigação de GEE ou estratégias para evitar desastres naturais. A partir da criação de uma matriz de qualidade, esses dados podem ser verificados na sua origem, proporcionando aos estudos posteriores um nível de credibilidade maior diante de uma análise de qualidade padronizada.

A necessidade então se encontra em definir quais dimensões serão utilizadas para formar a matriz com os indicadores de qualidade. A necessidade da opinião

especialista, nessa etapa, é evidente, uma vez que o conhecimento necessário para se estipular o nível de qualidade passa por um parecer subjetivo de avaliação. E essa matriz irá apresentar um índice de qualidade como exposto na Tabela 2.

TABELA 2 - Indicadores de qualidade (DQI)

Indicadores de Qualidade	Descrição
Acessibilidade	Disponibilidade e facilidade de conseguir os dados
Credibilidade	Como esses dados estão sendo gerados e a sua coleta é realizada de forma satisfatória;
Temporalidade	O período de representação dos dados;
Consistência	Esses dados são únicos para cada uma das leituras e o formato da saída é o esperado;
Integridade	Os dados não apresentam valores fora do esperado;

Fonte: Elaboração própria

Com o estabelecimento dos DQI, deve-se definir quais são os objetivos de qualidade que se espera atingir com cada um desses indicadores, ou seja, esses objetivos devem apresentar de forma clara as suas necessidades com cada um dos DQI (EDELLEN; INGWERSEN, 2016).

Com relação a credibilidade dos dados, procura-se mensurar para que o utilizador saiba como está o processo de geração e coleta, podendo garantir que os resultados estarão de acordo com o esperado. Já, na temporalidade, será apresentado o tempo de disponibilidade desses dados. A consistência passa para o usuário o índice da garantia de que os dados que foram gerados no equipamento são os mesmos que ele está trabalhando. A integridade reflete a capacidade do equipamento em realizar medidas dentro do intervalo fisicamente possível.

Já sobre a acessibilidade, utiliza-se uma classificação internacional para validar o nível de disponibilidade e facilidade de obtenção dos dados em análise. Essa classificação sugerida por Berners-Lee (2006) é a *5-star open data*, onde, quanto mais fácil e acessível esse conjunto de dados é, maior é a sua pontuação.

Logo, a Tabela 3 apresenta as pontuações, com as respectivas classificações dos DQI.

TABELA 3 - Classificação dos DQI

Pontuação dos DQI	1 – Ótimo	2 – Bom	3 – Regular	4 – Ruim	5 - Pésimo
DQI					
Credibilidade	Sensores de alta precisão e coleta automatizada	Sensores de média precisão e coleta automatizada ou manual	Sensores de baixa precisão e coleta automatizada	Sensores de baixa precisão e coleta manual	Sensores sem especificação de precisão
Temporalidade	Mais de 3 anos de dados	Entre 1 e 3 anos de dados	Até 1 ano de dados	< de 1 ano de dados	< de 6 meses de dados
Consistência	95-100% dados	85-94% dados	50-84% dados	49-30% dados	<30% dados
Integridade	95-100% dados	85-94% dados	50-84% dados	49-30% dados	<30% dados
Acessibilidade	Dados disponíveis ² e ligados a outros dados	Dados disponíveis por um URI ³	Dados disponíveis em um formato aberto	Dados disponíveis de forma estruturada	Dados disponíveis sem padronização ou indisponíveis

É importante que os resultados da análise de qualidade sejam demonstrados de forma clara e objetiva para quem está utilizando os dados, em sua versão final de disponibilização. Desta forma, a análise de qualidade se torna útil. (FOKEN et al., 2004). Por isso, a necessidade de se aplicar as pontuações após a construção e aplicação dos DQI.

² Nesse contexto, “Dados disponíveis” é aquele acessível de forma digital.

³ É um identificador uniforme de recurso, possibilitando que esse recurso possa ser utilizado pela internet.

3.3. MÉTODOS DE VALIDAÇÃO DA QUALIDADE DE VARIÁVEIS AMBIENTAIS

O desenvolvimento dos métodos para avaliar a qualidade dos dados deve passar por algumas validações básicas (DURRE et al. 2008; SCIUTO et al. 2013), como verificação de valores fisicamente impossíveis ou climatologicamente inconsistentes com a área estudada, verificação na ocorrência de *outliers*, verificar se existem zeros consecutivos em um determinado intervalo de tempo, validação dos valores máximos e mínimos e valores iguais consecutivos.

Várias pesquisas procuram entregar métodos capazes de auferir a qualidade de dados ambientais (NAPOLY et al., 2018, DURRE et al., 2008, MEIER et al., 2016, SHEN et al., 2012) e com todas essas pesquisas é possível elencar algumas validações que podem ser aplicadas na região do Pantanal Mato-grossense e nas torres de coleta automatizadas do INMET, uma vez que os métodos de controle de qualidade não apresentam uma generalização (STEPANEK et al., 2009) capaz de apresentar resultados satisfatórios com quaisquer tipos de variáveis ou em locais com características diferentes.

Atualmente grande parte dos portais de disponibilização de dados não se preocupa em realizar as atividades de qualidade de dados, deixando a critério dos pesquisadores, ou quem utilizar os dados, essa etapa importante antes de qualquer trabalho.

Bertrand et al. (2013) e Durre et al. (2008), sugerem algumas verificações básicas para o controle de qualidade. Todas as rotinas implementadas para executar as verificações de qualidade das variáveis ambientais estão disponíveis em um repositório público do GitHub⁴.

⁴ Rotinas na linguagem R disponível em: <https://github.com/AceMX/dadosqualidade>

3.3.1. VALORES FISICAMENTE IMPOSSÍVEIS OU CLIMATOLÓGICAMENTE INCONSISTENTES

Trata-se de uma verificação que deve ser feita, levando-se em consideração as características climatológicas do local avaliado. O teste procura verificar se os valores estão dentro de limites aceitáveis para o local estudado (BERTRAND et al., 2013), ou seja, essa verificação deve ser executada com parâmetros de estudos da região analisada.

Como exemplo pode-se citar as variações de temperatura que podem ocorrer no Pantanal mato-grossense, as temperaturas máximas alcançam 45°C e as mínimas são sempre superiores a 5°C. Nesse sentido, os intervalos foram definidos e analisados. Posteriormente, os valores que não estavam dentro do espectro do estudo são destacados, uma vez que esses dados podem comprometer os resultados de qualquer pesquisa.

A rotina foi implementada para receber como parâmetros de entrada um *data frame* (df) que contém as variáveis que serão analisadas com os valores medidos, a indicação numérica para que a rotina saiba a partir de qual coluna deve realizar a comparação de valores e um df que contém os limites físicos das variáveis em análise. Esse segundo df irá servir como a base das comparações da rotina, pois nele estão os limites que o usuário entende como balizadores entre um determinado valor ser considerado algo passível de verificação ou não mais detalhada.

O resultado da rotina será retornado em um df de saída, que irá conter as variáveis informadas para análise com uma coluna adicional para cada variável. Caso nessa coluna o valor apresentado é o número 1, quer dizer o dado analisado ficou fora dos limites físicos, ou seja, está acima ou abaixo desses limites definidos pelo usuário.

Com esse tipo de implementação, o df com as variáveis pode ser do tamanho que o usuário desejar, contendo n variáveis, não limitando a apenas uma variável por análise de limites físicos.

3.3.2. VERIFICAÇÃO DE OUTLIERS

A busca por *outliers* tenta minimizar os erros de leituras que podem estar dentro do intervalo da verificação apresentado na seção anterior, porém não estão condizentes com o conjunto de dados de forma geral. Como exemplo, é possível citar a presença de um determinado valor encontrado que foge dos padrões de leitura dos demais dados no período analisado, podendo identificar alguma ocorrência fora da rotina dos sensores utilizados. Como resultado, esses dados falsos ou interrupções repentinas do sistema de medição podem acontecer, resultando em uma influência negativa nas previsões climáticas (SHEN et al., 2012) que utilizam esses dados.

De forma geral, um valor pode ser considerado um *outlier* quando está distante do desvio padrão do conjunto em mais de três vezes (BODEN et al., 2013; VICKERS e MAHRT, 1997).

O código desenvolvido para implementação dessa verificação está armazenado no repositório público. Trabalhando com uma rotina de laço simples para percorrer todas as colunas do arquivo de dados, a função recebe como parâmetros de entrada o *data frame* que possui as variáveis para análise e também o valor numérico da coluna inicial para análise, isso é necessário para que a rotina não tente analisar dados que não façam sentido, como uma coluna descritiva ou de data. É feito o cálculo da média (*md*) e do desvio padrão (*sd*) da coluna em análise.

Logo em seguida, é feito o cálculo para saber a quantas vezes o dado está do desvio padrão calculado anteriormente. Essa rotina acontece realizando a subtração do dado em análise pela média do conjunto, depois, esse resultado é retornado com o seu valor absoluto positivo. Esse valor positivo é dividido pelo *sd*, por fim, esse resultado é retornado o seu valor inteiro arredondado para cima. Esse cálculo é demonstrado na equação 1.

$$\text{ceiling}(\text{abs}(df - md)/dp) \quad (1)$$

Onde:

ceiling é a função na linguagem R para retornar o valor inteiro arredondando para cima;

abs é função na linguagem R para retornar o valor absoluto positivo.

Ao final do processo a função retorna uma matriz contendo os conjuntos analisados com suas respectivas colunas informando a quantas vezes o dado está do desvio padrão do conjunto analisado.

Com isso em mãos, o pesquisador poderá verificar se aquela informação é condizente com a sua realidade ou se precisará realizar algum ajuste no conjunto devido a alguma falha.

3.3.3. VALORES ZERADOS OU IGUAIS CONSECUTIVOS

A verificação por valores consecutivos zerados ou iguais, pode indicar alguma falha no equipamento de medição, ou em alguma outra parte do sistema responsável pela coleta, armazenagem ou geração dos dados.

Esse tipo de verificação básica deve ocorrer para preparar o conjunto de dados para avaliações mais específicas e que podem ser influenciadas pelos erros básicos. Esse conjunto de validações já se provou efetivo de acordo com Durre et al. (2010), uma vez que esse conjunto de controle de qualidade é aplicado no portal da Rede Climatológica Histórica Global (GHCN).

Esse método também é citado por Boden et al. (2013) como sendo utilizado em uma das etapas de avaliação de dados para o ingresso na rede Ameriflux.

Sendo assim, a rotina criada em R tem como entrada três parâmetros, sendo o primeiro um df que contém os dados a serem analisados, o segundo parâmetro é um df em que o usuário precisa identificar qual valor ele será considerado como zerado para cada uma das variáveis que serão analisadas e o último valor de entrada é um número inteiro para identificar a partir de qual coluna do primeiro df a análise deve iniciar.

Após os parâmetros de entrada fornecidos, a função irá verificar se valores iguais consecutivos são localizados e alimentar um df novo com essa análise, caso algum valor igual seja encontrado, o valor 1 é inserido na respectiva linha do novo df. Após essa análise, a coluna com a variável é analisada novamente em busca de valores zerados, que leva em consideração o parâmetro de entrada definido pelo usuário, identificando se possui valores zerados e a coluna do novo df recebe o valor 1.

Dessa forma, o pesquisador poderá localizar facilmente quais os dados apontados pela rotina como possível falha e poderá atuar para eliminar ou corrigir o dado ou o equipamento, caso seja necessário.

3.3.4. IDENTIFICAÇÃO DOS PERÍODOS QUE FALTAM DADOS

Essa verificação será realizada no conjunto completo para reportar ao interessado quantos dados eram esperados e quantos dados o conjunto realmente possui, caso seja encontrada alguma diferença entre os dados esperados e encontrados, a rotina irá imediatamente identificar quais os dias que estão com uma possível falha na coleta.

Dessa forma, é possível reduzir o tempo de análise do conjunto, que por diversas vezes é maçante e repetitiva de ser executada manualmente, uma vez que o pesquisador já poderá identificar quais os dias que possuem alguma falha de coleta e aplicar algum tipo de preenchimento de falhas ou outra metodologia de correção necessária para sua análise.

O código realiza toda a verificação pelo conjunto de dados. Nessa função é necessário informar a quantidade de medidas esperadas por dia, bem como qual a coluna onde estão as informações da data, como parâmetros de entrada. Com essas informações, a rotina realiza a contagem das leituras do conjunto por dia e compara com a quantidade esperada, que foi informada pelo usuário.

Além das verificações elencadas nessa seção, outros cuidados, como um correto projeto de banco de dados relacional, utilizando de técnicas de normalização, restrições de integridade e controle de transações (CODD, 1970, 1986 e STONEBREAKER e KEMNITZ, 1991) também podem atuar como um controle da qualidade dos dados que serão analisados posteriormente, bem como a utilização de uma estrutura de armazém de dados dedicada a consultas desses dados, uma vez que essas técnicas podem ser aplicadas para qualquer tipo de estrutura relacional.

3.4. ESTRUTURA PARA ARMAZENAMENTO E DISPONIBILIZAÇÃO DOS DADOS

A estrutura de desenvolvimento do escopo proposto pode ser dividida em quatro etapas. A inicial é a definição das fontes de dados disponíveis a serem armazenadas na estrutura do banco de dados. A segunda parte é o pré-processamento que esses dados receberão, ou seja, um processo de extração, transformação e carga (ETL), onde podem ser encontradas inconsistências e notificadas aos técnicos responsáveis, para que possam verificar essas ocorrências e tomar providências, se necessário. A terceira parte é construir o banco de dados onde esse armazenamento ocorrerá. Aqui é necessário um projeto bem definido, garantindo conceitos importantes para a qualidade do armazenamento. E, finalmente, a apresentação desses dados, onde podemos consultar, gerar relatórios e gráficos para monitorar os dados gerados pelas fontes. Esta etapa final, apresentação e interação é altamente importante, pois a satisfação do usuário é uma medida fundamental em uma plataforma orientada a consulta (DEDIC e STAINER, 2017; DAVISON e DEEKS, 2007).

O banco de dados escolhido para a execução do armazenamento foi o Postgres SQL em sua versão 10. É um banco de dados de código aberto com uma estrutura relacional e que possui uma comunidade de suporte ativa e é constantemente atualizado.

O processo de tratamento dos dados para popular o banco de dados pode acontecer de diversas formas, como o tratamento manual pelos técnicos de coleta, pela criação de rotinas automatizadas em ferramentas específicas de ETL, ou até por programações diretamente carregadas nos *dataloggers*.

Como sugestão, a opção pelo tratamento utilizando uma solução livre de ETL é o mais indicado, já que possuem diversas formas de se conectar a diferentes fontes de dados e possuem a capacidade de se conectar com o banco de dados Postgres SQL de forma nativa, na maioria dos casos, fazendo com que os dados tratados já possam ser carregados diretamente no banco de dados final.

Para executar a visualização de dados, a opção de escolha seguirá alguns critérios:

- a) Ser uma ferramenta de código aberto, ou seja, *open source*;

- b) Possuir uma comunidade ativa e com atualizações constantes; e
- c) Ter documentação disponível para todos os processos, da parte técnica de instalação e configuração, até a utilização pelo usuário final

Dessa forma, é possível garantir que a decisão do visualizador dos dados tenha suporte ao longo do tempo de forma gratuita e que a equipe técnica responsável pela sua implantação na instituição consiga manter o seu funcionamento.

4. RESULTADOS

4.1. VALIDAÇÕES DE QUALIDADE DE VARIÁVEIS AMBIENTAIS

4.1.1. CASO DE ESTUDO COM DADOS DO PANTANAL

Foram utilizados dois conjuntos de dados da torre Sesc dos períodos de 2015 (conjunto α) e 2016 (conjunto β) para testes e validações dos métodos de mensuração da qualidade dos dados. Todos os dados são brutos, sem nenhum tipo de verificação prévia aos métodos qualitativos aqui explicados.

Com medições a cada 30 minutos, temos então o conjunto α com 17.520 registros de temperatura do ar. O conjunto β , contendo 1.488 registros, tem a mesma periodicidade de medições do conjunto α , porém o período de medição é menor, ocorrendo apenas durante o mês de janeiro de 2016.

A Tabela 4, com os resultados dos métodos de qualidade propostos, foi construída após a execução das rotinas de validações propostas.

TABELA 4 - Resultado de qualidade dos conjuntos analisados

Conjuntos de dados	IMPS	OUT	ZER	FAL
Conjunto α	0	1x sd ⁵ 11.598	1560	0
		2x sd 5.128		
		3x sd 794		
Conjunto β	0	1x sd 1.099	74	0
		2x sd 309		
		3x sd 80		

⁵ Nessa situação, sd significa desvio padrão.

Legenda: IMPS – Valores fisicamente impossíveis, OUT – Verificação de *outliers*, ZER – Valores zerados ou repetidos e FAL – Identificação dos períodos sem dados

A técnica de validação de valores fisicamente impossíveis foi consistente em todos os conjuntos, não apresentou nenhum resultado possível na análise de todos os registros. Dessa forma, podemos constatar que os valores estão todos dentro do esperado para a região que serviu como objeto de análise para a situação.

Após a execução da validação de busca por *outliers*, a maioria dos dados está dentro do conjunto de 1x do desvio padrão, ou seja, a maioria dos dados apresenta uma alta aderência à qualidade máxima dos conjuntos coletados pelos sensores. Dessa forma, podemos inferir que o trabalho do pesquisador seria avaliar de forma crítica apenas aproximadamente 32% dos dados, quando considerado todos os 3 conjuntos em análise, conforme a Tabela 5 mostra.

TABELA 5 - Quantidade de registros da verificação de *outliers*

Distância do desvio padrão	Quantidade de dados	Percentual
1x	12.697	66,79%
2x	5.437	28,60%
3x	874	4,59%
Total	19.008	100%

A técnica de verificação de valores zerados ou duplicados foi aplicada com sucesso e o resultado demonstra que uma pequena parcela dos dados precisa ser verificada individualmente, uma vez que, apresenta valores repetidos ou zerados, o que pode equivaler a algum problema no sensor de medição, na transmissão dos dados entre o

sensor e o *datalogger* ou até mesmo uma falha na comunicação entre o *datalogger* e o armazenamento final.

Como é possível verificar na Tabela 4, o conjunto α apresenta necessidade de verificação individual de apenas 8,9% do total de dados analisados e o conjunto β foi marcado em 4,9% dos dados.

A última técnica de validação, que busca por períodos sem dados, teve a proposta de identificar falhas de medição através da localização de valores não encontrados, mas que eram esperados. Nessa técnica, os conjuntos α e β não apresentaram nenhum problema.

4.1.2. CASO DE ESTUDO COM DADOS HISTÓRICOS DO INMET

Dentro da análise dessa sessão, serão considerados os números relativos à quantidade de possíveis problemas encontrados ao executar a rotina de verificação, não levando em consideração a proporcionalidade do total de dados disponíveis em cada região.

Os resultados da avaliação de qualidade aplicadas no conjunto de dados do INMET⁶, foi agrupado de duas formas diferentes. A primeira agregação aconteceu pelas regiões e estados do Brasil, conforme a Tabela 6.

TABELA 6 - Resultado das análises de qualidade para os dados do INMET

Região	Estado	Valores Zerados	Falhas de leitura	Outliers	TOTAL*
SE	MG	17738461	501393	935052	19174906
CO	MT	14520390	595650	446482	15562522
NE	BA	13946661	524072	712109	15182842
S	RS	13045111	382293	652088	14079492
SE	SP	10983257	421159	596150	12000566
CO	MS	10720441	510232	538838	11769511
N	PA	10571573	439833	337460	11348866
S	PR	9064412	341552	487886	9893850
N	AM	8783421	368026	255569	9407016

⁶ Disponível no repositório: <https://github.com/AceMX/dadosqualidade/tree/main/INMET>.

Originalmente em: <https://portal.inmet.gov.br/dadoshistoricos>

* Classificados por Total, do maior para o menor.

S	SC	7797859	223163	363581	8384603
CO	GO	6865970	236178	372833	7474981
NE	MA	6484553	262946	272538	7020037
SE	RJ	6403453	185264	365284	6954001
NE	PI	6135729	254209	426076	6816014
N	TO	4965908	223331	282771	5472010
NE	CE	4769704	207833	275460	5252997
NE	PE	3782494	138164	263330	4183988
N	AC	3669636	158245	85528	3913409
SE	ES	3204880	118502	184604	3507986
NE	PB	2902136	100302	113436	3115874
NE	RN	2636873	136934	195413	2969220
N	RO	2058457	93353	49792	2201602
NE	SE	1965060	81634	97834	2144528
NE	AL	1919848	82783	111327	2113958
N	AP	973791	44739	19849	1038379
CO	DF	811958	30540	47750	890248
N	RR	702807	27862	23873	754542

Não foi possível realizar de forma automática para o banco histórico do INMET devido a necessidade de se conhecer todas as características físicas de todas as variáveis de diferentes regiões do Brasil.

O estado de Minas Gerais apresentou o maior número de dados com algum tipo de problema, seguido por Mato Grosso, Bahia e Rio Grande do Sul. Já os estados com menos problemas identificados nas rotinas foram Roraima, Distrito Federal, Amapá e Alagoas.

Como uma segunda agregação, pode-se verificar na Tabela 7 uma visão mais individualizada sobre cada uma das variáveis ambientais, também divididas pelas regiões do Brasil.

TABELA 7 - Resultado das validações de qualidade por região e variável ambiental

Região	Variável	Valores Zerados	Falha de Leitura	Outliers	Total*
NE	PRECIPITACAO TOTAL HORARIO mm	11528612	93526	155731	11777869
SE	PRECIPITACAO TOTAL HORARIO mm	10712419	54769	155530	10922718
CO	PRECIPITACAO TOTAL HORARIO mm	7829067	79772	97413	8006252
S	PRECIPITACAO TOTAL HORARIO mm	7276415	45062	107164	7428641

N	PRECIPITACAO TOTAL HORARIO mm	6065022	75961	81255	6222238
NE	RADIACAO GLOBAL KJ m	4770368	495829	6095	5272292
SE	RADIACAO GLOBAL KJ m	4134394	390868	4516	4529778
CO	RADIACAO GLOBAL KJ m	3241098	307728	2865	3551691
S	RADIACAO GLOBAL KJ m	2912695	276489	2321	3191505
N	RADIACAO GLOBAL KJ m	2876298	270472	8073	3154843
NE	UMIDADE REL MAX NA HORA ANT AUT	2845089	88794	155393	3089276
SE	UMIDADE REL MAX NA HORA ANT AUT	2630949	56087	130213	2817249
NE	UMIDADE REL MIN NA HORA ANT AUT	2515909	89659	155440	2761008
NE	UMIDADE RELATIVA DO AR HORARIA	2407803	89606	155966	2653375
NE	VENTO VELOCIDADE HORARIA m s	2405448	79205	149706	2634359
CO	VENTO VELOCIDADE HORARIA m s	2351249	63359	100504	2515112
N	VENTO VELOCIDADE HORARIA m s	2306205	69951	67535	2443691
NE	VENTO RAJADA MAXIMA m s	2271250	79335	143382	2493967
S	UMIDADE REL MAX NA HORA ANT AUT	2252676	47735	97372	2397783
N	UMIDADE REL MAX NA HORA ANT AUT	2236266	69662	65247	2371175
SE	UMIDADE REL MIN NA HORA ANT AUT	2181236	56968	122768	2360972
SE	VENTO VELOCIDADE HORARIA m s	2176253	54010	141999	2372262
NE	TEMPERATURA ORVALHO MAX NA HORA ANT AUT C	2118992	88834	166674	2374500
SE	UMIDADE RELATIVA DO AR HORARIA	2116220	56813	129657	2302690
NE	TEMPERATURA ORVALHO MIN NA HORA ANT AUT C	2107658	88741	168668	2365067
CO	UMIDADE REL MAX NA HORA ANT AUT	2026815	69545	84204	2180564
CO	VENTO RAJADA MAXIMA m s	1975829	63871	94639	2134339
N	VENTO RAJADA MAXIMA m s	1920342	70769	67597	2058708
SE	VENTO RAJADA MAXIMA m s	1918322	54707	132592	2105621
N	UMIDADE RELATIVA DO AR HORARIA	1917200	69733	60087	2047020
N	UMIDADE REL MIN NA HORA ANT AUT	1911911	69075	60160	2041146
S	UMIDADE REL MIN NA HORA ANT AUT	1896512	48592	93825	2038929
S	UMIDADE RELATIVA DO AR HORARIA	1861416	48394	95363	2005173
NE	TEMPERATURA DO PONTO DE ORVALHO C	1823109	89215	167430	2079754
CO	UMIDADE REL MIN NA HORA ANT AUT	1779015	70001	82553	1931569
S	VENTO VELOCIDADE HORARIA m s	1759610	41674	101121	1902405

SE	TEMPERATURA ORVALHO MIN NA HORA ANT AUT C	1705446	57702	136587	1899735
CO	UMIDADE RELATIVA DO AR HORARIA	1688780	70049	83627	1842456
SE	TEMPERATURA ORVALHO MAX NA HORA ANT AUT C	1671346	56904	133574	1861824
S	VENTO RAJADA MAXIMA m s	1620900	41801	97356	1760057
N	TEMPERATURA ORVALHO MIN NA HORA ANT AUT C	1570531	69594	75501	1715626
N	TEMPERATURA ORVALHO MAX NA HORA ANT AUT C	1532530	69914	73275	1675719
NE	PRESSAO ATMOSFERICA MAX NA HORA ANT AUT mB	1532031	71452	147366	1750849
NE	PRESSAO ATMOSFERICA MIN NA HORA ANT AUT mB	1528840	71446	147062	1747348
CO	TEMPERATURA ORVALHO MIN NA HORA ANT AUT C	1518777	70262	95672	1684711
NE	TEMPERATURA MINIMA NA HORA ANT AUT C	1491506	70560	143205	1705271
CO	TEMPERATURA ORVALHO MAX NA HORA ANT AUT C	1486275	70034	95196	1651505
SE	TEMPERATURA DO PONTO DE ORVALHO C	1418092	57065	138231	1613388
NE	TEMPERATURA MAXIMA NA HORA ANT AUT C	1395446	70587	141981	1608014
N	TEMPERATURA DO PONTO DE ORVALHO C	1381075	70019	72812	1523906
S	TEMPERATURA ORVALHO MIN NA HORA ANT AUT C	1316339	47838	102417	1466594
CO	TEMPERATURA DO PONTO DE ORVALHO C	1312582	70251	95969	1478802
SE	PRESSAO ATMOSFERICA MAX NA HORA ANT AUT mB	1309552	46408	122858	1478818
NE	PRESSAO ATMOSFERICA AO NIVEL DA ESTACAO HORARIA mB	1306944	71250	146692	1524886
SE	PRESSAO ATMOSFERICA MIN NA HORA ANT AUT mB	1306241	46401	122064	1474706
S	TEMPERATURA ORVALHO MAX NA HORA ANT AUT C	1278544	47259	100193	1425996
NE	TEMPERATURA DO AR BULBO SECO HORARIA C	1276125	70596	141864	1488585
CO	PRESSAO ATMOSFERICA MAX NA HORA ANT AUT mB	1231843	63302	84287	1379432
N	TEMPERATURA MINIMA NA HORA ANT AUT C	1231432	62359	60398	1354189
CO	PRESSAO ATMOSFERICA MIN NA HORA ANT AUT mB	1229917	63302	83723	1376942
NE	VENTO DIRECAO HORARIA gr gr	1217928	80242	174868	1473038
N	PRESSAO ATMOSFERICA MAX NA HORA ANT AUT mB	1164291	64116	62817	1291224
N	TEMPERATURA MAXIMA NA HORA ANT AUT C	1159114	62372	59314	1280800
N	VENTO DIRECAO HORARIA gr gr	1158770	71301	58722	1288793
N	PRESSAO ATMOSFERICA MIN NA HORA ANT AUT mB	1155862	64108	62458	1282428
SE	TEMPERATURA MINIMA NA HORA ANT AUT C	1138708	45475	121450	1305633

S	TEMPERATURA DO PONTO DE ORVALHO C	1127988	47338	100668	1275994
CO	VENTO DIRECAO HORARIA gr gr	1103839	65304	79642	1248785
CO	TEMPERATURA MINIMA NA HORA ANT AUT C	1098257	61043	81462	1240762
S	PRESSAO ATMOSFERICA MAX NA HORA ANT AUT mB	1093534	35415	88785	1217734
N	TEMPERATURA DO AR BULBO SECO HORARIA C	1091415	62089	58372	1211876
SE	PRESSAO ATMOSFERICA AO NIVEL DA ESTACAO HORARIA mB	1085027	46090	125646	1256763
S	PRESSAO ATMOSFERICA MIN NA HORA ANT AUT mB	1083002	35414	88646	1207062
CO	PRESSAO ATMOSFERICA AO NIVEL DA ESTACAO HORARIA mB	1075120	63188	84856	1223164
N	PRESSAO ATMOSFERICA AO NIVEL DA ESTACAO HORARIA mB	1047329	63894	61219	1172442
SE	TEMPERATURA MAXIMA NA HORA ANT AUT C	1040084	45463	119066	1204613
CO	TEMPERATURA MAXIMA NA HORA ANT AUT C	1026290	60909	79305	1166504
S	TEMPERATURA MINIMA NA HORA ANT AUT C	971582	35600	85852	1093034
CO	TEMPERATURA DO AR BULBO SECO HORARIA C	944006	60680	79986	1084672
SE	TEMPERATURA DO AR BULBO SECO HORARIA C	932544	45150	123954	1101648
S	PRESSAO ATMOSFERICA AO NIVEL DA ESTACAO HORARIA mB	920534	35311	88319	1044164
S	TEMPERATURA MAXIMA NA HORA ANT AUT C	909677	35571	84107	1029355
SE	VENTO DIRECAO HORARIA gr gr	853218	55438	120385	1029041
S	TEMPERATURA DO AR BULBO SECO HORARIA C	819948	35574	84624	940146
S	VENTO DIRECAO HORARIA gr gr	806010	41941	85422	933373

*Classificado pelo Total, do maior para o menor

Nessa agregação individualizada pelas variáveis é possível identificar quais variáveis estão apresentando maior quantidade de dados com problemas. Os dados de precipitação total são os que mais apresentaram falhas em todas as regiões do Brasil e as variáveis que menos apresentaram problemas foram as de direção horária do vento e de temperatura do ar de bulbo seco na região Sul.

Realizando uma verificação mais detalhada e interanual das três variáveis que mais apresentaram marcações de problemas com a qualidade, é possível identificar que para os métodos de detecção de valores zerados e outliers os sensores apresentam uma

melhoria de qualidade nos dois últimos anos de análise, conforme as Figuras 4 e 5. Ou seja, não considerando o total de dados disponíveis para cada sensor, esses são os que mais receberam marcações para verificação mais detalhada.

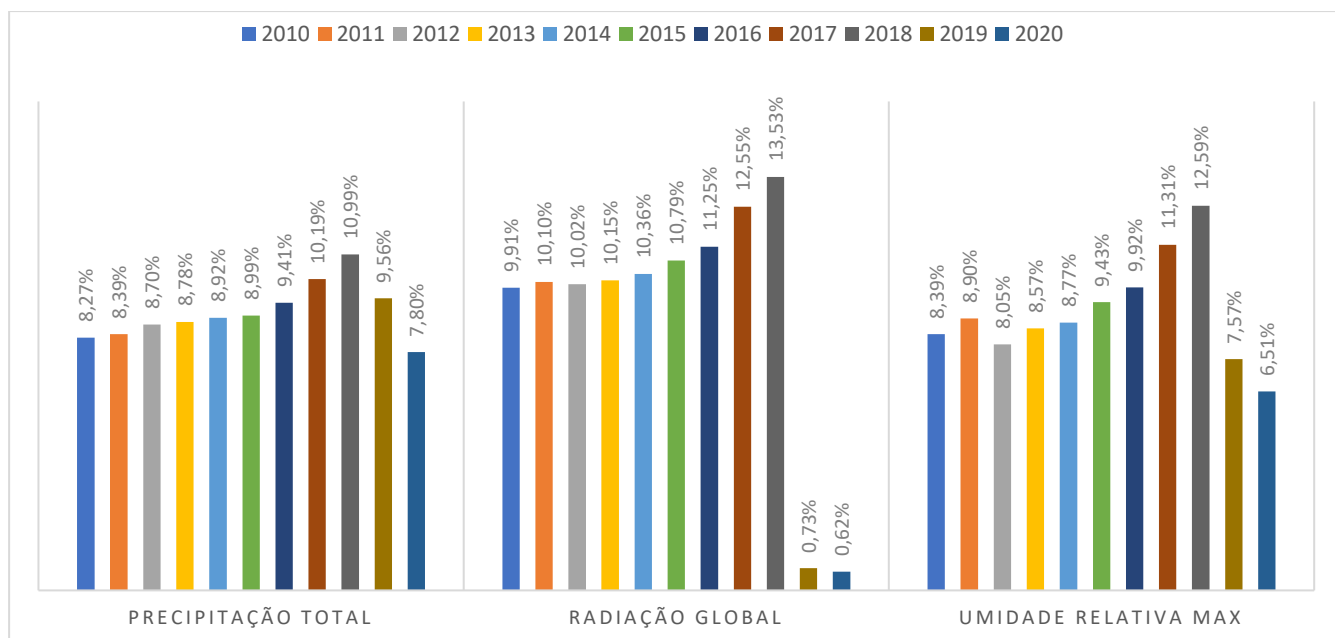


FIGURA 4 – As três variáveis com mais indicações de problemas para valores zerados

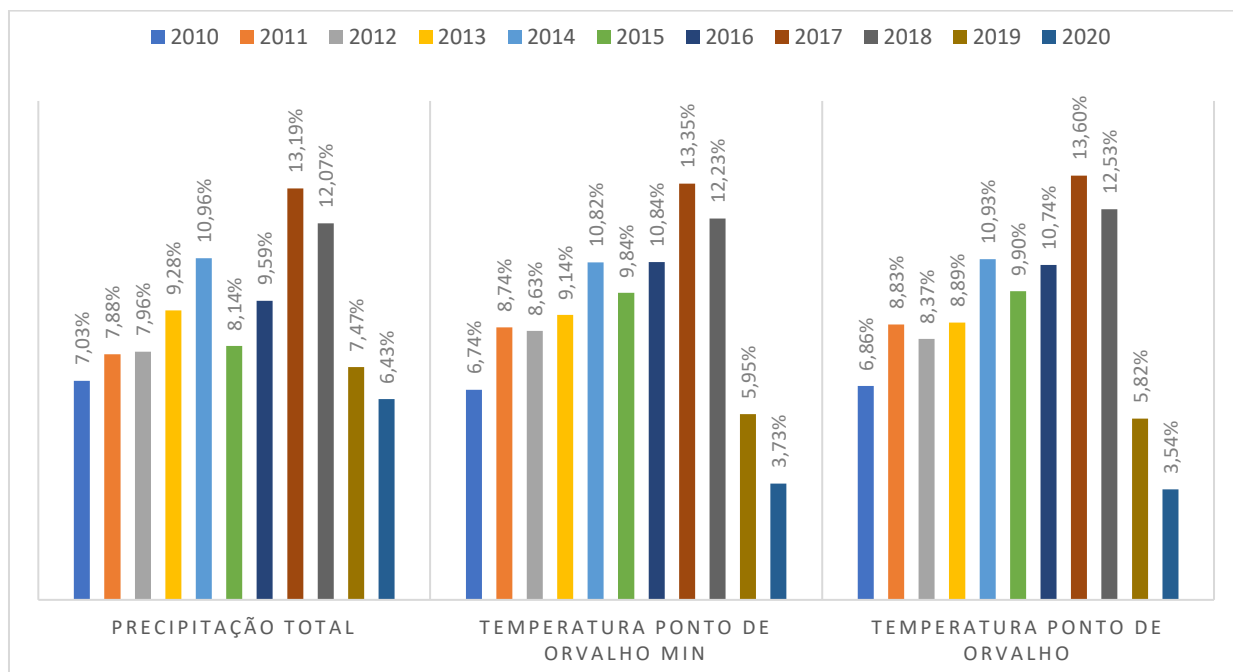


FIGURA 5 – As três variáveis com mais indicações de problemas para outliers

Para a verificação de falhas de leitura, o ano de 2020 apresentou um aumento nesse índice, como é possível identificar na Figura 6.

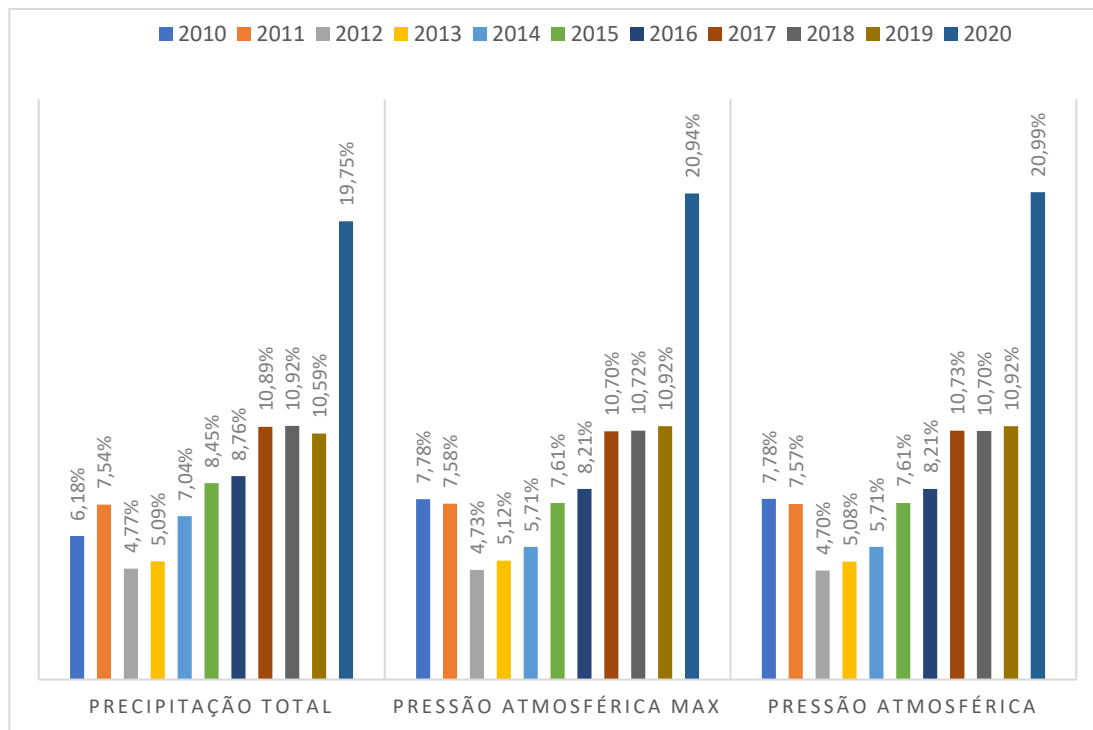


FIGURA 6 – As três variáveis com mais indicações de problemas para falhas de leitura

Continuando com a análise interanual, agora sendo regionalizada para cada método de qualidade aplicado, é possível identificar diferenças com o passar dos anos na quantidade de identificações de possíveis problemas nos dados. Nota-se que na pesquisa por possíveis *outliers*, os anos de 2019 e 2020 apresentam menos problemas que os anos anteriores. Já os anos de 2017 e 2018 apresentam os maiores níveis de identificação de possíveis *outliers*, conforme a Figura 7. Aqui continua o mesmo método de análise das figuras anteriores, ou seja, a avaliação acontece sobre a quantidade de possíveis erros encontrados, não considerando a proporção de dados gerados por cada sensor ou região.

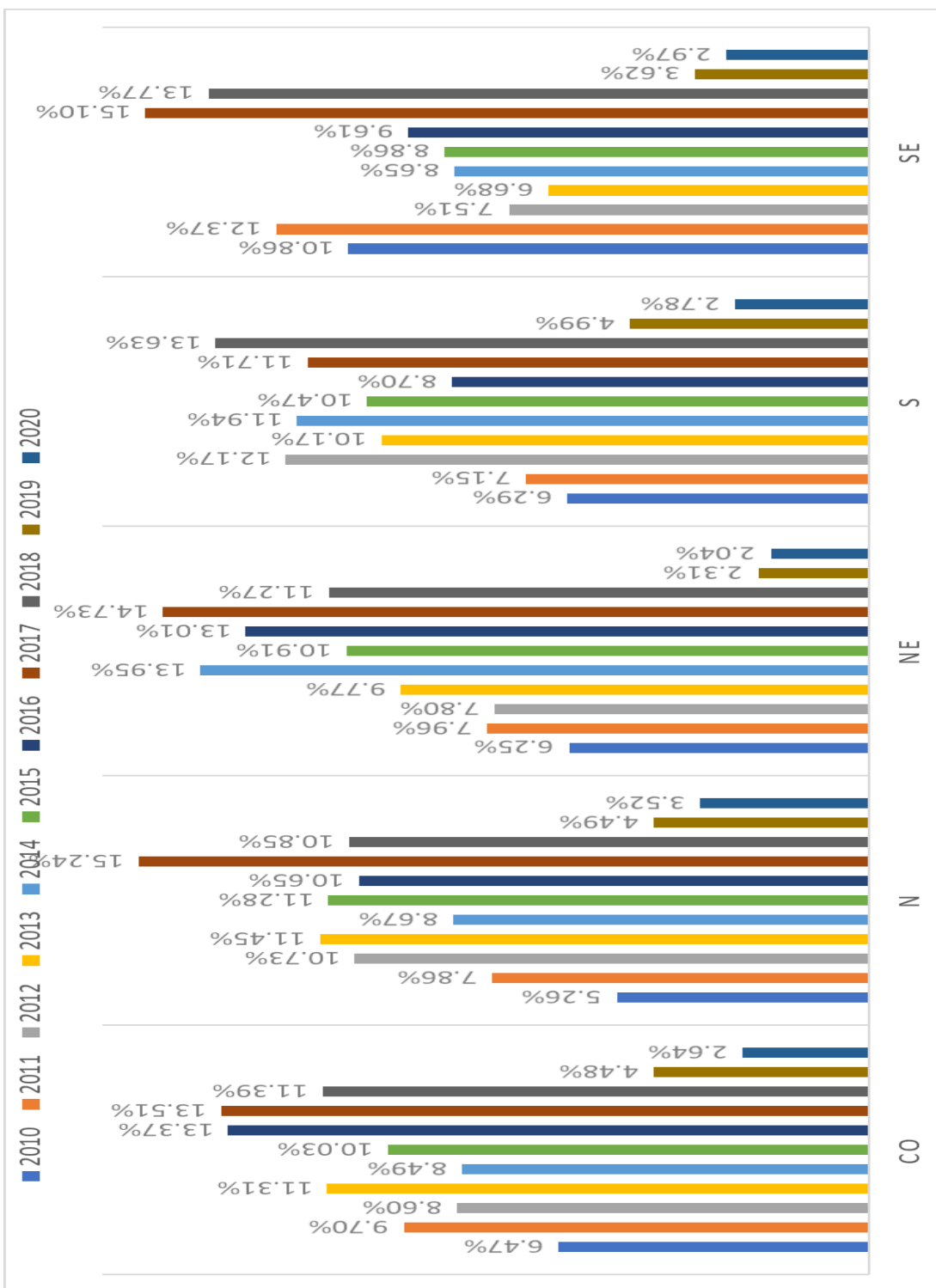


FIGURA 7 - Outliers por região e ano

Para o conjunto de análise sobre os valores zerados, o mesmo padrão verificado na análise de *outliers* é identificada aqui. Com a qualidade sendo melhorada com o passar

dos anos das estações avaliadas. Ou seja, 2019 e 2020 representam os melhores anos, com menor quantidade de dados identificados como zerados, conforme a Figura 8.

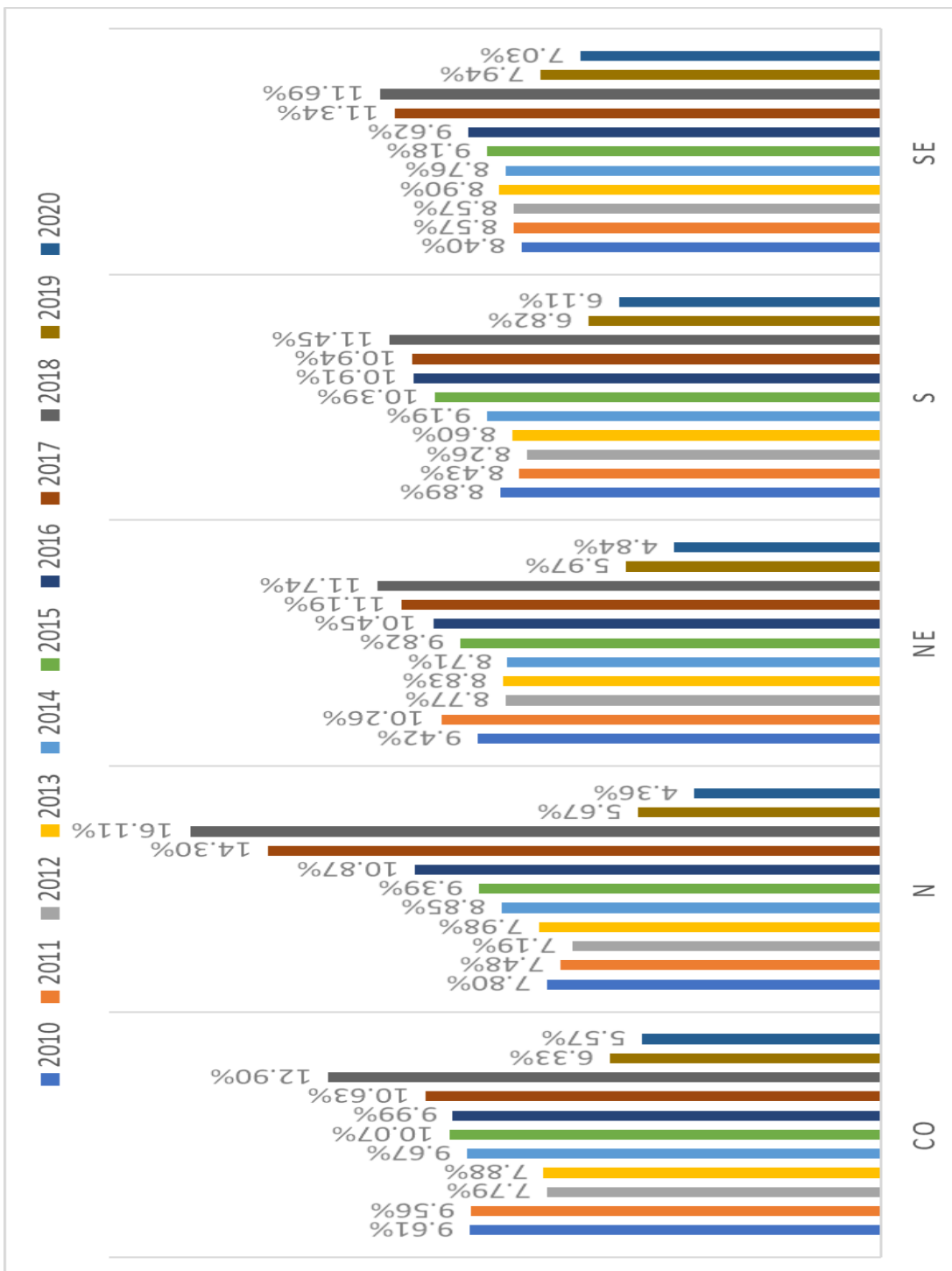


FIGURA 8 - Valores zerados por região e ano

Quanto a análise aplicada com o conjunto de falhas de leitura, foi identificado que os dados do ano de 2020 apresentaram uma maior frequência de falhas, ou seja, o conjunto de dados do último ano da série apresentou maior quantidade de dados faltantes no conjunto, conforme a Figura 9.

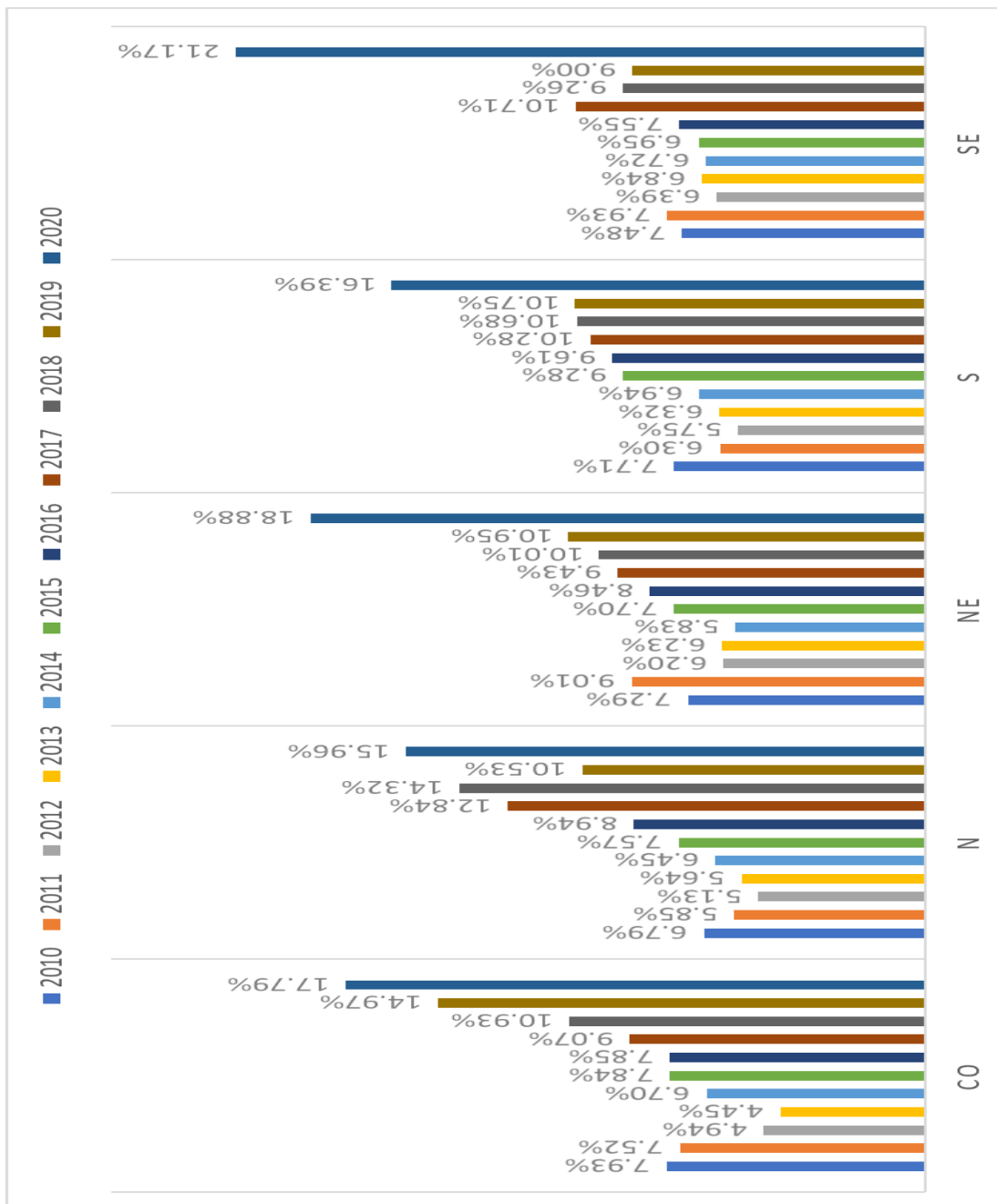


FIGURA 9 - Falhas de leitura por região e ano

4.2. APLICAÇÃO DOS INDICADORES DE QUALIDADE

Com a definição dos DQI, pode-se então aplicar o conjunto de regras definidos para a matriz de qualidade em um conjunto de dados para aferição de sua qualidade, claro, sempre de acordo com as métricas pré-estabelecidas.

A aplicação dos indicadores de qualidade ocorre sobre o conjunto completo dos dados, sendo assim, é importante que a análise de qualidade sobre os dados já tenha a sua execução realizada. Sendo assim, pode-se estabelecer os DQI para os conjuntos de dados do Pantanal e do Cerrado.

O conjunto do Pantanal é uma agregação dos conjuntos α e β das validações dos dados, que foi realizada anteriormente. Dessa forma, o novo conjunto criado possui dados de temperatura do ar da torre localizada no Pantanal mato-grossense no período de 2015 a janeiro de 2016.

O conjunto denominado Cerrado, corresponde ao conjunto analisado anteriormente com o nome de π . Esse conjunto apresenta dados de uma estação convencional do INMET, localizada em Cuiabá – MT.

TABELA 8 – Indicadores de Qualidade para o Pantanal e Cerrado

Conjunto de dados	CRD	TEP	CST	INT	ACE
Pantanal	1	2	1	1	5
Cerrado	5	2	1	2	3

Legenda: CRB – Credibilidade, TEP – Temporalidade, CST – Consistência, INT – Integridade e ACE – Acessibilidade

Agora é possível identificar os pontos de melhorias nos conjuntos de dados. No conjunto Pantanal percebe-se a necessidade de trabalhar a acessibilidade aos dados gerados. Uma vez que, esses dados não se encontram disponíveis na internet para que as pessoas possam utilizá-los quando quiserem.

Já, no conjunto Cerrado, a credibilidade precisa ser trabalhada de forma mais ativa, já que, apesar dos dados estarem disponíveis na internet em formato aberto, não se tem a

identificação dos sensores utilizados, o que pode levantar questionamentos sobre a forma de geração desses dados.

A Tabela 10 nos mostra os indicadores de qualidade aplicados aos conjuntos das estações automatizadas do INMET dos últimos dez anos. Para essa análise, utilizou-se dos dados agregados por região do Brasil.

TABELA 9 - Indicadores de Qualidade para as regiões do Brasil

Conjunto de dados	CRD	TEP	CST	INT	ACE
NE	5	1	3	1	3
SE	5	1	3	1	3
CO	5	1	3	1	3
S	5	1	3	1	3
N	5	1	3	1	3

Legenda: CRB – Credibilidade, TEP – Temporalidade, CST – Consistência, INT – Integridade e ACE – Acessibilidade

Como no conjunto do Cerrado, os dados das estações automáticas do INMET foram classificados com a credibilidade no pior valor do índice, uma vez que, não temos acesso as especificações dos sensores utilizados para realizar as medidas. A consistência obteve um índice regular, já que um número grande falhas e valores zerados foram identificados nos dados analisados.

4.3. ESTRUTURA DE ARMAZENAMENTO E DISPONIBILIZAÇÃO DOS DADOS

A modelagem aplicada à estrutura do banco de dados foi a estrela, como mostra a Figura 10, uma modelagem padrão em bancos de dados dedicados a pesquisa de informação, com esse tipo de modelagem, o banco de dados relacional concentra na busca de registros e na devolução dos resultados, o que melhora a sua performance. A

otimização da modelagem estrela é voltada para bases onde o principal recurso será a recuperação dos dados, ou as consultas, em vez da inserção deles.

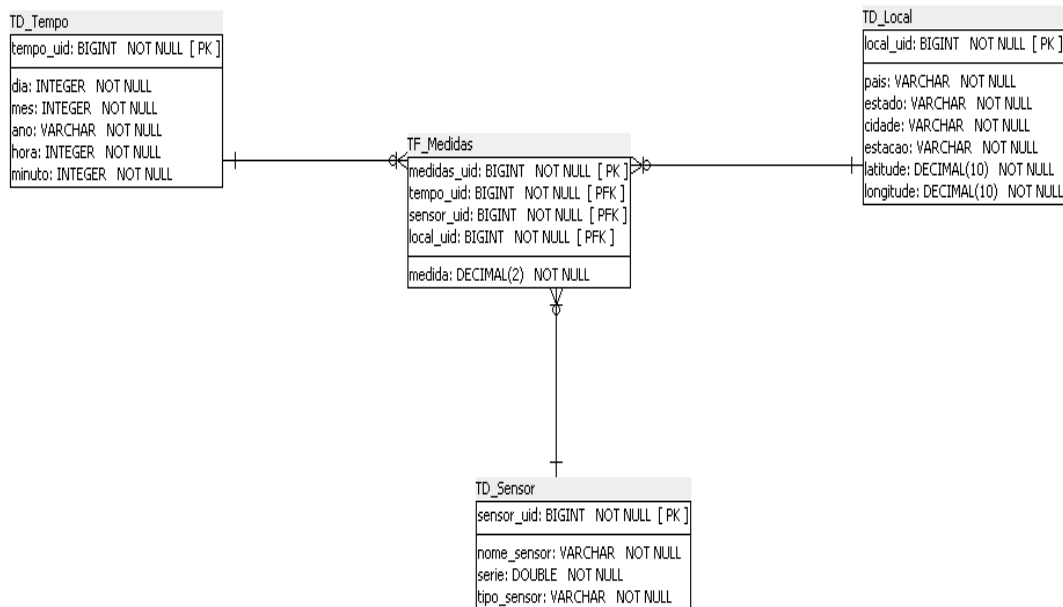


FIGURA 10- Modelo estrela para dados climatológicos

Dessa forma, toda a estrutura de carga de dados pode ser simplificada e ajustada por meio da padronização aplicada diretamente aos sensores que geram as informações, então, uma estrutura de ETL pode ser automatizada, reduzindo a complexidade das atividades envolvidas e tornando o processo mais amigável à manutenção quando necessário. Fazendo isso, não precisa ser um especialista na área de computação para poder alterar e ajustar os parâmetros de leitura dos dados de entrada do banco de dados. Com isso, optou-se pela utilização da ferramenta Talend Open Studio Community Edition, que é uma solução sem custos e que é capaz de gerar as rotinas automatizadas de ETL. Como exemplo, a Figura 11 mostra o ambiente e uma rotina para importar dados de uma planilha eletrônica para o DW.

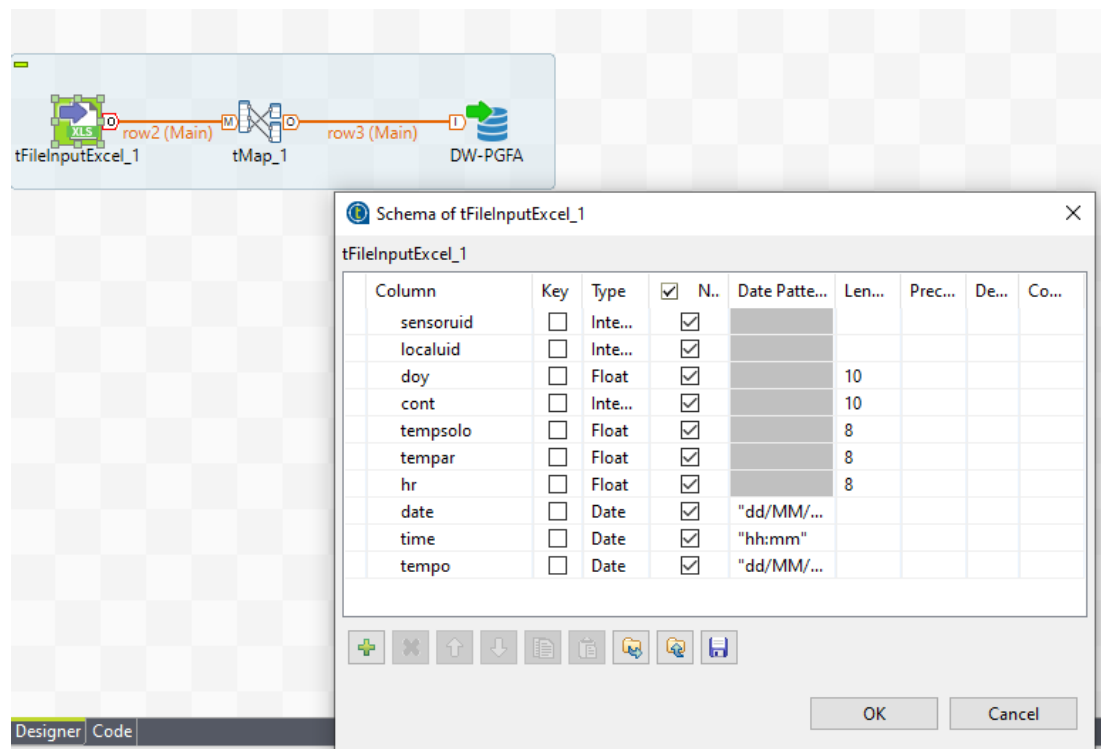


FIGURA 11- Rotina de carga no DW

O banco de dados escolhido para a execução da modelagem apresentada foi Postgres SQL em sua versão 10. É um banco de dados de código aberto com uma estrutura relacional e que possui uma comunidade de suporte ativo e é constantemente atualizado. Sabe-se que as bases de dados de código aberto (OSD) têm crescido em interesse nos últimos anos. De acordo com Schumacher (2010), mais de 80% das corporações se propunham a migrar para um OSD nos próximos anos.

A Figura 12 mostra o diagrama operacional da estrutura desenvolvida, com os dados de origem servindo como entrada para todo o processo, com a possibilidade de serem de fontes heterogêneas. O processamento de dados ocorre na fase de ETL, onde é possível aplicar um tratamento de qualidade e limpar os dados, ou seja, nessa etapa os dados têm os seus tipos analisados, se são o que realmente se espera e organizados de acordo com a estrutura de armazenamento do DW, recebendo os identificadores internos.

O banco de dados, representado pelo DW, que constitui um grande repositório de dados, busca auxiliar os pesquisadores na recuperação de informações o mais rápido possível e com tratamentos já aplicados. Finalmente, os usuários se conectam a este

DW através de algum aplicativo para realizar a recuperação das informações desejadas, a sugestão implementada aqui é o moderno software de exploração e visualização de dados chamado Superset[®], que é mantido e distribuído gratuitamente pela Apache Software Foundation⁷.

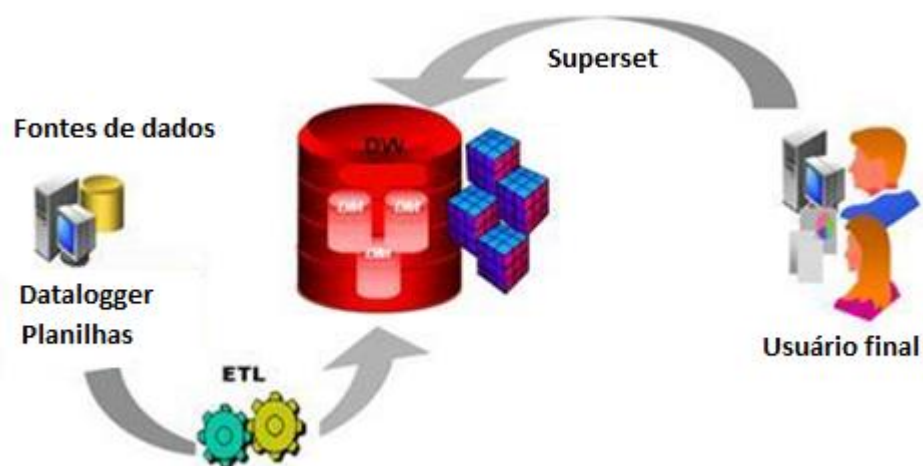


FIGURA 12 – Diagrama estrutural da PADC

A ferramenta possui a capacidade de realizar operações sobre os dados, como realizar a média, mediana, ou outras operações. Possibilita também disponibilizar publicamente alguns painéis de informações, através de atribuição pública de acesso aos conjuntos que o pesquisador desejar. Ou seja, os dados utilizados podem ser obtidos em um formato aberto para uso da comunidade. Isso mostra que a ferramenta atende à necessidade apresentada, consultando e oferecendo a possibilidade de se obter informações do DW de forma clara e objetiva.

Nas Figuras 13 e 14, com dados do conjunto Pantanal do ano de 2015, é possível analisar dados de temperatura com diferentes variáveis, por exemplo, a temperatura do ar e a temperatura do solo.

⁷ Disponível em: <https://github.com/apache/superset>

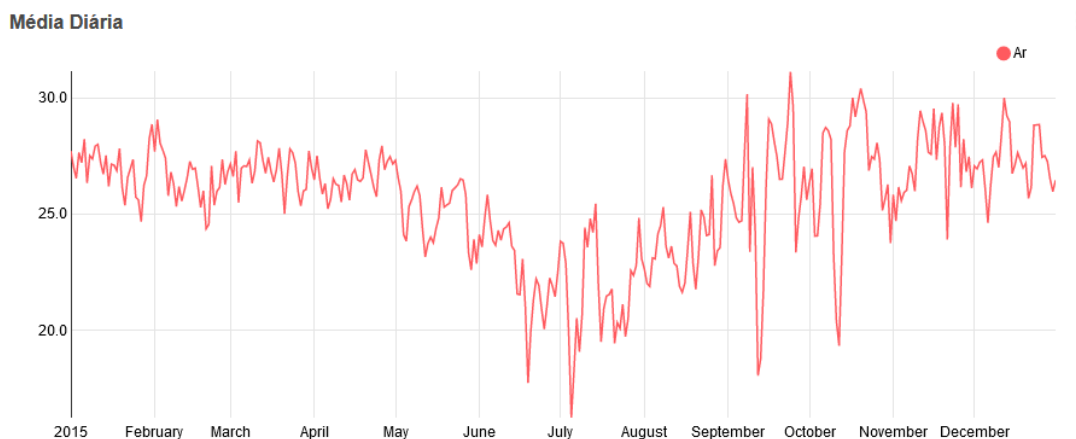


FIGURA 13 – Média diária da temperatura do ar em °C

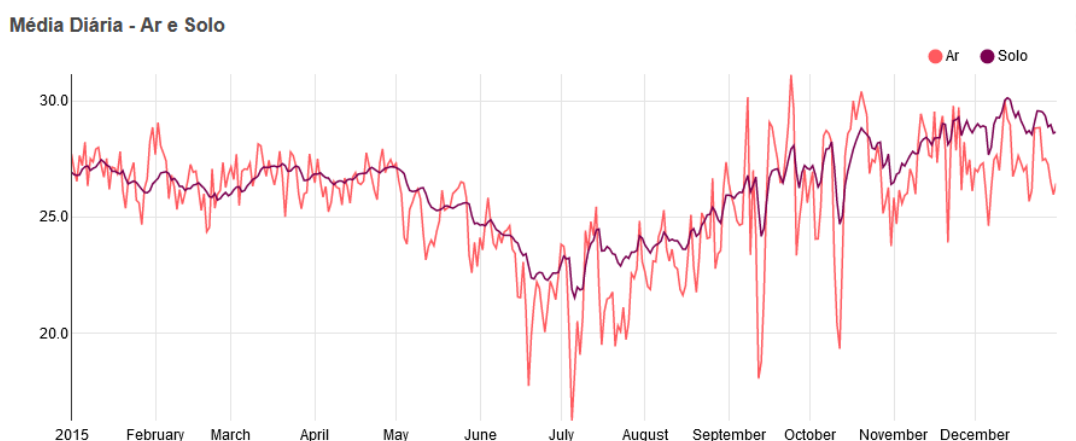


FIGURA 14 – Média diária das temperaturas do ar e do solo em °C

A escala utilizada foi uma média diária para ambas as variáveis. Como o conjunto de dados utilizado tem uma maior granularidade, com dados a cada meia hora, é possível obter mais detalhes, como a Figura 15 que mostra a média horária dos valores de temperatura do ar, do conjunto de dados Pantanal.

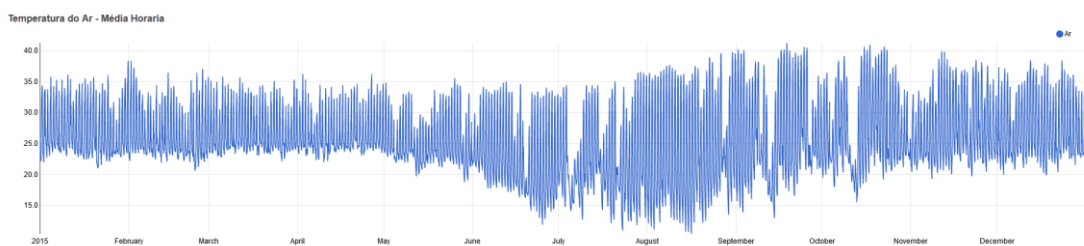


FIGURA 15 – Média horária da temperatura do ar em °C

As Figuras possibilitam analisar a dinamicidade da ferramenta, com a possibilidade de explorar os dados do armazém de várias formas diferentes, como visualizar as médias mensais e horárias. Ademais, foi possível criar a arquitetura da plataforma utilizando apenas software de código aberto, sem a necessidade de gastar recursos financeiros com licenciamento ou contratação de uma solução comercial. A base de dados Postgres, em sua versão 10, apresenta a robustez necessária para a concepção de um projeto de DW e estará pronto para o possível crescimento dos dados ambientais. A ferramenta de consulta Superset é amplamente utilizada por grandes empresas comerciais, como Airbnb, Udemy, Twitter, entre outras.

Um outro problema, resolvido indiretamente por essa plataforma, foi o nível de disponibilidade dos dados. Como essa plataforma opera *online*, a implementação pode ser aberta a todos como um serviço público. Fazendo isso, todos os dados exibidos nos painéis estão disponíveis para download em formatos abertos ou vinculando o painel desejado a algum conteúdo externo através de links fornecidos pelo proprietário da plataforma.

Para melhor visualização das informações inseridas na plataforma, foi elaborado um painel de gráficos que contém os dados utilizados na pesquisa. Desta forma, a Figura 16 mostra um exemplo de como os dados podem ser expostos de forma aberta⁸.

⁸ O painel construído com as variáveis climatológicas está disponível em: <http://sig.ufmt.br/superset/dashboard/pgfa/>.

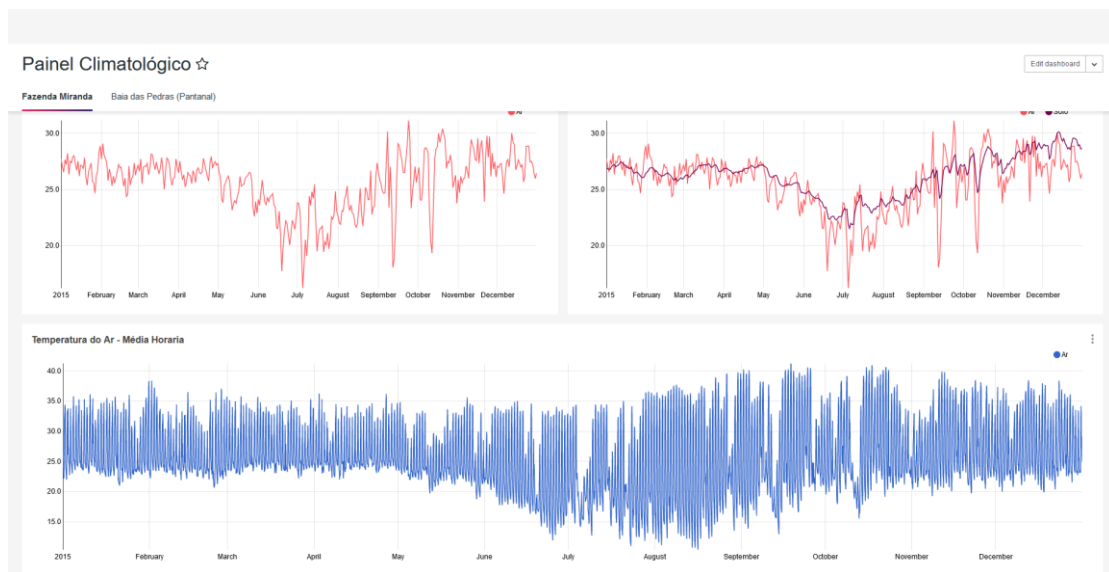


FIGURA 16 - Exemplo de painel construído com as variáveis climatológicas

5. CONCLUSÃO

Os resultados apresentados apontam que as técnicas apresentadas nessa pesquisa são fundamentais para validação dos dados e podem garantir a qualidade de diferentes conjuntos de dados de variáveis ambientais, com diferentes intervalos de medição, bem como em diferentes locais.

Com o direcionamento apresentado pela validação de qualidade, os técnicos que cuidam da coleta e armazenamento dos dados também são beneficiados, podendo identificar qualquer problema do conjunto de forma direta e atuar na correção necessária, deixando o conjunto pronto para futuros trabalhos.

É interessante notar que com a validação automatizada muitos problemas podem ser detectados de forma mais eficiente, uma vez que, seria possível a elaboração de rotinas para monitoramento dos resultados das validações, justamente para evitar um impacto maior no conjunto de dados. A validação automatizada possibilitou também estabelecer um monitoramento mais proativo em relação à qualidade dos sensores, efetividade da instalação, bem como da manutenção aplicada aos equipamentos e detecção de problemas nos locais de armazenamento, como possíveis falhas nos *dataloggers*.

A validação de qualidade nos dados públicos do INMET possibilitou mostrar que alguns sensores precisam de verificação para que possam melhorar a sua qualidade, diminuindo a quantidade de falhas apresentadas. Com as agregações realizadas por região e estado é possível obter um guia das regiões com dados mais confiáveis, possibilitando uma melhor qualidade nos resultados produzidos pelos estudos.

Ainda sobre a qualidade dos dados históricos do INMET, em uma análise interanual para cada um dos conjuntos de saída da metodologia, foi possível identificar que nos dois últimos anos da série histórica, 2019 e 2020, os dados apresentaram uma melhora na sua qualidade, porém, em 2020 nota-se um elevado número de falhas de leitura.

Na página onde os dados foram obtidos, o INMET não informa nada sobre a realização de manutenções ou problemas que ocorreram nos períodos para que o

pesquisador fique ciente de algum problema. O que demonstra a necessidade de aplicação das técnicas de qualidade aqui propostas.

Conforme Boden et al. (2013), as validações de qualidade já são exigidas em muitas redes internacionais de dados, como a rede Ameriflux, que trabalha com dados das Américas do Norte, Central e do Sul e possuem várias validações de qualidade para que os dados possam constar em sua rede.

Também foi possível centralizar toda a disponibilidade de dados em um único banco de dados, pois o armazém de dados é feito para suportar cargas de dados de diferentes fontes. Como a modelagem das tabelas foi feita em formato estrela, garantiu-se que os dados pudessem ser utilizados e analisados em conjunto, uma vez que características compartilhadas entre os dados tornem esse uso possível, como intervalos de tempo, localização, entre outros.

A plataforma elaborada se assemelha a uma estrutura para lidar com dados estratégicos para a tomada de decisões. Uma vez que as necessidades dos pesquisadores podem ser comparadas com as de gestores de grandes empresas, onde precisam do máximo de informações possível, objetividade, facilidade de visualização e personalização das informações para apoiar suas pesquisas ou decisões.

É necessário realizar a continuidade da disseminação dentro da comunidade científica sobre a necessidade de estabelecer padrões de disponibilidade de dados, bem como difundir conhecimentos sobre dados estratégicos. Isso pode ser feito apresentando portais como do Ameriflux e INMET, que possibilitam a realização de pesquisas com os dados produzidos por várias fontes diferentes.

É possível notar que muitas empresas comerciais utilizam as mesmas tecnologias apresentadas neste trabalho, cada uma com a sua particularidade, para previsões financeiras e planejamento de suas ações e isso pode ser facilmente levado à comunidade acadêmica e de pesquisa, uma vez que, decisões importantes são tomadas com base nas pesquisas realizadas por essas pessoas. Neste sentido, esse tipo de ferramenta à disposição dos pesquisadores, torna possível aplicar mais tempo no objetivo final da pesquisa do que na área intermediária, facilitando a obtenção de dados de qualidade.

Diferentes abordagens podem ser realizadas nesta área, explorando bancos de dados não convencionais (NoSQL), como Mongo DB, Maria DB, etc, ou mesmo bancos de dados colunares, como o SyBase IQ, para armazenar grandes massas de dados e onde há necessidade de realizar pesquisas em poucos segundos, com cruzamento maciço de dados entre diferentes estações.

Sugere-se também outros estudos que apliquem métodos de validação cruzada com outras torres microclimáticas, o que pode melhorar a validação de qualidade. No cenário atual não é possível aplicar tais validações, uma vez que, os conjuntos de dados analisados não possuem outras torres próximas.

REFERÊNCIAS BIBLIOGRÁFICAS

ALMANAQUE BRASIL SOCIOAMBIENTAL – ISA/2008. Disponível em: <<https://www.socioambiental.org/pt-br/o-isa/publicacoes/almanaque-brasil-socioambiental-2008>>. Acesso em: 29 de dezembro de 2020.

ALEXANDERSSON, H Moberg A. Homogenization of Swedish temperature data. Part 1: homogeneity test for linear trends. **Int J Climatology**, 17, 1997. 25–34 p.

ALVARES, Clayton Alcarde; STAPE, José Luiz; SENTELHAS, Paulo Cesar; GONÇALVES, José Leonardo de Moraes; SPAROVEK, Gerd. Koppen's climate classification map for Brazil. **Meteorologische Zeitschrift**, Vol.22, No. 6, 2014. 711–728 p.

ATENAS, J., Havemann, L. and PRIEGO, E. Open Data as Open Educational Resources: Towards transversal skills and global citizenship, **Open praxis**, Vol. 7, No. 4, 2015 pp. 377-389. ISSN 2304-070X. DOI 10.5944/openpraxis.7.4.233

AUSTRALIAN BUREAU OF METEOROLOGY e CSIRO. (2011). Climate change in the Pacific: scientific assessment and new research. Volume 1: Regional overview, 257. Disponível: <http://www.pacificclimatechangescience.org/publications/reports/> (Acesso em: 24 de novembro de 2019).

Azzam, T., Evergreen, S., Germuth, A. A., e Kistler, S. J. Data visualization and evaluation. In T. Azzam & S. Evergreen (Eds.), Data visualization, part 1. **New Directions for Evaluation**, 139, 2013, 7–32

BATINI, C.; CAPIELLO, C.; FRANCALANCI, C.; MAURINO, A. Methodologies for data quality assessment and improvement. **ACM Computing Surveys**, 41(3), 2009, 1–52. <https://doi.org/10.1145/1541880.1541883>

BERNERS-LEE, Tim. (2006) “5-star open data” (Online) – Disponível: <https://5stardata.info/en/>. (Acesso em: 10 de outubro de 2020).

BERTRAND, C.; GONZALEZ, S.; JOURNÉE, M. Quality control of 10-min air temperature data at RMI, **Adv. Sci. Res.** 10, 2013, 1-5 p, <https://doi.org/10.5194/asr-10-1-2013>.

BICALHO, T.; SAUER, I.; RAMBAUD, A.; ALTUKHOVA, Y. LCA data quality: A management science perspective. **Journal of Cleaner Production**, 156, 2017, 888–898. <https://doi.org/10.1016/j.jclepro.2017.03.229>.

BODEN, T. A., KRASSOVSKI, M., e YANG, B. The AmeriFlux data activity and data system: an evolving collection of data management techniques, tools, products and services, **Geosci. Instrum. Method. Data Syst.**, 2, 2013, 165–176, <https://doi.org/10.5194/gi-2-165-2013>.

BOULANGER, J.; AIZPURU, J.; LEGGIERI, L; MARINO, M. A procedure for automated quality control and homogenization of historical daily temperature and precipitation data (APACH): Part 1: Quality control and application to the Argentine weather service stations. **Climatic Change**. 98, 2010. 471-491 p.

CARMO, A. F. C.; SHIMABUKURO, M. H.; ALCANTARA, E. H. Avaliação da qualidade de dados ambientais por meio de técnicas de analítica visual. **Bol. Ciênc. Geod.**, Curitiba, v. 22, n. 3, p. 542-556, set. 2016. Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1982-21702016000300542&lng=pt&nrm=iso>. <http://dx.doi.org/10.1590/S1982-21702016000300031>. (Acesso em: 19 de março de 2021).

CAUSSINUS H; MESTRE, O. Detection and correction of artificial shifts in climate series. **Appl Stat**53, 2004. 405–425 p.

CODD, E. F. A relational model of data for large shared data banks. **Communications of the ACM**, 13(6), 1970. 377-387 p.

CODD, E. F. Relational Database: A Practical Foundation for Productivity. IBM San Jose Research Laboratory, 1982.

CODD, E. F. **An evaluation scheme for database management systems that are claimed to be relational**. The Second International Conference on Data Engineering, Los Angeles, CA. 1986. 720-729 p.

COURTNEY, R, WARE, W. Some Informal Comments About Integrity and the Integrity Workshop.º. In: Proc. Of the Invitational Workshop on Data Integrity, **National Institute of Standards and Technology**, Special Publication; 1989, p. 500-168.

CURADO, L. F. A. **Estudo da Intersazonalidade do Fluxo de Calor Latente e Sensível no Cerrado-Pantanal de Mato Grosso**. Tese (Doutorado em Física Ambiental) – Instituto de Física, UFMT, Cuiabá, 2013.

DAVISON, J.; DEEKS, D. Measuring the potential success of information system implementation. **Measuring Business Excellence**, 11(4), 2007, pp. 75–81.

DELVAUX, C; INGELS, R; VRÁBEL, V; JOURNÉE, M; BERTRAND, C. Quality control and homogenization of the Belgian historical temperature data. **Int J Climatol**, 39, 2019. 157– 171 p. <https://doi-org.ez52.periodicos.capes.gov.br/10.1002/joc.5792>

DURRE, I. M. J. M.; VOSE, R. S. Strategies for evaluating quality assurance procedures. **J. Appl. Meteor. Climatol.**, 47, 2008. 1785–1791 p.

DÜSTERHUS, A.; HENSE, A. Advanced information criterion for environmental data quality assurance. **Advances in Science and Research**, 8, 99–104, 2012. <https://doi.org/10.5194/asr-8-99-2012>

EDELEN, A.; INGWERSEN, W. **Guidance on Data Quality Assessment for Life Cycle Inventory Data**, (1), 2016, 37.

FINNVEDEN, G. On the limitations of life cycle assessment and environmental systems analysis tools in general. **The International Journal of Life Cycle Assessment**, 5(4), 2000, 229–238. <https://doi.org/10.1007/BF02979365>

FREE, M.; Creating climate reference datasets: CARDS workshop on adjusting radiosonde temperature data for climate monitoring. **Bull. Amer. Meteor. Soc.**, 83, 2002. 891–899 p.

Free Software Foundation. “Licenses” – GNU project – Free Software Foundation (FSF). 2011. Disponível: <http://www.gnu.org/licenses/> (Acesso em: 24 de Novembro de 2019).

FRIENDLY, M. **Milestones in the history of thematic cartography, statistical graphics, and data visualization**. 2009. Disponível em: <https://www.datavis.ca/milestones/>.

FOKEN, T., GOCKEDE, M., MAUDER, M., MAHRT, L. A. B.; MUNGER, W. Postfield data quality control. **Handbook of Micrometeorology**, 29(1988), 2004, 181–208. https://doi.org/10.1007/1-4020-2265-4_9

FRANÇOZO, R.D., BRANDÃO, R., NOGUEIRA, C.C. et al. Habitat loss and the effectiveness of protected areas in the Cerrado Biodiversity Hotspot. **Nat Conserv** 13:35–40. 2015. <https://doi-org.ez52.periodicos.capes.gov.br/10.1016/j.ncon.2015.04.001>

GIBBS, B.H.K., RAUSCH, L., MUNGER, J., et al. Brazil’s soy moratorium. **Science** 347:377–378. 2015. <https://doi-org.ez52.periodicos.capes.gov.br/10.1126/science.aaa0181>

GILBERT A.; ABRAHAM A.; PAPRZYCKI M. A system for ensuring data integrity in grid environments. In: **Information Technology: Coding and Computing**, 2004. Proceedings. ITCC 2004. International Conference on. vol. 1; p. 435-439 Vol.1.

HAMILTON, S. K.; SIPPEL, S. J.; MELACK J. M. Inundation patterns in the Pantanal of South America determined from passive microwave remote sensing. **Archiv Fur Hydrobiologie**,137(1), 1996. 1-23.

HARYOKO, U. International Workshop On The Digitization Of Historical Climate Data, The New Saca&D Database And Climate Analysis In The Asian Region 02 - 05 April 2012 Citeko, Bogor, Indonesia. 2012. Disponível: http://www.didah.org/images/day%201/UripHaryokoCLIMATE%20DATA%20MANAGEMENT_BMKG_Indonesia.pdf (Acesso em: 24 de novembro de 2019).

HASLAM, SM. **Understanding Wetlands: fen, bog and marsh**. London: Taylor & Francis. 2003.

IMRAN M.; HLAVACS H.; HAQ IU.; JAN B.; KHAN F. A.; AHMAD A. Provenance based data integrity checking and verification in cloud environments”. *PLoS ONE* 12(5): e0177576. 2017. <https://doi.org/10.1371/journal.pone.0177576>

INMET – Instituto Nacional de Meteorologia. 2021. Disponível em <https://portal.inmet.gov.br/sobre-meteorologia>. (Acesso em: 02 de fevereiro de 2021).

INMON, W., H. **Building the Data Warehouse**. 2002. 3rd edition John Wiley & Sons, Inc. ISBN 0-471-08130-2

JONES D. A. et al. An updated analysis of homogeneous temperature data at Pacific Island stations. **Aust. Meteorol. Mag.** 2013. 61: 285–302.

JAROLÍMEK J.; MARTINEC, R. Analysis of Open Data Availability in Czech Republic Agrarian Sector, AGRIS on-line Papers in **Economics and Informatics**, Vol. 8, No. 3, 2016. pp. 57 - 67. ISSN 1804-1930, DOI 10.7160/aol.2016.080306.

JAYAWARDENE, V., SADIQ, S., INDULSKA, M. **An analysis of data quality dimensions**. 2013. ITEE Technical Report.

JOHNSON, E. Handbook on Life Cycle Assessment Operational Guide to the ISO Standards. **Environmental Impact Assessment Review**, 23(1), 2003. 129–130. [https://doi.org/10.1016/S0195-9255\(02\)00101-4](https://doi.org/10.1016/S0195-9255(02)00101-4)

KIMBALL, R. **The data warehouse toolkit: the complete guide to dimensional modelling** — 2nd ed. p. cm. Wiley Computer Publishing. 2002. Includes index. ISBN 0-471-20024-7

LAPOLA, D.M.; MARTINELLI, L.A.; PERES, C.A. et. al. Pervasive transition of the Brazilian land - use system. **Nat Clim Chang**. 2013. 4:27–35. <https://doi-org.ez52.periodicos.capes.gov.br/10.1038/nclimate2056>

LIM, B.; BOILEAU, P. **Methods for assessment of inventory data quality: issues for an IPCC expert meeting**, 1999.

MARTIN, D. J.; HOWARD, A.; HUTCHINSON, R.; MCGREE, S.; JONES, D. A. Development and implementation of a climate data management system for western Pacific small island developing states. **Meteorological Applications** **22**, 2015. pag. 273 – 287, RMets. DOI: 10.1002/met.1461.

MCDOWALL, R.D. Data Integrity Focus, Part 1: Understanding the Scope of Data Integrity. 2019. LC-GC North America, vol. 37, no. 1, p. 44+. Gale Academic Onefile, Disponível:

<https://link.gale.com/apps/doc/A576523526/AONE?u=capes&sid=AONE&xid=6fb5e2a4>. (Acesso em: 03 de dezembro de 2020).

MEIER, F.; FENNER, D.; GRASSMANN, T.; OTTO, M.; SCHERER, D. Crowdsourcing air temperature from citizen weather stations for urban climate research. **Urban Clim**. **19**, 2017. 170–191 p. doi: 10.1016/j.uclim.2017.01.006

Ministério do Meio Ambiente (MMA) - **Florestas do Brasil em Resumo**. 2009. MMA, Federal.

Ministry of Interior, Školení otevřených dat VS ČR“. 2015. Disponível em: http://opendata.gov.cz/_media/edu:skoleni_open_data_web.pdf. (Acesso em: 04 Novembro 2020).

NAPOLY A.; GRASSMANN T.; MEIER F.; FENNER D. Development and Application of a Statistically-Based Quality Control for Crowdsourced Air Temperature Data. **Front. Earth Sci**. **6**, 2018. 118 p. doi: 10.3389/feart.2018.00118

OVERBECK, G.E.; VÉLEZ-MARTIN, E.; SCARANO, F.R. et al. Conservation in Brazil needs to include non-forest ecosystems. **Divers Distrib**. 2015. <https://doi-org.ez52.periodicos.capes.gov.br/10.1111/ddi.12380>.

R Core Team. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. 2018. URL <https://www.R-project.org/>.

RUSTICUCCI, M; BARRUCAND, M. Observed trends and changes in Temperature Extremes over Argentina. **J Climate** **17(20)**. 2004. 4099–4107 p.

SADIQ, S. **Handbook of Data Quality: Research and Practice**, Springer, 2013.

SADIQ, S.; INDULSKA, M. Open data: Quality over quantity. **International Journal of Information Management**, 2017. 37(3), 150–154. <https://doi.org/10.1016/j.ijinfomgt.2017.01.003>.

SANTOS, A.C.A. Absorção da Radiação Solar por partículas de aerossóis no Pantanal Mato-grossense, 107p. (Tese em Física Ambiental) - Instituto de Física, UFMT, Cuiabá, 2018.

SCHUMACHER, R. Dispelling the Top Five Open Source Database Myths. **Database Trends and Applications**. 2010. Disponível: <https://www.dbta.com/Columns/The-Open-DBA/Dispelling-the-Top-Five-Open-Source-Database-Myths-69748.aspx>. (Acesso em: 04 de novembro de 2020).

SCIUTO, G.; BONACCORSO, B.; CANCELLIERE, A.; ROSSI, G. Quality control of daily rainfall data through neural networks. **Journal of Hydrology** **364**. 2009. 13–22 p.

SCIUTO, G.; BONACCORSO, B.; CANCELLIERE, A.; ROSSI, G. Probabilistic quality control of daily temperature data. **International Journal of Climatology** **33**. 2013. 1211-1227 p.

SILVA, D.C.; VIEIRA, T.B.; SILVA, J.M. et al. Biogeography and priority areas for the conservation of bats in the Brazilian Cerrado. **Biodivers Conserv** **27**, 815–828 2018. <https://doi-org.ez52.periodicos.capes.gov.br/10.1007/s10531-017-1464-z>

SLATYER R.O.; BONNER W.N. Review of the Operation of the Bureau of Meteorology. Melbourne: **Australian Bureau of Meteorology**. 1996.

SHEN, J.; YANG, M.; ZOU, B.; WAN, N.; LIAO, Y. Outlier detection of air temperature series data using probabilistic finite state automata-based algorithm. **Complexity**, 17. 2012. 48-57 p. doi:10.1002/cplx.21390.

STEINACKER, R.; MAYER, D.; STEINER, A. Data Quality Control Based on Self-Consistency, **Monthly Weather Review**, 139(12). 2011. 3974–3991 p. doi: 10.1175/MWR-D-10-05024.1.

ŠTĚPÁNEK, P.; ZAHRADNÍČEK, P.; SKALÁK, P. Data quality control and homogenization of air temperature and precipitation series in the area of the Czech Republic in the period 1961–2007, **Adv. Sci. Res.**, 3. 2009. 23-26 p, <https://doi.org/10.5194/asr-3-23-2009>.

STONEBRAKER, M.; KEMNITZ, G. The Postgres Next-Generation Database Management System. **Communications of the ACM**, 34(10). 1991. 78-92 p.

THORNE, P. W.; D. E. P.; TETT, S. F. B.; JONES, P. D.; Mc-CARTHY M.; COLEMAN, H.; BROHAN, P. Revisiting radiosonde upper air temperatures from 1958 to 2002. **J. Geophys. Res.**, 110, 2005, D18105.

TONDATO, K. K.; MATEUS, L. A. F.; ZIOBER, S. R. Spatial and temporal distribution of fish larvae in marginal lagoons of Pantanal, Mato Grosso State, Brazil. **Neotrop. Ichth.** Porto Alegre, v. 8, n. 1, p. 123-134, Mar. 2010. Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1679-62252010000100015&lng=en&nrm=iso>. Epub Feb 05, 2010. <https://doi.org/10.1590/S1679-62252010005000002>.

VICKERS, D.; MAHRT, L. Quality Control and Flux Sampling Problems for Tower and Aircraft Data. **Journal of Atmospheric and Oceanic Technology** 14, 3, 1997. 512-526, Disponível em: < [https://doi.org/10.1175/1520-0426\(1997\)014<0512:QCAFSP>2.0.CO;2](https://doi.org/10.1175/1520-0426(1997)014<0512:QCAFSP>2.0.CO;2)> (Acesso em: 16 de Março de 2021).

WEIDEMA, B. P.; WESNAES, M. S. Data quality management for life cycle inventories-an example of using data quality indicators. **Journal of Cleaner Production**, 1996. 4(3–4), 167–174. [https://doi.org/10.1016/S0959-6526\(96\)00043-1](https://doi.org/10.1016/S0959-6526(96)00043-1)

APÊNDICE

Código das rotinas de qualidade de dados

```
# Impossível climatologicamente verifica se o registro está fora de limites físicos#
# x é a coluna inicial dos dados, já que o DF pode conter coluna de data
# df1 é o DataFrame com os limites a serem analisados
# limite é o DF de retorno da função
#' @export

fisico <- function(arq1,df1,x)

{

  limite <- arq1

  z<-1

  y<-1

  h<-ncol(arq1)

  c<-h+1

  limite[,c]<-0

  for (j in x:h)

  {

    for (i in 1:nrow(arq1))

    {

      if (arq1[i,j] > df1[z,y+1] || arq1[i,j] < df1[z,y])

      {

        limite[i,c] <- 1

      }

    }

  }

  z<-z+1
```

```
c<-c+1

if (j+1 <= h)
{
  limite[,c]<-0
}
}

return(limite)
}

#' @export
fisico2 <- function(arq1,df1,x)
{
  limite <- arq1
  z<-1
  y<-1
  h<-ncol(arq1)
  c<-h+1
  limite[,c]<-0
  for (j in x:h)
  {
    limite[arq1[,j] < df1[z,1],c] <- 1
    limite[arq1[,j] > df1[z,2],c] <- 1
  }
  z<-z+1
}
```

```

    c<-c+1
  }
  return(limite)
}

#####
#####

##### Verificação de valores zerados ou iguais consecutivos
#####

#' @export
equalval <- function(arq1,zero,x)
{
  zerado <- arq1
  h<-ncol(arq1)+1
  z<-1 # z controla o andamento do DF de valores zerados informado #
  for(j in x:ncol(arq1))
  {
    for (i in 1:(nrow(arq1)-1))
    {
      ## Verifica se os valores são zerados ##
      if (arq1[i,j] == zero[z])
      {
        zerado[i,h] <- 2
        next
      }
    }
  }
}

```

```

if (i == (length(arq1)-1))
{
  if (arq1[i,j] == zero[z])
  {
    zerado[i,h] <- 2
  }
}

## Verifica se os valores são iguais ao próximo ##

if (arq1[i,j] == arq1[i+1,j])
{
  zerado[i+1,h] <- 1
}

}

h<-h+1 ## Incrementa o índice para criar a próxima coluna para aplicar a marcação
da verificação ##

z<-z+1 ## Incrementa o índice para comprar o zerado da próxima variável ##

}

return(zerado)

}

##### Verificação de valores zerados ou iguais consecutivos
#####

#' @export

equalval2 <- function(arq1,zero,x)

```

```

{
  zerado <- arq1

  h<-ncol(arq1)+1

  z<-1 # z controla o andamento do DF de valores zerados informado #

  for(j in x:ncol(arq1))
  {
    zerado[,h] = NA

    pos = c(FALSE, diff(arq1[,j]) == 0)

    pos[is.na(pos)] = FALSE

    zerado[pos, h] = 1

    pos = zerado[,j] == zero[z]

    pos[is.na(pos)] = FALSE

    zerado[pos, h] = 2

    h<-h+1 ## Incrementa o índice para criar a próxima coluna para aplicar a marcação
da verificação ##

    z<-z+1 ## Incrementa o índice para comprar o zerado da próxima variável ##

  }

  return(zerado)

}

#####
#####

```

```
##### Verificação de Outliers #####

#' @export

outlier <- function(arq1,x) #x é a variável para identificar a partir de qual coluna a
análise deve acontecer #

{

  out <- arq1

  h<-ncol(arq1) # identifica quantas colunas o arquivo passado possui #

  c<-h+1 # controle para inserir colunas no final do DF de saída #

  for (j in x:h)

  {

    dp <- sd(arq1[,j],na.rm=TRUE)

    md <- mean(arq1[,j],na.rm=TRUE)

    out[,c]<-ceiling(abs(arq1[,j]-md)/dp) # realiza o cálculo para identificar a quantas
vezes o valor está do desvio padrão do conjunto em análise #

    c<-c+1

  }

  return(out)

}

#####

##### Verificar quais datas estão com registro falho #####

#' @export

falha <- function(arq1,dat,med)

{
```

```
int_medida <- med # Quantidade de medidas diárias esperada #
lista <- NA # Cria lista de datas com número de registros incompletos #
i <- 1
c <- dat # Coluna com a data
while (i < nrow(arq1))
{
  cont <- 0
  j <- i
  while (arq1[i,c] == arq1[j,c])
  {
    cont <- cont+1 # Contador de registros no dia #
    j <- j+1
    if (is.na(arq1[j,c]))
    {
      if (cont != int_medida)
      {
        if (lista == 0)
        {
          lista <- arq1[j,c]
        }
        if (lista != 0)
        {
          lista <- rbind(lista,arq1[j,c])
        }
      }
    }
  }
}
```

```

    }
  }
  break
}
}

if (cont != int_medida) # Se encontrou intervalo de data com menos medidas que o
esperado #
{
  if (is.na(lista)) # Verifica se a lista de falhas está vazia #
  {
    lista <- arq1[j-1,c] # lista de ocorrência recebe o registro anterior, identificando o
dia com falta de dados #
  }else
  {
    lista <- rbind(lista,arq1[j-1,c])
  }
}
i<-j
}

return(data.frame(lista)) # retorna o DF com a lista de datas com falta de dados #
}

```

```

#' @export

```

```

falha2 <- function(arq1,dat,med)

```

```

{

```

```
int_medida <- med # Quantidade de medidas diárias esperada #  
lista <- NA # Cria lista de datas com número de registros incompletos #  
i <- 1  
c <- dat # Coluna com a data  
h = dim(arq1)[2]+1  
res = ftable(arq1)  
soma = rowSums(res, na.rm = TRUE)  
  
colunas = colnames(arq1)  
dias = levels(sub[[colunas[c]]])  
  
return (dias[soma!=int_medida])  
  
#return(data.frame(lista)) # retorna o DF com a lista de datas com falta de dados #  
}
```