Universidade Federal de Juiz de Fora
Instituto de ciências exatas
Bacharelado em Sistemas de informação

BR-PVGen: Coleta, Sanitização e Caracterização de Dados de Geração Fotovoltaica Distribuída no Brasil

Rhuan Nascimento Ferreira

JUIZ DE FORA
JANEIRO, 2026

BR-PVGen: Coleta, Sanitização e Caracterização de Dados de Geração Fotovoltaica Distribuída no Brasil

Rhuan Nascimento Ferreira

Universidade Federal de Juiz de Fora
Instituto de ciências exatas
Departamento de ciência da computação
Bacharelado em Sistemas de informação

Orientador: Wagner Antonio Arbex
Coorientador: Eduardo Pestana de Aguiar

JUIZ DE FORA
JANEIRO, 2026

BR-PVGen: Coleta, Sanitização e Caracterização de Dados de Geração Fotovoltaica Distribuída no Brasil

Rhuan Nascimento Ferreira

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM SISTEMAS DE INFORMAÇÃO.

Aprovada por:

Wagner Antonio Arbex
Doutor em Engenharia de Sistemas e Computação

Eduardo Pestana de Aguiar
Doutor em Engenharia Elétrica

Luiz Maurílio da Silva Maciel
Doutor em Engenharia de Sistemas e Computação

Jose Maria Nazar David
Doutorado em Engenharia de Sistemas e Computação

JUIZ DE FORA

16 DE JANEIRO, 2026

Resumo

A rápida expansão da geração distribuída fotovoltaica no Brasil enfrenta um gargalo técnico significativo: a escassez de dados operacionais públicos e padronizados que reflitam as condições climáticas locais. Este trabalho apresenta a sanitização e a caracterização do conjunto de dados BR-PVGen, composto por informações de 51 usinas fotovoltaicas em operação comercial nas regiões Sudeste e Centro-Oeste do Brasil. A metodologia de aquisição utilizou um sistema SCADA baseado em nuvem para consolidar dados heterogêneos de inversores e estações solarimétricas. Para garantir a integridade das séries temporais, foi aplicado um processo de ETL que incluiu a padronização temporal em janelas de 15 minutos via Média Móvel Ponderada e mitigação de dados ausentes por interpolação linear. O resultado é um *dataset* com mais de 15 milhões de registros, abrangendo o período de março de 2024 a junho de 2025. O conjunto de dados encontra-se disponível publicamente em formatos CSV e JSON. Diferentemente de bases internacionais, este conjunto disponibiliza variáveis raras em dados públicos, como potência reativa, temperatura interna dos inversores e medições ambientais *on-site*. O trabalho contribui de forma relevante para a comunidade científica ao disponibilizar insumos fundamentais para o desenvolvimento de modelos de previsão de geração, diagnóstico de falhas, bem como para o desenvolvimento e a validação de novas tecnologias e a realização de estudos de qualidade de energia, considerando as particularidades da realidade tropical brasileira.
Palavras-chave: Energia Solar Fotovoltaica. Conjunto de Dados. Geração Distribuída. Séries Temporais. Qualidade de Dados.

Abstract

The rapid expansion of distributed photovoltaic generation in Brazil faces a significant technical bottleneck: the scarcity of publicly available, standardized operational data that accurately reflect local climatic conditions. This work presents the sanitization and characterization of the BR-PVGen dataset, composed of data from 51 photovoltaic power plants in commercial operation in the Southeast and Central-West regions of Brazil. The data acquisition methodology employed a cloud-based SCADA system to consolidate heterogeneous data from inverters and solarimetric stations. To ensure the integrity of the time series, a ETL process was applied, including temporal standardization into 15-minute intervals using a Weighted Moving Average and mitigation of missing data through linear interpolation. The resulting dataset comprises more than 15 million records, covering the period from March 2024 to June 2025. The dataset is publicly available in CSV and JSON formats. Unlike international databases, this dataset provides access to variables that are rarely available in public data sources, such as reactive power, internal inverter temperature, and on-site environmental measurements. This work contributes to the scientific community by providing robust and high-quality data that support the development of power generation forecasting models, fault diagnosis methodologies, as well as the development and validation of emerging technologies and power quality studies, explicitly accounting for the specific conditions of the Brazilian tropical climate.

Keywords: Photovoltaic Solar Energy. Dataset. Distributed Generation. Time Series. Data Quality.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

AC	Alternating Current (Corrente Alternada)
API	Application Programming Interface
CSV	Comma-Separated Values
DC	Direct Current (Corrente Contínua)
DCC	Departamento de Ciência da Computação
DKASC	Desert Knowledge Australia Solar Centre
GHI	Global Horizontal Irradiance (Irradiância Global Horizontal)
GRI	Global Radiation on Inclined plane (Irradiância Global no plano inclinado)
GW	Gigawatt
HTTPS	Hypertext Transfer Protocol Secure
IA	Inteligência Artificial
JSON	JavaScript Object Notation
kWp	Kilowatt-pico
LGPD	Lei Geral de Proteção de Dados
MW	Megawatt
MWp	Megawatt-pico
PICO	Problem, Intervention, Comparison, Outcome
POA	Plane of Array (Plano do Arranjo)
RSL	Revisão Sistemática da Literatura
SCADA	Supervisory Control and Data Acquisition
SQL	Structured Query Language
UFJF	Universidade Federal de Juiz de Fora

1 Introdução

A crescente demanda global por energia, impulsionada pelo crescimento populacional, avanços tecnológicos e mudanças no estilo de vida, levou as nações a diversificarem suas fontes de energia. Essa mudança é ainda mais alimentada pela necessidade urgente de mitigar as mudanças climáticas, causadas principalmente pela queima de combustíveis fósseis, bem como por preocupações sobre a disponibilidade limitada desses recursos e a instabilidade resultante de conflitos geopolíticos. Nesse contexto, a energia solar surgiu como uma solução de destaque devido à sua ampla acessibilidade e potencial.

De acordo com a Agência Internacional de Energia, a energia solar fotovoltaica foi a fonte de energia renovável de crescimento mais rápido em 2023, com mais de 270 GW de nova capacidade adicionada em todo o mundo, superando todas as outras tecnologias de geração de energia (International Energy Agency, 2023). Os investimentos em energia solar atingiram aproximadamente US$1.8 trilhão no mesmo ano, refletindo um aumento substancial em comparação com anos anteriores (BloombergNEF, 2024). Olhando para o futuro, as projeções indicam que a energia solar está pronta para se tornar a principal fonte de geração de eletricidade até 2050 (International Renewable Energy Agency, 2023). Esse crescimento é apoiado pela queda nos custos da tecnologia, estruturas regulatórias cada vez mais favoráveis e compromissos globais com a descarbonização.

O Brasil emergiu como um líder regional na adoção de energia renovável, graças à sua irradiação solar favorável e vasto território. Em 2023, as fontes renováveis representaram aproximadamente 89,2% da matriz elétrica do país, tornando-a uma das participações mais altas globalmente (Empresa de Pesquisa Energética (EPE), 2024). Entre essas fontes, a energia solar fotovoltaica experimentou o crescimento mais rápido, com uma capacidade instalada total de 37,8 GW em 2023. Isso representa um aumento de 54,8% em relação ao ano anterior (Empresa de Pesquisa Energética (EPE), 2024).

Um fator significativo nesse crescimento é o segmento de Geração Distribuída, regulamentado no Brasil pela Lei nº 14.300/2022, que permite aos consumidores gerar sua própria eletricidade e receber compensação por meio do sistema de compensação de

energia (Agência Nacional de Energia Elétrica (ANEEL), 2024).

Entretanto, a massiva integração de fontes intermitentes e descentralizadas, como a solar fotovoltaica, impõe novos desafios à operação e ao planejamento do sistema elétrico nacional. Para garantir a estabilidade da rede e otimizar a eficiência desses sistemas, torna-se imprescindível a realização de estudos aprofundados sobre o comportamento operacional real dessas usinas em condições climáticas brasileiras. A validação de tais estudos depende, fundamentalmente, da disponibilidade de registros históricos confiáveis e granulares.

1.1 Descrição do Problema

O problema central abordado neste trabalho consiste na ausência de um conjunto de dados público, abrangente e padronizado que represente o comportamento operacional real de usinas fotovoltaicas de geração distribuída no Brasil. A inexistência desse tipo de recurso limita o avanço de pesquisas que dependem de informações confiáveis para análises de desempenho, validação de modelos e desenvolvimento de soluções voltadas ao setor fotovoltaico nacional.

Superar essa lacuna exige mais do que a simples coleta de dados: demanda a aplicação de metodologias de sanitização e padronização de modo a garantir consistência, reprodutibilidade e utilidade científica ao conjunto de dados resultante (AHMAD et al., 2024; NIHAR et al., 2021).

1.2 Motivação e Justificativa

Apesar do crescimento notável, o setor enfrenta desafios significativos relacionados à disponibilidade de dados operacionais do mundo real. Esses dados são cruciais para realização de pesquisas na área.

Enquanto conjunto de dados internacionais como o DKASC (DKASC, 2024) e o FAIR PV (NIHAR et al., 2021) são amplamente referenciados, poucos conjunto de dados capturam a realidade operacional de países de clima tropical com condições regionais diversas como o Brasil. A ausência de um repositório de dados público limita o desenvol-

vimento de pesquisas aprofundadas e a inovação tecnológica adaptada às especificidades do cenário nacional. Este trabalho é motivado pela necessidade de preencher essa lacuna, fornecendo um recurso valioso para a comunidade científica e o setor de energia.

1.3 Objetivos

O objetivo deste trabalho é desenvolver, caracterizar e disponibilizar um conjunto de dados proveniente de usinas fotovoltaicas de geração distribuída em operação no Brasil. Esse recurso visa apoiar a pesquisa científica, impulsionar o desenvolvimento tecnológico e estimular a inovação no setor de energia solar nacional. Para atingir tal objetivo, o estudo envolve os seguintes desdobramentos:

- Coletar e centralizar dados operacionais de inversores e dados meteorológicos de estações solarimétricas de usinas fotovoltaicas distribuídas em diferentes estados do Brasil.
- Desenvolver e aplicar uma metodologia sistemática de sanitização e pré-processamento de dados para tratar valores ausentes e inconsistências.
- Estruturar o conjunto de dados em um formato lógico e modular, para facilitar a manipulação e a integração com fluxos de trabalho de ciência de dados.
- Caracterizar detalhadamente o conjunto de dados proposto, descrevendo suas variáveis, escopo geográfico e temporal, e realizando uma análise comparativa com conjunto de dados internacionais para destacar suas contribuições únicas.
- Disponibilizar o conjunto de dados resultante para a comunidade acadêmica e industrial, promovendo a ciência reprodutível e avanços em áreas como previsão de geração, diagnóstico de falhas e otimização de desempenho de sistemas fotovoltaicos.

1.4 Questões de Pesquisa

Para guiar o desenvolvimento deste trabalho, foram formuladas as seguintes questões de pesquisa.

A formulação das questões norteadoras deste trabalho decorre dos desafios práticos observados no contexto profissional do setor elétrico. A vivência operacional com dados da área permitiu constatar que, embora o volume de dados coletados seja crescente, a padronização dessas informações pode apresentar limitações, o que exige um esforço prévio de engenharia de dados para adequação.

Portanto, a investigação foca em como transpor a barreira entre a aquisição de dados brutos e a disponibilização. As questões a seguir foram estruturadas para cobrir o ciclo de vida do dado, visando resolver os entraves técnicos enfrentados tanto na academia quanto na indústria.

1.4.1 Pergunta Principal

Quais são os principais processos metodológicos envolvidos na construção, sanitização e organização estrutural de um conjunto de dados proveniente de usinas fotovoltaicas de geração distribuída, de modo a garantir uma base de informações para análises operacionais e estudos setoriais aprofundados?

1.4.2 Perguntas Secundárias

Sobre a Sanitização dos Dados

- Quais são os desafios metodológicos e as abordagens para identificar, tratar e validar valores ausentes de séries temporais coletados de múltiplos sensores?

Sobre a Organização e Estruturação dos Dados

- Qual modelo de organização e estruturação de dados é adequado para um conjuntos de dados com um grande volume de registros de diferentes fontes, de modo a facilitar sua manipulação, consulta e interoperabilidade com ferramentas de análise?
- Como informações provenientes de diversas fontes e equipamentos podem ser harmonizadas e integradas em uma estrutura de dados coesa e logicamente organizada?

Sobre a Contribuição e Caracterização do Conjunto de Dados

- Quais são as características distintivas do conjunto de dados resultante e como ele se posiciona para suprir lacunas identificadas em fontes previamente disponíveis?
- Como a metodologia empregada na construção deste conjunto de dados assegura sua utilidade como uma referência para futuras pesquisas?

2 Revisão Sistemática da Literatura

No âmbito deste trabalho, cujo foco reside no desenvolvimento e na caracterização de um novo e robusto conjunto de dados, aplicar uma RSL é fundamental para:

- Contextualizar a pesquisa, identificando a lacuna científica e a demanda por conjunto de dados públicos, abrangentes e detalhados sobre o desempenho de plantas fotovoltaicas de geração distribuída, especialmente no cenário brasileiro.
- Analisar conjunto de dados ou metodologias para coleta e tratamento de informações similares existentes na literatura, compreendendo suas principais características, limitações e oportunidades de aprimoramento que justificam a presente proposta.
- Identificar e discutir as potenciais aplicações do conjunto de dados proposto.

2.1 PICO

A clareza no escopo de um estudo é primordial. Para delinear de forma precisa os contornos desta pesquisa e o desenvolvimento do conjunto de dados associado, foi definida a seguinte estratégia PICO:

P (Population): Conjuntos de dados públicos ou devidamente documentados, que sejam abrangentes e devidamente sanitizados, o que dificulta o desenvolvimento de análises avançadas no setor de energia solar de geração distribuída.

I (Intervention): A intervenção principal aqui é o processo de desenvolvimento do conjunto de dados. Isso inclui:

- A sanitização dos dados brutos coletados (tratamento de erros, valores ausentes, outliers).
- A estruturação dessas informações de forma organizada e pronta para uso.

C (Comparison): Não aplicável. Devido à natureza deste trabalho, caracterizado pela proposição de um artefato e não por um estudo experimental comparativo, não foi estabelecido um grupo de controle ou uma técnica concorrente específica.

O (Outcome): O principal desfecho é a criação e caracterização de um novo e valioso conjunto de dados, especificamente:

- Um conjunto de dados sanitizado e detalhado, disponível publicamnete, pronto para ser utilizado em pesquisa e desenvolvimento.
- A demonstração do potencial desse conjunto de dados para aplicações futuras.

2.2 Estratégia de Busca e Seleção dos Estudos

A presente seção detalha o protocolo metodológico sistemático empregado para a identificação e seleção dos estudos que compõem esta revisão. Este protocolo é fundamental para garantir a rastreabilidade e reprodutibilidade do processo e compreende três etapas principais: a definição das palavras-chave a partir da questão de pesquisa e da estratégia PICO, a construção e aplicação da *string* de busca nas bases de dados científicas, e, por fim, o estabelecimento e aplicação dos critérios para a triagem e seleção dos artigos relevantes.

2.2.1 Definição das Palavras-Chave para a Busca

O primeiro passo para uma busca sistemática eficaz é a identificação precisa dos termos que representam os conceitos centrais da pesquisa. Neste estudo, as palavras-chave foram derivadas da estratégia PICO previamente definida na Seção 2.1, garantindo que a busca seja abrangente, mas focada na questão de pesquisa.

Para o "P", consideraram-se palavras relacionadas a plantas fotovoltaicas de geração distribuída, como "photovoltaic system*", "PV system*", "solar photovoltaic", "solar plant*", "PV plant*", "solar farm*", "distributed generation" ou "decentralized generation", bem como termos ligados a dados, incluindo "SCADA data", "sensor data", "operational data" e "time series data", e expressões associadas à lacuna de dados, como

"dataset*" ou "data set*", "open data" e "data quality".

Para o "I", incluíram-se palavras relacionadas à sanitização de dados, como "data sanitization", "error detection", "missing data" e "data validation", além de termos ligados à organização e estruturação de dados, como "data structure*", "data organization" e "data model*".

Finalmente, para o "O", consideraram-se palavras relacionadas à caracterização e utilidade do conjunto de dados, como "dataset characterization" e "data description".

Cabe destacar que o componente "C" da estratégia PICO não foi incorporado diretamente à definição das palavras-chave para a busca sistemática. Diferentemente dos demais elementos, o "C" não tem como objetivo auxiliar na identificação de estudos relacionados ao tema, mas sim fornecer um referencial para a avaliação e validação dos resultados obtidos neste trabalho.Assim, o componente "C" é aplicado em uma etapa posterior à busca bibliográfica, sendo fundamental para a análise crítica e contextualização das contribuições do conjunto de dados proposto, mas não para a recuperação inicial de estudos na literatura.

Com todas as palavras-chave definidas, a etapa seguinte consistiu na combinação lógica desses termos para formar uma *string* de busca robusta, visando maximizar a recuperação de estudos relevantes.

2.2.2 String de Busca e Aplicação nas Bases de Dados

Utilizando as palavras-chave definidas e operadores booleanos (AND, OR), foi construída a *string* de busca da Listagem 2.1. A aplicação desta *string* foi direcionada às bases de dados IEEE Xplore[1] e Scopus[2], escolhidas por sua ampla cobertura nas áreas de engenharia, tecnologia e ciências aplicadas, relevantes para este estudo.

Listagem 2.1: String de Busca

```
((
   "photovoltaic system*" OR "PV system*" OR
   "solar plant*" OR "PV plant*" OR
   "solar photovoltaic" OR "solar farm" OR
```

[1] https://ieeexplore.ieee.org
[2] https://www.scopus.com

```
   "distributed generation" OR
   "decentralized generation"
)) AND ((
   "dataset*" OR "data set*" OR
   "public data" OR "open data"
)) AND ((
   "monitoring data" OR "SCADA data" OR
   "sensor data" OR "operational data" OR
   "time series data"
)) AND ((
   "data clean*" OR "data sanitization" OR
   "error detection" OR "missing data" OR
   "outlier detection" OR "data validation"
   ) OR (
   "data structure*" OR "data organization" OR
   "data model*"
   ) OR (
   "dataset characterization" OR
   "data description" OR "data quality"
))
```

Aplicação da String:

- **Scopus:** Recomenda-se envolver toda a *string* com o operador `TITLE-ABS-KEY(...)`, aplicando-a assim aos campos de título, resumo e palavras-chave.
- **IEEE Xplore:** A *string* deve ser utilizada diretamente.

Após a execução desta *string* de busca nas bases de dados e a compilação dos resultados iniciais, os estudos identificados foram submetidos a um processo de triagem para determinar sua elegibilidade para inclusão na revisão, conforme os critérios de exclusão detalhados na Subsecção 2.2.3.

2.2.3 Critérios de Exclusão e Processo de Seleção

Para garantir a relevância e o foco dos estudos incluídos nesta revisão, estes critérios foram definidos para serem utilizados em duas fases principais de triagem: primeiramente, na

leitura de títulos e resumos dos artigos recuperados e, subsequentemente, na análise do texto completo dos estudos pré-selecionados que não foram excluídos na fase inicial. Os critérios de exclusão definidos para este estudo são:

EC1 Estudos não publicados nos idiomas inglês ou português.

EC2 Foco não relacionado a plantas fotovoltaicas.

EC3 Ausência de abordagem sobre conjunto de dados.

EC4 Uso de conjunto de dados de plantas fotovoltaicas apenas para aplicações finais, sem detalhamento metodológico da criação, tratamento ou características do conjunto de dados em si.

EC5 Foco exclusivo em hardware de sistemas de monitoramento ou comunicação, sem discussão sobre o conjunto de dados gerado ou suas métricas.

EC6 Tipos de publicação: teses, dissertações, capítulos de livros, resumos ou apresentações de conferência sem artigo completo associado, revisões sistemáticas, mapeamentos ou *surveys*.

EC7 Publicações duplicadas ou redundantes do mesmo estudo; neste caso, apenas a versão mais completa ou recente será considerada.

2.2.4 Processo de Identificação e Seleção de Estudos

O processo de revisão sistemática da literatura iniciou-se com a Identificação dos estudos (Estágio 1). A busca foi conduzida nas bases de dados IEEE Xploree e Scopus, utilizando a *string* de busca definida na Seção 2.2.2. Nesta etapa inicial, foram identificados os seguintes quantitativos de registros: 61 da base IEEE Xplore e 9 da base Scopus, totalizando 70 estudos potencialmente relevantes identificados pelas buscas em bases de dados.

Após a coleta, estes 70 registros foram importados para a ferramenta de auxílio à revisão sistemática Rayyan[3]. Com o suporte desta ferramenta, procedeu-se à identificação e remoção de duplicatas, resultando na exclusão de 5 estudos. Desta forma, um conjunto

[3]https://www.rayyan.ai

de 65 estudos únicos avançou para a subsequente fase de Triagem e Elegibilidade (Estágio 2).

No **Estágio 2 (Triagem e Elegibilidade)**, os 65 estudos únicos foram submetidos a uma análise criteriosa, dividida em duas fases:

- **Fase 1 (Triagem por Título e Resumo):** Realizou-se a leitura dos títulos e resumos dos 65 estudos. Com base nos Critérios de Exclusão (ECs) definidos anteriormente na Seção 2.2.3, 51 estudos foram excluídos nesta fase. Os 14 estudos restantes, que atenderam aos critérios ou geraram dúvidas, prosseguiram para a próxima fase.

- **Fase 2 (Avaliação do Texto Completo para Elegibilidade):** Os 14 estudos selecionados na Fase 1 tiveram seus textos completos recuperados e lidos na íntegra. Nesta fase, todos os Critérios de Exclusão (EC1 a EC7) foram rigorosamente aplicados. Como resultado desta análise detalhada, 10 estudos foram excluídos.

Ao final do Estágio 2, **4 estudos** foram considerados elegíveis e compõem o conjunto final de artigos incluídos nesta revisão sistemática. Estes seguirão para o **Estágio 3 (Extração e Síntese dos Dados)**. A quantidade de estudos em cada etapa do processo é visualmente detalhada no fluxograma apresentado na Figura 2.1.

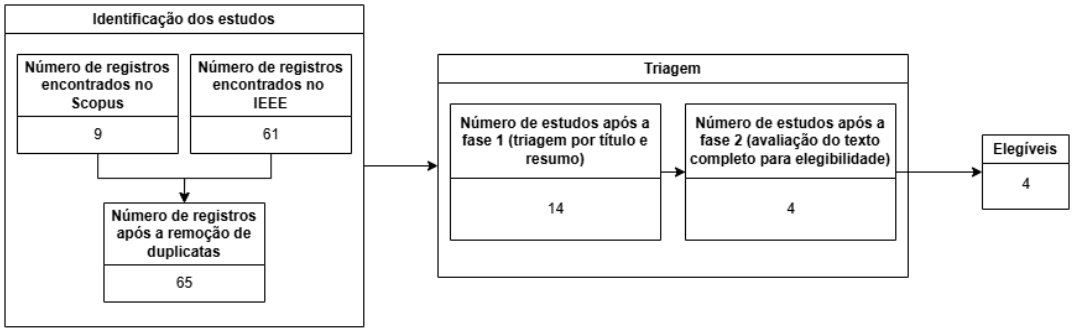

Figura 2.1: Fluxograma detalhando o processo de identificação, triagem, elegibilidade e inclusão dos estudos na revisão.

2.2.5 Estudos Selecionados

Após a aplicação dos critérios de inclusão e exclusão, os estudos selecionados nesta revisão são apresentados na Tabela 2.1. A tabela lista o título de cada estudo, seus respectivos

autores e o ano de publicação.

Tabela 2.1: Lista dos estudos selecionados.

Título do Estudo	**Autor(es)**	**Ano**
Performance Evaluation of AI-Driven Photovoltaic Output Forecasting	Haddad, H. and Jerbi, F. and Smaali, I.	2024
Toward Findable, Accessible, Interoperable and Reusable (FAIR) Photovoltaic System Time Series Data	Nihar, A. and Curran, A.J. and Karimi, A.M. and Braid, J.L. and Bruckman, L.S. and Koyuturk, M. and Wu, Y. and French, R.H.	2021
Integrated Tool for Cleaning Bulk Solar Power Data	Ahmad, E. and George, V. and Soman, D. and Kanthaliya, R. and Kumar, D. and Pateriya, V.	2024
Real-world Challenges and Opportunities in Degradation Rate Analysis for Commercial PV Systems	Sauer, K.J.	2011

A baixa quantidade de estudos recuperados (4) reforça a escassez de literatura focada especificamente na metodologia de construção de consjunto de dados públicos, evidenciando a motivação para o desenvolvimento deste trabalho.

2.3 Respostas às Perguntas Orientadoras

Esta seção contém as respostas das perguntas orientadoras apresentadas na Seção 1.4.1.

2.3.1 Resposta à Pergunta Principal

Em relação à pergunta principal, a construção de um conjunto de dados pede uma metodologia sistemática composta pelos seguintes processos fundamentais:

1. **Construção e Definição do Esquema:** Envolve o perfilamento dos dados para compreender cada variável, definindo tipos, unidades e significados.

2. **Sanitização e Tratamento de Dados:** Foca na consistência, tratando anomalias como dados ausentes e inconsistentes (AHMAD et al., 2024; HADDAD; JERBI; SMAALI, 2024).

3. **Organização Estrutural e Modelagem:** Define como os dados serão armazenados e relacionados, com uso de identificadores únicos e modelos lógicos (NIHAR et al., 2021).

2.3.2 Resposta à Pergunta sobre Sanitização dos Dados

Em resposta à pergunta 1.4.2, os principais desafios metodológicos no tratamento de valores ausentes em séries temporais são:

- **Escalabilidade e Desempenho Computacional:** Devido ao grande volume de dados, são necessárias abordagens eficientes (AHMAD et al., 2024).
- **Escolha e Validação do Método de Imputação:** A seleção inadequada pode introduzir vieses. Algumas abordagens incluem:
    1. Interpolação linear (HADDAD; JERBI; SMAALI, 2024).
    2. Uso de média ou mediana de períodos próximos (AHMAD et al., 2024).

2.3.3 Resposta às Perguntas sobre Organização e Estruturação dos Dados

No que se refere à definição de um modelo de organização e estruturação de dados adequado para conjuntos de dados com um grande volume de registros de diferentes fontes, este trabalho adota uma abordagem baseada na separação lógica em entidades distintas e inter-relacionadas, evitando a utilização de arquivos monolíticos (NIHAR et al., 2021). Essa estratégia favorece a modularidade, a escalabilidade e a interoperabilidade do conjunto de dados com diferentes ferramentas de análise.

Quanto à harmonização e integração de informações oriundas de múltiplas fontes e equipamentos, o processo é realizado por meio das seguintes etapas principais:

1. **Criação de Identificadores Únicos:** Utilizados como chaves de ligação entre diferentes entidades e fontes de dados, permitindo a correta associação das informações (NIHAR et al., 2021).

2. **Padronização de Dados:** Envolve a conversão de unidades de medida, a normalização de fusos horários e a adoção de uma nomenclatura uniforme para variáveis e campos, garantindo consistência e coerência ao conjunto de dados (NIHAR et al., 2021).

2.3.4 Resposta às Perguntas sobre Caracterização do Conjunto de Dados

No que diz respeito às características distintivas do conjunto de dados resultante, observa-se que ele apresenta:

- **Heterogeneidade:** Integra dados provenientes de múltiplas plantas fotovoltaicas e diferentes equipamentos, contemplando diversidade geográfica e operacional (HADDAD; JERBI; SMAALI, 2024).
- **Riqueza Granular:** Disponibiliza uma coleta abrangente de medições elétricas e climáticas, oferecendo elevado nível de detalhamento temporal e operacional (NIHAR et al., 2021; HADDAD; JERBI; SMAALI, 2024).

No que se refere à forma como a metodologia adotada assegura a utilidade do conjunto de dados como referência para pesquisas futuras, destacam-se os seguintes aspectos:

1. **Padronização e Transparência:** Garantidas por meio da documentação detalhada dos processos de aquisição, sanitização e estruturação, bem como pela disponibilização de um dicionário de dados (NIHAR et al., 2021).
2. **Reprodutibilidade e Extensibilidade:** A metodologia proposta pode ser replicada e expandida para a incorporação de novas plantas e fontes de dados (NIHAR et al., 2021).
3. **Base para Análises Avançadas:** O conjunto de dados oferece suporte a aplicações em inteligência artificial, estudos de *benchmarking* e subsídios para a formulação de políticas públicas (HADDAD; JERBI; SMAALI, 2024).

3 Metodologia

A base de dados foi construída a partir do registro sistemático de informações operacionais de usinas fotovoltaicas localizadas em diferentes estados do Brasil.

A coleta foi realizada por meio de um sistema SCADA[4] baseado em nuvem, responsável por receber, consolidar e transmitir continuamente dados provenientes de registradores instalados nas plantas. Essa infraestrutura permitiu a integração eficiente de medições em tempo real em um ambiente centralizado de armazenamento.

Em termos quantitativos, o monitoramento abrange o período contínuo de 26 de março de 2024 a 9 de junho de 2025. O volume de dados brutos processados totaliza mais de 49 milhões de registros, divididos entre medições de inversores e estações solarimétricas, conforme detalhado na Tabela 3.1.

Tabela 3.1: Resumo das características do conjunto de dados bruto.

Característica	**Inversor**	**Estação Solarimétrica**
Período do Conjunto de Dados	26 de março de 2024 – 9 de junho de 2025	
Número Total de Registros	44.092.970	5.669.460

3.1 Aquisição

O processo de aquisição dos dados abrange desde a leitura nos dispositivos de campo até o armazenamento seguro em nuvem, garantindo a integridade e a disponibilidade das séries temporais para as etapas subsequentes de processamento. A arquitetura de aquisição utilizda na construção deste conjunto está ilustrado na Figura 3.1.

Em cada usina, um registrador de dados local coleta continuamente medições provenientes dos dispositivos de campo, com resolução temporal de cinco minutos, utilizando o protocolo de comunicação Modbus[5] (A).

[4]É uma arquitetura de supervisão e controle composta por computadores, redes de comunicação e interfaces gráficas utilizadas para monitoramento e gerenciamento de processos em tempo real.

[5]Modbus é um protocolo de comunicação industrial amplamente utilizado em sistemas de automação para troca de dados entre controladores, sensores e dispositivos eletrônicos, operando em arquiteturas mestre–escravo e suportando diferentes meios físicos, como RS-485 e TCP/IP.

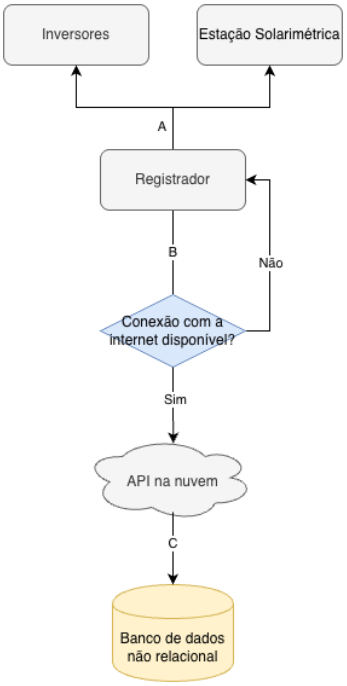

Figura 3.1: Diagrama da metodologia de aquisição.

Após a coleta, o registrador tenta estabelecer comunicação com uma API[6] hospedada em uma plataforma em nuvem (B). A transmissão das medições é realizada exclusivamente por meio do protocolo HTTPS[7], garantindo segurança e integridade na transferência dos dados.

Em situações de perda de conexão ou instabilidade de rede, o registrador armazena localmente as medições de forma temporária. Assim que a conectividade é restabelecida, os dados pendentes são automaticamente retransmitidos para a API, assegurando continuidade e minimizando lacunas na série temporal.

Uma vez recebidas pela API, as informações são processadas e armazenadas em um banco de dados não relacional em nuvem (C). Essa arquitetura de aquisição e transmissão promove tolerância a falhas, robustez operacional e entrega confiável dos dados.

Além disso, o sistema também possui uma instância de um banco de dados relacional na nuvem, responsável pelo armazenamento dos metadados das usinas, permitindo o cruzamento dessas informações.

[6]Um conjunto de rotinas e protocolos que permite a interação estruturada entre diferentes sistemas de software.

[7]Protocolo de transferência de hipertexto. É a base para a comunicação de dados na internet

3.2 Entidades

Nesta seção, apresentam-se as entidades disponibilizadas para este estudo, a saber: "Inversor", responsável pelos dados operacionais dos equipamentos; "Estação Solarimétrica", que registra os dados meteorológicos coletados localmente nas plantas; e os "Metadados das Usinas", contendo informações auxiliares relevantes para análise e integração dos dados.

A organização destas informações reflete a arquitetura de armazenamento apresentada na Seção 3.1, classificando os dados em dois grupos distintos de acordo com sua natureza:

- **Dados de Séries Temporais:** Englobam as entidades "Inversor" e "Estação Solarimétrica". Estes registros correspondem ao fluxo contínuo de leitura dos sensores. Devido ao elevado volume de ingestão e à característica sequencial, estes dados foram armazenados no banco de dados não relacional da arquitetura em nuvem descrita.
- **Dados Cadastrais:** Representados pela entidade "Metadados", que consolida as informações construtivas e geográficas de cada usina. Dada a natureza estruturada e estática dessas informações, elas encontram-se armazenadas em um banco de dados relacional, provendo o contexto necessário para a análise dos dados operacionais.

3.3 Extração

A presente seção detalha os procedimentos adotados para a extração dos registros da base de dados original. Esta etapa é fundamental para a integridade do estudo, assegurando a seleção de informações relevantes e a estruturação adequada para as análises subsequentes.

Como descrito na Seção 3.2, os metadados residem em um sistema de gerenciamento de banco de dados relacional. Consequentemente, a extração destes registros foi realizada mediante a execução de consultas SQL[8]. O procedimento foi operacionalizado por meio de uma rotina automatizada responsável pela conexão com o banco, execução

[8]Linguagem padronizada utilizada para comunicar-se com bancos de dados relacionais, permitindo a definição, manipulação e controle de acesso aos dados de forma estruturada.

da consulta e persistência dos dados em arquivos no formato JSON[9] .

Por outro lado, a natureza massiva dos dados de séries temporais exigiu uma estratégia de extração mais robusta. Para tanto, desenvolveu-se um algoritmo de automação otimizado para o processamento de grandes volumes de dados. Este procedimento executa além da recuperação dos documentos, uma etapa de pré-processamento com filtros para expurgar registros incompletos.

Um ponto igualmente importante refere-se à seleção dos campos extraídos. A base de dados original contém diversos atributos associados à lógica interna do sistema supervisório do qual os dados foram obtidos, mas que não possuem relevância analítica para o escopo deste trabalho. Assim, tais campos foram descartados, concentrando-se a extração exclusivamente nas variáveis operacionais necessárias para a caracterização e a composição do conjunto de dados final.

3.3.1 Filtros e Seleção de Campos

Para garantir a qualidade e relevância dos dados extraídos, foram aplicados filtros e realizada a seleção de atributos tanto para os registros dos inversores fotovoltaicos quanto para os dados meteorológicos. A definição dos campos apresentados a seguir foi pautada pela disponibilidade das informações no conjunto de dados original fornecido pela empresa parceira, sendo selecionadas apenas as variáveis que apresentavam pertinência direta com o escopo e os objetivos deste trabalho.

Filtros aplicados:

- Valores não nulos para registro de data/hora, ID da usina e, no caso dos inversores, ID do equipamento;

Campos selecionados:

- **Inversores:** potência ativa total (W), potência reativa total (VAR), potência DC total (W), temperatura interna do inversor (°C), ID da usina, ID do inversor, registro de data/hora;

[9]Um formato leve de intercâmbio de dados, estruturado em pares chave–valor e amplamente utilizado para armazenamento e transmissão de informações devido à sua simplicidade e legibilidade.

- **Estação solarimétrica:** registro de data/hora, ID da usina, irradiância global no plano inclinado (W/m²), irradiância no plano do arranjo (W/m²), irradiância global horizontal (W/m²), temperatura ambiente (°C), temperatura do painel (°C), velocidade e direção do vento, índice de albedo do tracker, índice de albedo, tensão da bateria (V), precipitação acumulada (mm).

3.3.2 Pseudocódigo do Procedimento de Extração

O pseudocódigo na Listagem 3.1 descreve o fluxo geral utilizado para exportação de documentos a partir da base de dados. Essa estrutura realiza conexão com a coleção desejada, executa a leitura em lotes, armazena cada conjunto em arquivos JSON independentes e encerra a operação somente após a leitura integral dos dados. Esse pseudocódigo serve como base para qualquer uma das duas modalidades de exportação (inversores ou estação solarimétrica), variando apenas os parâmetros de consulta apresentados na Seção 3.3.1.

Listagem 3.1: Procedimento Geral de Exportação

```
function ExportarDados(MONGO_URI, DB, COLLECTION, QUERY, PROJECTION,
   BATCH_SIZE):
   criarDiretorio(OUTPUT_DIR)
   client = ConectarMongo(MONGO_URI)
   collection = client[DB][COLLECTION]
   cursor = collection.find(
             query = QUERY,
             projection = PROJECTION,
             batch_size = 1000
          )

   lote = []
   indice = 0

   for doc in cursor:
      append(lote, doc)

      if tamanho(lote) == BATCH_SIZE:
         salvarJSON(lote, "export_" + indice + ".json")
```

```
        lote = []
        indice = indice + 1

  if tamanho(lote) > 0:
     salvarJSON(lote, "export_" + indice + ".json")
```

A adoção deste procedimento permitiu a extração de milhões de registros de forma segura, controlada e escalável, evitando a saturação de memória e reduzindo o tempo total do processo. O resultado desta etapa consistiu em múltiplos arquivos segmentados por lote, separados em diretórios de acordo com sua entidade, cuja reorganização estrutural será detalhada na Seção 3.4.

3.4 Organização

Esta seção apresenta o fluxo de processamento responsável pela segregação dos registros e pela sua organização temporal. Após a extração, os dados brutos passam por uma etapa de reestruturação lógica para facilitar o acesso e a análise. Nesse processo, os registros são separados em arquivos distintos, adotando como critérios principais de organização a usina de origem e o tipo de entidade (inversor ou estação solarimétrica).

Além da segregação, os dados contidos em cada arquivo são ordenados cronologicamente de acordo com o instante de registro. No caso dos arquivos de inversores, aplica-se ainda um critério adicional de desempate, ordenando os registros também pelo identificador do equipamento correspondente.

3.4.1 Pseudocódigo do Procedimento de Organização

O procedimento automatizado de organização e ordenação dos dados na Listagem 3.2 é estruturado em duas etapas principais.

Na primeira etapa, denominada “SegregarDados”, os arquivos de dados brutos são processados de forma incremental. Os registros são lidos e agrupados em memória de acordo com o identificador da usina e, no caso da entidade inversor, também pelo identificador do inversor. Em seguida, os registros são gravados em arquivos de saída

específicos para cada usina.

Na segunda etapa, denominada “OrdenarDados”, cada arquivo é carregado em memória para aplicação da ordenação. A ordenação é realizada com base na variável temporal “datetime” como chave primária e, para a entidade inversor, utiliza-se o `inverter_id` como critério secundário de desempate.

O mesmo procedimento é aplicado aos dados da estação solarimétrica, com algumas simplificações. Para essa entidade, não há necessidade de agrupamento ou ordenação secundária por identificador do inversor, mantendo-se apenas a segregação por usina e a ordenação cronológica principal pelo campo “datetime”.

Listagem 3.2: Pseudocódigo para Segregação e Ordenação de Dados

```
function ProcessarDados():
   DIRETORIO_ENTRADA = "../dados_brutos_json"
   DIRETORIO_SAIDA = "../dados_organizados_json"

   SegregarDados(DIRETORIO_ENTRADA, DIRETORIO_SAIDA)
   OrdenarDados(DIRETORIO_SAIDA)

function SegregarDados(DIRETORIO_ENTRADA, DIRETORIO_SAIDA):
   criarDiretorio(DIRETORIO_SAIDA)
   arquivos_entrada = buscarArquivosJSON(DIRETORIO_ENTRADA)

   for arquivo_entrada in arquivos_entrada:
      grupos_locais = {}
      for item in carregarItensJSON(arquivo_entrada):
         se item.ps_id e item.inverter_id existem:
            chave = (item.ps_id, item.inverter_id)
            adicionar item em grupos_locais[chave]

      for (ps_id, inverter_id), entradas in grupos_locais.items():
         arquivo_saida = DIRETORIO_SAIDA + "ps_" + ps_id + ".json"
         anexarJSON(arquivo_saida, entradas)

function OrdenarDados(DIRETORIO_SAIDA):
   arquivos_processados = buscarArquivosJSON(DIRETORIO_SAIDA)
```

```
for arquivo in arquivos_processados:
   registros = carregarJSON(arquivo)
   registros_ordenados = ordenar(registros,
                         chave_primaria=DATETIME,
                         chave_secundaria=INVERTER_ID)
   sobrescreverJSON(arquivo, registros_ordenados)
```

Após executar o processo para ambas as entidades, a estrutura de dados final consiste em dois diretórios distintos: um para inversores e outro para estações solarimétricas.

Dentro de cada um destes diretórios, os dados são segregados em arquivos individuais por usina. Todos os registros nestes arquivos encontram-se ordenados cronologicamente, sendo que, no caso dos inversores, aplica-se adicionalmente uma ordenação secundária pelo identificador do equipamento.

3.5 Sanitização

A sanitização aplicada ao conjunto de dados consiste em 2 etapas: (i) a padronização temporal por meio de uma média móvel ponderada, utilizada para reduzir a influência de oscilações abruptas e valores atípicos; e (ii) o tratamento de lacunas por meio de interpolação linear. Esses processos visam aprimorar a consistência temporal das séries analisadas.

3.5.1 Padronização das Séries Temporais e Tratamento de Outliers

Para criar uma base temporal uniforme, fundamental para análises subsequentes e para a modelagem em aprendizado de máquina, todas as séries temporais foram reamostradas em uma grade composta por intervalos de 15 minutos, cobrindo integralmente todo o período diário (SCHNIERER, 2023).

A transição da amostragem original (intervalos de aproximadamente 5 minutos) para a grade de 15 minutos foi realizada através de uma agregação baseada em Média Móvel Ponderada. Esta abordagem foi selecionada tanto para padronizar a resolução

quanto para atenuar a influência de flutuações de curto prazo e *outliers*.

O processo de agregação foi centrado em cada amostra de tempo alvo da nova grade, $T_k \in \{00{:}00, 00{:}15, \ldots, 23{:}45\}$. Para cada T_k, foi definida uma janela temporal simétrica de 15 minutos de largura (com $h = 7,5$ minutos). Apenas os pontos de dados brutos $x(t_i)$, cujo tempo de registro t_i se enquadra no intervalo $[T_k - h, T_k + h]$, foram considerados no cômputo do valor agregado $y(T_k)$.

Para priorizar os dados temporalmente mais próximos da amostra de tempo alvo, foi atribuído um peso w_i a cada observação x_i válida dentro da janela, definido como o inverso da distância temporal (em minutos) acrescida de uma unidade:

$$w_i = \frac{1}{1 + |t_i - T_k|}$$

Onde $|t_i - T_k|$ é a diferença absoluta em minutos entre o tempo da observação e a amostra de tempo central. O termo aditivo unitário no denominador previne indeterminações matemáticas quando a amostra coincide perfeitamente com o tempo alvo.

O valor agregado $y(T_k)$ foi então calculado como a soma ponderada normalizada, garantindo que a soma dos pesos seja unitária:

$$y(T_k) = \frac{\sum_i w_i \cdot x_i}{\sum_i w_i}$$

Esta formulação concede maior influência aos pontos de dados mais próximos do centro da janela, preservando tendências locais enquanto mitiga o impacto de ruídos e flutuações. Conforme demonstrado em Hassani, Kalantari e Ghodsi (2019), esta técnica melhora a estabilidade do sinal sem distorcer seus padrões essenciais, tornando-a adequada para tarefas subsequentes.

Pseudocódigo do Procedimento de Padronização das Séries Temporais e Tratamento de Outliers

O procedimento automatizado de regularização das séries temporais na Listagem 3.3 inicia pela construção de uma grade temporal uniforme com intervalos de 15 minutos. Para cada instante dessa grade, é calculada uma Média Móvel Ponderada, na qual os valores mais

próximos do tempo central recebem maior relevância.

A janela utilizada no cálculo é ajustada dinamicamente quando o centro se aproxima das extremidades da série, garantindo o uso adequado dos valores disponíveis. Além do valor suavizado, o procedimento registra, para cada variável processada, o número de pontos efetivamente utilizados no cálculo da média movel. Esse indicador, denominado `document_count`, atua como uma variável de controle que expressa a densidade de dados ao longo do tempo, permitindo avaliar a confiabilidade local da média ponderada obtida.

Listagem 3.3: Pseudocódigo do Procedimento de Padronização das Séries Temporais e Tratamento de Outliers

```
function CarregarItens(arquivo):
   retornar LerJSONStream(arquivo)

function CriarGradeTemporal(inicio, fim, intervaloMinutos):
   inicioAjustado = floorData(inicio)
   fimAjustado = ceilData(fim) - intervaloMinutos minutos
   grade = []
   t = inicioAjustado
   enquanto t <= fimAjustado:
      adicionar t em grade
      t = t + intervaloMinutos minutos
   retornar grade

function AjustarLimitesJanela(indice, centro, metade):
   minT = menor tempo em indice
   maxT = maior tempo em indice
   se centro - metade < minT:
      retornar [centro, centro + 2*metade]
   se centro + metade > maxT:
      retornar [centro - 2*metade, centro]
   retornar [centro - metade, centro + metade]

function CalcularWMA(serie, centro, janelaMinutos):
   metade = janelaMinutos / 2
```

```
   intervalo = AjustarLimitesJanela(serie.indice, centro, metade)
   valores = selecionar pontos da série no intervalo considerado
   valores = remover valores nulos de valores

   se tamanho(valores) = 0 então
      retornar (nulo, 0)

   distancias = |t_i - centro| em minutos
   pesos = 1 / (1 + distancias)
   wma = soma(pesos * valores) / soma(pesos)
   retornar (wma, tamanho(valores))

function ProcessarGrupo(itens):
   df = ConverterParaDataframe(itens)
   df = ordenar df por datetime
   df = definir datetime como indice de df
   grade = CriarGradeTemporal(df.minDatetime, df.maxDatetime, 15)
   resultados = []
   para cada t em grade:
      valores = {}
      contagens = {}
      para cada coluna em colunas_de_interesse:
         (y, c) = CalcularWMA(df[coluna], t, 15)
         valores[coluna] = y
         contagens[coluna] = c
      adicionar {
         datetime: t,
         valores...,
         document_count: contagens
      } em resultados
   retornar resultados

function Principal(diretorio):
   arquivos = ListarArquivos(diretorio, "*.json")
   para arquivo em arquivos:
      itens = CarregarItens(arquivo)
```

```
grupos = AgruparPor(itens, ["ps_id", "inverter_id"])


para cada grupo:
   salvar ProcessarGrupo(grupo)
```

No caso da entidade inversor, o agrupamento simultâneo por `ps_id` e `inverter_id` é necessário para garantir que o processamento seja realizado individualmente para cada equipamento, preservando a integridade das séries temporais específicas de cada unidade. Já para a entidade de estação solarimétrica, esse agrupamento adicional não é requerido, uma vez que os registros pertencem a um único equipamento por usina; assim, o agrupamento pode ser mantido apenas por `ps_id`.

Os campos submetidos ao procedimento no caso dos inversores foram: potência ativa total (W), potência reativa total (var), potência DC total (W) e temperatura interna do inversor (°C).

Para a estação solarimétrica, o tratamento foi aplicado às variáveis: irradiância global no plano inclinado (W/m²), irradiância no plano do arranjo (W/m²), irradiância global horizontal (W/m²), temperatura ambiente (°C), temperatura do painel (°C), velocidade e direção do vento, índice de albedo do *tracker*, albedo, e tensão da bateria (V).

3.5.2 Mitigação de Valores Ausentes

Apesar da aplicação da média móvel ponderada apresentada na Seção 3.5.1, que já contribui para reduzir a ocorrência de valores nulos ao transformar intervalos de 5 em 5 minutos em janelas de 15 minutos, empregou-se também a interpolação linear para complementar esse processo. Esse método permite preencher lacunas curtas de forma contínua, preservando a coerência temporal da série, sendo amplamente utilizado dado sua simplicidade e capacidade de manter a consistência entre pontos adjacentes (LEPOT; AUBIN; CLEMENS, 2017).

A interpolação linear foi aplicada às séries temporais resultantes da etapa descrita na Seção **??**, condicionada estritamente a interrupções de curta duração. A imputação de valores ocorreu somente quando satisfeitas, simultaneamente, as seguintes premissas de

adjacência temporal para cada registro ausente:

- Existência de uma medição válida imediatamente **anterior** (antecessor) com defasagem máxima de 15 minutos;
- Existência de uma medição válida imediatamente **posterior** (sucessor) com defasagem máxima de 15 minutos.

Tal restrição foi adotada para assegurar que a estimativa preservasse a fidelidade do comportamento real das variáveis, evitando a suavização artificial em períodos de indisponibilidade prolongada.

Dada uma variável y, considerando amostras de tempo $t_1 < t < t_2$ com valores conhecidos $y_1 = y(t_1)$ e $y_2 = y(t_2)$, o valor ausente em t foi estimado por:

$$y(t) = y_1 + \left(\frac{t - t_1}{t_2 - t_1} \right) (y_2 - y_1)$$

Ao restringir a interpolação a lacunas curtas e apoiar-se apenas em dados válidos circundantes, o método assegura a continuidade da série temporal sem introduzir padrões artificiais.

Pseudocódigo do Procedimento de Mitigação de Valores Ausentes

O procedimento de interpolação linear na Listagem 3.4 foi empregado para completar lacunas curtas nas séries temporais e opera sobre os registros previamente regularizados pelo processo descrito na Seção 3.5.1. A interpolação é aplicada exclusivamente quando há valores válidos imediatamente anteriores e posteriores ao ponto ausente, respeitando um limite máximo de quinze minutos entre essas medições. Esse critério impede o preenchimento de lacunas extensas e reduz o risco de indução de comportamentos artificiais na série, ao mesmo tempo em que preserva a coerência temporal dos dados tratados. Durante o processamento, cada registro pode receber um campo adicional denominado `interpolated_keys`, que funciona como variável de controle indicando explicitamente quais variáveis foram preenchidas por interpolação na respectiva amostra de tempo. Esse mecanismo reforça a rastreabilidade do processo e permite que análises posteriores diferenciem valores medidos de valores estimados.

Listagem 3.4: Pseudocódigo do Procedimento de Mitigação de Valores Ausentes

```
function ParseDatetime(texto):
   retornar ConverterTextoParaDatetime(texto)

function DiferencaMinutos(t1, t2):
   retornar |t2 - t1| em minutos

function InterpolarLinearmente(v1, v2, posicao):
   retornar v1 + posicao * (v2 - v1)

function InterpolarDadosInversor(registros):
   ordenar registros por datetime
   resultado = []
   para i de 0 ate tamanho(registros)-1:
      atual = registros[i]
      se i == 0 ou i == tamanho(registros)-1:
         adicionar atual em resultado
         continuar
      anterior = registros[i-1]
      proximo = registros[i+1]
      tA = ParseDatetime(anterior.datetime)
      tC = ParseDatetime(atual.datetime)
      tP = ParseDatetime(proximo.datetime)
      dA = DiferencaMinutos(tA, tC)
      dP = DiferencaMinutos(tC, tP)
      se dA <= 15 e dP <= 15:
         para cada chave em KEYS_TO_INTERPOLATE:
            se atual[chave] nulo e anterior[chave] nao nulo e proximo[chave
               ] nao nulo:
               pos = (tC - tA) / (tP - tA)
               atual[chave] = InterpolarLinearmente(anterior[chave],
                  proximo[chave], pos)
               marcar chave em atual.interpolated_keys
      adicionar atual em resultado
   retornar resultado
```

```
function ProcessarArquivo(caminho):
   dados = LerJSON(caminho)
   grupos = AgruparPor(dados, ["ps_id", "inverter_id"])
   interpolados = []
   para cada grupo em grupos:
      parcial = InterpolarDadosInversor(grupo)
      mesclar parcial em interpolados
   ordenar interpolados por datetime
   SalvarJSON(caminho, interpolados)

function EncontrarArquivosJSON(diretorio):
   retornar lista de arquivos .json em diretorio

function Principal():
   arquivos = EncontrarArquivosJSON(INPUT_DIR)
   para cada arquivo em arquivos:
      ProcessarArquivo(arquivo)
```

No caso da entidade inversor, o agrupamento simultâneo por `ps_id` e `inverter_id` é necessário para garantir que o preenchimento das lacunas seja realizado separadamente para cada equipamento, mantendo a integridade das séries temporais individuais. Para a entidade de estação solarimétrica, esse agrupamento adicional não é necessário e pode ser realizado apenas pelo `ps_id`.

As variáveis submetidas ao procedimento de interpolação para os inversores foram: potência ativa total (W), potência reativa total (var), potência DC total (W) e temperatura interna do inversor (°C).

Para a estação solarimétrica, o mesmo método foi aplicado às seguintes variáveis: irradiância global no plano inclinado (W/m²), irradiância no plano do arranjo (W/m²), irradiância global horizontal (W/m²), temperatura ambiente (°C), temperatura do painel (°C), velocidade do vento, direção do vento, índice de albedo do tracker, albedo e tensão da bateria (V).

3.6 Anonimização

Visando assegurar a conformidade com a LGPD e preservar o sigilo estratégico das informações, dada a natureza sensível e proprietária dos dados originais, foi desenvolvido um procedimento automatizado para a ofuscação dos identificadores únicos dos inversores e das usinas.

O processo consiste na substituição irreversível das chaves primárias originais por novas geradas aleatoriamente. Essa desvinculação garante que o conjunto de dados resultante deste trabalho não possa ser cruzado com o sistema de origem, impossibilitando a reidentificação das plantas ou a exposição de informações comerciais sensíveis.

4 Resultados

Este capítulo apresenta o BR-PVGen, um conjunto de dados resultante da aplicação da metodologia descrita no Capítulo 3. A exposição dos resultados está organizada de modo a detalhar primeiramente a abrangência física e temporal do estudo, seguida pela definição estrutural das entidades de dados e, por fim, a análise quantitativa da volumetria e da integridade das informações processadas.

Visando fomentar a pesquisa colaborativa e a transparência científica, o BR-PVGen encontra-se publicamente disponível no repositório Kaggle[10]. Para garantir a interoperabilidade com diferentes ferramentas de análise, os arquivos foram disponibilizados tanto em formato estruturado JSON quanto em formato tabular CSV.

4.1 Abrangência Espaço-Temporal e Técnica

A representatividade de um conjunto de dados ambientais depende intrinsecamente de sua distribuição geográfica e temporal. Neste estudo, a cobertura espacial concentra-se nas regiões Sudeste e Centro-Oeste do Brasil, conforme ilustrado na Figura 4.1. Esta dispersão por diferentes Unidades Federativas incorpora à base de dados variados microclimas e perfis de irradiação, enriquecendo o potencial analítico para modelos de desempenho fotovoltaico.

No domínio temporal, o BR-PVGen reflete a natureza dinâmica do sistema de monitoramento. Diferentemente de bases estáticas, houve uma incorporação progressiva de ativos ao longo do período estudado. A Tabela 4.1 detalha o horizonte de dados disponível para cada usina (identificada por `PS_ID`), evidenciando a entrada escalonada das plantas no sistema de coleta.

Para complementar a caracterização das plantas, a Tabela 4.2 apresenta as especificações técnicas essenciais, listando a potência nominal instalada e o tipo de estrutura de fixação (fixa ou rastreador) de cada ativo monitorado.

[10]Disponível em: ⟨https://www.kaggle.com/datasets/tecsci/brazilian-pv-dataset⟩

Tabela 4.1: Período de abrangência dos dados por usina (PS_ID).

Usina	**Início**	**Fim**	**Usina**	**Início**	**Fim**
PS_001	2024-03-26	2025-06-09	PS_027	2024-09-30	2025-06-09
PS_002	2024-03-26	2025-06-09	PS_028	2024-10-01	2025-06-09
PS_003	2024-03-26	2025-06-09	PS_029	2024-09-30	2025-06-09
PS_004	2024-03-26	2025-06-09	PS_030	2024-10-02	2025-06-09
PS_005	2024-04-11	2025-06-09	PS_031	2024-10-04	2025-03-22
PS_006	2024-05-09	2025-06-09	PS_032	2024-10-04	2025-06-09
PS_007	2024-05-09	2025-06-09	PS_033	2024-10-08	2025-06-09
PS_008	2024-05-10	2025-06-09	PS_034	2024-10-10	2025-06-09
PS_009	2024-06-10	2025-06-09	PS_035	2024-11-25	2025-06-09
PS_010	2024-06-12	2025-06-09	PS_036	2024-11-26	2025-06-09
PS_011	2024-06-20	2025-06-09	PS_037	2024-12-02	2025-06-09
PS_012	2024-06-21	2025-06-09	PS_038	2024-12-02	2025-06-09
PS_013	2024-07-03	2025-05-31	PS_039	2024-12-06	2025-06-09
PS_014	2024-08-20	2024-12-17	PS_040	2024-12-19	2025-06-09
PS_015	2024-08-21	2025-06-09	PS_041	2024-12-17	2025-06-09
PS_016	2024-09-02	2025-06-09	PS_042	2025-02-28	2025-06-09
PS_017	2024-09-20	2025-06-09	PS_043	2025-02-28	2025-06-09
PS_018	2024-09-20	2025-04-10	PS_044	2025-02-28	2025-06-09
PS_019	2024-09-23	2025-06-09	PS_045	2025-02-28	2025-06-09
PS_020	2024-09-23	2025-06-09	PS_046	2025-02-28	2025-06-09
PS_021	2024-09-23	2025-06-09	PS_047	2025-02-28	2025-06-09
PS_022	2024-09-23	2025-05-29	PS_048	2025-02-28	2025-06-09
PS_023	2024-09-23	2025-05-30	PS_049	2025-02-28	2025-06-09
PS_024	2024-10-09	2025-06-09	PS_050	2025-02-28	2025-06-09
PS_025	2024-10-02	2025-06-09	PS_051	2025-03-10	2025-06-09
PS_026	2024-09-27	2025-06-09		—	

Tabela 4.2: Características técnicas das usinas: Potência Nominal e Estrutura.

Usina	**Pot. (MW)**	**Estrutura**	**Usina**	**Pot. (MW)**	**Estrutura**
PS_001	5.0	TRACKER	PS_027	3.0	FIXED
PS_002	1.0	FIXED	PS_028	1.0	FIXED
PS_003	1.0	FIXED	PS_029	3.0	FIXED
PS_004	2.0	FIXED	PS_030	2.0	FIXED
PS_005	2.0	TRACKER	PS_031	3.0	FIXED
PS_006	4.0	TRACKER	PS_032	3.0	FIXED
PS_007	1.0	TRACKER	PS_033	3.0	TRACKER
PS_008	4.0	TRACKER	PS_034	4.0	TRACKER
PS_009	1.0	TRACKER	PS_035	4.0	TRACKER
PS_010	3.0	TRACKER	PS_036	1.9	TRACKER
PS_011	3.0	TRACKER	PS_037	2.0	TRACKER
PS_012	5.0	TRACKER	PS_038	2.0	TRACKER
PS_013	1.0	TRACKER	PS_039	4.0	TRACKER
PS_014	2.5	TRACKER	PS_040	1.0	TRACKER
PS_015	2.55	FIXED	PS_041	1.7	TRACKER
PS_016	2.55	FIXED	PS_042	2.5	TRACKER
PS_017	2.0	FIXED	PS_043	2.5	TRACKER
PS_018	3.0	FIXED	PS_044	2.5	TRACKER
PS_019	3.0	FIXED	PS_045	2.5	TRACKER
PS_020	2.0	FIXED	PS_046	2.5	TRACKER
PS_021	3.0	FIXED	PS_047	2.5	TRACKER
PS_022	2.0	FIXED	PS_048	2.5	TRACKER
PS_023	2.0	FIXED	PS_049	2.5	TRACKER
PS_024	2.0	FIXED	PS_050	2.5	TRACKER
PS_025	3.0	FIXED	PS_051	2.5	TRACKER
PS_026	2.0	FIXED		—	

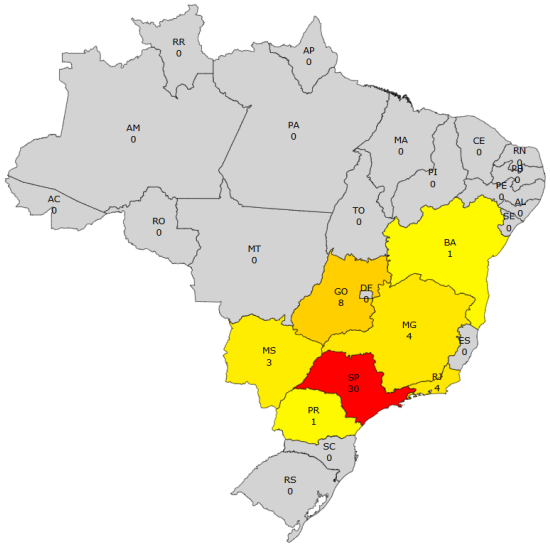

Figura 4.1: Distribuição geográfica das 51 usinas incluídas no BR-PVGen.

4.2 Dicionário de Dados

A estruturação lógica do BR-PVGen foi consolidada em três entidades fundamentais: "Metadados", "Inversor" e "Estação Solarimétrica". A Tabela 4.3 apresenta o dicionário de dados completo, descrevendo tecnicamente cada atributo e seu respectivo tipo de dado.

Visando facilitar a interoperabilidade e o uso pela comunidade científica internacional, a nomenclatura das variáveis foi padronizada em inglês utilizando o formato *snake case*[11], incluindo explicitamente a unidade de medida como sufixo do identificador.

Tabela 4.3: Atributos das Entidades

Atributo	**Tipo**	**Descrição**
	Entidade: Metadados	
`id`	Inteiro	Identificador único da usina.
`nominal_power_mw`	Float	Capacidade de potência nominal instalada (MW).
`is_panel_bifacial`	Bool	Indica se o painel é bifacial (`true`) ou monofacial.

Continua na próxima página...

[11] Estilo de escrita que substitui espaços por sublinhados (_) e utiliza apenas letras minúsculas (ex: `nominal_power`).

Continuação da Tabela 4.3

Atributo	**Tipo**	**Descrição**
`panel_temperature_coefficient`	Float	Perda percentual de potência por aumento de °C.
`panel_bifaciality_coefficient`	Float	Razão de eficiência traseira/frontal (0 a 1).
`panel_area_mm2`	Float	Área de um único módulo (mm^2).
`panel_efficiency_percentage`	Float	Eficiência de conversão do painel (%).
`number_of_panels`	Inteiro	Número total de módulos instalados.
`brazil_federative_unit`	String	Estado de localização (ex: "SP").
`structure_type`	String	Tipo: `TRACKER` ou `FIXED`.
Entidade: Inversor		
`datetime`	String	Data/hora ISO 8601 (`AAAA-MM-DDThh:mm:ssZ`).
`total_reactive_power_var`	Float	Potência reativa total (var).
`total_active_power_w`	Float	Potência ativa total (W).
`total_dc_power_w`	Float	Potência de entrada DC total (W).
`internal_temperature_celsius`	Float	Temperatura interna do inversor (°C).
`document_count`	Object<String, Int>	Contagem de amostras utilizadas no cálculo da média móvel.

Continua na próxima página...

Continuação da Tabela 4.3

Atributo	**Tipo**	**Descrição**
interpolated_keys	Object<String, Bool>	*Flag* indicativa de interpolação.
inverter_id	Inteiro	ID do inversor.
ps_id	String	ID da usina.
Entidade: Estação Solarimétrica		
datetime	String	Data/hora ISO 8601 (AAAA-MM-DDThh:mm:ssZ).
poa_irradiance_wm2	Float	Irradiância no Plano do Arranjo (POA) (W/m²).
battery_voltage	Float	Tensão da bateria da estação (V).
wind_speed_ms	Float	Velocidade do vento (m/s).
gri_irradiance_wm2	Float	Irradiância Refletida do Solo (GRI) (W/m²).
panel_temperature_celsius	Float	Temperatura do módulo (°C).
tracker_albedo_index	Float	Índice de albedo do solo.
ghi_irradiance_wm2	Float	Irradiância Horizontal Global (GHI) (W/m²).
wind_direction_degrees	Float	Direção do vento (graus).
ambient_temperature_celsius	Float	Temperatura ambiente (°C).
precipitation_accumulated_mm	Float	Precipitação acumulada (mm).
document_count	Object<String, Int>	Contagem de amostras utilizadas no cálculo da média móvel.

Continua na próxima página...

Continuação da Tabela 4.3

Atributo	**Tipo**	**Descrição**
`interpolated_keys`	Object<String, Bool>	*Flag* indicativa de interpolação.
`ps_id`	String	ID da usina.

4.3 Volumetria

A consolidação final do conjunto de dados resultou em um volume expressivo de informações, totalizando 14.400.480 registros válidos referentes aos inversores e 1.151.232 registros das estações solarimétricas, distribuídos entre as 51 plantas monitoradas.

A Figura 4.2 demonstra a evolução temporal deste volume. Observa-se um crescimento consistente na quantidade de registros mensais, correlacionado diretamente com a entrada de novas usinas no sistema (detalhada anteriormente na Tabela 4.1). Vale ressaltar a redução observada em junho de 2025, decorrente do encerramento da janela de coleta no nono dia do mês, caracterizando uma amostragem parcial para este período específico.

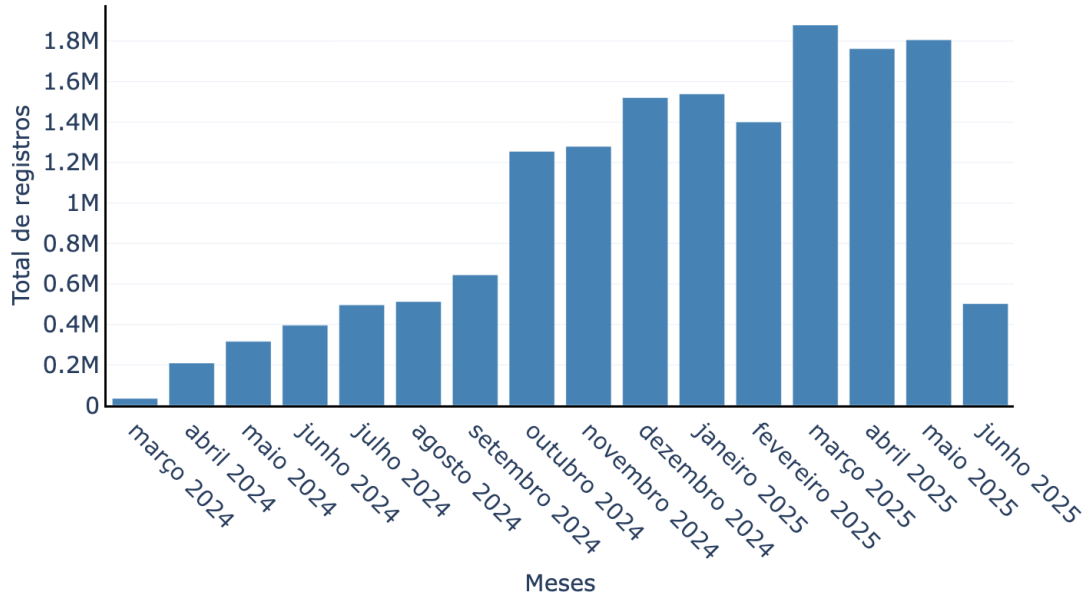

Figura 4.2: Total de registros mensais combinados (Inversores e Estações Solarimétricas).

4.4 Integridade dos Dados

A completude do conjunto de dados foi alcançada mediante a aplicação das técnicas de imputação descritas na Seção 3.5.2. A Tabela 4.4 apresenta o quantitativo de registros recuperados por cada metodologia.

Tabela 4.4: Total de registros preenchidos por metodologia e entidade.

Entidade	**Metodologia**	**Total Preenchido**
Inversor	Média Móvel	10.436.168
	Interpolação Linear	277.716
Estação Solar	Média Móvel	1.813.077
	Interpolação Linear	18.929

Para avaliar a distribuição dessas intervenções, a Figura 4.3 apresenta um mapa de calor que correlaciona as variáveis monitoradas com as usinas. Esta visualização permite identificar padrões de falhas de coleta, distinguindo comportamentos sistêmicos de ocorrências pontuais em ativos específicos.

Nota-se, por exemplo, que a usina `PS_016` (destacada em vermelho intenso) apresenta falhas sistêmicas em quase todas as variáveis, sugerindo problemas de conectividade do registrador local, enquanto usinas como a `PS_042` demonstram alta integridade (tons claros).

4.5 Análise Comparativa e Relevância

Para situar a contribuição deste trabalho frente ao estado da arte, realizou-se uma análise comparativa entre o BR-PVGen e outras bases de dados de energia solar amplamente citadas na literatura, como o DKASC (DKASC, 2024), o FAIR PV (NIHAR et al., 2021), o PVDAQ (DELINE et al., 2021) e o Pecan Street Dataport (PARSON et al., 2015). A Tabela 4.5 detalha as características técnicas e as variáveis disponíveis em cada uma destas bases em contraste com o conjunto de dados desenvolvido neste estudo.

A análise evidencia que o BR-PVGen preenche lacunas importantes, especialmente no contexto de geração distribuída em regiões tropicais. Enquanto o DKASC oferece uma longa série temporal (9 anos), ele restringe-se a zonas de clima desértico e

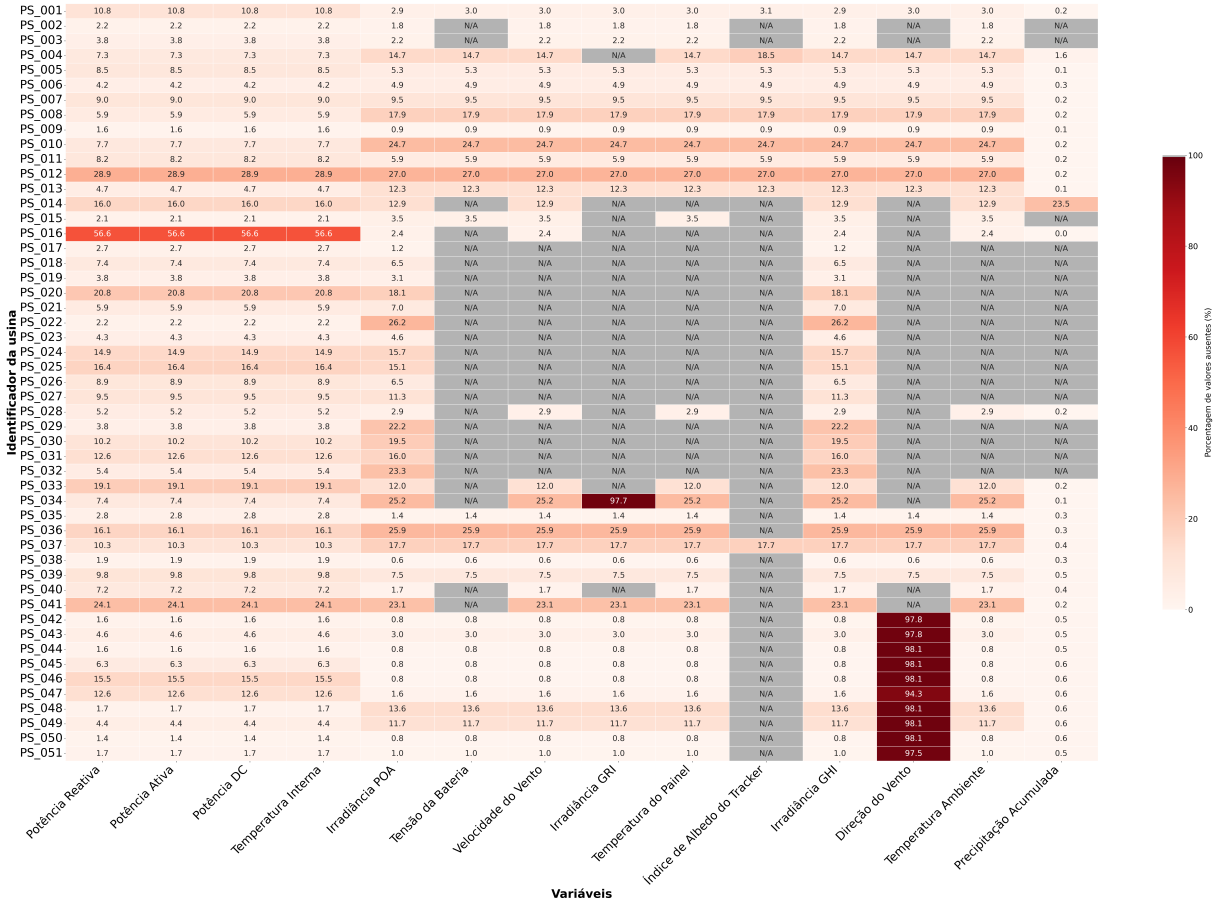

Figura 4.3: Percentual de dados ausentes por variável e usina durante o horário de geração efetiva (06:00–18:00, BRT).

a sistemas de menor porte (média de 416 kWp). Já o FAIR PV, embora abrangente em número de usinas (316), depende de dados meteorológicos de terceiros (APIs) e possui restrições de acesso e privacidade.

Em contrapartida, o BR-PVGen destaca-se pelos seguintes diferenciais:

- **Representatividade de Escala:** Composto por usinas de geração distribuída de grande porte (média de ∼2,5 MWp, variando de 1,7 a 5 MWp), o BR-PVGen reflete a realidade operacional de ativos comerciais robustos, diferindo das instalações menores ou residenciais comuns em outros *benchmarks*.

- **Aquisição de Dados *On-Site*:** Diferentemente de bases que estimam dados ambientais via satélite ou APIs, este estudo utiliza estações solarimétricas instaladas localmente em cada usina. Isso garante medições precisas de irradiância no plano do arranjo, irradiância global inclinada e temperatura dos módulos, fundamentais

para análises de desempenho de alta fidelidade.

- **Riqueza de Variáveis:** O BR-PVGen disponibiliza métricas raramente encontradas em bases públicas, como potência reativa, potência em corrente contínua, temperatura interna dos inversores e tensão das baterias das estações.
- **Granularidade e Sincronização:** A organização modular (Metadados, Inversor, Estação) aliada à sincronização temporal das medições elétricas e meteorológicas facilita a aplicação direta em modelos de aprendizado de máquina para detecção de falhas e previsão de geração.

A Tabela 4.5 sumariza as diferenças estruturais e o conteúdo informacional entre as bases analisadas, consolidando a relevância do BR-PVGen como uma ferramenta robusta para a pesquisa em sistemas fotovoltaicos.

Tabela 4.5: Comparativo entre o BR-PVGen e bases de dados consolidadas na literatura (DKASC, FAIR PV, PVDAQ e Pecan Street).

Propriedade / Característica	**BR-PVGen**	**DKASC**	**FAIR PV**	**PVDAQ**	**Pecan Street**
Medições com carimbo de tempo	✓	✓	✓	✓	✓
Potência Ativa (AC)	✓	✓	✓	✓	✓
Potência Reativa	✓	–	–	–	–
Potência DC	✓	–	✓	–	–
Temperatura Interna do Inversor	✓	–	–	–	–
Irradiância no Plano	✓	–	–	✓	–
Irradiância Global Inclinada	✓	–	–	✓	–
Irradiância Global Horizontal	✓	✓	✓	✓	✓
Temperatura do Painel	✓	✓	✓	–	–
Temperatura Ambiente	✓	✓	✓	–	✓
Velocidade do Vento	✓	✓	✓	–	✓
Direção do Vento	✓	✓	✓	–	✓
Precipitação	✓	✓	✓	–	✓
Tensão da Bateria (Estação)	✓	–	–	–	–
Índice de Albedo	✓	–	–	–	–
Indicador de Painel Bifacial	✓	–	–	–	–
Coeficiente de Temperatura	✓	–	–	–	–
Área do Painel	✓	–	–	✓	–
Tipo de Estrutura (Tracker/Fixa)	✓	–	–	✓	–
Granularidade por Inversor	✓	–	✓	✓	✓
Separação de Arquivos (Meta/Inv/Met)	✓	–	✓	✓	✓
Cobertura Temporal	1 ano	9 anos	2 anos	N.I.	3 anos
Resolução de Amostragem	15 min	5 min	5/15 min	15 min	15 min
Número de Usinas	51	9	316	N.I.	60
Capacidade Média	∼2.5 MWp	∼416 kWp	N.I.	N.I.	N.I.
Capacidade Mínima	1.7 MWp	22 kWp	N.I.	N.I.	N.I.
Capacidade Máxima	5 MWp	1.8 MWp	N.I.	N.I.	N.I.

5 Proposta de Aplicação e Potencial de Pesquisa

O desenvolvimento do conjunto de dados apresentado neste trabalho não se encerra na sua disponibilização; seu valor científico reside nas possibilidades de análise que ele habilita. Dada a alta granularidade temporal (15 minutos), a diversidade geográfica (51 usinas em diferentes estados) e a riqueza de variáveis (incluindo dados elétricos detalhados e ambientais *on-site*), este conjunto de dados oferece um terreno fértil para diversas linhas de pesquisa.

Este capítulo discute casos de uso práticos e teóricos para os quais o BR-PVGen é especialmente adequado, demonstrando como ele pode impulsionar o estado da arte em monitoramento e inteligência artificial aplicada a sistemas fotovoltaicos.

5.1 Previsão de Geração Fotovoltaica (Forecasting)

O BR-PVGen permite o desenvolvimento e o treinamento de modelos de aprendizado de máquina para previsão de sériess temporais. Diferentemente de bases que utilizam apenas dados de satélite, a presença de dados meteorológicos locais sincronizados com a produção elétrica permite a criação de modelos com maior acurácia (CARVALHO; FANTINI; SIQUEIRA, 2022).

Pesquisadores podem utilizar estes dados para comparar o desempenho de algoritmos de *Deep Learning* versus métodos estatísticos clássicos.

5.2 Detecção e Diagnóstico de Falhas

A operação e manutenção eficiente depende da capacidade de identificar anomalias antes que elas causem paradas críticas. A disponibilização de dados tanto do lado da corrente contínua (Potência DC) quanto da corrente alternada (Potência Ativa e Reativa), juntamente com a temperatura interna dos inversores, oferece uma visão da saúde dos

equipamentos (GHERARDI et al., 2018).

Aplicações potenciais incluem:

- **Análise de Curva de Eficiência:** O cálculo da eficiência de conversão (DC/AC) permite detectar degradação prematura de inversores ou perdas por *clipping* (ceifamento de potência).
- **Detecção de Sujidade (*Soiling*):** Correlacionando a precipitação acumulada com a recuperação da eficiência dos painéis, é possível estimar o nível de sujidade.

5.3 Estudos de Qualidade de Energia e Suporte à Rede

Um diferencial significativo do BR-PVGen em relação a outros conjunto de dados (como o DKASC) é a inclusão da variável de **Potência Reativa**. Com a evolução das normativas de geração distribuída (como a revisão da Lei 14.300 no Brasil), os inversores inteligentes estão sendo cada vez mais requisitados para fornecer suporte de tensão à rede.

Este conjutno de dados viabiliza estudos sobre:

- O impacto da injeção de potência ativa nos perfis de tensão da rede de distribuição.
- A capacidade dos inversores fotovoltaicos de atuar como compensadores estáticos de reativos durante períodos de baixa e alta irradiância.
- Análise do fator de potência real em operação comercial e sua conformidade com os requisitos de rede.

6 Conclusão

Este trabalho abordou a carência de dados públicos e padronizados sobre a geração distribuída fotovoltaica no Brasil, apresentando o desenvolvimento do BR-PVGen, um conjunto de dados sanitizado e documentado, proveniente de 51 usinas em operação comercial. A iniciativa alinha-se à tendência global de ciência aberta, fornecendo insumos essenciais para o avanço da pesquisa em energias renováveis em climas tropicais.

O objetivo principal de prover um recurso à comunidade científica materializou-se na estruturação de um acervo que, após o processamento, reúne mais de 15 milhões de registros. Beneficiando-se de uma aquisição de dados estável provida por um sistema SCADA em nuvem, o trabalho concentrou-se na metodologia de tratamento dessas informações, integrando medições de inversores e estações solarimétricas e mitigando ruídos através da Média Móvel Ponderada e interpolação linear.

É importante ressaltar que o trabalho apresenta limitações inerentes ao escopo definido. A cobertura temporal restringe-se a aproximadamente um ano e três meses (março de 2024 a junho de 2025), período suficiente para análises de sazonalidade anual, porém inferior às séries mais longas ideais para estudos de degradação de longo prazo. Adicionalmente, observa-se uma concentração geográfica nas regiões Sudeste e Centro-Oeste, resultando em uma menor representatividade das regiões Norte e Nordeste.

Apesar dessas restrições, as contribuições deste estudo são significativas, destacando-se pela análise comparativa frente a bases internacionais consolidadas (como DKASC e FAIR PV) e pelo preenchimento de lacunas específicas na literatura. As principais contribuições podem ser sintetizadas como:

- **Disponibilização de Variáveis Inéditas:** Inclusão de dados de potência reativa, temperatura interna de inversores e outras variáveis frequentemente ausentes em repositórios similares;
- **Medições meteorológicas *On-site*:** Dados meteorológicos medidos localmente permitindo correlações mais precisas do que dados de satélite;

Como desdobramentos desta pesquisa, para trabalhos futuros, sugere-se a expansão da base através da continuidade da coleta para ampliar a série temporal, permitindo estudos sobre o envelhecimento de componentes. Recomenda-se, ainda, o desenvolvimento de *benchmarks* oficiais de previsão de geração para o Brasil utilizando este conjunto de dados, estabelecendo métricas padronizadas para a comunidade acadêmica nacional.

Assim sendo, conclui-se que o BR-PVGen não serve apenas como um registro histórico, mas constitui uma ferramenta habilitadora para inovações que visam tornar a matriz energética brasileira mais eficiente, inteligente e sustentável.

Bibliografia

Agência Nacional de Energia Elétrica (ANEEL). *Sistema de Informações de Geração Distribuída (SIG-D)*. 2024. Disponível em: ⟨https://www.aneel.gov.br⟩.

AHMAD, E. et al. Integrated tool for cleaning bulk solar power data. In: *2024 Second International Conference on Smart Technologies for Power and Renewable Energy (SPECon)*. [S.l.: s.n.], 2024. p. 1–5.

BloombergNEF. *Energy Transition Investment Trends 2024*. [S.l.], 2024. Global energy transition investment reached US$1.8trillion in 2023, a 17% increase over 2022. Disponível em: ⟨https://assets.bbhub.io/professional/sites/24/Energy-Transition-Investment-Trends-2024.pdf⟩.

CARVALHO, M. J. d.; FANTINI, D. G.; SIQUEIRA, M. B. B. d. Métodos de previsão solar intra-hora: uma revisão da literatura. *Anais do Congresso Brasileiro de Energia Solar - CBENS*, 2022. Disponível em: ⟨https://doi.org/10.59627/cbens.2022.1179⟩.

DELINE, C. et al. *Photovoltaic Data Acquisition (PVDAQ) Public Datasets*. 2021. Disponível em: ⟨https://www.osti.gov/biblio/1846021⟩.

DKASC. 2024. Https://dkasolarcentre.com.au/locations/alice-springs. Alice Springs, Australia. Desert Knowledge Australia Solar Centre.

Empresa de Pesquisa Energética (EPE). *Balanço Energético Nacional 2024: Relatório Síntese - Ano Base 2023*. Rio de Janeiro: Ministério de Minas e Energia, 2024. Disponível em: ⟨http://www.epe.gov.br⟩.

GHERARDI, R. et al. Anomaly detection and predictive maintenance for photovoltaic systems. *Neurocomputing*, Elsevier, v. 310, p. 59–68, 2018. Disponível em: ⟨https://doi.org/10.1016/j.neucom.2018.05.017⟩.

HADDAD, H.; JERBI, F.; SMAALI, I. Performance evaluation of ai-driven photovoltaic output forecasting. In: *2024 IEEE PES/IAS PowerAfrica*. [S.l.: s.n.], 2024. p. 1–5.

HASSANI, H.; KALANTARI, M.; GHODSI, Z. Evaluating the performance of multiple imputation methods for handling missing values in time series data: A study focused on East Africa, soil-carbonate-stable isotope data. *Stats*, v. 2, n. 4, p. 457–467, 2019. ISSN 2571-905X.

International Energy Agency. *Renewables 2023: Analysis and forecast to 2028*. 2023. Disponível em: ⟨https://www.iea.org/reports/renewables-2023⟩.

International Renewable Energy Agency. *World Energy Transitions Outlook 2023*. 2023. Disponível em: ⟨https://www.irena.org/Publications/2023/Jun/World-Energy-Transitions-Outlook-2023⟩.

LEPOT, M.; AUBIN, J.-B.; CLEMENS, F. H. Interpolation in time series: An introductive overview of existing methods, their performance criteria and uncertainty assessment. *Water*, v. 9, n. 10, 2017. ISSN 2073-4441.

NIHAR, A. et al. Toward findable, accessible, interoperable and reusable (fair) photovoltaic system time series data. In: *2021 IEEE 48th Photovoltaic Specialists Conference (PVSC)*. [S.l.: s.n.], 2021. p. 1701–1706.

PARSON, O. et al. Dataport and nilmtk: A building data set designed for non-intrusive load monitoring. In: IEEE. *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. [S.l.], 2015. p. 210–214.

SCHNIERER, B. *The pros and cons of 1-minute, 15-minute, and 60-minute solar data*. 2023. Https://solargis.com/resources/blog/best-practices/the-pros-and-cons-of-1-minute-15-minute-and-60-minute-solar-data. Solargis blog.