MINISTÉRIO DA CIÊNCIA, TECNOLOGIA, INOVAÇÕES E COMUNICAÇÕES
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

aa/bb/cc/dd

APLICAÇÕES DE FERRAMENTAS COMPUTACIONAIS NA QUALIDADE DE DADOS METEOROLÓGICOS OBSERVACIONAIS DE MULTI-SENSORES SOBRE A REGIÃO AMAZÔNICA

Thomaz Assaf Pougy

Relatório Final de Iniciação Científica PIBIC, orientada pelos Dr. Alan James Peixoto Calheiros e Prof. Dr. Pedro Luiz Pizzigatti Corrêa, submetido em 13 de agosto de 2021.

URL do documento original:
<http://urlib.net/xx/yy>

INPE
São José dos Campos
2021

PUBLICADO POR:

Instituto Nacional de Pesquisas Espaciais - INPE
Gabinete do Diretor (GB)
Serviço de Informação e Documentação (SID)
Caixa Postal 515 - CEP 12.245-970
São José dos Campos - SP - Brasil
Tel.:(012) 3945-6923/6921
Fax: (012) 3945-6919
E-mail: pubtc@sid.inpe.br

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA, INOVAÇÕES E COMUNICAÇÕES
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

aa/bb/cc/dd

APLICAÇÕES DE FERRAMENTAS COMPUTACIONAIS NA QUALIDADE DE DADOS METEOROLÓGICOS OBSERVACIONAIS DE MULTI-SENSORES SOBRE A REGIÃO AMAZÔNICA

Thomaz Assaf Pougy

Relatório Final de Iniciação Científica PIBIC, orientada pelos Dr. Alan James Peixoto Calheiros e Prof. Dr. Pedro Luiz Pizzigatti Corrêa, submetido em 13 de agosto de 2021.

URL do documento original:
<http://urlib.net/xx/yy>

INPE
São José dos Campos
2021

Dados Internacionais de Catalogação na Publicação (CIP)

Sobrenome, Nomes.

Cutter Aplicações de Ferramentas Computacionais na Qualidade de Dados Meteorológicos Observacionais de Multi-Sensores Sobre a Região Amazônica / Nome Completo do Autor1; Nome Completo do Autor2. – São José dos Campos : INPE, 2021.

xiv + 49 p. ; (aa/bb/cc/dd)

Dissertação ou Tese (Mestrado ou Doutorado em Nome do Curso) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, AAAA.

Orientador : José da Silva.

1. Palavra chave. 2. Palavra chave 3. Palavra chave. 4. Palavra chave. 5. Palavra chave I. Título.

CDU 000.000

ATENÇÃO! A FOLHA DE APROVAÇÃO SERÁ INCLUIDA POSTERIORMENTE.
Mestrado ou Doutorado em Nome do Curso

RESUMO

O INPE produz importantes pesquisas que ajudam a compreender as dinâmicas climáticas e de tempo no Brasil e no mundo, com impactos significativos no planejamento estratégico público e privado nacional. Entre as informações essenciais para os estudos supracitados estão os dados pluviométricos. Nesse contexto, garantir a qualidade desses dados impacta diretamente sobre a confiabilidade das previsões e análises geradas a partir deles. Dessa forma, este trabalho, que é uma parceria entre o INPE, o Laboratório de Física Atmosférica e Escola Politécnica da USP e o ARM-DoE (Atmospheric Radiation Measurement Climate Research Facility), visou estabelecer ferramentas computacionais que pudessem tratar da qualidade de dados de chuva em conformidade com as principais diretivas internacionais. Assim, propôs-se para este estudo o desenvolvimento de um toolkit específico para dados do instrumento Micro Rain Radar (MRR) que auxiliasse pesquisadores do INPE, USP e parceiros a: padronizar a preparação de dados brutos para formatos internacionalmente aceitos; processar figuras para subsidiar análises rápidas; analisar e tratar a qualidade de dados e, por fim, registar metadados e análises de qualidade para publicação em repositórios de dados internacionais, como o do ARM (EUA) e do instituto Max Planck (Alemanha). Foram desenvolvidos scripts e bibliotecas em Python que convertem os dados brutos do instrumento para o formato netCDF4, em conformidade com as diretrizes de estrutura e qualidade de dados do ARM para alguns experimentos de campo no Brasil. Produziu-se também algoritmos de visualização interativas e estáticas dos dados, que auxiliam principalmente na análise rápida da integridade dos dados pelos mentores dos equipamentos e pesquisadores. Outro aspecto importante desta pesquisa foi a elaboração de documentos python do tipo notebook explicativos e pré-organizados para apoiar a exploração e análise dos dados, com destaque para o cálculo de estatísticas analíticas (e.g., RMSE, correlações e outros) e diagramas que auxiliam na interpretação destas informações. Ainda, tendo em vista o registro de metadados e análises de qualidade foi elaborada uma proposta de arquitetura para um sistema de coleta, armazenamento e apresentação de relatórios de qualidade de dados, que foi descrita em termos de requisitos de interface, sistema e banco de dados. Por fim, com as ferramentas desenvolvidas, foi possível inicialmente avaliar a performance das medidas do MRR durante o experimento de campo SOSCHUVA. Observou-se que o MRR apresenta alta correlação com as medidas de taxa de chuva capturadas pelos pluviômetros (0,74) e disdrômetros (0,87). Contudo, foi observado uma subestimativa máxima de 0,3 mm/h, indicando que o instrumento apresentou boa performance.

Palavras-chave: Ciência dos dados. Microfísica de nuvens. Qualidade de dados.

COMPUTATIONAL TOOLS APPLICATION IN DATA QUALITY FOR OBSERVATIONAL MULTI-SENSOR METEOROLOGICAL DATA ON THE AMAZON BASIN

ABSTRACT

INPE produces important researches that help to understand climate and weather dynamics in Brazil and in the world, with significant impacts on national public and private strategic planning. Among the essential information for the aforementioned studies there is rainfall data. In this context, ensuring the quality of these data directly impacts the reliability of the forecasts and analysis conclusions generated from them. Thus, this work, which is a partnership between INPE, the Laboratory of Atmospheric Physics and Polytechnic School of USP and the ARM-DoE (Atmospheric Radiation Measurement Climate Research Facility), aimed to establish computational tools that could address the quality of data from rain in accordance with the main international directives. Thus, it was proposed for this study to develop a specific toolkit for data from the Micro Rain Radar (MRR) instrument that would help researchers from INPE, USP and partners to: standardize the preparation of raw data for internationally accepted formats; process figures to support quick analysis; analyze and treat data quality and, finally, register metadata and quality analysis for publication in international data repositories, such as the ARM (USA) and the Max Planck Institute (Germany). Python scripts and libraries that convert the instrument's raw data to netCDF4 format were developed, in compliance with the ARM data structure and quality guidelines for some field experiments in Brazil. Interactive and static data visualization algorithms were also produced, which mainly help in the quick analysis of data integrity by equipment mentors and researchers. Another important aspect of this research was the elaboration of explanatory and pre-organized notebook-type Python documents to support the exploration and analysis of the data, with emphasis on the calculation of analytical statistics (eg, RMSE, correlations and others) and diagrams that assist in the interpretation of this information. Also, with a view to recording metadata and quality analysis, an architecture proposal was developed for a data quality report collection, storage and reporting system, which was described in terms of interface, system and database requirements . Finally, with the developed tools, it was initially possible to evaluate the performance of MRR measurements during the SOSCHUVA field experiment. It was observed that the MRR has a high correlation with the rainfall rate measurements captured by the pluviometers (0.74) and disdrometers (0.87). However, a maximum underestimation of 0.3 mm/h was observed, indicating that the instrument performed well.

Keywords: Data Science. Cloud microphysics. Data Quality.

LISTA DE FIGURAS

Pág.

LISTA DE TABELAS

Pág.

SUMÁRIO

Pág.

1 INTRODUÇÃO

O estudos meteorológicos conduzido por centros de excelência em pesquisa no Brasil é fundamental para evoluir o conhecimento da ciência sobre fenômenos tempo e clima de biomas brasileiros que podem ter consequência globais ou locais. Nesse contexto, pesquisas baseadas na coleta de dados pluviométricos e de aerossóis são importantes para modelar de forma acurada os fenômenos climáticos estudados. Contudo, a validade desses modelos depende diretamente da qualidade dos dados utilizados.

Garantir a qualidade dos dados é um requisito essencial para a gestão de dados. Esse requisito é um desafio importante no cenário da pesquisa atmosférica, dado o grande fluxo arquivos recebidos por campanha de coleta de dados e informações operacionais.

Assim, o desenvolvimento de novas ferramentas para aprimorar e otimizar a gestão de qualidade dados, não só trará agilidade ao fluxo de trabalho, mas garantirá também boas condições para a futura disponibilidade dos dados para outras pesquisas. A gestão de qualidade dos dados auxilia também no acesso a metadados robustos de qualidade, o qual é condição importante para garantir a fácil recuperação, o acesso e a reutilização dos dados de pesquisa (WILKINSON, 2016).

A gestão de dados, com atenção especial à qualidade, já é empregada no cenário internacional de instituições de pesquisa atmosféricas, no qual destaca-se o *Atmospheric Radiation Measurement Climate Research Facility* (ARM/ARM-DoE) (ARM, 2017). O ARM é um programa do governo norte-americano que desenvolve estudos sobre o clima e possui um departamento específico para determinar diretrizes para gestão de qualidade dos dados. O ARM recebe dados de pesquisas conduzidas pelo próprio instituto ou por instituições parceiras em outros países em seu sistema de *Data Delivery* e agrega ao total cerca de 10PB de dados.

1.1 Objetivos

Estamos na era dos dados, big data se tornou uma das palavras mais citadas por institutos de pesquisa na atualidade. O INPE realiza junto com seus parceiros uma série de experimentos de campo para entender as características físicas da atmosfera de modo a melhorar suas previsões de tempo e clima, assim como, para validar dados de satélites. Contudo, ainda existe um longo caminho até que esta gama de dados esteja pronta para responder questões científicas. Deste modo, este projeto

tem como objetivo propor e implementar ferramentas de gestão, controle e análise da qualidade de dados obtidos por sensores que coletam dados tanto na superfície, com dentro da atmosfera. Neste caso refere-se a informações sobre os aerossóis e a chuva na Bacia Amazônica principalmente.

Os dados utilizados neste trabalho são baseados no projeto temático GOAmazon (Green Ocean Amazon) que estuda as características físicas da atmosfera na região amazônica, baseado numa parceria internacional da qual o INPE é um dos líderes. E informações de outros experimentos, como o SOS-CHUVA (MACHADO, 2015).

Os objetivos específicos deste projeto são propor e implementar ferramentas computacionais que melhorem o fluxo de trabalho para gestão de qualidade dos dados, obtidos através de diversos sensores instalados em diversos sítios experimentais, principalmente o GoAmazon. A princípio serão implementadas as três ferramentas apresentadas na listagem abaixo, contudo é possível que novas demandas sejam identificadas e mais ferramentas sejam implementadas. De modo a seguir os padrões internacionais de qualidade das informações geradas por sensores para ciências atmosféricas, este trabalho será baseado no sistema de coleta, processamento e qualidade de dados do ARM.

A. Ferramenta de coleta, armazenamento e apresentação de Relatórios de Qualidade dos Dados.

   Os Relatórios de Qualidade dos Dados, são um dispositivo já largamente utilizado no ARM para garantir o acesso do usuário final à dados robustos quanto ao tratamento, além das validações de qualidade os quais um determinado conjunto de dados foi submetido.

   i. Foi implementada uma ferramenta para que o pesquisador, responsável pela validação de qualidade de um determinado sensor, possa registrar informações sobre quais verificações de qualidade foram aplicadas, as conclusões obtidas e eventuais correções utilizadas. Note que a implementação da ferramenta perpassa pela modelagem dos metadados de qualidade a serem coletados.
   ii. Foi modelado e implementado um banco de dados para armazenar os metadados de qualidade submetidos pelos pesquisadores através ferramenta implementada.
   iii. Foi implementada uma ferramenta que entrega ao usuário final os metadados de qualidade para o dado escolhido por ele para download. A princípio será

desenvolvida uma versão beta desta ferramenta, haja vista que não há um portal de *Data Delivery* ainda disponível para tal propósito.

B. Ferramenta para aplicação de *flags* de qualidade e condensação de informações de verificação de qualidade por meio de palavras binárias.

   Conhecida como *Flagging by Bit Packing* (ARM, 2015), essa técnica de armazenamento de informações já é utilizada atualmente no ARM e permite que um conjunto de dados passe por diversos níveis de validação de qualidade e receba *flags* a cada etapa de processamento para linhas de dado aprovadas ou reprovadas. Essa técnica evita a exclusão de dados entre níveis de validação, evita a criação demasiada de colunas no conjunto de dados e também permite mais controle do usuário sobre os dados que utilizará.

   i. Foi implementada uma ferramenta que permite a inclusão de *flags* de qualidade por linhas em um determinado conjunto de dados. A ferramenta utilizou, como susodito, a técnica de *Bit Packing* para a criação das *flags*.

   ii. Foi implementada uma ferramenta que permite o registro, no cabeçalho do conjunto de dados, da descrição sobre os bits das *flags* incorporadas nas linhas.

C. Ferramenta de geração de arquivo NetCDF

   O formato de arquivo NetCDF (*Network Common Data Form*) foi desenvolvido pela organização UNIDATA e é largamente utilizado para o compartilhamento de arquivos de dados científicos. O formato permite encapsular os dados junto com cabeçalhos mais robustos que descrevem com mais detalhes características importantes dos dados (REW; DAVIS, 1990). Atualmente é utilizado no ARM como principal formato de arquivo para distribuição de dados.

   i. Foi desenvolvida uma ferramenta que recebe um arquivo de metadados e um arquivo CSV ou ascii (dependendo do instrumento) e encapsula as informações de ambos em um mesmo arquivo NetCDF.

   ii. Foi implementada uma ferramenta que recebe um arquivo NetCDF e retorna um arquivo de metadados e um arquivo CSV contendo os dados de medição em si.

Os dados utilizados nesta pesquisa foram aqueles associados a sensores que determinam a distribuição de aerossóis e gotas de chuva na superfície (disdrômetros de

superfície e um micro rain radar). Estes dados estão disponíveis nos repositórios do ARM, USP e INPE.

2 DESENVOLVIMENTO

2.1 Revisão Bibliográfica

A fim de subsidiar a elaboração das metodologias que guiaram os trabalhos deste projeto e também contribuir no desenvolvimento delas, foram consultadas referências para auxiliar na exploração de conceitos fundamentais para a correta compreensão de desafios e tarefas a serem abordadas neste estudo. Nesse cenário, é possível citar a gestão de qualidade de dados, estrutura dos dados do instrumento Micro Rain Radar e também relações de dados de aerossol com dados de precipitação. Os tópicos 2.2 registram os principais pontos da revisão bibliográfica.

2.2 Gestão de Dados e Gestão de Qualidade dos Dados

Garantir boas práticas de gerenciamento em todo o ciclo de vida de dados desde a ingestão até o processamento e análise perpassa por adotar um claras diretrizes de Gestão de Dados. Essa premissa se mostra ainda mais importante no cenário da pesquisa científica, no qual os dados capturado subsidiam conclusões importantes sobre sociedade e meio ambiente orientando a tomada de decisão de indivíduos e instituições.

Como abordado por Silva (2020), a gestão de dados:

> Refere-se àquelas atividades relacionadas à gestão ativa de dados durante o tempo que continuam a ter interesse acadêmico, científico, administrativo e pessoal, a fim de favorecer sua reprodução, reutilização e agregação de valor, os dados são gerenciados desde a sua criação até que é determinado que eles não são mais úteis, garantindo a sua acessibilidade a longo prazo, sua conservação, sua autenticidade e sua integridade.

Assim, adotar boas práticas de curadoria, gerenciamento e armazenamento dos dados faz parte da definição de uma estratégia robusta de Gestão de Dados a ser adotada.

Como apresentado por Anjos G. A. Dias (2017) no contexto acadêmico brasileiro ainda há uma parcela significativa de pesquisadores, correspondente a 31% da comunidade, que não aplicam plenamente diretrizes robustas de gestão de dados sobre as informações capturadas em suas pesquisas. Dessa forma, iniciativas que consideram a definição de uma estratégia de gestão de dados como parte do objetivo final de projeto contribuem para disseminar a metodologia de gestão de dados no cenário da pesquisa Brasileira.

Ademais, é importante esclarecer que um dos aspectos a ser considerado na elabora-

ção de estratégia de gestão de dados deve ser a gestão de qualidade de dados. Ela é constituída pelas práticas de análise de qualidade e gerenciamento de metadados de qualidade. Como analisado pela referência Kwon Ohbyung; Lee (2014) o emprego de diretrizes claras de gestão de qualidade de dados em ambientes com abundância de dados é um fator que potencializa a execução de análises mais complexas sobre os dados, incentivando a avaliação de relações múltiplas entre variáveis de origem distintas.

2.3 Dados de precipitação

Dentre os dados utilizados neste trabalho estão aqueles associados a medidas de distribuição de gostas de chuva. Entre eles estão o disdrômetro joss-waldvoguel (JOSS; WALDVOGEL, 1967) e o equipamento pluviômetro, os quais foram utilizados para validar o principal objeto de estudo desta pesquisa, o MRR (Micro Rain Radar), que será descrito com maiores detalhes na seção 2.3.1.

2.3.1 Equipamento Micro Rain Radar

O Micro Rain Radar (MRR) consiste em um equipamento de estimativa do perfil de parâmetros de chuva por meio da emissão de ondas no espectro do micro-ondas. O MRR é capaz de registrar dados da taxa de chuva, distribuição do tamanho de gotas, refletividade de radar, conteúdo de água líquida e velocidade de queda (METEK, 2011).

O princípio de funcionamento do MRR se baseia no espalhamento da radiação emitida centrada na frequência de 24GHz. Dessa forma, dado que a radiação é emitida verticalmente para a atmosfera, parte dela é espalhada pelas gotas de chuva e retorna à antena do radar. Uma vez que as gotículas responsáveis pelo espalhamento das micro ondas estão em movimento de queda, há um desvio na frequência das ondas de retorno devido ao efeito Doppler. Como as gotas com tamanho diferentes possuem velocidade de queda distintas, o sinal refletido consiste em uma distribuição de diferentes frequências Doppler. Portanto, a análise espectral do sinal recebido permite identificar a potência captada para cada uma das frequências da distribuição e assim calcular os parâmetros supracitados (METEK, 2011).

A resolução temporal das medidas do equipamento é de 10 segundos. Ademais, a emissão de radiação ocorre com modulação de frequência, de forma que o sinal refletido para cada frequência modulada permite calcular os parâmetros susoditos para diferentes perfis de altura da atmosfera. A resolução vertical do feixe para o

equipamento varia de 10 metros a 300 metros no perfil (METEK, 2011).

2.3.1.1 Estrutura dos dados do MRR

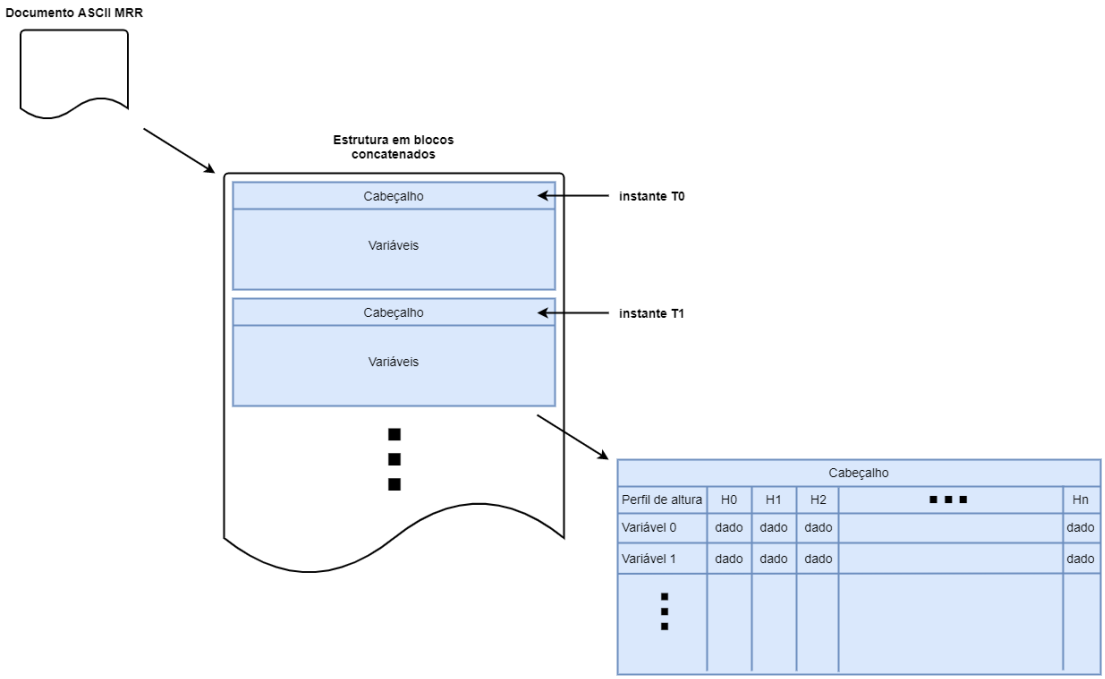
O MRR é capaz de exportar 3 tipos de arquivos que registram as medidas capturadas: arquivos *average* (extensão .ave) que armazenam as medidas integradas em um período de 1 minuto; arquivos *processed* (extensão .pro) que contém as medidas diretamente na resolução temporal nativa do equipamento; e também arquivos *raw* (extensão .raw) que salvam as medidas brutas de refletividade, sem incluir as variáveis calculadas.

O objeto principal de estudos deste trabalho são os arquivos *average*, que são o principal produto do MRR utilizado em pesquisa para caracterizar a precipitação (PETERS et al., 2005). Porém, os arquivos *processed* e *raw* também foram considerados na elaboração dos scripts de processamento de dados de forma a potencializar o uso real do *toolkit* proposto em trabalhos futuros.

Para que seja possível armazenar os dados capturados pelo instrumento em um arquivo netCDF, se faz necessário extrair os dados de cada variável a partir dos formatos de arquivo específicos do equipamento. Nesse contexto, é fundamental entender a estrutura desse formato específico.

Todos os arquivos de dados exportados pelo MRR são arquivos de texto ASCII e a estrutura geral é baseada em blocos registrados para cada instante medido. Nesses, a primeira linha corresponde ao cabeçalho com informações temporais do instante da medida, bem como informações do próprio equipamento. A segunda linha corresponde aos perfis de altura configurados para a operação do radar e cada qual das linhas seguintes correspondem a uma variável medida nos diferentes perfis de altura. Cada linha de variável divide as medidas de cada perfil com espaços. A Figura 2.1 apresenta um diagrama que resume a estrutura do arquivos de dados do MRR.

Figura 2.1 - Estrutura do arquivo bruto do equipamento MRR



Fonte: Adaptada de METEK (2011).

A partir da referencia METEK (2011) foi possível elaborar a Tabela 2.1 que apresenta as principais informações de cada variável registrada nas linhas de cada bloco do arquivo.

Os arquivos do tipo *raw* seguem um arranjo idêntico à estrutura supracitada, porém, como dito, não incluem as variáveis calculadas. Dessa forma, nesses arquivos há apenas as variáveis TF e Fnn.

2.4 Dados de Aerossol

Dados de aerossol possuem possuem significado e exigem uma análise diferente de dados de precipitação, os quais são o principal objeto de estudos deste trabalho. Contudo, a natureza do dados é idêntica (dados numéricos) e o formato dos arquivos brutos produzidos pelos equipamentos de coleta apresentam semelhanças significativas, de forma que trechos do processo de preparação de dados (*data prep*) de ambos utilizam lógicas parecidas.

Assim, este trabalho aproveita essas similitudes e utiliza como referência scripts de preparação de dados de aerossol para basear o desenvolvimento de ferramentas de

Tabela 2.1 - Variáveis por linha em bloco de arquivo bruto do equipamento MRR

Linha	Identificador	Significado
1	MRR	*Header Line*
2	H	*Height*
3	TF	*Transfer function*
4 - 68	Fnn (F00 - F63)	*Spectral Reflectivities*
69 - 132	Dnn (D00 - D63)	*Drop Size*
133 - 196	Nnn (N00 - N63)	*Spectral Drop Size*
197	PIA	*Path Integrated Attenuation*
198	Z	*Radar Reflectivity*
199	z	*Attenuated Radar Reflectivity*
200	RR	*Rain Rate*
201	LWC	*Liquid Water Contents*
202	W	*Fall Velocity*

Fonte: Adaptada de METEK (2011).

preparação de dados de precipitação (com atenção especial ao equipamento MRR). Não há reaproveitamento de código diretamente, porém estratégias de leitura de dados, conversão de datas, concatenação de tabelas e outros aspectos gerais serviram de ponto de partida para os trabalhos realizados.

2.4.1 Evolução dos scripts anteriores

Este projeto se insere como uma parceria do INPE com o Laboratório de Física Atmosférica do Instituto de Física da USP, Grupo de Gestão de Dados do ARM-DoE e o Grupo de Pesquisa em Big Data liderado pelo Prof. Dr. Pedro Luiz Pizzigatti Corrêa (PCS/EPUSP). Nesse contexto, considerando as condições explicitadas no tópico 2.4 os trabalhos propostos neste estudo se caracterizam como uma evolução dos trabalhos realizados no período de 2019 a 2020 pelo autor, no cenário de processamento de dados de aerossol para o Laboratório de Física Atmosférica do Instituto de Física da USP.

O trabalho anterior produziu scripts Python de preparação de dados para 4 sensores de aerossol atmosférico e esses resultados foram de grande valia para acelerar os desenvolvimento das ferramentas propostas neste estudo servindo de ponto de partida na elaboração da lógica de conversão de formato de arquivos e outras questões de manejo de dados.

Ademais, o projeto anterior também teve importância na medida em que explorou

pela primeira vez as práticas de gestão de dados utilizadas no ARM-DoE e identificou aquelas que possuíam maior potencial de impacto positivo se implementadas no cenário da pesquisa atmosférica brasileira. Essa análise foi utilizada como referência na elaboração dos objetivos deste trabalho, que incluem ferramentas de gerenciamento de relatórios de qualidade de dados e incorporação de *flags* de qualidade nos arquivos processados, ambos presentes na análise supracitada.

2.5 Metodologia

A elaboração de cada ferramenta seguiu diretrizes de metodologia bem estabelecidas, a fim de garantir a eficácia no processo de desenvolvimento. Assim, cada aspecto principal relacionado a produção das ferramentas teve práticas e métodos definidos.

2.5.1 Diretrizes gerais

Para guiar a elaboração das ferramentas propostas neste estudo foi considerada uma diretriz geral para alinhar as diversas frentes de trabalho. Os dois pontos principais que compõem a diretriz são apresentados abaixo.

- O produto processado pelas ferramentas deve estar em acordo com padrões e normas internacionais, tendo em vista a facilitação da publicação destes em repositórios internacionais. Nesse contexto, o padrão de referência a ser utilizado são as práticas e estrutura de gestão de dados do ARM-DoE.

- O desenvolvimento das ferramentas seguiu a metodologia em cascata, tradicional para o desenvolvimento de software. Assim, o desenvolvimento das ferramentas que compõem o Toolkit proposto partiu da prototipação da lógica para seguir à implementação dos scripts em um formato amigável à sua execução automática.

Os próximos tópicos apresentam questões específicas do desenvolvimento do projeto que se baseiam na diretriz geral supracitada.

2.5.2 Gestão de dados

Para garantir robustez e confiabilidade aos scripts de processamento de dados desenvolvidos foi elaborada uma metodologia de gestão de dados, a fim de especificar tanto o processamento a qual os dados são submetidos quanto padronizar aspectos fundamentais como nome de arquivos, variáveis e formato de metadados.

O registro da gestão de dados para o cenário específico de cada equipamento foi feito pela elaboração de um *Handbook* (Desenvolvido pelo orientador principal em parceira com o bolsista deste projeto) do equipamento de forma a suportar as atividades do mentor e usuários do equipamento durante campanhas de coleta de dados. Este *Handbook* foi inicialmente desenvolvido para as atividades associadas as medidas realizadas no sítio da campina a 10km da torre ATTO na região amazônica.

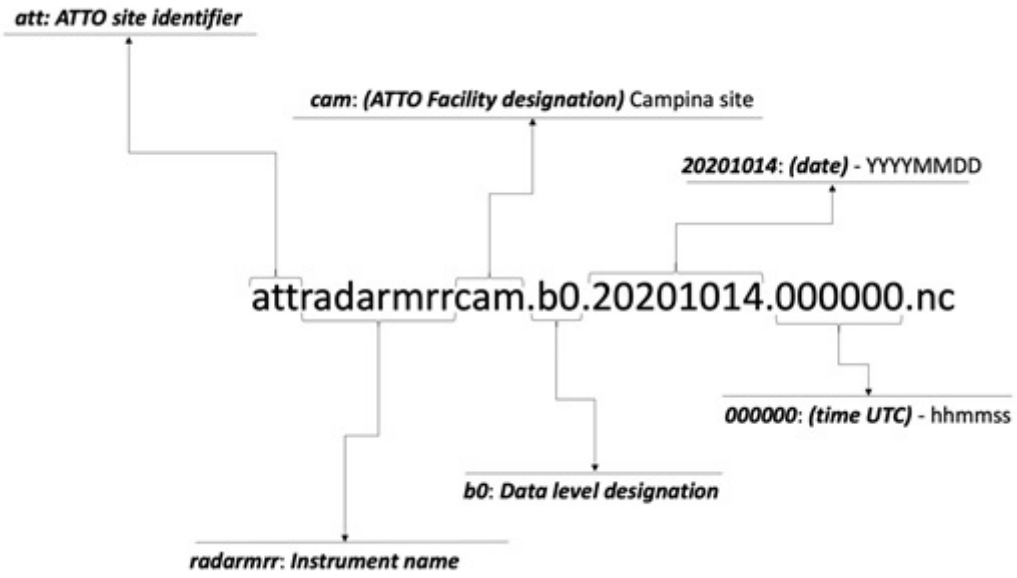
2.5.2.1 Formato de arquivo

O formato de arquivo proposto por este trabalho para o armazenamento e distribuição de dados é o netCDF, versão 4. Este formato de arquivo, criado pela Unidata na década 90, possui grande relevância na comunidade científico-atmosférica internacional e implementa uma abstração de dados que modela um *dataset* como uma coleção de variáveis multidimensionais, acompanhadas de seus metadados complementares (REW; DAVIS, 1990).

O formato netCDF4 é o principal padrão de arquivo utilizado para a distribuição de dados no ARM-DoE. Ele permite a integração direta de dados com informações adicionais, característica especialmente importante que garante maior confiabilidade na entrega de metadados.

Ademais, baseado no modelo implementado pela referência ARM (2016) foi elaborado um esquema de nomenclatura de arquivos que segue a especificação apresentada pela figura 2.2.

Figura 2.2 - Esquema de nomenclatura de arquivos para o instrumento MRR



Fonte: Elaborada pelo autor.

2.5.2.2 Dimensões e variáveis

A definição das variáveis de interesse a serem incluídas em cada arquivo e suas informações complementares foi elaborada pelo orientador deste trabalho que também atua como mentor do equipamento MRR, como citado no item 2.5.2. A Tabela 2.2 apresenta as dimensões das variáveis capturadas pelo instrumento e a tabela 2.3 exibe as variáveis, suas dimensões, resolução temporal, unidades e descrição.

Tabela 2.2 - Dimensões das variáveis capturadas pelo instrumento MRR

Dimensão	**Descrição**
time	Dimensão temporal
profile	Dimensão de perfil de altura
Ndrop	Dimensão de perfil espectral

Fonte: Elaborada pelo autor.

Tabela 2.3 - Variáveis incluídas no arquivo netCDF para dados do equipamento MRR

Símbolo	**Nome**	**Unidade**
lat	*Latitude*	*degrees east*
lon	*Longitude*	*degrees north*
alt	*Altitude*	*meters above sea level*
base_time	*Base time in epoch*	*seconds since 1907-1-1 00:00:00*
time_offset	Time offset from base time	*seconds since base time*
time	*Time offset from midnight GMT*	*seconds since midnight*
mdq	*MRR Data Quality*	*%*
height_mrr	*Height above ground*	*M*
TF_mrr	*Transfer Function*	*none*
Fnn_mrr	*Spectral Reflectivities*	*dB*
Dnn_mrr	*Drop Size Diameter*	*mm*
Nnn_mrr	*Spectral Drop Densities*	*m^(-3)*mm^(-1)*
PIA_mrr	*Path Integrated Attenuation*	*dB*
Zdb_mrr	*Radar Reflectivity*	*dBZ*
Zdb_att_mrr	*Attenuated Radar Reflectivity*	*dBZ*
rain_rate_mrr	*Rain Intensity*	*mm*h^(-1)*
liq_water_mrr	*Liquid Water Content*	*g*m^(-3)*
fall_vel_mrr	*Fall Velocity*	*m*s^(-1)*

Fonte: Elaborada pelo autor.

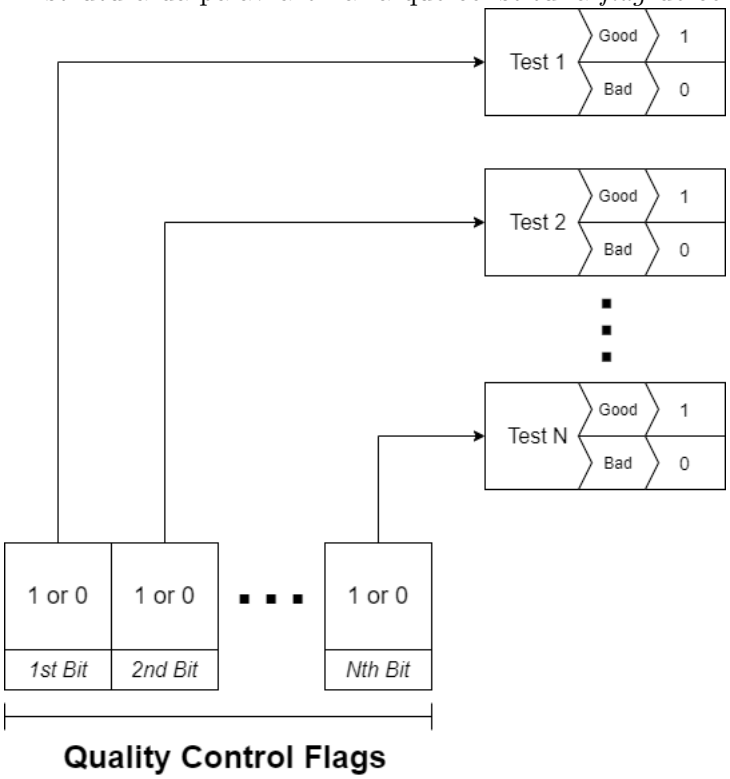
2.5.3 Metadados

Os metadados são dados sobre dados, eles trazem informações complementares relevantes ao dado de interesse, como localização, detalhes do equipamento de medida e outros. Os metadados de qualidade são declarações que trazem informações sobre a qualidade dos dados a que se referem. No esquema de gestão de dados proposto neste trabalho são previstos os metadados gerais do arquivo e também dois principais metadados de qualidade: *flags* de qualidade e os relatórios de qualidade de dados. A distribuição dos metadados gerais do arquivo será implementada diretamente no arquivo netCDF.

As *flags* de qualidade representam o resultado de um teste de qualidade aplicado linha a linha e foram incluídas diretamente no arquivo netCDF. Vale destacar que, baseado nas práticas do ARM DoE, optou-se por fazer a inclusão de multiplas *flags* por meio da técnica de *bit-packing*. Quando aplicada, esta técnica consiste na inclusão de uma nova coluna de dados e em cada linha desta é inserida uma palavra

binária cujos bits registram, individualmente, o resultado de um teste de controle de qualidade aplicado à linha (0 representa reprovado e 1 aprovado). Assim, por exemplo, é possível com uma palavra de 3 bits registrar o resultado da aplicação de 3 testes de qualidade. Note que neste caso, apenas uma coluna é inserida no conjunto de dados, de forma a otimizar o armazenamento e organização dos arquivos. A Figura [2.3 apresenta um diagrama que resume a estrutura de cada *flag* de controle de qualidade.

Figura 2.3 - Estrutura da palavra binária que constitui a *flag* de controle de qualidade



Fonte: Elaborada pelo autor.

Os relatórios de qualidade de dados são documentos que congregam um conjunto de informações que suporta conclusões tomadas sobre a qualidade de dados, isto em um determinado intervalo temporal associado a um arquivo. A proposta de utilização desses documentos é baseada nas práticas atuais do ARM-DoE que faz uso de relatórios de qualidade para registrar classificações obtidas a partir da análise da qualidade das informações medidas por cada instrumento.

O a disponibilização desses documentos ao usuário final facilita seu acesso aos de-

talhes da análise que levou à avaliação positiva ou negativa sobre a qualidade dos dados que está utilizando. É necessário existir um mecanismo que garanta a coleta, armazenamento e distribuição desses documentos, assim neste trabalho será proposta uma ferramenta que realize essas funções. Contudo, devido à limitações técnicas não será possível implementar esse dispositivo, portanto para elaborar a proposta foi adotada como metodologia realizar apenas a descrição da ferramenta em termos de requisitos funcionais, de sistema, interface a banco de dados.

2.5.4 Preparação de dados

A preparação de dados constitui uma ou mais etapas desafiadoras porém imprescindíveis ao processamento de dados e tomada de conclusões. Ela é responsável por administrar tanto os diferentes formatos de dados específicos de cada instrumento, quanto os erros instrumentais que podem se apresentar ocasionalmente em linhas de dados. Isto garante que os dados que chegam efetivamente ao processamento e análise estão de acordo com as diretrizes de formato e qualidade.

A garantia dos aspectos supracitados é fundamental para fortalecer o reaproveitamento de código nos *scripts* de tratamento e avaliação, além de potencializar a otimização da lógica utilizada.

O instrumento MRR, objeto de estudo deste trabalho, exporta os dados capturados em um formato complexo. Ainda, nesses arquivos há a possibilidade de existirem problemas de dados faltantes associados a falhas no sistema de medida. Assim, a metodologia para implementação da lógica para conversão do formato de dados do instrumento em dados netCDF levou em consideração possíveis erros ainda não detectáveis, a partir da suposição do que é esperado de um arquivo sem problemas. Logo, as lógicas que compõem o *script* completo de preparação de dados se deu em uma instância de protótipo para posteriormente ser incluída em um arquivo final para implantação.

2.5.4.1 Prototipação e documentação

A fase de prototipação das lógicas de conversão do formato de dados do MRR para o netCDF se deu por meio de documentos do tipo notebook editados por meio da plataforma Jupyter. Esse documentos, são compostos por células de código em Python e permitem a execução delas de forma não linear e seletiva, assim é possível acelerar o processo de desenvolvimento, facilitando a identificação e correção de bugs entre outros problemas. Ademais, além da conversão do formato de dados, os

códigos de geração de figuras para acompanhamento rápido da integridade dos dados também foram prototipados utilizando-se documentos notebook.

Os documentos notebook além de se apresentarem como uma útil ferramenta de prototipação, também constituem um formato amigável de apresentação da lógica de código para usuários menos experientes em Python. Neles é possível a inclusão de células de textos *Markdown* e figuras explicativas em conjunto das células de código, assim a leitura e interpretação do código pelo usuário pode ser guiado e suportado por materiais auxiliares incluídos diretamente no arquivo.

Tendo em vista essas características, a estratégia de documentação dos códigos deste trabalho previu a elaboração de notebooks explicativos, utilizando-se a plataforma Jupyter para tal, de forma à apresentar a lógica implementada no código de preparação de dados. Vale destacar que o objetivo desta estratégia é facilitar evoluções posteriores das ferramentas e também incentivar o uso dessa metodologia e linguagem por mais pesquisadores.

2.5.4.2 Implantação

Para a elaboração da versão final dos *scripts* de preparação de dados e geração de figuras de análise rápida foi considerado como principal requisito de funcionalidade que permitissem o uso de *scripts shell*, de forma a potencializar a automatização da preparação de dados e permitir que a execução ocorra periodicamente.

Assim, para atender à essa funcionalidade a metodologia de implantação das lógicas na versão final do *script* em Python previu que a elaboração de programa que sejam executados pelo CLI (*command line interface*) que permitisse a inclusão de *flags* para execução condicional. Assim evitá-se-a necessidade de interação do usuário com o programa depois que a execução é iniciada.

Por fim, optou-se por utilizar o paradigma de programação funcional para o desenvolvimento das lógicas.

2.5.5 Controle de qualidade dos dados

O controle de qualidade dos dados visa garantir que aqueles distribuídos aos usuários estejam de acordo com as diretrizes de qualidade especificadas. A metodologia de controle de qualidade dos dados adotada neste trabalho considera que o mesmo se insere em diferentes etapas do processamento dos dados.

Primeiramente, durante a preparação de dados, há a identificação de falhas e aplicação de filtros para erros instrumentais. Em um segundo momento, já durante o processamento efetivo dos dados, há a inserção de *flags* de controle de qualidade de dados que classificam as linhas de medida com relação a testes de qualidade aplicados sobre elas. Esses testes podem ser simples, que consideram métricas intrínsecas ao arquivo estudado, ou de maior complexidade, que utilizam validação cruzada com instrumentos similares ou que medem diretamente as variáveis estimadas, como é o caso do disdrômetros e pluviômetros, para identificar inconsistências nas medidas.

2.5.5.1 Dados faltantes

No contexto do instrumento MRR, dados faltantes podem ter duas origens:

a) Erros instrumentais que afetam uma ou mais variáveis, mas não impacta a geração de arquivos com as medidas capturadas. Assim, é gerado um arquivo diário, porém em alguns intervalos de tempo nem todas as variáveis possuem valores capturados, por isso caracterizam-se como dados faltantes que precisam ser reprocessados.

b) Arquivo bruto limitado temporalmente. Nesse cenário o arquivo diário gerado pelo instrumento não cobre todos os períodos do dia, assim as linhas de medida que não estão incluídas no arquivo exportado também se qualifica como dado faltante (também conhecido por *missing data*).

Uma vez que existem dados faltantes é necessário definir uma diretriz para tratamento destes nos scripts de processamento. A metodologia adotada neste estudo prevê a atribuição da constante -999.99 às variáveis faltantes em linhas de dados afetadas.

2.5.5.2 Validação Cruzada

A validação cruzada é um passo importante na análise de qualidade de dados da pesquisa atmosférica. Dessa forma, ela permite identificar efeitos secundários que podem implicar em erros nas medidas do equipamento estudado.

No contexto deste trabalho, no qual o instrumento objeto de estudos é o MRR temos que os instrumentos pluviômetro e disdrômetro são adequados para realizar a validação cruzada e assim tomar conclusões com relação ao desempenho do MRR. Isso é possível pois o MRR é um equipamento de estimativa de precipitação, uma

vez que as medidas registradas em suas variáveis são obtidas de forma indireta através de relações matemáticas para a grandeza efetivamente capturada pelo sensor (potência espectral do sinal espalhado refletido) e o disdrômetro e pluviômetro são equipamentos que capturam medidas diretamente da precipitação. Assim, observar as estatísticas comparativas entre os dados registrados nesses equipamentos em um mesmo intervalo de tempo permite avaliar a qualidade das medidas do MRR.

Para explorar a validação cruzada, foi definido com diretriz geral a elaboração de documentos Python do tipo *notebook* uma vez que permitem a execução interativa do código e facilita ajustes na lógica de cálculo caso necessário, contribuindo para a flexibilidade no processo de validação. Ainda a metodologia prevê que os *notebooks* elaborados incluam o cálculo de estatísticas de comparação entre os dados de interesse, de forma a subsidiar a análise cruzada entre instrumentos.

2.5.5.3 Estatísticas para validação cruzada

Segundo Wilks (2011) e os estudos realizados durante o SOS-CHUVA (CALHEIROS, 2018) alguns índices estatísticos podem ser aplicados para avaliar a comparação de dados que medem variáveis similares, de modo a realizar a validação cruzada de interesse para este projeto (i.e., estatísticas adequadas para comparar estimativas e dados de referência). A metodologia definida para implementar o cálculo de cada uma delas consistiu em desenvolver uma função que recebe dois *dataframes* unidimensionais da biblioteca Pandas do Python, um deles contendo a variável de interesse do MRR e outro a variável de referência (i.e. disdrômetro e/ou pluviômetro), e retorna um valor correspondente à estatística calculada para os dados. As estatísticas escolhidas são apresentadas na Tabela 2.4 (WILKS, 2011).

Tabela 2.4 - Estatísticas escolhidas para a validação cruzada

Identificador	**Nome**	**Estatística**	**Cálculo**	**Intervalo válido**	**Valor ideal**
MAE	*Mean Absolute Error*	Erro absoluto médio	$MAE = \frac{1}{N} \bullet \sum_{i=1}^{N} \lvert F_i - O_i \rvert$	$0 \leq MAE \leq \infty$	0
RMSE	*Root Mean Square Error*	Raiz do erro quadrático médio	$RMSE = \sqrt{\frac{1}{N} \bullet \sum_{i=1}^{N} (F_i - O_i)^2}$	$0 \leq RMSE \leq \infty$	0
CORR	*Correlation Coefficient*	Coeficiente de Correlação	$CORR = \frac{\sum (F-\bar{F}) \bullet (O-\bar{O})}{\sqrt{\sum (F-\bar{F})^2} \bullet \sqrt{\sum (O-\bar{O})^2}}$	$-1 \leq CORR \leq 1$	1
BIAS	Bias	Desvio sistemático	$BIAS = \frac{\frac{1}{N} \bullet \sum_{i=1}^{N} F_i}{\frac{1}{N} \bullet \sum_{i=1}^{N} O_i}$	$-\infty \leq BIAS \leq \infty$	1

Fonte: Adaptada de WMO (2017).

Por fim, é importante mencionar que realizar a validação cruzada utilizando-se os dados do MRR, preparados pela ferramenta proposta neste trabalho, também se mostra conveniente para avaliar qualitativamente o desempenho da ferramenta de preparação de dados. Isso ocorre uma vez que o processo de validação de dados exigirá a exploração dos arquivos netCDF e, portanto, permitirá examinar a experiência de usuário com relação ao aspecto de acesso aos dados.

2.6 Resultados e análises

Este trabalho tem como objetivo geral propor um *toolkit* específico para dados do instrumento MRR. A elaboração dessas ferramentas perpassa pelo desenvolvimento de dispositivos para cada um dos 3 objetivos designados no tópico 1.1. A listagem abaixo resume os resultados obtidos para cada objetivo.

A. Ferramenta de coleta, armazenamento e apresentação de Relatórios de Qualidade dos Dados.

   O sistema de coleta e armazenamento de relatórios de qualidade dos dados foi especificado em termos de requisitos de interface, sistema e banco de dados.

B. Ferramenta para aplicação de *flags* de qualidade e condensação de informações de verificação de qualidade por meio de palavras binárias.

   Documentos Python do tipo notebook foram elaborados para subsidiar o acesso de novos pesquisadores às ferramentas e formato netCDF. Nesses documento há seções com código pré-pronto que permite a inclusão de *flags* de qualidade por meio da técnica de *bit-packing*.

C. Ferramenta de geração de arquivo NetCDF

   Foram elaborados dois *scripts* principais para geração de netCDF a partir de dados brutos do sensor, além da elaboração de figuras interativas para consulta rápida. Ainda, foi desenvolvido uma biblioteca de funções importantes ao código e que pode ser reutilizada em evoluções das ferramentas ou novas ferramentas incorporadas ao *toolkit*.

2.6.1 Códigos

Para desenvolver as ferramentas necessárias aos objetivos do projeto foram obtidos 4 principais produtos em código: 2 *scripts* em Python e 2 documentos Python do tipo notebook. A Tabela 2.5 apresenta características gerais dos códigos.

Tabela 2.5 - Características gerais dos códigos produzidos

Documento	Extensão	Tipo	Objetivo
MRR_gen_netCDF	.py	Script python puro	Gerar arquivos netCDF a partir de arquivos brutos do equipamento
MRR_gen_figures	.py	Script python puro	Gerar figuras estáticas e interativas a partir dos netCDFs gerados
utils	.py	Script python puro	Contém as principais funções utilização na geraÇão de netCDF
MRR_explore_netCDF	.ipynb	Notebook	Subsidiar a exploração dos arquivos netCDF gerados
MRR_validations	.ipynb	Notebook	Subsidiar a validação cruzada do MRR com outros equipamentos

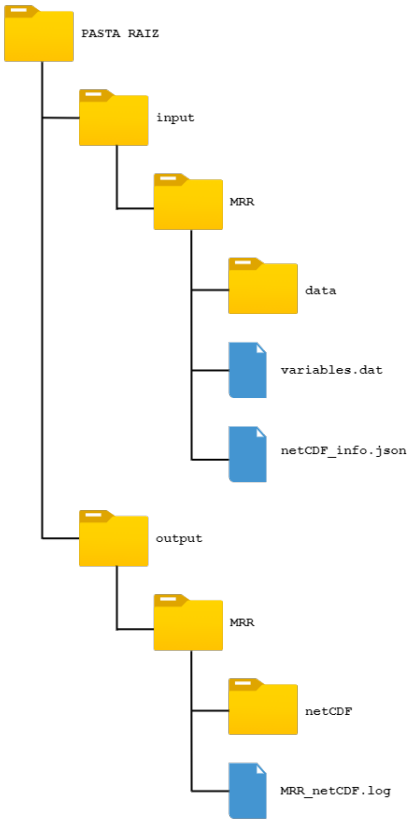
Fonte: Elaborada pelo autor.

2.6.1.1 *Scripts*

Seguindo a metodologia de prototipação e implantação especificadas nos tópicos 2.5.4.1 e 2.5.4.2 foi possível desenvolver um *script* Python principal para a conversão do formato de dados do MRR para netCDF e outro *script* secundário que a partir dos netCDFs exportados gera figuras de análise rápida.

O *script* de conversão de dados foi elaborado de tal forma que seu funcionamento depende do *input* de arquivos em uma estrutura de pastas apresentada no diagrama da Figura 2.4. Nessa estrutura espera-se como entrada dois arquivos complementares e o arquivo bruto do MRR cujas variáveis devem ser extraídas e encapsuladas em formato netCDF. Os arquivos complementares especificam os metadados a serem incluídos no arquivo bem como a lista de variáveis a serem retiradas do arquivo bruto. Ademais, a saída para os produtos processados pelo programa também são exportados em uma estrutura de arquivos também explícita na Figura 2.4, nela vale destacar o *log* do programa que registra a data e hora do momento no qual cada netCDF foi gerado.

Figura 2.4 - Estrutura de arquivos de entrada e saída para o *script* de geração de arquivos netCDF



Fonte: Elaborada pelo autor.

A execução do *script* principal é feita pelo terminal utilizando-se Python 3. No comando de execução é necessário inserir as *flags* de execução que permitem especificar qual tipo de arquivo será processado na sessão, além de possibilitar a indicação de quais arquivos devem ser incluídos no processamento. Alguns exemplos de comandos de execução do *script* no terminal é apresentado abaixo.

```
$ .python3 MRR_gen_netCDF.py --help
$ .python3 MRR_gen_netCDF.py --ave
$ .python3 MRR_gen_netCDF.py -p
```

As *flags* de execução e seus efeitos são listadas a seguir.

- -h ou –help

  Imprime no terminal uma mensagem de ajuda explicando as *flags* de execução seus efeitos

- -a ou –ave

  Extrai as variáveis dos arquivos .ave incluídos na pasta de entrada de dados e encapsula os dados em um netCDF.

- -p ou –pro

  Extrai as variáveis dos arquivos .pro incluídos na pasta de entrada de dados e encapsula os dados em um netCDF.

- -r ou –raw

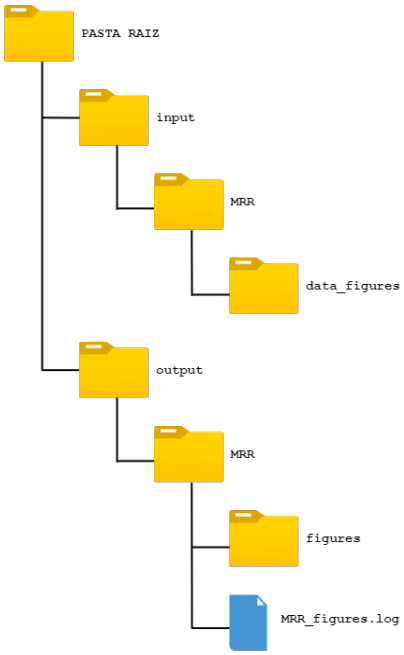
  Extrai as variáveis dos arquivos .raw incluídos na pasta de entrada de dados e encapsula os dados em um netCDF.

Uma vez executado, o *script* produz um netCDF para cada arquivo inserido na pasta de *input* de dados. Nesse contexto vale destacar que o nome do arquivo gerado é elaborado no código seguindo a metodologia especificada no tópico 2.5.2.1. Ainda, em conformidade com o modelo de dados do formato, as variáveis extraídas do arquivo bruto são incluídas no netCDF cada qual como uma matriz multidimensional, seguindo os nomes e dimensões descritos na Tabela [ref tabela de variáveis e dimensões]. Vale comentar que, naturalmente, as variáveis incluídas no netCDF dependem das variáveis presentes no arquivo bruto, portanto, enquanto um netCDF produto, da conversão de um arquivo *average* ou *processed*, contém todas as variáveis da tabela, um netCDF resultado do processamento de um arquivo *raw* contém apenas as variáveis Fnn com nn de 01 a 64 (referentes ao tamanho das gotas de chuva).

Dado que o paradigma de programação funcional foi utilizado no desenvolvimento, todos os processos lógicos foram incluídos em funções. Dessa forma, visando a melhor organização e leitura do código optou-se por elaborar uma biblioteca denominada *utils*, que reúne as principais funções utilizadas na conversão de dados.

O *script* que gera as figuras para análise rápida de integridade dos dados também tem seu funcionamento condicionado ao *input* de arquivos em um estrutura de pastas especificada na Figura 2.6. Nela, é necessário inserir os netCDFs a serem explorados na pasta de entrada de dados para figuras e o produto da execução do *script* é apresentado na pasta de saída de figuras. Todas essas estruturas visam a rápida instalação e execução dos algoritmos sem maiores esforços e de modo amigável.

Figura 2.5 - Estrutura de arquivos de entrada e saída para o *script* de geração de figuras



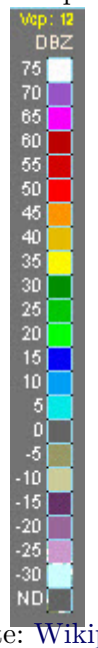
Fonte: Elaborada pelo autor.

Para cada netCDF explorado pelo programa são geradas 4 figuras, as quais são descritas na lista abaixo e apresentadas, respectivamente, nas Figuras 2.7, 2.8, 2.9 e 2.10. É importante destacar que para cada figura são produzidos dois formatos, um formato estático (png) e outro formato interativo (html) que permite uma exploração mais minuciosa dos dados e facilita análise de pesquisadores .

- Gráfico do tipo *heatmap* para a variável Zdb_att_mrr
- Gráfico do tipo *scatter* para a variável rain_rate_mrr
- Gráfico do tipo *heatmap* para a variável liq_water_mrr
- Gráfico do tipo *heatmap* para a variável fall_vel_mrr

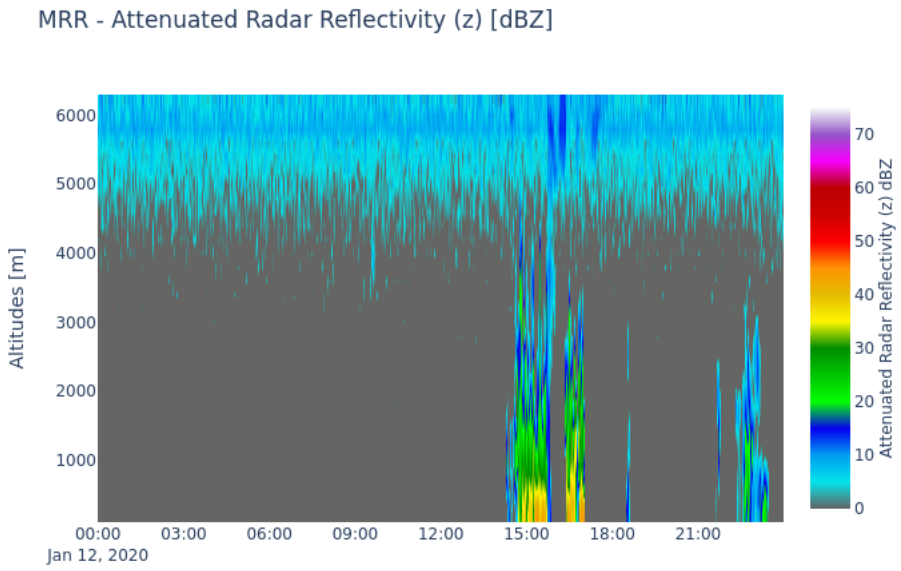
Vale mencionar que a escala de cores utilizada nos gráficos foi baseada nas cores para valores positivos da escala dBZ definida pela NOAA (*National Oceanic and Atmospheric Administration*, orgão americano ligado a pesquisa e operação oceanográfica e meteorológica) e disponível para consulta na referência Wikipedia (2014). A Figura 2.6 apresenta a escala que foi utilizada como modelo.

Figura 2.6 - Escala para radares atmosféricos



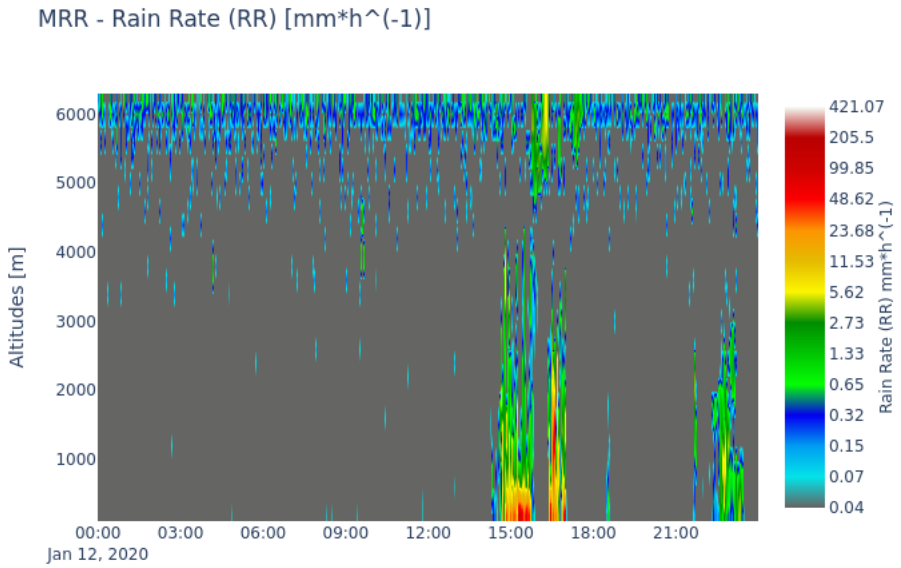
Fonte: Wikipedia (2014)

Figura 2.7 - Gráfico elaborado para dados do MRR com relação à variável z



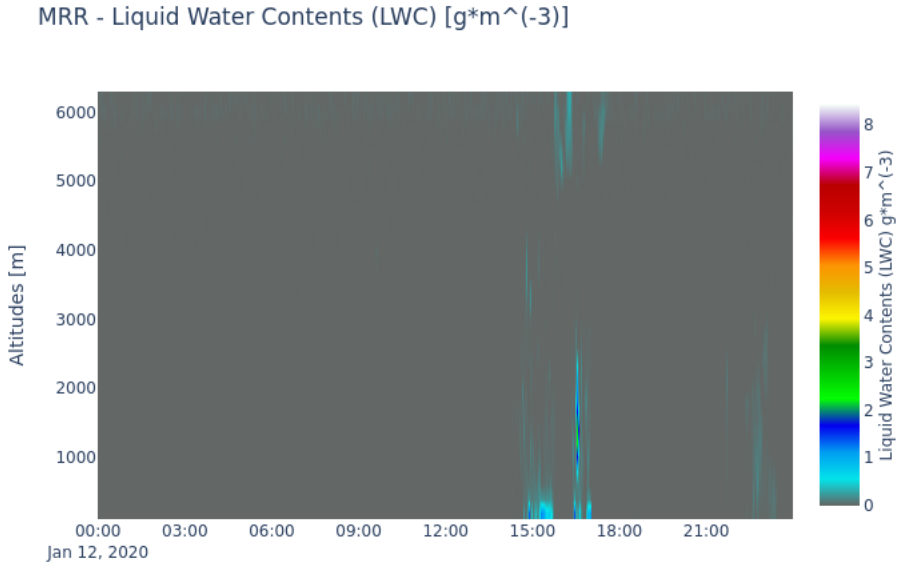
Fonte: Elaborada pelo autor.

Figura 2.8 - Gráfico elaborado para dados do MRR com relação à variável RR



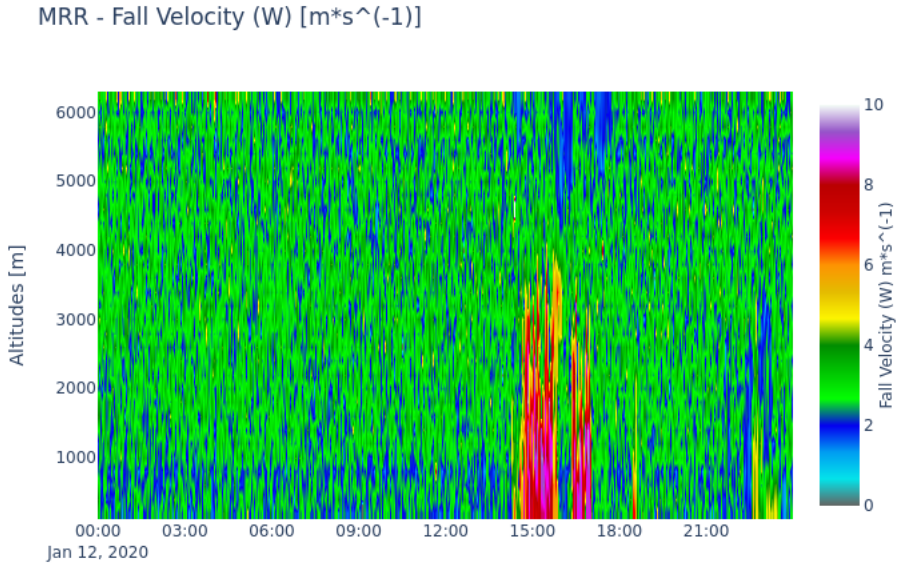
Fonte: Elaborada pelo autor.

Figura 2.9 - Gráfico elaborado para dados do MRR com relação à variável LWC



Fonte: Elaborada pelo autor.

Figura 2.10 - Gráfico elaborado para dados do MRR com relação à variável W



Fonte: Elaborada pelo autor.

A execução do *script* pode ocorrer de três formas distintas, ativadas por *flags* de execução. As *flags* e o funcionamento para cada modo é especificado a seguir.

- `-a ou --auto`

  Executa o *script* (i.e., gera as figuras) para todos os arquivos incluídos na pasta de entrada de dados para figuras, porém acessa os *logs* de geração de arquivo e de geração de figuras e garante que serão geradas figuras apenas para aqueles arquivos que ainda não foram explorados.

- `-f <nome do arquivo netCDF> ou --file <nome do arquivo netCDF>`

  Executa o *script* (i.e., gera as figuras) apenas para o arquivo especificado no comando de execução.

- `-l <nome do arquivo txt> ou --list <nome do arquivo txt>`

  Executa o *script* (i.e., gera as figuras) para os arquivos listados no arquivo txt especificado no comando.

Vale comentar que foram utilizadas diversas bibliotecas de Python para elaborar a lógica dos *scripts* como NumPy e Pandas. O anexo A apresenta as bibliotecas

necessárias para a execução dos scripts e suas respectivas versões conforme gerado pela ferramenta *pip* de gerenciamento de pacotes Python. Por fim, é importante mencionar que os códigos continuam em desenvolvimento mesmo após a publicação deste relatório, visando a correção de bugs identificados na utilização em situações reais das ferramentas e também contribuições futuras de usuários.

2.6.1.2 Métricas de validação de *script*

Para sustentar as ferramentas propostas neste trabalho é fundamental explorar comparativamente o desempenho delas em relação aos métodos utilizados atualmente. Primeiramente em relação às ferramentes de conversão de formato de dados, é importante:

i Verificar quantitativamente a equivalência dos produtos gerados pelas ferramentas propostas em comparação com as ferramentas em uso

ii Avaliar qualitativamente a facilidade de distribuição e uso das ferramentas propostas em comparação com as ferramentas em uso

Para explorar o primeiro ponto, foi importante considerar a diferença de formato entre os produtos gerados pelas ferramentas atuais e aquela proposta neste trabalho. Enquanto o proposto congrega todas as matrizes de dados em um arquivo único, os produtos da preparação de dados até então separavam as matrizes de cada variável em arquivos ASCII diferentes. Considerando essa particularidade, estabeleceu-se como metodologia para ler os dados de uma variável conveniente, tanto dos arquivos processados pelos programas vigentes, quanto pelo proposto em um *dataframe* utilizando-se a biblioteca Pandas. Nesse cenário, ainda é importante estabelecer o tipo e precisão numérica a ser utilizada na leitura dos dados, a qual, considerando a precisão dos dados registrados nos arquivos ASCII e netCDF, foi definida como ponto flutuante com 32 bits.

Dessa forma, embora com origens diferentes, os dados nos *dataframes* podem ser comparados diretamente e portanto é possível estabelecer uma estratégia para obter estatísticas de comparação direta. A estratégia identificou duas métricas correlacionadas e também um método já implementado na biblioteca pandas que servem ao propósito da comparação desejada.

As duas métricas correspondem ao número de linhas e colunas de cada *dataframe* contendo os dados lidos. De modo a indicar que os dados são equivalentes elas devem

Tabela 2.6 - Métricas para avaliação da equivalência entre os produtos processados pelas ferramentas em usso e aquelas propostas.

Dataframe	Linhas	Colunas	Equals
dados processados pela ferramenta vigente	43191	31	True
dados processados pela ferramenta proposta	43191	31	

Fonte: Elaborada pelo autor.

ser idênticas para ambos os *dataframes*. Já o método é o "pandas.DataFrame.equals" que executa a comparação entre cada célula de dois *dataframes* e retorna como saída *True* caso as tabelas sejam idênticas e *False* caso contrário.

Assim, as métricas supracitadas foram exploradas e o método identificado foi executado para os dados preparados pelas ferramentas correntes em comparação com aqueles preparados pela ferramenta proposta. Para isso foram utilizados os dados da variável *Rain Rate* (taxa de chuva, mm/h) do equipamento MRR oriundos da companha SOS-Chuva cujos dados brutos e processados estavam à disposição e já validados. Sobre todos os dados disponíveis dessa variável para a campanha selecionou-se o período de 1 de junho de 2017 a 30 de junho de 2017 para a análise. A Tabela 2.6 apresenta os resultados obtidos. As métricas obtidas demonstram a equivalência dos produtos da ferramenta proposta em comparação àquelas em uso.

Com relação à análise qualitativa da facilidade de uso e distribuição das ferramentas a avaliação considerou dois pontos principais:

- Praticidade de execução
- Acessibilidade dos requisitos de execução

Nesse contexto foi elaborada a Tabela 2.7 que resume as característica das ferramentas em relação aos aspectos supracitados.

A comparação entre as características das ferramentas evidencia que a qual é proposta por este trabalho é vantajosa em relação a ferramenta atual nas duas métricas de comparação. Contudo, cabe ressaltar que a proposta apresenta em IDL não utilizou todos os recursos sofisticados que a linguagem disponibiliza, o que aumentaria sua praticidade. Contudo, é no quesito *open-source* que o Python se destaca ainda mais, de forma potencializar seu uso pela comunidade científico-atmosférica do INPE e USP sem custos as instituições e um maior número de usuários da linguagem.

Tabela 2.7 - Características qualitativas das ferramentas de processamento de dados

Ferramenta	Praticidade	Acessibilidade
Ferramenta vigente (IDL)	Necessita ser ajustada em código para cada campanha de coleta de dados a ser processada. Execução na IDE específica.	A linguagem utilizada requer um software de licença paga para ser executada.
Ferramenta proposta (Python)	Ferramenta parametrizada em função de arquivos de metadados e estrutura de pastas. Execução no terminal com flags condicionais.	A linguagem Python é open-source e distribuída gratuitamente.

Fonte: Elaborada pelo autor.

2.6.1.3 Notebook de Exploração de Dados

Visando potencializar a utilização não apenas dos *scripts* propostos, mas como também incentivar o uso da tecnologia Python na gestão de dados em pesquisa atmosférica pela comunidade do INPE e USP, e seus colaboradores, foi elaborado um documento do tipo notebook interativo e explicativo que aborda em mais detalhes as lógicas implementadas nos códigos Python.

A plataforma Jupyter foi utilizada no desenvolvimento como descrito na metodologia abordada no tópico 2.5.4.1.

No cenário de exploração dos dados, algumas ferramentas já estão funcionais e prontas para testes com os dados do MRR. Como por exemplo, a análise de *flags* associados a dados faltantes são facilmente implementadas, uma vez que são criados dados diários que levam em consideração a resolução temporal das medidas. Na Tabela 2.8 é possível verificar o quanto de dados válidos existiram durante o experimento GoAmazon para o período de medida em 2014.

Tabela 2.8 - Dados válidos do MRR para o experimentos GoAmazon.

Mês	**Jan.**	**Fev.**	**Mar.**	**Ago.**	**Set.**	**Out.**
MRR UEA	100%	99.92%	82.69%	16.79%	100%	36.17%

Fonte: Elaborada pelo autor.

2.6.2 Estatísticas de avaliação de dados

Como descrito no item 2.5.5.2, a avaliação minuciosa dos dados do MRR perpassa pela comparação deles com dados capturados por outros dois sensores no mesmo intervalo de tempo (i.e. os disdrômetro e pluviômetro). As estatísticas de compara-

ção explicitadas na metodologia foram calculadas para cada dupla de dados entre MRR e os outros sensores, a Tabela 2.9 apresenta os resultados calculados para o experimento SOS-CHUVA. Essas avaliações para um experimento fora da região amazônica se deu devido ao fato de estudos já terem sido realizados pelo orientador deste projeto (CALHEIROS, 2018) e serviram de referência para verificar a acurácia dos algoritmos implementados. Contudo, cabe ressaltar que o processo de criação levou em consideração dados gerados na região Amazônica como exemplo e para testes preliminares. Os resultados mostraram que estes algoritmos em Python estão de acordo com os estudos preliminares durante este experimento. O valores da estatística descritiva mostram que o instrumento se mostrou apto a realizar as estimativas de chuva com alto grau acurácia, onde apresentou alta correlação e um erro médio baixo em comparação aos instrumentos de referência.

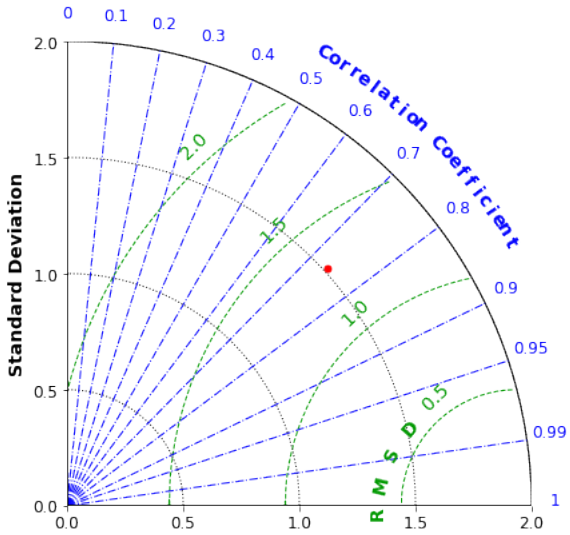
Tabela 2.9 - Resultados calculados para as estatísticas de validação cruzada aplicadas aos dados da campanha SOS-Chuva

Estatística	**MRR x JWD**	**MRR x Pluviômetro**
MAE	0.306	0.601
RMSE	0.727	1.349
CORR	0.874	0.739
BIAS	-0.191	-0.294
SDEV	0.702	1.311

Fonte: Elaborada pelo autor.

Além da apresentação das estatísticas no formato numérico, uma possibilidade gráfica de explicitar o conjunto de estatísticas SDEV, CORR e RSME é o diagrama de Taylor. Este diagrama possui 3 eixos que representam cada uma das grandezas supracitadas e nele é possível apresentar diversos pontos permitindo observar estatísticas que comparam diversos instrumentos em um mesmo gráfico, facilitando a análise. A Figura 2.11 apresenta um exemplo de diagrama de Taylor que inclui apenas um ponto (ponto vermelho) correspondente às estatísticas de comparação entre MRR e pluviômetro. Este tipo de gráfico irá permitir ao mentor dos dados, assim como seus usuários, inserir mais informações referentes a outros instrumentos e comparações em diferentes períodos de medidas, validando assim se os dados continuam aptos a serem disponibilizados ao público.

Figura 2.11 - Diagrama de Taylor com um ponto referente à resultados para estatísticas comparativas entre MRR e Pluviômetro



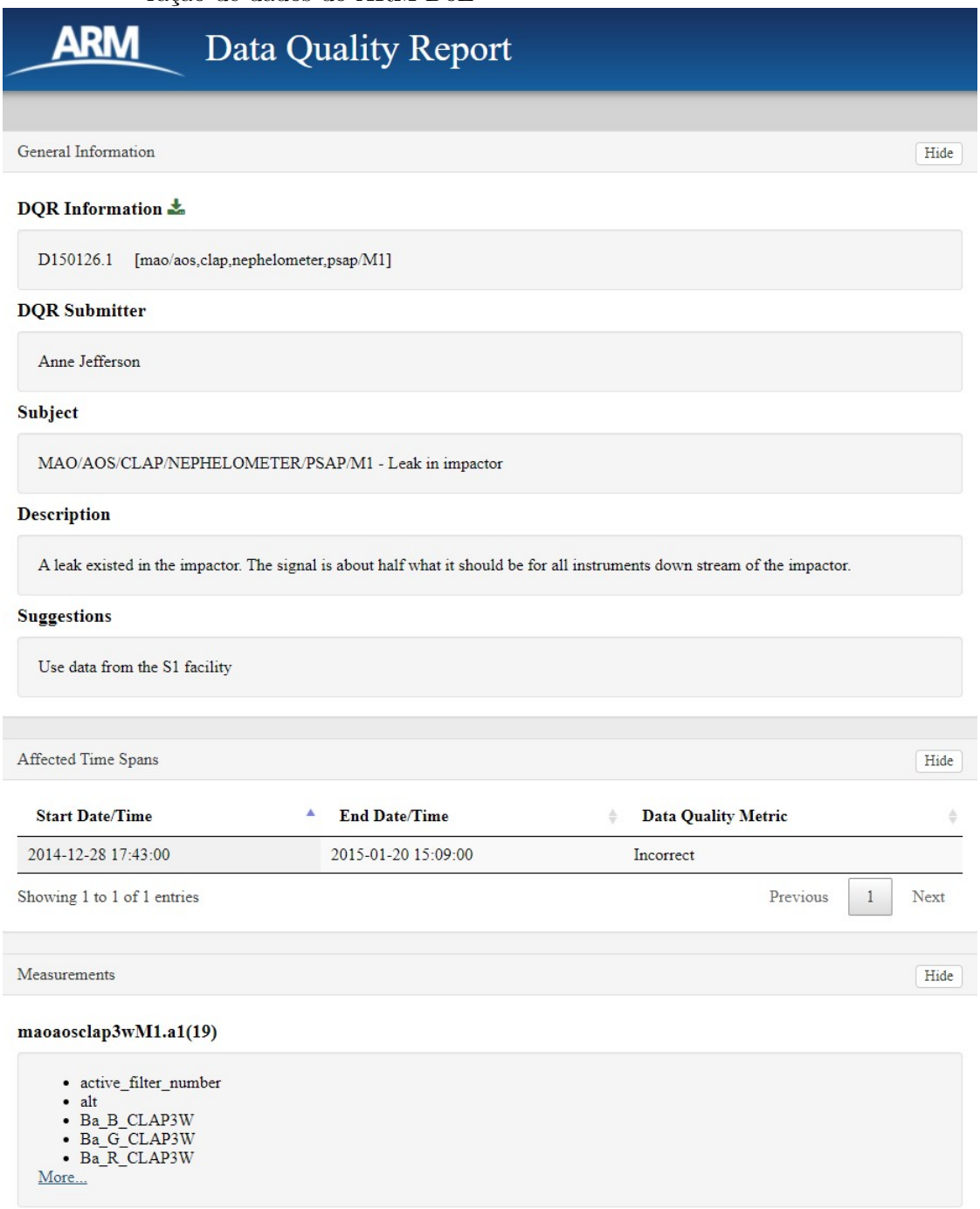
Fonte: Elaborada pelo autor.

2.6.3 Relatórios de Qualidade de Dados

Como especificado no tópico 2.5.3 foi elaborada uma descrição de um sistema de submissão, armazenamento e distribuição de Relatórios de Qualidade de Dados (DQR). Cada DQR registra conclusões sobre algum teste de qualidade realizado sobre um arquivo de dados e portanto se refere a apenas um arquivo. Contudo, diferentes variáveis dentro destes arquivos podem ser registradas no documento sob distintas classificações de qualidade. Considerando essas características a descrição foi construída em termos de requisitos funcionais, sistema, interface e banco de dados.

Ademais, para o desenvolvimento da descrição supracitada foi utilizado como referência o sistema de DQRs utilizado no ARM-DoE. Assim, utilizando-se a ferramenta de exploração de dados da instituição foi localizado um relatório de qualidade de dados disponibilizado para dados da campanha GOAmazon, já submetidos na plataforma. O acesso a este documento permitiu identificar campos de informações importantes que deveriam ser considerados na descrição do sistema a ser proposto. A Figura 2.12 apresenta uma captura de tela do documento acessado. A análise do relatório disponibilizado permite identificar que o primeiro campo exibido ("DQR Information") representa o nome do documento, gerado automaticamente pelo sistema, e portanto não corresponde a uma entrada do usuário a ser incluída no formulário de submissão a ser proposto.

Figura 2.12 - Captura de tela da interface de consulta de DQRs da plataforma de exploração de dados do ARM-DoE



Fonte: Plataforma ARM para exploração de dados.

Dessa forma, utilizando como base o sistema do ARM, atentando-se às características de um relatório de qualidade de dados, e visando potencializar o acesso prático

e rápido de pesquisadores à metadados e análises de qualidade, foi possível elaborar tanto a estrutura do formulário de submissão de um relatório de qualidade de dados, como apresentado na tabela 2.6.3, quanto os pontos que descrevem o sistema de gerenciamento de relatórios de qualidade de dados.

Tabela 2.10 - Estrutura de campos a serem incluídos no formulário de submissão de DQRs

Campo	**Tipo**	**Descrição**
Responsável	Texto	Nome do responsável pelo relatório.
Email responsável	Texto	Email do responsável pelo relatório
Arquivo analisado	Texto	Nome do arquivo analisado
Assunto	Texto	Sumarização do problema de qualidade encontrado.
Descrição	Texto	Explicação detalhada da problemática encontrada.
Metodologia	Texto	Explicação detalhada da análise dos dados conduzida que identificou os problemas apresentados.
Sugestão	Texto	Sugestões para corrigir ou contornar o problema identificado.
Períodos afetados	Datas	Conjunto de períodos nos quais o problema de qualidade ocorre.
Classificação de qualidade	Seleção	Classificação da qualidade para os dados no período afetado pelo problema
Variáveis afetadas	Texto	Nome das variáveis afetadas pelo problema.

Fonte: Elaborada pelo autor.

i. **Requisitos Funcionais**

- **Receber as informações do relatório por meio de um formulário de submissão. Permitir a submissão de versão em português e inglês do documento.**
  Seguir a estrutura apresentada na tabela 2.6.3. Basear interface no ARM-DoE.

- **Armazenar os relatórios.**
  Banco de dados compatível com o tipo de informação a ser armazenada. Definir requisitos de banco de dados.

- **Permitir a consulta de relatórios.**
  Ferramenta de busca deve ser implementada. Ademais, é fundamental existir uma forma de compartilhar o acesso direto à um documento específico.

- **Permitir a alteração de informações em um relatório.**
  Garante a praticidade na correção de eventuais erros na submissão de informações.

ii. **Requisitos de Sistema**

- **Sistema escalável.**
  Utilizar tecnologias de hospedagem e processamento que permitam a escalabilidade horizontal do sistema.

- **Plataforma web.**
  Desenvolver as ferramentas para plataforma web, tendo em vista potencializar o acesso de qualquer usuário ao sistema.

- **Ferramenta de submissão.**
  Implementar uma ferramenta de submissão de relatórios. Constitui um formulário com os campos de entrada para o responsável.

- **Ferramenta de consulta.**
  Implementar ferramenta de consulta de relatórios. Permite gerar um link de compartilhamento direto do acesso a um documento, possibilidade de desenvolver API web, que gerencie a solicitação de dados.

iii. **Requisitos de Interface**

- **Exibir nome dos campos em português ou inglês**
  Necessário para implementar a submissão das versões em português e inglês do documento.

- **Interface responsiva para diferentes tamanhos de tela.**
  Altera o tamanho dos campos exibidos conforme o tamanho da tela do dispositivo em que ocorre o acesso. Garante melhor experiência de usuário. Possibilidade de utilizar um *framework* de CSS como Bootstrap ou *framework* de JavaScript como React.

- **Otimização da interface para acesso em dispositivos *mobile*.**
  Potencializa a utilização da ferramenta por mais pesquisadores.

iv. **Requisitos de Banco de Dados**

- **Esquema de dados flexível**
  Importante para facilitar eventuais evoluções do sistema (adição de novos campos de informações). Indicado uso de banco de dados não relacional [inserir footnote].

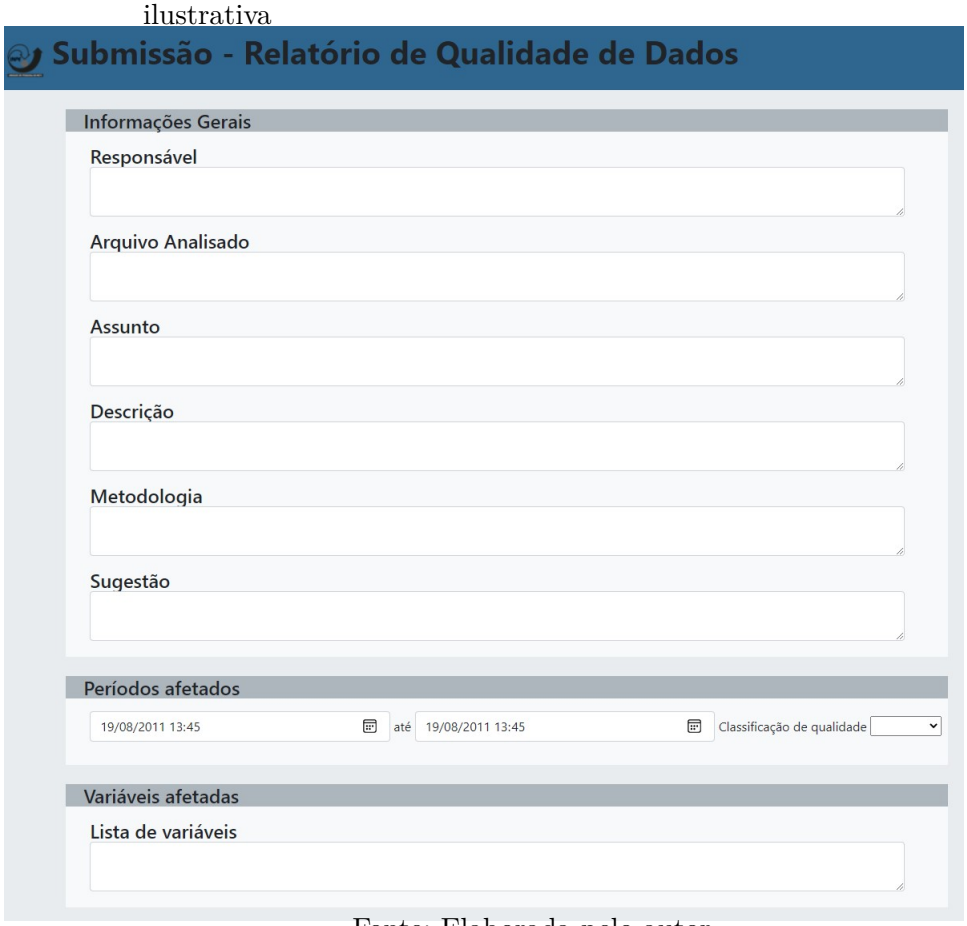
- **Escalável horizontalmente**
  Utilizar um sistema de gerenciamento de banco de dados (SGBD) que permita a escalabilidade horizontal do banco de dados. Facilita a expansão do sistema.

- **Compatibilidade com implementação de consultas por API Restfull**
  Possibilita que a consulta de um relatório seja acessada pelo acesso a um link da API com as flags de pesquisa desejadas incluídas.

É importante mencionar que uma vez que não foi possível implementar o sistema proposto, também se mostrou inviável definir os requisitos de segurança para as ferramentas. Isso ocorreu pois, o estudo e análise do cenário de segurança a ser instalado no ambiente da plataforma depende da definição efetiva das tecnologias a serem utilizadas no desenvolvimento. Assim, admite-se a elaboração dos requisitos de segurança como uma etapa concomitante à implementação da plataforma e portanto será executada em uma próxima etapa deste trabalho. Uma vez que tanto o INPE como uma de seus parceiros disponibilize tal plataforma. A USP por meio do IFUSP já trabalha neste tipo de plataforma para disponibilizar dados da torre ATTO, atualmente em teste. Espera-se que até o final do corrente ano esta plataforma esteja disponivél para que possamos implementar este sistema de DQRs.

Ainda que a implementação da ferramenta fora determinada como um encaminhamento para uma etapa posterior do projeto, foi explorado de forma inicial uma possibilidade de elaboração de interface de submissão dos relatórios. Assim, foi desenvolvido um exemplo ilustrativo de interface estática com alguns dos campos previstos para o formulário utilizando-se HTML, JavaScript e CSS com o *framework* Bootstrap. A Figura 2.13 exibe uma captura da interface ilustrativa desenvolvida.

Figura 2.13 - Captura de tela da interface de submissão de DQRs elaborada de forma ilustrativa



Fonte: Elaborada pelo autor.

3 CONCLUSÕES

Acerca da ferramenta de geração de netCDF a partir dos arquivos brutos do instrumento MRR foi possível atingir plenamente os objetivos estabelecidos. Diante dos resultados obtidos foi possível concluir que as ferramentas propostas desenvolvidas em Python produzem dados com plena equivalência em comparação com os produtos das ferramentas em utilização atualmente. Ademais, o *script* em Python apresenta vantagens importantes sobre a tecnologia em uso, dado sua distribuição gratuita e *open-source* bem como a possibilidade de elaboração de documentos de execução interativa e guiada do tipo *notebook* que facilitam a documentação das ferramentas e potencializam seu uso por pesquisadores com pouca experiência com a linguagem Python.

Com relação aos documentos *notebooks* desenvolvidos na pataforma Jupyter foi possível implementar esta ferramenta para facilitar a leitura e plotagem dos dados gerados em netCDF por este projeto. Com o link a seguir é possível realizar o download do repositório com um exemplo de dado do MRR, instalado próximo a torre ATTO, na região amazônica e o notebook para ler e visualizar as informações (verificar o documento readme presente no diretório): https://github.com/tpougy/relatorio-IC-2021. Note que o usuário terá mais facilidade em utilizar os dados em suas pesquisas e visualizar as informações, que é de suma importância para definir a qualidade das mesmas.

Os algoritmos de definição de *flags* de controle de qualidade foram contruídos e testados com dados de medidas de alguns experimentos aqui analisados. De modo preliminar, as análises mostraram que estes estão aptos ao uso em dados de maior volume no que diz respeito a problemas sistemáticos, como falta de medidas e erros decorrentes de problemas associados ao equipamento, em específico o MRR. Já as *flags* associadas a medidas não-realísticas a partir da correlação cruzadas ainda precisam ser implementadas de modo eficiente, contudo, este se caracteriza como uma etapa de análise pós-processamento que pode ser realizada posteriormente no ciclo de vida do dado. Neste caso, algoritmos de nível 2 de dados estão em preparação e maiores detalhes são apresentados no capítulo seguinte.

Com relação aos aerossóis vale comentar que no período anterior ao início deste trabalho foi desenvolvida uma ferramenta de conversão de formato de dados específicos de 4 sensores de aerossol atmosférico para um formato padrão ASCII, a ferramenta possui uma interface acessada pelo terminal e permite a alteração de metadados gerais do arquivo. Ela foi elaborada utilizando-se a linguagem Python,

porém não implementa a conversão para o formato netCDF, objetivo o qual foi reservado como um encaminhamento dos trabalhos realizados a ser executado futuramente. Nesse contexto, as bases construídas para lógica de conversão de formato de arquivo, bem como o desenvolvimento em conformidade com os métodos e práticas do ARM serviram como importante referência para aspectos importantes do rápido desenvolvimento das ferramentas propostas neste estudo como já abordado no tópico 2.4.

Cabe ressaltar que as ferramentas aqui desenvolvidas para controle de dados do MRR, também serão implementadas para os dados de aerossol atmosférico medidos em experimentos na região amazônica sob a tutela de pesquisadores responsáveis na USP e em parceria com o INPE assim que a plataforma lá desenvolvida for disponibilizada. Esse encaminhamento constituirá uma evolução da ferramenta supracitada e permitirá a geração de arquivos netCDF em conformidade com as diretrizes internacionais para dados de aerossol dos 4 sensores estudados.

As atividades aqui realizadas foram complexas e exigiram alto empenho e comprometimentos, além de conhecimentos avançados de computação e dos dados aqui analisados. Algumas das atividades propostas ficaram pendentes devido a grande quantidade de aspectos ainda incipientes que demandaram mais tempo do que o esperado, na seção seguinte são apresentados alguns desses pontos que serão endereçados em trabalhos futuros.

3.1 Trabalhos Futuros

Nesta seção são explicitadas as etapas que se referem as ferramentas que ainda precisam ser testadas e implementadas seguindo a trajetória especificada pelos órgãos internacionais, como o ARM-DoE. Estes trabalhos podem ser realizados caso haja a renovação da bolsa por mais um ano.

Com relação a ferramenta de coleta, armazenamento e distribuição de relatórios de qualidade de dados as análises conduzidas permitem identificar que para se atingir as funcionalidades necessárias, deverá ser elaborado um sistema para fazer o gerenciamento dos relatórios. Ademais, embora a implementação não tenha sido conduzida, diversos aspectos relevantes ao desenvolvimento da plataforma foram levantados e servirão de encaminhamento para os trabalhos que seguirão esta proposta.

Nesse contexto, vale destacar que foi possível depreender que o sistema deverá ser desenvolvido para plataforma web, de forma a universalizar o acesso. Além disso,

também se mostrou promissor utilizar o paradigma de banco de dados não relacional para armazenamento dos documentos, facilitando a escalabilidade do sistema e flexibilidade no esquema de dados.

Ademais, no que diz respeito à ferramenta de aplicação de *flags* de qualidade sua implementação também se mostrou inviável nesta etapa do projeto. Essa ferramenta permite classificar as linhas de dados conforme um teste de qualidade é aplicado no conjunto, e, no cenário deste estudo, com enfoque nos dados gerados pelo equipamento MRR, os testes de qualidade citados correspondem à validação cruzada especificada na metodologia e resultados. Contudo, a exploração das estatísticas dessa validação se mostrou um processo mais complexo e longo do que o planejado, de forma a inviabilizar que seus resultados fossem então aproveitados a tempo para implementar a ferramenta de aplicação de *flags* de qualidade.

É importante mencionar, que o desenvolvimento dessa ferramenta não se resume apenas à implementação final, existem etapas complementares anteriores importantes para a elaboração do projeto da ferramente, as quais foram completadas. Nesse contexto, se insere a elaboração do código dos *scripts* e *notebooks* que foram concebidos tendo em vista facilitar a implantação das *flags* e também a própria exploração e análise das estatísticas de validação cruzada. Dessa forma, podemos concluir que mesmo sem a implementação efetuada o objetivo foi parcialmente cumprido restando apenas a ultima etapa de todo o processo de desenvolvimento para ser executada.

REFERÊNCIAS BIBLIOGRÁFICAS

ANJOS G. A. DIAS, A. A. R. R. L. dos. Dados científicos: As práticas de gestão dos pesquisadores brasileiros na ciência da informação. In: ENCONTRO NACIONAL DE PESQUISA EM CIêNCIA DA INFORMAçãO, 18., 2017, Marília, SP, Brasil. Marília, SP, Brasil: ENANCIB, 2017. 5

ATMOSPHERIC RADIATION MEASUREMENT CLIMATE RESEARCH FACILITY. **Embedded quality control fields**. 2015. Disponível em: <https://code.arm.gov/docs/QC-flag-examples/-/wikis/Introduction>. Acesso em: 22 mai. 2020. 3

_____. **ARM Data File Standards**. [S.l.], ago. 2016. 1–14 p. 11

ATMOSPHERICRADIATION MEASUREMENT CLIMATE RESEARCH FACILITY. **Management Structure**. 2017. Disponível em: <https://www.arm.gov/about/management-structure>. Acesso em: 19 mai. 2020. 1

CALHEIROS, A. J. P. Relatório sobre os dados do projeto sos-chuva (pluviômetros, disdrômetros, mrr e mp3000a). 2018. Disponível em: <http://chuvaproject.cptec.inpe.br/portal/pdf/relatorios/Rel_dados_sos_chuva.pdf>. 18, 30

JOSS, J.; WALDVOGEL, A. Ein spektrograph für niederschlagstropfen mit automatischer auswertung. **pure and applied geophysics**, v. 68, n. 1, p. 240–246, 12 1967. Disponível em: <https://doi.org/10.1007/BF00874898>. 6

KWON OHBYUNG; LEE, N. S. B. Data quality management, data usage experience and acquisition intention of big data analytics. **International Journal of Information Management**, Elsevier Science, v. 34, 06 2014. 6

MACHADO, L. A. T. Previsão imediata de tempestades intensas e entendimento dos processos físicos no interior das nuvens: O sos- chuva (sistema de observação e previsão de tempo severo). 2015. Disponível em: <http://chuvaproject.cptec.inpe.br/portal/pdf/relatorios/Rel_dados_sos_chuva.pdf>. 2

METEOROLOGISCHE MESSTECHNIK GMBH. **MRR-2**: Micro rain radar user manual. Germany: METEK GmbH, 2011. 6, 7, 8, 9

PETERS, G.; FISCHER, B.; MUNSTER, H.; CLEMENS, M.; WAGNER, A. Profiles of raindrop size distributions as retrieved by microrain radars. **Journal of Applied Meteorology**, American Meteorological Society, Boston MA, EUA, v. 44, n. 12, p. 1930 – 1949, 2005. 7

REW, R.; DAVIS, G. Netcdf: an interface for scientific data access. **IEEE Computer Graphics and Applications**, v. 10, n. 4, p. 76–82, 1990. 3, 11

SILVA, F. C. C. da. **Gestão de Dados Científicos**. Rio de Janeiro: Editora Interciência Ltda, 2020. 3 p. ISBN 978-65-990252-2-8. 5

WIKIPEDIA. **dBZ (meteorology)**. 2014. Disponível em: <https://en.wikipedia.org/wiki/DBZ_(meteorology)>. Acesso em: 17 dez. 2020. 23, 24

WILKINSON, M. D. e. a. The fair guiding principles for scientific data management and stewardship. **Scientific Data**, v. 3, 3 2016. 1

WILKS, D. **Statistical methods in the atmospheric sciences**. Amsterdam Boston: Elsevier/Academic Press, 2011. ISBN 978-0-12-385022-5. 18

WORLD METEREOLOGICAL ORGANIZATION. **Forecast Verification methods Across Time and Space Scales**. 2017. Disponível em: <https://cawcr.gov.au/projects/verification/>. Acesso em: 29 mai. 2021. 18

ANEXO A - BIBLIOTECAS PYTHON NECESSÁRIAS PARA A EXECUÇÃO DAS FERRAMENTAS PROPOSTAS

```
alembic==1.1.0.dev0
appdirs==1.4.4
astroid==2.4.2
attrs==19.3.0
Automat==0.8.0
Babel==2.6.0
backcall==0.1.0
bcrypt==3.1.7
black==20.8b1
bleach==3.1.1
blinker==1.4
certifi==2019.11.28
chardet==3.0.4
click==7.1.2
cloud-init==21.2
colorama==0.4.3
command-not-found==0.3
configobj==5.0.6
constantly==15.1.0
cryptography==2.8
cycler==0.10.0
dbus-python==1.2.16
decorator==4.4.2
defusedxml==0.6.0
distlib==0.3.1
distro==1.4.0
distro-info===0.23ubuntu1
entrypoints==0.3
filelock==3.0.12
Flask==1.1.1
Flask-BabelEx==0.9.3
Flask-Compress==1.4.0
Flask-Gravatar==0.4.2
Flask-Login==0.4.1
Flask-Mail==0.9.1
```

```
Flask-Migrate==2.5.2
Flask-Paranoid==0.2.0
Flask-Principal==0.4.0
Flask-Security-Too==3.4.2
Flask-SQLAlchemy==2.1
Flask-WTF==0.14.2
html5lib==1.0.1
httplib2==0.14.0
hyperlink==19.0.0
idna==2.8
importlib-metadata==1.5.0
incremental==16.10.1
ipykernel==5.2.0
ipython==7.13.0
ipython-genutils==0.2.0
ipywidgets==6.0.0
isort==5.7.0
itsdangerous==1.1.0
jedi==0.15.2
Jinja2==2.10.1
jsonpatch==1.22
jsonpointer==2.0
jsonschema==3.2.0
jupyter-client==6.1.2
jupyter-console==6.2.0
jupyter-core==4.6.3
jupyterthemes==0.20.0
keyring==18.0.1
kiwisolver==1.3.1
language-selector==0.1
launchpadlib==1.10.13
lazr.restfulclient==0.14.2
lazr.uri==1.0.3
lazy-object-proxy==1.4.3
ldap3==2.4.1
lesscpy==0.14.0
Mako==1.1.0
```

```
MarkupSafe==1.1.0
matplotlib==3.4.1
mccabe==0.6.1
mistune==0.8.4
more-itertools==4.2.0
mypy-extensions==0.4.3
nbconvert==5.6.1
nbformat==5.0.4
netifaces==0.10.4
notebook==6.0.3
numpy==1.19.5
oauthlib==3.1.0
pandas==1.2.0
pandocfilters==1.4.2
paramiko==2.6.0
parso==0.5.2
passlib==1.7.2
pathspec==0.8.1
pexpect==4.6.0
pickleshare==0.7.5
Pillow==8.2.0
ply==3.11
prometheus-client==0.7.1
prompt-toolkit==2.0.10
psutil==5.5.1
psycopg2==2.8.4
pyasn1==0.4.2
pyasn1-modules==0.2.1
Pygments==2.3.1
PyGObject==3.36.0
PyHamcrest==1.9.0
pyinotify==0.9.6
PyJWT==1.7.1
pylint==2.6.0
pymacaroons==0.13.0
PyNaCl==1.3.0
pyOpenSSL==19.0.0
```

```
pyparsing==2.4.7
pyrsistent==0.15.5
pyserial==3.4
python-apt==2.0.0+ubuntu0.20.4.5
python-dateutil==2.7.3
python-debian===0.1.36ubuntu1
pytz==2020.5
PyYAML==5.3.1
pyzmq==18.1.1
QtPy==1.9.0
regex==2021.4.4
requests==2.22.0
requests-unixsocket==0.2.0
SecretStorage==2.3.1
Send2Trash==1.5.0
service-identity==18.1.0
simplejson==3.16.0
six==1.14.0
sos==4.1
speaklater==1.4
SQLAlchemy==1.3.12
sqlparse==0.2.4
ssh-import-id==5.10
sshtunnel==0.1.4
systemd-python==234
terminado==0.8.2
testpath==0.4.4
toml==0.10.2
tornado==5.1.1
traitlets==4.3.3
Twisted==18.9.0
typed-ast==1.4.2
typing-extensions==3.7.4.3
ubuntu-advantage-tools==27.0
ufw==0.36
unattended-upgrades==0.1
urllib3==1.25.8
```

```
virtualenv==20.2.2
wadllib==1.3.3
wcwidth==0.1.8
webencodings==0.5.1
Werkzeug==0.16.1
widgetsnbextension==2.0.0
wrapt==1.12.1
WTForms==2.2.1
zipp==1.0.0
zope.interface==4.7.1
```

PUBLICAÇÕES TÉCNICO-CIENTÍFICAS EDITADAS PELO INPE

Teses e Dissertações (TDI)

Teses e Dissertações apresentadas nos Cursos de Pós-Graduação do INPE.

Manuais Técnicos (MAN)

São publicações de caráter técnico que incluem normas, procedimentos, instruções e orientações.

Notas Técnico-Científicas (NTC)

Incluem resultados preliminares de pesquisa, descrição de equipamentos, descrição e ou documentação de programas de computador, descrição de sistemas e experimentos, apresentação de testes, dados, atlas, e documentação de projetos de engenharia.

Relatórios de Pesquisa (RPQ)

Reportam resultados ou progressos de pesquisas tanto de natureza técnica quanto científica, cujo nível seja compatível com o de uma publicação em periódico nacional ou internacional.

Propostas e Relatórios de Projetos (PRP)

São propostas de projetos técnico-científicos e relatórios de acompanhamento de projetos, atividades e convênios.

Publicações Didáticas (PUD)

Incluem apostilas, notas de aula e manuais didáticos.

Publicações Seriadas

São os seriados técnico-científicos: boletins, periódicos, anuários e anais de eventos (simpósios e congressos). Constam destas publicações o Internacional Standard Serial Number (ISSN), que é um código único e definitivo para identificação de títulos de seriados.

Programas de Computador (PDC)

São a seqüência de instruções ou códigos, expressos em uma linguagem de programação compilada ou interpretada, a ser executada por um computador para alcançar um determinado objetivo. Aceitam-se tanto programas fonte quanto os executáveis.

Pré-publicações (PRE)

Todos os artigos publicados em periódicos, anais e como capítulos de livros.