



UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Engenharia Elétrica e Computação

Luiz Antônio de Sousa Ferreira

**UTILIZAÇÃO DE TÉCNICAS DE
PRÉ-PROCESSAMENTO PARA CONSTRUÇÃO DE BANCO DE
DADOS SOLARIMÉTRICO A PARTIR DE DADOS AMBIENTAIS NA
AUSÊNCIA (DE CONFIABILIDADE) DA INFORMAÇÃO**

**USE OF PRE-PROCESSING TECHNIQUES FOR THE
CONSTRUCTION OF A SOLARIMETRIC DATABASE FROM
ENVIRONMENTAL DATA IN THE ABSENCE OF (RELIABILITY) OF
THE INFORMATION**

Campinas
2018



Luiz Antônio de Sousa Ferreira

**UTILIZAÇÃO DE TÉCNICAS DE
PRÉ-PROCESSAMENTO PARA CONSTRUÇÃO DE BANCO DE
DADOS SOLARIMÉTRICO A PARTIR DE DADOS AMBIENTAIS NA
AUSÊNCIA (DE CONFIABILIDADE) DA INFORMAÇÃO**

Dissertação apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia Elétrica, na Área de Telecomunicações e Telemática.

Orientador: Prof. Dr. Yuzo Iano

ESTE TRABALHO CORRESPONDE À
VERSÃO FINAL DA DISSERTAÇÃO
DEFENDIDA PELO ALUNO LUIZ
ANTÔNIO DE SOUSA FERREIRA E
ORIENTADA PELO PROF. DR. YUZO
IANO

Campinas
2018

Agência(s) de fomento e nº(s) de processo(s): Não se aplica.

ORCID: <https://orcid.org/0000-0002-3788-6971>

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Rose Meire da Silva - CRB 8/5974

F413u Ferreira, Luiz Antônio de Sousa, 1990-
Utilização de técnicas de pré-processamento para a construção de banco de dados solarimétrico a partir de dados ambientais na ausência (de confiabilidade) da informação / Luiz Antônio de Sousa Ferreira. – Campinas, SP : [s.n.], 2018.

Orientador: Yuzo Iano.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Processamento de dados. 2. Banco de dados. 3. Análise de dados. 4. Mineração de dados (Computação). I. Iano, Yuzo, 1950-. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Use of pre-processing techniques for the construction of a solarimetric database from environmental data in the absense of (reliability) of the information

Palavras-chave em inglês:

Data processing

Database

Data analysis

Data mining (computing)

Área de concentração: Telecomunicações e Telemática

Titulação: Mestre em Engenharia Elétrica

Banca examinadora:

Yuzo Iano [Orientador]

Silvio Renato Messias de Carvalho

Rogério Seiji Higa

Data de defesa: 14-12-2018

Programa de Pós-Graduação: Engenharia Elétrica

COMISSÃO JULGADORA - DISSERTAÇÃO DE MESTRADO

Candidato: Luiz Antônio de Sousa Ferreira

RA: 190750

Data da Defesa: 14 de dezembro de 2018

Título da Tese: “Utilização de técnicas de pré-processamento para construção de banco de dados solarimétricos a partir de dados ambientais, na ausência de registros ou confiabilidade da informação”

Prof. Dr. Yuzo Iano (Presidente)

Prof. Dr. Silvio Renato Messias de Carvalho

Prof. Dr. Rogério Seiji Higa

A ata de defesa, com as respectivas assinaturas dos membros da Comissão Julgadora, encontra-se no SIGA (Sistema de Fluxo de Dissertação/Tese) e na Secretaria de Pós-Graduação da Faculdade de Engenharia Elétrica e de Computação.

*"Por vezes sentimos que aquilo que fazemos não é senão uma gota de água no mar.
Mas o mar seria menor se lhe faltasse uma gota".
(Madre Teresa de Calcutá)*

Dedico essa dissertação à minha família.

Agradecimentos

Agradeço principalmente a Deus, por me dar a disposição e motivação em dias turbulentos que insistiam em me desafiar. Também aos meus pais, por serem os transmissores de energias positivas e incentivadores do realizar. Meu irmão que sempre se mostrou como o melhor amigo, sempre disposto a ouvir e aconselhar nos momentos em que mais necessitei.

Agradeço aos amigos que fiz durante essa etapa, principalmente os companheiros de morada, por muitas vezes serem o apoio que a família, distante fisicamente, talvez não conseguia. Posso chamá-los de disseminadores de ciência, seja ela atribuída à pesquisa feita nas salas desta universidade, seja ela ao aprendizado que a convivência proporcionou, tempos curtos que desejo que se prolonguem pela existência terrena.

Agradeço à minha namorada, por todo apoio, paciência, carinho e amor que foi fundamental nessa caminhada.

Agradeço ao professor Yuzo Iano, por sempre acreditar que o caminho que trilhamos dentro dessa universidade é, também, o caminho que pode levar nosso país a um novo patamar, que a ciência não está morta, mesmo quando diminuída, resiste em cada ser que deseja o melhor pra si e para o próximo. Agradeço toda a confiança depositada desde o primeiro dia em que estive nessa universidade.

Agradeço à banca de avaliação da minha qualificação, professor Rangel e Silvio, cujo conhecimento passado norteou o resultado final.

Agradeço aos companheiros de laboratório, a destacar o de maior convívio, cuja sabedoria e disposição para compartilhar experiências e conhecimento foram e são uma grande fonte de riqueza, onde habitam os maiores desejos de descoberta e curiosidade, elementos que fizeram e ainda fazem despertar a ambição de buscar patamares cada vez mais elevados.

Agradeço, também, aos meus amigos, que mesmo com minha ausência em diversas ocasiões, foram capazes de compreender minha busca por novas inspirações e que sempre torceram por mim, desejando o melhor durante essa caminhada.

Agradeço, por fim, ao Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas Gerais (IFSULDEMINAS), que proporcionou o período necessário para a realização das atividades do curso concomitante ao trabalho que desenvolvo junto a esta instituição. Somado a todos os citados, relembro aqueles que direta ou indiretamente, cruzaram meu caminho durante essa etapa de minha vida, positiva ou negativamente contribuíram para meu crescimento como pessoa, seja com suas histórias ou experiências adquiridas com esses eventos.

Resumo

A possibilidade da manipulação da informação surge como uma das alternativas para explorar novas soluções em diversas áreas do conhecimento, tanto para a descoberta de fatores não explorados, quanto para a elaboração de novos mecanismos que possam aumentar a produtividade e a eficiência onde aplicados. Técnicas de *Big Data* estão se tornando ferramentas poderosas quando falamos de análise exploratória, principalmente pela capacidade de manipulação de dados não-estruturados, como redes sociais, bancos com grande volume de dados, que podem ser colhidos de forma dinâmica e contínua, cujo valor agregado pode ser considerado uma riqueza dentro de qualquer ambiente organizacional. *Big Data* também pode ser utilizado em uma das maiores dificuldades encontrados no mundo, que é a produção de energia, cuja demanda tem se tornado crescente com o passar dos anos, fazendo com que a busca por alternativas dos métodos atuais seja necessária. Destacando a energia fotovoltaica, diversas aplicações podem ajudar a melhorar a eficiência e maximizar os ganhos com a produção. Nesse sentido, a utilização de métodos estatísticos e ferramentas de *Big Data* podem ser alternativas para a compensação de erros que os bancos de dados públicos e privados podem possuir, permitindo a realização de análises sobre dados consolidados. Nesse contexto, esta pesquisa atendeu a essas expectativas, trabalhando com maneiras de disponibilização de dados confiáveis, sejam eles coletados ou estimados, de forma pública, para que mais regiões possam ter suas demandas sobre aplicações de dados solares satisfeitas.

Palavras-chaves: processamento de dados; banco de dados; análise de dados; mineração de dados (computação).

Abstract

The possibility of manipulating information emerges as one of the alternatives to explore new solutions in several areas of knowledge, both for the discovery of unexplored factors and for the elaboration of new mechanisms that can increase productivity and efficiency where applied. Big data techniques are becoming powerful tools when we talk about exploratory analysis, mainly because of the ability to manipulate unstructured data, such as social networks, banks with large data volumes, that can be collected dynamically and continuously, whose added value can be considered a wealth within any organizational environment. Big Data can also be used in one of the greatest difficulties encountered in the world, which is the production of energy, whose demand has become increasing with the passing of the years, causing the search by alternatives of the current methods is required. Highlighting photovoltaic energy, several applications can help improve efficiency and maximize production gains. In this sense, the use of statistical methods and Big Data tools can be alternatives for the compensation of errors that public and private databases can have, allowing the analysis of consolidated data. In this context, this research has met these expectations, working with ways to make reliable data available, whether collected or estimated publicly, so that more regions can have their demands on solar data applications satisfied.

Keywords: data processing; database; data analysis; data mining (computing).

Lista de ilustrações

Figura 1 – Representação gráfica dos níveis de formação do conhecimento . . .	28
Figura 2 – Exemplificação de pré-processamento	55
Figura 3 – Gráfico de ausência de valores	51
Figura 4 – Gráfico de dados ausentes da base SONDA	67
Figura 5 – Boxplot da variável <i>global_irradiance</i> na base SONDA	68
Figura 6 – Boxplot da variável <i>humidity</i> na base SONDA	68
Figura 7 – Boxplot da variável <i>temperature</i> na base SONDA	69
Figura 8 – Boxplot da variável <i>wind_speed</i> na base SONDA	70
Figura 9 – Gráfico de dados ausentes da base SINDA	71
Figura 10 – Boxplot da variável <i>global_irradiance</i> na base SINDA	71
Figura 11 – Boxplot da variável <i>humidity</i> na base SINDA	72
Figura 12 – Boxplot da variável <i>precipitation</i> na base SINDA	72
Figura 13 – Boxplot da variável <i>wind_speed</i> na base SINDA	73
Figura 14 – Gráfico de dados ausentes da base POWER	74
Figura 15 – Boxplot da variável <i>humidity</i> na base POWER	74
Figura 16 – Boxplot da variável <i>precipitation</i> na base POWER	75
Figura 17 – Boxplot da variável <i>temperature</i> na base POWER	75
Figura 18 – Boxplot da variável <i>wind_speed</i> na base POWER	76
Figura 19 – Boxplot da variável <i>clarity_index</i> na base POWER	76
Figura 20 – Boxplot da variável <i>global_irradiance</i> na base POWER	76
Figura 21 – Boxplot da variável <i>clear_sky</i> na base POWER	77
Figura 22 – Gráfico de dados ausentes da base BDMEP horário	77
Figura 23 – Boxplot da variável <i>temperature</i> na base BDMEP horário	78
Figura 24 – Boxplot da variável <i>humidity</i> na base BDMEP horário	78
Figura 25 – Boxplot da variável <i>cloud_cover</i> na base BDMEP horário	79
Figura 26 – Boxplot da variável <i>wind_speed</i> na base BDMEP horário	79
Figura 27 – Gráfico de dados ausentes da base BDMEP diário	80
Figura 28 – Boxplot da variável <i>humidity</i> na base BDMEP diário	81
Figura 29 – Boxplot da variável <i>precipitation</i> na base BDMEP diário	81
Figura 30 – Boxplot da variável <i>wind_speed</i> na base BDMEP diário	82
Figura 31 – Variável <i>air_temperature</i> pré-processada	86
Figura 32 – Variável <i>clarity_index</i> pré-processada	86
Figura 33 – Variável <i>cloud_cover</i> pré-processada	87
Figura 34 – Variável <i>global_irradiance</i> pré-processada	87
Figura 35 – Variável <i>humidity</i> pré-processada	87

Figura 36 – Variável <i>precipitation</i> pré-processada	88
Figura 37 – Variável <i>temperature</i> pré-processada	88
Figura 38 – Variável <i>wind_speed</i> pré-processada	88

Lista de quadros

Quadro 1 – Descrição dos atributos e seus tipos	30
Quadro 2 – Estações principais do projeto com sua localização e altitude . . .	34
Quadro 3 – Estações parceiras do projeto com sua localização e altitude . . .	34
Quadro 4 – Exemplo de dados estatísticos de temperatura de uma região K . .	46
Quadro 5 – Relação de variáveis	60
Quadro 6 – Variáveis da base SINDA	62
Quadro 7 – Variáveis da base POWER	63
Quadro 8 – Variáveis da base BDMEP	64
Quadro 9 – Relação de variáveis e bases de dados	65

Lista de abreviaturas e siglas

AERONET	<i>Aerosol Robotic Network</i>
AnEn	<i>Analog Ensemble</i>
ANA	Agência Nacional de Água
ANEEL	Agência Nacional de Energia Elétrica
ANN	<i>Artificial Neural Networks</i>
BDMEP	Banco de Dados Meteorológicos para Ensino e Pesquisa
DCS	<i>Transponder</i> de coleta de dado
EUA	Estados Unidos da América
GA	<i>Genetic Algorithm</i>
GISC	Centro de Sistema de Informação Mundial
INMET	Instituto Nacional de Meteorologia
INPE	Instituto Nacional de Pesquisas Espaciais
ISOMAP	<i>Isometric Feature Mapping</i>
<i>hs</i>	Horas
<i>kgkg – 1</i>	Quilograma Quilograma-litro
<i>kPa</i>	Quilopascal
<i>kW – hr/m²/day</i>	Quilowatt-hora por Metro Quadrado por Dia
LLE	<i>Locally Linear Embedding</i>
<i>mps</i>	Metros por Segundo
MA	Maranhão
MAE	<i>Mean Absolute Error</i>
<i>max</i>	Máximo
MB	<i>Megabytes</i>
<i>mB</i>	<i>Milibar</i>

<i>mbar</i>	Milésimo de <i>Bar</i>
MICE	<i>Multivariate Imputation by Chained Equations</i>
<i>min</i>	Mínimo
<i>mm</i>	Milímetro
<i>mmday</i> – 1	Mílitro por Dia-litro
<i>m/s</i>	Metro por Segundo
NaN	<i>Not a Number</i>
NASA	<i>National Aeronautics and Space Administration</i>
NIST	<i>National Institute of Standarts and Technology</i>
NOAA	<i>National Centers for Environmental Information</i>
NWP	<i>Numerical Weather Prediction</i>
OMM	Organização Meteorológica Mundial
<i>oNV</i>	Graus em Norte Verdadeiro
PCA	<i>Principal Components Analysis</i>
PCD's	Plataformas de Coletas de Dados
POWER	<i>Prediction of Worldwide Energy Resource</i>
<i>qtd</i>	Quantidade
range	Variação
RMSE	<i>Root Mean Square Error</i>
SIAR	<i>Servicio Integral de Asesoramiento al Regante</i>
SINDA	Sistema Integrado de Dados Ambientais
SMS	<i>Short Message Service</i>
SONDA	Sistema de Organização Nacional de Dados Ambientais
SQL	<i>Structured Query Language</i>
SWERA	<i>Solar and Wind Energy Resource Assessment</i>
TB	<i>Terabytes</i>

UNEP	Programa das Nações Unidas para o Ambiente
W/m^2	Watt por Metro Quadrado
XML	<i>Extensible Markup Language</i>
°C	Graus Celsius

Lista de Publicações

- [1] Douglas Aguiar do Nascimento, Yuzo Iano, Hermes José Loschi, Luiz Antônio de Sousa Ferreira, José Antônio Donizete Rossi, and Clayton Duarte Pessoa. Evaluation of partial discharge signatures using inductive coupling at on-site measuring for instrument transformers. *International Journal of Emerging Electric Power Systems*, 19(1), 2018.
- [2] Paulo Eduardo dos Reis Cardoso, Yuzo Iano, Silvio Renato Messias de Carvalho, Hermes José Loschi, Fabiano Gustavo Silveira Magrin, Diego Arturo Pajuelo Castro, and Luiz Antônio de Sousa Ferreira. Reasons for sfn failure in broadcast. *SET INTERNATIONAL JOURNAL OF BROADCAST ENGINEERING*, 3:7, 2018.
- [3] Luiz Antônio de Sousa Ferreira, L Y Schwarzstein, Yuzo Iano, and Camila Santana Domingues. Alexa's case: vulnerability issues in iot devices in residential automation. *Brazilian Technology Symposiumm Proceedings*, 1:3, 2018.
- [4] Luiz AS Ferreira, Hermes J Loschi, Abel AD Rodriguez, Yuzo Iano, and Douglas A do Nascimento. A solar tracking system based on local solar time integrated to photovoltaic systems. *Journal of Solar Energy Engineering*, 140(2):021010, 2018. DOI: 10.1115/1.4039094.
- [5] Kelem Jordao, Yuzo Iano, Hermes José Loschi, Luiz Antônio de Sousa Ferreira, and David Bianchini. Smart city iot model: a prospective analysis considering indicators of social, technical and operational aspects. *International Journal of Energy Optimization and Engineering (IJE OE)*, 19(1), 2018.
- [6] Hermes José Loschi, Luiz Antônio de Sousa Ferreira, Yuzo Iano, Silvio R Messias Carvalho, Paulo Eduardo R Cardoso, and Douglas Aguiar Nascimento. Evaluation of solar position and angle of incidence calculation of the din5034 and nrel solpos algorithms. *Brazilian Technology Symposiumm Proceedings*, 1, 2017.
- [7] Hermes José Loschi, Yuzo Iano, Luiz Antônio F de Sousa, Douglas Aguiar, and Reinaldo Padilha França do Nascimento. An algorithm for development of transition probabilities matrices to predicting ramp events of the global solar irradiance, based on markov model. *International Journal of Development and Sustainability*, 6(11):20, 2017.
- [8] HJ Loschi, LAS Ferreira, Y Iano, DA Nascimento, PER Cardoso, FD Conte, and SRM Carvalho. Emc evaluation of off-grid and grid-tied photovoltaic systems for

the brazilian scenario. *Journal of clean Energy Tecnnologies*, 6(2):125–133, 2017.
DOI: 10.13140/RG.2.2.18086.63040.

Sumário

1	INTRODUÇÃO	21
1.1	Justificativa	22
1.2	Objetivos	24
1.2.1	Objetivos gerais	24
1.2.2	Objetivos específicos	24
2	CONCEITUAÇÃO TEÓRICA	26
2.1	Dado	26
2.2	Informação	26
2.3	Conhecimento	27
2.4	Caracterização de dados	28
2.4.1	Tipagem de dados	28
2.4.1.1	Quantidade	29
2.4.1.2	Qualidade	29
2.4.2	Escala	29
2.5	Dados ambientais e meteorológicos	31
2.5.1	Projeto SWERA	32
2.5.2	Projeto SONDA	33
2.6	Disponibilidade de dados públicos	34
2.7	Volume dos dados	35
2.8	Caracterização de <i>Big Data</i>	36
2.8.1	Os cinco V's	37
3	PRÉ-PROCESSAMENTO DE DADOS	40
3.1	Integração	41
3.2	Limpeza	42
3.2.1	Dados ausentes	43
3.2.2	Dados ruidosos	44
3.2.3	Dados inconsistentes	44
3.2.4	Dados duplicados	45
3.3	Transformação	45
3.3.1	Transformação por normalização	46
3.3.1.1	Normalização <i>Min-Max</i>	46
3.3.1.2	Normalização por <i>z-score</i>	47
3.3.1.3	Normalização por escala decimal	48

3.3.2	Agregação	49
3.3.3	Generalização	50
3.4	Redução	50
3.4.1	Redução de dimensionalidade	51
3.4.2	Redução por amostragem de dados	51
3.4.3	Redução de cardinalidade	53
3.4.4	Discretização	53
3.5	Resumo do capítulo	54
4	METODOLOGIA	56
4.1	Base teórica para a seleção das variáveis	56
4.1.1	Cervone, G. et al.	57
4.1.2	Gagne II, D. J. et al.	57
4.1.3	Urraca, R. et al.	58
4.1.4	Lima, F. J. L. et al.	59
4.1.5	Zhu, G. et al.	59
4.1.6	Resumo da seção	60
4.2	Fontes de dados ambientais para extração	60
4.2.1	SINDA	62
4.2.2	POWER	62
4.2.3	BDMEP	64
4.3	Uso de técnicas de pré-processamento	65
5	RESULTADOS	66
5.1	As incongruências das bases de dados ambientais	66
5.1.1	Base de dados SONDA	66
5.1.2	Base de dados SINDA	70
5.1.3	Base de dados POWER	73
5.1.4	Base de dados BDMEP - Horário	77
5.1.5	Base de dados BDMEP - Diário	80
5.2	Adequação e montagem da base	82
5.2.1	Variáveis calculadas	82
5.2.2	Variáveis extraídas das bases de dados	83
6	CONSIDERAÇÕES FINAIS	89
6.1	Trabalhos futuros	90
	REFERÊNCIAS	91

APÊNDICES	95
APÊNDICE A – TABELA DE DADOS PRÉ-PROCESSADOS	96
APÊNDICE B – CÓDIGO FONTE	112

1 Introdução

Atualmente, o mercado entende que as soluções energéticas disponíveis e convencionalmente utilizadas podem se tornar um problema futuro, o qual pode estar relacionado aos custos na produção de energia elétrica que podem se tornar elevados, dependendo da forma com que é produzida e também os custos relacionados à sustentabilidade, visto que, a maior parcela da produção energética utilizada por todo o mundo baseia-se em meios que afetam consideravelmente o meio-ambiente.

Somado a isso, a escassez de recursos, como o petróleo, aliada à busca pela diminuição de custos na produção da energia elétrica, apresentam um cenário favorável às soluções renováveis, em que a energia limpa pode ser adquirida através de recursos naturais, sem grandes prejuízos à biodiversidade em que o sistema está inserido, e que ainda não são explorados de forma significativa pelo setor.

Porém, a imprevisibilidade das condições ambientais necessárias para a produção surge como um dos maiores problemas nos meios de produção de energia elétrica de forma sustentável. A instabilidade constante das condições climáticas faz com que as soluções sustentáveis sejam utilizadas como matrizes de produção energética secundária, por não apresentarem um grau de certeza aceitável para uma previsão de tamanha importância. Essa informação é considerada importante, principalmente pelas distribuidoras de energia que, geralmente, possuem contratos com estratégias de crescimento pré-estabelecidas, fazendo com que a incerteza seja um fator chave para a resistência na adoção dessas soluções, mesmo que isso signifique um custo maior para produzir.

Com vista à predição da produção de energia solar fotovoltaica, podemos citar a utilização de técnicas de *big data* para a resolução de problemas desse tipo, de forma eficiente, utilizadas em diversas áreas, com destaque às pesquisas em áreas médicas.

Dentre as técnicas que são utilizadas para o desenvolvimento de uma predição, podem ser mencionados o uso de Redes Neurais Artificiais (doravante ANN, do inglês, *Artificial Neural Networks*), Algoritmos Genéticos (doravante GA, do inglês, *Genetic Algorithm*) e modelos de regressão, os quais são mais utilizados para a reali-

zação desse tipo de atividade [10].

Nesse sentido, a soma de grandes quantidades de dados provenientes de diversas fontes são características de problemas de *big data* e, muitas vezes, a união das individualidades de cada base de dados podem resultar em uma nova informação valiosa na resolução de problemas [33].

Com isso, a quantidade de dados empíricos pode ser relacionada à precisão do método, tendo em vista que a variabilidade dos dados trabalhados por essas técnicas permite identificar padrões distintos conforme o período observado. Portanto, a coleta de dados pode ser importante para o sucesso de uma predição, assim como para se alcançar uma quantidade de dados adequada, técnicas como *Data Fusion* podem ser aplicadas quando é possível coletar os mesmos dados em interfaces distintas.

No contexto abordado, a realização de atividades ligadas a dados empíricos na área de sistemas de energia fotovoltaica tem sido uma alternativa à imprevisibilidade presente ao analisar o meio ambiente, mas, conforme podemos observar em estudos recentes, a falta dos dados coletados e a restrita distribuição dos sistemas de captura de tal informação ainda é um obstáculo a ser superado. Mesmo iniciativas como o projeto SONDA (Sistema de Organização Nacional de Dados Ambientais), coordenado pelo Instituto Nacional de Pesquisas Espaciais (INPE) e o projeto SWERA (*Solar and Wind Energy Resource Assessment*), desenvolvido pelo Programa das Nações Unidas para o Ambiente (UNEP), cujos objetivos centram em compilar e tornar disponíveis informações confiáveis sobre o recurso solar, com o objetivo de atrair investimentos de empresas governamentais e privadas na área [19], ainda não são capazes de apresentar dados suficientes para elaboração de predições com alto grau de confiabilidade e outras atividades do setor energético.

1.1 Justificativa

Entre os problemas relatados por pesquisadores da área de aplicações em sistemas de energia fotovoltaica, a dificuldade de encontrar as informações necessárias de forma confiável se destaca. Observamos que o acesso a estações de coleta de dados solarimétricos ainda é restrito a poucos lugares no mundo, devido às complexi-

dades políticas, financeiras e logísticas, do mesmo modo que as estações existentes não possuem capacidade de cobrir uma área territorial significativa. Com isso, propostas de utilização de métodos estatísticos surgem como opção à falta da coleta em tempo real.

Mesmo com o uso de modelos estatísticos, muitas vezes, os dados não são suficientes para que o problema possa ser solucionado; situações estas que forçam a busca de mais alternativas. Nesse sentido, o uso de dados empíricos para realizar extrapolações permite a criação de padrões que podem ser utilizados na solução, mas quando o dado é faltante, em energia fotovoltaica, o uso de dados ambientais podem ser uma alternativa a preencher essa lacuna.

Entre os projetos que captam dados ambientais, a base de dados do projeto SONDA é a mais conhecida no Brasil por seu trabalho com dados solarimétricos. Apesar de ser atuante em apenas alguns municípios espalhados ao longo das regiões do país, sua contribuição pode ser um ponto a se somar a desenvolvimentos de aplicações desse tipo. Porém, mesmo com os dados coletados a todo momento pelos equipamentos especializados, a confiabilidade da informação pode sofrer variações que podem comprometer o funcionamento das aplicações, uma vez que, ou o dado suspeito ou faltante podem desequilibrar as equações utilizadas. Existem também problemas decorrentes do não repasse de recursos públicos que afeta a disponibilização dos dados, fazendo com que exista um intervalo obscuro que compromete os projetos nos quais está relacionado.

Com pesquisas relacionadas à manipulação de dados, podemos obter uma maior abrangência da área coberta por projetos já existentes, ou, até mesmo, podemos utilizar como uma comprovação da veracidade da informação disponibilizada nos locais atendidos. Os dados ambientais coletados de forma isolada também podem ter seu valor agregado de forma positiva ao pensar no detalhamento de cada microrregião, o que ampliaria as minúcias do mapa solarimétrico no país, desde que consolidado em uma base de dados robusta e de qualidade.

1.2 Objetivos

Este estudo tem como foco a construção de um banco de dados utilizando técnicas de pré-processamento, cuja finalidade é produzir um meio de obtenção de dados confiáveis colhidos através de projetos abertos, governamentais ou não, para servir como base em pesquisas e estudos que possuam relação com o tema armazenado.

1.2.1 Objetivos gerais

Como objetivo geral, buscamos coletar bases de dados solarimétricos disponíveis com informações do território brasileiro, sendo o mais conhecido o Sistema de Organização Nacional de Dados Ambientais (SONDA), iniciativa do governo brasileiro, através do Instituto Nacional de Pesquisas Espaciais (INPE), que investe em pesquisas relacionadas à clima e produção de energia em diversos projetos.

A partir das informações coletadas, almejamos elaborar um banco de dados com informações completas com alternativas aos dados considerados suspeitos ou ausentes. Buscamos usar fontes alternativas de dados públicos, utilizando técnicas de pré-processamento de dados para elaboração de uma base única e de qualidade considerável.

Nesse íterim, pretendemos fazer um levantamento histórico de obtenção de variáveis de um *dataset* solarimétrico com base em cálculos estatísticos, especificamente no preenchimento de informações ausentes ou dados suspeitos.

1.2.2 Objetivos específicos

A pesquisa tem como objetivos específicos:

- Executar buscas de fontes de dados públicos relevantes para a área de estudo;
- Unificar bancos de dados solarimétricos com informações sobre o território brasileiro;
- Elaborar métodos para complementação do banco de dados unificados, substituindo os dados considerados ausentes e/ou suspeitos por dados obtidos de outras fontes;

- Construir uma base de dados com ênfase em dados para atividades relacionadas com construção de aplicações na área de estudo, principalmente, elaboração de predições utilizando aprendizado de máquina;
- Desenvolver uma ferramenta de consulta aos dados consolidados, através de um navegador de internet, com intenção de ser disponibilizada publicamente como fonte de pesquisa para analistas da área.

2 Conceituação teórica

Em muitas áreas, é comum encontrar referências à importância dos dados para as mais diversas atividades, sejam elas analíticas, econômicas, linguísticas, espaciais, ou seja, começa a surgir um consenso sobre a importância do tema na sociedade e na academia. Porém, podemos notar que nem sempre a utilização do termo se enquadra corretamente na situação.

No artigo de Setzer [26], são apresentadas definições de termos desde a aquisição do dado até a geração do conhecimento que as informações obtidas podem revelar.

2.1 Dado

Dados são itens capturados de forma independente, cuja coleta pode ser feita através de sensores, de coletores de informação ou qualquer meio informatizado (ou não) que representa alguma situação, mas que, se analisado separadamente, não possui significado claro, ou seja, não possui uma semântica aparente.

Segundo Setzer [26], dado é uma sequência de símbolos quantificados ou quantificáveis. Definição que nos leva a pensar que o dado é o termo mínimo de um cenário em observação.

2.2 Informação

A informação será gerada a partir da contextualização dos dados. A abstração gerada com a interpretação dos dados é chamada de informação. Notamos que, mesmo o entendimento processado dos dados não compõe uma abstração formal, isso porque não é possível armazenar, a nível de computação, uma informação, usando uma teoria lógica ou matemática. Para que possa ser processada de maneira formal é necessário reduzir a fase de dados.

O armazenamento de informações somente é possível através de representação em forma de dados. Destacando que apenas a forma de dados é armazenada,

uma maneira de gerar uma representação e não apenas a informação em si, exige um entendimento informal dos dados, conforme mencionado anteriormente. Isso pode ser ratificado com as palavras de Setzer [26], que afirma que "Uma distinção fundamental entre dado e informação é que o primeiro é puramente sintático e o segundo contém necessariamente semântica".

O uso da semântica é uma habilidade exclusiva do pensar, assim, embora as máquinas tenham a capacidade de armazenar e processar dados adquiridos das mais diversas formas, elas não possuem ainda a capacidade de interpretá-los.

Isto posto, podemos fazer um comparativo com os exemplos levantados por Searle [25], que fala sobre o "Quarto Chinês" para tornar mais compreensível a noção de semântica, quando no contexto computacional, em que temos tabelas com nomes e localizações que podem ser escritos em chinês e em português, sendo que, se estiver escrito em chinês, com o uso de seus ideogramas, para quem não tem conhecimento da língua, torna-se impossível interpretar a informação descrita. Da mesma forma que os mesmos dados escritos em português adquirem sentido para os falantes nativos dessa língua e demais conhecedores do idioma, mas impossibilita a interpretação daqueles que possuem fluência somente no chinês (Searle[25] apud Setzer [26]).

2.3 Conhecimento

O conhecimento é a experimentação da informação com vista a entender os dados de forma objetiva; um olhar subjetivo para as informações obtidas, destacando-se a experiência. Uma informação que já foi trabalhada por alguém que teve a nuance de criar interpretações com base no que lhe foi oferecido. O conhecimento não pode ser inserido em um computador, pois depende, exclusivamente, do tátil de quem o manipula. Nesse sentido, enquanto a informação é relacionada à semântica computacional, o conhecimento está ligado à pragmática da mesma [26].

Tomando por base as informações passadas, podemos imaginar uma pirâmide de dados, onde, em nível hierárquico, sendo dado o elemento com mais quantidade no ambiente exploratório, capaz, ou não, de gerar informações pertinentes que após analisados podem ser o ingrediente principal de um conhecimento real do ambiente em que está inserido, como podemos ver na Figura 1.

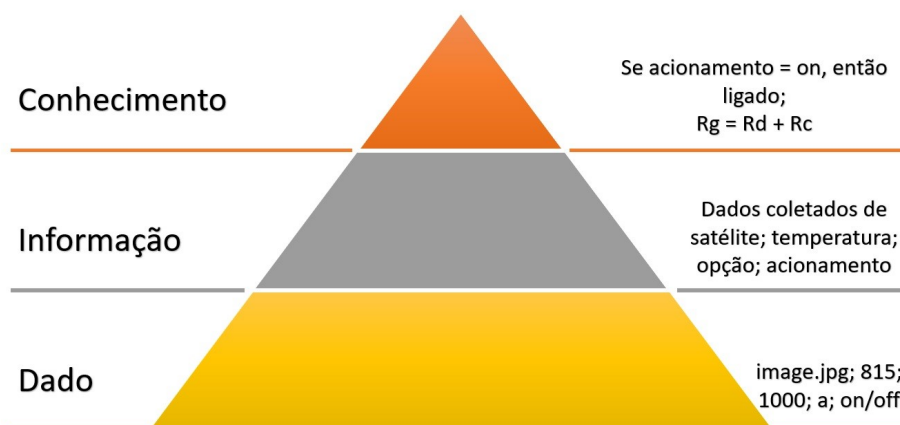


Figura 1 – Representação gráfica dos níveis de formação do conhecimento

2.4 Caracterização de dados

Para representar formalmente os dados, imaginamos uma matriz de objetos X_n^d , onde n representa a quantidade de objetos a serem tratados nessa matriz e d o número de atributos de cada objeto na mesma matriz. Para visualizar, podemos afirmar que cada elemento X_i^j ou X_{ij} possui o valor da j -ésima características para o i -ésimo objeto, onde os atributos d estão dispostos nos eixos ortogonais [6].

2.4.1 Tipagem de dados

Podemos utilizar dados para abstrair quase todos os elementos do mundo em que vivemos, essa característica nos faz observar melhor o mundo ao nosso redor, fazendo com que produzamos informações reais e nos permita alcançar o conhecimento quanto à nossa natureza, atitudes, comportamentos, entre outros.

A partir desse conceito, dispomos de uma gama infinita de atributos que, dispersos, não possuem significado aparente. Uma forma de começarmos a analisar essa informação está em dividi-los conforme o uso de seus atributos.

Na literatura, encontramos diversas maneiras e conotações para distinguir atributos conforme a sua tipagem, porém, muitos autores convergem para a distinção entre quantidade e qualidade, como podemos observar a partir de Faceli [6]:

2.4.1.1 Quantidade

Também encontrado como atributos quantitativos ou numéricos, são conjuntos de dados que possuem a característica de poder ser manipulado aritmeticamente, além de também poder ser ordenado conforme o problema, pois seu significado não necessariamente carece de uma ordem específica para a análise.

Os atributos quantitativos ainda possuem duas distinções que são cruciais para o melhor entendimento do problema, são elas: contínuo e discreto.

Atributos contínuos são aqueles que podem assumir um valor infinito, uma das formas de se obter esse tipo de atributo é através de coleta de informações em um sensor. Como exemplo, encontramos atributos quantitativos contínuos quando nos deparamos com medidas de peso, tamanho e distância. Esse tipo de dado aparece ligado a uma unidade, o que é fundamental para a distinção de qualquer ambiguidade dentre os dados coletados.

Atributos discretos são aqueles que estão em um intervalo contável de valores, um caso desse tipo de atributo, geralmente muito utilizado, são os dados binários ou booleanos, cujo valor pode ser chaveado entre dois valores: sim/não, 0/1, verdadeiro/falso, entre outros. Os atributos quantitativos discretos podem ser números, porém, nenhuma operação aritmética fará sentido, uma vez que o uso de soma, subtração multiplicação e divisão, por exemplo, não trará nenhuma informação válida relevante.

2.4.1.2 Qualidade

Quando encontrado qualitativo, simbólico e categórico, também falamos do tipo de atributo em questão, eles possuem valores que podem ser ligados à categorização da informação buscada, como exemplo: pequeno, médio, grande, interno, externo, residencial, industrial, comercial, entre outros. Os atributos qualitativos também podem ser uma lista de números, porém, assim como os atributos quantitativos discretos, iterações matemáticas não relevaram significados relevantes.

2.4.2 Escala

Antes de realizar uma análise dos dados de uma forma mais eficiente, a utilização de uma separação dos dados conforme sua escala é uma ferramenta muito

importante para que não haja um trabalho executado de forma equivocada, restringindo a interpretação de dados de forma com que cálculos errôneos não aconteçam. Os termos nominal, ordinal, intervalar e racional foram utilizados por Stevens [27] apud Velleman [31] por volta dos anos 40, onde estabeleceu uma forma de obter essa escala de maneira hierárquica.

Os resultados obtidos pelo autor, cujo trabalho inicial foi na área de psicologia, estabeleceu as operações que poderiam ser utilizadas em cada atributo dentro da escala [28].

Quadro 1 – Descrição dos atributos e seus tipos

Tipo do atributo	Descrição	Exemplos	Operações
Categórico (Qualitativo)			
Nominal	Os valores de um atributo nominal são apenas nomes diferentes; isto é, os valores nominais fornecem apenas informações suficientes para distinguir um objeto do outro. (=, =)	CEP, número do ID de funcionários, cor do olho, gênero	Modo, entropia, contingência, correlação, teste X2
Ordinal	Os valores de um atributo ordinal fornecem informações suficientes para ordenar objetos. (<, >)	Dureza de materiais, bom, melhor, superior, séries, número de rua	Mediana, percentis, rank de correlação, teste de corrida, testes de sinais
Numérico (Quantitativo)			
Intervalar	A diferença entre valores é significativa, isto é, uma unidade de medida existente. (+, -)	Datas de calendário, temperatura em Celsius ou Fahrenheit	Média, desvio padrão, correlação de Pearson's, teste t e F
Racional	Para variáveis racionais, ambas diferenças e proporções são significativas. (*, /)	Temperatura em Kelvin, quantidades monetárias, conta, idade, massa, tamanho, corrente elétrica	Média geométrica, média harmônica, variação percentual

2.5 Dados ambientais e meteorológicos

Dados meteorológicos e ambientais são aqueles que interagem entre si e produzem reações dentro da atmosfera do planeta, tendo a irradiação solar como a principal variável, sendo as demais presentes no meio ambiente influenciadas pela presença ou ausência desse fator solar [17].

Ao entrar na atmosfera terrestre, a energia emanada pelo sol pode sofrer alterações, tanto no sentido quanto no potencial energético. Dentre as moléculas e partículas responsáveis por esse fenômeno podemos citar: vapor de água, dióxido de carbono, ozônio, compostos nitrosos, entre outros.

Para a montagem de um banco de dados robusto para o setor devemos levar em consideração diversos aspectos para que o funcionamento do sistema de produção seja realizado da melhor forma possível. Dentre eles, a correlação dessas variáveis que interferem na produção energética a partir do sol, para que seja possível explorar os dados para obter novas informações ou modelos relevantes para o setor, a consistência do banco de dados de onde as informações devem ser extraídas deve ser elevada. Porém, devido a altos custos de montagem e manutenção de tais estações de coleta, é comum encontrarmos localidades onde tal coleta é errante ou, até mesmo, inexistente.

Levando em consideração esse cenário nebuloso de falta de confiabilidade ou disponibilidade de dados do tipo por projetos como o SONDA (Sistema Nacional de Organização de Dados Ambientais) ou o projeto SWERA (*Solar and Wind Energy Resource Assessment*), cuja atividade no território brasileiro findou-se em 2005, mas que continua em diversos territórios pelo globo, pesquisadores trabalham na descoberta de formas para se obter dados preditivos da atividade energética solar utilizando dados provenientes de outras fontes, porém, sempre esbarrando em questões como a adequação dos dados e atributos específicos para o desenvolvimento de sua pesquisa. Dificuldades estas que passam também pelo volume e quantidade de dados disponíveis por organizações, geralmente governamentais, encontradas de forma pública e acessível.

O armazenamento de dados ambientais consiste na coleta de todas as informações que formam um conjunto com observação meteorológica e de clima, que se utiliza de diversos sensores responsáveis por captar cada momento em um deter-

minado local.

As informações climáticas são utilizadas nos mais diversos setores da sociedade, seja para conhecimento do ambiente onde habita ou elaborar planejamentos estratégicos pela indústria. Entre muitos exemplos podemos citar a agricultura, em que conhecer a situação climática pode fazer com que uma colheita seja mais proveitosa ou um defensivo possa ser mais eficiente. Assim como se faz importante no monitoramento do clima por responsáveis sobre voos, em que as condições climáticas podem ser determinantes para a realização de uma viagem com segurança.

Nesta perspectiva, o uso de dados ambientais climáticos representa um fator chave na elaboração de estratégias, também no setor energético, que pode ajudar a calcular a variação da capacidade de geração de energia, adotando planejamentos para balancear a produção em uma estiagem, em que as hidrelétricas podem ter um baixo nível de sua represa, como, por exemplo, seria um painel solar cujo desempenho pode ser afetado com a quantidade de nuvens no céu.

2.5.1 Projeto SWERA

Em 1972, foi criado o Programa das Nações Unidas para o Meio Ambiente (UNEP, do inglês United Nations Environment Programme), para coordenar esforços de todos os países membros com objetivo de promover ações voltadas para o desenvolvimento sustentável. Como um dos projetos do Programa, o projeto SWERA (Projeto de Avaliação de Recursos de Energia Solar e Eólica, do inglês, Solar and Wind Energy Resource Assessment Project) foi criado com o objetivo de tornar disponível de forma pública um levantamento das informações coletadas em países em desenvolvimento, como forma de atrair investimentos públicos e privados no setor de energia, focando em fontes renováveis ¹.

O resultado final do projeto foi uma base inicial com dados capazes de proporcionar a criação de mapas de composição solar e eólica dos países atendidos, bem como, apresentação de relatórios socioeconômicos, inclusive questões de infraestrutura da matriz energética e sua distribuição [19].

O Brasil esteve entre os países em desenvolvimento escolhidos para o pro-

¹ ONU. Disponível em <https://www.unenvironment.org/about-un-environment/what-we-do>. Acesso em 01/2018.

grama, que teve seu início no ano de 2002, concluindo os trabalhos em 2005 [24]. Dentre os interesses do país no projeto estavam:

- a elaboração de mapas de potencial energético renovável de alta resolução do território nacional;
- a elaboração de mapas de potencial energético renovável de baixa resolução para a América do Sul;
- o levantamento de cenários, onde a utilização de energia solar poderia ser viável;
- a ampla divulgação dos resultados de forma pública, principalmente em seu website.

No decorrer do programa, o Instituto Nacional de Pesquisas Espaciais (INPE), órgão vinculado ao Ministério de Ciência, Tecnologia, Inovações e Comunicações, criou um projeto paralelo com vistas à coleta e validação dos dados apresentados pelo SWERA.

2.5.2 Projeto SONDA

O projeto SONDA, projeto de iniciativa do INPE em parceria com diversos institutos de pesquisas e universidades brasileiras, tem como principal objetivo implementar e gerenciar infraestrutura de uma base de dados necessária pelo levantamento de recursos de energia de fontes renováveis, principalmente, eólica e solar no Brasil, bem como, formas de planejar o uso de informações adquiridas com o passar do tempo para uma melhor exploração do potencial energético do país.

O projeto reúne dados e informações de projetos extintos ou obsoletos de redes de coleta de informações, provenientes de outros programas do governo federal, iniciativa privada e outros órgãos como o UNEP (Programa das Nações Unidas para o Meio Ambiente, do inglês, *United Nations Environment Programme*) com o projeto SWERA. A elaboração de Atlas Brasileiro de Energia Solar é uma das formas de atingir os objetivos de forma concreta, com a exposição de mapas solarimétricos do território nacional, cuja primeira edição foi desenvolvida dentro do projeto SWERA, como um dos resultados do mesmo. O Atlas propõe impulsionar planejamento e de-

envolvimento de políticas públicas a projetos nacionais de energia solar e eólica, com vista a atrair investimentos da iniciativa privada [22].

Na sua edição mais recente, lançada no ano de 2017, o Atlas Brasileiro de Energia Solar, apontou a existência de 17 estações de observação espalhadas de forma estratégica por todo o território nacional, sendo 12 operadas pelo próprio INPE (Quadro 2) e as demais administradas por parceiras do programa (Quadro 3).

Quadro 2 – Estações principais do projeto com sua localização e altitude

Estação	UF	Latitude	Longitude	Altitude
Brasília (BRB)	DF	-15,601	-47,713	1023
Cachoeira Paulista (CPA)	SP	-22,690	-45,006	574
Caicó (CAI)	RN	-6,467	-37,085	176
Campo Grande (CGR)	MS	-20,438	-54,538	677
Cuiabá (CBA)	MT	-15,555	-56,070	185
Natal (NAT)	RN	-5,837	-35,206	58
Ourinhos (ORN)	SP	-23,000	-49,844	502
Palmas (PMA)	TO	-10,178	-48,362	216
Petrolina (PTR)	PE	-9,069	-40,320	387
Rolim de Moutra (RLM)	RO	-11,582	-61,774	252
São Luiz (SLZ)	MA	-2,593	-44,212	40
São Martinho da Serra (SMS)	RS	-29,443	-53,823	489

Quadro 3 – Estações parceiras do projeto com sua localização e altitude

Estação	UF	Latitude	Longitude	Altitude
Chapecó (CHP)	SC	-27,080	-52,614	700
Curitiba (CTB)	PR	-25,495	-49,331	891
Florianópolis (FLN)	SC	-27,602	-48,518	31
Joinville (JOI)	SC	-26,253	-48,858	48
Sombrio (SBR)	SC	-29,096	-49,813	15

A localização dessas estações foram escolhidas a fim de representar as características climáticas de cada região do território nacional, levando em consideração as possibilidades logísticas para a implantação de sistemas de geração de energia renovável, compreendendo cada estação a um raio inferior a 40 quilômetros dentro da região em que está localizado [23].

2.6 Disponibilidade de dados públicos

A valorização do dado, atualmente, tem trazido diversos casos sobre sua propriedade e disponibilidade, afinal, no mundo de hoje, "dado é dinheiro", e o inte-

resse cresce conforme surgem novas formas de se utilizar os mesmos. A propriedade dos dados, costumeiramente, se transformam em problemas, em que o usuário acredita que seus dados estão sobre sua posse, porém, a todo momento, grandes empresas como Facebook, Twitter, Instagram e muitas outras usufruem dessas informações para obtenção de lucros, o mesmo vale para os dados em nuvem [12].

Quanto ao acesso desses dados, algumas companhias restringem seu uso, fazendo com que apenas possam ser utilizadas internamente e que, quando liberado, somente se dê através do pagamento de uma taxa para obtenção do privilégio, ainda podemos citar algumas empresas que permitem o acesso a uma amostra de seus dados para o desenvolvimento de pesquisa na academia. Esse cenário favorece o surgimento de situações em que apenas universidades com parcerias fortes ou com mais recursos possam realizar seus trabalhos com acesso aos dados necessários. Como consequência, surgem divisões quanto as possibilidades, uma vez que seriam necessários recursos, nem sempre disponíveis, para acessar tais dados.

Citando Virginia Rometty, em seu discurso na 157^a formatura da Northwestern University,

O que o vapor foi para o século 18, a eletricidade para o século 19 e hidrocarbonetos para o século 20, os dados serão para o século 21. É por isso que eu chamo os dados de um novo recurso natural [5].

A importância que essa situação suscita, deve estimular a discussão sobre os valores e vieses dessa nova onda [2].

2.7 Volume dos dados

Com o advento da tecnologia móvel, o surgimento de dispositivos chamados IoT (*Internet of Things*, do inglês, Internet das coisas) e com a expansão contínua da internet, o hábito de armazenar dados para traçar planejamentos em torno de sua atividade passou a ser importante para o crescimento dos ganhos de quem o faz. Seja pela coleta dos dados de navegação do usuário ou as compras feitas em uma loja, obter conhecimento sobre seu cliente é uma das estratégias para atrair consumo em seu estabelecimento. Com o surgimento da prática de análise desses dados, esse negócio apresentou-se um fator definitivo para a expansão dos lucros, tendo em vista a pos-

sibilidade de fornecer propaganda, promoções e descontos direcionados de acordo com o perfil traçado.

Grandes companhias passaram a traçar suas estratégias de venda direcionadas ao histórico de cada consumidor, criando a sensação de personalização em cada indivíduo cercado por esse método, que hoje apresenta mais claramente em propagandas de banners em sites de maior uso, que são capazes de oferecer um produto baseado no perfil de buscas na rede mundial de computadores.

A partir desse cenário, diversos outros setores passaram a observar com bons olhos a análise de dados realizada com a junção de seus bancos de dados que, em diversos casos, eram tratados de forma isolada, para atividades específicas. Uma visão macro de seu negócio possibilitou a identificação de padrões nunca antes percebidos por seus administradores, abrindo um leque de oportunidades para exploração de novas fontes de recursos.

Exemplo disso é que a grande rede de varejo Walmart, em 2012, coletava cerca de 1 petabyte de informações a cada hora sobre suas transações, um pacote de dados que incluía o comportamento, localização e dispositivos de quem realizava compras [3].

2.8 Caracterização de *Big Data*

Laney [15], ao destacar o novo conceito chamado de *Big Data*, o classificou em três dimensões, os chamados 3 V's, são eles: Volume, Variedade e Velocidade, como fatores chaves na identificação de um problema do tipo. Com o avançar do tempo, mais duas características foram adicionadas as referidas por ele como uma das formas de detalhar ainda mais o funcionamento dessa tecnologia de uma maneira mais eficiente e valorosa possível: Valor e Veracidade. Em meados dos anos 2000, surgiram os primeiros trabalhos acadêmicos utilizando esse termo pra especificar pesquisa sobre coletas de dados, gerando, através de análise exploratória, relatórios que poderiam servir para construção de decisões que poderiam alavancar os trabalhos em diversas áreas.

Originalmente, *Big Data* era tratado como um montante elevado de dados cujo processamento não poderia ser feito de forma eficiente com bancos de dados e

ferramentas tradicionais e, conforme surgiam novos meios de armazenar a quantidade de dados, também crescia a facilidade de se acessar [12]. Com a crescente geração de dados no mundo todo, somado ao advento do IoT, também sensores e instrumentos de medição puderam evoluir e, graças ao compartilhamento da rede, tornaram-se uma fonte de riqueza a ser explorada, onde a geração de conhecimento é o carro chefe do desenvolvimento de centros de pesquisas e da indústria.

Essa tendência pode ser confirmada com a evolução do crescimento dos dados nos anos subsequentes. Por exemplo, em 2012, durante debate presidencial nos EUA, quando comentários na rede social Twitter atingiram a casa dos 10 milhões de interações em apenas duas horas. Outro exemplo do grande tráfego de informações que a internet proporcionou foi a rede de compartilhamento de fotos Flickr que, declarou ter recebido cerca de 1,8 milhões de fotos por dia, em média, entre fevereiro e março de 2012, assumindo uma qualidade baixa de 2 MB por foto, o armazenamento desses dados chegou à casa de 3,6 TB de informações em um dia, o que exigiu uma evolução em questões de armazenamento dentre as técnicas disponíveis na época [34].

Portanto, adotamos, de acordo com o *National Institute of Standards and Technology* (NIST), a definição de *Big Data* como: "Dados dos quais a velocidade de aquisição do volume de dados ou a representação de dados limitam a capacidade de usar métodos relacionais tradicionais para realizar análises efetivas"².

2.8.1 Os cinco V's

O primeiro, e talvez o mais importante "V" trata-se do Volume, com o acréscimo da produção de dados ao redor do mundo, um dos maiores desafios computacionais nos últimos anos é tratar a quantidade de dados a serem armazenados.

Na área de distribuição de energia, um item que está ganhando destaque no mercado é o medidor de energia inteligente, que permite conhecer melhor o perfil de consumo da localidade com base nos dados extraídos pelo equipamento. De acordo com Tu [29], o número de dispositivos desse tipo subiu para 29,9 milhões de unidades no ano de 2017, sendo que, em 2011, esse número era de 10,3 milhões de

² Informação extraída do site do NIST *Big Data Program*, disponível em <http://bigdatawg.nist.gov>. Acesso em 01/2018.

unidades. Para estimarmos a dimensão desse crescimento, um milhão desses dispositivos são capazes de coletar cerca de até 2920 TB de dados, desde que medidos em intervalos de 15 minutos, durante 1 ano [14].

Com o advento de tecnologias que permitem ter sensores e comunicação em quase todo lugar do mundo, um bombardeio de informações pode acontecer a todo momento, sejam dados de monitoramento pessoal com vistas ao controle da saúde, sejam sobre tráfego, sobre economia e todos os demais setores existentes, tornando cara e trabalhosa a missão de armazenar toda essa informação gerada nos dias de hoje.

Apesar da imensa quantidade de dados capturados, a velocidade (segundo "V") com que a captura é feita também pode ser considerada um fator chave para a manipulação das informações desejadas. Da mesma forma que os dados são gerados, eles podem se tornar obsoletos com o passar do tempo, por exemplo, um streaming de vídeo de um canal de notícias, além de dados complexos exigidos para uma boa transmissão, tais como a qualidade do som e áudio, a principal característica trabalhada é o imediatismo com que deve ser tratado, tornando assim uma característica iminente de problemas de *Big Data*.

A adequação de diversas formas de dados provenientes de diversas fontes também é uma das questões que chamam a atenção em problemas de *Big Data*. Esses dados podem ser provenientes de sensores, smartphones, vídeos, áudios, textos, redes sociais, data logs e de tantos outros dispositivos que estão sendo inseridos no mercado. *Big Data* trata da variedade de dados como uma mistura de diversos tipos de estruturas de dados, sejam elas estruturadas, no caso de dados coletados por sensores, semi-estruturadas, no caso de dados proveniente da troca de informação entre diferentes dispositivos através de XML e dados não estruturados, que podem ser mensagens de redes sociais, SMS, dentre outros, que, uma vez trabalhados de forma conjunta, podem trazer grande significância e complexidade às aplicações que irão utilizá-la [14].

O quarto "V" seria o Valor. Levando em consideração o potencial exploratório de uma base de dados, este fator, de fato, aponta uma característica fundamental de aplicações em *Big Data*, uma vez que a utilização de técnicas de análise de dados permite a criação de informações valiosas de negócios, aumentando o poder de in-

vestimento no setor. Por exemplo, para o consumidor de energia elétrica, uma gestão inteligente dos dados permitiria entender padrões quanto ao uso do produto, tornando uma ferramenta crucial para a personalização da prestação de serviço, podendo ser um fator chave para a economia de recursos e, possivelmente, uma melhora na distribuição, já que a companhia conhecerá o perfil de cada usuário [14].

O último “V” a ser bem aceito na conceituação de *Big Data* é o da Veracidade, uma vez que uma análise de dados feita de forma equivocada pode gerar uma interpretação falsa sobre o cenário, acarretando em prejuízos para o usuário final. Muitas vezes, os dados que compõem o banco de análise de uma empresa são integrados à outras bases ou necessitam de informações que não são visíveis em um primeiro momento. Investimentos em validação e interpretação desses dados são necessários para que o risco em utilizar esses dados seja cada vez menor [32].

3 Pré-Processamento de dados

Para que a extração de conhecimento a partir de dados seja feita de forma com que se obtenha relevância durante uma análise, as entradas devem passar por uma etapa de refinamento, dentro de um ambiente organizacional, a geração de dados provém de diversas fontes, sejam elas estruturadas ou não, cada fonte pode conter características próprias e seu tamanho pode ser um fator muito relevante no trabalho. Além dessa peculiaridade, os dados podem possuir inconsistências, ruídos, imperfeições, valores preenchidos de forma incorreta, valores duplicados ou até mesmo ausentes [6].

A chamada fase de pré-processamento é composta de atividades que vão desde a captação do dado em questão, organização e tratamento até a preparação, com a finalidade de tornar o dataset viável para execução de tarefas [9]. Como trabalhar com geração de conhecimento sobre a base de dados requer um nível mínimo elevado de confiabilidade, a base deve ser a mais completa possível. O que não acontece com o que encontramos na análise de um caso real, cuja baixa qualidade afeta diretamente na performance da aplicação [8].

O pré-processamento surge como um facilitador para minimizar, ou até eliminar os problemas que possam surgir e consiste em diferentes estratégias e técnicas que podem ser utilizadas de forma conjunta ou específica, porém, com o mesmo objetivo de tornar o processamento ainda mais eficiente. Não é comum, nem necessária, a aplicação de todas as técnicas possíveis, uma vez que a análise deve levar em conta o *know-how* do profissional responsável pela tarefa. Dentre as mais utilizadas, podemos citar:

- Integração
- Limpeza
- Transformação
- Redução

Cabe ao analista responsável pela execução do pré-processamento, definir quais etapas deverão ser aplicadas nos dados em questão, bem como, a ordem de

execução, uma vez que, sabendo da variedade de métodos disponíveis, o analista estará apto para decidir, conforme o estado que se encontra as suas fontes e sabendo qual informação deseja obter, quais métodos adotar de maneira mais eficaz.

3.1 Integração

O uso de bases de dados oriundas de diversas fontes é uma situação comum em diversos cenários de análise de dados, uma vez que o grupo de atributos desejado, raramente, estará centralizado em apenas um local, a não ser que a base seja pensada para esse fim. Mesmo que seja baseado em uma fonte de dados apenas, existe a possibilidade de agregar outras bases que podem existir dentro do mesmo cenário comum, caso não exista a centralização da informação.

Caso não haja uma estratégia para integrar essas bases, existe uma enorme possibilidade de ocorrer redundâncias e inconsistências que poderão acarretar em uma diminuição da precisão e velocidade de processos de análise, fazendo com que o conhecimento produzido não corresponda, em sua totalidade, com a realidade.

Em sua obra, Garcia [8] afirmam que, uma parte essencial do processo de integração é a definição de como cada instância será colocada na estrutura a ser montada, em uma estrutura comum que represente o mundo real dos dados. Com base no dito, podemos afirmar que a integração de dados não se trata apenas da união dos dados em si, mas sim do acréscimo de atributos que possam enriquecer o trabalho de exploração dos dados, evitando que, mais cedo ou mais tarde, surjam inconsistências decorrentes de dados duplicados ou não relevantes para a construção dessa nova base.

Como exemplo, podemos citar conjuntos de dados empresariais que possuem atributos que identificam um cliente que, com a integração, faz que com os objetos das diferentes bases sofram uma combinação a fim de unificar as informações em um único conjunto integrado, fazendo com que exista um valor único completo para cada objeto. A complexidade da operação de integração de base de dados pode se elevar a medida que atributos de mesma significância possuem identificações diferentes em suas bases originais. Uma maneira de identificar problemas de duplicidade de informações dentro de uma base de dados é através da utilização de análise de cor-

relação entre duas variáveis, onde podemos identificar se existe uma forte implicância de uma sobre a outra [1].

3.2 Limpeza

Durante o processo de coleta ou armazenamento em uma base de dados comum, existe fatores que podem acarretar em falhas que impossibilitam o registro correto do dado. Esse tipo de problema causa inconsistência na base fazendo com que a credibilidade do processamento seja prejudicada, revelando informações que não condizem com a realidade amostrada pelos dados.

Essas falhas provocam uma lacuna no preenchimento da base, fazendo com que não seja possível identificar uma informação com total precisão e surgem, por exemplo, quando um equipamento de coleta, por motivação alheia à vontade do analista, não consegue realizar seu trabalho de forma adequada; durante a fase de armazenamento, um erro de comunicação ou de disco impede que o dado coletado seja registrado de forma concisa ou até mesmo quando uma integração de bases não apresenta uma relação perfeita entre seus atributos, fazendo com que surjam dados “sujos” que podem ser descritos como aqueles que apresentam falhas no preenchimento, dados errôneos e representações fora do escopo que representa, comprometendo a qualidade dos dados, criando assim uma base de dados pobre (do inglês, *poor data*) [8].

A qualidade da base é um fator fundamental para a descoberta de novas informações relevantes a partir da análise de dados. A limpeza de dados é uma das etapas fundamentais para o processo, tendo em vista que, é nessa fase que descobrimos falhas que podem afetar a realização de operações em sua coleta. Os problemas que podem ser tratados nessa etapa são: dados ausentes, dados ruidosos, dados inconsistentes e dados duplicados [28].

De uma maneira global, a escolha de métodos para o tratamento dessas falhas varia conforme a natureza, objetivo e forma de coleta dos dados, fazendo com que não exista uma solução única e que realize todas as correções de forma satisfatória.

3.2.1 Dados ausentes

Dentre os problemas relacionados à poluição de uma base de dados, o problema mais encontrado é a ausência de atributos em um objeto. Esse problema na construção de uma base de dados pode ocorrer tanto em sistemas que coletam informações automaticamente, quanto em situações em que uma pessoa é a responsável pela coleta, ou seja, por falha humana ou operacional. Algumas possíveis causas que podem gerar esse problema são listadas por Faceli[6], a saber:

- Dados considerados irrelevantes no momento da coleta;
- Falta de conhecimento do valor do atributo durante um preenchimento;
- Falha humana no decorrer do preenchimento de alguma informação;
- Não obrigação de preenchimento de formulários;
- Impossibilidade de coleta de um determinado valor;
- Falhas no processo de coleta, transmissão e armazenamento de dados.

Muitas técnicas são empregadas para a resolução deste problema, no meio acadêmico algumas técnicas se destacam, dentre elas: exclusão de casos, preenchimento manual, preenchimento com variáveis globais, preenchimento com medidas estatísticas e métodos de mineração de dados [9].

A exclusão de casos é a alternativa mais simples de ser empregada. Ela consiste na retirada de registros que possuam casos de ausência de dados, essa alternativa deve ser utilizada com cautela, uma vez que a extração de muitos registros em uma base pode comprometer ainda mais a sua interpretação.

Já a técnica de preenchimento manual necessita do conhecimento prévio da informação a ser inserida. A inserção manual se torna uma tarefa enfadante, pois se trata de uma inserção feita completamente por meio de digitação, o que causa um gasto de tempo inviável ao se tratar de grandes bases de dados.

A técnica de preenchimento com variáveis globais propõe o uso de uma constante para a definição de valores, como, por exemplo, a imputação de palavras-chave como “desconhecido” ou “ausente”, porém, pouco contribui para o preenchimento da base de dados, pois os valores não utilizáveis deverão ser desconsiderados

e podem afetar outros procedimentos, como aprendizado de máquina, que pode considerar o valor como um dado real.

Expostas todas as formas de imputação manual, temos o preenchimento com medidas estatísticas como alternativa que requer menos esforço manual de quem a utiliza. Esse preenchimento se trata de utilização de cálculos numéricos com vista à obtenção de um valor coerente e dentro do contexto em que a variável está contida. O seu uso oferece, na maioria dos casos, um ganho sobre as demais técnicas citadas anteriormente, pois estima um valor factível de processamento e pode ser aplicado em grandes bases de dados, esses valores podem ser obtidos, por exemplo, através do cálculo da média ou mediana entres os valores do atributo.

Outra abordagem, através de mineração de dados, compreende o uso de métodos preditivos como forma de obter um valor mais real do que as demais técnicas, isso porque utiliza a exploração dos próprios dados para propor uma solução através de Redes Neurais, Estatística ou Árvores de Decisão. Goldschmidt [9] cita o fato de, mesmo que na fase de preprocessamento, esse tipo de técnica pode ser utilizado para preenchimento dos dados ausentes.

3.2.2 Dados ruidosos

Dados que apresentam problemas aleatórios durante a coleta e que podem apresentar valores distorcidos da realidade são chamados de dados ruidosos. Mesmo que o atributo apresente valores, um dado ruidoso pode prejudicar a análise de dados ou até mesmo o funcionamento de algoritmos de limpeza, por exemplo, algoritmos de descoberta de dados ausentes, baseados em operações de média, podem ter seu valor afetado por desvios oriundos de tal problema [8].

O termo ruidoso é decorrente da origem do problema que costuma ser associado a falhas na transmissão do dado. Para o tratamento de tais erros, podemos utilizar desde tratamento de sinal, processamento de imagem ou algoritmos de aprendizado de máquina.

3.2.3 Dados inconsistentes

A limpeza de dados inconsistentes se trata de identificar e eliminar valores que possuem divergência dentro da base de dados a que pertencem, onde requer

um conhecimento prévio da base para que seja aplicado com eficácia. Alguns problemas são de fácil identificação, como ilustra Tan [28], onde uma base de dados traz informações de endereço de um registro e percebe-se que dois registros possuem o mesmo número de código postal, porém, registram cidades diferentes, ou o peso de uma pessoa registrado na base como negativo. Outros não são tão fáceis de serem percebidos e necessitam de alguma informação adicional para seu reconhecimento e posterior tratamento.

3.2.4 Dados duplicados

Uma base de dados pode apresentar dados duplicados quando encontrado um valor ou um atributo idêntico, com alguns valores diferentes ou dados que podem ser obtidos de uma maneira simples.

Para exemplificar o conceito, podemos citar um registro que aparece duas ou mais vezes na base de dados, em que um algoritmo de aprendizado de máquina pode atribuir uma importância maior de sua representação real, o que levaria a destacar-se dentre as demais como uma tendência. Um atributo idêntico pode surgir quando a fusão de duas bases de dados é capaz de representar duas colunas com os mesmos valores. Essas duas situações também podem ocorrer, não necessariamente, com igualdade de todos os valores. Em um conjunto de atributos, podem ocorrer alguns valores que merecem uma atenção maior.

Por último, dados que podem ser obtidos através de estimativa simples, como, por exemplo, a idade de uma pessoa pode ser obtida com a data de seu nascimento decrescida da data de realização da tarefa, desde que, os dois estejam armazenados em uma mesma base [28].

3.3 Transformação

A transformação de dados pode ser considerada a etapa em que os dados são ajustados para que tenham um melhor desempenho durante a fase de análise. Uma vez que, dados podem ser obtidos de diversas fontes, um padrão entre eles se torna necessário para que possamos compreender os eventos com a mesma intensidade e

magnitude, garantindo, assim, a razoabilidade durante uma identificação proveniente dessa análise.

3.3.1 Transformação por normalização

O método de normalização de dados trabalha em uma forma de estabelecer um intervalo em que possa representar limites para a validação do valor presente no atributo. Definimos o intervalo entre 0 e 1 ou (-1) e 1. O objetivo do estabelecimento dessa escala tem vistas a evitar surgimento de tendências errôneas dentro da base de dados, pois, com o estabelecimento de limites, evita-se que valores discrepantes se estabeleçam dentre a busca de padrões [9].

Os métodos mais utilizados na realização de normalização de dados podem ser encontrados na literatura como: normalização linear, z-score e escala decimal [1].

3.3.1.1 Normalização *Min-Max*

Esse tipo de normalização, conhecida também por interpolação linear, consiste na utilização de limites, superior e inferior, como escala. Necessário o uso de todos os valores, não apenas amostras, esse método deve ser utilizado quando se tem a certeza de que o dado real se encontra em um domínio discreto [16]. Para a realização desse processo, temos a equação 3.1:

$$X_{mm}^* = \frac{X - \min(X)}{\text{range}(X)} = \frac{X - \min(X)}{\max(X) - \min(X)} \quad (3.1)$$

Onde X representa o valor a ser subtraído do valor mínimo definido ($\min(X)$) que deverá ser dividido pelo intervalo entre os valores mínimo e máximo.

Em um exemplo de análise de temperatura de uma determinada região, temos o cenário visto no Quadro 4:

Quadro 4 – Exemplo de dados estatísticos de temperatura de uma região K

Informações sobre temperatura da região K, em graus Celsius	
Valor Médio	27
Valor Mínimo	18
Valor Máximo	35
Intervalo	17

Para determinar o valor Min-Max dessa região, podemos analisar a temperatura de três dias, que apresenta os valores mínimo, médio e máximo, temos:

$$X_{mm}^* = \frac{X - \min(X)}{\text{range}(X)} = \frac{18 - 18}{17} = 0 \quad (3.2)$$

Na equação 3.2, temos um exemplo de quando o dado a ser normalizado coincide com o valor mínimo encontrado. Notamos que o resultado da equação é 0, o que significa que o dado está dentro do intervalo em seu limite mínimo.

$$X_{mm}^* = \frac{X - \min(X)}{\text{range}(X)} = \frac{27 - 18}{17} = 0,53 \quad (3.3)$$

Por sua vez, a equação 3.3 apresenta um valor que se encontra, aproximadamente, no centro do intervalo definido. Nota-se que o valor é próximo a $X_{mm} = 0,5$ que seria o valor médio representado por 28 graus.

$$X_{mm}^* = \frac{X - \min(X)}{\text{range}(X)} = \frac{35 - 18}{17} = 1 \quad (3.4)$$

A equação 3.4 define o limite máximo da equação, em que o valor de Min-Max equivale a 1.

Imaginemos uma situação em que um valor fora do intervalo proposto surge em nossa base, talvez em decorrência de uma falha no equipamento de coleta.

$$X_{mm}^* = \frac{X - \min(X)}{\text{range}(X)} = \frac{5 - 18}{17} = -0,76 \quad (3.5)$$

Como podemos observar na 3.5, o valor se apresenta fora de escala, conforme proposto. Esse valor tem alta possibilidade de ser uma falha a ser corrigida. Uma forma de correção seria, nesse sentido, a substituição do mesmo pela temperatura mínima, já que apresenta valores negativos.

3.3.1.2 Normalização por z-score

Esse tipo de normalização é utilizado quando não é possível determinar os limites mínimos e máximos dos valores do intervalo. Também chamada de normalização por desvio padrão, pois utiliza dessa medida estatística para definir o intervalo. Utiliza-se da média dos valores de um atributo e a dispersão dos valores sobre essa

média [9].

Larose et al. [16] apresentam a equação do Z-score como:

$$Z - score = \frac{X - media(X)}{desviopadrao(X)} \quad (3.6)$$

Com a fórmula apresentada em 3.6, podemos ter valores positivos e negativos em relação ao valor da média, como definidos nos exemplos, utilizando os dados do Quadro 4.

$$Z - score = \frac{X - media(X)}{desviopadrao(X)} = \frac{18 - 27}{8,03} = -1,12 \quad (3.7)$$

Vemos na equação 3.7 que, os valores negativos, são os valores que são menores que a média dos dados.

$$Z - score = \frac{X - media(X)}{desviopadrao(X)} = \frac{27 - 27}{8,03} = 0 \quad (3.8)$$

Como esperado, o cálculo médio apresenta o valor 0 na equação 3.8, pois é o responsável por estabelecer os padrões do modelo.

$$Z - score = \frac{X - media(X)}{desviopadrao(X)} = \frac{35 - 27}{8,03} = 0,99 \quad (3.9)$$

Por último, na equação 3.9, os valores positivos são os que são maiores que a média.

Utilizando o cálculo do Z-score, com os conceitos do cálculo do desvio padrão, podemos estabelecer quais serão os valores factíveis para nosso problema, tanto superior quanto inferior.

3.3.1.3 Normalização por escala decimal

A normalização por escala decimal, consiste no estabelecimento de uma faixa de valores que são compreendidos entre -1 e 1. Como apresentado por Larose et al. [16]:

$$X_{decimal}^* = \frac{X}{10^d} \quad (3.10)$$

Onde, na equação 3.10, X representa o valor e d representa o número de casas decimais que serão acrescentadas ao número. Larose et al. [16] afirmam que para estabelecer os valores mínimo e máximo do intervalo, temos de calcular a equação 3.10, utilizando a quantidade de dígitos do maior valor absoluto. Que, aplicando ao nosso exemplo de temperatura, apresentado no Quadro 4, temos:

$$Min : X_{decimal}^* = \frac{X}{10^{(d)}} = \frac{18}{10^2} = 0,18 \quad (3.11)$$

$$Max : X_{decimal}^* = \frac{X}{10^{(d)}} = \frac{25}{10^2} = 0,35 \quad (3.12)$$

Definindo assim o intervalo dos dados utilizáveis com os valores 0,18 (equação 3.11) para mínimo e 0,35 (equação 3.12) para máximo.

3.3.2 Agregação

Em uma base de dados padrão podemos encontrar diversos tipos de incoerências que se firmam em um cenário que necessita de operações sobre os dados como, por exemplo, a soma de valores totais de vendas de uma empresa. Esse tipo de ocorrência acontece, geralmente, pelo uso de SQL (Linguagem de Consulta Estruturada, do inglês Structured Query Language) para realização de operações, lógicas ou matemáticas, que visa evitar a duplicidade de informações dentro de um ambiente organizacional que utiliza banco de dados relacional. Portanto, o analista de dados deve desempenhar um papel mais crítico quanto à observação de quais dados utilizar, uma vez que, em uma base de dados relacional, é comum existir tabelas de dados auxiliares para que ocorra esse complemento proposto pelo uso de operações com base em SQL. Essa análise pode ser chamada de relevância do atributo em análise, pois cabe ao analista levantar os atributos que mais agregarão valor ao seu trabalho [20].

Uma variação que também pode ser destacada é a sumarização de dados, porém, diversas literaturas apontam que o mesmo também se trata de uma agregação. Em diferentes dimensões, a soma de atributos em uma base é determinante para seu significado, onde um valor diário agregado pode se tornar um valor mensal, adequando a necessidade do analista. Também utiliza-se o termo agregação para a junção de

elementos diferentes a uma base de dados [13].

3.3.3 Generalização

Seguindo a lógica da técnica de transformação de dados onde prima a adequação de uma base de dados para um formato que seja mais amigável a operações de mineração de dados e extração de conhecimentos, Jiang et al. [11] propõem que o processo de generalização consiste em duas etapas: uma etapa de mesclagem de dados semelhantes e uma segunda etapa em que transformam os dados generalizados em um segundo tipo de dado.

Dessa forma, temos uma adequação dos dados, quanto ao seu tipo e uma restrição de dados semelhantes, o que pode reduzir as chances de duplicidade de informação dentro de uma base pré-processada e, também, a observação de novos valores em dimensões e escalas diferentes.

A primeira fase propõe mesclar os dados que possuem características semelhantes suficientes para que se perceba que se tratam de uma mesma medida. Para realizar a mesclagem é necessário percorrer a base de dados relacional de origem, vasculhando seus relacionamentos em busca de peculiaridades que possam descaracterizar tal semelhança.

Já a segunda fase consiste em reduzir o número de valores distintos de um atributo discreto, isso irá reduzir também a complexidade de execução de um algoritmo de mineração de dados. O foco dessa fase seria na criação de categorizações aproximadas de valores discretos, como se fossem intervalos de valores, o que iria diminuir a quantidade de dados a serem processados, porém, observando o mínimo de alteração possível na tradução final do conhecimento que deseja ser alcançado.

3.4 Redução

O uso de dados de grandes volumes implica na dificuldade de se trabalhar com todos ao mesmo tempo, uma vez que o tempo e recursos de processamento são fatores que podem ser cruciais para a perda de eficácia e qualidade do resultado final. Pensando nessa situação, as técnicas de redução de dados trabalham em uma alternativa reduzida de se representar o mesmo conjunto de dados de forma com que

se mantenha o máximo possível das características da base. Seu objetivo é atingir o mesmo resultado (ou o mais próximo possível) com um conjunto de dados reduzido, economizando recursos e tempo de processamento, mantendo a mesma qualidade. Segundo Garcia et al. [8], geralmente, categoriza-se a técnica de redução de dados em três eixos principais, são eles: redução de dimensionalidade, redução por amostragem de dados e redução de cardinalidade.

3.4.1 Redução de dimensionalidade

Como afirmado anteriormente, problemas de análise de dados estão relacionados à sua densidade durante a fase de processamento, o que dificulta sua execução, esse problema também é encontrado na literatura como "Maldição da dimensionalidade", conforme vemos nos trabalhos de Facelo [6], Garcia [8] e Tan [28], entre outros. Faceli [6] afirmam que, percebendo um atributo como um ponto em um espaço dimensional, em que cada um é uma dimensão, temos um crescimento exponencial com a adição de novos atributos. Como tratamento de um problema de crescimento que possa prejudicar o trabalho do analista, assim como outras técnicas, a redução de dimensionalidade foca em combinar e eliminar os chamados dados irrelevantes, cuja existência é necessária para outro contexto sem ser o dado para o processamento.

Os benefícios trazidos pela redução de dimensionalidade é listada por Garcia [8]: ganho na velocidade de processamento, ganho de qualidade, ganho de performance na execução e simplificação de entendimento dos resultados. Essas reduções são feitas através de métodos estatísticos e se diferem quanto à linearidade dos dados, em que, dados lineares são tratados com técnicas de análise de fator e PCA (Análise de Componentes Principais, do inglês *Principal Components Analysis*), enquanto os dados não lineares podem utilizar os chamados LLE (Mapeamento Topológico Localmente Linear, do inglês *Locally Linear Embedding*) e ISOMAP (Mapeamento de atributo isométrico, do inglês *Isometric feature mapping*).

3.4.2 Redução por amostragem de dados

Segundo Garcia [8], a amostragem de dados tem como papel facilitar a análise e a modelagem de grande volumes de dados. Ele ainda cita as utilidades da amostragem em um algoritmo de mineração de dados, como:

- A redução do número de dados, que propõe diminuir o custo de armazenamento e processamento com dados que podem ser descartados sem alteração do resultado, possibilitando a agregação de novos dados que podem enriquecer a base de dados e auxiliar no balanceamento de dados;
- Ocorrência de eventos raros, análises preditivas são sensíveis a dados desbalanceados, algoritmos como árvore de decisão ou rede neural artificial, portanto, manter a base balanceada permite explorar os dados de forma consistente e trabalhar com resultados comuns e raros em suas reais proporções.

Técnicas de amostragem de dados são aplicadas para atender esses propósitos, Garcia [8] levantam diversas técnicas, principalmente da área estatística. Imaginemos uma base de dados D que possui N exemplos, quando aplicadas tais técnicas temos:

Amostra aleatória simples sem substituição de tamanho t : Nesse tipo de amostragem, é criada uma amostra t a partir de N de D , onde D ($s < N$) e a probabilidade de obter qualquer exemplo em D é $1/N$, ou seja, todos os exemplos possuem a mesma chance de ocorrência.

Amostra aleatória simples com substituição de tamanho t : Técnica semelhante a anterior, porém, quando retirada uma amostra, ela é novamente preenchida dentro do espaço total, permitindo que o item possa se repetir na escolha da amostra, ao contrário da técnica anterior.

Amostra balanceada : Essa técnica extrai amostras de uma variável alvo de acordo com critérios pré-definidos, por exemplo, uma amostra em que se deseja extrair dados diurnos e noturnos, em que a ocorrência de coleta de dados diurnos corresponde a 70% da base analisada, restando assim, 30% de dados coletados no período noturno.

Amostra por conglomerados : Já a técnica de amostra por conglomerados considera que os dados sejam retirados de seus determinados grupos ou conglomerados e que os mesmos possuem características heterogêneas, Garcia [8] ainda rememoram que o conteúdo extraído deverá ser menor que o grupo que foi retirado.

Amostra estratificada : Nessa técnica, a base de dados deve ser dividida nos chamados estratos razoavelmente homogêneos e sua amostra é coletada especificando a quantidade de itens que serão retirados dentre os mesmos. Após a coleta de cada estrato, constituímos, assim, a amostra estratificada.

3.4.3 Redução de cardinalidade

Este processo consiste em combinar duas ou mais categorias a fim de obter uma nova categoria de variáveis, sejam essas nominais e ordinais. Esse processo se torna importante, principalmente, para as variáveis nominais em que o número de categorias pode tornar o processo de análise problemático.

O processo de redução de cardinalidade é simples e, geralmente, realizado de forma manual, onde o analista observa a variação de valores dentro de um escopo específico e propõe uma junção de valores, baseado em sua semântica, buscando o mais próximo possível do resultado adequado com uma amostra de dados menor.

Garcia [8] cita o processo chamado de “binning” como uma técnica de aplicação de redução de cardinalidade, essa técnica consiste em converter variáveis contínuas em conjuntos de intervalos, processo semelhante ao que acontece na proposta de Jiang et al. [11], em sua segunda etapa na realização do processo de generalização.

3.4.4 Discretização

Um outro método difundido pela literatura é a discretização de dados, esse método de pré-processamento visa a auxiliar algoritmos que trabalham com dados categóricos. Seu objetivo principal é tornar um atributo contínuo em discreto, a fim de auxiliar os algoritmos de aprendizado de máquina a terem ganhos em processamento e velocidade. Essa técnica que prima a transformação de dados quantitativos em qualitativos, oferece ganhos em redução de dimensionalidade de um atributo em sua base e também uma simplificação de dados, trazendo mais precisão às análises. Garcia et al. [8] definem a discretização com o seguinte exemplo:

$$D = \{[d_0, d_1], [d_1, d_2], \dots, [d_{m-1}, d_m]\} \quad (3.13)$$

Em que, dado uma base de dados com N atributos, realizamos a discretização em m intervalos, onde d_0 possui o valor mínimo, d_m possui o valor máximo do intervalo discreto e $d_i < d_{i+1}$ para $i = \{0, 1, \dots, m - 1\}$. Com isso, temos D como o esquema de discretização dos atributos de N e podemos afirmar que existe um domínio $P = \{d_1, d_2, d_{m-1}\}$ que contém os valores de intervalos corte dos atributos em N .

3.5 Resumo do capítulo

Neste capítulo, expomos e discutimos acerca de como os dados são importantes e presentes no meio em que vivemos, seja em uma simples anotação do usuário a possibilidade de obtenção de métricas ainda não conhecidas que podem alavancar as vendas de uma empresa, diagnosticar doenças, melhorar a qualidade de vida em comunidade de uma determinada região, infinitas possibilidades que surgem através de análises de dados. Muitos dados já são coletados por empresas ou especialistas no ramo, seja para simples auditoria de seus trabalhos, como uma base histórica de suas atividades, o que torna ainda mais factível sua exploração para novas possibilidades.

Com essas observações, podemos notar que a qualidade dos dados podem ser fundamentais como base em uma tomada de decisão. Portanto, eliminar as mazelas encontradas em uma base de dados é de importância ímpar e notória, para que não sejam cometidos erros durante essa análise. Uma das formas de se garantir essa qualidade é através do pré-processamento.

Vimos que existe uma variância quanto à conceituação de cada etapa, uma vez que muitas técnicas se assemelham, causando assim uma ambiguidade em sua explanação, variando com os autores que encontramos, o que não significa erro, apenas uma caracterização diferente e/ou complementar para que a base seja ainda mais uniforme. Nem todas as fases são aplicáveis em todos os casos, uma vez que cada base possui características próprias que se diferem umas das outras, sendo assim, levando em consideração uma base de dados genérica, em que todos os passos foram seguidos, temos uma representação na figura 2.

Com esses passos aplicados, passa a existir a garantia de uma qualidade mínima suficiente para um bom andamento de um trabalho de análise de dados co-

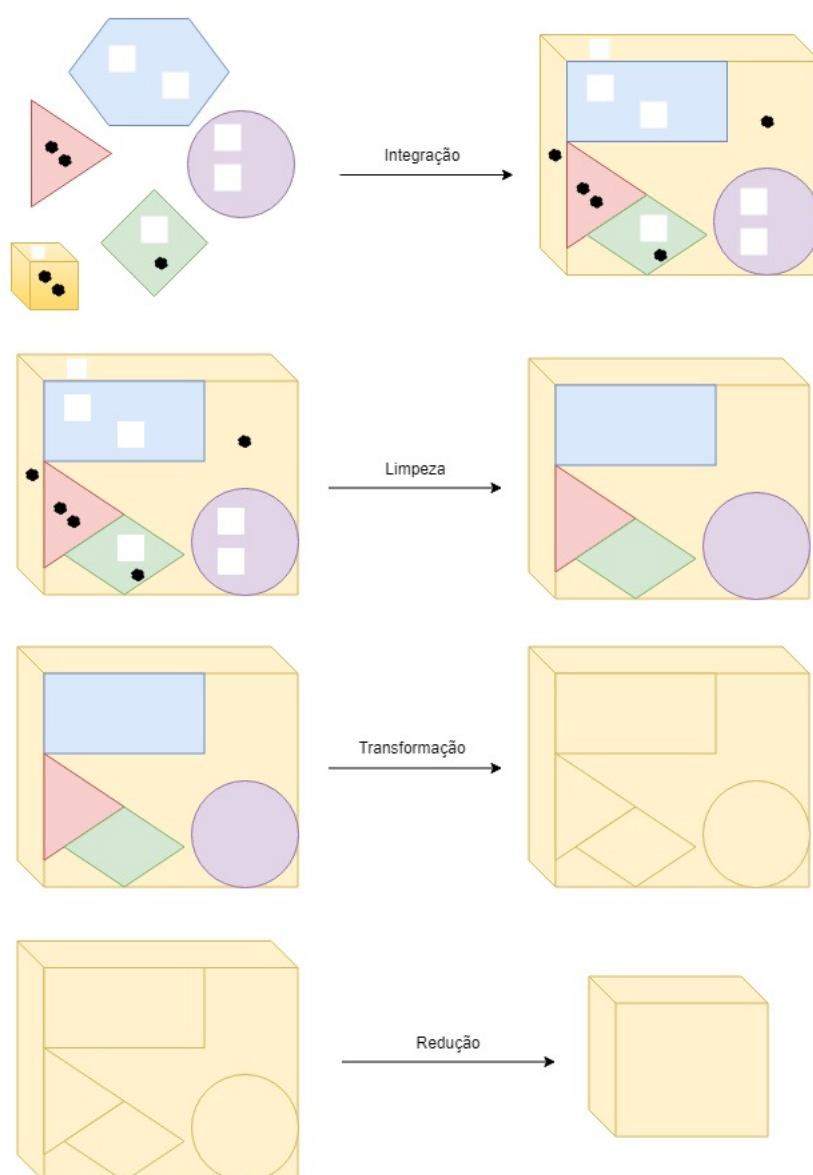


Figura 2 – Exemplificação de pré-processamento

erente, em que se possa extrair informações confiáveis, desde que, o algoritmo de mineração de dados seja adequado para satisfazer os objetivos do analista ou solicitante de tal trabalho.

4 Metodologia

Neste capítulo, abordaremos as idéias que foram empregadas no desenvolvimento desta pesquisa. Primeiramente, iremos discutir a base teórica que foi utilizada na seleção das variáveis para a montagem do banco de dados. Em seguida, serão apresentadas as bases de dados públicas que contêm as coletas da região observada e, por último, os métodos de pré-processamento que cada base necessitará para que possamos realizar a construção da proposta deste estudo.

4.1 Base teórica para a seleção das variáveis

Pesquisas apontam que o uso de dados para a obtenção de conhecimento com base em dados coletados buscam alternativas mais eficientes dos modelos encontrados atualmente. Dentre os estudos analisados, notamos que não é utilizado um projeto de padronização dessas bases de dados, uma vez que os trabalhos focam no cenário local de cada um dos pesquisadores, os quais levam em conta a realidade regionalizada do problema. Além do fator localidade, temos o fato de que não existe um método perfeito de uso desses dados, em que diversas características são capazes de influenciar no resultado esperado pelo pesquisador, levando, então, ao uso de diferentes variáveis para a resolução do mesmo problema.

Nessa perspectiva, notamos uma ocorrência sistêmica de cenário onde as pesquisas estão focadas em uma determinada região e restritas a dados específicos para atendimento da demanda, vezes disponibilizados publicamente, vezes necessária aquisição ou, até mesmo, fontes próprias de coleta para tal projeto.

Nossa proposta se situa na utilização das variáveis contidas em pesquisas recentes cujos resultados foram convincentes, com vistas a possibilitar sua aplicação no território brasileiro. Essa nova base gerada engloba os tipos de dados de cinco estudos, permitindo a utilização de estudos existentes e incentivando novas relações provenientes de estudos que possam vir a surgir.

Dentre as pesquisas analisadas estão Cervone(2017), Gagne(2017), Lima(2016), Urraca(2018) e Zhu(2018) que serão detalhadas nas seções a seguir.

4.1.1 Cervone, G. et al.

Em estudo, publicado em 2017, denominado "*Short-term photovoltaic power forecasting using Artificial Neural Networks and an Analog Ensemble*", Cervone et al. [4] utilizaram algoritmos de redes neurais artificiais (ANN, do inglês, *Artificial Neural Networks*) e algoritmos de conjuntos análogos (AnEn, do inglês, *Analog Ensemble*) para realizar previsões de energia gerada por painéis solares fotovoltaicos.

Conforme problema relatado anteriormente, a localização do seu trabalho foi focada em dados colhidos especificamente de três regiões da Itália, são elas: Lombardia, Calábria e Sicília, buscando representar os extremos das condições climáticas desse país. Os autores levantam a necessidade de explorar as características de cada lugar para aumentar a eficiência de produção de energia, mencionando, entre outros países, o Brasil como um dos locais com destaque potencial nesse tipo de estudo, mitigando os gastos em grandes centros urbanos ou como uma forma de resolver problemas oriundos de conexão com localidades remotas, principalmente, nos estados que compreendem a floresta amazônica.

Para o estudo, Cervone et al. [4] exploraram as variáveis de irradiância global horizontal, a cobertura de nuvens, a temperatura do ar a 2 metros da superfície, ângulo azimute e elevação. Dados esses coletados pelas próprias fazendas de energia de cada região, em formato estabelecido pelo NOAA (do inglês, *National Centers for environmental Information*) chamado de NWP (do inglês, *Numerical Weather Prediction*).

Os resultados foram baseados no chamado "*Short-term*" que compreende período de 3 dias para a previsão. Cervone et al. [4] sugerem resultados com alta taxa de confiabilidade, uma vez que, através do uso do cálculo do erro médio quadrático (RMSE, do inglês, *Root Mean Square Error*) obtiveram um índice compreendido entre 7,07% e 9,07% de informações errôneas em todas as variações propostas pelo uso de ANN e AnEn.

4.1.2 Gagne II, D. J. et al.

Gagne et al. [7] publicaram um artigo denominado "*Evaluation of statistical learning configurations for gridded solar irradiance forecasting*" cuja proposta era avaliar diferentes modelos estatísticos e configurações que permitissem a realização de

predições a partir de dados ambientais fornecidos pela Oklahoma Mesonet, centrando a pesquisa no estado de mesmo nome localizado nos EUA.

Nesse estudo foram testadas onze técnicas baseadas em uso de dados para exploração de informações e conhecimento sobre a base, sendo reportado pelos autores que, dentre todas elas, houve uma taxa de erro máximo de 5%. Para a obtenção desse fator, observa-se o uso do método de Média Absoluta (MAE). As variáveis exploradas foram: horas de sol, cobertura de nuvens, azimute solar, zênite solar, elevação solar, irradiância solar de céu limpo e o índice Kt.

Suas predições foram baseadas em curto prazo, especificamente a cada 5 minutos, porém, para a obtenção de tais predições, os autores não utilizaram nenhuma medição de dias nublados e/ou dados fora da curva, o que pode impactar na qualidade da base. Os autores não mencionam, em seu artigo, a quantidade de dados descartados para a realização do trabalho.

4.1.3 Urraca, R. et al.

Com foco no leste da Espanha, Urraca et al. [30] realizaram pesquisas com ênfase na predição de irradiação solar, utilizando bancos de dados da *Servicio Integral de Asesoramiento al Regante* (SIAR), cuja coleta tem resolução de 30 minutos e, adicionalmente, utilizaram o banco de dados da *Aerosol Robotic Network* (AERONET).

As variáveis extraídas para a utilização foram: temperatura (em graus Celsius), umidade relativa do ar (em porcentagem), velocidade do vento (em metros por segundo), precipitação (em litros) e informações sobre claridade (Kt). O período de predição, nessa pesquisa, foi de curtíssimo prazo, especificamente, predições de uma hora.

A taxa de erros na predição utilizando os métodos de RMSE e MAE em todos os modelos e variações apresentados variou entre 12,93% e 16,29%. Porém, uma particularidade dessa pesquisa vale a pena ser citada, houve demasiados descartes de coletas, uma vez que empregou-se uma política de descartar o dia do ano em que houvesse perda de 25% dos dados, repetindo na desconsideração de 35 dias para o período observado, que foi durante todo o ano de 2013, o que daria, aproximadamente, 9,25% da amostra total.

4.1.4 Lima, F. J. L. et al.

O artigo intitulado *"Forecast for surface solar irradiance at the Brazilian Northeastern region using NWP model and artificial neural networks"* publicado em 2016, na revista *Renewable Energy*, por Lima et al. [18], cita um estudo acerca do território brasileiro, explorando apenas a base de dados do Instituto Nacional de Meteorologia (INMET), durante os anos de 2009 e 2011, utilizando dados de diversas estações de coletas espalhadas pela região nordeste.

O foco desse trabalho foi a produção de previsões de irradiação solar para a região, utilizando a resolução de curto prazo, ou seja, 24 horas, apresentando estudos que variam entre 9,58% a 43,85% e utilizando MSE para o período de tempo seco e 21,06% a 47,46% para os períodos chuvosos. Para o método RMSE obtiveram entre 15,07% a 58,85% para o tempo seco e 30,25% a 66,59% para o tempo úmido. Nesse estudo, foram utilizadas as variáveis umidade relativa do ar, temperatura, velocidade do vento, cobertura de nuvens e precipitação.

Uma observação a fazer com relação a esse trabalho é o fato de o banco de dados não possuir todas as informações necessárias, fazendo com que seja apresentada uma variação muito grande entre os erros mínimos e máximos e, também, falhas em medidas, como a presença do termo "NaN" em alguns registros da tabela de resultados.

4.1.5 Zhu, G. et al.

No ano de 2018, Zhu et al. [35] propuseram um sistema de previsão de curto prazo (um dia) para monitoramento da produção em ambientes de uma universidade localizada em Hong Kong, que utilizou do banco de dados do observatório da própria cidade para trabalhar os modelos desenvolvidos.

As variáveis escolhidas para análise foram: umidade relativa do ar, temperatura, precipitação, velocidade do vento, irradiação solar global, cobertura de nuvens e índice de claridade. Dentre todos os métodos e variações trabalhadas, o cálculo de erros de previsão se deu através da média absoluta (MAE), sendo testado em dois ambientes diferentes, resultando em variações para o primeiro em: 5,57% a 29,41% e para o segundo ambiente: 7,34% e 24,33%.

4.1.6 Resumo da seção

Essa seção trata de pesquisas que utilizam dados para realização de predições de diferentes formas, sendo necessárias diversas variáveis que não, necessariamente, são as mesmas em todas elas. Também é possível analisar falhas decorridas da ausência ou falta de confiabilidade dos dados, fazendo com que a porcentagem de erros seja maior ou até mesmo, como no caso de Lima et al. [18], onde o relatório final apresentou essa falha em sua base de análise.

A metodologia proposta engloba o maior número de soluções possível para o uso de dados na região que analisarmos. Assim sendo, a seleção de variáveis que serão utilizadas passou pelo levantamento de todas essas pesquisas, como passo inicial de um projeto, a fim de compor uma base completa e eficiente para servir de base para novos estudos e soluções.

As variáveis selecionadas podem ser verificadas no Quadro 5, fazendo uma relação com os autores que a propuseram.

Quadro 5 – Relação de variáveis

Relação de variáveis por utilização						
Variável traduzida	Variável	(CERVONE, 2017)	(GAGNE-2017)	(LIMA, 2016)	(URRACA-2016)	(ZHU, 2018)
air_temperature	Temperatura do ar	x				
azimuth	Ângulo Azimute	x	x			
clarity_index	Índice de claridade (Kt)		x		x	x
clear_sky_global_irradiance	Irradiação Global de céu limpo		x			
cloud_cover	Área Coberta por nuvens	x	x	x		x
elevation	Elevação	x	x			
global_irradiance	Irradiância solar global	x				x
humidity	Umidade relativa do ar			x	x	x
latitude	Latitude					
longitude	Longitude					
precipitation	Precipitação pluvial			x	x	
prediction_hour	Horário da predição					
solar_day_duration	Duração do dia solar		x			
temperature	Temperatura			x	x	x
wind_speed	Velocidade do vento			x	x	x
zenith	Ângulo Zenith		x			

4.2 Fontes de dados ambientais para extração

Após a seleção de variáveis, a seleção das fontes de dados é um fator determinante para a qualidade de nossa base. Bases públicas provenientes de fontes governamentais, ditas oficiais, são um dos caminhos para obtenção de um dado real

e de qualidade adequada, porém, mesmo nesse cenário considerado ideal, podemos encontrar falhas que influenciam no desenvolvimento de soluções.

A base do projeto SONDA é a responsável por gerar o Atlas Solarimétricos do Brasil, mas, em análises simples, podemos identificar falhas comprometedoras, como podemos ver na figura 3.

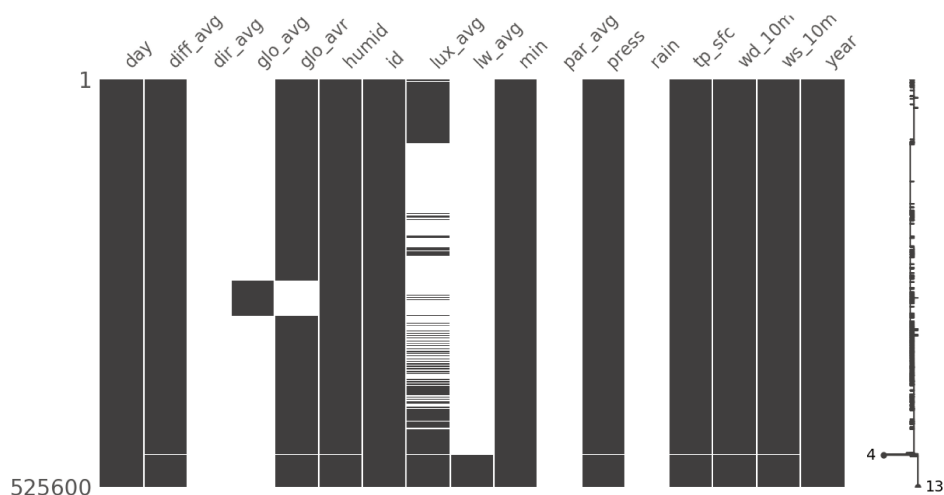


Figura 3 – Gráfico de ausência de valores

É possível perceber a partir da figura 3, que diversos valores são faltantes nas coletas da base SONDA, especificamente nessa análise, para a cidade de São Luís do Maranhão, no ano de 2015. A interpretação do gráfico segue o raciocínio de que a cada dado ausente não é impressa a linha correspondente de tal valor.

Isso não significa o descarte da base pois os dados contidos podem refletir a realidade. Porém, necessitamos de uma atenção maior aos dados coletados. Os erros contidos na base, geralmente, se tratam de falhas de comunicação entre a estação de coleta e o armazenamento, uma descalibragem ou um defeito de um equipamento, não sendo, necessariamente, fatores contínuos e duradouros; os erros podem ser corrigidos, e, tão logo aconteça, os valores adquirem novamente a qualidade desejada.

Portanto, bases auxiliares podem suprir a falha ou a ausência de confiabilidade de uma outra, ou até mesmo ajustar detalhes que por precisão ou conversão de medidas possam surgir. Dessa forma, em nosso projeto, utilizaremos as seguintes bases: SONDA (como já especificado anteriormente), SINDA, NASA, BDMEP diário e BDMEP horário, que serão melhor detalhadas a seguir.

4.2.1 SINDA

Como podemos consultar no seu site, o projeto SINDA (Sistema Integrado de Dados Ambientais) é o responsável por compilar as informações obtidas através do sistema de satélites que carregam o sistema DCS (transponder de coleta de dados) e da rede de plataformas de coletas de dados (PCD's) instaladas no solo brasileiro.

Esse projeto é mantido pelo INPE (Instituto Nacional de Pesquisas Espaciais) e têm por objetivo atender as demandas de dados ambientais em todas as instituições do país, dentre elas, a ANA (Agência Nacional de Água), Petrobrás - Oceanografia, ANEEL (Agência Nacional de Energia Elétrica), órgãos de Defesa Civil, entre outros ¹.

A base utiliza regiões onde possui estações de coletas como parâmetro de consulta, utilizando-se das informações de latitude e longitude, as variáveis disponibilizadas estão listadas no quadro 6.

Quadro 6 – Variáveis da base SINDA

Variáveis
Direção do vento (oNV)
Pluviometria (mm)
Pressão Atmosférica (mB)
Radiação Solar Global (W/m2)
Temperatura do ar (°C)
Temperatura máxima (°C)
Temperatura mínima (°C)
Umidade Absoluta do Ar (\%)
Umidade Relativa do Ar (\%)
Velocidade do Vento (m/s)
Velocidade Máxima do Vento (m/s)

4.2.2 POWER

A base de dados POWER (do inglês, *Prediction of Worldwide Energy Resource*) é um projeto estabelecido pela NASA (do inglês, *National Aeronautics and Space Administration*), agência do Governo Federal dos Estados Unidos responsável por pesquisas e desenvolvimento de tecnologias e programas de exploração espacial. Seu objetivo é entender e modelar a estrutura climática global e suas variações. O

¹ Informação extraída do site do SINDA, disponível em <http://sinda.crn.inpe.br/PCD/SITE/novo/site/sobre.php>. Acesso em 01/2018.

projeto foi pensado para fornecer à população, dados a partir de sistemas de satélite focado em três comunidades: Energia renovável, edifícios sustentáveis e agroclimatologia ².

O projeto disponibiliza dados de todo o globo, baseado na localização do local. Os dados são extraídos de imagens adquiridas pelos satélites próprios e complementada com coletas de dados parciais de alguns outros projetos cujo acesso é público. A modelagem de dados voltada a energia renovável dispõe das variáveis levantadas no quadro 7

Quadro 7 – Variáveis da base POWER

Variáveis
Meteorologia (Umidade e outro):
Precipitação (mm day-1) Umidade Específica a 2 metros (kg kg-1) Umidade Relativa a 2 metros (%) Pressão atmosférica (kPa)
Meteorologia (Temperaturas)
Variação de temperatura a 2 metros (°C) Temperatura da pele da terra (°C) Ponto de orvalho/geada a 2 metros (°C) Temperatura máxima a 2 metros (°C) Temperatura mínima a 2 metros (°C) Temperatura a 2 metros (°C)
Meteorologia (Vento)
Velocidade mínima do vento a 10 e 50 metros (m/s) Velocidade máxima do vento a 10 e 50 metros (m/s) Variação da velocidade do vento a 10 metros (m/s) Direção do vento a 10 e 50 metros (Convenção meteorológica)(Graus) Velocidade do vento a 10 e 50 metros (m/s)
Aplicações térmicas-solares
Índice de claridade atmosférica com céu limpo Índice de claridade atmosférica Fluxo radiativo infravermelho para baixo (onda longa) (kW-hr/m ² /day) Índice de insolação de céu claro em superfície horizontal (kW-hr/m ² /day) Insolação incidente \textit{All-Sky} em superfície horizontal (kW-hr/m ² /day)
Solar
Insolação incidente \textit{All-Sky} em uma superfície horizontal (kW-hr/m ² /day) Insolação incidente de céu limpo em superfície horizontal (kW-hr/m ² /day)
Painéis solares inclinados
Índice de claridade

² Informação extraída do site NASA POWER, disponível em <https://power.larc.nasa.gov/>. Acesso em 01/2018.

4.2.3 BDMEP

O governo brasileiro dispõe de um instituto responsável por medições meteorológicas, conhecido como INMET, representando o país junto à Organização Meteorológica Mundial (OMM). Dentro dessa organização, o INMET tem por obrigação compartilhar as coletas relacionados a meteorologia da América do Sul, além de sediar o GISC (Centro de Sistema de Informação Mundial). O órgão possui estações de coletas conhecidas como radiosonda, estações meteorológicas de superfície manuais e automáticas. Essas informações são disponibilizadas de forma gratuita no site da instituição, com o objetivo de auxiliar a tomada de decisão de diversos setores ³.

O INMET possui registros, em forma digital, coletados desde 1961 e em processo de incorporação de outros dados históricos registrados por documentos da época do império. O sistema responsável pela disponibilização das informações é chamado de BDMEP (Banco de Dados Meteorológicos para Ensino e Pesquisa) e dispõe de diferentes formas de compartilhamento, mensal, diária e horária. As variáveis são diferentes para cada tipo de métrica, conforme quadro 8.

Quadro 8 – Variáveis da base BDMEP

Quadro de variáveis do BDMEP		
Mensal	Diária	Horária
Direção do Vento Predominante (tabela)	Evaporação do Piche(mm)	Direção do Vento
Evaporação do Piche (mm)	Insolação(horas)	Nebulosidade
Evapotranspiração Potencial BH (mm)	Precipitação(mm)	Pressão Atm nível Estação(mbar)
Evapotranspiração Real BH (mm)	Temp Máxima(°C)	Temp Bulbo Seco(°C)
Insolação Total (hs)	Temp Mínima(°C)	Temp Bulbo Úmido(°C)
Nebulosidade Média (décimos)	Temperatura Compensada Média(°C)	Umidade Relativa(%)
Número de Dias com Precipitação (qtd)	Umidade Relativa Média(%)	Velocidade do Vento(m/s)
Precipitação Total (mm)	Velocidade Vento Média(mps)	
Pressão Atm Média (mbar)		
Pressão Atm nível Mar Média (mbar)		
Temp Compensada Média(°C)		
Temp Máxima Média(°C)		
Temp Mínima Média(°C)		
Umidade Relativa Média (%)		
Velocidade do Vento Máxima Média (mps)		
Velocidade do Vento Média (mps)		
Visibilidade Média (%)		

³ Informação extraída do site do INMET, disponível em <http://www.inmet.gov.br/portal/index.php>. Acesso em 01/2018.

4.3 Uso de técnicas de pré-processamento

As bases de dados seguem a afirmação de Faceli [6] afirma ao dizer que cada fonte de dados contém características próprias que dificultam uma simples soma de conteúdo. Sendo assim um tratamento deve ser feito afim de adequar as bases para uma forma que permita a ocorrência de uma agregação com aumento de significância e robustez, não apenas isso, mas também o uso de técnicas de pré-processamento para tratar casos que correspondam a um desvio de informações ou a ausência de valores de uma coleta que não tenha tido um comportamento esperado.

O primeiro passo é localizar as variáveis selecionadas dentro das bases públicas citadas a fim de saber como utilizar corretamente as ferramentas propostas. Nesse sentido, temos o quadro 9.

Quadro 9 – Relação de variáveis e bases de dados

Tabela de variáveis por base de dados						
Variável traduzida	Variável	SONDA	SINDA	NASA	BDMEP horário	BDMEP diário
air_temperature	Temperatura do ar		x			
azimuth	Ângulo Azimute	Calculado				
clarity_index	Índice de claridade (Kt)			x		
clear_sky_global_irradiance	Irradiação Global de céu limpo			x		
cloud_cover	Área Coberta por nuvens				x	
elevation	Elevação	Calculado				
global_irradiance	Irradiância solar global	x	x	x		
humidity	Umidade relativa do ar	x	x	x	x	x
latitude	Latitude	Calculado				
longitude	Longitude	Calculado				
precipitation	Precipitação pluvial	x	x	x		x
solar_day_duration	Duração do dia solar	Calculado				
temperature	Temperatura	x	x	x	x	x
wind_speed	Velocidade do vento	x	x	x	x	x
zenith	Ângulo Zenith	Calculado				

As variáveis marcadas como "Calculado" correspondem a informações imutáveis, caso das variáveis latitude e longitude, e de variáveis obtidas através de cálculos que podem ser realizados de forma independente aos dados das demais variáveis.

5 Resultados

Os resultados obtidos nesse trabalho serão tratados em duas seções. Primeiramente, serão apresentadas análises das fontes de dados, inseridas na metodologia, em seu estado original, identificando as características das mesmas. Na segunda subseção, será demonstrado a construção das variáveis calculadas dentre as selecionadas e, após, será aplicado métodos de pré-processamento, com vistas a adequação das bases para a agregação e construção da nova proposta.

5.1 As incongruências das bases de dados ambientais

As análises sobre bases de dados são fatores indispensáveis na hora de obter conhecimento sobre as mesmas. Uma vez iterado sobre como cada base trata as informações ali armazenadas, podemos passar para a próxima fase que a de realização do pré-processamento, cuja finalidade principal é disponibilizar um conjunto com características semelhantes a fim de obter uma agregação de seus valores. Executamos assim as análises realizadas em cada base citada até então.

Para a análise, utilizou-se dados de São Luís, cidade do estado de Maranhão, localizada ao norte do Brasil, com latitude $02^{\circ}31'48''S$ e longitude $44^{\circ}18'10''O$. O ano de 2015 foi escolhido para observação, uma vez que é o último ano com medidas contabilizadas como confiáveis pela base de referência SONDA.

5.1.1 Base de dados SONDA

Observando as informações especificadas anteriormente, podemos levantar os dados da base SONDA conforme proposta na seleção de variáveis, tendo assim expostas as seguintes variáveis: Irradiação solar global, umidade relativa do ar, precipitação, temperatura e velocidade do vento.

O gráfico 4 apresenta a ocorrência de coleta de valores para o ano de 2015 em São Luís - MA.

Nesse gráfico podemos extrair informações que são relevantes para a escolha de quais métodos de pré-processamento devemos utilizar. As análises apontam

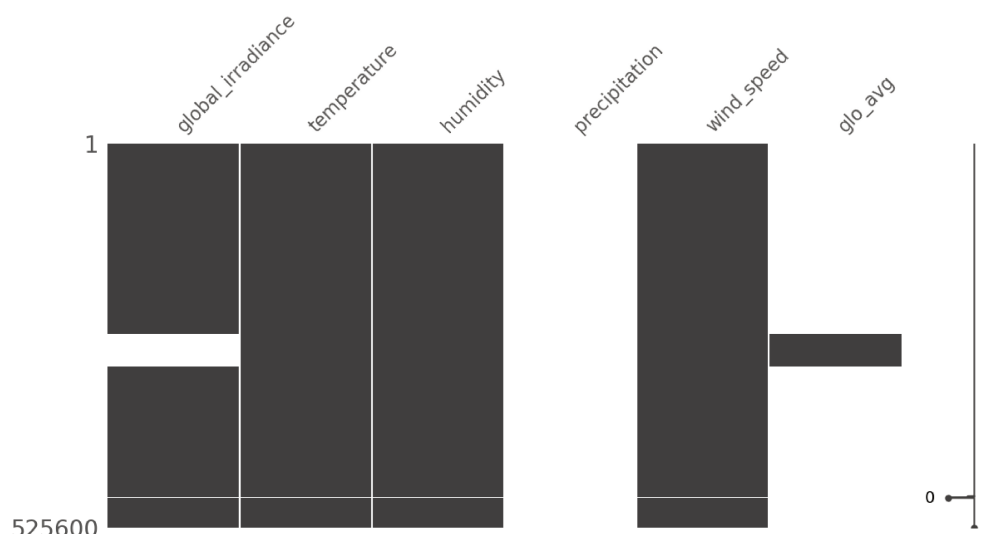


Figura 4 – Gráfico de dados ausentes da base SONDA

que foram executadas 525.600 coletas em um ano, o que representa a frequência de coleta em minutos, dentre essas coletas, percebe-se que existem dados ausentes em todas as variáveis analisadas, sendo que a variável *precipitation* não apresenta coletas para o ano.

Antes de realizar as análises, é válido relembrar seu conceito. O gráfico do tipo *boxplot* necessita das seguintes informações sobre a variável analisada: valor mínimo, valor do primeiro quartil (Q1), a mediana, valor do terceiro quartil (Q3) e valor máximo. Essas variáveis são necessárias para construir o gráfico, desenha-se um retângulo alinhado vertical ou horizontalmente com duas semiretas, em ambos lados opostos do retângulo. A altura do retângulo é definida pelos quartis 1 e 3, a mediana é apresentada dentro do retângulo representando o centro da distribuição, os valores máximo e mínimo ligadas ao retângulo representam os extremos dos dados expostos. Com essas medidas podemos obter informações acerca da localização, dispersão, assimetria, comprimento da cada e valores discrepantes (*outliers*).

Os valores coletados mostram a ausência de 45.838 coletas, os valores máximo e mínimo são 1.446 e -524, respectivamente. O quartil 1 tem valor de -1,33 e o quartil 3 255,5 e a mediana dos valores ficou em 0,56.

Ao analisar a página da base SONDA, no item Validação de dados¹, temos questões relacionadas à qualidade dos dados para cada variável. Pesquisando

¹ Informação extraída do site do projeto SONDA, disponível em <http://sonda.ccst.inpe.br/infos/validacao.html>. Acesso em 06/2018.

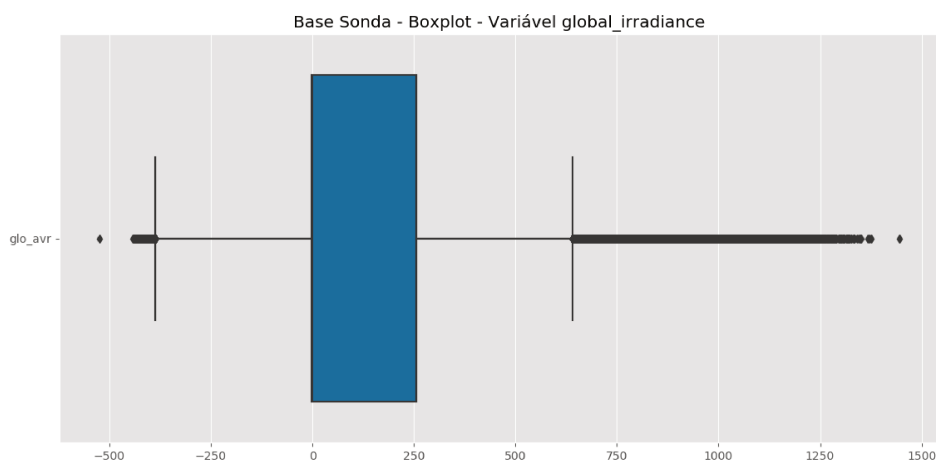


Figura 5 – Boxplot da variável *global irradiance* na base SONDA

a variável *global_irradiance* temos a informação de que seu valor mínimo deve ser de $-4W/m^2$, ou seja, bem acima dos -524 encontrado na análise. Podendo ser constatado um erro no preenchimento da base, pode-se considerar que os valores errôneos afetam a interpretação do dado considerado *outlier*.

A próxima variável é a umidade relativa do ar, gráfico 6.

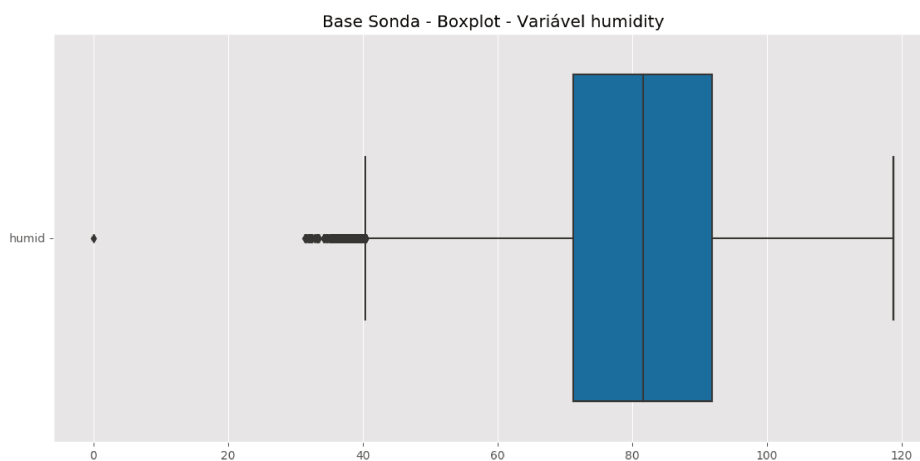


Figura 6 – Boxplot da variável *humidity* na base SONDA

O gráfico foi construído com base em 524.404 coletas de dados, apresentando então ausência de 1.196 coletas. Não obstante à análise da variável *global_irradiance*, temos valores discrepantes apontados como máximo, uma vez que por se tratar de dados percentuais, o gráfico aponta a existência de valores acima de 100% sendo assim

necessário um tratamento de tais dados, que, por sua vez, afeta as informações de *outliers*.

As coletas sobre precipitação não necessitarão de análise, uma vez que nenhuma coleta foi realizada para esse período na região de São Luís - MA. Já as coletas sobre temperatura são representadas no gráfico 7.

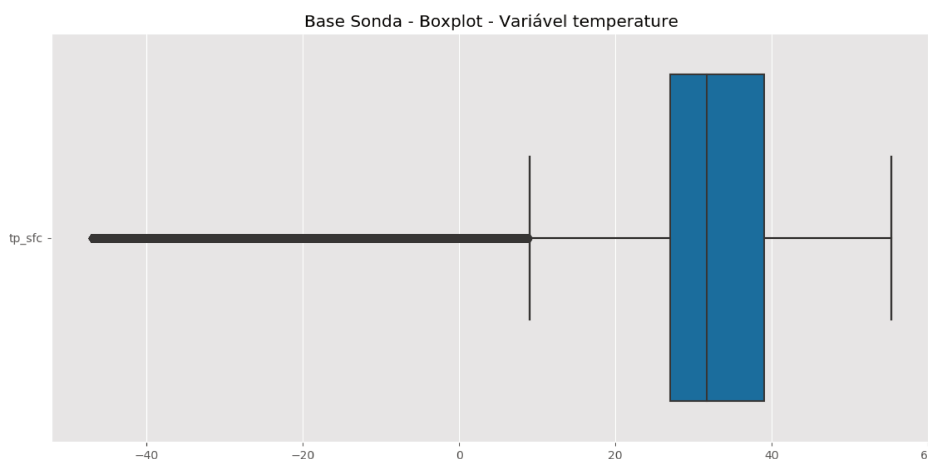


Figura 7 – Boxplot da variável *temperature* na base SONDA

Novamente temos coletas ausentes nessa variável, precisamente 1.196. Por se tratar de variável climatológica, podemos utilizar os dados históricos levantados pelo INMET, autoridade na área meteorológica no país, que apontam valores mínimos e máximos absolutos para identificar incongruências das coletas, onde temos que:

- Para valores mínimos, temos informações sobre a mínima histórica ocorrida em março de 1987, cujo valor registrado fora de $17,9^{\circ}C$;
- Para valores máximos, temos informações sobre a máxima histórica ocorrida em outubro de 1997, cujo valor registrado fora de $37,2^{\circ}C$.

Já os dados apresentados na figura 7, temos valores mínimos de $-47^{\circ}C$ e máximos de $55.38^{\circ}C$ o que destoa dos valores históricos, sendo assim então, também, dados errôneos.

Por último, temos as informações sobre velocidade do vento, cuja representação se dá pelo gráfico 8.

Seguindo a análise, encontramos a ausência de 1.196 coletas, assim como os anteriores. As informações de validação disponíveis na página da base apontam

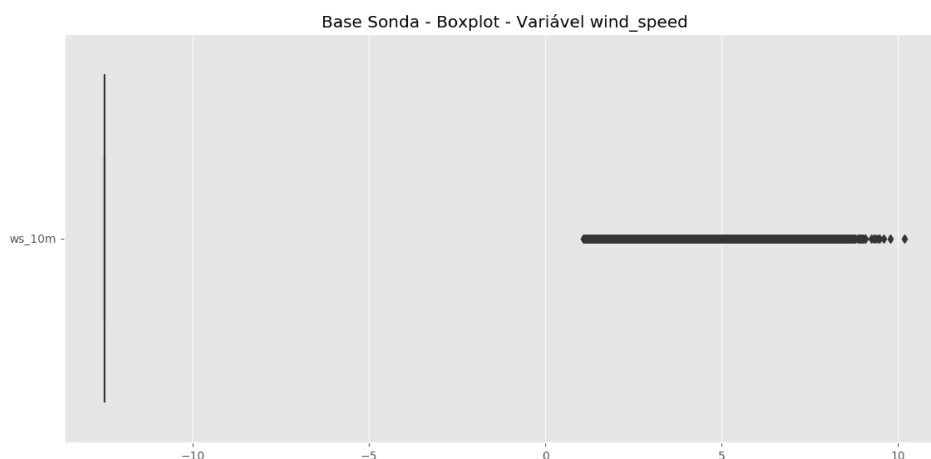


Figura 8 – Boxplot da variável *wind_speed* na base SONDA

os valores validos para essa variável, que são: $0m/s$ de valor mínimo e $25m/s$ de valor máximo. Novamente encontramos valores discrepantes, uma vez que os dados apontados pelo gráfico apresenta $-12.5m/s$ de valor mínimo, o que afeta consideravelmente a montagem do mesmo.

5.1.2 Base de dados SINDA

Levando em consideração a mesma ambientação para a análise da base SINDA, que é de informações de São Luís - MA no ano de 2015. A base SINDA apresenta medidas com resolução de 3 horas de intervalo entre as coletas. A base SINDA não disponibiliza publicamente as características de validação de seus dados, sendo assim, necessário a utilização de fontes de validações já utilizadas pro mesmo local e condições. Como as utilizadas na base SONDA, por exemplo.

No gráfico apresentado na figura 38 podemos analisar a ausência de valores nas coletas de suas respectivas variáveis. Para ser específico, do total de 2.920 coletas possíveis, apenas 2.684 coletas estão representadas, o que impacta na qualidade da base, principalmente nos dados sobre temperatura e velocidade do vento.

Uma característica notável sobre a base SINDA é o fato de ela trazer valores de temperatura máximo e mínimo para cada coleta, fazendo com que seja necessário uma adequação para obter um único valor.

No gráfico apresentado na figura 10 temos o boxplot da variável cuja con-

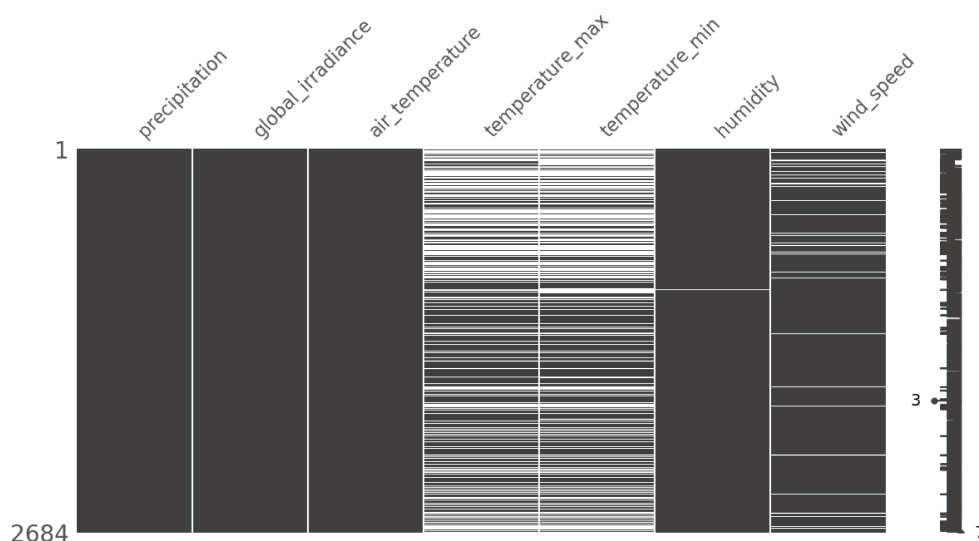


Figura 9 – Gráfico de dados ausentes da base SINDA

tabilização de coletas fora a destacada pelo gráfico de ausência. Os valores mínimo e máximo ficaram entre 0 e 51.10, bem diferente do apontado na base SONDA. Apotando inclusive, dados fora da curva de forma excessiva.

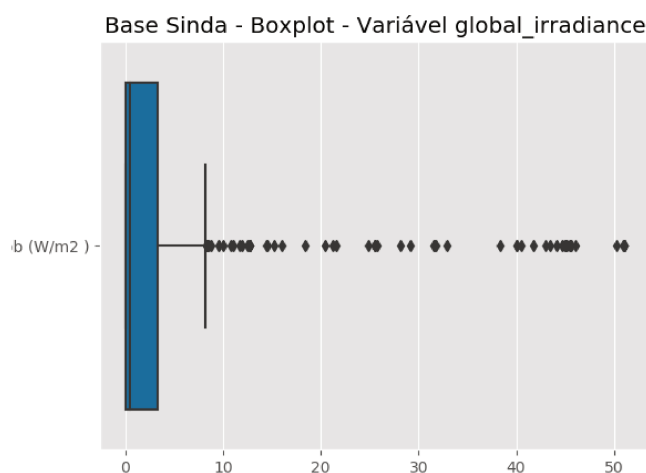


Figura 10 – Boxplot da variável *global_irradiance* na base SINDA

Já para a variável *humidity*, o gráfico apresentado na figura 11 mostra um equilíbrio maior de informações. Os valores mínimo e máximo ficaram entre 16% e 100%, tendo a mediana fixada em 82% revelando poucos valores fora da curva.

A variável *precipitation*, representada no gráfico da figura 12, apresenta uma formação não muito comum em relação aos seus dados, uma vez que a caixa do boxplot ficou muito próximo do valor 0, mostrando que existem valores zerados em

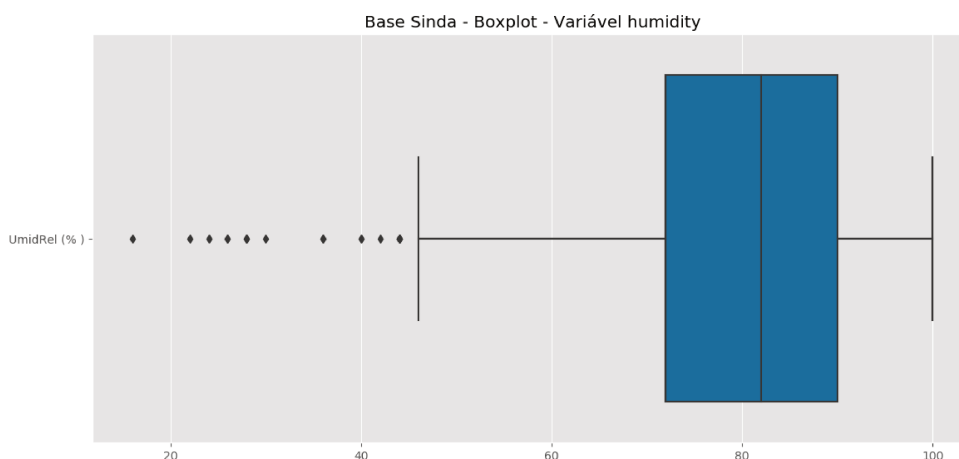


Figura 11 – Boxplot da variável *humidity* na base SINDA

demasia, forçando um valor muito abaixo para a mediana da mesma, além disso, o fato de existirem dados *outliers*, pode significar um erro durante a coleta.

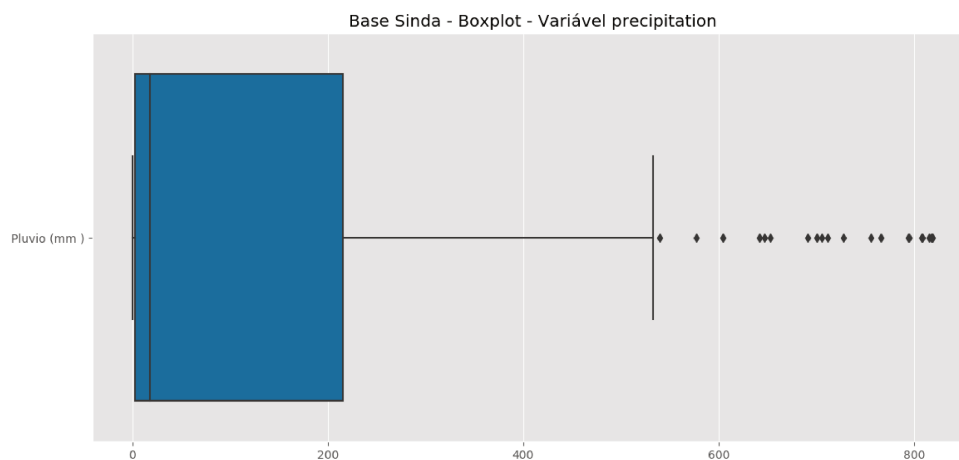


Figura 12 – Boxplot da variável *precipitation* na base SINDA

A próxima variável selecionada dessa base é a *wind_speed*, apresenta muitos valores 0, sendo o que desequilibrou a o gráfico apresentado na figura 13, os valores mínimo e mediana estão zerados, sendo o valor máximo representado por $51.1m/s$, valores que se comparados ao da base SONDA, apresenta discrepância tanto na quantidade de valores não coletados, quanto no valor máximo.

Os valores de temperatura, necessitam de um tratamento antes de serem analisadas, ficando assim para a próxima subseção, lembrando que os limites histó-

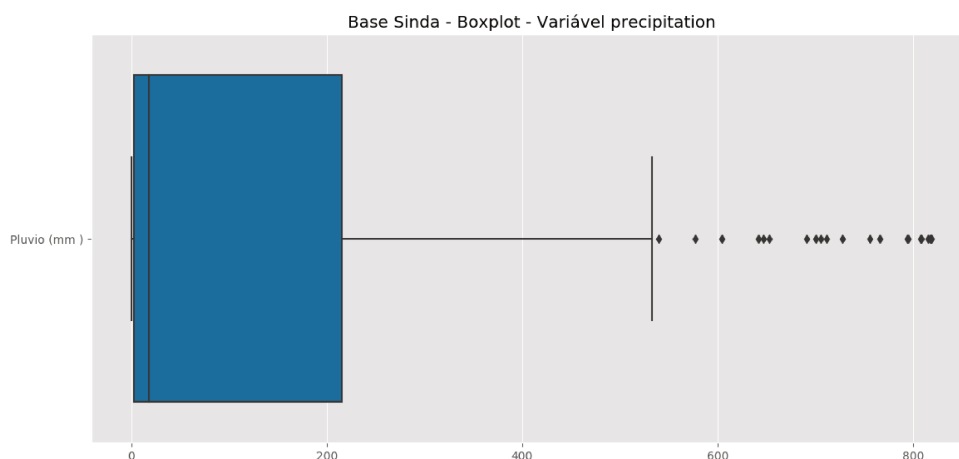


Figura 13 – Boxplot da variável *wind_speed* na base SINDA

ricos de temperatura utilizados como validação da base SONDA também devem ser utilizados nesse caso.

5.1.3 Base de dados POWER

A base POWER, mantida pela NASA, disponibiliza informações em três séries temporais diária, interanual e climatológico (onde é possível extrair longas amostras mensais). Essa base possui uma alta gama de variáveis, compreendendo a maior parte das variáveis selecionadas para este projeto, a saber: Índice de claridade, irradiação solar global de céu limpo, irradiância solar global, umidade relativa, precipitação, temperatura e velocidade do vento.

O gráfico de ausência de dados na figura 14 retrata um cenário ideal, onde não existem valores ausentes na base POWER, o que, a princípio pode ser considerado perfeito para realização de tarefas ligadas a dados, porém, é necessário mais de uma avaliação para que tenhamos certeza sobre a realidade que ela apresenta.

Analisando variáveis individualmente, começamos pela variável de umidade relativa do ar, anpalse essa que podemos ver na figura 15. Os valores colhidos demonstram boa qualidade, uma vez que, o boxplot apresenta uma boa construção, com valores que oscilam entre 56,89% e 92,6% e não apresenta dados fora da curva, ou seja, dentro da margem esperada.

A variável *precipitation*, tem sua representação gráfica através do boxplot

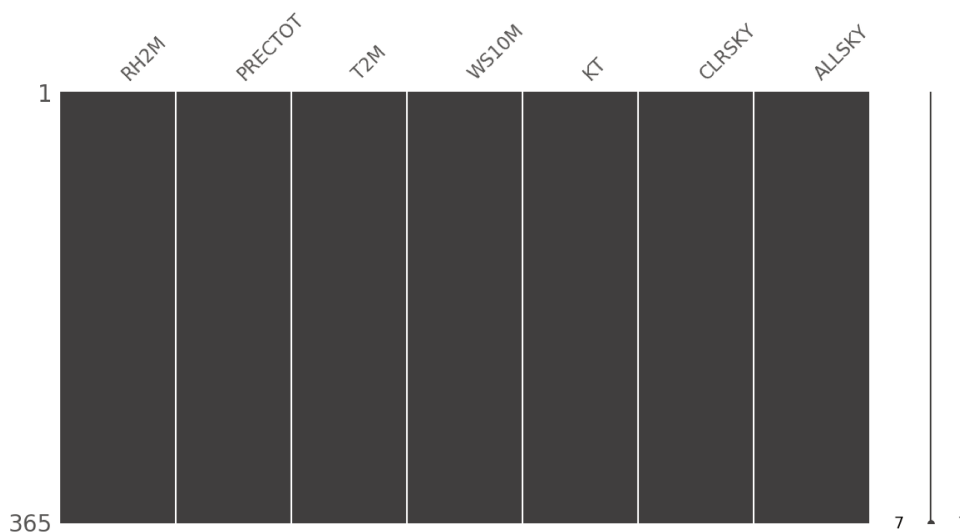


Figura 14 – Gráfico de dados ausentes da base POWER

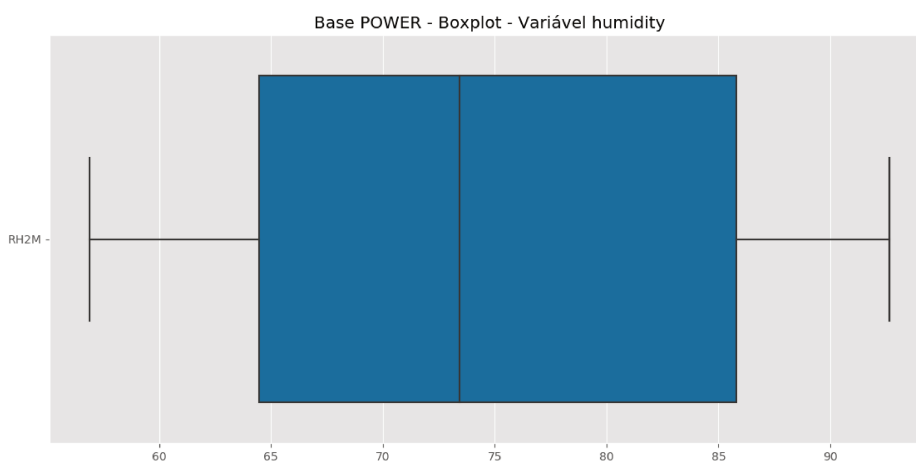


Figura 15 – Boxplot da variável *humidity* na base POWER

da figura 16. Nela podemos notar uma alta ocorrência de valores zerados, o que fez com que o boxplot não ficasse tão balanceado quanto a variável anterior. O destaque aos itens fora de curva se aplica uma vez que o desbalanceamento pode ser por conta de erros de coleta. Os valores máximo e mínimo para essa variável, para o ano de 2015, foram de 0 e 33.15 respectivamente.

A próxima variável a ser analisada é a que trata de valores de temperatura. Para isso, utilizaremos o boxplot mostrado pela figura 17. Como podemos notar, não apontam dados que estejam fora da curva, ou grandes discrepâncias, uma vez que o registro apresentou valores entre $24,93^{\circ}\text{C}$ e $30,65^{\circ}\text{C}$, totalmente dentro dos normais climatológicos propostos pelo INMET (mínima de $17,9^{\circ}\text{C}$ e máxima de $37,2^{\circ}\text{C}$).

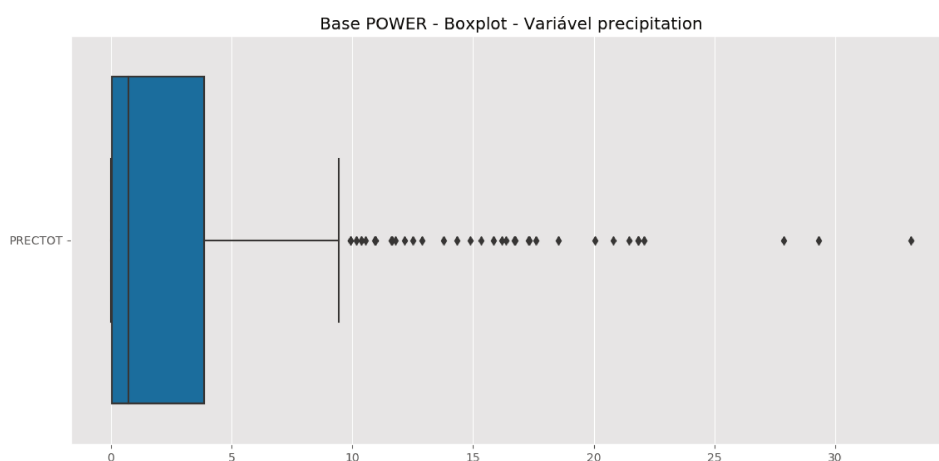


Figura 16 – Boxplot da variável *precipitation* na base POWER

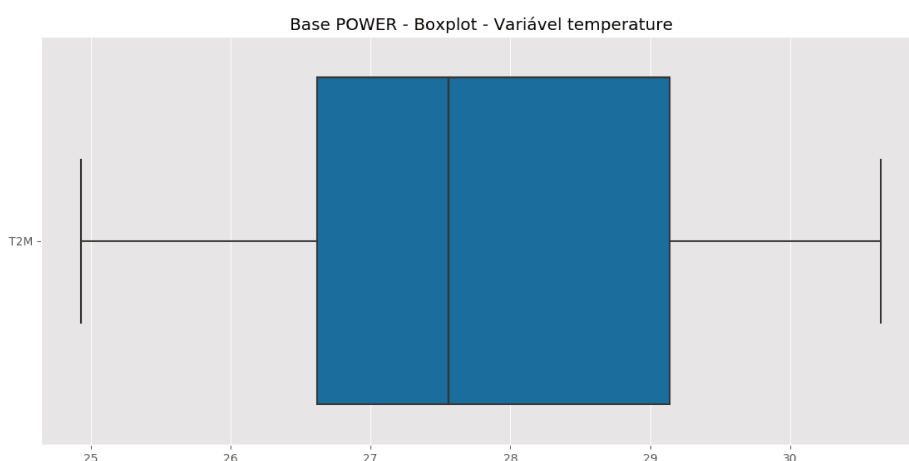


Figura 17 – Boxplot da variável *temperature* na base POWER

Analisando a variável velocidade do vento, notamos que também encontramos uma boa montagem de dados, uma vez que o boxplot, figura 19, não apresenta dados *outliers* e seus valores de quartis estão simétricos. Também comparando à validação da base SONDA, onde temos os valores válidos compreendidos entre $0m/s$ de valor mínimo e $25m/s$ de valor máximo, observamos que os registros da base POWER se encaixa dentro desse intervalo, onde os valores estão entre $1,12m/s$ e $4,55m/s$.

Já a análise da variável índice de claridade, não pode ser realizada, uma vez que os valores, em sua maioria, encontram-se em -999 significando ocorrência de erro ou ausência de coleta. Com isso o boxplot, figura 15, torna-se totalmente inválido. O mesmo ocorre com as demais variáveis dessa base: irradiação solar global

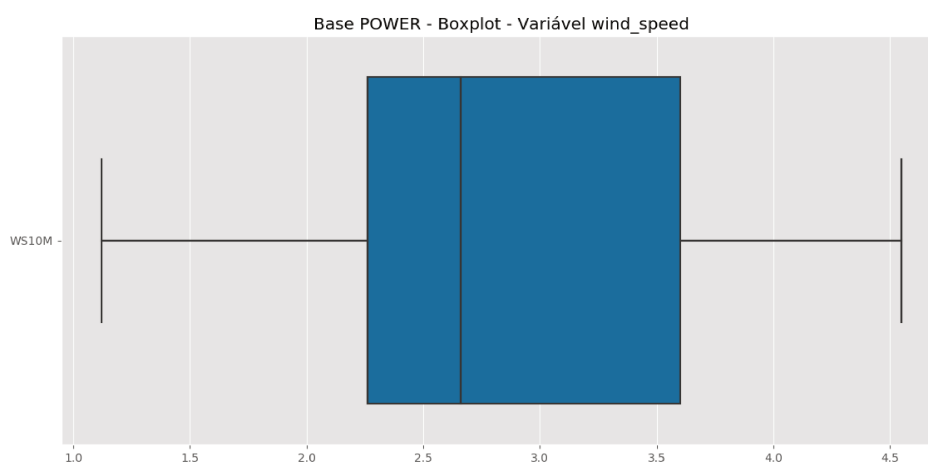


Figura 18 – Boxplot da variável *wind_speed* na base POWER

e irradiação solar de céu limpo, representadas pelos boxplots das figuras 20 e 21 respectivamente.

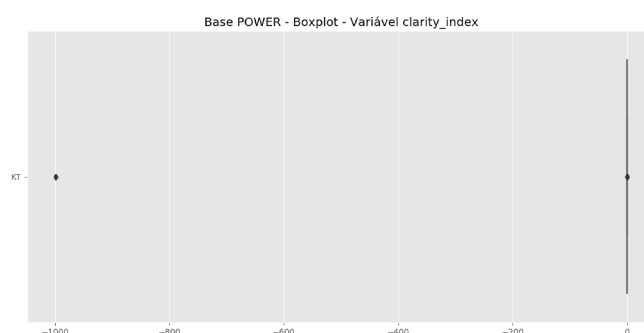


Figura 19 – Boxplot da variável *clarity_index* na base POWER

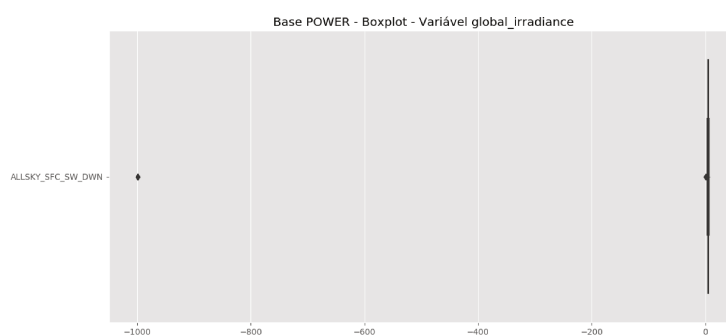


Figura 20 – Boxplot da variável *global_irradiance* na base POWER

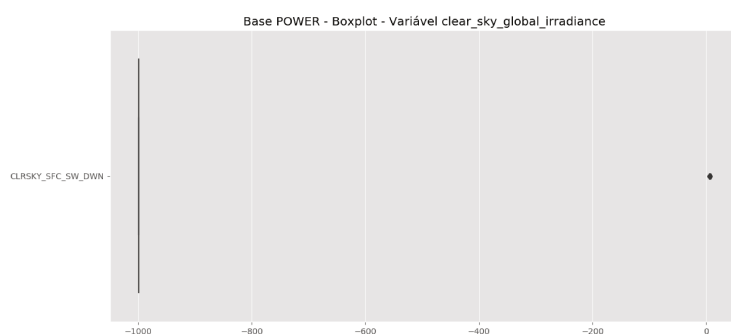


Figura 21 – Boxplot da variável *clear_sky* na base POWER

5.1.4 Base de dados BDMEP - Horário

A base de dados BDMEP Horária possui coletas diárias em três períodos, sendo eles as 00h00, 12h00 e 18h00. O INMET, responsável sobre a hospedagem de tal base de dados, não disponibiliza os mecanismos de validação de seus dados, assim como nos casos anteriores, sendo, então, necessário buscar outras fontes de avaliação, como os parâmetros da base SONDA e outros complementos históricos.

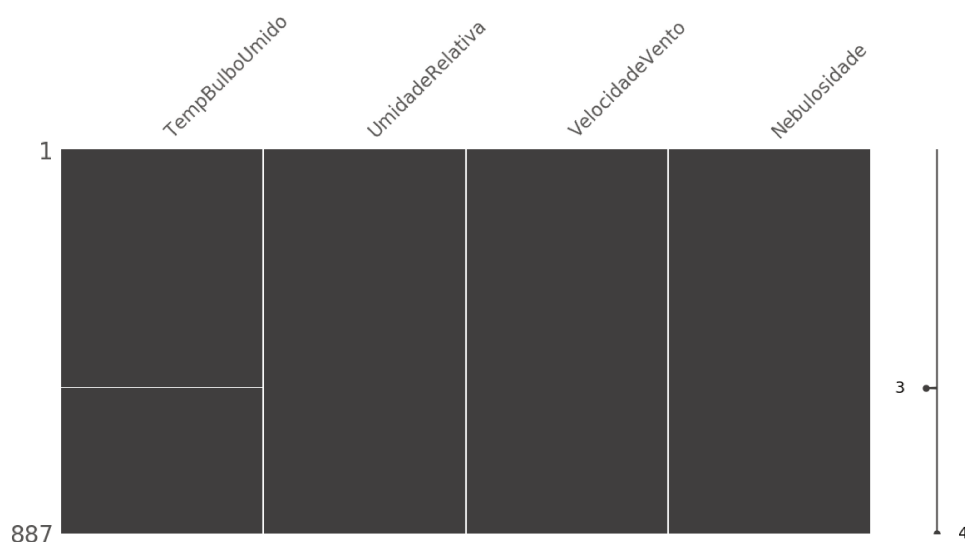


Figura 22 – Gráfico de dados ausentes da base BDMEP horário

Como podemos observar no gráfico da figura 22, a base possui um número de pequeno de anomalias, a não ser pelo fato de que coletas não realizadas foram totalmente ignoradas. Com três coletas em um dia, a base deveria conter exatas 1.095 coletas, porém, observamos que apenas 887 foram identificadas, fazendo que seja necessário uma interpretação mais precisa das informações.

Analisando o boxplot das variáveis, iniciando por *temperature*, temos o gráfico da figura 23, onde notamos a presença de dados *outliers*, mesmo assim, a base encontra-se dentro dos valores de validação da base SONDA.

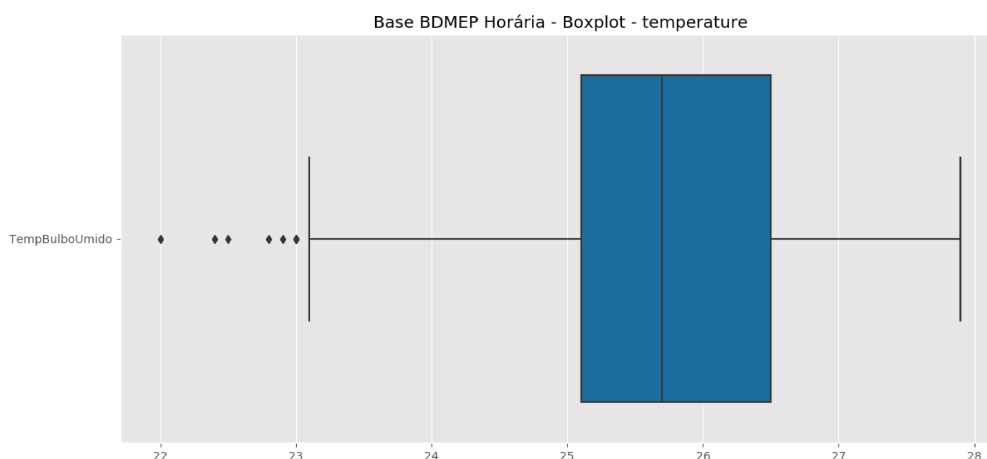


Figura 23 – Boxplot da variável *temperature* na base BDMEP horário

Quando tratamos da variável *humidity*, temos apenas uma ocorrência de dado *outlier*, apresentando uma boa construção da base, observando que ele está dentro do limite estabelecido por se tratar de um dado percentual.

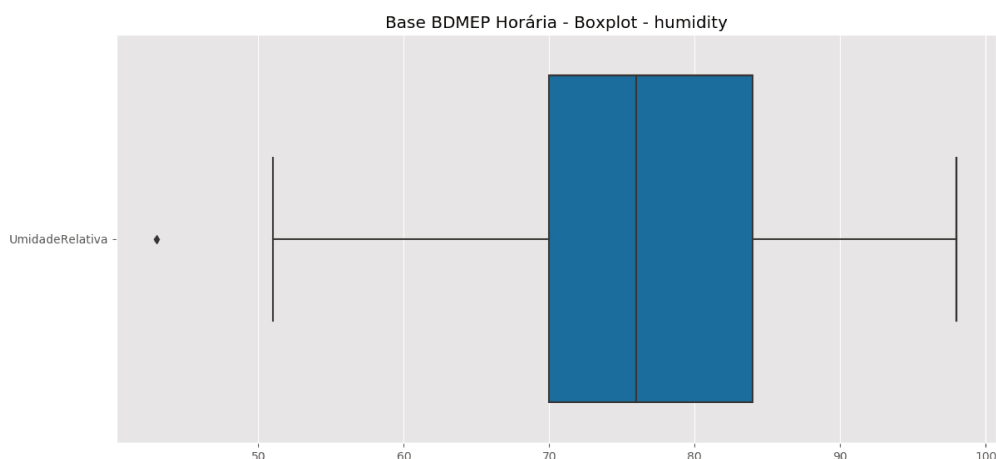


Figura 24 – Boxplot da variável *humidity* na base BDMEP horário

Uma variável exclusiva da base BDMEP horária, é a *cloud_cover*, que trata de informações relacionado a presença de nuvens no momento de coleta de dados, essa informação é medida em uma escala que varia de 0 a 10, onde 0 representa

um céu limpo e 10 um céu totalmente coberto por nuvens densas. O boxplot dessa variável não possui nenhuma ocorrência de *outlier* e sua construção é bem próxima ao ideal.

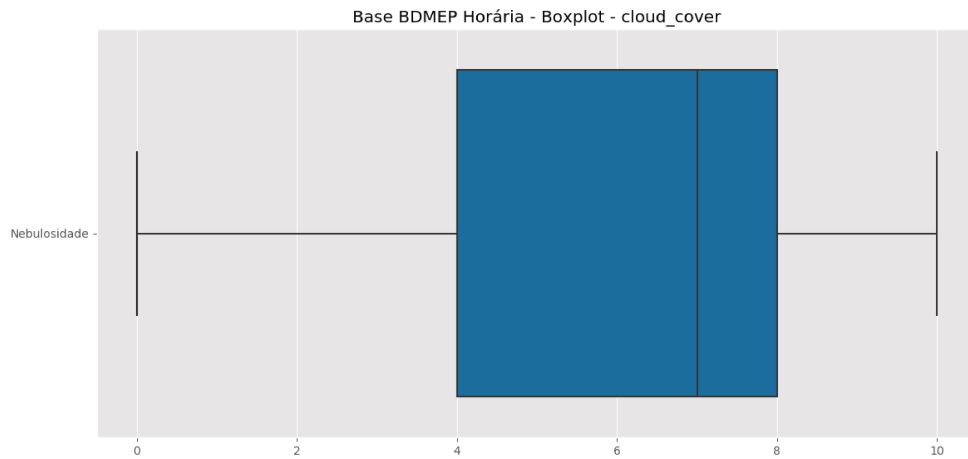


Figura 25 – Boxplot da variável *cloud_cover* na base BDMEP horário

Por último, temos a variável *wind_speed*, que se apresenta como a variável mais complicada de se analisar, principalmente, por conta de seus dados zerados, que fazem com que surja a dúvida sobre a existência ou não de valores fora da curva, sendo necessário interpretá-las através de bases auxiliares.

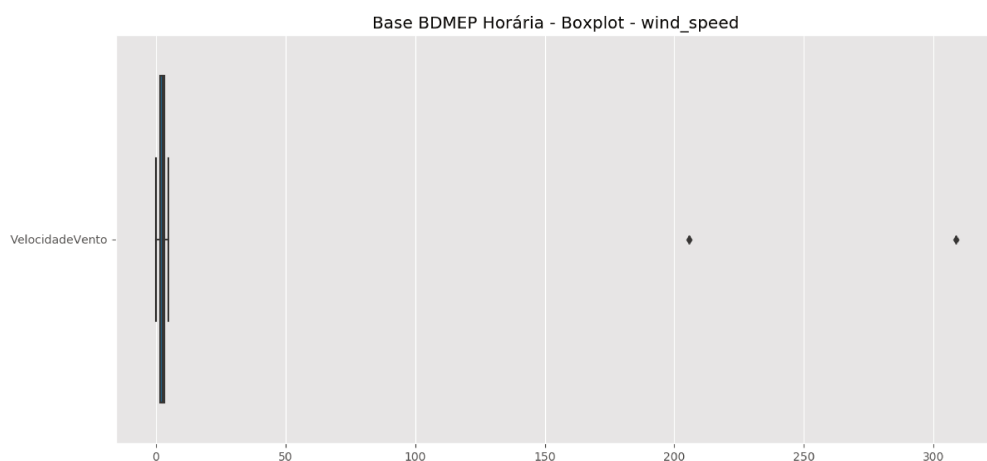


Figura 26 – Boxplot da variável *wind_speed* na base BDMEP horário

5.1.5 Base de dados BDMEP - Diário

Quando alteramos a dimensionalidade da base de dados BDMEP, obtemos variáveis diferentes, com algumas análises relevantes, os valores diários dessa base oferece as variáveis *precipitation*, *humidity*, *wind_speed* e valores de temperatura, mínimo e máximo. A base não possui métodos de validação disponibilizados em seu site, mesmo assim, pode ser utilizada como ferramenta auxiliar para a resolução de problemas. Os valores de temperatura necessitam de tratamento para a análise, haja visto que os valores estão desmembrados em valores máximo e mínimo.

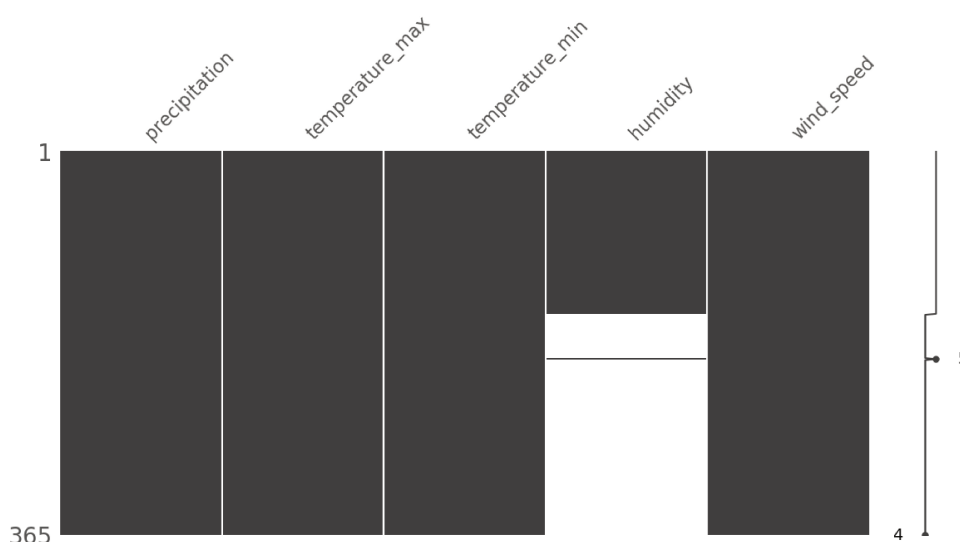


Figura 27 – Gráfico de dados ausentes da base BDMEP diário

Analisando as variáveis pelo gráfico de ausência na figura 27, notamos que os valores estão presentes, exceto a da variável *humidity*, onde a coleta foi menor do que a metade das outras variáveis. Mesmo assim, podemos analisar a qualidade dessa variável, conforme podemos ver na figura 28.

A qualidade da informação sobre umidade é alta, podemos ver que a base não possui valores *outliers*, porém, a ausência de coleta fez com que a informação contida em tal variável não pudesse ser utilizada em sua totalidade.

Prosseguindo a análise, temos as variáveis *precipitation* e *wind_speed*, cujos boxplot podem ser vistos, respectivamente, nas figuras 29 e 30.

Os dados colhidos sobre velocidade do vento (*wind_speed*) estão completos e possuem qualidade considerável, uma vez que não apresentam valores discrepantes.

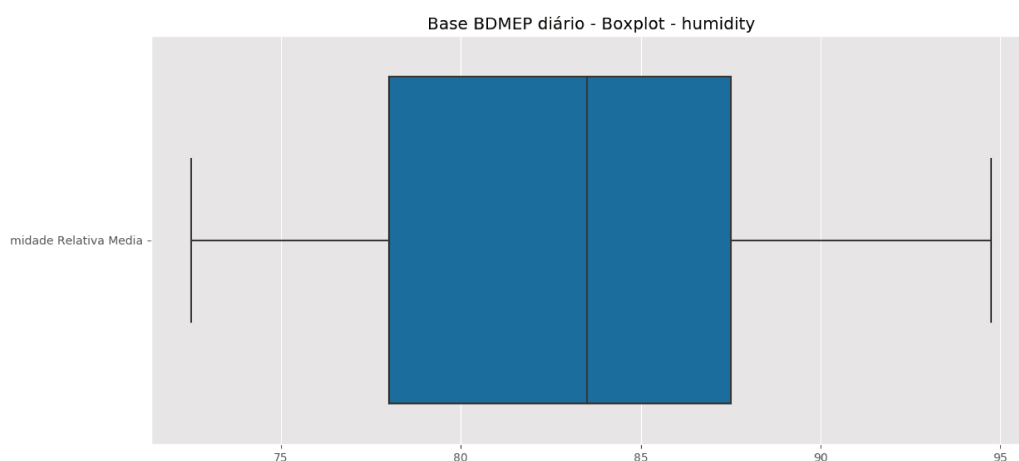


Figura 28 – Boxplot da variável *humidity* na base BDMEP diário

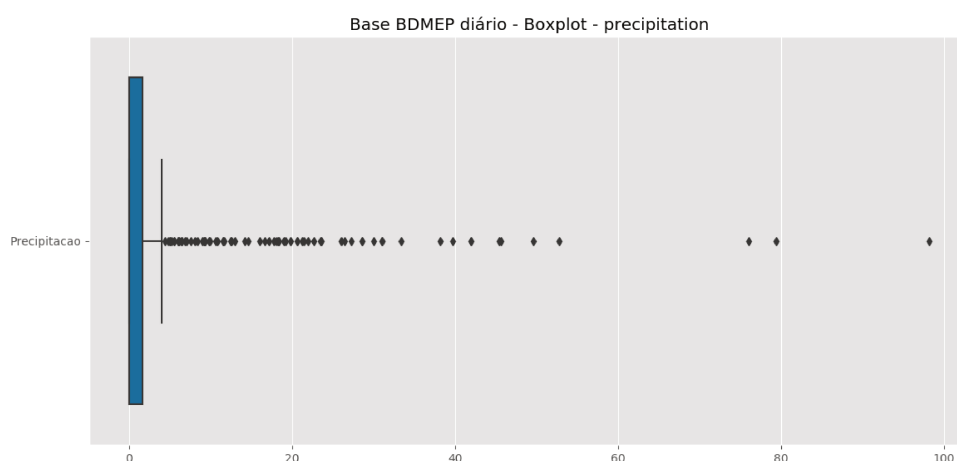


Figura 29 – Boxplot da variável *precipitation* na base BDMEP diário

Diferente da variável precipitação (*precipitation*) onde os valores com anomalias são a maioria dentre as coletas representadas. Porém, como se trata de dados meteorológicos, uma análise mais profunda pode apontar anomalias climáticas que torna os valores fiéis, porém, não condizentes com padrões esperados, como nosso caso. Um estudo sobre as demais bases, que possuem os mesmos dados, pode apontar esse cenário anômalo como padrão.

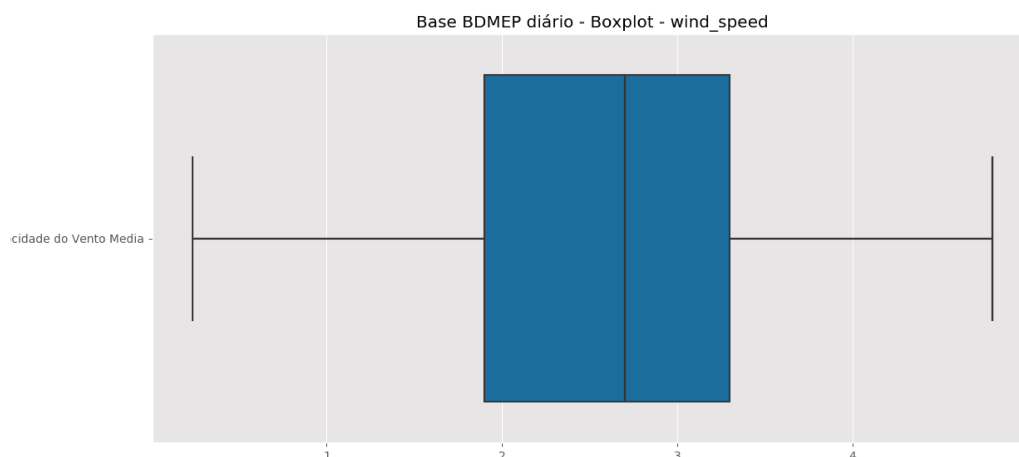


Figura 30 – Boxplot da variável *wind_speed* na base BDMEP diário

5.2 Adequação e montagem da base

Em todo o levantamento de variáveis para serem utilizadas nesse estudo, identificamos duas formas de obtenção das variáveis selecionadas, calculadas e extraídas de sensores que serão detalhadas nas subseções a seguir.

Para ambas as formas de tratamento, a linguagem de programação *Python*² foi escolhida, como ferramenta amplamente divulgada na comunidade científica de alta qualidade no tratamento de dados e aplicação de aprendizado de máquina. A comunidade que ampara a linguagem é muito ativa, disponibilizando diversas ferramentas produzidas para fins específicos a fim de evitar reinvenções de soluções, simples ou complexas, em novos projetos.

5.2.1 Variáveis calculadas

As bibliotecas *pvlb*³, *pandas*⁴, *suntime*⁵ e *pysolar*⁶ foram necessárias para a implementação dessa solução. Todas as bibliotecas são públicas, disponibilizadas como projeto livre pela plataforma de compartilhamento de códigos *GitHub*⁷.

A partir das informações de latitude e longitude é possível calcular os valo-

² *Python Software Foundation*. Disponível em <https://www.python.org/>. Acesso em 09/2018.

³ *Pysolar*. Disponível em <https://github.com/pingswept/pysolar>. Acesso em 04/2018.

⁴ *Pandas*. Disponível em <https://pandas.pydata.org/>. Acesso em 04/2018.

⁵ *SunTime*. Disponível em <https://github.com/SatAgro/suntime>. Acesso em 04/2018.

⁶ *PySolar*. Disponível em <https://github.com/pingswept/pysolar>. Acesso em 04/2018.

⁷ *GitHub*. Disponível em <https://github.com/>. Acesso em 04/2018.

res de elevação, azimute e zênite utilizando recursos da biblioteca *pvl*. Essas variáveis são levantadas com informações de fuso-horário do local, calculadas diariamente ao meio-dia.

Ainda com as informações da data, utilizamos a biblioteca *pysolar* para a aquisição de duas outras variáveis: horário do nascer e pôr do sol. Esses valores são necessários para a obtenção da duração do dia solar.

Ainda utilizando a biblioteca *pysolar*, conseguimos extrair informações sobre a radiação solar global com céu limpo, uma das variáveis extraídas do banco de dados *POWER*, cuja análise aponta dificuldades em obter confiabilidade em suas coletas.

5.2.2 Variáveis extraídas das bases de dados

Com as variáveis calculadas definidas, passamos para as que possuem registros armazenados em bases de dados. Essas variáveis selecionadas são: temperatura do ar, índice de claridade, irradiação global de céu limpo, área coberta por nuvens, irradiância solar global, umidade relativa do ar, precipitação pluvial, temperatura e velocidade do vento. Conforme exposto no quadro 9, não necessariamente todas as variáveis podem ser encontradas em todas as bases, uma vez que cada fonte contém objetivos e métodos específicos, fazendo necessário um tratamento para nosso cenário conforme a necessidade.

Após o carregamento dos dados oriundos das bases de dados citadas anteriormente, passamos, primeiramente, um "pente fino" em todas elas eliminando as variáveis que não correspondessem com as selecionadas, mantendo as variáveis listadas a seguir:

- Sonda: *global_irradiance*, *humidity*, *precipitation*, *temperature* e *wind_speed*;
- Sinda: *air_temperature*, *global_irradiance*, *humidity*, *precipitation*, *temperature* e *wind_speed*;
- Power: *clarity_index*, *clear_sky_global_irradiance*, *global_irradiance*, *humidity*, *precipitation*, *temperature* e *wind_speed*;
- BDMEP - horário: *cloud_cover*, *humidity*, *temperature* e *wind_speed*;

- BDMEP - diário: *humidity, precipitation, temperature* e *wind_speed*.

A seguir, os dados serão manipulados afim de obter um ganho de robustez, passando por todas as etapas de um pré-processamento, mas não cumprindo, necessariamente todos os métodos. As análises realizadas anteriormente permitem que conheçamos melhor cada variável e a necessidade de trabalho a ser aplicado a cada uma.

Para a base SONDA, houve a necessidade de formatação quanto à dimensionalidade das variáveis, fixando os dados em escala diária, sendo a mais adequada devido à disponibilidade das demais bases em relação aos projetos que serviram como alicerce para a seleção de todas as variáveis. Houve a necessidade de mescla dos dados fornecidos pela base, uma vez que, as mesmas são fornecidas em grupos mensais. Ainda sobre a base SONDA, a indexação por dia foi necessária, como parâmetro de relação entre as bases, item importante para a mescla de valores entre as bases.

A base POWER, fornecida pela NASA, possui grupos diferentes de extração, o grupo que trata de dados solarimétricos, necessitou de tratamentos nas variáveis *clear_sky_global_irradiance* e *global_irradiance*, que foram apresentados em uma unidade diferente das utilizadas pelos projetos que serviram de base, a base POWER apresentou, até essa data, as informações em $kWh/m^2/dia$, enquanto a exigência é de W/m^2 . A indexação por dia do ano também se fez necessária.

Quando falamos da base SINDA, que necessitou de indexação por dia do ano também, destaca-se o fato da base não apresentar uma variável com a média de temperatura, mas sim, valores para máxima e mínima diária, fazendo necessário a criação de tal. Assim como a base SONDA, foi necessário uma redução de dimensionalidade, quando reduzido a dados diários, pois a mesma apresenta coletas a cada 3 horas como padrão.

A base BDMEP em sua escala diária, apresentou as mesmas necessidades da base SINDA em relação as variáveis de temperatura, máxima e mínima. Não foi necessária a adequação diária, uma vez que a escala, assim como a base POWER, está em conformidade com o que se pede.

Por último a base BDMEP em sua escala horária, cuja necessidade se resumiu em indexação pro dia do ano, haja visto que todas as métricas estão corretas.

Após a modificação da base como um todo, o foco foi transferido para as variáveis individualmente, visando agregar confiabilidade nos valores apresentados, buscando a amostragem de dados sem *outliers*, onde valores que possuem parâmetros de medidas estejam adequadas, ou seja, confiável.

O padrão de tratamento foi aplicado a todas as variáveis e consiste na utilização do método z-score para que se eliminasse as inconsistências apresentadas na base e visíveis através dos boxplots de cada variável, sendo excluído os valores *outliers*. Após a adequação, utilizou-se métodos de aprendizado de máquina para imputação de novos valores, tanto os já ausentes, quanto os removidos pelo ajuste com o z-score.

O método utilizado foi a imputação interativa, disponibilizada pela biblioteca *fancyimpute*⁸ cuja implementação é inspirado no método MICE (Do inglês, *Multivariate Imputation by Chained Equations*), cuja implementação permite que seja realizada múltiplas imputações através de múltiplas variações de execução, possibilitando obter um conjunto de predições para um único preenchimento, inserindo o de menor padrão de erro. Uma das maiores vantagens desse método é o fato de que esse tipo de imputação respeita a variabilidade dos dados, não permitindo a eliminação completa das incertezas, que muitas vezes demonstra ser o motivo que despertou o interesse sobre a base analisada.

Após a imputação, os valores obtidos de cada base foram agregados de forma a obter uma única amostra que, concatenada as das demais variáveis, trazem a nova base de dados completa como resultado.

Vale ressaltar situações encontradas nas bases recolhidas para o cenário estabelecido, onde, as variáveis *clear_sky_global_irradiance*, retirada da base POWER, a *precipitation* e *wind_speed*, retiradas da base SONDA, e a variável *global_irradiance*, retirada da base SINDA, onde foram constatados ausências ou falhas que não puderam ser corrigidas, e assim, descartadas.

As ausências ficam por conta das variáveis retiradas da base SONDA, onde os dados se basearam nas demais bases. Já a variável da base POWER, possui valores ausentes em número superior aos presentes, o que impossibilita o uso, fazendo-se necessário o cálculo da mesma, conforme apontado na subseção anterior. O mesmo

⁸ Fancyimputer. Disponível em <https://github.com/iskandr/fancyimpute>. Acesso em 09/2018.

caso foi encontrado na variável da base SINDA, porém, assim como as variáveis da base SONDA, o uso das outras bases foram o suficiente para o preenchimento.

Após o explanado, obteve-se os boxplots que seguem nas figuras 31 a 38.

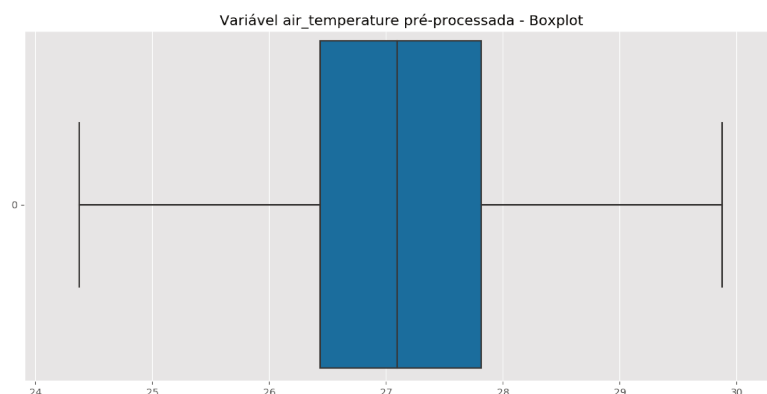


Figura 31 – Variável *air_temperature* pré-processada

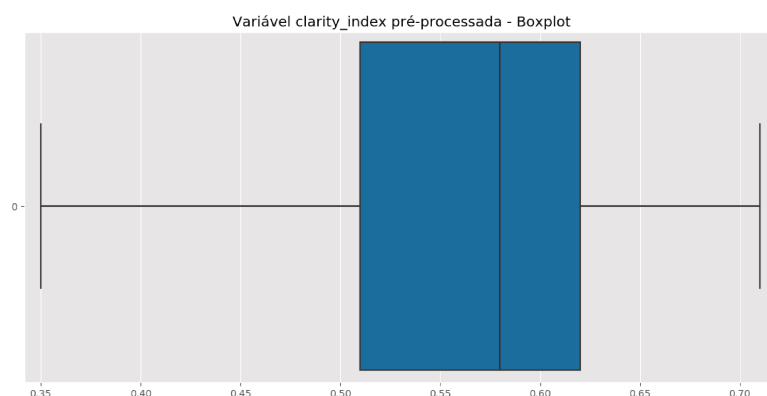


Figura 32 – Variável *clarity_index* pré-processada

Como podemos analisar nos novos gráficos boxplot gerados, as variáveis estão dentre valores razoáveis e não apresentam anomalias quanto aos seus dados outliers, com exceção da variável precipitação, que mesmo após todo o pré-processamento não obteve o resultado esperado, o que pode significar desvios reais que merecem uma atenção maior na hora de sua utilização.

Em resumo, podemos destacar os trabalhos em integração como adequação de unidades métricas e junção de bases, em limpeza com a imputação de novos valores a coletas ausentes e, também, tratamento dos dados ruidosos, em transformação podemos citar o uso de normalização utilizando o método de z-score e, por

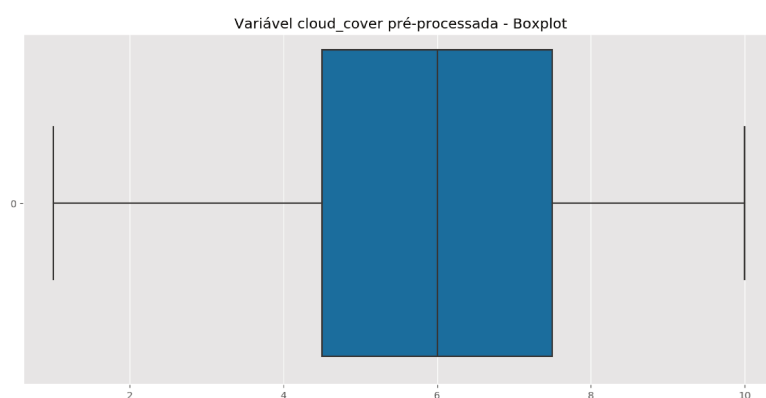


Figura 33 – Variável *cloud_cover* pré-processada

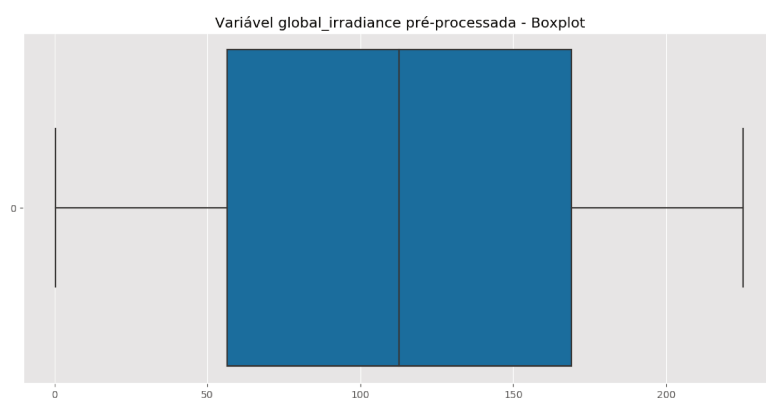


Figura 34 – Variável *global_irradiance* pré-processada

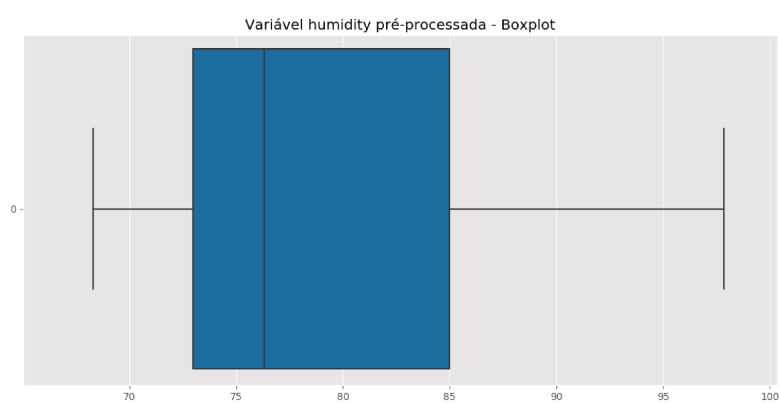


Figura 35 – Variável *humidity* pré-processada

último, quando lembramos de redução fizemos o uso de redução de dimensionalidade para adequar os dados a um mesmo padrão mais enxuto.

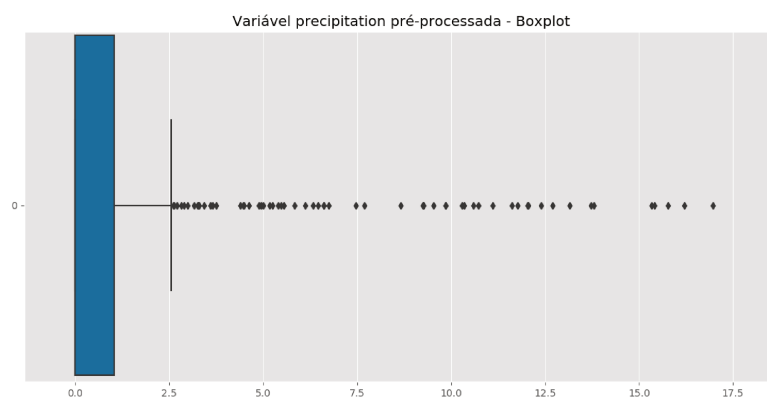


Figura 36 – Variável *precipitation* pré-processada

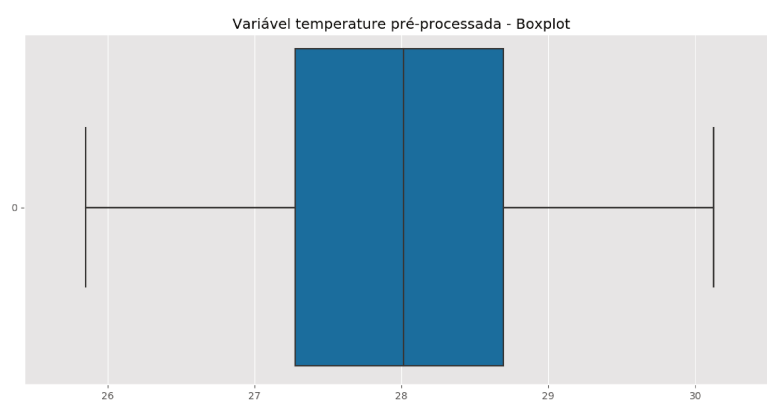


Figura 37 – Variável *temperature* pré-processada

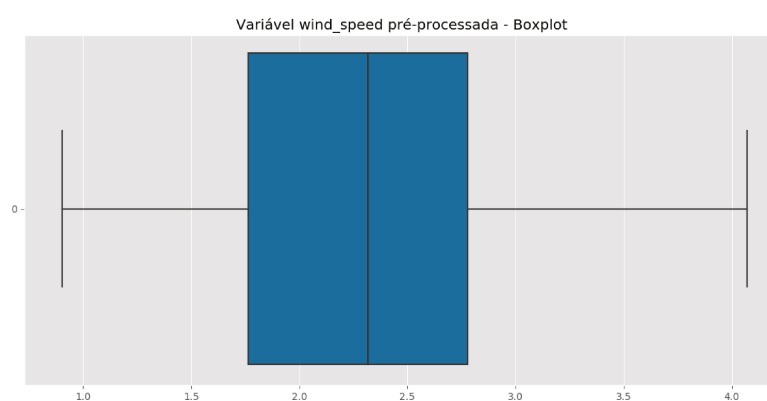


Figura 38 – Variável *wind_speed* pré-processada

6 Considerações finais

Esse projeto construiu uma base de dados robusta, cuja intenção seria possibilitar a aplicação e o desenvolvimento de pesquisas relacionadas à área solarimétrica no território brasileiro. O que atinge diversas camadas da população, não somente os pesquisadores, mas o público em geral que poderá ser beneficiado com uma prestação de serviço eficiente, proveniente de uma fonte não poluente.

A apoio fornecido por levantamentos históricos e predições é fundamental para atração de maior investimento para a área, pois possibilita ao investidor conhecer adequadamente o local onde receberá o aporte para o estabelecimento de uma nova planta solar ou a implementação de uma nova tecnologia.

Podemos comparar ao trabalho feito pelos pesquisadores Parkinson et al. [21], que deram início a um projeto que buscava armazenar informações confiáveis sobre perfis de expressão de genes, onde os dados armazenados necessitam ter um certo grau de confiabilidade para que possam ser utilizados como suporte a novas pesquisas de área biomédica. Como está evoluída, essa base de dados permite que pesquisadores enviem seus dados para que, após análise e tratamento, os dados sejam disponibilizados para a comunidade acadêmica, afim de disseminar o conhecimento na área em que está atribuída. O mesmo podemos ver em relação a essa pesquisa, que é o início de uma nova construção confiável e comunitária, cuja intenção é se estabelecer como uma base confiável para o desenvolvimento tecnológico na área energética.

O objetivo foi atingido, uma vez que foi possível extrair informações concretas acerca de uma determinada região, mesmo que para isso, pratiquemos o uso de diversas bases com dados suspeitos, disponibilizados por fontes de pesquisas públicas e/ou governamentais. O nível de confiabilidade, conforme evidenciado, aumentou consideravelmente, uma vez que as dúvidas foram eliminadas durante as análises compreendidas nesse estudo.

6.1 Trabalhos futuros

Assim como a base *ArrayExpress*, sugerida por Parkinson [21], este trabalho permite que evoluções aconteçam ao seu em torno, fazendo com que novos padrões de coletas e armazenamento de dados solarimétricos brasileiro possam surgir, possibilitando o acréscimo de qualidade devido à quantidade de informação coletada ou submetida pela comunidade. Também, a aplicação de alguns métodos preditivos, até então possíveis apenas em determinadas regiões, possam ser utilizadas em nosso território. Bem como a criação de uma plataforma, onde será possível buscar dados sobre qualquer localidade do território brasileiro, podendo definir inclusive, qual método preditivo deseja analisar. Com a garantia de qualidade adquirida, o uso de ferramentas IoT é um fator relevante no crescimento do cenário, uma vez que dispositivos de coletas em curto prazo e em maior volume passam a ter um depósito confiável para agregar valor ao projeto no qual pertence, ou qualquer outro projeto, cuja base pode ser construída através de dados.

Referências

- [1] Suad A Alasadi and Wesam S Bhaya. Review of data preprocessing techniques in data mining. *Journal of Engineering and Applied Sciences*, 12(16):4102–4107, 2017.
- [2] Danah Boyd and Kate Crawford. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5):662–679, 2012.
- [3] Eric T Bradlow, Manish Gangwar, Praveen Kopalle, and Sudhir Voleti. The role of big data and predictive analytics in retailing. *Journal of Retailing*, 93(1):79–95, 2017.
- [4] Guido Cervone, Laura Clemente-Harding, Stefano Alessandrini, and Luca Delle Monache. Short-term photovoltaic power forecasting using Artificial Neural Networks and an Analog Ensemble. *Renewable Energy*, 108:274–286, 2017.
- [5] Ian L Dryden and David J Hodge. Journeys in big data statistics. *Statistics & Probability Letters*, 136:121–125, 2018.
- [6] Katti Faceli, Ana Carolina Lorena, João Gama, and André C. P. L. F. Carvalho. *Inteligência Artificial: Uma abordagem de aprendizado de máquina*. Grupo Gen-LTC, 2011.
- [7] David John Gagne, Amy McGovern, Sue Ellen Haupt, and John K. Williams. Evaluation of statistical learning configurations for gridded solar irradiance forecasting. *Solar Energy*, 150:383–393, 2017.
- [8] Salvador García, Julián Luengo, and Francisco Herrera. *Data preprocessing in data mining*. Springer, 2015.
- [9] Ronaldo Goldschmidt, Eduardo Bezerra, and E Passos. *Data mining: conceitos, técnicas, algoritmos, orientações e aplicações*, volume 1. Elsevier Editora, 2 edition, 2015.

- [10] Giorgio Graditi, Sergio Ferlito, Giovanna Adinolfi, Giuseppe Marco Tina, and Cristina Ventura. Energy yield estimation of thin-film photovoltaic plants by using physical approach and artificial neural networks. *Solar Energy*, 130:232–243, 2016.
- [11] Mon-Fong Jiang, Shian-Shyong Tseng, and Shan-Yi Liao. Data types generalization for data mining algorithms. In *Systems, Man, and Cybernetics, 1999. IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on*, volume 3, pages 928–933. IEEE, 1999.
- [12] Stephen Kaisler, Frank Armour, J Alberto Espinosa, and William Money. Big data: Issues and challenges moving forward. In *System sciences (HICSS), 2013 46th Hawaii international conference on*, pages 995–1004. IEEE, 2013.
- [13] Mehmed Kantardzic. *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons, 2011.
- [14] Natalija Koseleva and Guoda Ropaite. Big data in building energy efficiency: understanding of big data and main challenges. *Procedia Engineering*, 172:544–549, 2017.
- [15] Doug Laney. 3d data management: Controlling data volume, velocity and variety. *META group research note*, 6(70):1, 2001.
- [16] Daniel T Larose and Chantal D Larose. *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons, 2014.
- [17] F. J. L. de Lima. *Previsão de irradiação solar no nordeste do Brasil empregando o modelo WRF ajustado porredes neurais artificiais (RNAs)*. PhD thesis, Tese (Doutorado em Meteorologia). São Jose dos Campos, INPE, 2015.
- [18] Francisco J.L. Lima, Fernando R. Martins, Enio B. Pereira, Elke Lorenz, and Dettlev Heinemann. Forecast for surface solar irradiance at the Brazilian Northeastern region using NWP model and artificial neural networks. *Renewable Energy*, 87:807–818, 2016.
- [19] Fernando Ramos Martins, Enio Bueno Pereira, and MP de S ECHER. Levantamento dos recursos de energia solar no brasil com o emprego de saté-

- lite geoestacionário—o projeto swera. *Revista Brasileira de Ensino de Física*, 26(2):145–159, 2004.
- [20] Carlos Ordonez. Data set preprocessing and transformation in a database system. *Intelligent Data Analysis*, 15(4):613–631, 2011.
- [21] H Parkinson, Misha Kapushesky, Mohammadreza Shojatalab, Niran Abeygunawardena, Richard Coulson, Anna Farne, Ele Holloway, Nikolay Kolesnykov, P Lilja, Margus Lukk, et al. Arrayexpress—a public database of microarray experiments and gene expression profiles. *Nucleic acids research*, 35(suppl_1):D747–D750, 2006.
- [22] Enio Bueno Pereira, Fernando Ramos Martins, Samuel Luna de Abreu, and Ricardo Rüther. *Atlas brasileiro de energia solar*, volume 1. INPE - São José dos Campos, 1 edition, 2006.
- [23] Enio Bueno Pereira, Fernando Ramos Martins, Samuel Luna de Abreu, and Ricardo Rüther. *Atlas brasileiro de energia solar*, volume 1. INPE - São José dos Campos, 2 edition, 2017.
- [24] Dave Renné, Ray George, Bill Marion, Richard Perez, Christoph Schillings, Franz Trieb, Richard Meyer, Enio B Pereira, Fernando R Martins, Samuel Luna De Abreu, et al. Results of solar resource assessments in the unep/swera project. In *Proceedings of the Solar World Congress, Orlando, FL, International Solar Energy Soc. Citeseer*, 2005.
- [25] John R Searle. *Minds, brains and science*. Harvard University Press, 1984.
- [26] Valdemar W Setzer. Dado, informação, conhecimento e competência. *DataGramZero Revista de Ciência da Informação*, n. 0, page 28, 1999.
- [27] Stanley Smith Stevens et al. On the theory of scales of measurement. 1946.
- [28] Pang-Ning Tan et al. *Introduction to data mining*. Pearson Education India, 2007.
- [29] Chunming Tu, Xi He, Zhikang Shuai, and Fei Jiang. Big data issues in smart grid—a review. *Renewable and Sustainable Energy Reviews*, 79:1099–1107, 2017.

- [30] R. Urraca, J. Antonanzas, M. Alia-Martinez, F. J. Martinez-De-Pison, and F. Antonanzas-Torres. Smart baseline models for solar irradiation forecasting. *Energy Conversion and Management*, 108:539–548, 2016.
- [31] Paul F Velleman and Leland Wilkinson. Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician*, 47:65–72, 1993.
- [32] Martin White. Digital workplaces: Vision and reality. *Business information review*, 29(4):205–214, 2012.
- [33] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding. Data mining with big data. *IEEE transactions on knowledge and data engineering*, 26(1):97–107, 2014.
- [34] Shui Yu, Meng Liu, Wanchun Dou, Xiting Liu, and Sanming Zhou. Networking for big data: A survey. *IEEE Communications Surveys & Tutorials*, 19(1):531–549, 2017.
- [35] Guangya Zhu, Tin Tai Chow, and Norman Tse. Short-term load forecasting coupled with weather profile generation methodology. *Building Services Engineering Research and Technology*, 39(3):310–327, 2018.

Apêndices

APÊNDICE A – Tabela de dados pré-processados

index	latitude	longitude	elevation	azimuth	zenith	hour	solar_day	clear_sky
0	-21,788	-46,561	-45,124	183,095	135,124	12:00:00	48360	1018,705
1	-21,788	-46,561	-45,200	183,255	135,200	12:00:00	48360	1018,418
2	-21,788	-46,561	-45,285	183,414	135,285	12:00:00	48300	1018,095
3	-21,788	-46,561	-45,376	183,573	135,376	12:00:00	48300	1017,739
4	-21,788	-46,561	-45,475	183,730	135,475	12:00:00	48300	1017,348
5	-21,788	-46,561	-45,581	183,887	135,581	12:00:00	48240	1016,923
6	-21,788	-46,561	-45,694	184,043	135,694	12:00:00	48180	1016,464
7	-21,788	-46,561	-45,815	184,198	135,815	12:00:00	48180	1015,970
8	-21,788	-46,561	-45,942	184,353	135,942	12:00:00	48180	1015,443
9	-21,788	-46,561	-46,077	184,505	136,077	12:00:00	48120	1014,883
10	-21,788	-46,561	-46,219	184,657	136,219	12:00:00	48060	1014,288
11	-21,788	-46,561	-46,367	184,808	136,367	12:00:00	48060	1013,661
12	-21,788	-46,561	-46,523	184,957	136,523	12:00:00	48000	1013,000
13	-21,788	-46,561	-46,686	185,105	136,686	12:00:00	47940	1012,306
14	-21,788	-46,561	-46,855	185,252	136,855	12:00:00	47940	1011,579
15	-21,788	-46,561	-47,031	185,397	137,031	12:00:00	47880	1010,819
16	-21,788	-46,561	-47,214	185,540	137,214	12:00:00	47820	1010,027
17	-21,788	-46,561	-47,403	185,682	137,403	12:00:00	47820	1009,203
18	-21,788	-46,561	-47,599	185,822	137,599	12:00:00	47760	1008,347
19	-21,788	-46,561	-47,801	185,960	137,801	12:00:00	47700	1007,459
20	-21,788	-46,561	-48,010	186,097	138,010	12:00:00	47640	1006,539
21	-21,788	-46,561	-48,225	186,231	138,225	12:00:00	47580	1005,589
22	-21,788	-46,561	-48,446	186,363	138,446	12:00:00	47520	1004,608
23	-21,788	-46,561	-48,673	186,493	138,673	12:00:00	47460	1003,596
24	-21,788	-46,561	-48,907	186,620	138,907	12:00:00	47460	1002,554
25	-21,788	-46,561	-49,146	186,745	139,146	12:00:00	47400	1001,481
26	-21,788	-46,561	-49,391	186,868	139,391	12:00:00	47280	1000,379
27	-21,788	-46,561	-49,642	186,988	139,642	12:00:00	47280	999,247
28	-21,788	-46,561	-49,899	187,106	139,899	12:00:00	47220	998,086
29	-21,788	-46,561	-50,161	187,221	140,161	12:00:00	47100	996,895
30	-21,788	-46,561	-50,429	187,333	140,429	12:00:00	47100	995,675
31	-21,788	-46,561	-50,702	187,443	140,702	12:00:00	47040	994,425
32	-21,788	-46,561	-50,980	187,550	140,980	12:00:00	46920	993,147
33	-21,788	-46,561	-51,264	187,654	141,264	12:00:00	46920	991,841
34	-21,788	-46,561	-51,553	187,756	141,553	12:00:00	46800	990,506
35	-21,788	-46,561	-51,847	187,854	141,847	12:00:00	46800	989,142
36	-21,788	-46,561	-52,145	187,950	142,145	12:00:00	46680	987,750
37	-21,788	-46,561	-52,449	188,043	142,449	12:00:00	46620	986,330
38	-21,788	-46,561	-52,757	188,133	142,757	12:00:00	46560	984,882
39	-21,788	-46,561	-53,070	188,221	143,070	12:00:00	46500	983,407

40	-21,788	-46,561	-53,387	188,305	143,387	12:00:00	46440	981,903
41	-21,788	-46,561	-53,709	188,387	143,709	12:00:00	46380	980,372
42	-21,788	-46,561	-54,035	188,465	144,035	12:00:00	46260	978,813
43	-21,788	-46,561	-54,365	188,541	144,365	12:00:00	46260	977,226
44	-21,788	-46,561	-54,699	188,614	144,699	12:00:00	46140	975,613
45	-21,788	-46,561	-55,037	188,684	145,037	12:00:00	46080	973,973
46	-21,788	-46,561	-55,379	188,751	145,379	12:00:00	46020	972,306
47	-21,788	-46,561	-55,725	188,814	145,725	12:00:00	45960	970,612
48	-21,788	-46,561	-56,074	188,875	146,074	12:00:00	45840	968,893
49	-21,788	-46,561	-56,426	188,932	146,426	12:00:00	45840	967,147
50	-21,788	-46,561	-56,782	188,987	146,782	12:00:00	45720	965,376
51	-21,788	-46,561	-57,141	189,038	147,141	12:00:00	45660	963,579
52	-21,788	-46,561	-57,504	189,085	147,504	12:00:00	45600	961,758
53	-21,788	-46,561	-57,869	189,130	147,869	12:00:00	45540	959,911
54	-21,788	-46,561	-58,237	189,171	148,237	12:00:00	45420	958,039
55	-21,788	-46,561	-58,608	189,208	148,608	12:00:00	45360	956,143
56	-21,788	-46,561	-58,982	189,242	148,982	12:00:00	45300	954,223
57	-21,788	-46,561	-59,358	189,273	149,358	12:00:00	45240	952,278
58	-21,788	-46,561	-59,736	189,300	149,736	12:00:00	45120	950,309
59	-21,788	-46,561	-60,117	189,324	150,117	12:00:00	45060	948,316
60	-21,788	-46,561	-60,500	189,344	150,500	12:00:00	45000	946,300
61	-21,788	-46,561	-60,885	189,361	150,885	12:00:00	44880	944,260
62	-21,788	-46,561	-61,272	189,375	151,272	12:00:00	44880	942,197
63	-21,788	-46,561	-61,660	189,385	151,660	12:00:00	44760	940,110
64	-21,788	-46,561	-62,051	189,392	152,051	12:00:00	44700	938,000
65	-21,788	-46,561	-62,443	189,396	152,443	12:00:00	44640	935,867
66	-21,788	-46,561	-62,836	189,396	152,836	12:00:00	44520	933,711
67	-21,788	-46,561	-63,231	189,393	153,231	12:00:00	44460	931,531
68	-21,788	-46,561	-63,627	189,387	153,627	12:00:00	44340	929,329
69	-21,788	-46,561	-64,024	189,378	154,024	12:00:00	44280	927,105
70	-21,788	-46,561	-64,422	189,366	154,422	12:00:00	44280	924,857
71	-21,788	-46,561	-64,821	189,351	154,821	12:00:00	44160	922,587
72	-21,788	-46,561	-65,221	189,333	155,221	12:00:00	44100	920,295
73	-21,788	-46,561	-65,622	189,312	155,622	12:00:00	44040	917,981
74	-21,788	-46,561	-66,022	189,288	156,022	12:00:00	43920	915,645
75	-21,788	-46,561	-66,424	189,262	156,424	12:00:00	43860	913,287
76	-21,788	-46,561	-66,826	189,232	156,826	12:00:00	43800	910,909
77	-21,788	-46,561	-67,227	189,200	157,227	12:00:00	43680	908,509
78	-21,788	-46,561	-67,629	189,165	157,629	12:00:00	43620	906,089
79	-21,788	-46,561	-68,031	189,127	158,031	12:00:00	43560	903,648
80	-21,788	-46,561	-68,433	189,087	158,433	12:00:00	43440	901,187
81	-21,788	-46,561	-68,834	189,043	158,834	12:00:00	43380	898,707
82	-21,788	-46,561	-69,235	188,997	159,235	12:00:00	43380	896,207
83	-21,788	-46,561	-69,636	188,948	159,636	12:00:00	43260	893,688
84	-21,788	-46,561	-70,035	188,896	160,035	12:00:00	43200	891,151
85	-21,788	-46,561	-70,434	188,842	160,434	12:00:00	43140	888,595
86	-21,788	-46,561	-70,833	188,785	160,833	12:00:00	43020	886,022
87	-21,788	-46,561	-71,230	188,725	161,230	12:00:00	42960	883,430
88	-21,788	-46,561	-71,626	188,663	161,626	12:00:00	42900	880,821

89	-21,788	-46,561	-72,021	188,599	162,021	12:00:00	42780	878,195
90	-21,788	-46,561	-72,414	188,532	162,414	12:00:00	42720	875,552
91	-21,788	-46,561	-72,807	188,463	162,807	12:00:00	42660	872,892
92	-21,788	-46,561	-73,197	188,393	163,197	12:00:00	42540	870,216
93	-21,788	-46,561	-73,587	188,321	163,587	12:00:00	42480	867,525
94	-21,788	-46,561	-73,974	188,247	163,974	12:00:00	42420	864,817
95	-21,788	-46,561	-74,360	188,172	164,360	12:00:00	42360	862,094
96	-21,788	-46,561	-74,743	188,095	164,743	12:00:00	42300	859,356
97	-21,788	-46,561	-75,125	188,018	165,125	12:00:00	42240	856,602
98	-21,788	-46,561	-75,504	187,941	165,504	12:00:00	42120	853,835
99	-21,788	-46,561	-75,882	187,863	165,882	12:00:00	42060	851,053
100	-21,788	-46,561	-76,257	187,785	166,257	12:00:00	42000	848,258
101	-21,788	-46,561	-76,629	187,707	166,629	12:00:00	41880	845,449
102	-21,788	-46,561	-76,999	187,630	166,999	12:00:00	41880	842,628
103	-21,788	-46,561	-77,366	187,554	167,366	12:00:00	41820	839,794
104	-21,788	-46,561	-77,731	187,480	167,731	12:00:00	41700	836,949
105	-21,788	-46,561	-78,093	187,407	168,093	12:00:00	41640	834,093
106	-21,788	-46,561	-78,452	187,336	168,452	12:00:00	41580	831,227
107	-21,788	-46,561	-78,808	187,268	168,808	12:00:00	41460	828,351
108	-21,788	-46,561	-79,160	187,203	169,160	12:00:00	41460	825,466
109	-21,788	-46,561	-79,510	187,141	169,510	12:00:00	41400	822,573
110	-21,788	-46,561	-79,856	187,083	169,856	12:00:00	41280	819,672
111	-21,788	-46,561	-80,199	187,029	170,199	12:00:00	41220	816,766
112	-21,788	-46,561	-80,538	186,981	170,538	12:00:00	41220	813,853
113	-21,788	-46,561	-80,874	186,938	170,874	12:00:00	41100	810,936
114	-21,788	-46,561	-81,206	186,903	171,206	12:00:00	41040	808,015
115	-21,788	-46,561	-81,534	186,875	171,534	12:00:00	40920	805,090
116	-21,788	-46,561	-81,858	186,856	171,858	12:00:00	40920	802,164
117	-21,788	-46,561	-82,179	186,848	172,179	12:00:00	40860	799,236
118	-21,788	-46,561	-82,495	186,851	172,495	12:00:00	40740	796,307
119	-21,788	-46,561	-82,807	186,868	172,807	12:00:00	40740	793,379
120	-21,788	-46,561	-83,115	186,900	173,115	12:00:00	40680	790,451
121	-21,788	-46,561	-83,418	186,950	173,418	12:00:00	40560	787,526
122	-21,788	-46,561	-83,717	187,021	173,717	12:00:00	40560	784,604
123	-21,788	-46,561	-84,011	187,115	174,011	12:00:00	40440	781,686
124	-21,788	-46,561	-84,301	187,237	174,301	12:00:00	40440	778,773
125	-21,788	-46,561	-84,586	187,390	174,586	12:00:00	40380	775,865
126	-21,788	-46,561	-84,866	187,579	174,866	12:00:00	40320	772,965
127	-21,788	-46,561	-85,142	187,810	175,142	12:00:00	40260	770,072
128	-21,788	-46,561	-85,412	188,090	175,412	12:00:00	40140	767,189
129	-21,788	-46,561	-85,677	188,428	175,677	12:00:00	40140	764,316
130	-21,788	-46,561	-85,937	188,832	175,937	12:00:00	40020	761,455
131	-21,788	-46,561	-86,191	189,315	176,191	12:00:00	40020	758,606
132	-21,788	-46,561	-86,439	189,891	176,439	12:00:00	40020	755,772
133	-21,788	-46,561	-86,682	190,577	176,682	12:00:00	39900	752,953
134	-21,788	-46,561	-86,918	191,395	176,918	12:00:00	39900	750,152
135	-21,788	-46,561	-87,148	192,371	177,148	12:00:00	39780	747,369
136	-21,788	-46,561	-87,372	193,539	177,372	12:00:00	39780	744,606
137	-21,788	-46,561	-87,588	194,942	177,588	12:00:00	39720	741,866

138	-21,788	-46,561	-87,797	196,632	177,797	12:00:00	39660	739,149
139	-21,788	-46,561	-87,997	198,681	177,997	12:00:00	39660	736,457
140	-21,788	-46,561	-88,189	201,179	178,189	12:00:00	39600	733,792
141	-21,788	-46,561	-88,370	204,241	178,370	12:00:00	39540	731,157
142	-21,788	-46,561	-88,539	208,016	178,539	12:00:00	39480	728,552
143	-21,788	-46,561	-88,694	212,684	178,694	12:00:00	39480	725,980
144	-21,788	-46,561	-88,833	218,451	178,833	12:00:00	39420	723,442
145	-21,788	-46,561	-88,951	225,510	178,951	12:00:00	39360	720,940
146	-21,788	-46,561	-89,044	233,955	179,044	12:00:00	39360	718,476
147	-21,788	-46,561	-89,108	243,652	179,108	12:00:00	39300	716,051
148	-21,788	-46,561	-89,140	254,122	179,140	12:00:00	39300	713,668
149	-21,788	-46,561	-89,140	264,606	179,140	12:00:00	39240	711,328
150	-21,788	-46,561	-89,110	274,331	179,110	12:00:00	39240	709,032
151	-21,788	-46,561	-89,057	282,795	179,057	12:00:00	39180	706,784
152	-21,788	-46,561	-88,986	289,842	178,986	12:00:00	39120	704,583
153	-21,788	-46,561	-88,903	295,555	178,903	12:00:00	39120	702,433
154	-21,788	-46,561	-88,813	300,123	178,813	12:00:00	39060	700,334
155	-21,788	-46,561	-88,719	303,752	178,719	12:00:00	39060	698,288
156	-21,788	-46,561	-88,623	306,624	178,623	12:00:00	39060	696,298
157	-21,788	-46,561	-88,528	308,888	178,528	12:00:00	39000	694,365
158	-21,788	-46,561	-88,434	310,663	178,434	12:00:00	39000	692,490
159	-21,788	-46,561	-88,342	312,038	178,342	12:00:00	38940	690,676
160	-21,788	-46,561	-88,254	313,085	178,254	12:00:00	38940	688,924
161	-21,788	-46,561	-88,168	313,858	178,168	12:00:00	38940	687,236
162	-21,788	-46,561	-88,087	314,399	178,087	12:00:00	38880	685,613
163	-21,788	-46,561	-88,009	314,742	178,009	12:00:00	38940	684,058
164	-21,788	-46,561	-87,936	314,911	177,936	12:00:00	38940	682,571
165	-21,788	-46,561	-87,867	314,927	177,867	12:00:00	38940	681,155
166	-21,788	-46,561	-87,803	314,807	177,803	12:00:00	38880	679,812
167	-21,788	-46,561	-87,742	314,562	177,742	12:00:00	38880	678,543
168	-21,788	-46,561	-87,687	314,203	177,687	12:00:00	38880	677,350
169	-21,788	-46,561	-87,636	313,738	177,636	12:00:00	38820	676,234
170	-21,788	-46,561	-87,589	313,171	177,589	12:00:00	38880	675,198
171	-21,788	-46,561	-87,547	312,506	177,547	12:00:00	38880	674,242
172	-21,788	-46,561	-87,509	311,747	177,509	12:00:00	38880	673,367
173	-21,788	-46,561	-87,475	310,896	177,475	12:00:00	38880	672,576
174	-21,788	-46,561	-87,446	309,952	177,446	12:00:00	38880	671,868
175	-21,788	-46,561	-87,420	308,916	177,420	12:00:00	38880	671,245
176	-21,788	-46,561	-87,398	307,787	177,398	12:00:00	38880	670,708
177	-21,788	-46,561	-87,379	306,563	177,379	12:00:00	38880	670,258
178	-21,788	-46,561	-87,363	305,244	177,363	12:00:00	38940	669,895
179	-21,788	-46,561	-87,351	303,825	177,351	12:00:00	38940	669,620
180	-21,788	-46,561	-87,341	302,305	177,341	12:00:00	38940	669,434
181	-21,788	-46,561	-87,333	300,681	177,333	12:00:00	39000	669,370
182	-21,788	-46,561	-87,327	298,950	177,327	12:00:00	38940	669,361
183	-21,788	-46,561	-87,323	297,108	177,323	12:00:00	38940	669,441
184	-21,788	-46,561	-87,319	295,154	177,319	12:00:00	39000	669,610
185	-21,788	-46,561	-87,316	293,085	177,316	12:00:00	39000	669,869
186	-21,788	-46,561	-87,313	290,899	177,313	12:00:00	39000	670,216

187	-21,788	-46,561	-87,309	288,596	177,309	12:00:00	39060	670,652
188	-21,788	-46,561	-87,304	286,177	177,304	12:00:00	39060	671,177
189	-21,788	-46,561	-87,297	283,643	177,297	12:00:00	39120	671,790
190	-21,788	-46,561	-87,287	280,999	177,287	12:00:00	39180	672,490
191	-21,788	-46,561	-87,273	278,251	177,273	12:00:00	39180	673,277
192	-21,788	-46,561	-87,255	275,406	177,255	12:00:00	39240	674,151
193	-21,788	-46,561	-87,233	272,474	177,233	12:00:00	39240	675,110
194	-21,788	-46,561	-87,205	269,469	177,205	12:00:00	39240	676,155
195	-21,788	-46,561	-87,171	266,403	177,171	12:00:00	39300	677,285
196	-21,788	-46,561	-87,129	263,293	177,129	12:00:00	39300	678,498
197	-21,788	-46,561	-87,081	260,156	177,081	12:00:00	39420	679,794
198	-21,788	-46,561	-87,024	257,008	177,024	12:00:00	39420	681,172
199	-21,788	-46,561	-86,959	253,868	176,959	12:00:00	39420	682,631
200	-21,788	-46,561	-86,886	250,753	176,886	12:00:00	39480	684,169
201	-21,788	-46,561	-86,803	247,679	176,803	12:00:00	39540	685,785
202	-21,788	-46,561	-86,711	244,661	176,711	12:00:00	39600	687,478
203	-21,788	-46,561	-86,609	241,712	176,609	12:00:00	39600	689,246
204	-21,788	-46,561	-86,499	238,844	176,499	12:00:00	39660	691,088
205	-21,788	-46,561	-86,378	236,064	176,378	12:00:00	39720	693,001
206	-21,788	-46,561	-86,248	233,380	176,248	12:00:00	39720	694,985
207	-21,788	-46,561	-86,109	230,798	176,109	12:00:00	39780	697,037
208	-21,788	-46,561	-85,961	228,318	175,961	12:00:00	39840	699,156
209	-21,788	-46,561	-85,804	225,944	175,804	12:00:00	39900	701,340
210	-21,788	-46,561	-85,638	223,675	175,638	12:00:00	39960	703,587
211	-21,788	-46,561	-85,463	221,510	175,463	12:00:00	40020	705,894
212	-21,788	-46,561	-85,280	219,447	175,280	12:00:00	40020	708,260
213	-21,788	-46,561	-85,089	217,482	175,089	12:00:00	40140	710,683
214	-21,788	-46,561	-84,890	215,613	174,890	12:00:00	40140	713,161
215	-21,788	-46,561	-84,684	213,835	174,684	12:00:00	40260	715,691
216	-21,788	-46,561	-84,469	212,145	174,469	12:00:00	40260	718,271
217	-21,788	-46,561	-84,248	210,538	174,248	12:00:00	40320	720,900
218	-21,788	-46,561	-84,019	209,011	174,019	12:00:00	40380	723,575
219	-21,788	-46,561	-83,784	207,559	173,784	12:00:00	40440	726,295
220	-21,788	-46,561	-83,542	206,177	173,542	12:00:00	40500	729,056
221	-21,788	-46,561	-83,293	204,862	173,293	12:00:00	40560	731,858
222	-21,788	-46,561	-83,039	203,611	173,039	12:00:00	40620	734,699
223	-21,788	-46,561	-82,778	202,418	172,778	12:00:00	40680	737,576
224	-21,788	-46,561	-82,512	201,281	172,512	12:00:00	40740	740,488
225	-21,788	-46,561	-82,240	200,196	172,240	12:00:00	40800	743,433
226	-21,788	-46,561	-81,962	199,160	171,962	12:00:00	40860	746,408
227	-21,788	-46,561	-81,679	198,170	171,679	12:00:00	40920	749,413
228	-21,788	-46,561	-81,391	197,224	171,391	12:00:00	40980	752,446
229	-21,788	-46,561	-81,098	196,319	171,098	12:00:00	41040	755,503
230	-21,788	-46,561	-80,800	195,452	170,800	12:00:00	41160	758,585
231	-21,788	-46,561	-80,498	194,621	170,498	12:00:00	41220	761,689
232	-21,788	-46,561	-80,191	193,824	170,191	12:00:00	41220	764,812
233	-21,788	-46,561	-79,879	193,060	169,879	12:00:00	41340	767,955
234	-21,788	-46,561	-79,563	192,326	169,563	12:00:00	41400	771,114
235	-21,788	-46,561	-79,244	191,621	169,244	12:00:00	41460	774,288

236	-21,788	-46,561	-78,920	190,944	168,920	12:00:00	41580	777,475
237	-21,788	-46,561	-78,592	190,293	168,592	12:00:00	41580	780,675
238	-21,788	-46,561	-78,260	189,667	168,260	12:00:00	41640	783,885
239	-21,788	-46,561	-77,925	189,064	167,925	12:00:00	41700	787,103
240	-21,788	-46,561	-77,586	188,484	167,586	12:00:00	41820	790,329
241	-21,788	-46,561	-77,244	187,926	167,244	12:00:00	41880	793,560
242	-21,788	-46,561	-76,898	187,388	166,898	12:00:00	41940	796,795
243	-21,788	-46,561	-76,549	186,870	166,549	12:00:00	42060	800,033
244	-21,788	-46,561	-76,197	186,371	166,197	12:00:00	42120	803,272
245	-21,788	-46,561	-75,842	185,891	165,842	12:00:00	42180	806,511
246	-21,788	-46,561	-75,485	185,428	165,485	12:00:00	42180	809,748
247	-21,788	-46,561	-75,124	184,983	165,124	12:00:00	42300	812,982
248	-21,788	-46,561	-74,761	184,553	164,761	12:00:00	42360	816,213
249	-21,788	-46,561	-74,395	184,139	164,395	12:00:00	42420	819,439
250	-21,788	-46,561	-74,027	183,740	164,027	12:00:00	42480	822,658
251	-21,788	-46,561	-73,657	183,356	163,657	12:00:00	42600	825,871
252	-21,788	-46,561	-73,284	182,985	163,284	12:00:00	42660	829,075
253	-21,788	-46,561	-72,909	182,627	162,909	12:00:00	42720	832,270
254	-21,788	-46,561	-72,533	182,283	162,533	12:00:00	42780	835,456
255	-21,788	-46,561	-72,155	181,950	162,155	12:00:00	42900	838,630
256	-21,788	-46,561	-71,775	181,630	161,775	12:00:00	42960	841,793
257	-21,788	-46,561	-71,393	181,321	161,393	12:00:00	43020	844,944
258	-21,788	-46,561	-71,010	181,023	161,010	12:00:00	43080	848,081
259	-21,788	-46,561	-70,626	180,736	160,626	12:00:00	43200	851,203
260	-21,788	-46,561	-70,240	180,459	160,240	12:00:00	43260	854,311
261	-21,788	-46,561	-69,853	180,192	159,853	12:00:00	43320	857,403
262	-21,788	-46,561	-69,466	179,935	159,466	12:00:00	43440	860,479
263	-21,788	-46,561	-69,077	179,688	159,077	12:00:00	43500	863,538
264	-21,788	-46,561	-68,688	179,449	158,688	12:00:00	43560	866,579
265	-21,788	-46,561	-68,298	179,220	158,298	12:00:00	43620	869,601
266	-21,788	-46,561	-67,908	179,000	157,908	12:00:00	43740	872,605
267	-21,788	-46,561	-67,517	178,788	157,517	12:00:00	43800	875,588
268	-21,788	-46,561	-67,126	178,584	157,126	12:00:00	43860	878,552
269	-21,788	-46,561	-66,735	178,389	156,735	12:00:00	43920	881,494
270	-21,788	-46,561	-66,343	178,202	156,343	12:00:00	44040	884,415
271	-21,788	-46,561	-65,952	178,023	155,952	12:00:00	44100	887,313
272	-21,788	-46,561	-65,561	177,852	155,561	12:00:00	44100	890,188
273	-21,788	-46,561	-65,170	177,689	155,170	12:00:00	44220	893,040
274	-21,788	-46,561	-64,780	177,534	154,780	12:00:00	44280	895,868
275	-21,788	-46,561	-64,390	177,386	154,390	12:00:00	44340	898,672
276	-21,788	-46,561	-64,001	177,245	154,001	12:00:00	44460	901,450
277	-21,788	-46,561	-63,612	177,112	153,612	12:00:00	44520	904,203
278	-21,788	-46,561	-63,224	176,986	153,224	12:00:00	44580	906,930
279	-21,788	-46,561	-62,837	176,867	152,837	12:00:00	44700	909,632
280	-21,788	-46,561	-62,452	176,755	152,452	12:00:00	44760	912,306
281	-21,788	-46,561	-62,067	176,649	152,067	12:00:00	44820	914,954
282	-21,788	-46,561	-61,684	176,551	151,684	12:00:00	44940	917,574
283	-21,788	-46,561	-61,303	176,458	151,303	12:00:00	45000	920,167
284	-21,788	-46,561	-60,923	176,372	150,923	12:00:00	45000	922,732

285	-21,788	-46,561	-60,544	176,293	150,544	12:00:00	45120	925,268
286	-21,788	-46,561	-60,168	176,219	150,168	12:00:00	45180	927,777
287	-21,788	-46,561	-59,793	176,151	149,793	12:00:00	45240	930,256
288	-21,788	-46,561	-59,421	176,090	149,421	12:00:00	45360	932,706
289	-21,788	-46,561	-59,050	176,034	149,050	12:00:00	45420	935,126
290	-21,788	-46,561	-58,682	175,983	148,682	12:00:00	45480	937,518
291	-21,788	-46,561	-58,316	175,938	148,316	12:00:00	45540	939,879
292	-21,788	-46,561	-57,953	175,899	147,953	12:00:00	45600	942,210
293	-21,788	-46,561	-57,593	175,865	147,593	12:00:00	45720	944,510
294	-21,788	-46,561	-57,235	175,836	147,235	12:00:00	45780	946,780
295	-21,788	-46,561	-56,881	175,813	146,881	12:00:00	45840	949,019
296	-21,788	-46,561	-56,529	175,794	146,529	12:00:00	45900	951,226
297	-21,788	-46,561	-56,180	175,781	146,180	12:00:00	46020	953,402
298	-21,788	-46,561	-55,835	175,773	145,835	12:00:00	46020	955,547
299	-21,788	-46,561	-55,493	175,770	145,493	12:00:00	46140	957,659
300	-21,788	-46,561	-55,155	175,772	145,155	12:00:00	46200	959,738
301	-21,788	-46,561	-54,820	175,779	144,820	12:00:00	46260	961,785
302	-21,788	-46,561	-54,488	175,790	144,488	12:00:00	46320	963,799
303	-21,788	-46,561	-54,161	175,807	144,161	12:00:00	46440	965,780
304	-21,788	-46,561	-53,837	175,828	143,837	12:00:00	46440	967,727
305	-21,788	-46,561	-53,518	175,854	143,518	12:00:00	46560	969,640
306	-21,788	-46,561	-53,202	175,885	143,202	12:00:00	46560	971,520
307	-21,788	-46,561	-52,891	175,921	142,891	12:00:00	46680	973,365
308	-21,788	-46,561	-52,584	175,961	142,584	12:00:00	46680	975,177
309	-21,788	-46,561	-52,282	176,005	142,282	12:00:00	46800	976,954
310	-21,788	-46,561	-51,984	176,054	141,984	12:00:00	46860	978,696
311	-21,788	-46,561	-51,691	176,107	141,691	12:00:00	46920	980,404
312	-21,788	-46,561	-51,403	176,165	141,403	12:00:00	46980	982,077
313	-21,788	-46,561	-51,119	176,226	141,119	12:00:00	47040	983,714
314	-21,788	-46,561	-50,841	176,292	140,841	12:00:00	47100	985,317
315	-21,788	-46,561	-50,568	176,362	140,568	12:00:00	47160	986,885
316	-21,788	-46,561	-50,299	176,435	140,299	12:00:00	47220	988,417
317	-21,788	-46,561	-50,036	176,512	140,036	12:00:00	47280	989,913
318	-21,788	-46,561	-49,779	176,593	139,779	12:00:00	47280	991,375
319	-21,788	-46,561	-49,527	176,677	139,527	12:00:00	47400	992,800
320	-21,788	-46,561	-49,281	176,765	139,281	12:00:00	47460	994,190
321	-21,788	-46,561	-49,040	176,856	139,040	12:00:00	47460	995,544
322	-21,788	-46,561	-48,805	176,951	138,805	12:00:00	47580	996,863
323	-21,788	-46,561	-48,576	177,048	138,576	12:00:00	47640	998,145
324	-21,788	-46,561	-48,353	177,149	138,353	12:00:00	47640	999,391
325	-21,788	-46,561	-48,136	177,253	138,136	12:00:00	47700	1000,602
326	-21,788	-46,561	-47,925	177,360	137,925	12:00:00	47760	1001,775
327	-21,788	-46,561	-47,720	177,470	137,720	12:00:00	47760	1002,913
328	-21,788	-46,561	-47,521	177,583	137,521	12:00:00	47820	1004,013
329	-21,788	-46,561	-47,329	177,698	137,329	12:00:00	47880	1005,077
330	-21,788	-46,561	-47,144	177,817	137,144	12:00:00	47880	1006,105
331	-21,788	-46,561	-46,965	177,938	136,965	12:00:00	47940	1007,095
332	-21,788	-46,561	-46,792	178,062	136,792	12:00:00	48000	1008,048
333	-21,788	-46,561	-46,626	178,188	136,626	12:00:00	48000	1008,964

334	-21,788	-46,561	-46,467	178,317	136,467	12:00:00	48060	1009,843
335	-21,788	-46,561	-46,314	178,449	136,314	12:00:00	48120	1010,685
336	-21,788	-46,561	-46,169	178,583	136,169	12:00:00	48120	1011,490
337	-21,788	-46,561	-46,030	178,719	136,030	12:00:00	48180	1012,257
338	-21,788	-46,561	-45,898	178,857	135,898	12:00:00	48240	1012,988
339	-21,788	-46,561	-45,774	178,998	135,774	12:00:00	48240	1013,680
340	-21,788	-46,561	-45,656	179,140	135,656	12:00:00	48300	1014,336
341	-21,788	-46,561	-45,546	179,285	135,546	12:00:00	48300	1014,954
342	-21,788	-46,561	-45,442	179,431	135,442	12:00:00	48300	1015,535
343	-21,788	-46,561	-45,346	179,579	135,346	12:00:00	48360	1016,079
344	-21,788	-46,561	-45,257	179,729	135,257	12:00:00	48360	1016,586
345	-21,788	-46,561	-45,176	179,880	135,176	12:00:00	48360	1017,056
346	-21,788	-46,561	-45,101	180,032	135,101	12:00:00	48420	1017,489
347	-21,788	-46,561	-45,034	180,186	135,034	12:00:00	48420	1017,885
348	-21,788	-46,561	-44,975	180,341	134,975	12:00:00	48420	1018,244
349	-21,788	-46,561	-44,922	180,497	134,922	12:00:00	48420	1018,567
350	-21,788	-46,561	-44,878	180,653	134,878	12:00:00	48420	1018,853
351	-21,788	-46,561	-44,840	180,811	134,840	12:00:00	48480	1019,103
352	-21,788	-46,561	-44,811	180,969	134,811	12:00:00	48480	1019,316
353	-21,788	-46,561	-44,788	181,128	134,788	12:00:00	48480	1019,493
354	-21,788	-46,561	-44,773	181,288	134,773	12:00:00	48480	1019,634
355	-21,788	-46,561	-44,766	181,448	134,766	12:00:00	48480	1019,739
356	-21,788	-46,561	-44,766	181,608	134,766	12:00:00	48480	1019,808
357	-21,788	-46,561	-44,774	181,769	134,774	12:00:00	48480	1019,841
358	-21,788	-46,561	-44,790	181,929	134,790	12:00:00	48480	1019,838
359	-21,788	-46,561	-44,812	182,091	134,812	12:00:00	48420	1019,798
360	-21,788	-46,561	-44,843	182,252	134,843	12:00:00	48420	1019,723
361	-21,788	-46,561	-44,880	182,413	134,880	12:00:00	48420	1019,613
362	-21,788	-46,561	-44,926	182,574	134,926	12:00:00	48420	1019,466
363	-21,788	-46,561	-44,979	182,735	134,979	12:00:00	48420	1019,284
364	-21,788	-46,561	-45,039	182,896	135,039	12:00:00	48360	1019,067

air_temperature	clarity_index	cloud_cover	global_irradiance	humidity	precipitation	temperature	wind_speed
27,438	0,600	5,000	225,224	75,059	0,153	27,938	2,899
25,800	0,580	4,000	234,197	75,351	0,347	27,758	2,814
26,250	0,270	8,000	111,280	79,629	0,267	27,116	2,087
26,875	0,560	6,333	219,572	77,865	1,263	27,509	2,356
26,625	0,560	6,000	246,497	80,747	9,665	27,462	2,316
25,750	0,440	6,667	181,201	79,046	0,027	27,255	1,866
27,688	0,560	4,000	237,593	77,844	0,823	27,803	2,918
27,929	0,600	4,000	248,965	73,291	0,133	28,029	3,427
27,500	0,580	4,333	234,755	75,188	0,020	27,911	2,847
26,125	0,520	6,000	191,894	80,905	0,440	27,860	2,493
27,357	0,510	7,667	211,007	81,643	3,167	27,784	3,140
28,833	0,550	6,333	247,147	72,400	0,043	28,314	3,569
29,500	0,580	4,333	247,036	73,490	0,037	28,362	3,688
28,300	0,360	9,000	153,299	74,295	0,020	28,050	3,219
28,750	0,460	6,000	217,850	73,204	0,013	27,980	3,292
28,700	0,570	4,333	247,315	72,956	0,090	28,132	3,121

28,750	0,630	4,333	248,949	72,247	0,380	27,730	3,292
27,071	0,470	7,000	193,492	77,660	1,169	28,105	3,082
27,093	0,430	6,833	177,579	75,634	0,481	28,186	2,477
27,093	0,560	5,000	241,069	69,956	0,057	28,337	3,279
27,093	0,590	5,333	240,838	71,751	0,247	28,014	2,853
27,093	0,550	5,667	211,446	77,226	0,544	27,962	1,946
29,875	0,390	7,333	155,547	75,288	0,583	27,715	3,153
27,093	0,550	8,000	206,167	74,524	1,054	28,423	2,415
27,093	0,340	7,333	141,457	74,112	0,384	27,876	2,603
27,093	0,430	10,000	191,008	72,496	0,134	28,227	2,135
25,600	0,660	3,333	271,262	72,757	0,063	27,853	2,633
28,143	0,590	5,667	227,928	71,582	0,037	28,626	2,960
28,357	0,600	5,333	231,320	69,666	0,010	28,305	3,231
27,875	0,600	4,333	230,294	70,581	0,010	28,385	3,375
27,688	0,550	5,000	222,670	72,130	0,167	28,352	3,123
27,938	0,460	6,333	200,783	73,961	0,957	28,373	3,110
27,625	0,460	8,333	201,595	79,469	4,883	27,870	2,620
26,357	0,290	9,000	137,603	82,309	4,448	26,785	1,944
27,438	0,590	8,000	237,211	78,828	2,258	27,443	2,584
26,750	0,460	8,667	177,593	80,267	3,843	26,939	2,426
26,125	0,500	8,333	197,143	81,434	0,993	27,433	2,487
27,250	0,490	8,000	211,263	82,511	4,023	27,235	2,367
28,143	0,590	7,333	232,043	75,232	0,413	28,480	2,588
29,375	0,650	5,000	239,342	72,931	0,237	28,679	3,033
28,500	0,480	8,000	191,627	75,602	0,640	28,470	2,919
28,700	0,530	6,667	202,993	73,049	0,687	28,306	3,030
27,900	0,560	6,000	245,421	72,693	0,280	28,463	3,150
28,643	0,570	4,333	250,505	70,919	0,040	29,037	3,451
27,833	0,580	5,667	226,219	71,879	0,070	28,792	3,190
28,000	0,550	5,000	226,190	72,626	0,620	28,726	3,131
27,250	0,490	5,000	207,998	74,735	1,053	28,464	3,214
28,400	0,530	6,000	215,908	75,821	0,737	28,464	3,446
27,125	0,270	8,000	150,979	81,289	1,237	28,202	2,797
24,625	0,150	10,000	78,518	91,305	11,923	28,912	1,873
24,800	0,360	9,000	143,887	91,659	5,920	26,424	1,746
24,375	0,400	9,667	149,778	93,338	6,377	26,591	1,433
25,083	0,350	9,333	155,537	87,501	3,613	26,629	1,481
26,125	0,470	7,000	203,835	85,201	4,268	26,847	1,721
25,438	0,360	8,667	159,236	90,561	10,245	27,042	1,758
26,500	0,380	6,667	170,270	87,379	0,868	27,264	2,272
27,429	0,550	9,000	205,211	85,865	0,728	27,535	2,479
27,214	0,530	5,333	229,356	85,009	0,586	27,765	2,491
26,688	0,430	7,667	166,118	87,518	1,143	27,393	1,652
25,750	0,370	8,667	151,227	90,647	3,277	27,216	1,739
25,688	0,600	8,667	212,312	90,634	5,261	26,849	2,261
25,313	0,260	10,000	123,361	92,723	10,075	26,700	1,615
25,813	0,490	6,333	200,856	87,633	1,172	26,747	1,992
24,875	0,430	10,000	146,393	92,234	2,668	26,784	1,866
25,786	0,400	9,333	152,311	93,113	15,298	26,533	1,447

27,093	0,080	10,000	45,011	96,735	9,666	28,450	1,662
26,357	0,420	7,667	192,638	90,880	4,861	26,345	1,939
25,688	0,480	9,000	181,520	93,038	2,038	26,994	1,775
26,250	0,520	6,333	197,163	89,837	6,077	27,336	2,000
26,563	0,540	7,667	226,766	88,667	4,233	27,444	2,067
27,875	0,680	5,000	272,178	81,996	0,228	27,861	2,412
27,125	0,490	6,333	189,646	83,405	0,570	27,992	2,520
27,813	0,610	5,667	236,351	80,758	0,347	28,055	2,655
27,917	0,550	6,667	228,000	83,256	0,673	28,325	2,648
27,000	0,580	8,667	246,325	85,670	2,348	27,572	2,328
27,375	0,460	7,667	195,157	83,313	1,020	27,528	2,492
26,875	0,250	9,333	102,741	86,136	1,635	27,403	1,609
26,300	0,360	8,333	148,995	88,856	1,513	27,416	2,086
25,750	0,120	8,000	53,012	93,700	7,653	26,774	1,441
25,375	0,400	10,000	160,047	93,190	6,750	26,465	1,749
24,563	0,190	9,000	76,508	97,847	1,882	26,739	2,360
24,563	0,500	8,667	200,142	92,110	6,827	26,072	1,413
24,813	0,320	8,333	123,392	94,844	6,180	26,560	1,449
25,688	0,560	8,333	232,320	92,113	6,162	26,529	1,373
25,688	0,490	7,333	183,581	93,938	6,227	26,973	1,656
26,625	0,570	7,667	222,469	90,736	11,557	27,126	1,983
27,188	0,590	7,000	237,173	87,779	0,800	27,338	2,250
26,875	0,650	6,000	267,478	85,678	2,082	27,438	2,237
26,250	0,570	7,000	214,344	87,363	2,432	27,320	1,906
26,286	0,450	9,667	177,914	92,057	9,456	27,156	1,706
25,438	0,310	8,667	135,741	93,316	8,912	26,857	1,589
26,375	0,420	9,333	201,570	89,749	12,430	26,644	1,703
26,188	0,550	8,333	215,321	89,735	4,569	26,990	1,955
25,313	0,400	10,000	167,079	88,982	6,098	26,824	1,702
25,000	0,310	8,333	154,358	90,267	9,931	26,243	1,068
25,938	0,640	6,333	259,274	85,587	3,150	27,120	2,198
26,625	0,580	9,000	209,810	87,046	3,035	26,987	1,598
25,750	0,490	10,000	178,642	89,738	6,603	26,880	1,753
25,938	0,410	8,333	167,439	89,384	4,486	26,865	1,546
25,167	0,320	10,000	128,200	92,363	2,600	26,376	1,384
26,286	0,640	7,000	237,159	87,989	8,793	26,667	0,905
25,188	0,310	9,000	122,128	92,324	7,888	26,293	1,263
26,500	0,520	8,333	190,179	87,535	6,857	26,380	1,213
25,625	0,240	8,000	137,872	88,047	2,863	26,604	1,019
25,500	0,530	6,667	186,080	91,375	1,470	26,592	1,070
25,875	0,510	8,667	184,573	90,789	14,402	26,838	1,568
25,813	0,630	6,000	222,955	87,976	5,155	27,040	2,433
25,938	0,480	6,667	156,610	90,279	12,955	26,849	1,723
26,000	0,510	8,000	173,993	90,212	10,528	26,318	1,635
25,000	0,540	10,000	179,479	89,865	13,084	25,850	1,590
26,250	0,370	7,667	168,372	87,976	11,347	26,679	1,700
27,093	0,562	7,667	205,575	87,693	3,522	26,876	1,545
26,563	0,480	6,333	210,438	86,890	6,725	26,937	1,922
26,938	0,420	6,000	205,878	83,932	2,427	27,012	2,008

26,875	0,590	8,000	227,831	83,436	1,070	27,215	2,019
26,688	0,640	6,333	236,336	82,926	0,570	27,456	2,151
26,875	0,510	5,667	184,399	86,865	1,702	27,547	1,786
27,500	0,650	5,667	251,884	82,352	3,998	27,877	2,501
27,938	0,560	7,000	227,608	80,380	0,453	27,891	2,651
27,750	0,630	4,667	240,942	79,913	0,120	28,001	2,102
27,357	0,550	6,667	193,887	81,364	0,623	27,913	2,114
26,286	0,510	7,333	175,373	88,356	3,050	27,356	1,484
24,438	0,380	10,000	129,306	94,061	1,882	26,137	1,002
26,688	0,460	8,667	200,372	85,678	4,220	26,888	1,360
27,250	0,590	6,000	236,127	83,642	0,567	27,366	1,692
25,833	0,580	5,000	207,252	84,574	3,073	27,087	1,753
26,214	0,530	7,000	181,870	86,287	2,177	26,796	1,533
26,438	0,530	8,000	179,854	88,701	4,720	26,926	1,332
25,750	0,240	7,000	105,327	89,881	5,612	26,724	1,803
25,375	0,610	6,333	222,181	86,751	5,437	26,713	2,107
25,900	0,610	6,333	224,040	85,146	0,983	27,137	2,333
27,125	0,640	2,333	245,032	82,553	0,147	27,561	2,266
26,688	0,630	5,667	242,182	83,105	0,080	27,687	2,019
27,188	0,620	5,000	234,213	81,885	1,473	27,275	2,088
27,438	0,520	7,000	216,579	83,801	2,827	27,607	1,913
27,063	0,510	3,333	193,928	84,577	0,782	27,685	1,804
25,438	0,370	8,000	147,096	89,696	4,315	27,197	1,085
26,063	0,450	8,667	170,453	88,963	9,742	27,155	1,223
26,625	0,530	7,667	197,087	87,693	2,262	27,398	1,855
26,313	0,470	7,667	168,413	89,044	2,500	27,080	1,551
26,125	0,600	5,333	216,727	88,311	14,727	27,058	1,402
27,000	0,590	7,000	214,932	87,344	8,447	27,067	1,949
26,375	0,530	8,333	189,620	86,606	7,065	27,208	1,332
27,063	0,610	4,667	234,054	85,943	3,067	28,627	1,422
25,875	0,410	8,000	142,098	92,086	14,810	28,181	1,397
25,500	0,480	8,333	176,140	90,427	9,893	27,424	1,723
26,938	0,470	7,667	169,487	85,027	1,930	27,129	1,502
25,813	0,590	8,000	205,004	88,704	0,383	28,694	1,045
26,875	0,550	5,000	215,009	85,526	1,790	27,435	1,399
26,875	0,650	5,000	244,311	83,965	1,197	27,574	1,300
26,500	0,280	7,333	107,686	88,869	2,670	27,465	1,346
26,563	0,640	5,333	122,648	86,935	7,540	27,531	1,501
26,688	0,580	7,000	110,524	85,148	0,463	28,707	1,466
27,438	0,660	6,000	123,453	83,086	2,587	28,392	1,597
26,063	0,620	6,333	115,716	87,783	5,302	28,235	1,424
24,625	0,250	8,333	48,713	87,790	5,262	28,498	1,515
25,813	0,610	6,000	113,882	86,377	1,961	28,559	1,503
26,000	0,590	7,000	112,550	87,129	6,072	28,836	2,148
26,813	0,548	8,000	113,950	84,601	2,353	28,420	1,668
25,250	0,620	8,000	114,652	84,287	1,558	27,837	1,580
25,875	0,550	6,500	104,934	80,962	0,560	28,666	1,765
26,188	0,620	8,000	116,682	86,154	1,660	28,988	1,218
24,563	0,560	8,500	107,604	87,006	1,787	29,029	1,855

27,250	0,550	7,000	103,322	83,998	1,120	28,936	1,997
26,625	0,610	7,000	113,886	82,917	3,368	29,070	1,870
27,313	0,710	3,000	129,016	81,457	0,410	29,256	1,643
26,438	0,610	6,500	116,578	82,313	0,093	29,068	1,582
27,313	0,640	7,000	120,723	81,821	0,093	29,158	1,783
26,875	0,620	5,500	119,223	79,690	0,020	28,682	1,728
25,938	0,554	5,500	114,035	80,443	0,830	27,151	1,422
26,438	0,630	6,000	118,052	82,524	2,415	27,651	2,054
26,063	0,380	6,000	72,728	85,193	1,677	27,357	1,773
26,500	0,640	7,000	118,931	82,586	7,747	26,821	1,910
26,563	0,620	6,000	114,286	81,938	0,835	27,233	1,755
26,929	0,560	7,000	104,521	81,789	0,387	27,287	1,745
26,250	0,630	5,000	117,883	79,963	1,187	27,005	1,917
25,625	0,610	7,000	114,107	83,456	10,088	26,562	1,720
25,313	0,540	7,500	99,676	80,342	0,287	26,493	1,313
25,850	0,590	7,500	111,275	81,285	0,207	26,818	1,575
26,500	0,610	7,000	115,043	81,582	0,857	27,053	2,010
26,500	0,500	9,000	94,689	83,117	1,470	27,234	1,595
26,375	0,540	7,000	197,539	82,250	3,632	27,464	1,658
26,125	0,290	9,000	139,213	85,051	10,423	27,005	1,695
27,000	0,460	7,500	170,884	83,314	2,547	27,177	1,773
25,375	0,590	9,000	194,637	83,870	3,257	26,714	1,420
26,000	0,570	7,000	191,510	83,126	1,780	26,926	1,514
26,000	0,590	7,500	197,551	82,792	5,137	27,005	1,659
27,563	0,620	7,000	201,931	80,809	0,223	27,408	1,487
26,563	0,620	6,500	201,510	80,023	0,233	27,373	1,838
26,688	0,600	6,000	196,090	79,967	0,133	27,375	1,998
26,875	0,630	5,000	203,178	78,623	0,000	27,424	2,090
27,063	0,640	3,000	202,970	77,619	0,000	27,374	2,082
26,875	0,640	3,500	202,971	77,538	0,000	27,351	2,841
27,188	0,640	4,000	206,921	76,554	2,363	27,315	2,258
27,125	0,660	2,000	212,546	75,797	0,570	27,378	1,843
26,875	0,680	3,000	211,301	74,907	0,057	27,399	1,935
26,250	0,600	6,500	199,422	77,431	0,103	27,347	1,500
26,688	0,600	7,750	197,970	79,627	0,387	26,968	1,771
26,563	0,600	8,000	199,223	79,434	2,307	27,014	1,632
26,625	0,530	8,500	184,218	76,882	3,117	27,343	2,573
26,313	0,480	8,000	175,674	77,035	0,103	27,198	1,828
26,875	0,630	6,500	205,258	79,416	0,767	27,270	1,770
26,938	0,610	5,500	199,223	79,656	0,277	27,578	2,013
26,875	0,580	7,000	195,258	79,729	2,387	27,378	1,913
26,813	0,590	6,500	197,344	76,356	0,243	27,521	2,038
26,313	0,420	9,000	165,261	83,863	1,923	26,901	1,373
26,750	0,570	6,500	193,589	80,582	4,753	27,047	1,348
25,813	0,550	5,500	192,348	79,261	0,120	27,140	1,860
27,063	0,600	5,500	204,009	78,973	0,293	27,559	2,008
27,093	0,630	5,000	204,012	79,790	0,437	27,867	2,781
27,500	0,630	2,000	208,805	77,245	0,901	27,520	1,965
26,750	0,680	4,000	214,634	76,766	4,247	27,649	1,850

26,375	0,630	4,500	250,057	75,829	0,035	27,235	1,570
26,438	0,670	3,500	257,899	75,989	0,057	27,122	1,880
26,917	0,580	5,000	222,708	77,525	0,153	27,370	1,995
27,313	0,630	6,000	234,927	75,926	0,053	27,701	2,302
27,188	0,620	4,000	256,879	75,230	0,013	27,759	1,880
27,563	0,620	3,500	242,570	75,607	0,013	27,793	2,035
27,125	0,630	4,500	236,358	75,700	0,010	27,821	2,133
27,188	0,630	3,500	248,104	74,274	0,000	27,859	2,138
26,938	0,640	3,000	255,975	74,576	0,027	27,735	2,068
27,125	0,590	6,000	217,551	75,796	0,067	27,891	1,987
26,813	0,600	7,500	225,892	76,045	0,030	27,839	1,930
27,250	0,600	4,500	243,809	72,227	0,067	27,918	1,840
26,750	0,660	5,500	247,297	72,214	0,060	27,625	2,480
26,688	0,590	3,500	225,436	75,167	0,020	27,437	1,860
27,063	0,630	5,500	241,570	73,591	0,103	27,928	1,797
27,125	0,650	1,500	247,630	74,855	0,010	28,464	1,910
27,938	0,640	4,500	250,325	73,677	0,547	28,659	1,746
27,438	0,660	4,000	260,266	72,836	0,090	28,648	1,878
26,938	0,640	4,500	229,227	75,777	0,413	28,025	1,700
26,938	0,660	7,000	240,954	75,122	0,543	27,963	2,813
26,813	0,670	5,500	267,871	73,362	0,120	27,315	2,213
26,625	0,640	6,000	232,888	72,341	0,073	27,854	2,113
26,688	0,570	8,000	207,181	76,811	0,013	28,478	1,925
27,000	0,590	8,000	207,602	76,311	0,023	28,566	2,333
27,375	0,560	6,500	215,646	73,382	0,013	28,312	3,024
27,786	0,660	4,500	264,645	71,989	0,003	28,365	2,638
27,313	0,620	5,000	249,365	74,290	0,027	27,914	2,323
28,000	0,570	5,000	224,938	73,844	0,093	28,308	2,688
27,313	0,610	4,500	238,130	71,690	0,257	28,380	2,592
27,357	0,610	3,500	247,805	71,159	0,027	28,333	2,805
27,093	0,590	8,000	209,590	75,371	0,197	28,018	2,492
26,700	0,590	8,000	203,928	76,596	0,563	28,193	2,173
27,093	0,600	5,500	235,740	76,422	0,377	28,329	2,967
27,093	0,610	5,500	239,356	74,682	0,157	28,368	2,652
27,093	0,590	7,000	241,393	75,153	0,121	28,452	2,660
27,093	0,610	7,500	225,043	75,187	0,107	28,204	2,788
27,093	0,570	5,500	225,358	74,153	0,107	28,424	2,690
27,093	0,570	6,500	228,285	74,175	0,114	28,298	2,387
27,093	0,600	4,500	260,859	71,820	0,104	28,142	2,762
28,833	0,620	5,500	246,442	71,248	0,003	28,193	2,683
27,313	0,640	3,000	265,224	71,949	0,013	28,314	2,917
27,500	0,600	3,000	243,679	71,542	0,123	28,655	2,923
27,188	0,620	3,500	254,732	73,774	0,003	28,711	2,380
27,125	0,620	4,500	260,328	73,576	0,073	28,331	2,850
27,063	0,620	6,000	244,705	73,985	0,340	28,314	2,515
27,875	0,640	3,500	271,593	70,709	0,010	28,957	2,470
27,750	0,610	5,000	258,204	70,944	0,018	29,084	2,295
27,188	0,640	3,000	259,396	71,513	0,013	28,867	2,673
27,375	0,670	3,000	279,731	71,470	0,000	28,682	2,805

27,750	0,580	4,500	248,020	71,093	0,008	28,795	2,538
27,375	0,550	4,500	241,441	71,161	0,000	28,779	2,775
27,375	0,650	3,500	259,228	71,377	0,008	28,771	2,615
27,563	0,550	7,000	223,986	72,425	0,000	29,090	2,695
27,750	0,610	4,000	268,499	69,929	0,000	28,780	2,730
28,125	0,540	5,000	230,119	70,105	0,033	28,908	2,985
27,750	0,590	3,000	260,030	70,641	0,017	29,010	2,638
27,250	0,630	4,500	251,694	71,822	0,003	28,662	2,320
26,813	0,570	4,000	230,800	71,819	0,000	28,123	2,758
27,000	0,530	6,000	208,389	72,984	0,517	28,274	2,258
27,125	0,590	4,000	240,496	71,055	0,217	28,468	3,279
27,563	0,620	5,000	257,438	70,661	1,690	28,532	2,508
27,500	0,590	3,500	263,295	69,940	3,047	28,418	2,843
27,500	0,600	3,500	257,421	70,866	0,137	28,695	2,625
27,625	0,620	5,500	270,350	72,222	0,000	28,642	2,175
27,688	0,590	7,500	240,082	73,559	0,000	28,744	3,515
27,375	0,630	7,500	258,303	72,236	0,000	28,420	2,923
27,875	0,600	4,000	250,655	71,232	0,000	28,825	2,415
27,875	0,630	5,000	258,864	72,856	0,000	28,971	2,370
27,938	0,560	5,500	225,739	74,147	0,053	29,028	2,630
27,938	0,650	3,500	277,866	70,207	0,480	29,059	2,598
28,063	0,640	3,000	276,836	69,545	1,440	29,054	2,923
27,750	0,590	3,500	246,492	71,143	0,000	29,058	2,648
27,875	0,610	4,500	258,614	71,504	0,067	29,001	2,838
28,438	0,660	4,500	273,112	68,773	1,067	29,304	2,823
28,125	0,610	4,500	260,526	70,773	0,000	29,521	3,039
28,500	0,520	4,000	241,666	70,226	0,017	29,894	2,810
28,500	0,660	1,000	270,340	72,624	0,000	29,256	3,849
28,000	0,620	4,000	256,833	73,748	0,025	29,266	2,938
27,188	0,620	5,500	241,842	75,761	0,003	28,766	2,782
27,500	0,560	7,000	223,145	74,935	0,128	28,593	3,599
27,188	0,550	7,000	204,635	76,103	0,137	28,383	2,787
27,750	0,610	3,500	256,258	74,279	0,003	28,724	2,573
27,688	0,550	3,500	219,010	74,995	0,007	28,786	2,782
27,938	0,539	5,500	219,833	74,352	0,007	28,654	2,570
28,125	0,630	4,000	253,027	72,895	1,067	28,707	2,855
27,750	0,670	1,500	279,634	72,983	0,000	28,802	2,918
27,250	0,520	6,000	218,767	75,274	0,007	28,508	2,485
27,688	0,640	5,000	258,236	74,439	0,037	28,678	2,927
27,688	0,610	6,000	241,725	73,811	0,060	28,696	2,763
27,438	0,620	4,500	245,717	74,700	0,020	28,694	3,536
27,643	0,590	7,000	224,745	75,492	0,183	28,644	3,332
27,688	0,540	5,500	230,049	75,958	0,570	28,466	2,648
28,188	0,650	3,500	141,753	74,116	0,678	28,643	2,565
27,813	0,620	4,500	133,896	74,498	0,000	28,802	2,903
27,857	0,570	4,000	124,061	73,012	0,000	28,763	2,617
27,938	0,600	6,000	130,133	72,802	0,000	28,857	2,635
27,688	0,650	4,500	140,877	73,152	0,008	28,542	3,130
27,813	0,650	5,000	140,327	73,545	0,000	28,715	2,273

27,625	0,500	3,500	109,223	74,645	0,010	28,757	2,500
27,938	0,570	5,000	124,290	74,845	0,013	28,737	2,945
27,875	0,450	6,000	96,844	73,999	0,017	28,888	2,585
27,938	0,590	4,000	128,477	72,773	0,015	28,954	2,778
27,500	0,600	5,000	130,721	74,857	0,037	28,740	2,680
27,563	0,490	6,500	105,317	76,067	0,283	28,776	2,350
27,375	0,570	5,500	122,687	74,777	0,157	28,441	2,628
28,125	0,600	5,500	129,948	72,466	0,020	28,793	3,395
28,000	0,650	4,000	140,338	70,927	0,007	28,661	2,293
27,875	0,590	6,000	127,315	75,253	0,060	28,824	2,505
28,000	0,590	5,500	126,593	74,412	0,173	28,180	2,720
28,125	0,610	4,000	132,120	72,235	0,017	28,738	2,848
27,875	0,660	3,500	141,684	72,829	0,007	28,682	2,793
27,938	0,580	6,000	124,579	73,867	0,023	28,640	2,553
27,688	0,590	5,000	127,203	73,712	0,007	28,604	2,933
27,938	0,550	9,000	118,741	75,280	0,000	28,752	2,725
28,188	0,640	4,000	136,947	74,148	0,000	28,774	2,598
28,125	0,610	5,000	129,900	73,357	0,003	28,622	2,530
28,438	0,550	5,000	118,504	73,262	0,003	28,988	2,407
28,313	0,640	3,500	136,901	72,789	0,000	29,008	4,071
28,250	0,620	4,000	132,338	72,343	0,000	28,891	3,125
28,313	0,600	4,500	127,431	73,321	0,000	28,778	2,578
28,563	0,620	3,500	132,570	71,998	0,000	29,074	2,835
28,875	0,630	5,500	135,392	71,501	0,000	29,282	3,137
29,188	0,590	4,000	169,344	72,760	0,000	29,530	2,936
29,125	0,570	3,000	204,222	69,151	0,000	30,127	3,210
28,813	0,560	4,000	313,473	69,562	0,000	29,682	2,973
28,750	0,580	5,500	257,252	70,874	0,033	29,065	2,975
28,250	0,540	6,000	248,469	70,455	0,000	28,945	2,435
27,938	0,480	4,500	216,132	71,182	0,000	28,573	2,808
28,250	0,530	3,000	237,777	70,829	0,000	28,809	3,095
28,313	0,530	4,500	240,455	71,552	0,013	28,755	2,985
28,563	0,550	3,500	233,773	72,016	0,083	29,019	1,648
28,500	0,460	7,500	184,391	73,280	0,530	29,165	1,631
28,625	0,500	3,000	236,877	71,130	0,173	28,431	2,793
27,917	0,520	5,500	226,674	70,619	0,000	28,383	2,553
29,417	0,570	5,000	248,482	72,239	0,000	29,166	2,820
28,875	0,510	4,500	226,519	72,027	0,000	29,284	2,830
29,400	0,540	4,000	237,652	70,642	0,000	28,365	2,848
28,625	0,500	6,000	219,222	73,234	0,160	27,702	2,762
28,875	0,540	5,500	236,984	73,347	1,193	29,173	2,735
28,688	0,540	5,500	237,899	74,216	3,780	28,836	3,088
28,214	0,520	4,000	236,077	73,112	0,100	29,151	2,925
27,250	0,530	7,000	223,768	74,924	0,043	29,250	2,630
29,214	0,510	6,500	229,325	72,083	0,123	29,008	2,703
28,625	0,510	7,500	210,290	73,043	0,010	28,928	2,755
28,643	0,490	3,500	207,724	70,972	0,003	29,106	2,393
28,688	0,534	6,500	231,280	70,774	0,057	29,128	2,700
28,750	0,580	3,500	264,146	69,853	0,030	29,163	2,530

27,093	0,590	6,500	241,259	72,405	0,063	29,072	2,815
27,093	0,550	4,500	251,571	70,344	0,203	29,374	2,778
29,063	0,440	6,500	200,312	68,322	0,633	28,721	2,623
28,938	0,490	6,500	221,011	68,291	0,007	29,061	2,990
28,688	0,530	7,000	222,275	72,685	0,063	29,245	2,893
27,813	0,330	9,500	135,124	78,450	2,547	28,670	2,250

APÊNDICE B – Código fonte

Análise de todas as bases

```
1 # -*- coding: utf-8 -*-
2 """
3 @author: Luiz Antônio de Sousa Ferreira
4 """
5
6 #%%
7
8 import pandas as pd
9 import numpy as np
10 import matplotlib.pyplot as plt
11 plt.style.use("ggplot")
12 import seaborn as sns
13 import missingno as msno
14
15 #%% Import SONDA data
16
17 cols = ['dir_avg', 'par_avg', 'rain']
18
19 sonda_jan = pd.read_csv('SONDA/SLZ1501ED.csv', sep=";", index_col=False)
20 sonda_jan[cols] = sonda_jan[cols].applymap(lambda x: x if not isinstance(x,
    str) else np.nan)
21 sonda_feb = pd.read_csv('SONDA/SLZ1502ED.csv', sep=";", index_col=False)
22 sonda_feb[cols] = sonda_feb[cols].applymap(lambda x: x if not isinstance(x,
    str) else np.nan)
23 sonda_mar = pd.read_csv('SONDA/SLZ1503ED.csv', sep=";", index_col=False)
24 sonda_mar[cols] = sonda_mar[cols].applymap(lambda x: x if not isinstance(x,
    str) else np.nan)
25 sonda_apr = pd.read_csv('SONDA/SLZ1504ED.csv', sep=";", index_col=False)
26 sonda_apr[cols] = sonda_apr[cols].applymap(lambda x: x if not isinstance(x,
    str) else np.nan)
27 sonda_may = pd.read_csv('SONDA/SLZ1505ED.csv', sep=";", index_col=False)
28 sonda_may[cols] = sonda_may[cols].applymap(lambda x: x if not isinstance(x,
```

```

1     str) else np.nan)
2
3 sonda_jun = pd.read_csv('SONDA/SLZ1506ED.csv', sep=";", index_col=False)
4 sonda_jun[cols] = sonda_jun[cols].applymap(lambda x: x if not isinstance(x,
5     str) else np.nan)
6
7 sonda_jul = pd.read_csv('SONDA/SLZ1507ED.csv', sep=";", index_col=False)
8 sonda_jul[cols] = sonda_jul[cols].applymap(lambda x: x if not isinstance(x,
9     str) else np.nan)
10
11 sonda_aug = pd.read_csv('SONDA/SLZ1508ED.csv', sep=";", index_col=False)
12 sonda_aug[cols] = sonda_aug[cols].applymap(lambda x: x if not isinstance(x,
13     str) else np.nan)
14
15 sonda_sep = pd.read_csv('SONDA/SLZ1509ED.csv', sep=";", index_col=False)
16 sonda_sep[cols] = sonda_sep[cols].applymap(lambda x: x if not isinstance(x,
17     str) else np.nan)
18
19 sonda_oct = pd.read_csv('SONDA/SLZ1510ED.csv', sep=";", index_col=False)
20 sonda_oct[cols] = sonda_oct[cols].applymap(lambda x: x if not isinstance(x,
21     str) else np.nan)
22
23 sonda_nov = pd.read_csv('SONDA/SLZ1511ED.csv', sep=";", index_col=False)
24 sonda_nov[cols] = sonda_nov[cols].applymap(lambda x: x if not isinstance(x,
25     str) else np.nan)
26
27 sonda_dec = pd.read_csv('SONDA/SLZ1512ED.csv', sep=";", index_col=False)
28 sonda_dec[cols] = sonda_dec[cols].applymap(lambda x: x if not isinstance(x,
29     str) else np.nan)
30
31
32 frames = [sonda_jan, sonda_feb, sonda_mar, sonda_apr, sonda_may, sonda_jun,
33     sonda_jul, sonda_aug, sonda_sep, sonda_oct, sonda_nov, sonda_dec]
34
35 sonda = pd.concat(frames, sort=False)
36
37 # Format dataframe sonda, remove not utilized and reset index
38 sonda_columns_drop = ['id', 'dir_avg', 'diff_avg', 'glo_avg', 'press', 'par_avg',
39     'lw_avg', 'lux_avg', 'wd_10m']
40 sonda_formatted = sonda.drop(sonda_columns_drop, axis=1)
41 sonda_formatted.reset_index(drop=True, inplace=True)
42
43 # Rename sonda cols for project default
44 sonda_formatted.rename(columns={'glo_avr': 'global_irradiance',
45     'humid': 'humidity',
46     'rain': 'precipitation',
47     'tp_sfc': 'temperature',

```

```

58         'ws_10m': 'wind_speed'}, inplace=True)
59
60 #%% Import SINDA data
61
62 sinda = pd.read_csv('SINDA/sinda.csv', sep=";", index_col=False)
63
64 # Format dataframe sinda, remove not utilized and reset index
65 sinda_columns_drop = ['PressaoAtm (mB )', 'DataHora', '(GMT)', 'Bateria (Volts
        )', 'CorrPSol (Logico )', 'DirVelVentoMax (oNV )', 'DirVento (oNV )', '
        UmidInt (%)', 'VelVentoMax (m/s )']
66 sinda_formatted = sinda.drop(sinda_columns_drop, axis=1)
67 sinda_formatted.reset_index(drop=True, inplace=True)
68 # Calculate medium temperature data for all collects
69 sinda_formatted.loc[:, 'temperature'] = (sinda_formatted['TempMax (oC )'] +
        sinda_formatted['TempMin (oC )']) / 2
70
71 sinda_formatted.rename(columns={'day': 'day',
72                                'Pluvio (mm )': 'precipitation',
73                                'RadSolGlob (W/m2 )': 'global_irradiance',
74                                'TempAr (oC )': 'air_temperature',
75                                'TempMax (oC )': 'temperature_max',
76                                'TempMin (oC )': 'temperature_min',
77                                'UmidRel (%)': 'humidity',
78                                'VelVento (m/s )': 'wind_speed'
79                                }, inplace=True)
80
81 for i in range(1, 366):
82     if sinda_formatted[sinda_formatted['day'] == i].empty:
83         sinda_formatted.loc[len(sinda_formatted), ['day']] = i
84
85 sinda_formatted = sinda_formatted.sort_values(by=['day'])
86 sinda_formatted = sinda_formatted.reset_index(drop=True)
87
88 #%% Import POWER data
89 nasa = pd.read_csv('NASA/nasa.csv', sep=";", index_col=False)
90
91 # Format dataframe nasa, remove not utilized and reset index
92 nasa_columns_drop = ['LAT', 'LON', 'PS', 'KT_CLEAR', 'T2M_RANGE', 'TS', '
        ALLSKY_SFC_LW_DWN', 'T2MDEW', 'T2M_MAX', 'T2M_MIN', 'WD10M']

```

```

93 nasa_formatted = nasa.drop(nasa_columns_drop, axis=1)
94 nasa_formatted.reset_index(drop=True, inplace=True)
95
96 # Rename nasa cols for project default
97 nasa_formatted.rename(columns={'YEAR': 'year',
98                               'MO': 'month',
99                               'DY': 'day',
100                              'PRECTOT': 'precipitation',
101                              'T2M': 'temperature',
102                              'ALLSKY_SFC_SW_DWN': 'global_irradiance',
103                              'KT': 'clarity_index',
104                              'RH2M': 'humidity',
105                              'WS10M': 'wind_speed',
106                              'CLRSKY_SFC_SW_DWN': '
                                clear_sky_global_irradiance'}, inplace=
                                True)
107
108 #%% Import BDMEP (diario and horario) data
109 bdmeep_diario = pd.read_csv('BDMEP/bdmeep_diario.csv', sep=";", index_col=False
    )
110 bdmeep_horario = pd.read_csv('BDMEP/bdmeep_horario.csv', sep=";", index_col=
    False)
111
112 bdmeep_diario = bdmeep_diario.drop(bdmeep_diario.columns[bdmeep_diario.columns.
    str.contains('Unnamed', case = False)], axis = 1)
113 bdmeep_horario = bdmeep_horario.drop(bdmeep_horario.columns[bdmeep_horario.
    columns.str.contains('Unnamed', case = False)], axis = 1)
114
115 # Format dataframe nasa, remove not utilized and reset index
116 bdmeep_horario_columns_drop = ['Data', 'Estacao', 'Hora', 'TempBulboSeco', '
    PressaoAtmEstacao']
117 bdmeep_diario_columns_drop = ['Data', 'Hora', 'Estacao']
118 bdmeep_horario_formatted = bdmeep_horario.drop(bdmeep_horario_columns_drop,
    axis=1)
119 bdmeep_diario_formatted = bdmeep_diario.drop(bdmeep_diario_columns_drop, axis
    =1)
120
121 # Rename bdmeep cols for project default
122 bdmeep_horario_formatted.rename(columns={'Nebulosidade': 'cloud_cover',

```

```

123         'TempBulboUmido': 'temperature',
124         'UmidadeRelativa': 'humidity',
125         'VelocidadeVento': 'wind_speed'},
126         inplace=True)
127 bdmep_diario_formatted.rename(columns ={'Precipitacao': 'precipitation',
128         'TempMaxima': 'temperature_max',
129         'TempMinima': 'temperature_min',
130         'Umidade Relativa Media': 'humidity
131         'Velocidade do Vento Media': '
132         wind_speed'}, inplace=True)
133 # Calculate medium temperature data for all collects
134 bdmep_diario_formatted = bdmep_diario_formatted.groupby(
135     bdmep_diario_formatted['day'], as_index=False).max()
136 bdmep_horario_formatted = bdmep_horario_formatted.groupby(
137     bdmep_horario_formatted['day'], as_index=False).mean()
138 bdmep_horario_formatted = bdmep_horario_formatted.drop('day', axis=1)
139 bdmep_formatted = pd.concat([bdmep_diario_formatted,
140     bdmep_horario_formatted], axis=1, join='inner')
141 bdmep_formatted.loc[:, 'temperature'] = (bdmep_formatted['temperature_max'] +
142     bdmep_formatted['temperature_min']) / 2
143
144 #%% #####
145 # Missingno – Sonda – all data
146
147 msno.matrix(sonda_formatted)
148
149 #%% Boxplot – Sonda – glo_avr
150
151 sonda_bplot_drop = ['day', 'diff_avg', 'dir_avg', 'glo_avg', 'humid',
152     'id', 'lux_avg', 'lw_avg', 'min', 'par_avg', 'rain',
153     'tp_sfc', 'wd_10m', 'ws_10m', 'year', 'press']
154 sonda_bplot = sonda.drop(sonda_bplot_drop, axis=1)
155
156 bplot = sns.boxplot(data=sonda_bplot, width=0.8, orient="h", palette="
157     colorblind")
158 bplot.set\char' _title ("Base Sonda – Boxplot – Variãqvel global\char '
159     _irradiance")

```

```

153 plt.show()
154
155 #%% Boxplot – Sonda – tp_sfc
156
157 sonda_bplot_drop = [ 'day', 'diff_avg', 'dir_avg', 'glo_avg', 'humid',
158                     'id', 'lux_avg', 'lw_avg', 'min', 'par_avg', 'rain',
159                     'glo_avr', 'wd_10m', 'ws_10m', 'year', 'press' ]
160 sonda_bplot = sonda.drop(sonda_bplot_drop, axis=1)
161
162 bplot = sns.boxplot(data=sonda_bplot, width=0.8, orient="h", palette="
    colorblind")
163 bplot.set\char' _title ("Base Sonda – Boxplot – VariÃavel temperature")
164 plt.show()
165
166 #%% Boxplot – Sonda – humid
167
168 sonda_bplot_drop = [ 'day', 'diff_avg', 'dir_avg', 'glo_avg', 'tp_sfc',
169                     'id', 'lux_avg', 'lw_avg', 'min', 'par_avg', 'rain',
170                     'glo_avr', 'wd_10m', 'ws_10m', 'year', 'press' ]
171 sonda_bplot = sonda.drop(sonda_bplot_drop, axis=1)
172
173 bplot = sns.boxplot(data=sonda_bplot, width=0.8, orient="h", palette="
    colorblind")
174 bplot.set\char' _title ("Base Sonda – Boxplot – VariÃavel humidity")
175 plt.show()
176
177 #%% Boxplot – Sonda – rain
178
179 sonda_bplot_drop = [ 'day', 'diff_avg', 'dir_avg', 'glo_avg', 'tp_sfc',
180                     'id', 'lux_avg', 'lw_avg', 'min', 'par_avg', 'press',
181                     'glo_avr', 'wd_10m', 'ws_10m', 'year', 'humid' ]
182 sonda_bplot = sonda.drop(sonda_bplot_drop, axis=1)
183
184 bplot = sns.boxplot(data=sonda_bplot, width=0.8, orient="h", palette="
    colorblind")
185 bplot.set\char' _title ("Base Sonda – Boxplot – VariÃavel precipitation")
186 plt.show()
187
188 #%% Boxplot – Sonda – ws_10m

```

```

189
190 sonda_bplot_drop = [ 'day', 'diff_avg', 'dir_avg', 'glo_avg', 'tp_sfc',
191                     'id', 'lux_avg', 'lw_avg', 'min', 'par_avg', 'rain',
192                     'glo_avr', 'wd_10m', 'rain', 'year', 'humid', 'press' ]
193 sonda_bplot = sonda.drop(sonda_bplot_drop, axis=1)
194
195 bplot = sns.boxplot(data=sonda_bplot, width=0.8, orient="h", palette="
        colorblind")
196 bplot.set\char' _title("Base Sonda – Boxplot – VariÃavel wind_speed")
197 plt.show()
198
199 #%% #####
200 # Missingno – Sinda – all data
201
202 msno.matrix(sinda_formatted)
203
204 #%% Boxplot – Sinda – Pluvio
205
206 sinda_bplot_drop = [ 'DataHora', '(GMT)', 'day', 'Bateria (Volts )', 'CorrPSol (
        Logico )',
207 'DirVelVentoMax (oNV )', 'DirVento (oNV )', 'PressaoAtm (mB )', 'RadSolGlob (
        W/m2 )',
208 'TempAr (oC )', 'TempMax (oC )', 'TempMin (oC )', 'UmidInt (%)', 'UmidRel (%)
        )', 'VelVento (m/s )',
209 'VelVentoMax (m/s )' ]
210
211 sinda_bplot = sinda.drop(sinda_bplot_drop, axis=1)
212
213 bplot = sns.boxplot(data=sinda_bplot, width=0.8, orient="h", palette="
        colorblind")
214 bplot.set\char' _title("Base Sinda – Boxplot – VariÃavel precipitation")
215 plt.show()
216
217 #%% Boxplot – Sinda – RadSolGlob
218
219 sinda_bplot_drop = [ 'DataHora', '(GMT)', 'day', 'Bateria (Volts )', 'CorrPSol (
        Logico )',
220 'DirVelVentoMax (oNV )', 'DirVento (oNV )', 'Pluvio (mm )', 'PressaoAtm (mB )
        ',

```

```

221 'TempAr (oC )', 'TempMax (oC )', 'TempMin (oC )', 'UmidInt (%)', 'UmidRel (%)
    )', 'VelVento (m/s )',
222 'VelVentoMax (m/s )']
223
224 sinda_bplot = sinda.drop(sinda_bplot_drop, axis=1)
225
226 bplot = sns.boxplot(data=sinda_bplot, width=0.8, orient="h", palette="
    colorblind")
227 bplot.set\char'\_title("Base Sinda – Boxplot – Variável global_irradiance")
228 plt.show()
229
230 #%% Boxplot – Sinda – UmidRel
231
232 sinda_bplot_drop = [ 'DataHora', '(GMT)', 'day', 'Bateria (Volts )', 'CorrPSol (
    Logico )',
233 'DirVelVentoMax (oNV )', 'DirVento (oNV )', 'Pluvio (mm )', 'PressaoAtm (mB )
    ',
234 'TempAr (oC )', 'TempMax (oC )', 'TempMin (oC )', 'RadSolGlob (W/m2 )', '
    UmidInt (%)', 'VelVento (m/s )',
235 'VelVentoMax (m/s )']
236
237 sinda_bplot = sinda.drop(sinda_bplot_drop, axis=1)
238
239 bplot = sns.boxplot(data=sinda_bplot, width=0.8, orient="h", palette="
    colorblind")
240 bplot.set\char'\_title("Base Sinda – Boxplot – Variável humidity")
241 plt.show()
242
243
244 #%% Boxplot – Sinda – VelVento
245
246 sinda_bplot_drop = [ 'DataHora', '(GMT)', 'day', 'Bateria (Volts )', 'CorrPSol (
    Logico )',
247 'DirVelVentoMax (oNV )', 'DirVento (oNV )', 'Pluvio (mm )', 'PressaoAtm (mB )
    ',
248 'TempAr (oC )', 'TempMax (oC )', 'TempMin (oC )', 'RadSolGlob (W/m2 )', '
    UmidInt (%)', 'UmidRel (%)',
249 'VelVentoMax (m/s )']
250

```

```

251 sinda_bplot = sinda.drop(sinda_bplot_drop, axis=1)
252
253 bplot = sns.boxplot(data=sinda_bplot, width=0.8, orient="h", palette="
    colorblind")
254 bplot.set\char' _title("Base Sinda – Boxplot – VariÃqvel wind_speed")
255 plt.show()
256
257 #%% #####
258 # Missingno – Power – all data
259
260 msno.matrix(sinda_formatted)
261
262 #%% Boxplot – Power – PRECTOT
263
264 nasa_bplot_drop = ['LAT', 'LON', 'YEAR', 'MO', 'DY', 'RH2M',
265 'PS', 'T2M_RANGE', 'TS', 'T2MDEW', 'T2M_MAX', 'T2M_MIN', 'T2M', 'WD10M', 'WS10M', '
    KT_CLEAR',
266 'KT', 'ALLSKY_SFC_LW_DWN', 'CLRSKY_SFC_SW_DWN', 'ALLSKY_SFC_SW_DWN']
267
268 nasa_bplot = nasa.drop(nasa_bplot_drop, axis=1)
269
270 bplot = sns.boxplot(data=nasa_bplot, width=0.8, orient="h", palette="
    colorblind")
271 bplot.set\char' _title("Base POWER – Boxplot – VariÃqvel precipitation")
272 plt.show()
273
274 #%% Boxpolot – Power – T2M
275
276 nasa_bplot_drop = ['LAT', 'LON', 'YEAR', 'MO', 'DY', 'RH2M', 'PRECTOT',
277 'PS', 'T2M_RANGE', 'TS', 'T2MDEW', 'T2M_MAX', 'T2M_MIN', 'WD10M', 'WS10M', '
    KT_CLEAR',
278 'KT', 'ALLSKY_SFC_LW_DWN', 'CLRSKY_SFC_SW_DWN', 'ALLSKY_SFC_SW_DWN']
279
280 nasa_bplot = nasa.drop(nasa_bplot_drop, axis=1)
281
282 bplot = sns.boxplot(data=nasa_bplot, width=0.8, orient="h", palette="
    colorblind")
283 bplot.set\char' _title("Base POWER – Boxplot – VariÃqvel temperature")
284 plt.show()

```

```

285
286 #%% Boxplot – Power – = ALLSKY_SFC_SW_DWN
287
288 nasa_bplot_drop = [ 'LAT', 'LON', 'YEAR', 'MO', 'DY', 'RH2M', 'PRECTOT',
289 'PS', 'T2M_RANGE', 'TS', 'T2MDEW', 'T2M_MAX', 'T2M', 'T2M_MIN', 'WD10M', 'WS10M',
290 'KT_CLEAR',
291 'KT', 'ALLSKY_SFC_LW_DWN', 'CLRSKY_SFC_SW_DWN' ]
292
293 nasa_bplot = nasa.drop(nasa_bplot_drop, axis=1)
294
295 bplot = sns.boxplot(data=nasa_bplot, width=0.8, orient="h", palette="
296 colorblind")
297
298 bplot.set\char' _title("Base POWER – Boxplot – Variável global_irradiance")
299 plt.show()
300
301 #%% Boxplot – Power – = KT
302
303 nasa_bplot_drop = [ 'LAT', 'LON', 'YEAR', 'MO', 'DY', 'RH2M', 'PRECTOT',
304 'T2M', 'T2M_RANGE', 'TS', 'T2MDEW', 'T2M_MAX', 'T2M_MIN', 'WD10M', 'WS10M', 'PS',
305 'KT_CLEAR', 'ALLSKY_SFC_LW_DWN', 'CLRSKY_SFC_SW_DWN', 'ALLSKY_SFC_SW_DWN' ]
306
307 nasa_bplot = nasa.drop(nasa_bplot_drop, axis=1)
308
309 bplot = sns.boxplot(data=nasa_bplot, width=0.8, orient="h", palette="
310 colorblind")
311
312 bplot.set\char' _title("Base POWER – Boxplot – Variável clarity_index")
313 plt.show()
314
315 #%% Boxplot – Power – = RH2M
316
317 nasa_bplot_drop = [ 'LAT', 'LON', 'YEAR', 'MO', 'DY', 'KT', 'PRECTOT',
318 'T2M', 'T2M_RANGE', 'TS', 'T2MDEW', 'T2M_MAX', 'T2M_MIN', 'WD10M', 'WS10M', 'PS',
319 'KT_CLEAR', 'ALLSKY_SFC_LW_DWN', 'CLRSKY_SFC_SW_DWN', 'ALLSKY_SFC_SW_DWN' ]
320
321 nasa_bplot = nasa.drop(nasa_bplot_drop, axis=1)
322
323 bplot = sns.boxplot(data=nasa_bplot, width=0.8, orient="h", palette="
324 colorblind")

```

```

320 bplot.set\char\_title("Base POWER – Boxplot – Variável humidity")
321 plt.show()
322
323 #%% Boxplot – Power – WS10M
324
325 nasa_bplot_drop = ['LAT', 'LON', 'YEAR', 'MO', 'DY', 'KT', 'PRECTOT',
326 'T2M', 'T2M_RANGE', 'TS', 'T2MDEW', 'T2M_MAX', 'T2M_MIN', 'WD10M', 'RH2M', 'PS',
327 'KT_CLEAR', 'ALLSKY_SFC_LW_DWN', 'CLRSKY_SFC_SW_DWN', 'ALLSKY_SFC_SW_DWN']
328
329 nasa_bplot = nasa.drop(nasa_bplot_drop, axis=1)
330
331 bplot = sns.boxplot(data=nasa_bplot, width=0.8, orient="h", palette="
    colorblind")
332 bplot.set\char\_title("Base POWER – Boxplot – Variável wind_speed")
333 plt.show()
334
335 #%% Boxplot – Power – CLRSKY_SFC_SW_DWN
336
337 nasa_bplot_drop = ['LAT', 'LON', 'YEAR', 'MO', 'DY', 'KT', 'PRECTOT',
338 'T2M', 'T2M_RANGE', 'TS', 'T2MDEW', 'T2M_MAX', 'T2M_MIN', 'WD10M', 'RH2M', 'PS',
339 'KT_CLEAR', 'ALLSKY_SFC_LW_DWN', 'WS10M', 'ALLSKY_SFC_SW_DWN']
340
341 nasa_bplot = nasa.drop(nasa_bplot_drop, axis=1)
342
343 bplot = sns.boxplot(data=nasa_bplot, width=0.8, orient="h", palette="
    colorblind")
344 bplot.set\char\_title("Base POWER – Boxplot – Variável
    clear_sky_global_irradiance")
345 plt.show()
346
347 #%% #####
348 # Missingno – BDMEP – Horário – all data
349
350 msno.matrix(bdmep_horario_formatted)
351
352 #%% Boxplot – BDMEP – Horário – PressaoAtmEstacao
353
354 bdmep_horario_bplot_drop = ['Estacao', 'Data', 'day', 'Hora', 'TempBulboSeco', '
    TempBulboUmid',

```

```

355 'UmidadeRelativa', 'VelocidadeVento', 'Nebulosidade']
356
357 bdmeq_horario_bplot = bdmeq_horario.drop(bdmeq_horario_bplot_drop, axis=1)
358
359 bplot = sns.boxplot(data=bdmeq_horario_bplot, width=0.8, orient="h",
    palette="colorblind")
360 bplot.set_title("Base Nasa – Boxplot – PressaoAtmEstacao")
361 plt.show()
362
363 #%% Boxplot – BDMEP – Horario – Nebulosidade
364
365 bdmeq_horario_bplot_drop = ['Estacao', 'Data', 'day', 'Hora', 'TempBulboSeco', '
    TempBulboUmido',
366 'UmidadeRelativa', 'VelocidadeVento', 'PressaoAtmEstacao']
367
368 bdmeq_horario_bplot = bdmeq_horario.drop(bdmeq_horario_bplot_drop, axis=1)
369
370 bplot = sns.boxplot(data=bdmeq_horario_bplot, width=0.8, orient="h",
    palette="colorblind")
371 bplot.set_title("Base BDMEP Horária – Boxplot – cloud_cover")
372 plt.show()
373
374 #%% Boxplot – BDMEP – Horario – TempBulboUmido
375
376 bdmeq_horario_bplot_drop = ['Estacao', 'Data', 'day', 'Hora', 'TempBulboSeco', '
    Nebulosidade',
377 'UmidadeRelativa', 'VelocidadeVento', 'PressaoAtmEstacao']
378
379 bdmeq_horario_bplot = bdmeq_horario.drop(bdmeq_horario_bplot_drop, axis=1)
380
381 bplot = sns.boxplot(data=bdmeq_horario_bplot, width=0.8, orient="h",
    palette="colorblind")
382 bplot.set_title("Base BDMEP Horária – Boxplot – temperature")
383 plt.show()
384
385 #%% Boxplot – BDMEP – Horario – UmidadeRelativa
386
387 bdmeq_horario_bplot_drop = ['Estacao', 'Data', 'day', 'Hora', 'TempBulboSeco', '
    Nebulosidade',

```

```

388 'TempBulboUmido', 'VelocidadeVento', 'PressaoAtmEstacao']
389
390 bdmeq_horario_bplot = bdmeq_horario.drop(bdmeq_horario_bplot_drop, axis=1)
391
392 bplot = sns.boxplot(data=bdmeq_horario_bplot, width=0.8, orient="h",
    palette="colorblind")
393 bplot.set\char' _title("Base BDMEP Hor  ria – Boxplot – humidity")
394 plt.show()
395
396 #%% Boxplot – BDMEP – Horario – VelocidadeVento
397
398 bdmeq_horario_bplot_drop = ['Estacao', 'Data', 'day', 'Hora', 'TempBulboSeco', '
    Nebulosidade',
399 'TempBulboUmido', 'UmidadeRelativa', 'PressaoAtmEstacao']
400
401 bdmeq_horario_bplot = bdmeq_horario.drop(bdmeq_horario_bplot_drop, axis=1)
402
403 bplot = sns.boxplot(data=bdmeq_horario_bplot, width=0.8, orient="h",
    palette="colorblind")
404 bplot.set\char' _title("Base BDMEP Hor  ria – Boxplot – wind_speed")
405 plt.show()
406
407 #%% #####
408 # Missingno – BDMEP – Diario – all data
409
410 msno.matrix(bdmeq_diario_formatted)
411
412 #%% Boxplot – BDMEP – Precipitacao
413
414 bdmeq_diario_bplot_drop = ['Estacao', 'Data', 'day', 'Hora', 'TempMaxima',
    'TempMinima',
415 'Umidade Relativa Media', 'Velocidade do Vento Media']
416
417 bdmeq_diario_bplot = bdmeq_diario.drop(bdmeq_diario_bplot_drop, axis=1)
418
419 bplot = sns.boxplot(data=bdmeq_diario_bplot, width=0.8, orient="h", palette
    ="colorblind")
420 bplot.set\char' _title("Base BDMEP di  rio – Boxplot – precipitation")
421 plt.show()

```

```

422
423 #%% Boxplot – BDMEP – Umidade Relativa Media
424
425 bdmep_diario_bplot_drop = ['Estacao', 'Data', 'day', 'Hora', 'TempMaxima',
426                             'TempMinima',
427                             'Precipitacao', 'Velocidade do Vento Media']
428
429 bdmep_diario_bplot = bdmep_diario.drop(bdmep_diario_bplot_drop, axis=1)
430
431 bplot = sns.boxplot(data=bdmep_diario_bplot, width=0.8, orient="h", palette
432                     ="colorblind")
433 bplot.set\char' _title("Base BDMEP diÃrio – Boxplot – humidity")
434 plt.show()
435
436 #%% Boxplot – BDMEP – Velocidade Vento Media
437
438 bdmep_diario_bplot_drop = ['Estacao', 'Data', 'day', 'Hora', 'TempMaxima',
439                             'TempMinima',
440                             'Precipitacao', 'Umidade Relativa Media']
441
442 bdmep_diario_bplot = bdmep_diario.drop(bdmep_diario_bplot_drop, axis=1)
443
444 bplot = sns.boxplot(data=bdmep_diario_bplot, width=0.8, orient="h", palette
445                     ="colorblind")
446 bplot.set\char' _title("Base BDMEP diÃrio – Boxplot – wind_speed")
447 plt.show()

```

Criação de variáveis calculadas

```

1 import pvlb as pv
2 import pandas as pd
3 import datetime
4 import suntime
5 from IPython.display import HTML
6 from pysolar.solar import *
7
8 #%%

```

```

9 latitude = -21.787778
10 longitude = -46.560833
11
12 initial = '2015-01-01'
13 final = '2015-12-31'
14
15 dates = pd.date_range(initial, final).tolist()
16
17 rows = pd.DataFrame(columns=['latitude', 'longitude', 'elevation', 'azimuth',
18                             'zenith', 'hour', 'solar_day'])
19
20 #%%
21
22 for date in dates:
23
24     dayofyear = pd.to_datetime([date]).tz_localize('America/Recife')
25
26     spa = pv.solarposition.spa_python(dayofyear, latitude, longitude, how=
27         'numpy')
28
29     deltat = pv.spa.calculate_deltat(2015,1)
30
31     sun = suntime.Sun(latitude, longitude)
32     day = dayofyear.date
33
34     today_sr = sun.get_sunrise_time(day.item(0)) - datetime.timedelta(
35         hours=3)
36     today_ss = sun.get_sunset_time(day.item(0)) - datetime.timedelta(hours
37         =3)
38     solar_day = today_ss - today_sr
39     date_formatted = datetime.datetime(date.year,
40         date.month,
41         date.day, 12, 0, 0, 0, tzinfo=datetime.
42         timezone.utc)
43
44     altitude = get_altitude(latitude, longitude, date_formatted)
45     clear_sky = radiation.get_radiation_direct(date, altitude)
46
47     data = {'latitude': latitude, 'longitude': longitude,

```

```

44         'elevation': spa.elevation.values.item(0),
45         'azimuth': spa.azimuth.values.item(0),
46         'zenith': spa.zenith.values.item(0),
47         'hour': '12:00:00',
48         'solar_day': solar_day.total_seconds(),
49         'clear_sky_global_irradiance': clear_sky}
50
51     rows = rows.append(data, ignore_index=True)

```

Integração

```

1  # -*- coding: utf-8 -*-
2
3  #%% Sonda - Remove valores 'year' e 'min'
4
5  sonda_formatted_drop = ['year', 'min']
6
7  sonda_formatted = sonda_formatted.drop(sonda_formatted_drop, axis=1)
8  #%% Sonda - Agrupa valores por mês, a fim de obtenção de valores
   diários
9
10 sonda_formatted = sonda_formatted.groupby(sonda_formatted['day'], as_index=
   False).mean()
11
12 #%% Sinda - Agrupa valores por mês, a fim de obtenção de valores
   diários
13
14 sinda_formatted = sinda_formatted.groupby(sinda_formatted['day'], as_index=
   False).mean()
15
16 #%% POWER - Remove valores 'year' e 'month'
17
18 nasa_formatted_drop = ['year', 'month']
19 nasa_formatted = nasa_formatted.drop(nasa_formatted_drop, axis=1)
20
21 #%% POWER - Criar dia do ano
22 for x in range(1,366):

```

```

23     nasa_formatted['day'][x] = x
24
25 #%% POWER – Conversão de unidade da variável global_irradiance
26     # KWh/mÂš/day para W/mÂš
27
28 def convert_global(global_irradiance):
29     value = (global_irradiance*1000)/24
30     return value
31
32 nasa_formatted['convert_global_irradiance']= nasa_formatted.
    global_irradiance.apply(convert_global)
33
34 nasa_formatted_drop = ['global_irradiance']
35 nasa_formatted = nasa_formatted.drop(nasa_formatted_drop, axis=1)
36
37 nasa_formatted.rename(columns={'convert_global_irradiance': '
    global_irradiance'}, inplace=True)
38
39
40 #%% POWER – Conversão de unidade da variável clear_sky_global_irradiance
41     # KWh/mÂš/day para W/mÂš
42
43 def convert_clear_sky_global(clear_sky_global_irradiance):
44     value = (clear_sky_global_irradiance*1000)/24
45     return value
46
47 nasa_formatted['convert_clear_sky_global_irradiance']= nasa_formatted.
    clear_sky_global_irradiance.apply(convert_clear_sky_global)
48
49 nasa_formatted_drop = ['clear_sky_global_irradiance']
50 nasa_formatted = nasa_formatted.drop(nasa_formatted_drop, axis=1)
51
52 nasa_formatted.rename(columns={'convert_clear_sky_global_irradiance': '
    clear_sky_global_irradiance'}, inplace=True)
53
54 #%% BDMEP – horario – Adiciona coluna 'day' com valores
55
56 bdmep_horario_formatted['day'] = 1
57

```

```

58 for x in range(1,366):
59     bdmeq_horario_formatted['day'][x] = x
60
61 #%% BDMEP – horario – Altera posição das variáveis
62 bdmeq_horario_formatted = bdmeq_horario_formatted[['day', 'temperature',
63     'humidity', 'wind_speed', 'cloud_cover']]
64
65 #%% BDMEP – diário – Cria a variável temperature, com base em
66     temperature_max e temperature_min
67 bdmeq_diario_formatted.loc[:, 'temperature'] = (bdmeq_diario_formatted['
68     temperature_max'] + bdmeq_diario_formatted['temperature_min']) / 2
69
70 bdmeq_diario_formatted_drop = ['temperature_max', 'temperature_min']
71 bdmeq_diario_formatted = bdmeq_diario_formatted.drop(
72     bdmeq_diario_formatted_drop, axis=1)

```

air_temperature

```

1
2 # -*- coding: utf-8 -*-
3
4 sinda_air_temperature = sinda_formatted.loc[:, "air_temperature"]
5 sinda_air_temperature = sinda_air_temperature.to_frame()
6
7 #%% Cálculo do zscore de global_irradiance das bases SONDA, SINDA e NASA
8
9 sinda_air_temperature['zscore'] = (sinda_air_temperature -
10     sinda_air_temperature.mean()) / sinda_air_temperature.std(ddof=0)
11
12 #%% Normaliza Zscore
13
14 for index, row in sinda_air_temperature.iterrows():
15     if row['zscore'] > 2.5:
16         row['zscore'] = np.nan
17         row['air_temperature'] = np.nan
18     elif row['zscore'] < -2.5:
19         row['zscore'] = np.nan

```

```

19         row['air_temperature'] = np.nan
20     else:
21         continue
22
23 #%%
24 XY_incomplete = sinda_air_temperature
25
26 n_imputations = 5
27 XY_completed = []
28 for i in range(n_imputations):
29     imputer = IterativeImputer(n_iter=5, sample_posterior=True,
30                                random_state=i)
31     XY_completed.append(imputer.fit_transform(XY_incomplete))
32
33 XY_completed_mean_sinda_air_temperature = np.mean(XY_completed, 0)
34 XY_completed_std = np.std(XY_completed, 0)
35
36 #%%
37 result_air_temperature = np.mean( np.array([
38     XY_completed_mean_sinda_air_temperature]), axis=0 )
39
40 #%%
41 result_air_temperature = pd.DataFrame(data=result_air_temperature ,
42                                       index=np.array(range(0, 365)),
43                                       columns=['air_temperature', 'zscore'])
44
45 #%%
46 bplot = sns.boxplot(data=result_air_temperature['air_temperature'], width
47                     =0.8, orient="h", palette="colorblind")
48 bplot.set_title("Vari vel air_temperature pr processada – Boxplot")
49 plt.show()

```

clarity_index

```

1 # -*- coding: utf-8 -*-
2
3 nasa_clarity_index = nasa_formatted.loc[:, "clarity_index"]
4 nasa_clarity_index = nasa_clarity_index.to_frame()

```

```

5
6 ### Cálculo do zscore de global_irradiance das bases SONDA, SINDA e NASA
7
8 nasa_clarity_index['zscore'] = (nasa_clarity_index - nasa_clarity_index.
    mean())/nasa_clarity_index.std(ddof=0)
9
10 ### Normalizaçãõ Zscore Power
11
12 for index, row in nasa_clarity_index.iterrows():
13     if row['zscore'] > 3:
14         row['zscore'] = np.nan
15         row['clarity_index'] = np.nan
16     elif row['zscore'] < -3:
17         row['zscore'] = np.nan
18         row['clarity_index'] = np.nan
19     else:
20         continue
21
22 ###
23 XY_incomplete = nasa_clarity_index
24
25 n_imputations = 5
26 XY_completed = []
27 for i in range(n_imputations):
28     imputer = IterativeImputer(n_iter=5, sample_posterior=True,
    random_state=i)
29     XY_completed.append(imputer.fit_transform(XY_incomplete))
30
31 XY_completed_mean_nasa_clarity_index = np.mean(XY_completed, 0)
32 XY_completed_std = np.std(XY_completed, 0)
33
34 ###
35 result_clarity_index = np.mean( np.array([
    XY_completed_mean_nasa_clarity_index ]), axis=0 )
36
37 ###
38 result_clarity_index = pd.DataFrame(data=result_clarity_index,
39     index=np.array(range(0, 365)),
40     columns=['clarity_index', 'zscore'])

```

```

41 #%%
42 bplot = sns.boxplot(data=result_clarity_index , width=0.8, orient="h",
    palette="colorblind")
43 bplot.set_title("Variável clarity_index pré-processada – Boxplot")
44 plt.show()

```

cloud_cover

```

1 # -*- coding: utf-8 -*-
2
3 bdmap_horario_cloud_cover = bdmap_horario_formatted.loc[:, "cloud_cover"]
4 bdmap_horario_cloud_cover = bdmap_horario_cloud_cover.to_frame()
5
6 #%% Cálculo do zscore de global_irradiance das bases SONDA, SINDA e NASA
7
8 bdmap_horario_cloud_cover['zscore'] = (bdmap_horario_cloud_cover -
    bdmap_horario_cloud_cover.mean()) / bdmap_horario_cloud_cover.std(ddof=0)
9
10
11 #%% Normalizações Zscore
12
13 for index, row in bdmap_horario_cloud_cover.iterrows():
14     if row['zscore'] > 3.0:
15         row['zscore'] = np.nan
16         row['cloud_cover'] = np.nan
17     elif row['zscore'] < -3.0:
18         row['zscore'] = np.nan
19         row['cloud_cover'] = np.nan
20     else:
21         continue
22
23
24 #%%
25 XY_incomplete = bdmap_horario_cloud_cover
26
27 n_imputations = 5
28 XY_completed = []

```

```

29 for i in range(n_imputations):
30     imputer = IterativeImputer(n_iter=5, sample_posterior=True,
31                                random_state=i)
32     XY_completed.append(imputer.fit_transform(XY_incomplete))
33 XY_completed_mean_bdmep_horario_cloud_cover = np.mean(XY_completed, 0)
34 XY_completed_std = np.std(XY_completed, 0)
35
36 #%%
37 result_cloud_cover = np.mean( np.array([
38     XY_completed_mean_bdmep_horario_cloud_cover]), axis=0 )
39
40 #%%
41 result_cloud_cover = pd.DataFrame(data=result_cloud_cover,
42                                   index=np.array(range(0, 365)),
43                                   columns=['cloud_cover', 'zscore'])
44
45 #%%
46 bplot = sns.boxplot(data=result_cloud_cover['cloud_cover'], width=0.8,
47                     orient="h", palette="colorblind")
48 bplot.set_title("Vari  vel cloud_cover pr  -processada – Boxplot")
49 plt.show()

```

global_irradiance

```

1 # -*- coding: utf-8 -*-
2
3 sonda_glo = sonda_formatted.loc[:, "global_irradiance"]
4 sonda_glo = sonda_glo.to_frame()
5
6 sinda_glo = sinda_formatted.loc[:, "global_irradiance"]
7 sinda_glo = sinda_glo.to_frame()
8
9 nasa_glo = nasa_formatted.loc[:, "global_irradiance"]
10 nasa_glo = nasa_glo.to_frame()
11
12 #%% C  lculo do zscore de global_irradiance das bases SONDA, SINDA e NASA

```

```

13
14 sonda_glo['zscore'] = (sonda_glo - sonda_glo.mean()) / sonda_glo.std(ddof=0)
15 sinda_glo['zscore'] = (sinda_glo - sinda_glo.mean()) / sinda_glo.std(ddof=0)
16 nasa_glo['zscore'] = (nasa_glo - nasa_glo.mean()) / nasa_glo.std(ddof=0)
17
18 #%% NormalizaŃŃŃŃ Zscore Sonda
19
20 for index, row in sonda_glo.iterrows():
21     if row['zscore'] > 3:
22         row['zscore'] = np.nan
23         row['global_irradiance'] = np.nan
24     elif row['zscore'] < -3:
25         row['zscore'] = np.nan
26         row['global_irradiance'] = np.nan
27     else:
28         continue
29
30 #%% NormalizaŃŃŃŃ Zscore Sinda
31
32 for index, row in sinda_glo.iterrows():
33     if row['zscore'] > 3:
34         row['zscore'] = np.nan
35         row['global_irradiance'] = np.nan
36     elif row['zscore'] < -3:
37         row['zscore'] = np.nan
38         row['global_irradiance'] = np.nan
39     else:
40         continue
41
42 #%% NormalizaŃŃŃŃ Zscore Power
43
44 for index, row in nasa_glo.iterrows():
45     if row['zscore'] > 3:
46         row['zscore'] = np.nan
47         row['global_irradiance'] = np.nan
48     elif row['zscore'] < -3:
49         row['zscore'] = np.nan
50         row['global_irradiance'] = np.nan
51     else:

```

```

52         continue
53
54     #%%
55     XY_incomplete = sonda_glo
56
57     n_imputations = 5
58     XY_completed = []
59     for i in range(n_imputations):
60         imputer = IterativeImputer(n_iter=5, sample_posterior=True,
61                                   random_state=i)
62         XY_completed.append(imputer.fit_transform(XY_incomplete))
63
64     XY_completed_mean_sonda_glo = np.mean(XY_completed, 0)
65     XY_completed_std = np.std(XY_completed, 0)
66
67     #%%
68
69     XY_incomplete = nasa_glo
70
71     n_imputations = 5
72     XY_completed = []
73     for i in range(n_imputations):
74         imputer = IterativeImputer(n_iter=5, sample_posterior=True,
75                                   random_state=i)
76         XY_completed.append(imputer.fit_transform(XY_incomplete))
77
78     XY_completed_mean_nasa_glo = np.mean(XY_completed, 0)
79     XY_completed_std = np.std(XY_completed, 0)
80
81     #%%
82     result_global = np.mean( np.array([ XY_completed_mean_sonda_glo,
83                                       XY_completed_mean_nasa_glo ]), axis=0 )
84
85     #%%
86
87     result_global = pd.DataFrame(data=result_global,
88                                index=np.array(range(0, 365)),
89                                columns=['global_irradiance', 'zscore'])
90
91     #%%
92
93     bplot = sns.boxplot(data=result_global['global_irradiance'], width=0.8,

```

```

    orient="h", palette="colorblind")
88 bplot.set_title("Vari vel global_irradiance pr cessada – Boxplot")
89 plt.show()

```

humidity

```

1 # -*- coding: utf-8 -*-
2
3 sonda_humidity = sonda_formatted.loc[:, "humidity"]
4 sonda_humidity = sonda_humidity.to_frame()
5
6 sinda_humidity = sinda_formatted.loc[:, "humidity"]
7 sinda_humidity = sinda_humidity.to_frame()
8
9 nasa_humidity = nasa_formatted.loc[:, "humidity"]
10 nasa_humidity = nasa_humidity.to_frame()
11
12 bdmap_horario_humidity = bdmap_horario_formatted.loc[:, "humidity"]
13 bdmap_horario_humidity = bdmap_horario_humidity.to_frame()
14
15 bdmap_diario_humidity = bdmap_diario_formatted.loc[:, "humidity"]
16 bdmap_diario_humidity = bdmap_diario_humidity.to_frame()
17
18 ### C lculo do zscore de global_irradiance das bases SONDA, SINDA e NASA
19
20 sonda_humidity['zscore'] = (sonda_humidity - sonda_humidity.mean()) /
    sonda_humidity.std(ddof=0)
21 sinda_humidity['zscore'] = (sinda_humidity - sinda_humidity.mean()) /
    sinda_humidity.std(ddof=0)
22 nasa_humidity['zscore'] = (nasa_humidity - nasa_humidity.mean()) /
    nasa_humidity.std(ddof=0)
23 bdmap_horario_humidity['zscore'] = (bdmap_horario_humidity -
    bdmap_horario_humidity.mean()) / bdmap_horario_humidity.std(ddof=0)
24 bdmap_diario_humidity['zscore'] = (bdmap_diario_humidity -
    bdmap_diario_humidity.mean()) / bdmap_diario_humidity.std(ddof=0)
25
26

```

```

27 #%% Normalização Zscore
28
29 for index, row in sonda_humidity.iterrows():
30     if row['zscore'] > 3.0:
31         row['zscore'] = np.nan
32         row['humidity'] = np.nan
33     elif row['zscore'] < -3.0:
34         row['zscore'] = np.nan
35         row['humidity'] = np.nan
36     else:
37         continue
38
39 for index, row in sinda_humidity.iterrows():
40     if row['zscore'] > 3.0:
41         row['zscore'] = np.nan
42         row['humidity'] = np.nan
43     elif row['zscore'] < -3.0:
44         row['zscore'] = np.nan
45         row['humidity'] = np.nan
46     else:
47         continue
48
49 for index, row in nasa_humidity.iterrows():
50     if row['zscore'] > 3.0:
51         row['zscore'] = np.nan
52         row['humidity'] = np.nan
53     elif row['zscore'] < -3.0:
54         row['zscore'] = np.nan
55         row['humidity'] = np.nan
56     else:
57         continue
58
59 for index, row in bdmap_diario_humidity.iterrows():
60     if row['zscore'] > 3.0:
61         row['zscore'] = np.nan
62         row['humidity'] = np.nan
63     elif row['zscore'] < -3.0:
64         row['zscore'] = np.nan
65         row['humidity'] = np.nan

```

```

66         else:
67             continue
68
69 for index, row in bdmap_horario_humidity.iterrows():
70     if row['zscore'] > 3.0:
71         row['zscore'] = np.nan
72         row['humidity'] = np.nan
73     elif row['zscore'] < -3.0:
74         row['zscore'] = np.nan
75         row['humidity'] = np.nan
76     else:
77         continue
78
79 #%%
80 XY_incomplete = sonda_humidity
81
82 n_imputations = 5
83 XY_completed = []
84 for i in range(n_imputations):
85     imputer = IterativeImputer(n_iter=5, sample_posterior=True,
86                               random_state=i)
87     XY_completed.append(imputer.fit_transform(XY_incomplete))
88
89 XY_completed_mean_sonda_humidity = np.mean(XY_completed, 0)
90 XY_completed_std = np.std(XY_completed, 0)
91 #%%
92 XY_incomplete = sinda_humidity
93
94 n_imputations = 5
95 XY_completed = []
96 for i in range(n_imputations):
97     imputer = IterativeImputer(n_iter=5, sample_posterior=True,
98                               random_state=i)
99     XY_completed.append(imputer.fit_transform(XY_incomplete))
100
101 XY_completed_mean_sinda_humidity = np.mean(XY_completed, 0)
102 XY_completed_std = np.std(XY_completed, 0)

```

```
103 #%%
104 XY_incomplete = nasa_humidity
105
106 n_imputations = 5
107 XY_completed = []
108 for i in range(n_imputations):
109     imputer = IterativeImputer(n_iter=5, sample_posterior=True,
110                                random_state=i)
111     XY_completed.append(imputer.fit_transform(XY_incomplete))
112
113 XY_completed_mean_nasa_humidity = np.mean(XY_completed, 0)
114 XY_completed_std = np.std(XY_completed, 0)
115
116 #%%
117 XY_incomplete = bdmep_diario_humidity
118
119 n_imputations = 5
120 XY_completed = []
121 for i in range(n_imputations):
122     imputer = IterativeImputer(n_iter=5, sample_posterior=True,
123                                random_state=i)
124     XY_completed.append(imputer.fit_transform(XY_incomplete))
125
126 XY_completed_mean_bdmep_diario_humidity = np.mean(XY_completed, 0)
127 XY_completed_std = np.std(XY_completed, 0)
128
129 #%%
130 XY_incomplete = bdmep_horario_humidity
131
132 n_imputations = 5
133 XY_completed = []
134 for i in range(n_imputations):
135     imputer = IterativeImputer(n_iter=5, sample_posterior=True,
136                                random_state=i)
137     XY_completed.append(imputer.fit_transform(XY_incomplete))
138
139 XY_completed_mean_bdmep_horario_humidity = np.mean(XY_completed, 0)
140 XY_completed_std = np.std(XY_completed, 0)
```

```

139 #%%
140 result_humidity = np.mean( np.array([ XY_completed_mean_sonda_humidity ,
141                                     XY_completed_mean_sinda_humidity ,
142                                     XY_completed_mean_nasa_humidity ,
143                                     XY_completed_mean_bdmep_diario_humidity ,
144                                     XY_completed_mean_bdmep_horario_humidity]) ,
                             axis=0 )
145
146 #%%
147 result_humidity = pd.DataFrame(data=result_humidity ,
148                               index=np.array(range(0, 365)) ,
149                               columns=['humidity' , 'zscore' ])
150
151 #%%
152 bplot = sns.boxplot(data=result_humidity['humidity'] , width=0.8 , orient="h"
153                    , palette="colorblind")
154 bplot.set_title("Variável humidity processada – Boxplot")
155 plt.show()

```

precipitation

```

1
2
3 # -*- coding: utf-8 -*-
4
5 sonda_precipitation = sonda_formatted.loc[:, "precipitation"]
6 sonda_precipitation = sonda_precipitation.to_frame()
7
8 sinda_precipitation = sinda_formatted.loc[:, "precipitation"]
9 sinda_precipitation = sinda_precipitation.to_frame()
10
11 nasa_precipitation = nasa_formatted.loc[:, "precipitation"]
12 nasa_precipitation = nasa_precipitation.to_frame()
13
14 bdmep_diario_precipitation = bdmep_diario_formatted.loc[:, "precipitation"
15 ]
16 bdmep_diario_precipitation = bdmep_diario_precipitation.to_frame()

```

```

16
17 #%% C  lculo do zscore de global_irradiance das bases SONDA, SINDA e NASA
18
19 sonda_precipitation['zscore'] = (sonda_precipitation - sonda_precipitation.
    mean()) / sonda_precipitation.std(ddof=0)
20 sinda_precipitation['zscore'] = (sinda_precipitation - sinda_precipitation.
    mean()) / sinda_precipitation.std(ddof=0)
21 nasa_precipitation['zscore'] = (nasa_precipitation - nasa_precipitation.
    mean()) / nasa_precipitation.std(ddof=0)
22 bdmap_diario_precipitation['zscore'] = (bdmap_diario_precipitation -
    bdmap_diario_precipitation.mean()) / bdmap_diario_precipitation.std(ddof
    =0)
23
24
25 #%% Normaliza  o Zscore
26
27 for index, row in sonda_precipitation.iterrows():
28     if row['zscore'] > 3.0:
29         row['zscore'] = np.nan
30         row['precipitation'] = np.nan
31     elif row['zscore'] < -3.0:
32         row['zscore'] = np.nan
33         row['precipitation'] = np.nan
34     else:
35         continue
36
37 for index, row in sinda_precipitation.iterrows():
38     if row['zscore'] > 3.0:
39         row['zscore'] = np.nan
40         row['precipitation'] = np.nan
41     elif row['zscore'] < -3.0:
42         row['zscore'] = np.nan
43         row['precipitation'] = np.nan
44     else:
45         continue
46
47 for index, row in nasa_precipitation.iterrows():
48     if row['zscore'] > 3.0:
49         row['zscore'] = np.nan

```

```

50         row['precipitation'] = np.nan
51     elif row['zscore'] < -3.0:
52         row['zscore'] = np.nan
53         row['precipitation'] = np.nan
54     else:
55         continue
56
57 for index, row in bdmap_diario_precipitation.iterrows():
58     if row['zscore'] > 3.0:
59         row['zscore'] = np.nan
60         row['precipitation'] = np.nan
61     elif row['zscore'] < -3.0:
62         row['zscore'] = np.nan
63         row['precipitation'] = np.nan
64     else:
65         continue
66
67
68 #%%
69 XY_incomplete = sonda_precipitation
70
71 n_imputations = 5
72 XY_completed = []
73 for i in range(n_imputations):
74     imputer = IterativeImputer(n_iter=5, sample_posterior=True,
75                                random_state=i)
76     XY_completed.append(imputer.fit_transform(XY_incomplete))
77
78 XY_completed_mean_sonda_precipitation = np.mean(XY_completed, 0)
79
80 #%%
81 XY_incomplete = sinda_precipitation
82 n_imputations = 5
83 XY_completed = []
84 for i in range(n_imputations):
85     imputer = IterativeImputer(n_iter=5, sample_posterior=True,
86                                random_state=i)
87     XY_completed.append(imputer.fit_transform(XY_incomplete))

```

```

87
88 XY_completed_mean_sinda_precipitation = np.mean(XY_completed, 0)
89 XY_completed_std = np.std(XY_completed, 0)
90
91 #%%
92 XY_incomplete = nasa_precipitation
93
94 n_imputations = 5
95 XY_completed = []
96 for i in range(n_imputations):
97     imputer = IterativeImputer(n_iter=5, sample_posterior=True,
98                               random_state=i)
99     XY_completed.append(imputer.fit_transform(XY_incomplete))
100
101 XY_completed_mean_nasa_precipitation = np.mean(XY_completed, 0)
102 XY_completed_std = np.std(XY_completed, 0)
103
104 #%%
105 XY_incomplete = bdmap_diario_precipitation
106
107 n_imputations = 5
108 XY_completed = []
109 for i in range(n_imputations):
110     imputer = IterativeImputer(n_iter=5, sample_posterior=True,
111                               random_state=i)
112     XY_completed.append(imputer.fit_transform(XY_incomplete))
113
114 XY_completed_mean_bdmap_diario_precipitation = np.mean(XY_completed, 0)
115 XY_completed_std = np.std(XY_completed, 0)
116
117 #%%
118 result_precipitation = np.mean( np.array([
119     XY_completed_mean_sinda_precipitation,
120     XY_completed_mean_nasa_precipitation,
121     XY_completed_mean_bdmap_diario_precipitation]),
122                                axis=0 )
123
124 #%%

```

```

122 result_precipitation = pd.DataFrame(data=result_precipitation ,
123                                     index=np.array(range(0, 365)),
124                                     columns=['precipitation', 'zscore'])
125
126 ###
127 bplot = sns.boxplot(data=result_precipitation['precipitation'], width=0.8,
128                    orient="h", palette="colorblind")
129 bplot.set_title("Variável precipitation pré-processada – Boxplot")
130 plt.show()

```

temperature

```

1  # -*- coding: utf-8 -*-
2
3  sonda_temperature = sonda_formatted.loc[:, "temperature"]
4  sonda_temperature = sonda_temperature.to_frame()
5
6  sinda_temperature = sinda_formatted.loc[:, "temperature"]
7  sinda_temperature = sinda_temperature.to_frame()
8
9  nasa_temperature = nasa_formatted.loc[:, "temperature"]
10 nasa_temperature = nasa_temperature.to_frame()
11
12 bdmap_horario_temperature = bdmap_horario_formatted.loc[:, "temperature"]
13 bdmap_horario_temperature = bdmap_horario_temperature.to_frame()
14
15 bdmap_diario_temperature = bdmap_diario_formatted.loc[:, "temperature"]
16 bdmap_diario_temperature = bdmap_diario_temperature.to_frame()
17
18 ### Removendo valores cujo mínimo e máximo superam os registros
19     históricos
20 for index, row in sonda_temperature.iterrows():
21     if row['temperature'] > 37.2:
22         row['temperature'] = np.nan
23
24     elif row['temperature'] < 17.9:

```

```

25         row[ 'temperature' ] = np.nan
26     else:
27         continue
28
29 for index, row in sinda_temperature.iterrows():
30     if row[ 'temperature' ] > 37.2:
31         row[ 'temperature' ] = np.nan
32     elif row [ 'temperature' ] < 17.9:
33         row[ 'temperature' ] = np.nan
34     else:
35         continue
36
37 for index, row in nasa_temperature.iterrows():
38     if row[ 'temperature' ] > 37.2:
39         row[ 'temperature' ] = np.nan
40     elif row [ 'temperature' ] < 17.9:
41         row[ 'temperature' ] = np.nan
42     else:
43         continue
44
45 for index, row in bdmap_diario_temperature.iterrows():
46     if row[ 'temperature' ] > 37.2:
47         row[ 'temperature' ] = np.nan
48     elif row [ 'temperature' ] < 17.9:
49         row[ 'temperature' ] = np.nan
50     else:
51         continue
52
53 for index, row in bdmap_horario_temperature.iterrows():
54     if row[ 'temperature' ] > 37.2:
55         row[ 'temperature' ] = np.nan
56     elif row [ 'temperature' ] < 17.9:
57         row[ 'temperature' ] = np.nan
58     else:
59         continue
60
61 #%% C  lculo do zscore de global_irradiance das bases SONDA, SINDA e NASA
62
63 sonda_temperature[ 'zscore' ] = (sonda_temperature - sonda_temperature.mean())

```

```

        )/sonda_temperature.std(ddof=0)
64 sinda_temperature['zscore'] = (sinda_temperature - sinda_temperature.mean()
    )/sinda_temperature.std(ddof=0)
65 nasa_temperature['zscore'] = (nasa_temperature - nasa_temperature.mean())/
    nasa_temperature.std(ddof=0)
66 bdmap_horario_temperature['zscore'] = (bdmap_horario_temperature -
    bdmap_horario_temperature.mean())/bdmap_horario_temperature.std(ddof=0)
67 bdmap_diario_temperature['zscore'] = (bdmap_diario_temperature -
    bdmap_diario_temperature.mean())/bdmap_diario_temperature.std(ddof=0)
68
69
70 #%% Normaliza Zscore
71
72 for index, row in sonde_temperature.iterrows():
73     if row['zscore'] > 3.0:
74         row['zscore'] = np.nan
75         row['temperature'] = np.nan
76     elif row['zscore'] < -3.0:
77         row['zscore'] = np.nan
78         row['temperature'] = np.nan
79     else:
80         continue
81
82 for index, row in sinda_temperature.iterrows():
83     if row['zscore'] > 3.0:
84         row['zscore'] = np.nan
85         row['temperature'] = np.nan
86     elif row['zscore'] < -3.0:
87         row['zscore'] = np.nan
88         row['temperature'] = np.nan
89     else:
90         continue
91
92 for index, row in nasa_temperature.iterrows():
93     if row['zscore'] > 3.0:
94         row['zscore'] = np.nan
95         row['temperature'] = np.nan
96     elif row['zscore'] < -3.0:
97         row['zscore'] = np.nan

```

```

98         row[ 'temperature ' ] = np.nan
99     else:
100         continue
101
102 for index , row in bdmap_diario_temperature.iterrows():
103     if row[ 'zscore ' ] > 3.0:
104         row[ 'zscore ' ] = np.nan
105         row[ 'temperature ' ] = np.nan
106     elif row [ 'zscore ' ] < -3.0:
107         row[ 'zscore ' ] = np.nan
108         row[ 'temperature ' ] = np.nan
109     else:
110         continue
111
112 for index , row in bdmap_horario_temperature.iterrows():
113     if row[ 'zscore ' ] > 3.0:
114         row[ 'zscore ' ] = np.nan
115         row[ 'temperature ' ] = np.nan
116     elif row [ 'zscore ' ] < -3.0:
117         row[ 'zscore ' ] = np.nan
118         row[ 'temperature ' ] = np.nan
119     else:
120         continue
121
122 #P/%
123 XY_incomplete = sonda_temperature
124
125 n_imputations = 5
126 XY_completed = []
127 for i in range(n_imputations):
128     imputer = IterativeImputer(n_iter=5, sample_posterior=True,
129                                random_state=i)
129     XY_completed.append(imputer.fit_transform(XY_incomplete))
130
131 XY_completed_mean_sonda = np.mean(XY_completed, 0)
132 XY_completed_std = np.std(XY_completed, 0)
133
134 #P/%
135 XY_incomplete = sinda_temperature

```

```

136
137 n_imputations = 5
138 XY_completed = []
139 for i in range(n_imputations):
140     imputer = IterativeImputer(n_iter=5, sample_posterior=True,
141                               random_state=i)
142     XY_completed.append(imputer.fit_transform(XY_incomplete))
143
144 XY_completed_mean_sinda = np.mean(XY_completed, 0)
145 XY_completed_std = np.std(XY_completed, 0)
146
147 #%%
148 XY_incomplete = nasa_temperature
149
150 n_imputations = 5
151 XY_completed = []
152 for i in range(n_imputations):
153     imputer = IterativeImputer(n_iter=5, sample_posterior=True,
154                               random_state=i)
155     XY_completed.append(imputer.fit_transform(XY_incomplete))
156
157 XY_completed_mean_nasa = np.mean(XY_completed, 0)
158 XY_completed_std = np.std(XY_completed, 0)
159
160 #%%
161 XY_incomplete = bdmep_diario_temperature
162
163 n_imputations = 5
164 XY_completed = []
165 for i in range(n_imputations):
166     imputer = IterativeImputer(n_iter=5, sample_posterior=True,
167                               random_state=i)
168     XY_completed.append(imputer.fit_transform(XY_incomplete))
169
170 XY_completed_mean_bdmep_diario = np.mean(XY_completed, 0)
171 XY_completed_std = np.std(XY_completed, 0)
172
173 #%%
174 XY_incomplete = bdmep_horario_temperature

```

```

172
173 n_imputations = 5
174 XY_completed = []
175 for i in range(n_imputations):
176     imputer = IterativeImputer(n_iter=5, sample_posterior=True,
177                                random_state=i)
178     XY_completed.append(imputer.fit_transform(XY_incomplete))
179
180 XY_completed_mean_bdmep_horario = np.mean(XY_completed, 0)
181 XY_completed_std = np.std(XY_completed, 0)
182
183 #%%
184 result_temperature = np.mean( np.array([ XY_completed_mean_sonda,
185                                           XY_completed_mean_sinda,
186                                           XY_completed_mean_nasa,
187                                           XY_completed_mean_bdmep_diario,
188                                           XY_completed_mean_bdmep_horario]), axis=0 )
189
190 #%%
191 result_temperature = pd.DataFrame(data=result_temperature,
192                                   index=np.array(range(0, 365)),
193                                   columns=['temperature', 'zscore'])
194
195 #%%
196 bplot = sns.boxplot(data=result_temperature['temperature'], width=0.8,
197                     orient="h", palette="colorblind")
198 bplot.set_title("Variável temperature pré-processada – Boxplot")
199 plt.show()

```

wind_speed

```

1 # -*- coding: utf-8 -*-
2
3 sonda_wind_speed = sonda_formatted.loc[:, "wind_speed"]
4 sonda_wind_speed = sonda_wind_speed.to_frame()
5
6 sinda_wind_speed = sinda_formatted.loc[:, "wind_speed"]

```

```

7 sinda_wind_speed = sinda_wind_speed.to_frame()
8
9 nasa_wind_speed = nasa_formatted.loc[:, "wind_speed"]
10 nasa_wind_speed = nasa_wind_speed.to_frame()
11
12 bdmap_horario_wind_speed = bdmap_horario_formatted.loc[:, "wind_speed"]
13 bdmap_horario_wind_speed = bdmap_horario_wind_speed.to_frame()
14
15 bdmap_diario_wind_speed = bdmap_diario_formatted.loc[:, "wind_speed"]
16 bdmap_diario_wind_speed = bdmap_diario_wind_speed.to_frame()
17
18 #%% C  lculo do zscore de global_irradiance das bases SONDA, SINDA e NASA
19
20 sonda_wind_speed['zscore'] = (sonda_wind_speed - sonda_wind_speed.mean()) /
    sonda_wind_speed.std(ddof=0)
21 sinda_wind_speed['zscore'] = (sinda_wind_speed - sinda_wind_speed.mean()) /
    sinda_wind_speed.std(ddof=0)
22 nasa_wind_speed['zscore'] = (nasa_wind_speed - nasa_wind_speed.mean()) /
    nasa_wind_speed.std(ddof=0)
23 bdmap_horario_wind_speed['zscore'] = (bdmap_horario_wind_speed -
    bdmap_horario_wind_speed.mean()) / bdmap_horario_wind_speed.std(ddof=0)
24 bdmap_diario_wind_speed['zscore'] = (bdmap_diario_wind_speed -
    bdmap_diario_wind_speed.mean()) / bdmap_diario_wind_speed.std(ddof=0)
25
26 #%% Normaliza  o Zscore
27
28 for index, row in sonda_wind_speed.iterrows():
29     if row['zscore'] > 2:
30         row['zscore'] = np.nan
31         row['wind_speed'] = np.nan
32     elif row['zscore'] < -2:
33         row['zscore'] = np.nan
34         row['wind_speed'] = np.nan
35     else:
36         continue
37
38 for index, row in sinda_wind_speed.iterrows():
39     if row['zscore'] > 2:
40         row['zscore'] = np.nan

```

```

41         row[ 'wind_speed' ] = np.nan
42     elif row [ 'zscore' ] < -2:
43         row[ 'zscore' ] = np.nan
44         row[ 'wind_speed' ] = np.nan
45     else:
46         continue
47
48 for index, row in nasa_wind_speed.iterrows():
49     if row[ 'zscore' ] > 2:
50         row[ 'zscore' ] = np.nan
51         row[ 'wind_speed' ] = np.nan
52     elif row [ 'zscore' ] < -2:
53         row[ 'zscore' ] = np.nan
54         row[ 'wind_speed' ] = np.nan
55     else:
56         continue
57
58 for index, row in bdmap_diario_wind_speed.iterrows():
59     if row[ 'zscore' ] > 2:
60         row[ 'zscore' ] = np.nan
61         row[ 'wind_speed' ] = np.nan
62     elif row [ 'zscore' ] < -2:
63         row[ 'zscore' ] = np.nan
64         row[ 'wind_speed' ] = np.nan
65     else:
66         continue
67
68 for index, row in bdmap_horario_wind_speed.iterrows():
69     if row[ 'zscore' ] > 2:
70         row[ 'zscore' ] = np.nan
71         row[ 'wind_speed' ] = np.nan
72     elif row [ 'zscore' ] < -2:
73         row[ 'zscore' ] = np.nan
74         row[ 'wind_speed' ] = np.nan
75     else:
76         continue
77
78 #%%
79 XY_incomplete = sonda_wind_speed

```

```

80
81 n_imputations = 5
82 XY_completed = []
83 for i in range(n_imputations):
84     imputer = IterativeImputer(n_iter=5, sample_posterior=True,
85                                random_state=i)
86     XY_completed.append(imputer.fit_transform(XY_incomplete))
87
88 XY_completed_mean_sonda_wind_speed = np.mean(XY_completed, 0)
89 XY_completed_std = np.std(XY_completed, 0)
90
91 #%%
92 XY_incomplete = sinda_wind_speed
93
94 n_imputations = 5
95 XY_completed = []
96 for i in range(n_imputations):
97     imputer = IterativeImputer(n_iter=5, sample_posterior=True,
98                                random_state=i)
99     XY_completed.append(imputer.fit_transform(XY_incomplete))
100
101 XY_completed_mean_sinda_wind_speed = np.mean(XY_completed, 0)
102 XY_completed_std = np.std(XY_completed, 0)
103
104 #%%
105 XY_incomplete = nasa_wind_speed
106
107 n_imputations = 5
108 XY_completed = []
109 for i in range(n_imputations):
110     imputer = IterativeImputer(n_iter=5, sample_posterior=True,
111                                random_state=i)
112     XY_completed.append(imputer.fit_transform(XY_incomplete))
113
114 XY_completed_mean_nasa_wind_speed = np.mean(XY_completed, 0)
115 XY_completed_std = np.std(XY_completed, 0)
116
117 #%%
118 XY_incomplete = bdmeep_diario_wind_speed

```

```

116
117 n_imputations = 5
118 XY_completed = []
119 for i in range(n_imputations):
120     imputer = IterativeImputer(n_iter=5, sample_posterior=True,
121                                random_state=i)
122     XY_completed.append(imputer.fit_transform(XY_incomplete))
123
124 XY_completed_mean_bdmep_diario_wind_speed = np.mean(XY_completed, 0)
125 XY_completed_std = np.std(XY_completed, 0)
126
127 #%%
128 XY_incomplete = bdmep_horario_wind_speed
129
130 n_imputations = 5
131 XY_completed = []
132 for i in range(n_imputations):
133     imputer = IterativeImputer(n_iter=5, sample_posterior=True,
134                                random_state=i)
135     XY_completed.append(imputer.fit_transform(XY_incomplete))
136
137 XY_completed_mean_bdmep_horario_wind_speed = np.mean(XY_completed, 0)
138 XY_completed_std = np.std(XY_completed, 0)
139
140 #%%
141 result_wind_speed = np.mean( np.array([
142     XY_completed_mean_sinda_wind_speed,
143     XY_completed_mean_nasa_wind_speed,
144     XY_completed_mean_bdmep_diario_wind_speed,
145     XY_completed_mean_bdmep_horario_wind_speed]),
146                               axis=0 )
147
148 #%%
149 result_wind_speed = pd.DataFrame(data=result_wind_speed,
150                                  index=np.array(range(0, 365)),
151                                  columns=['wind_speed', 'zscore'])
152
153 #%%
154 bplot = sns.boxplot(data=result_wind_speed['wind_speed'], width=0.8, orient

```

```
    ="h", palette="colorblind")  
152 bplot.set_title("Variável wind_speed pré-processada – Boxplot")  
153 plt.show()
```