INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENERGIAS SUSTENTÁVEIS




VALIDAÇÃO DE UMA METODOLOGIA DE CÁLCULOS DE ESTIMATIVAS DE PRODUÇÃO DE ENERGIA EM PARQUES FOTOVOLTAICOS EM OPERAÇÃO

JESSICA TATIANA GARCIA FERREIRA
outubro de 2025

POLITÉCNICO
DO PORTO

ISEP INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO

VALIDAÇÃO DE UMA METODOLOGIA DE CÁLCULOS DE ESTIMATIVAS DE PRODUÇÃO DE ENERGIA EM PARQUES FOTOVOLTAICOS EM OPERAÇÃO

Jessica Tatiana Garcia Ferreira

Relatório de Estágio para obtenção do Grau de Mestre em Energias Sustentáveis

Orientador no ISEP: Professor Doutor Jorge Manuel Pires Mendonça

Coorientadora no ISEP: Professora Doutora Rosa Maria Barbosa Rodrigues Pilão

Supervisor na Megajoule: Engenheiro Mestre Agostinho Paulo Ferreira Pinto

MEGAJOULE
Rethinking Energy Worldwide

Júri:

Presidente:

Olga dos Remédios Sobral Castro, Professor Adjunto, Instituto Superior de Engenharia do Politécnico do Porto

Vogais:

Jorge Manuel Pires Mendonça, Professor Adjunto, Instituto Superior de Engenharia do Politécnico do Porto

Adélio Cavadas, Professor Adjunto, Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Viana do Castelo

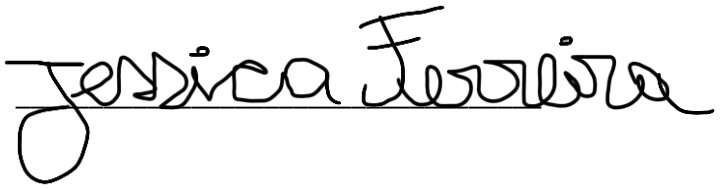
Porto, 17 de setembro de 2025

Declaração de Integridade

Declaro ter conduzido este trabalho académico com integridade. Não plagiei ou apliquei qualquer forma de uso indevido de informações ou falsificação de resultados ao longo do processo que levou à sua elaboração.

Declaro que o trabalho apresentado neste documento é original e de minha autoria, não tendo sido utilizado anteriormente para nenhum outro fim.

Declaro ainda que tenho pleno conhecimento do Código de Conduta Ética do P.PORTO.



ISEP, Porto, 17 de setembro de 2025

Dedicatória

À minha querida avó Francisca, que partiu antes de eu concluir esta etapa da minha vida. Foi e continuará a ser uma das pessoas mais influentes na minha vida. Para si, "mamãe": *eu sou porque tu foste*.

Resumo

A transição energética é um dos grandes desafios do nosso tempo, e a energia solar fotovoltaica tem-se afirmado como catalisador desta transformação. Torna-se, por isso, indispensável dispor de metodologias robustas para estimar a produção a longo prazo em parques já em operação. Ao contrário das previsões de pré-construção, as estimativas pós-operacionais permitem recalibrar as projeções futuras de produção anual de energia, incorporando dados operacionais reais. Esta abordagem possibilita a redução das incertezas associadas a estudos energéticos preliminares, decisões de investimento, *Power Purchase Agreements (*PPAs) e avaliações de risco.

Neste enquadramento, a presente investigação pretende validar a aplicabilidade da metodologia *Measure–Correlate–Predict* (MCP), originalmente desenvolvida no setor eólico, ao setor fotovoltaico, adaptando-a a dados mensais de produção real e a séries de radiação de longo prazo provenientes de bases reconhecidas. A abordagem assentou na regressão linear simples entre a produção medida e a série de radiação de referência, após uma análise comparativa de três bases (Solargis, PVGIS e ERA5) com o intuito de identificar a mais adequada para caracterizar cada parque. Esta etapa foi complementada por um processo de deteção de outliers, assegurando uma maior robustez das correlações. Em paralelo, foram explorados modelos alternativos como: regressão múltipla e algoritmos de *machine learning* (Random Forest e XGBoost) para comparar o respetivo desempenho preditivo com o MCP clássico.

Os resultados evidenciam que, quando suportado por seleção criteriosa da base de radiação e por tratamento sistemático de outliers, o MCP apresenta desempenho consistente e fiável em contexto pós-operacional. As abordagens de *machine learning* revelam ganhos pontuais em cenários de maior variabilidade, constituindo um complemento pertinente à clareza e parcimónia da regressão linear. Em síntese, o estudo reforça a confiança nas estimativas de longo prazo em parques fotovoltaicos em operação, contribui para a redução de incertezas e informa decisões de investimento.

Palavras-chave: Energia fotovoltaica; Measure–Correlate–Predict (MCP); Estimativa de produção de longo prazo; Dados pós-operacionais; Bases de radiação; *Machine learning* (ML).

Abstract

The energy transition is one of the great challenges of our time, and photovoltaic solar energy has established itself as a catalyst for this transformation. It is therefore essential to have robust methodologies for estimating long-term production in parks already in operation. Unlike pre-construction forecasts, post-operational estimates allow future annual energy production projections to be recalibrated, incorporating actual operational data. This approach reduces the uncertainties associated with preliminary energy studies, investment decisions, power purchase agreements (PPAs), and risk assessments.

In this context, the present research validates the applicability of the Measure–Correlate–Predict (MCP) methodology, originally developed in the wind sector, to the photovoltaic sector, adapting it to monthly actual production data and long-term radiation series from recognized databases. The approach was based on simple linear regression between measured production and the reference radiation series, after a comparative analysis of the three databases (Solargis, PVGIS, and ERA5) to identify the most appropriate one to characterize each farm. This step was complemented by a process of outlier detection and management, ensuring a greater robustness of the correlations.

In parallel, alternative models were explored, such as multiple regression and machine learning algorithms (Random Forest and XGBoost), to compare their predictive performance with that of the classic MCP.

The results show that, when supported by careful selection of the radiation base and systematic treatment of outliers, MCP performs consistently and reliably in a post-operational context. Machine learning approaches reveal occasional gains in scenarios of greater variability, constituting a relevant complement to the clarity and parsimony of linear regression. In summary, the study reinforces confidence in long-term estimates for photovoltaic parks in operation, contributes to reducing uncertainties, and informs investment decisions.

Keywords: Photovoltaic energy; Measure–Correlate–Predict (MCP); Long-term production estimation; post-operational data; Radiation bases; Machine learning (ML).

Agradecimentos

A jornada até à conclusão deste trabalho foi, sem dúvida, exigente e repleta de desafios. A sua realização não teria sido possível sem o apoio, a amizade, o incentivo e o carinho das pessoas que caminharam ao meu lado ajudando-me a enfrentar as dúvidas e a vencer os desafios.

Em primeiro lugar, agradeço a Deus pela vida e por me ter permitido trilhar este percurso, colocando as pessoas certas ao meu lado em cada etapa.

Expresso a minha gratidão ao Engenheiro Paulo Pinto, pela disponibilidade e pelas oportunidades de aprendizagem e acompanhamento, e à Doutora Rosa Pilão, pelas conversas essenciais, críticas construtivas e apoio constante ao longo de todo este processo.

Ao meu orientador, Doutor Jorge Mendonça, deixo um agradecimento muito especial. Obrigado pela paciência, pela dedicação e pela disponibilidade constante, mesmo em momentos menos oportunos. A sua orientação, críticas e ensinamentos foram determinantes para ultrapassar as dificuldades e para o enriquecimento deste trabalho.

Agradeço ainda à equipa da Megajoule, em particular à Mariana, pela colaboração e apoio prestados.

À minha família, o meu profundo reconhecimento: aos meus pais e à minha irmã, pelo amor incondicional e pela presença constante. À minha avó, que já não se encontra entre nós, mas que permanece como fonte de inspiração e de força em cada etapa da minha vida.

Às minhas melhores amigas Diana e Sofia, pela vossa amizade, por nunca me deixarem desistir, pelas palavras de encorajamento e até pelas lágrimas partilhadas, que se transformaram sempre em motivação renovada.

A todos os que, de forma direta ou indireta, contribuíram para a realização deste trabalho, deixo a minha mais sincera gratidão.

Sem vós, estas páginas estariam certamente em branco.

Índice Geral

Declaração de Integridade ..... *iii*

Dedicatória ..... *v*

Resumo ..... *vii*

Abstract ..... *ix*

Agradecimentos ..... *xi*

Índice Geral ..... *xiii*

Lista de Figuras ..... *xvii*

Lista de Tabelas ..... *xix*

Acrónimos e Símbolos ..... *xxi*

Lista de Acrónimos ..... **xxi**

Lista de Símbolos ..... **xxii**

1. Introdução *1*

1.1 Enquadramento **1**

1.2 Justificação e relevância do estudo **4**

1.3 Objetivos da investigação **5**

1.4 Apresentação da empresa **6**

1.5 Estrutura do relatório **6**

2. Revisão da literatura *7*

2.1 Recurso solar **7**

2.1.1 Variabilidade do recurso solar 9

2.1.2 Energia fotovoltaica 10

2.1.3 Bases de Dados de Radiação Solar 12

2.2 Metodologias Pós-Construção em Sistemas Fotovoltaicos **13**

2.3 Revisão Crítica do MCP **15**

2.3.1 Evolução histórica do MCP no setor solar 16

2.3.2 Abordagens GHI–GHI vs. Produção–GHI 17

2.3.3 Lacunas na literatura 18

3. Metodologia *21*

3.1 Caracterização dos Dados Utilizados 21

3.2 Adaptação da Metodologia MCP 23

3.3 Métricas Estatísticas de Avaliação de Desempenho 25

3.4 Análise de outliers 27

3.4.1 Validação estatística da regressão após remoção dos outliers 28

3.4.2 Agrupamento dos parques por sensibilidade à presença de outliers 28

3.4.3 Distribuição e sazonalidade dos outliers 29

3.4.4 Causas técnicas e operacionais associadas aos outliers 30

3.5 Escolha da Base de Dados de Radiação 33

3.6 Validação da Metodologia MCP Adaptada 34

3.7 Modelos Alternativos ao MCP 36

4. Resultados 39

4.1 Escolha da melhor base 39

4.1.1 Alternativas na Escolha da Base de Dados 41

4.2 Análise de outliers 43

4.2.1 Impacto estatístico dos outliers nos modelos MCP 43

4.2.2 Classificação dos parques por sensibilidade a outliers 45

4.2.3 Distribuição temporal e padrão sazonal dos outliers 46

4.2.4 Causas técnicas e operacionais associadas aos outliers 49

4.3 Validação da metodologia adotada 52

4.4 Modelos alternativos ao MCP 54

5. Conclusão 59

5.1 Recomendações e Trabalhos Futuros 61

Bibliografia 63

Apêndice A – Escolha da melhor base 69

Apêndice B – ANÁLISES DE OUTLIERS 71

Apêndice C – VALIDAÇÃO DA METODOLOGIA 77

Apêndice D – MODELOS ALTERNATIVOS AO MCP 96

Apêndice E – SCRIPT EM PHYTON PARA TREINO E VALIDAÇÃO DO MODELO XGBOOST EM SERIES MENSAIS 103

Lista de Figuras

Figura 1 - Evolução da contribuição das diferentes tecnologias renováveis na geração elétrica entre 2000 e 2030, evidenciando a rápida ascensão do fotovoltaico (IEA, 2024)..... 2
Figura 2 - Empregos globais nas energias renováveis por tecnologia (IRENA). 3
Figura 3 - Radiação solar incidente e seus componentes principais (Duffie & Beckman, 2013). 8
Figura 4 - Esquema simplificado do efeito fotovoltaico (Eletrónica PT, s.d.). 11
Figura 5 - Localização geográfica dos parques selecionados (rbiedermann, n.d.). 22
Figura 6 - Distribuição mensal de outliers..... 47
Figura 7 - Heatmap de outliers por parque e mês. 48
Figura 8 - Histogramas da disponibilidade e do PR nos meses classificados como outliers. 50
Figura 9 - Relação entre disponibilidade e PR para os outliers identificados. 51
Figura 10- Comparação produção real vs previsões (PV13). 56
Figura 11 - Comparação produção real vs previsões (PV17). 56
Figura 12 - Comparação produção real vs previsões (PV07). 57
Figura 13 - Comparação produção real vs previsões (PV08). 57

Lista de Tabelas

Tabela 1 - Síntese comparativa de bases de dados de radiação solar de longo prazo. 13
Tabela 2 - Informação genérica dos parques fotovoltaicos. 23
Tabela 3 - Valores de coeficiente de determinação (R^2) para todas as sete bases de dados de radiação. 40
Tabela 4 - Valores médios de R^2, MBE e RMSE por base de dados. 41
Tabela 5 - Indicadores estatísticos com e sem remoção de outliers. 44
Tabela 6 - Classificação de sensibilidade para cada parque. 45
Tabela 7 - Comparação das médias resumida dos modelos. 55

Acrónimos e Símbolos

Lista de Acrónimos

ANOVA	*Analysis of Variance*
CPV	*Concentrated Photovoltaic*
CSP	*Concentrated Solar Power*
DHI	*Diffuse Horizontal Irradiation*
DNI	*Direct Normal Irradiation*
ECDF	*Empirical Cumulative Distribution Function*
ECMWF	*European Centre for Medium-Range Weather Forecasts*
ERA5	*ECMWF Reanalysis 5th Generation*
GHI	*Global Horizontal Irradiation*
IEA	*International Energy Agency*
IEA-PVPS	*IEA Photovoltaic Power Systems Programme*
IEC	*International Electrotechnical Commission*
IRENA	*International Renewable Energy Agency*
LCOE	*Levelised Cost of Electricity*
MBE	*Mean Bias Error*
MCP	*Measure–Correlate–Predict*
ML	*Machine Learning*
NREL	*National Renewable Energy Laboratory*
O&M	Operação e Manutenção
OMIE	Operador do Mercado Ibérico de Energia
PCA	*Principal Component Analysis*

PPA	*Power Purchase Agreement*
PR	*Performance Ratio*
PVGIS	*Photovoltaic Geographical Information System*
PVsyst	*Photovoltaic System Software*
PVUSA	*Photovoltaics for Utility-Scale Applications*
RMSE	*Root Mean Square Error*
SAM	*System Advisor Model*
SCADA	*Supervisory Control And Data Acquisition*
STL	*Seasonal-Trend decomposition using Loess*

Lista de Símbolos

ΔR^2	Variação do coeficiente de determinação	
σ	Desvio padrão	
μ	Média estatística	
P	Produção mensal bruta corrigida por disponibilidade	[MWh]
A	Coeficiente angular da regressão linear	
B	Termo independente (interceção do modelo)	
Pi	Valor observado de produção no instante i	[MWh]
fi	Valor previsto pelo modelo no instante i	[MWh]
$\bar{P}$	Média dos valores observados de produção	[MWh]
N	Número total de observações	
Q1	Primeiro quartil	
Q3	Terceiro quartil	

Zi	*Z-score* da observação i
xi	Observação individual no cálculo de Z-score
$\bar{x}$	Média da amostra utilizada no cálculo de Z-score
F	Estatística do teste ANOVA
p	Valor-p (nível de significância)
DW	Estatística de Durbin–Watson
H	Estatística do teste Kruskal–Wallis
R^2	Coeficiente de determinação
ΔE	Distância entre grupos (métrica de clusterização)
$\|\|uA - uB\|\|^2$	Norma quadrática da diferença entre vetores médios A e B
uA	Vetor médio do grupo A
uB	Vetor médio do grupo B
nA	Número de elementos do grupo A
nB	Número de elementos do grupo B
i	Índice temporal / observação
j	Índice de grupo / classe

1. Introdução

1.1 Enquadramento

O sistema energético global está a atravessar uma transformação rápida e profunda, impulsionada pela necessidade urgente de combater as alterações climáticas e garantir um futuro energético mais seguro. No centro desta transição encontra-se a expansão de fontes de energia renovável, com a tecnologia fotovoltaica a emergir como um elemento-chave para os sistemas energéticos do futuro.

A energia solar fotovoltaica está projetada para se tornar a maior fonte de energia renovável até 2029, devido à drástica redução do custo das tecnologias fotovoltaicas e também pelo reforço das políticas públicas de apoio à sua implementação em todo o mundo. Entre 2018 e 2023, a capacidade global de energia fotovoltaica instalada triplicou, atingindo um valor recorde de 320 TWh.

A Figura 1 evidencia que a trajetória da capacidade solar acompanha de perto o cenário *Net Zero Emissions* até 2050 (IEA, 2024). A energia fotovoltaica tem vindo a afirmar-se como peça decisiva nos compromissos climáticos globais, sustentada tanto pela redução acentuada dos custos tecnológicos como pelo efeito cumulativo das políticas públicas de incentivo. Só na União Europeia, em 2023, foram adicionados cerca de 61 GW, um crescimento anual de 45% (IEA, 2024) que confirma o papel da região como um dos principais motores da expansão mundial da energia solar.

Nenhuma outra fonte de energia registou um crescimento tão rápido e consistente como a energia solar, que se tornou um pilar estratégico para enfrentar simultaneamente os desafios climáticos e de segurança de abastecimento.

Para além dos benefícios ambientais, o setor fotovoltaico tem demonstrado ser um condutor impressionante no que se refere ao avanço económico e à criação de novos empregos. A redução consistente do *Levelised Cost of Electricity* (LCOE) dos sistemas fotovoltaicos tem-se revelado um dos principais motores da sua adoção em larga escala. Este decréscimo de custos contribuiu para que a tecnologia solar se posicionasse como a alternativa energética de menor preço unitário e, simultaneamente, a de maior dinamismo no crescimento global (IEA, 2024).

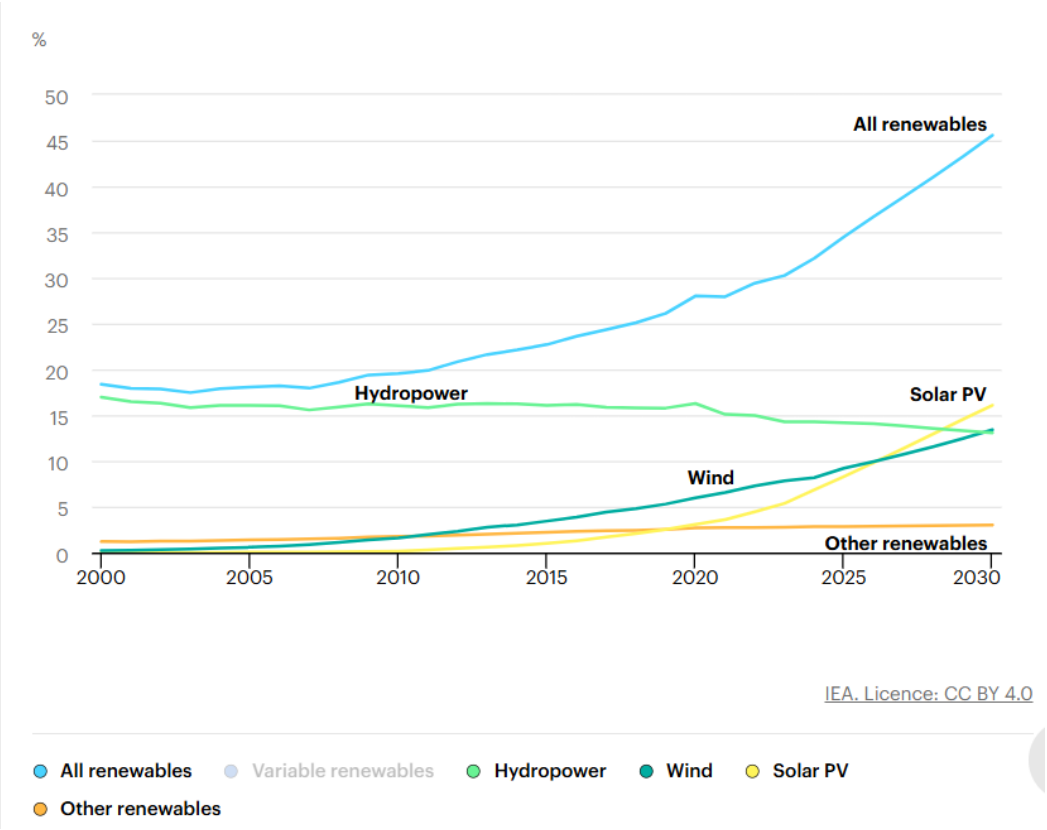


Figura 1 - Evolução da contribuição das diferentes tecnologias renováveis na geração elétrica entre 2000 e 2030, evidenciando a rápida ascensão do fotovoltaico (IEA, 2024).

Do ponto de vista socioeconómico, em 2023 as energias renováveis atingiram um volume recorde de cerca de 16,2 milhões de empregos, o que representa um crescimento de 18% face a 2022. Dentro deste universo, o setor fotovoltaico destacou-se como principal motor, concentrando aproximadamente 7,2 milhões de empregos (Figura 2), segundo dados da IRENA (2024). Estes números evidenciam não apenas a expansão tecnológica, mas também o impacto direto da transição energética na criação de novas oportunidades de trabalho a nível global. Este crescimento é particularmente notório na Europa meridional. Na Península Ibérica, em particular, o potencial das energias renováveis é elevado, com destaque para o setor fotovoltaico. Atualmente, a região ultrapassa a capacidade instalada de 50 GW, posicionando a energia solar fotovoltaica a tecnologia com o crescimento mais acelerado. As previsões apontam para que, até 2030, a capacidade solar instalada na Península Ibérica ultrapasse a da energia eólica, passando a representar mais de metade da produção renovável na região (Rystad Energy, 2023). Este cenário traduz uma inversão histórica na hierarquia tecnológica,

refletindo não só a contínua redução de custos do solar, mas também a sua maior adaptabilidade em termos de expansão territorial.

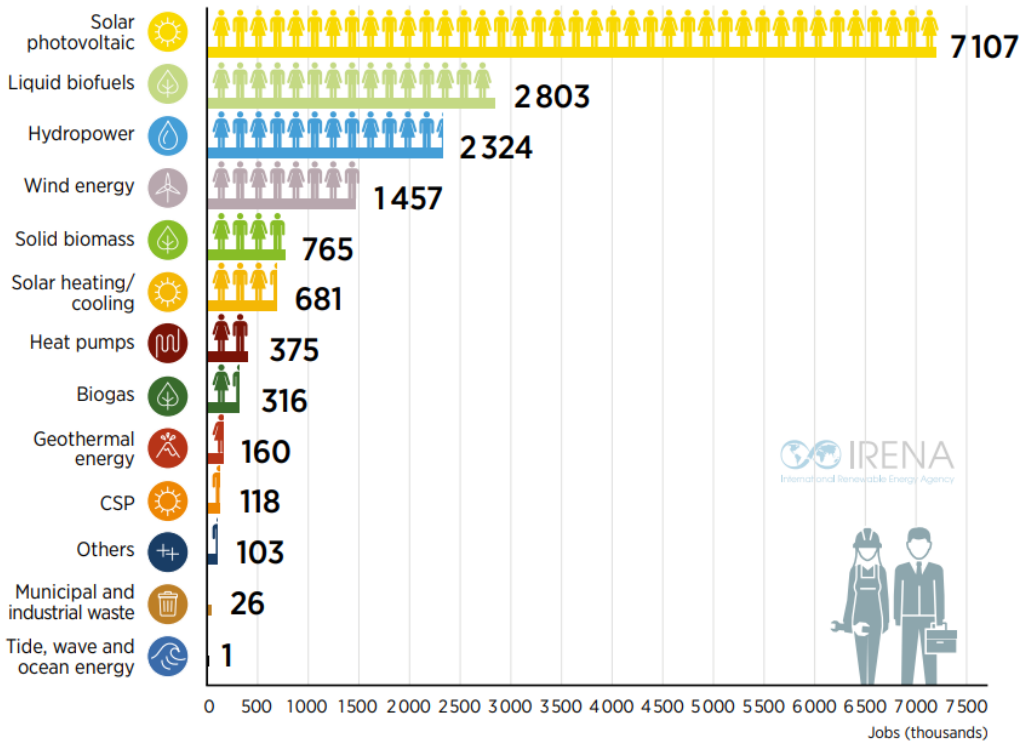


Figura 2 - Empregos globais nas energias renováveis por tecnologia (IRENA, 2024).

À medida que os parques solares se tornam mais maduros e a sua escala aumenta, torna-se cada vez mais crítico garantir previsões de produção fiáveis a longo prazo. Surge, então uma questão central: com que precisão se pode estimar a produção energética de um parque em operação ao longo de décadas de funcionamento? No setor eólico, esta problemática foi enfrentada através do desenvolvimento de metodologias para extrapolar medições de curto prazo com base em séries meteorológicas de longo termo. No setor fotovoltaico, observa-se uma adoção crescente de abordagens semelhantes.

Em particular, a metodologia *Measure–Correlate–Predict* (MCP), originalmente desenvolvida para estimar o recurso eólico em locais com medições limitadas, tem vindo a emergir como uma técnica promissora para a estimativa da produção em parques fotovoltaicos. A metodologia MCP constitui um enquadramento sistemático que integra séries locais de curta duração com registos históricos de radiação solar. Esta integração permite aumentar a

consistência estatística das estimativas e reduzir incertezas em análises de longo prazo, aspeto especialmente relevante em aplicações pós-construção (Carta et al., 2013).

1.2 Justificação e relevância do estudo

A rápida expansão da capacidade de energia solar fotovoltaica em todo o mundo tem reforçado a necessidade de estimativas de produção fotovoltaica de longo prazo em parques em operação mais fiáveis. Ao contrário das estimativas de pré-construção, as estimativas de pós-construção fornecem previsões mais ajustadas à realidade. As estimativas de produção de longo prazo assumem especial importância para as diferentes partes interessadas, desde entidades envolvidas em *Power Purchase Agreements* (PPAs) até investidores que avaliam riscos financeiros, ativos ou indicadores de desempenho operacional. No sul da Europa, a rápida expansão da capacidade instalada tem evidenciado a maturidade crescente do setor fotovoltaico, aumentando em paralelo a procura por metodologias de previsão mais fiáveis e fundamentadas em dados (Rystad Energy, 2023).

Neste contexto, a confiança nos indicadores probabilísticos P50 e P90[1] é determinante, uma vez que estes servem de referência em processos de financiamento e na avaliação da bancabilidade[2] dos projetos (IEA-PVPS, 2020). Contudo, as práticas atuais continuam a revelar fragilidades: a utilização de bases de radiação solar com viés, a escassa adaptação às condições locais e a ausência de tratamento adequado de eventos operacionais anómalos podem resultar em estimativas inflacionadas ou excessivamente conservadoras. As consequências são evidentes, traduzindo-se em distorções de modelos financeiros e numa erosão progressiva da confiança dos investidores.

Face a estes desafios, o setor tem procurado metodologias alternativas, muitas delas inspiradas em práticas já consolidadas na energia eólica. Entre elas destaca-se o método MCP, concebido originalmente para extrapolação de recursos eólicos e, mais recentemente, aplicado ao contexto solar. A sua lógica é simples mas eficaz: correlacionar dados de curto prazo medidos

---

[1] P50 e P90 são indicadores probabilísticos usados em previsões de energia. O P50 corresponde a um valor que tem 50% de probabilidade de ser ultrapassado, enquanto o P90 representa um valor mais conservador, com 90% de probabilidade de ser excedido. São usados com frequência em análises de risco e financiamento.

[2] Bancabilidade é o grau de confiança que investidores e entidades financeiras atribuem às previsões de produção de um projeto, determinando se podem servir de base fiável para financiamento.

localmente com séries de irradiância de longo prazo obtidas por satélite, o que abre caminho para previsões mais precisas e fundamentadas.

Ao validar o método MCP no contexto fotovoltaico, esta investigação visa reduzir a incerteza de rendimento e reforçar a base analítica para o planeamento a longo prazo, a mitigação de riscos e o investimento em projetos solares de grande dimensão.

1.3 Objetivos da investigação

Considerando o contexto apresentado anteriormente, o objetivo principal deste estudo é validar uma metodologia de cálculo para estimar a produção de longo prazo de parques fotovoltaicos em operação. A abordagem proposta baseia-se na correlação entre séries históricas de radiação solar e dados de produção mensal real medidos em 19 parques fotovoltaicos, recorrendo a técnicas adaptadas do modelo MCP.

Para alcançar esse objetivo principal, foram estabelecidos os seguintes objetivos específicos:

- Comparar três bases de dados de radiação solar de longo prazo: Solargis, PVGIS e ERA5, com o objetivo de identificar qual delas possui maior robustez estatística e representa de forma mais precisa a variabilidade observada nas produções reais, servindo como referência para a aplicação da metodologia.
- Identificar e analisar outliers nos dados operacionais de produção, utilizando critérios estatísticos e operacionais que permitam compreender o impacto dos desvios e garantir a qualidade e fiabilidade da modelação.
- Validar a aplicação do modelo de regressão linear (MCP) no contexto fotovoltaico, através da verificação dos pressupostos estatísticos fundamentais e da análise da qualidade do ajuste em diferentes parques.
- Explorar metodologias alternativas ao MCP tradicional, com especial foco na comparação do seu desempenho face a modelos baseados em machine learning , avaliando a sua robustez, capacidade preditiva e aplicabilidade prática.

1.4 Apresentação da empresa

Este trabalho foi desenvolvido em colaboração com a empresa Megajoule, uma consultora independente de engenharia sediada em Portugal, com vasta experiência na avaliação de recursos renováveis. Desde a sua fundação, tem prestado serviços técnicos a promotores, investidores e entidades financeiras, operando em mercados internacionais.

Embora historicamente associada ao setor eólico, a Megajoule tem expandido a sua atuação para a energia solar, realizando estudos técnicos, modelação de produção e auditorias em projetos fotovoltaicos. A sua experiência abrange desde campanhas de medição até à gestão e supervisão de projetos.

A parceria com a Megajoule foi fundamental para esta investigação, permitindo o acesso a dados reais de produção de parques solares, bem como ao conhecimento técnico aplicado, contribuindo para a robustez metodológica da análise desenvolvida.

1.5 Estrutura do relatório

Este trabalho encontra-se organizado em cinco capítulos. No primeiro capítulo é feito o enquadramento teórico e são apresentadas as motivações da investigação, com ênfase nos desafios da estimativa de produção de longo prazo em parques fotovoltaicos em operação. No capítulo seguinte desenvolvemos uma revisão crítica da literatura, abordamos o recurso solar, as bases de dados climáticas e a evolução da metodologia Measure–Correlate–Predict (MCP), destacando também abordagens alternativas, como modelos estatísticos e de machine learning. No terceiro capítulo, apresenta-se a metodologia aplicada, desde a caracterização das bases de dados à implementação dos modelos preditivos e respetiva validação, incluindo a análise de outliers. No quarto capítulo é feita uma exposição e discussão dos resultados obtidos para os vários objetivos definidos, com base em métricas quantitativas e representações gráficas. Por fim, no quinto capítulo sintetizamos as conclusões deste estudo e propomos direções futuras de investigação.

2. Revisão da literatura

Esta secção é dedicada à analise dos conceitos teóricos e científicos essenciais ao desenvolvimento deste trabalho. Serão explorados o recurso solar e a metodologia aplicada, com uma revisão crítica.

2.1 Recurso solar

A energia solar é uma das fontes de energia renovável mais abundantes e previsíveis à escala global, sendo muito relevante para a transição para sistemas energéticos mais sustentáveis. A disponibilidade de energia solar resulta da radiação eletromagnética emitida pelo Sol. Parte dessa radiação chega à superfície terrestre e pode ser convertida em eletricidade ou calor através de diferentes tecnologias, entre as quais se destacam os sistemas fotovoltaicos pela sua ampla difusão. Vale salientar, contudo, que o aproveitamento eficaz deste recurso exige mais do que a simples captação, mas também requer a compreensão dos princípios físicos que o regem e da variabilidade que condiciona a sua utilização prática.

O termo recurso solar, de uma forma simples, refere-se, à quantidade de energia emitida pelo Sol que chega à superfície terrestre e que pode ser convertida em formas úteis de energia, mas, essencialmente elétrica e térmica. É expressa geralmente por meio de grandezas como a irradiância ou a irradiação. Em termos físicos, a irradiância corresponde à taxa instantânea de radiação solar recebida numa superfície de referência, expressa em W/m^2. Já a irradiação traduz a energia total acumulada sobre essa superfície ao longo de um intervalo temporal definido, sendo habitualmente expressa em kWh/m^2 ou MJ/m^2. Esta distinção conceptual, fundamental na literatura clássica (Duffie and Beckman, 2013), é particularmente relevante em estudos fotovoltaicos, dado que a produção elétrica depende da energia acumulada e não apenas da intensidade momentânea.

Antes de alcançar a superfície terrestre, a radiação solar passa por diferentes processos atmosféricos: pode ser transmitida, absorvida, refletida ou dispersa em proporções variadas pelos componentes da atmosfera, tais como vapor de água, aerossóis, ozono e diversos outros gases, consoante o comprimento de onda da radiação. Resultam como consequência destas

interações, três principais componentes da radiação solar incidente: *Global Horizontal Irradiance* (GHI), *Direct Normal Irradiance* (DNI) e *Diffuse Horizontal Irradiance* (DHI).

Com base no trabalho de (Duffie and Beckman, 2013), vamos definir cada uma destas componentes. A Figura 3 resume graficamente os principais fenómenos que condicionam a distribuição da radiação solar antes de atingir a superfície terrestre, ilustrando a complexidade da sua decomposição em GHI, DNI e DHI.

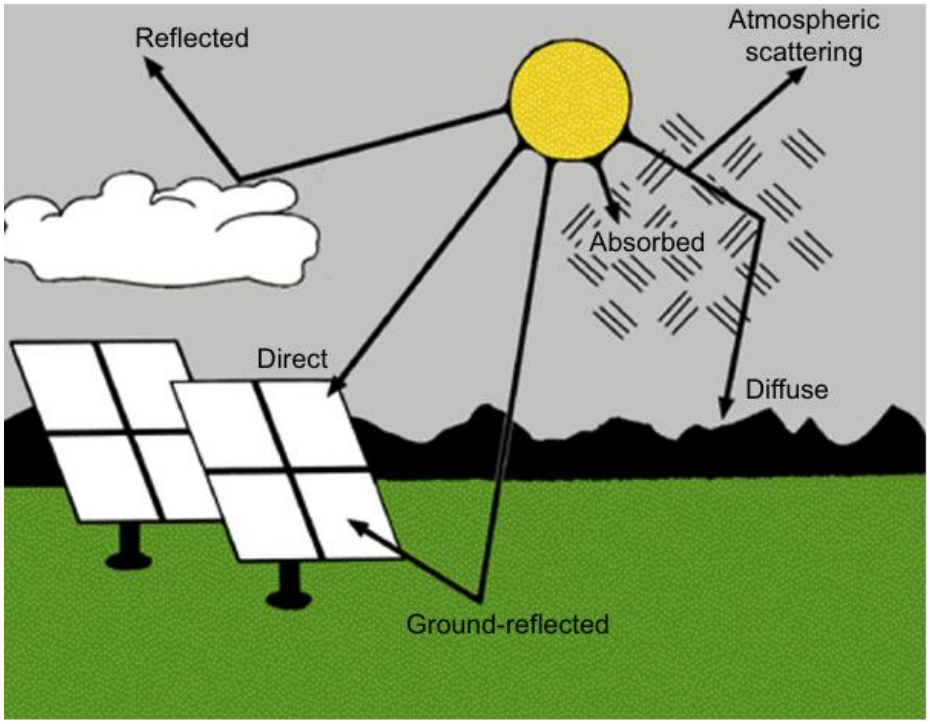


Figura 3 - Radiação solar incidente e seus componentes principais (Duffie and Beckman, 2013).

• *Global Horizontal Irradiance* (GHI): é a radiação total que chega ao solo terrestre num plano horizontal à superfície da Terra, isto é, a soma da radiação solar direta e difusa que incide sobre uma superfície horizontal. É a métrica mais relevante para sistemas fotovoltaicos fixos, com orientação não rastreada.

• *Direct Normal Irradiance* (DNI): representa uma parte da radiação solar que atinge uma superfície perpendicular à direção do Sol, sem ser espalhado pela atmosférica. É particularmente relevante para tecnologias de concentração solar (CSP e CPV).

• *Diffuse Horizontal Irradiance* (DHI): é a outra parte da radiação solar refletida pelos diversos constituintes atmosféricos, como moléculas de ar, aerossóis e nuvens, e representa a radiação recebida por uma superfície horizontal a partir de todas as direções do céu.

A decomposição destas componentes permite modelar de forma mais rigorosa o comportamento da radiação solar à escala local, sendo fundamental para o dimensionamento, simulação e previsão do desempenho energético de instalações solares. A distribuição relativa de GHI, DNI e DHI depende fortemente das condições atmosféricas, da posição do Sol e da localização geográfica do sítio de interesse.

2.1.1 Variabilidade do recurso solar

A radiação solar que chega à superfície terrestre é naturalmente variável, tanto no tempo como no espaço, dada a combinação complexa de fatores astronómicos e atmosféricos. Contudo, torna-se evidente a necessidade de compreender esta variabilidade que é crucial para o planeamento, operação e avaliação dos sistemas solares. Esta variabilidade repercute-se no desempenho energético dos sistemas fotovoltaicos e condiciona o nível de incerteza presente nas estimativas de produção a longo prazo (Duffie and Beckman, 2013; Rennée, 2016). Importa salientar que tal variabilidade tem origem em múltiplos fatores, começando ainda antes de a radiação solar atingir a atmosfera terrestre.

O Sol emite energia de forma praticamente constante. No entanto, a excentricidade da órbita terrestre introduz uma variação anual próxima de 3,3% na radiação extraterrestre, enquanto a inclinação axial do planeta é responsável pela sucessão das estações (Gören, 2021). Para além destes efeitos astronómicos, a quantidade de radiação recebida em cada local e instante é também modulada por parâmetros geométricos como a elevação solar, a declinação, o zénite e o azimute, grandezas bem descritas pela geometria solar clássica (Duffie and Beckman, 2013).

Ao passar pela atmosfera, a radiação solar sofre uma atenuação e dispersão provocadas pelas nuvens, aerossóis, vapor de água e outros constituintes, o que dá origem aos seus três principais componentes: direta, difusa e refletida. A composição relativa destas frações determina a natureza do recurso solar num determinado lugar e momento.

A variabilidade temporal do recurso solar manifesta-se em diversas escalas. Em escalas temporais curtas (de segundos a poucas horas), a variabilidade do recurso solar resulta sobretudo da passagem de nuvens e da dinâmica atmosférica local. Estes fenómenos afetam diretamente os sistemas fotovoltaicos, cuja produção elétrica acompanha de forma quase imediata as flutuações da irradiância (Duffie and Beckman, 2013). Já em horizontes de maior duração, como os ciclos sazonais ou interanuais, a variabilidade adquire importância distinta, uma vez que condiciona a análise de viabilidade financeira e o desempenho energético sustentado ao longo de toda a vida útil dos projetos (Renné, 2016).

Do ponto de vista espacial, a variabilidade do recurso solar tende a ser menos pronunciada em escalas regionais. Ainda assim, pode influenciar de forma relevante a escolha de locais para novos projetos fotovoltaicos. Nas regiões equatoriais, a irradiância mantém-se relativamente estável ao longo do ano, enquanto em latitudes médias e altas surgem contrastes sazonais bem mais evidentes (Renné, 2016). Para além destes padrões de ordem astronómica, fatores locais como a altitude, a topografia, a cobertura vegetal ou a proximidade de grandes massas de água podem alterar significativamente a radiação disponível, seja pela nebulosidade persistente, seja por efeitos orográficos ou de sombreamento (Duffie and Beckman, 2013). Acresce que, embora a componente direta (DNI) seja particularmente vulnerável às flutuações meteorológicas, a componente difusa contribui para suavizar essas variações, conferindo ao GHI uma maior estabilidade temporal e espacial (Wild, 2012). Portanto, a variabilidade do recurso solar não é apenas uma característica física a ser descrita mas também é um fator crítico na modelação do desempenho energético e na gestão do risco em sistemas solares. A consideração desta variabilidade em séries temporais extensas permite obter estimativas mais robustas do potencial de produção de um parque fotovoltaico, fornecendo suporte quantitativo essencial ao planeamento técnico e às decisões de investimento (Tavares et al., 2024).

2.1.2 Energia fotovoltaica

Na transição para um futuro energético mais limpo e sustentável, a energia solar fotovoltaica tem-se afirmado como uma das tecnologias renováveis mais promissoras, constituindo uma alternativa concreta aos combustíveis fósseis. A energia solar fotovoltaica, ou apenas energia fotovoltaica, como será designada no contexto desta tese, destaca-se como uma energia "verde" alimentada exclusivamente pelo sol. A tecnologia fotovoltaica pode ser aplicada em diferentes configurações, desde sistemas autónomos (*off-grid*) até soluções integradas na rede

elétrica (*on-grid*). Esta versatilidade permite a sua utilização em contextos tão diversos como instalações residenciais de pequena escala ou centrais de grande dimensão dedicadas à produção comercial de eletricidade (Duffie and Beckman, 2013).

A base física por trás da energia fotovoltaica pode ser explicada pelo efeito fotovoltaico. Este efeito foi descoberto pela primeira vez em 1839 pelo físico francês Alexandre Edmond Becquerel e consiste na capacidade que certos materiais semicondutores possuem de converter a luz solar directamente em eletricidade. Este fenómeno ocorre a nível microscópico, ou seja, os fótons provenientes da radiação solar incidem sobre o material semicondutor, fazendo com que os elétrons se desloquem da banda de valência para a banda de condução. O processo resulta na formação de pares eletrão-lacuna que, sob a influência do campo elétrico interno da junção p–n, são separados e conduzidos para contactos distintos, originando uma corrente elétrica contínua, conforme ilustrado na Figura 4 (Duffie and Beckman, 2013). A intensidade da corrente gerada está diretamente relacionada com a densidade de energia incidente, o que justifica a forte correlação entre o recurso solar local e a produção de eletricidade.

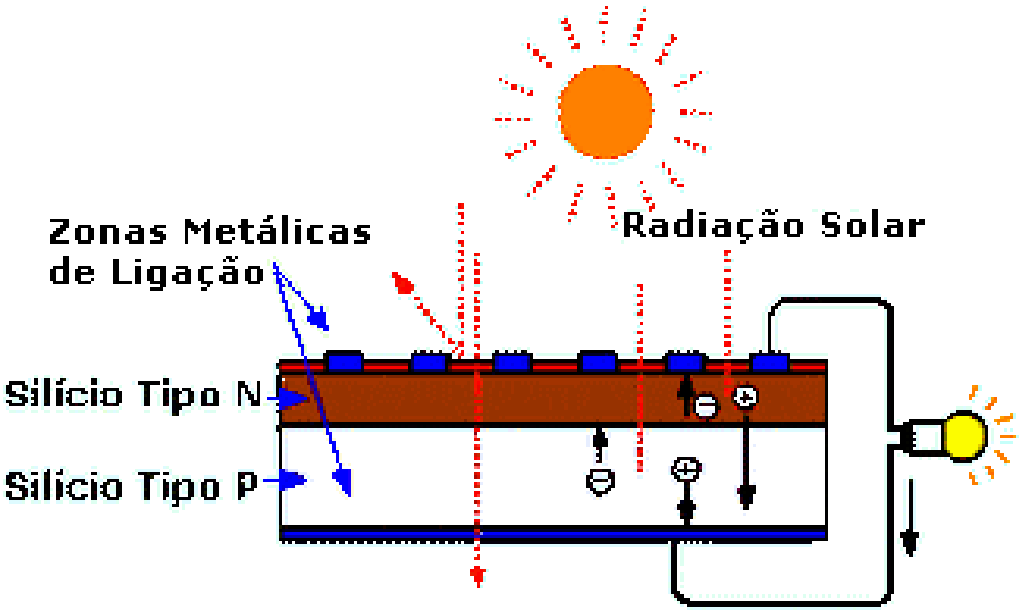


Figura 4 - Esquema simplificado do efeito fotovoltaico (Eletrónica PT, s.d.).

A conversão da luz solar em energia útil, neste caso fotovoltaica, não depende exclusivamente dos módulos fotovoltaicos, mas sim de um conjunto de tecnologias que formam um sistema

fotovoltaico. Um sistema fotovoltaico típico inclui, para além dos módulos, inversores (que convertem a corrente contínua em corrente alternada), dispositivos de proteção e monitorização, baterias no caso de sistemas isolados da rede, regulador de tensão, estruturas de fixação e componentes de cablagem. Em parques fotovoltaicos de grande dimensão é fundamental compreender que a produção de energia medida reflete o desempenho conjunto de todos estes elementos. Assim, o sinal utilizado na modelação estatística da produção elétrica reflete não apenas a radiação incidente, mas também o efeito acumulado de perdas térmicas, processos de conversão eletrónica, ineficiências de operação e até possíveis falhas do sistema (Wild, 2012; Duffie and Beckman, 2013). Além disso, a geração de eletricidade a partir de sistemas fotovoltaicos é afetada por vários fatores externos, como a temperatura do módulo, o ângulo de incidência da radiação, a acumulação de humidade e as condições de funcionamento do investidor. Consequentemente, importa sublinhar que os registos de produção não traduzem unicamente a disponibilidade do recurso solar, mas incorporam igualmente a influência de fatores técnicos e ambientais associados ao funcionamento do sistema (Wild, 2012; Tavares et al., 2024).

2.1.3 Bases de Dados de Radiação Solar

Atualmente, uma variedade de bases de dados de radiação solar encontram-se disponíveis e são amplamente utilizadas tanto em investigação científica como em aplicações industriais. Estas bases diferem significamente em termos de origem dos dados, resolução espacial e temporal e métodos de validação adotados. Na Tabela 1, apresenta-se um resumo comparativo das características mais importantes das bases de dados, (PVGIS, Solargis e ERA5), normalmente utilizadas em estudos de avaliação de recursos solares. Estas bases desempenham um papel crucial na previsão da produção de energia em sistemas fotovoltaicos, fornecendo séries estáveis e de longo prazo de radiação solar, essenciais para estudos de viabilidade e modelação estatística.

Tabela 1 - Síntese comparativa de bases de dados de radiação solar de longo prazo.

Característica	**PVGIS**	**Solargis**	**ERA5**
Origem dos dados	Satélite + modelos climatológicos	Satélite (MSG, geostacionários) + medição de solo	Reanálise climática (ECMWF)
Período disponível	2005 até ao presente	~1994 até ao presente	1950 até ao presente (ERA5),
Resolução temporal	Horária	15-30 minutos	Horária (ERA5)
Resolução espacial	~5 km	~1 km	~31 km (ERA5),
Variáveis principais	GHI, DNI, DHI, temperatura, vento	GHI, DNI, DHI, temperatura, albedo	GHI, nebulosidade, vento, pressão, etc
Método de validação	Comparação com estações de medição	Alta densidade de validação cruzada com redes solares de qualidade	Validação generalizada com dados de superfície
Vantagens	Gratuito, fácil acesso, metodologia transparente	Alta resolução e baixa incerteza, adequada para estudos financiáveis	Longa série temporal, adequado para análise climática
Limitações	Menor resolução e cobertura histórica limitada	Serviço pago (versão completa), metodologia proprietária	Resolução espacial grosseira, maior viés sistemático

2.2 Metodologias Pós-Construção em Sistemas Fotovoltaicos

Após o comissionamento, a produção efetiva dos parques fotovoltaicos pode revelar desvios face às estimativas pré-construtivas. Esses desvios decorrem de diferentes fatores, incluindo simplificações nos modelos físicos, processos de degradação natural dos equipamentos, variabilidade do recurso solar e até constrangimentos operacionais (Deceglie et al., 2023). Com o objectivo de recalibrar estas previsões iniciais e optimizar o desempenho energético ao longo do ciclo de vida útil, têm vindo a ser desenvolvidas e aplicadas diversas metodologias de avaliação pós-construtiva no sector solar fotovoltaico. Enquanto no setor eólico já existem iniciativas de sistematização, como o relatório técnico *Post-Construction Production Assessment of Wind Farms* (Lindvall, 2016), que organiza e descreve os principais métodos em uso, no domínio fotovoltaico continua a faltar uma revisão académica abrangente capaz de enquadrar estas abordagens numa perspetiva estatística, metodológica ou normativa comum.

Embora a literatura apresente casos aplicados relevantes e abordagens emergentes, não existe, até ao momento, uma tipologia formal amplamente aceite que permita classificar os métodos pós-operacionais em uso corrente. É neste contexto que esta secção, embora sucinta, apresenta uma visão geral dos principais métodos de estimativa de produção pós-construtiva aplicados no sector fotovoltaico.

A monitorização contínua do *Performance Ratio* (PR) ou razão de desempenho é uma das abordagens mais utilizadas na fase pós-construtiva de centrais fotovoltaicas. O PR, é definido como a relação entre a energia real injetada na rede e a energia que seria produzida pelo sistema, se operasse nas suas condições de referência , ou seja, considerando o recurso solar medido mas assumindo que a instalação converte esse recurso de acordo com o seu tamanho nominal em condições padrão de ensaio (irradiância de 1.000 W/m², temperatura de célula de 25 °C e espectro AM 1.5). Ao normalizar o rendimento para o recurso solar disponível e o tamanho do sistema, o PR capta o efeito agregado de todas as perdas (ineficiências do módulo, perdas do inversor e da cablagem, tempo de inatividade, etc.) Os operadores da central monitorizam frequentemente os valores de PR mensais e anuais e comparam-nos com o PR esperado do modelo de pré-construção. Se o PR estabilizar abaixo do previsto, as estimativas de produção futura podem ser ajustadas para refletir essa nova realidade. De igual modo, o acompanhamento prolongado do PR ao longo de vários anos permite determinar taxas efetivas de degradação, constituindo uma informação fundamental para a calibração e atualização de modelos de produção a longo prazo (Deceglie et al., 2023; IEA-PVPS, 2023).

Outro método é a utilização de modelos de regressão empírica que caracterizam e fazem a previsão de produção de energia com base nos dados meteorológicos. Um exemplo clássico da aplicação desta abordagem é a técnica de *Photovoltaics for Utility Scale Applications* (PVUSA) que faz o ajuste de regressão multilinear das produções de energia numa central em relação à radiação, temperatura ambiente e velocidade do vento. Ao coletar dados de curto termo normalmente dados de algumas semanas, quando a produção é estável e é possível estimar coeficientes que representam a capacidade efetiva e as sensibilidades de perda do sistema fotovoltaico. A equação de regressão é então usada para prever a potência esperada sob quaisquer condições ou para extrapolar o desempenho para condições de referência padrão. É dessa forma que se obtém uma estimativa baseada em dados de desempenho energético a longo prazo. Pela sua simplicidade, a regressão linear é frequentemente utilizada em testes de aceitação e na verificação do desempenho de novos parques fotovoltaicos (Swami et al., 2024).

A metodologia mais abrangente é a calibração dos modelos baseados nos dados da SCADA. Por exemplo, reparametrização de modelos físicos com base em dados reais de operação. Neste caso, modelos de simulação como o *Photovoltaic System Software* (PVsyst) ou o *System Advisor Model* (SAM) são ajustados com dados do desempenho real medidos (como irradiância, temperatura, produção elétrica) para que a simulação reproduza o comportamento observado da central. Os autores (Mubari and Ramahana, 2024) exemplificam esta abordagem ao recalibrar parâmetros de perdas e de desempenho numa instalação fotovoltaica real, tomando como referência o primeiro ano de operação. A seguir, o modelo é executado com dados de longo prazo (ex.: ano meteorológico típico) para atualizar a estimativa de produção futura. Este processo ajuda a aproximar o modelo da realidade do sistema em operação e a reduzir o desvio entre a previsão inicial e o desempenho observado. Apesar de ser prática corrente entre engenheiros e consultores, a investigação académica dedicada ao tema permanece relativamente escassa. Nos últimos anos começaram a ganhar destaque novas aborda gens, como a atualização Bayesiana (*Bayesian updating*) e diferentes algoritmos de *Machine Learning* (ML), sobretudo em estudos de caráter experimental. A atualização Bayesiana, em particular, permite incorporar progressivamente dados operacionais ao longo do tempo, o que contribui para diminuir a incerteza nas estimativas, embora a sua aplicação prática em projetos comerciais ainda seja limitada (Jadidi et al., 2020).

No domínio do ML, algoritmos como Random Forest, XGBoost ou redes neuronais têm evidenciado melhorias apenas marginais em termos de precisão, sobretudo em cenários marcados por elevada variabilidade ou ruído (Markovics and Mayer, 2022). Apesar disso, a sua adoção comercial continua limitada, em grande parte devido à complexidade metodológica e às dificuldades de interpretabilidade.

Todas estas metodologias partilham um objetivo comum: utilizar os dados reais de produção para recalibrar as estimativas de longo prazo. Nesse sentido, a secção seguinte aprofunda uma abordagem adicional inspirada em práticas consolidadas no sector eólico que visa precisamente este objetivo, com ênfase na sua adaptação ao contexto fotovoltaico.

2.3 Revisão Crítica do MCP

Na presente secção vamos analisar criticamente a aplicação da metodologia MCP no setor fotovoltaico. São analisadas as principais abordagens utilizadas, os modelos estatísticos

aplicados, as lacunas identificadas na literatura e as tendências metodológicas mais recentes, com o intuito de fundamentar o enquadramento científico do presente trabalho.

2.3.1 Evolução histórica do MCP no setor solar

O método MCP surgiu no setor eólico no final da década de 1990 com o intuito de suprir a necessidade de estimar recursos de longo prazo a partir de medições de curto prazo. Os dados medidos são recolhidos a partir de uma torre anemométrica instalada temporariamente (6 a 24 meses) no local de interesse. Estes dados são correlacionados com uma série histórica de longo prazo proveniente de uma estação de referência próxima.

A aplicação do MCP permite gerar uma série reconstruída que caracteriza o regime de vento de longo prazo no local de interesse, contribuindo para reduzir a incerteza das estimativas energéticas e os custos associados às campanhas de medição na fase pré-construtiva (Carta et al., 2013). Esta metodologia consolidou-se como referência no setor eólico, estando incorporada em normas como a IEC 61400-15-1:2022 e em diretrizes de qualidade, como as do MEASNET (2016), que estabelecem critérios formais para medições, correlações e quantificação de incertezas.

No setor fotovoltaico, a adoção do MCP ocorreu mais tarde e de forma gradual. Os primeiros estudos na década de 2010 exploraram técnicas de *site adaptation* (“adaptação ao sítio”) para melhorar as estimativas de irradiância solar de longo termo usando poucos dados medidos no local. Um dos primeiros exemplos desta adaptação encontra-se numa revisão preliminar conduzida por (Polo et al., 2016), que referencia estudos de 2014 nos quais foi sugerida uma metodologia análoga ao MCP aplicado à eólica. Nessa proposta, procedeu-se à calibração de séries de Irradiância Global Horizontal (GHI) obtidas por satélite com medições locais de curta duração, visando a correção de vieses presentes nos dados solares. Na época, utilizavam-se muitas terminologias diferentes, como bias correction, calibração de recurso, adaptação local ou dataset merging, que têm em comum, combinar medições in situ de curta duração com dados modelados de longa duração (satélite ou reanálise) para obter uma série solar de longo termo mais acurada. A partir de 2020, surgem estudos onde a metodologia passa a correlacionar diretamente dados reais de produção fotovoltaica de curto prazo com séries de GHI históricas, em vez de apenas GHI vs. GHI, que se baseia na relação entre séries de irradiância global horizontal curtas, medidas localmente, e séries históricas de GHI de longo prazo obtidas

via satélite, reanálise ou estações vizinhas. Essa evolução para a abordagem produção–GHI abriu novos caminhos a métodos híbridos que combinam informações de energia gerada e irradiância, introduzindo novos desafios, como a variabilidade operacional e perdas específicas do sistema (sujidade nos módulos, falhas de componentes, etc).

2.3.2 Abordagens GHI–GHI vs. Produção–GHI

A literatura atual tem destacado duas abordagens principais de aplicação do MCP no sector solar. O primeiro método, GHI–GHI, baseia-se na relação entre medições locais de GHI durante um curto período de tempo e dados históricos de satélites, reanálises ou estações meteorológicas vizinhas. Esta estratégia tenta criar um padrão de radiação a longo prazo que seja representativo da área de interesse. Estudos conduzidos por Polo et al. (2016) e Cebecauer et al. (2016) demonstraram que a calibração de dados de satélite com medições locais pode reduzir substancialmente os erros, alcançando diminuições no RMSE de até 70%.
A segunda abordagem, mais recente, é a produção - GHI, em que os dados de curto prazo não são irradiâncias, mas, a energia efetivamente gerada por um sistema fotovoltaico em operação (por exemplo, produção mensal ou horária, medida). Essa produção real é então correlacionada com séries de GHI de longo termo da mesma localização. Este método tem a vantagem de incorporar diretamente os efeitos operacionais reais (como perdas por sujidade, falhas ou degradação), que não estão presentes nas séries puramente meteorológicas. Narváez et al. (2021) ilustram esta mudança metodológica ao recorrer a algoritmos de ML para ajustar séries de irradiância modelada com base em dados de operação real, obtendo melhorias significativas na precisão das estimativas.

Neste sentido, a abordagem produção - GHI é bastante interessante do ponto de vista inovador, pois pode ser utilizada como uma ferramenta pós-construção. Quando se refere "inovador", entenda-se que este método pode ser usado não só para a calibração do recurso solar num determinado local, mas também para a estimativa de energia de longo prazo, permitindo recalibrar as projeções de produção futura (valores P50 e P90). O valor P50 representa uma estimativa com 50% de probabilidade de ser superada, enquanto o valor P90 é uma estimativa conservadora, com 90% de probabilidade de ser ultrapassada.

Na sua investigação, Reich et al. (2015) observaram que a incorporação de um ano adicional de dados operacionais numa central fotovoltaica permitiu reduzir a diferença entre os indicadores

P90 e P50 em mais de quatro pontos percentuais. De forma consistente, o relatório IEA-PVPS (2020) evidencia que o uso de dados reais na calibração das previsões tende a aumentar a precisão, sobretudo em contextos com elevada variabilidade climática ou exposição a riscos operacionais, como acumulação de sujidade nos módulos, degradação acelerada ou falhas técnicas. O avanço desta metodologia representa mais um passo na aplicação do modelo estatístico à realidade operacional dos sistemas fotovoltaicos, sendo reconhecido cada vez mais como uma alternativa viável à abordagem tradicional GHI–GHI nos estudos de estimativa energética.

2.3.3 Lacunas na literatura

Apesar dos avanços, identificam-se várias lacunas importantes na literatura sobre MCP solar:

(1) Falta de estudos com dados agregados mensais de produção. Na maioria dos trabalhos o MCP é aplicado em escala horária ou diária, aproveitando a alta resolução temporal para maximizar a correlação com a irradiância. No entanto, na prática pós-construção muitas vezes só se dispõem de dados mensais de energia, por exemplo, dados provenientes dos relatórios operacionais. A viabilidade de aplicação do MCP em dados mensais agregados permanece ainda pouco explorada. Não se encontraram referências que avaliem explicitamente um MCP calibrado com produção mensal indicando tratar-se de um ponto em aberto para investigação.

(2) Falta de validação sistemática da abordagem pós-operacional. Diversos autores sugeriram e alguns demonstraram reduções de incerteza ao calibrar modelos com dados reais de performance. Entretanto, ainda não existe um corpo consistente de literatura quantificando esses benefícios de forma comparativa em diferentes projetos, climas e tecnologias. De forma clara, ainda não existe, um estudo abrangente, ou um conjunto de estudos, que comprove em que medida a redução da incerteza por exemplo, nos valores P50/P90 proporcionada pelo método MCP após a construção, é generalizável. Enquanto tal evidência consolidada não existir, a abordagem pós-operacional permanecerá menos bancável perante investidores reforçando a urgência de pesquisas que preencham esse vazio metodológico.

3) Escassez de comparações entre diferentes bases de dados de irradiância. Existem hoje várias fontes de séries de longo prazo, desde modelos satelitais comerciais como o Solargis até dados de reanálise como o ERA5, ou produtos *open-source* como o PVGIS, o HelioClim e o NASA-CAMS, cada uma com níveis próprios de viés e de precisão regional. Apesar dessa diversidade, poucos

estudos avaliam de forma sistemática até que ponto a escolha da base de referência de GHI condiciona os resultados obtidos pelo MCP solar. Seria relevante compreender, por exemplo, se calibrar a produção de um parque com o Solargis conduz a estimativas diferentes das obtidas com outras bases, o que teria implicações diretas na avaliação de risco e na fiabilidade de estudos de bancabilidade.

Até ao momento, não são conhecidos estudos robustos comparativos nesse sentido. Relatórios técnicos, como o Global Solar Atlas 2.0 Validation Report do Banco Mundial, trazem avaliações gerais da exatidão dessas bases (indicando, por exemplo, que dados satelitais anteriores a 2005 possuem maior incerteza, ou quantificando RMSE de cada fonte em várias localidades). Porém, falta transpor essas diferenças para o contexto do MCP. Em suma, a seleção da base de irradiância de longo prazo nos estudos MCP tem sido feita de forma não padronizada e variável, abrindo espaço para pesquisas que comparem sistematicamente múltiplas bases dentro de um mesmo framework MCP estabelecendo quais fornecem menores erros em distintos cenários temporais e geográficos.

3. Metodologia

Neste capítulo descreve-se a metodologia adotada para a aplicação e validação da abordagem MCP no contexto pós-construção de parques fotovoltaicos. O processo foi estruturado em etapas sequenciais, abrangendo a seleção da base de dados de radiação mais adequada, a verificação estatística dos pressupostos do modelo, a deteção e tratamento de outliers e a comparação com modelos preditivos alternativos.

No que diz respeito às ferramentas analíticas, a utilização da regressão linear simples combinada com a metodologia MCP foi executada no Microsoft Excel, por meio de uma ferramenta interna criada pela empresa Megajoule, especializada em consultoria energética. As análises restantes, como a deteção de outliers, regressão múltipla, testes estatísticos, aplicação de modelos de ML e geração de gráficos, foram realizados em Python[3], utilizando o ambiente Jupyter Notebook[4] e empregando bibliotecas como pandas, numpy, scikit-learn, matplotlib, seaborn e statsmodels. Essa combinação de ferramentas teve como objetivo garantir a reprodutibilidade científica, a precisão estatística e a conformidade com as práticas profissionais do setor.

3.1 Caracterização dos Dados Utilizados

Para a realização do presente trabalho, considerou-se um conjunto de 19 parques fotovoltaicos em operação. Estes parques estão todos localizados em Espanha, distribuídos pelo norte e centro do país, mais especificamente nas regiões de Aragón e Castilla La Mancha (Figura 5). Estes 19 parques foram agrupados em quatro clusters de acordo com a respetiva proximidade geográfica. A sua seleção foi baseada na qualidade dos dados operacionais disponíveis, resultantes de estudos pós-construção realizados previamente pela empresa Megajoule.

---

[3] *Python* é uma linguagem de programação amplamente utilizada em ciência de dados, estatística e modelação.

[4] *Jupyter Notebook* é um ambiente interativo que permite executar código, visualizar resultados e documentar análises de forma integrada

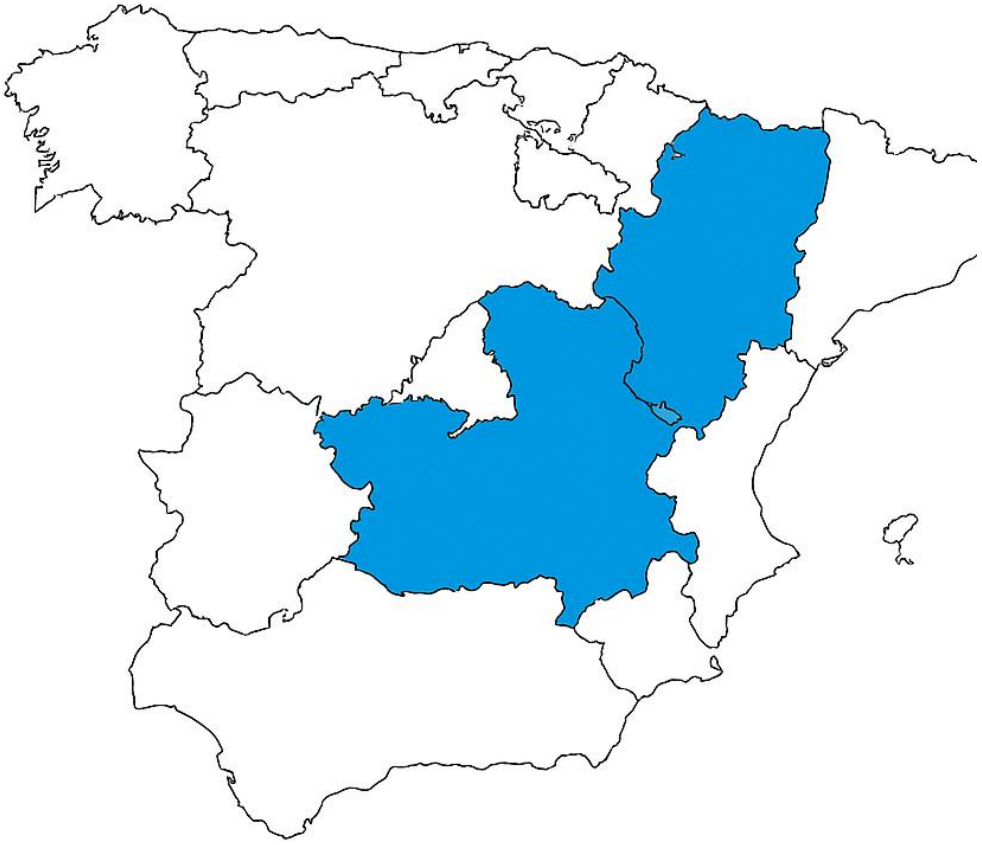


Figura 5 - Localização geográfica dos parques selecionados (rbiedermann, n.d.).

O período coberto pelos dados operacionais é de aproximadamente três anos e meio, desde o início da operação dos parques em 2020 até meados de 2023. As potências instaladas variam entre 40,2 MWp e 49,9 MWp, abrangendo assim uma gama representativa de parques solares comerciais de média dimensão. Na Tabela 2, apresentam-se as informações genéricas relativas a cada um dos parques fotovoltaicos considerados neste estudo.

Por outro lado, foram também utilizadas séries de radiação solar de longo termo provenientes de três bases de dados distintas: PVGIS, ERA5 e Solargis, com resolução mensal, cobrindo o período compreendido entre 2004 e 2023.

No âmbito deste trabalho, estas bases foram subdivididas em sete séries independentes, identificadas desde S1 a S7, com a seguinte correspondência:

- Solargis é representada por uma única série (S1);
- ERA5 deu origem a quatro séries diferentes (S2, S3, S4 e S5);
- PVGIS foi desdobrada em duas séries (S6 e S7).

Tabela 2 - Informação genérica dos parques fotovoltaicos.

Cluster	**Parques**	**Capacidade instalada (MWp)**	**Localização geral**
Cluster-1	PV01[5]	45,0	Ciudad Real
	PV02	45,0	Ciudad Real
	PV17	49,9	Ciudad Real
	PV19	49,9	Ciudad Real
Cluster-2	PV05	40,2	Zaragoza
	PV04	45,6	Zaragoza
	PV11	40,2	Zaragoza
	PV06	49,9	Zaragoza
	PV09	49,9	Zaragoza
	PV03	49,9	Zaragoza
	PV10	49,9	Zaragoza
Cluster-3	PV14	49,9	Zaragoza
	PV13	49,9	Zaragoza
	PV18	49,9	Zaragoza
	PV15	49,9	Zaragoza
Cluster-4	PV07	49,9	Zaragoza
	PV08	49,9	Zaragoza
	PV16	49,9	Zaragoza
	PV12	49,9	Zaragoza

3.2 Adaptação da Metodologia MCP

A metodologia MCP tradicional, descrita previamente na secção 2.3, foi adaptada no presente estudo de forma a permitir a sua aplicação ao contexto da energia fotovoltaica. Em termos

[5] Cada parque fotovoltaico incluído neste estudo foi identificado por um código sequencial de PV, abreviatura de 'Photovoltaics', seguido de um número de 1 a 19 (PV01 a PV19), de modo a preservar a confidencialidade dos projetos analisados.

gerais, a metodologia baseia-se na utilização de dois conjuntos de dados distintos: um conjunto de curto prazo, baseado em dados reais medidos durante o período operacional dos parques, e um conjunto de longo prazo, obtido a partir de séries históricas de radiação solar provenientes de bases de dados de reanálise e satélite.

No caso dos dados de curto prazo, foram utilizados registos mensais de produção elétrica medida em parques fotovoltaicos com cerca de três anos e meio de operação. Para garantir a comparabilidade com dados de longo prazo e compensar eventuais indisponibilidades parciais durante o período operacional, os valores de produção foram corrigidos por disponibilidade. Esta correção foi realizada através do quociente entre a produção mensal registada e o valor percentual de disponibilidade técnica do parque nesse mesmo mês, resultando numa série de produção bruta corrigida para uma disponibilidade ideal de 100%.

A escolha da regressão linear simples deve-se ao facto de esta ser a técnica tradicional mais utilizada no contexto da metodologia MCP, amplamente reconhecida na literatura especializada. A sua adoção neste trabalho foi motivada não apenas pela sua simplicidade, aplicabilidade prática e facilidade de interpretação, mas também pela sua robustez estatística em aplicações com séries temporais mensais. A GHI constitui um dos principais fatores que explicam a produção de energia em sistemas fotovoltaicos. Em muitos casos, a relação entre a GHI e a produção mensal pode ser descrita de forma suficientemente aproximada por um modelo linear simples. Quando os dados são analisados em séries temporais mensais, esta modelação permite identificar a tendência média entre radiação e geração elétrica, revelando-se uma ferramenta prática sobretudo em análises de longo prazo, ainda que com algumas limitações conhecidas.

A IEA-PVPS (2022), entre outros organismos de referência, recomenda igualmente esta abordagem, que tem sido amplamente utilizada tanto em estudos de pré-viabilidade técnica como em análises pós-operacionais de ativos fotovoltaicos. Os dados reais de produção mensal de cada parque fotovoltaico foram correlacionados, através da regressão linear, com cada uma das sete séries de radiação históricos de longo termo para o período simultâneo de dados, obtendo assim sete correlações para cada parque.

A Equação 1 representa a estrutura da regressão linear utilizada neste estudo:

$$P = a \times GHI + b \quad (1)$$

Onde:

- P representa a produção mensal bruta corrigida por disponibilidade;
- GHI é o valor mensal de radiação global horizontal proveniente da base de dados selecionada;
- a é o coeficiente angular da regressão, que traduz a sensibilidade da produção em relação à radiação solar;
- b é o termo independente (interceção), que pode refletir perdas constantes, desvios sistemáticos ou efeitos de fundo no sistema.

Para melhorar a robustez da correlação linear e reduzir o efeito de outliers (valores atípicos que se desviam significativamente do padrão esperado dos dados), foi aplicada uma metodologia híbrida de remoção de outliers, cuja descrição completa encontra-se detalhada na secção 3.4. Esta etapa é de clara importância, uma vez que a regressão linear é sensível à presença de valores extremos, podendo assim enviesar significativamente a relação estatística entre as variáveis. Após a filtragem inicial, procedeu-se à análise das correlações entre produção e cada base de radiação, de forma a selecionar a série de referência mais adequada para o modelo MCP. O critério de seleção considerou simultaneamente o coeficiente de determinação (R^2) e o viés médio (MBE), conforme descrito na secção 3.5, privilegiando a minimização do viés sistemático pela sua relevância para estimativas de longo prazo.

Após escolher a base mais apropriada, a equação de regressão linear obtida foi aplicada à série de dados de GHI selecionados ao longo de 20 anos, permitindo assim a reconstrução de uma série mensal de produção estimada para o mesmo horizonte temporal. As estimativas das séries mensais reconstruídas foram posteriormente convertidas, numa série anual somando as produções mensais correspondentes. Finalmente o valor médio anual obtido para o período de 20 anos constituiu a estimativa da produção energética de longo prazo do parque.

3.3 Métricas Estatísticas de Avaliação de Desempenho

A utilização de métricas estatísticas adequadas é fundamental para a análise da qualidade de ajustamento dos modelos e da sua capacidade preditiva no contexto deste estudo.

As métricas utilizadas foram o *Coefficient of Determination* (R^2), o *Root Mean Square Error* (RMSE) e o *Mean Bias Error* (MBE). Estas permitem uma avaliação complementar: o R^2 mede a

proporção da variabilidade explicada pelo modelo, enquanto o RMSE quantifica a magnitude dos erros e o MBE indica a existência de viés sistemático entre os valores previstos e os reais.

Importa salientar que, embora a nomenclatura tradicional seja mantida ao longo do texto, todas as métricas de erro foram utilizadas na sua forma normalizada, ou seja, expressas em percentagem (%) da produção média real. Esta abordagem correspondente às designações nRMSE e nMBE na literatura foi adotada de forma deliberada para assegurar a comparabilidade entre parques com diferentes escalas de produção e garantir consistência estatística. A normalização é considerada boa prática em estudos aplicados a séries reais de produção fotovoltaica (Markovics and Mayer, 2022).

***Coefficient of Determination* (R^2)**

O coeficiente de determinação mede a proporção da variância nos dados dependentes explicada pelo modelo. Um valor próximo de 1 indica forte capacidade explicativa, embora não traduza a magnitude dos erros.

$$R^2 = 1 - \frac{\sum_{i=0}^{N}(P_i - f_i)^2}{\sum_{i=0}^{N}\left(P_i - \widehat{P_i}\right)^2} \quad (2)$$

***Root Mean Square Error* (RMSE)**

O RMSE fornece uma medida da magnitude média dos erros de previsão, sendo particularmente sensível a grandes desvios individuais.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(f_i - P_i)^2} \quad (3)$$

***Mean Bias Error* (MBE)**

O MBE indica a tendência média do modelo em sobrestimar ou subestimar os valores reais, revelando a presença de viés sistemático.

$$\mathrm{MBE} = \frac{1}{N}\sum_{i=0}^{N}(f_i - P_i) \tag{4}$$

onde f_i representa o valor previsto, P_i o valor observado, $\widehat{P_i}$ a média dos valores observados, e N o número total de observações.

As métricas selecionadas permitiram avaliar de forma quantitativa a correlação entre a radiação solar e a produção elétrica nos parques fotovoltaicos em análise aplicadas em dois momentos-chave: (i) na seleção da base de dados climática mais adequada e (ii) na comparação do desempenho preditivo dos modelos MCP e dos modelos alternativos baseados em aprendizagem automática.

3.4 Análise de outliers

A identificação e tratamento de outliers constitui uma etapa essencial para garantir a robustez da metodologia MCP. Estes valores anómalos podem comprometer a qualidade da regressão entre radiação e produção, afetando a fiabilidade dos modelos de previsão.

Tendo isto em consideração, foi desenvolvida uma abordagem integrada que combina critérios estatísticos, operacionais e visuais. O objetivo é assegurar que apenas dados representativos do funcionamento real dos sistemas sejam utilizados na calibração. A metodologia descrita a seguir está estruturada em quatro objetivos complementares:

- Validar estatisticamente a melhoria da regressão após remoção dos outliers;
- Agrupar os parques por sensibilidade à presença de outliers;
- Explorar a distribuição e sazonalidade dos outliers;
- Estudar as causas técnicas e operacionais associadas aos outliers;

3.4.1 Validação estatística da regressão após remoção dos outliers

O objetivo inicial consistiu em analisar o impacto da remoção de outliers na robustez estatística dos modelos MCP. Desta forma, foram elaboradas regressões lineares simples entre a produção mensal medida e a radiação estimada pela base Solargis, utilizando-se a ferramenta de análise MCP usada na empresa Megajoule, a qual foi desenvolvida em Excel, com os dados brutos como base de referência. Nesta fase, a identificação de outliers baseou-se numa combinação de análise visual e critérios operacionais, tendo sido assinalados os pontos mais afastados da linha de tendência da regressão e os meses com disponibilidade inferior a 85%, em conformidade com as orientações da norma IEC (2021). Este parâmetro é muito utilizado em estudos pós-construção por consultorias como a empresa Megajoule, por assegurar que apenas períodos representativos do funcionamento normal sejam considerados na calibração. A análise dos modelos antes e depois da filtragem possibilitou a verificação de melhorias consistentes em R^2, RMSE e MBE, corroborando a eficácia do processo de limpeza na estabilidade do ajuste.

3.4.2 Agrupamento dos parques por sensibilidade à presença de outliers

Para além da melhoria estatística que foi o primeiro objetivo, o segundo objetivo procurou-se em cada parque quantificar a sensibilidade dos modelos MCP. A abordagem escolhida consistiu em calcular a variação percentual do coeficiente de determinação referida aqui como ΔR^2, que traduz a diferença relativa no desempenho do modelo antes e depois da remoção desses valores anómalos .

Com base nos valores de ΔR^2 observados, os parques fotovoltaicos foram classificados em três níveis de sensibilidade estatística: Tipo A (estáveis, $\Delta R^2 < 2$ %), Tipo B (sensibilidade moderada, $2\ \% \leq \Delta R^2 < 10\ \%$) e Tipo C (alta sensibilidade, $\Delta R^2 \geq 10$ %). Estes intervalos foram definidos de forma empírica, tendo como referência a literatura especializada em regressão linear, onde se reconhece que variações superiores a 10 % em métricas de ajuste, como o coeficiente de determinação (R^2), podem indiciar instabilidade estatística significativa e sinais de sobreajuste (Chatterjee and Hadi, 2006). Por outro lado, variações inferiores a 2 % tendem a ser interpretadas como oscilações dentro da margem de erro esperada, não alterando de forma relevante a estrutura do modelo. Esta classificação tornou possível uma comparação mais

consistente entre os diferentes ativos, evidenciando aqueles cuja performance se mostrou mais sensível à presença de valores anómalos e aproximando-se das práticas habitualmente seguidas na modelação estatística aplicada ao setor energético. Esta abordagem possibilitou uma análise comparativa clara entre os diferentes ativos avaliados, contribuindo para a identificação daqueles mais vulneráveis à presença de dados anómalos.

Por outro lado, variações inferiores a 2 % tendem a ser interpretadas como oscilações dentro da margem de erro esperada, não alterando de forma relevante a estrutura do modelo. Esta classificação tornou possível uma comparação mais consistente entre os diferentes ativos, evidenciando aqueles cuja performance se mostrou mais sensível à presença de valores anómalos e aproximando-se das práticas habitualmente seguidas na modelação estatística aplicada ao setor energético.

3.4.3 Distribuição e sazonalidade dos outliers

Com o intuito de melhorar os resultados relativos aos dois primeiros objetivos, a partir do terceiro objetivo que visa explorar a distribuição e sazonalidade dos outliers introduziu-se técnicas estatísticas para a detecção de outliers, as quais foram aplicadas aos resíduos da regressão MCP, obtidos a partir da diferença entre a produção medida e estimada. A análise foi realizada utilizando o software Python, com a utilização de bibliotecas especializadas, tais como pandas, numpy, scipy, statsmodels, seaborn e scikit-learn, assegurando assim rigor estatístico na detecção, análise e visualização gráfica dos outliers. Foram aplicados dois métodos para a identificação de outliers: o IQR, que é sensível a distribuições assimétricas, e o Z-score, que se mostra apropriado para variáveis que apresentam uma distribuição aproximadamente normal. Foi realizada uma análise da distribuição temporal dos outliers por mês, consolidando os dados de todos os parques numa série mensal. A representação visual baseou-se em gráficos de barras, os quais indicam a quantidade mensal de outliers, e um heatmap que ilustra a distribuição dos outliers por parque e mês.

Para testar a significância das variações mensais, usou-se o teste não paramétrico de Kruskal–Wallis. Recorreu-se ainda à decomposição STL (Seasonal-Trend decomposition using Loess) para separar tendência, sazonalidade e ruído na série mensal. Por fim, aplicaram-se os

algoritmos de K-means e Ward para agrupar os parques por padrões sazonais de outliers, permitindo identificar perfis estatísticos distintos de comportamento anómalo.

3.4.4 Causas técnicas e operacionais associadas aos outliers

Prosseguindo com a identificação estatística dos outliers, passamos para o quarto objetivo: perceber quais as causas que estiveram na origem desses desvios do ponto de vista técnico e operacional. Para isso, a análise centrou-se em dois indicadores fundamentais mais especificamente, a disponibilidade operacional e o *performance ratio* (PR). A disponibilidade traduz o tempo em que os equipamentos estiveram tecnicamente capazes de produzir energia, enquanto o PR avalia o grau de eficiência do sistema, tendo em conta as condições meteorológicas reais. A relevância destes indicadores é sublinhada por normas e entidades internacionais, incluindo a IEC (2021), o Departamento de Energia dos Estados Unidos (DOE, 2022) e o NREL (2024), que os destacam como ferramentas fundamentais para a deteção de falhas, a quantificação de perdas e a garantia da fiabilidade dos dados em sistemas solares. A sua utilização conjunta permite contextualizar os desvios observados, distinguindo falhas operacionais de variações climáticas. Foram elaborados gráficos de dispersão e histogramas que permitiram cruzar estes indicadores com os meses anteriormente classificados como outliers. Recorreu-se também aos relatórios técnicos de *due diligence*[6] de cada parque para identificar ocorrências específicas como avarias, *curtailments*[7] e operações de manutenção coincidentes com os períodos em análise. Na ausência de evidência operacional suficiente, avaliou-se a ocorrência de condições meteorológicas extremas, comparando a irradiância mensal observada com a distribuição de longo prazo para o mesmo mês.

---

[6] Due diligence: processo de auditoria técnica, financeira e documental destinado a avaliar de forma sistemática os riscos, a viabilidade e as condições de um projeto ou ativo antes de decisões estratégicas, como financiamento, aquisição ou investimento.

[7] Curtailment designa a limitação deliberada da produção de uma central renovável, geralmente imposta por restrições da rede elétrica ou por motivos operacionais, implicando redução direta da energia entregue e das receitas.

3.4.6 Testes estatísticos

A seguir apresentam-se os testes estatísticos aplicados na análise de outliers, tal como descrito anteriormente na secção 3.4. A seleção destes métodos baseou-se nas características das séries de dados e na sua adequação ao contexto fotovoltaico.

Interquartile Range (IQR)

O método do Intervalo Interquartil é uma técnica robusta e muito aplicada para identificar valores atípicos em distribuições assimétricas ou com presença de extremos. Baseia-se na diferença entre o terceiro e o primeiro quartil ($IQR = Q_3 - Q_1$). Consideram-se outliers os valores que se situam abaixo de $Q_1 - 1.5 \times \mathrm{IQR}$ ou acima de $Q_3 + 1.5 \times \mathrm{IQR}$ (Tukey, 1977).

Z-score (pontuação padronizada)

O Z-score quantifica a distância de uma observação em relação à média da amostra x_i, em unidades de desvio-padrão σ, segundo a Equação 5. Valores com $|Z_i| > 3$ são geralmente considerados outliers em distribuições aproximadamente normais (Montgomery and Runger, 2014).

$$Z_i = \frac{x_i - \bar{x}}{\sigma} \tag{5}$$

Teste de Kruskal–Wallis

Este teste estatístico não paramétrico permite comparar as medianas de três ou mais grupos independentes, sem assumir normalidade ou homogeneidade de variâncias. A estatística de teste é dada por (Equação 6), onde n_j é o tamanho do grupo, R_j a média dos ranks, e N o

número total de observações. Foi utilizado para avaliar variações sazonais na distribuição mensal dos outliers (Kruskal and Wallis, 1952).

$$\mathrm{H} = \frac{12}{N(N+1)} \sum_{j=0}^{k} n_j \left(R_j - \bar{R}\right)^2 \tag{6}$$

Decomposição STL

A técnica STL (Seasonal-Trend decomposition using Loess) permite decompor uma série temporal Y_t em três componentes: tendência T_t , sazonalidade S_t e resíduo R_t , tal que $Y_t = T_t + S_t + R_t$ .Utilizando regressão local (Loess), mostrou-se especialmente eficaz na identificação de padrões sazonais persistentes nas séries de outliers mensais (Cleveland et al., 1990).

Clusterização (K-means e Ward)

A clusterização foi utilizada para agrupar os parques solares com base nos padrões mensais de outliers. O algoritmo K-means minimiza a soma das distâncias quadráticas dentro dos grupos (Kanungo et al., 2002) e método de Ward por sua vez, agrupa iterativamente os elementos minimizando a variância intra-cluster (Ward, 1963). Correspondentes as Equações 7 e 8.

$$min \sum_{i=1}^{k} \sum_{x \in C_i} \|x - u_i\| \, 1^2 \tag{7}$$

$$\Delta E = \frac{n_A n_B}{n_A + n_B} \|u_A - u_B\|^2 \tag{8}$$

3.5 Escolha da Base de Dados de Radiação

A seleção da base de dados climáticos mais adequada constitui uma etapa essencial para assegurar a fiabilidade do modelo de correlação MCP, particularmente no contexto da estimativa da produção energética de longo prazo. Dada a existência de múltiplas fontes de dados, quer reanalíticos, quer semi-empíricos, com diferentes resoluções temporais e espaciais, um dos objetivos desta tese é precisamente realizar uma análise comparativa entre três bases de radiação específicas: PVGIS, ERA5 e Solargis. Pretende-se compreender, em diferentes localizações, qual das bases traduz uma melhor explicação da variabilidade da produção energética observada em cada parque.

Cada base de radiação foi representada através de diferentes versões, perfazendo um total de sete séries distintas. Estas séries foram correlacionadas individualmente com os dados mensais de produção medida em cada um dos 19 parques fotovoltaicos em estudo, através de regressões lineares. Com base nesta abordagem, foi possível gerar séries independentes para posterior avaliação comparativa.

A determinação da base de radiação mais apropriada foi fundamentada em três métricas estatísticas amplamente utilizadas na literatura: o *Coefficient of Determination* (R^2), o *Root Mean Square Error* (RMSE) e o *Mean Bias Error* (MBE). Estas métricas (ver equações de 2, 3 e 4), foram calculadas individualmente para todos os parques, permitindo comparações diretas entre o desempenho das sete versões das bases de radiação.

Com o intuito de verificar se as diferenças observadas entre as bases eram estatisticamente significativas, foi selecionada a melhor versão dentro de cada uma das bases, tendo como critério de seleção a versão com maior número de valores mais favoráveis de cada métrica estatística.

Posteriormente, as diferenças entre bases de radiação foram avaliadas para cada métrica de desempenho (R^2, RMSE e MBE). Para este efeito, aplicou-se uma análise de variância (ANOVA) quando os pressupostos de homogeneidade e normalidade se encontravam satisfeitos e, em alternativa, recorreu-se ao teste não paramétrico de Kruskal–Wallis.

Para complementar esta análise estatística quantitativa, realizou-se também uma verificação exploratória da aleatoriedade dos resíduos gerados por cada modelo base-parque através de

gráficos de dispersão. Esta avaliação qualitativa visou identificar quaisquer tendências sistemáticas ou padrões nos resíduos que pudessem indicar instabilidade na relação entre radiação e produção. Contudo, optou-se por reservar a análise estatística formal da independência dos resíduos para a secção seguinte (secção 3.6), onde os pressupostos da regressão linear são abordados com maior profundidade.

Finalmente, a escolha definitiva da base de radiação baseou-se numa combinação equilibrada entre robustez estatística (R^2 mais elevado), desempenho global (menor RMSE) e viés sistemático reduzido (MBE mais próximo de zero). Esta estratégia contribuiu para estabelecer uma metodologia equilibrada, que concilia qualidade de ajuste, precisão e imparcialidade nas previsões, em consonância com as melhores práticas internacionais e com estudos recentes sobre modelação solar pós-operacional (Markovics and Mayer, 2022; Lindvall, 2016).

3.6 Validação da Metodologia MCP Adaptada

O objetivo principal desta tese é validar a metodologia MCP para estimativas de produção energética em parques fotovoltaicos em operação. Por isso, esta secção desempenha um papel fundamental, assegurando que os resultados obtidos sejam confiáveis e cientificamente robustos. Para alcançar este objetivo, é imprescindível verificar rigorosamente se os pressupostos básicos da regressão linear estão satisfeitos nas correlações obtidas com a base de longo prazo selecionada.

A validação foi conduzida através de uma análise detalhada dos resíduos gerados pelos modelos de regressão linear desenvolvidos para cada parque fotovoltaico. Cada uma das sete séries de radiação solar de longo prazo (quatro da ERA5, duas da PVGIS e uma da Solargis) foi analisada separadamente, com o intuito adicional de fornecer mais evidências sobre a escolha da base mais adequada. Os pressupostos da regressão linear: normalidade, independência e homocedasticidade dos resíduos, bem como a significância global do modelo foram avaliados através de testes estatísticos reconhecidos. Concretamente, utilizou-se o teste de ANOVA, o teste de Shapiro–Wilk, o teste de Durbin–Watson e o teste de Levene.

A *Analysis of Variance* (ANOVA) foi usada para determinar a significância global do modelo de regressão linear ajustado. Este teste permite perceber se as variáveis independentes (neste

caso, a radiação solar histórica) conseguem explicar adequadamente a variabilidade observada na produção real dos parques fotovoltaicos. Para esta avaliação é utilizado o valor de prova-p como indicador estatístico: caso este valor seja inferior ao nível crítico (geralmente 0,05), conclui-se que o modelo é estatisticamente significativo e que a regressão linear fornece um ajuste apropriado aos dados estudados (Department of Statistics and Data Science Yale University, 1998).

Para verificar a normalidade dos resíduos, recorreu-se ao teste de Shapiro–Wilk. Este teste avalia se os dados seguem uma distribuição normal ao comparar os valores observados com os esperados sob essa hipótese. Um valor-p inferior ao nível de significância estipulado ($\alpha = 0,05$) indica a rejeição da hipótese nula de normalidade, sugerindo que os resíduos não se distribuem normalmente e podendo, por conseguinte, comprometer a validade dos pressupostos da regressão linear (Shapiro and Wilk, 1965).

Outro aspeto crucial analisado foi a independência dos resíduos, avaliada pelo teste de Durbin–Watson. Este teste verifica a presença de autocorrelação entre resíduos consecutivos, isto é, se o erro num dado mês está correlacionado com o erro dos meses anteriores. A estatística resultante varia de 0 a 4, onde um valor próximo de 2 sugere ausência de autocorrelação significativa. Valores substancialmente distantes deste ponto indicam dependências nos resíduos, comprometendo a validade do modelo e, por conseguinte, a robustez das previsões de produção energética a longo prazo (Kenton, 2023).

Finalmente, a homocedasticidade, ou seja, a constância das variâncias dos resíduos, foi avaliada pelo teste de Levene ($\alpha = 0.05$). Este pressuposto é importante para assegurar que os erros do modelo de regressão têm uma variância constante ao longo do intervalo estudado. O não cumprimento deste pressuposto pode indicar que a regressão apresenta desempenhos diferentes consoante as faixas de produção ou radiação, o que limitaria o seu uso generalizado (Gastwirth and Gel, 2009) .

Uma vez verificados e satisfeitos estes pressupostos estatísticos fundamentais, torna-se possível validar formalmente a adequação da metodologia MCP adaptada ao setor fotovoltaico e assegurar a robustez e confiabilidade das estimativas energéticas geradas ao longo do horizonte temporal pretendido.

3.7 Modelos Alternativos ao MCP

A literatura científica revela uma lacuna significativa no que diz respeito à aplicação de técnicas de ML para previsão de produção energética com dados mensais em sistemas fotovoltaicos. A maioria dos estudos concentra-se em séries horárias ou diárias, deixando pouco explorado o potencial preditivo de modelos ML em cenários com resolução temporal maior, como é o caso da produção mensal. Esta análise assume particular importância no setor fotovoltaico, sobretudo quando aplicada a estimativas de produção de longo prazo que servem de base a estudos de bancabilidade, avaliações de risco (P50/P90), auditorias técnicas e processos de *due diligence* (IEA PVPS, 2023).

Neste sentido, após a validação da metodologia MCP com recurso à regressão linear simples, constatou-se que, embora os resultados globais fossem estatisticamente robustos, houve incumprimento parcial dos pressupostos de regressão em alguns parques. Este resultado motivou a exploração de alternativas metodológicas mais flexíveis e robusta com o objetivo de verificar se outros modelos poderiam melhorar o desempenho preditivo sem comprometer a robustez estatística.

A estratégia metodológica foi estruturada em dois grupos principais de modelos:

- Modelos lineares clássicos, em particular a regressão linear múltipla, aplicada com o objetivo de avaliar o contributo de variáveis meteorológicas adicionais (GHI, temperatura, vento, pressão atmosférica, humidade relativa e precipitação). Pretendeu-se ainda verificar se a regressão múltipla poderia mitigar as violações dos pressupostos estatísticos anteriormente identificados, funcionando como complemento à regressão linear simples;
- Modelos não lineares de *machine learning*, com ênfase no XGBoost (*Extreme Gradient Boosting*) e no Random Forest Regressor, ambos amplamente referenciados pela literatura pela sua capacidade de capturar interações complexas, lidar com colinearidade e identificar padrões não lineares, mantendo robustez mesmo em contextos com séries de dados relativamente limitadas (Friedman, 2001; Breiman, 2001).

Para garantir uma comparação equitativa, as regressões lineares simples e múltipla foram aplicadas ao mesmo conjunto de dados mensais de produção. Na regressão múltipla, foram utilizadas como variáveis independentes todas as variáveis meteorológicas disponíveis.

Os modelos foram treinados sem validação cruzada, mantendo coerência com a lógica da metodologia MCP, baseada em regressão determinística simples. As métricas de avaliação utilizadas foram R^2, RMSE e MBE na sua forma normalizada (secção 3.2), tal como definido pela norma IEC 61724-1:2021.

No que toca aos modelos ML, foi adotada uma abordagem metodológica mais exigente, com foco na mitigação de sobreajuste (overfitting) e validação estatística rigorosa. O processo seguiu os seguintes passos:

- Pré-processamento dos dados: remoção de valores em falta, tratamento de outliers e exclusão de parques com menos de 35 observações mensais, assegurando bases adequadas para treino e teste.
- Seleção de variáveis: foram utilizadas todas as variáveis meteorológicas disponíveis, replicando a estrutura da regressão linear múltipla para permitir comparação direta entre modelos.
- Validação cruzada temporal: aplicou-se o método *TimeSeriesSplit* com 5 *folds*, respeitando a ordem cronológica, como recomendado em séries com baixa frequência temporal (Bergmeir and Benítez, 2012).
- Métricas de desempenho: usaram-se R^2, RMSE e MBE, mantendo consistência com os modelos lineares.
- Análise de *overfitting*: foram sinalizados os casos em que a diferença entre o R^2 de treino e teste excedeu 0.15.
- Configuração de hiperparâmetros: adotaram-se valores baseados na literatura (Géron, 2022; Probst et al., 2019), sem recurso a *grid search*, privilegiando configurações simples e replicáveis.

Com esta abordagem, procurou-se garantir um rigor metodológico compatível com os standards da literatura científica recente em previsão energética com ML, mesmo em contextos com dados mensais limitados. A validação cruzada temporal representou um passo metodológico inovador nesta área de aplicação, contribuindo para colmatar a escassez de estudos que tratem especificamente deste tipo de granularidade temporal. Adicionalmente, foram produzidos gráficos comparativos entre os valores previstos e observados para cada modelo e parque, como forma de análise visual complementar da performance preditiva. Estes

gráficos permitiram verificar o alinhamento global das previsões e identificar padrões de sub ou sobrestimação.

A escolha do XGBoost e do Random Forest assentou em três fundamentos centrais:

- Capacidade de generalização amplamente documentada na literatura aplicada à energia (Meng and Song, 2020; Huang et al., 2021);
- Robustez perante séries temporais de dimensão reduzida;
- Relevância prática na engenharia preditiva e na análise de incertezas em sistemas fotovoltaicos.

Este conjunto de testes representa, por conseguinte, uma contribuição relevante para o avanço da metodologia de previsão em parques fotovoltaicos operacionais, propondo modelos alternativos ao MCP com validade estatística, consistência preditiva e potencial de adoção prática, sobretudo em contextos com limitações de dados.

4. Resultados

Os resultados obtidos com a aplicação da metodologia descrita no capítulo 3 são aqui apresentados, respeitando a estrutura dos objetivos definidos para cada etapa da investigação. Os dados foram analisados de forma integrada e discutidos à luz da literatura e das normas técnicas relevantes, com o intuito de validar a metodologia MCP, identificar limitações e propor soluções robustas. Os resultados estatísticos foram complementados por visualizações gráficas, evidência documental e análise técnica dos ativos. Cada secção corresponde a um objetivo metodológico específico, conforme explicitado previamente.

4.1 Escolha da melhor base

Antes de avançarmos com a validação formal da metodologia MCP aplicada aos 19 parques fotovoltaicos em estudo, procedeu-se à análise das sete séries de dados de radiação de longo termo disponíveis (S1 da Solargis, S2 a S5 da Era 5 e S6, S7 da PVGIS) com o intuito de identificar, para cada parque, a série que melhor representasse a variabilidade da produção elétrica observada. Ou seja, procurou-se determinar qual das séries melhor justificava as variações registadas na produção mensal, sendo assim indicativa de uma caracterização mais fidedigna do recurso solar local.

A fase inicial desta análise, centrada na avaliação do coeficiente de determinação (R^2), revelou resultados inesperados (Tabela 3). De forma surpreendente, a base PVGIS obteve o melhor R^2 em 12 dos 19 parques analisados, ao passo que a base Solargis apenas se destacou em 7 parques. A base ERA5, por sua vez, não registou qualquer caso em que apresentasse a melhor correlação. Este resultado contraria, em parte, o que é frequentemente reportado na literatura, onde a Solargis tende a apresentar, em média, melhor desempenho do ponto de vista do ajustamento estatístico às séries medidas de produção.

Tabela 3 - Valores de coeficiente de determinação (R^2) para todas as sete bases de dados de radiação.

Parque	S1 (%)	S2 (%)	S3 (%)	S4 (%)	S5 (%)	S6 (%)	S7 (%)
PV01	97.76	97.44	97.47	97.48	97.42	97.48	97.89
PV02	97.86	97.66	97.71	97.58	97.62	97.65	**98.03**
PV03	**98.19**	**98.39**	**98.14**	**98.34**	**98.25**	97.28	97.83
PV04	97.11	96.33	96.69	96.60	96.79	96.31	96.62
PV05	**98.19**	97.44	97.48	97.49	97.40	97.28	97.83
PV06	96.92	96.28	95.96	96.32	96.17	96.17	96.50
PV07	**99.11**	**98.34**	**98.25**	**98.36**	**98.20**	**98.45**	**99.00**
PV08	**98.51**	**98.23**	**98.02**	**98.24**	**98.08**	**98.39**	**98.63**
PV09	**98.82**	97.90	97.88	97.80	97.86	97.93	**98.86**
PV10	**98.64**	**98.58**	**98.54**	**98.60**	**98.59**	**98.47**	**98.77**
PV11	**98.58**	97.83	97.83	97.77	97.77	97.74	**98.54**
PV12	**99.06**	**98.24**	**98.25**	**98.34**	**98.47**	**98.12**	**99.01**
PV13	97.27	96.55	96.49	96.38	96.35	96.54	97.05
PV14	**98.86**	**98.52**	**98.42**	**98.55**	**98.53**	**98.91**	**99.18**
PV15	**98.26**	97.43	97.24	97.40	97.34	97.61	**98.45**
PV16	**98.83**	**98.27**	**98.23**	**98.33**	**98.05**	**98.50**	**98.92**
PV17	97.15	97.08	97.10	97.09		97.09	97.49
PV18	**98.19**	97.61	97.46	97.60	97.55	97.99	**98.45**
PV19	96.93	96.78	96.78	96.74	96.69	96.76	97.34

No entanto, a análise aprofundada das métricas de erro, complementada por um teste de ANOVA, revelou que o MBE foi a única métrica a apresentar diferenças estatisticamente significativas entre as três séries representativas S1 (Solargis), S4 (ERA5) e S7 (PVGIS). O teste ANOVA revelou uma estatística $F = 5.66$ e $p = 0.0058 < 0.05$, indicando que as diferenças observadas entre as bases não são atribuíveis ao acaso. Este resultado reforça a relevância crítica do MBE nesta comparação, dado o impacto que o viés sistemático pode exercer sobre as estimativas de longo prazo. Observou-se que, apesar da base PVGIS ter evidenciado melhor correlação em mais parques, apresentou, em todos os casos, os valores de MBE mais elevados. Em contrapartida, a Solargis revelou consistentemente o menor MBE, o que indicia maior fiabilidade nas previsões de longo termo, sobretudo em cenários onde a imparcialidade das estimativas assume particular importância. A Tabela 4 resume os valores médios de R^2, RMSE e MBE para cada base de dados selicionada, obtidos a partir da média dos resultados individuais dos 19 parques analisados. Os resultados completos por parque encontram-se no Apêndice A, na Tabela A.1.

Tabela 4 - Valores médios de R^2, MBE e RMSE por base de dados.

Base de dados	R^2 (%)	MBE (%)	RMSE (%)
Solargis (S1)	98,12	0,67	1,88
ERA5 (S4)	97,63	0,95	2,37
PVGIS (S7)	98,13	0,89	1,84

Do ponto de vista estatístico, estes resultados sustentam que a base Solargis é a mais adequada para aplicação da metodologia MCP no contexto específico desta investigação. A escolha do MBE como principal critério de comparação justifica-se, sobretudo, quando se considera o horizonte temporal da previsão 20 anos e a importância de garantir estimativas isentas de enviesamento. Um modelo com R^2 elevado, mas MBE significativo tenderá a sobrestimar ou subestimar sistematicamente a produção, comprometendo a validade de estimativas do tipo P50/P90, bem como a fiabilidade de análises económicas e financeiras com base nessas previsões.

Complementarmente, realizou-se uma análise exploratória da aleatoriedade dos resíduos, com recurso a diagramas de dispersão. Esta avaliação não revelou a existência de padrões sistemáticos graves que pudessem comprometer a estabilidade estatística dos modelos base-parque.

Assim, tendo em conta o conjunto de evidências reunidas, a base Solargis foi selecionada como referência para a aplicação da metodologia MCP nesta tese, em alinhamento com os critérios de robustez estatística e minimização do viés sistemático, essenciais para garantir previsões de longo prazo fiáveis e tecnicamente sustentáveis.

4.1.1 Alternativas na Escolha da Base de Dados

Apesar da base Solargis ter sido selecionada como a mais adequada para a aplicação da metodologia MCP neste trabalho, importa reconhecer que a escolha da base de dados de radiação não é universal nem imutável. Em contextos reais de projeto, a decisão pode ser influenciada por fatores operacionais, económicos e estratégicos, os quais devem ser ponderados em função dos objetivos específicos da análise energética a realizar.

Do ponto de vista técnico, a Solargis revelou um desempenho superior em termos de viés sistemático (MBE), aspeto particularmente relevante quando o objetivo é a previsão de longo

prazo com elevado grau de imparcialidade, como é o caso da presente investigação. No entanto, trata-se de uma base de acesso comercial, sujeita a custos associados à sua utilização, o que pode representar uma limitação em determinados contextos. Em contrapartida, a base PVGIS, de acesso gratuito disponibilizada por instituições públicas europeias, embora apresente um MBE mais elevado, demonstrou melhor correlação com a produção observada em grande parte dos parques analisados (Tabela A.1 no Apêndice A).

Esta distinção entre fontes pagas e gratuitas é especialmente relevante em projetos de menor escala, estudos de viabilidade preliminares, ou em geografias com restrições orçamentais e institucionais. Nesses casos, a adoção da base PVGIS pode constituir uma escolha perfeitamente legítima, desde que se reconheçam os seus limites em termos de viés sistemático e se incorpore essa incerteza nos fatores de correção ou margens de segurança das estimativas.

Do ponto de vista económico-financeiro, a escolha da base de dados deve ser compatível com os requisitos de precisão impostos por entidades financiadoras, seguradoras ou promotores. Em estimativas bancáveis, por exemplo, o controlo rigoroso do viés assume especial importância, dado o seu impacto direto na definição dos cenários de produção (P50, P90) e no cálculo de receitas esperadas ao longo da vida útil da central fotovoltaica. Nestes casos, bases como a Solargis ou outras equivalentes de elevada qualidade poderão justificar o seu custo, dado o retorno em confiança e credibilidade que aportam às previsões.

Por fim, a base ERA5, embora não tenha evidenciado o melhor desempenho em R^2 ou MBE, apresentou maior consistência nos testes estatísticos de pressupostos da regressão linear. Os resultados desta análise encontram-se descritos na Tabela 3. Esta característica sugere que, em contextos onde a validação estatística formal e a rastreabilidade científica sejam prioritárias, como em estudos académicos ou modelações estatísticas robustas, o uso da ERA5 pode também ser considerado uma alternativa válida ou complementar.

Assim, recomenda-se que a escolha da base de dados, em aplicações futuras da metodologia MCP, seja feita de forma ponderada, tendo em consideração não apenas os critérios estatísticos de desempenho, mas também os objetivos do estudo, as condicionantes económicas, e o contexto de aplicação. Esta abordagem flexível permite garantir a utilidade prática da metodologia em diferentes cenários, sem comprometer a sua validade científica.

4.2 Análise de outliers

As análises de outliers têm uma função relevante para sustentar a validação da robustez da metodologia adotada. Como instrumento de previsão, o modelo MCP requer um entendimento aprofundado do comportamento dos dados, logo, é imprescindível analisar a influência de valores atípicos, dado o seu potencial para distorcer os resultados estatísticos e operacionais. Assim, os resultados obtidos nesta análise complementar são aqui apresentados, seguindo a mesma lógica dos quatro objetivos descritos na metodologia (secção 3.4), abrangendo a avaliação da influência dos outliers nos indicadores de desempenho, a sua distribuição temporal e respetivas causas técnicas.

4.2.1 Impacto estatístico dos outliers nos modelos MCP

A análise teve como primeiro objetivo verificar em que medida a presença de outliers comprometia a robustez estatística dos modelos elaborados com base na metodologia MCP. Para isso, adotou-se uma abordagem baseada na análise visual dos gráficos de dispersão que relacionam a produção medida à radiação estimada, complementada por critérios operacionais que remove os meses cuja disponibilidade seja inferior a 85%, de acordo com as orientações estabelecidas pela norma IEC 61724-1 (2021).

A aplicação dessa filtragem possibilitou notar um aumento considerável em R^2 em 18 dos 19 parques estudados, evidenciando aprimoramentos na qualidade do ajuste dos modelos de regressão linear. Os parques PV06 e PV10 mostraram aumentos superiores a 60%, mostrando melhorias na qualidade do ajustamento dos modelos de regressão linear. Ao mesmo tempo, identificaram-se diminuições consistentes em RMSE e MBE, o que revela não apenas um aumento na precisão, mas também uma diminuição do viés sistemático dos modelos após a exclusão das observações atípicas. Os resultados estão de acordo com a regressão linear, onde valores atípicos tendem a alterar a linha de tendência, diminuindo a correlação e aumentando os erros preditivos (Chatterjee and Hadi, 2006). Na Tabela 5, são apresentados os valores de R^2, RMSE e MBE antes e após a eliminação dos outliers, possibilitando uma comparação direta do desempenho dos modelos ajustados.

Estudos de Nguyen e Müsgens (2022) indicam que a remoção de valores atípicos e a avaliação da disponibilidade operacional antes da modelação estatística têm um papel essencial na diminuição do viés e no aumento do coeficiente de determinação, mostrando que os

resultados obtidos estão alinhados com a evidência científica destacando a relevância de aplicar filtros operacionais numa fase inicial da análise, como forma de melhorar a qualidade dos dados e aumentar a precisão de modelos preditivos. Essa observação evidencia a relevância de medidas de controle de qualidade fundamentadas em critérios técnicos, mesmo na falta de metodologias estatísticas formais. Desta forma, o primeiro objetivo validou que a exclusão fundamentada de outliers ainda que se baseie apenas em critérios visuais e operacionais exerce um efeito positivo significativo na robustez estatística dos modelos MCP, criando uma base confiável para as análises que se seguirão. Além disso, a redução consistente dos valores de RMSE e MBE observada após este processo demonstra que a metodologia contribui para diminuir a incerteza associada às estimativas de longo prazo, reforçando a fiabilidade das previsões energéticas obtidas.

Tabela 5 - Indicadores estatísticos com e sem remoção de outliers.

Parque	**Com outliers (%)**			**Sem outliers (%)**		
	R^2	RMSE	MBE	R^2	RMSE	MBE
PV01	96,42	3,58	0,85	97,76	2,24	0,65
PV02	96,60	3,40	0,80	97,86	2,14	0,57
PV03	62,41	37,59	86,20	98,19	1,81	0,38
PV04	82,98	17,02	6,36	97,11	2,89	0,92
PV05	85,63	14,37	5,75	98,19	1,81	0,62
PV06	32,82	67,18	87,27	96,92	3,08	0,29
PV07	99,11	0,89	0,63	99,11	0,89	0,63
PV08	96,09	3,91	1,48	98,51	1,49	0,49
PV09	86,70	13,30	4,07	98,82	1,18	0,63
PV10	31,49	68,51	65,11	98,64	1,36	0,71
PV11	74,49	25,31	5,78	98,58	1,42	0,77
PV12	62,05	37,95	49,83	99,06	0,94	0,76
PV13	88,05	11,95	6,08	97,27	2,73	1,09
PV14	93,11	6,89	5,26	98,86	1,14	0,50
PV15	93,69	6,31	4,48	98,26	1,74	0,82
PV16	90,58	9,42	1,64	98,83	1,17	0,47
PV17	96,30	3,70	0,88	97,15	2,85	0,79
PV18	91,09	8,09	7,15	98,19	1,81	0,69
PV19	95,05	4,76	1,07	96,93	3,07	0,91

4.2.2 Classificação dos parques por sensibilidade a outliers

Na sequência da análise anterior, que, confirmou o efeito benéfico da exclusão de outliers na robustez estatística dos modelos MCP, o segundo objetivo teve a finalidade de examinar a sensibilidade de cada parque em relação à presença dessas anomalias. Com esse intuito, foi calculada a variação percentual do coeficiente de determinação (ΔR^2) antes e após a filtragem, o que possibilitou a avaliação do impacto dos dados extremos sobre os modelos.

Com base nos valores de ΔR^2, os parques foram agrupados em três categorias: Tipo A ($\Delta R^2 < 2\ \%$), Tipo B ($2\ \% \leq \Delta R^2 < 10\ \%$) e Tipo C ($\Delta R^2 \geq 10\ \%$).

Como se pode observar na Tabela 5, oito parques foram categorizados como Tipo C, evidenciando alta sensibilidade à presença de outliers, os quais resultaram em aumentos significativos no R^2 após a filtragem. Seis parques foram classificados como Tipo B, apresentando impacto moderado, enquanto cinco foram considerados estatisticamente estáveis, pertencendo ao Tipo A. Na Tabela 6 são apresentados os valores de R^2 antes e depois da remoção e também a classificação correspondente.

Os resultados indicam que, em diversas situações, os modelos MCP apresentam vulnerabilidades em relação à presença de dados anómalos, sobretudo em parques cujo histórico operacional é menos estável. Foi identificada uma associação entre a elevada sensibilidade estatística (Tipo C) e a alta incidência de eventos técnicos, como interrupções, falhas de comunicação e disfunções em componentes críticos, conforme relatado nos documentos de O&M e *due diligence*.

Por outro lado, os parques do Tipo A demonstraram um desempenho mais sólido, frequentemente vinculado a práticas de manutenção regulares e a uma infraestrutura confiável. Esta classificação demonstrou-se eficaz na segmentação do risco estatístico-operacional dos ativos fotovoltaicos, atuando como um apoio à priorização de intervenções técnicas, à realização de inspeções direcionadas e ao fortalecimento das rotinas de controle de qualidade.

Tabela 6 - Classificação de sensibilidade para cada parque.

Parque	**R^2 C/outliers (%)**	**R^2 S/outliers (%)**	**ΔR^2 (%)**	**Classificação**
PV10	31,49	98,64	67,15	Tipo C – Alta sensibilidade
PV06	32,82	96,92	64,10	Tipo C – Alta sensibilidade

PV12	62,05	99,06	37,01	Tipo C – Alta sensibilidade
PV03	62,41	98,19	35,78	Tipo C – Alta sensibilidade
PV11	74,49	98,58	24,09	Tipo C – Alta sensibilidade
PV04	82,98	97,11	14,13	Tipo C – Alta sensibilidade
PV05	85,63	98,19	12,56	Tipo C – Alta sensibilidade
PV09	86,70	98,82	12,12	Tipo C – Alta sensibilidade
PV13	88,05	97,27	9,22	Tipo B – Sensibilidade moderada
PV16	90,58	98,83	8,25	Tipo B – Sensibilidade moderada
PV18	91,09	98,19	7,10	Tipo B – Sensibilidade moderada
PV14	93,11	98,86	5,75	Tipo B – Sensibilidade moderada
PV15	93,69	98,26	4,57	Tipo B – Sensibilidade moderada
PV08	96,09	98,51	2,42	Tipo B – Sensibilidade moderada
PV19	95,05	96,93	1,88	Tipo A – Estável
PV01	96,42	97,76	1,34	Tipo A – Estável
PV02	96,60	97,86	1,26	Tipo A – Estável
PV17	96,30	97,15	0,85	Tipo A – Estável
PV07	99,11	99,11	000	Tipo A – Estável

4.2.3 Distribuição temporal e padrão sazonal dos outliers

O terceiro objetivo pretendia identificar se existia uma lógica temporal na ocorrência dos outliers detetados por meio dos métodos estatísticos IQR e Z-score, especialmente para verificar a presença de algum padrão sazonal. Este tema é de grande importância, devido ao fato de que a presença sistemática de anomalias em determinadas épocas do ano pode comprometer um dos princípios fundamentais dos modelos de MCP: a independência temporal dos resíduos.

A avaliação mensal demonstrou que os outliers não se distribuem de forma aleatória ao longo do ano. Pelo contrário, verificou-se uma acumulação expressiva no mês de março, representando aproximadamente 28% do total, com aumentos adicionais em junho e outubro (Figura 6). Este padrão sugere uma forte associação com períodos de transição sazonal nomeadamente do inverno para a primavera e do verão para o outono tradicionalmente caracterizados por maior instabilidade atmosférica e flutuações abruptas nas condições de irradiância e temperatura. Tal fenómeno foi igualmente identificado por (Okorieimoh, Norton and Conlon, 2024), os quais evidenciaram que estas transições sazonais provocam variações significativas no desempenho energético de sistemas fotovoltaicos, interferindo na estabilidade

dos modelos de previsão. De forma complementar, Mohanasundaram e Rangaswamy (2025) reforçam a importância da decomposição sazonal como ferramenta essencial para isolar picos sistemáticos de anomalias durante estes períodos, salientando a necessidade de integrar a componente temporal na análise de desempenho. Assim, os presentes resultados indicam que parte significativa dos outliers detetados pode estar associada a efeitos sazonais previsíveis, o que reforça a importância de mecanismos de correção estacional na modelação MCP.

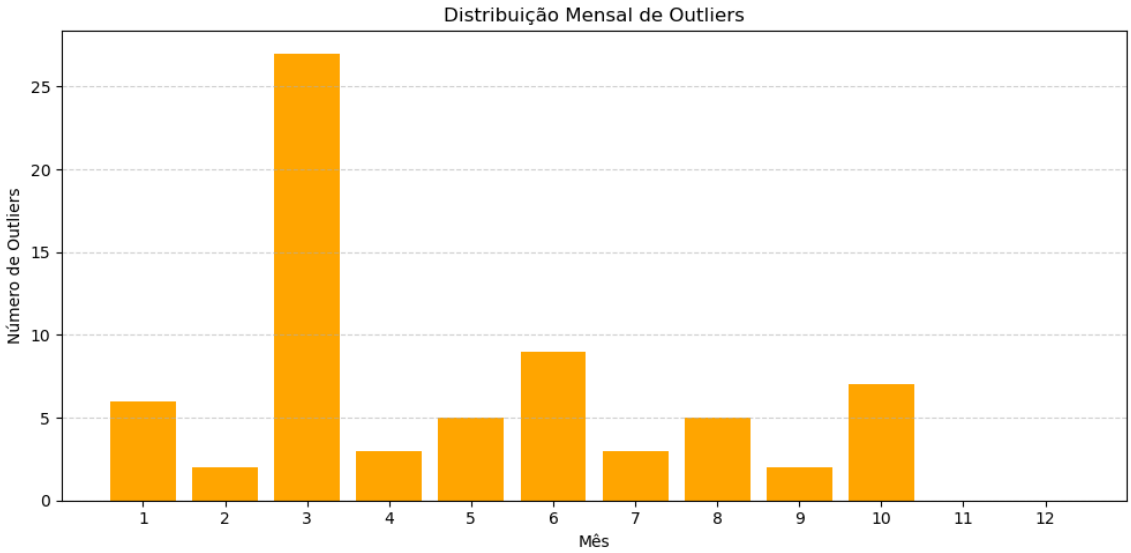


Figura 6 - Distribuição mensal de outliers.

A observação da Figura 7 permite uma identificação clara da sazonalidade dos outliers ao apresentar um mapa de calor da ocorrência de outliers por parque e por mês. Esta visualização permite distinguir a distribuição heterogénea entre os parques, revelando que os padrões identificados anteriormente não são uniformes em toda a amostra. Parques como PV03, PV06 e PV10 evidenciam concentrações sistemáticas em março, sugerindo uma maior exposição operacional ou climática. Esta informação é particularmente relevante para a posterior análise de agrupamento, onde se pretende avaliar se estes padrões são replicáveis em grupos homogéneos de parques.

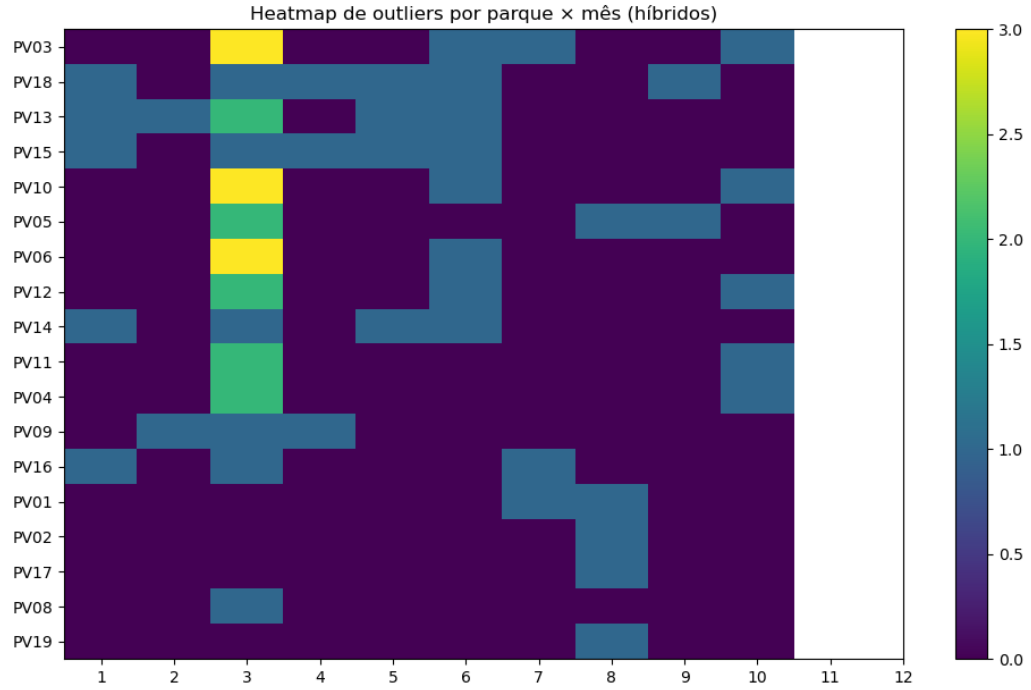


Figura 7 - Heatmap de outliers por parque e mês.

Para validar a hipótese de que esta distribuição não era aleatória, recorreu-se ao teste não paramétrico de Kruskal–Wallis, cujos resultados (H = 21.844; p = 0.0256) confirmaram diferenças estatisticamente significativas entre os meses. Esta evidência reforça a ideia de que a sazonalidade influencia diretamente a presença de valores atípicos. Torna-se, por isso, recomendável a inclusão de filtros temporais nas análises preditivas, como também defendem Mohanasundaram e Rangaswamy (2025), que demonstraram melhorias claras na previsão e deteção de anomalias quando se faz uma decomposição sazonal prévia à modelação.

Para uma melhor compreensão dos padrões identificados, utilizou-se a técnica de decomposição STL, que possibilita a separação de uma série temporal em três componentes principais como tendência, sazonalidade e ruído. Os resultados referentes aos outliers mensais evidenciou uma expressiva variação sazonal anual, apresentando picos recorrentes em março, o que corrobora os padrões previamente identificados. Em relação à tendência, observou-se uma leve redução ao longo do tempo (Apêndice B), a qual pode estar vinculada a aprimoramentos operacionais, maior maturidade dos sistemas ou a um controle mais eficiente dos processos de manutenção.

Para aprofundar a análise, fez- se o uso de técnicas de clusterização, notadamente o K-means e o agrupamento hierárquico segundo o método de Ward. Essas ferramentas possibilitaram a

classificação dos parques solares em três categorias distintas: (i) parques com reduzida incidência de outliers e distribuição dispersa ao longo do ano; (ii) parques que apresentam uma concentração sazonal significativa nos períodos mais críticos; (iii) um subconjunto de parques com padrões mais extremos, sugerindo maior exposição a fenómenos específicos. Estes agrupamentos indicam que determinados ativos apresentam maior suscetibilidade a fenómenos sazonais específicos, o que fundamenta a implementação de estratégias de monitorização adaptadas às particularidades de cada parque (Apêndice B).

Um outro resultado muito interessante é que aproximadamente 90% dos outliers identificados por meio de análise estatística corresponderam àqueles já detetados previamente por meio de inspeção visual e critérios operacionais. A razoável concordância entre diferentes abordagens reforça a confiança na deteção inicial feita pela ferramenta interna MCP da empresa, alinhando-se com os resultados de Mohanasundaram e Rangaswamy (2025), que mostram como os métodos híbridos são eficazes para identificar anomalias em sistemas fotovoltaicos. Resumindo, destaca-se a relevância de incorporar a variável temporal nos modelos de previsão fundamentados em MCP, especialmente no âmbito da energia solar. Identificar padrões sazonais não apenas aprimora a deteção de outliers, mas também contribui diretamente para a otimização da manutenção, a administração de riscos e o aprimoramento dos modelos empregados na previsão de desempenho.

4.2.4 Causas técnicas e operacionais associadas aos outliers

A análise das causas técnicas e operacionais associadas aos outliers mostrou ter uma forte correlação entre as anomalias estatisticamente identificadas e os episódios críticos documentados nos registos operacionais dos parques. A intersecção entre os outliers e os períodos de disponibilidade limitada, juntamente com a avaliação paralela dos indicadores de *Performance Ratio* (PR), corroborou que uma parcela significativa das anomalias pode ser relacionada a falhas técnicas, intervenções corretivas prolongadas ou situações de operação não representativa.

Na Figura 8 são apresentados os histogramas da disponibilidade operacional (ver à esquerda) e do PR (ver à direita), correspondentes aos meses classificados como outliers. Observa-se uma forte concentração de ocorrências com disponibilidade inferior a 85 %, o que está em conformidade com o limiar proposto pela norma IEC 61724-1:2021 para a exclusão de períodos

com funcionamento irregular ou interrupções significativas. Já no caso do PR, a distribuição revela uma assimetria acentuada, com a presença de valores extremamente baixos, indicativos de subdesempenho operacional, mas também de alguns registos anormalmente elevados, que podem estar associados a falhas de medição, erros de calibração ou inconsistências nos sistemas de monitorização. Esta dualidade reforça a complexidade dos fenómenos que originam os outliers e a importância de uma análise cruzada entre métricas técnicas e estatísticas.

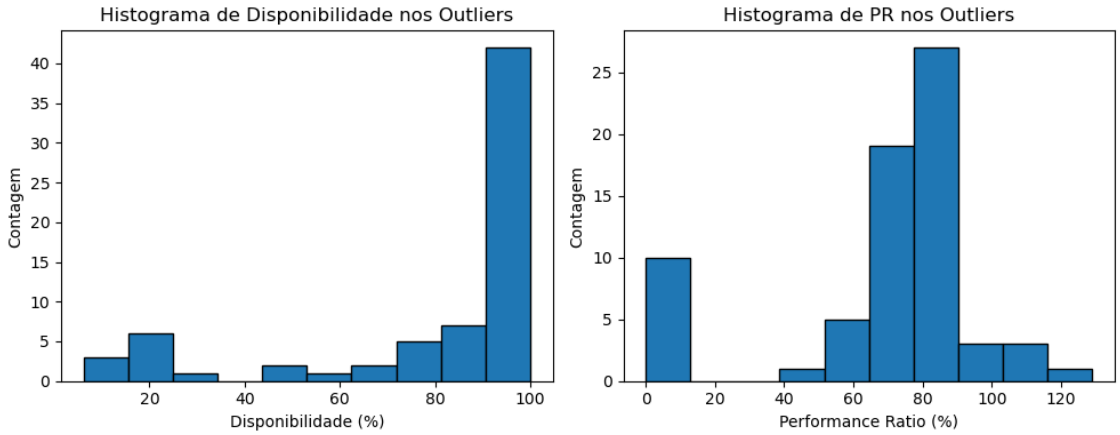


Figura 8 - Histogramas da disponibilidade e do PR nos meses classificados como outliers.

Adicionalmente, a Figura 9 mostra um gráfico de dispersão entre a disponibilidade e o PR, identificando categorias de outliers operacionais (em vermelho), por PR (em laranja) e híbridos (em cinza), evidenciando a infinidade de padrões, com diversos casos de alta disponibilidade correlacionada a PR anormalmente baixo e o inverso. A presente análise gráfica foi adicionada à revisão dos relatórios de due diligence disponibilizados para cada parque. No parque PV03, em março de 2022, a análise documental verificou a troca de um inversor central no decorrer do mês, o que coincidiu com uma redução acentuada na produção e uma disponibilidade inferior a 60%. Situações parecidas foram observadas em PV04 (outubro de 2020 e março de 2021) e PV06 (março de 2021), nas quais os relatórios técnicos apontam para interrupções prolongadas e manutenções extraordinárias, resultando em PRs distorcidos e outliers que se mostram consistentes com falhas técnicas.

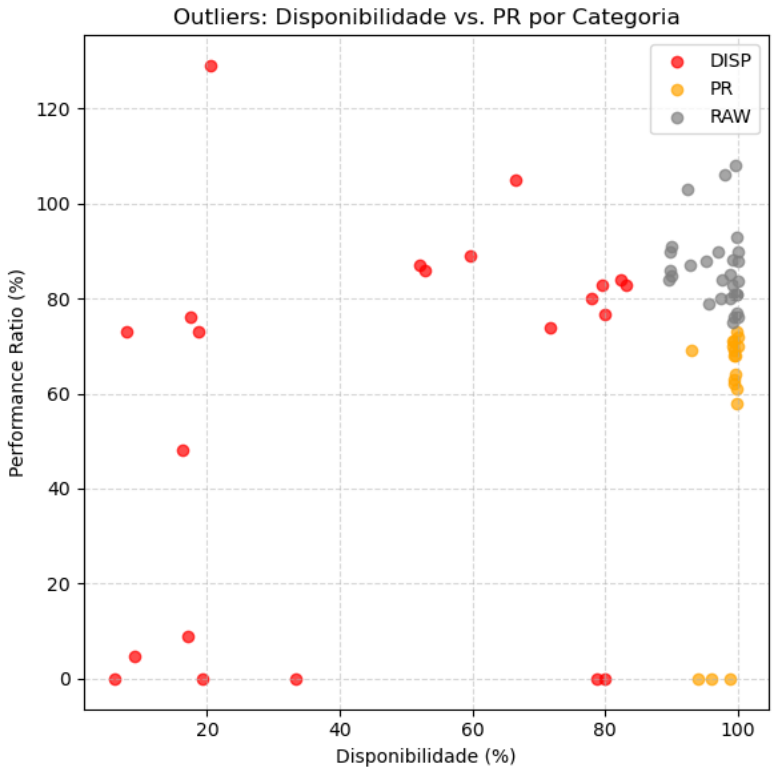


Figura 9 - Relação entre disponibilidade e PR para os outliers identificados.

Outros casos, como PV10 (março de 2021) e PV11 (outubro de 2022), demonstraram interrupções operacionais substanciais, com reflexo direto na produção, não havendo relação com condições meteorológicas. Nos dados apresentados pela Solargis, não foram observadas variações significativas na irradiância, o que indica que a origem da anomalia é exclusivamente técnica. Em contrapartida, os parques situados na área de Cluster-3 como PV13, PV18, PV14 e PV15 revelaram uma distinta categoria de outliers, relacionada a fenômenos climáticos adversos. No mês de janeiro de 2023, foram observados níveis de irradiância global horizontal que se situaram consideravelmente abaixo da média histórica, conforme atestado pela base de dados Solargis e confirmado por profissionais em campo. Essa condição climática pouco comum foi capaz de provocar diferenças significativas entre a produção projetada e a produção efetivamente registada, sendo identificada adequadamente como um outlier estatístico. Um dos padrões mais relevantes identificados foi a divergência entre a disponibilidade e o desempenho efetivo dos sistemas. No parque PV14, verificou-se um PR elevado mesmo com

baixa disponibilidade, o que sugere um funcionamento eficiente nos breves períodos ativos. Por outro lado, em PV15, observaram-se valores de PR anormalmente baixos apesar de uma disponibilidade elevada, o que aponta para possíveis falhas na medição, erros de calibração de sensores ou defeitos técnicos em strings ou inversores. Segundo a norma IEC 61724-1:2021, erros de calibração em sensores, como pirómetros e sensores de temperatura, podem distorcer o PR, mesmo quando o sistema aparenta estar operacional. Esta distinção entre métricas é também sublinhada pelo DOE (2022), que define disponibilidade como o tempo de operação e o PR como um indicador de eficiência real, sensível a falhas internas. O NREL (2024) reforça que falhas em inversores afetam o PR sem alterar significativamente a disponibilidade, dificultando a identificação de anomalias se os indicadores forem analisados isoladamente.

Esta evidência justifica a necessidade de separar os outliers de origem operacional associados a avarias, curtailment ou manutenção dos de origem climática, resultantes de variações ambientais. Tal segmentação é fundamental para orientar estratégias diferenciadas: reforço de protocolos de O&M[8] para causas técnicas e ajuste sazonal nos modelos MCP para eventos climáticos.

Em síntese, tal objetivo possibilitou a integração eficaz da dimensão técnica e estatística na análise de outliers. Ao articular as informações com a documentação técnica e os índices de desempenho, potencializa-se a habilidade interpretativa dos modelos e a aplicabilidade da metodologia MCP em cenários operacionais concretos, favorecendo uma maior confiabilidade e robustez nas previsões energéticas.

4.3 Validação da metodologia adotada

Antes de se proceder à avaliação do desempenho preditivo dos modelos utilizados, é essencial garantir que os pressupostos subjacentes à regressão linear são adequadamente verificados. Esta verificação é essencial para garantir a validade estatística da metodologia adotada. Assim, foram realizados uma série de testes estatísticos, incluindo teste de ANOVA, testes de normalidade, independência dos resíduos e homocedasticidade, e aplicados a todos os parques fotovoltaicos considerados. Os resultados completos estão organizados a partir da Tabela C.1

[8] O termo *Operação e Manutenção (O&M)* designa o conjunto de atividades destinadas a garantir o funcionamento eficiente, seguro e contínuo de um parque fotovoltaico, incluindo inspeções, reparações, limpezas e monitorização de desempenho.

do Apêndice C. É importante referir que esta análise foi conduzida apenas para as três bases de radiação que apresentaram o melhor desempenho global (S1, S4 e S7).

A ANOVA permitiu verificar a significância estatística da regressão linear para cada uma das três bases. Estas evidências sugerem que a variável independente (GHI) tem um poder explicativo relevante quanto à variabilidade da produção elétrica medida nos parques fotovoltaicos em funcionamento. Na prática, os valores de p obtidos foram consistentemente inferiores ao limite de significância adotado ($\alpha = 0,05$), o que confirma a adequação do modelo sem reconhecer a sua significância global.

Relativamente à normalidade dos resíduos, os resultados do teste de Shapiro-Wilk mostraram que, na maioria dos casos, se adequaram a esta hipótese, revelando uma distribuição de resíduos aproximadamente normal para a generalidade dos parques e bases. As exceções foram registadas nos parques PV07 e PV16 (base S1), PV11 (base S4) e PV11 e PV06 (base S7). Relativamente à independência dos erros, aferida através do teste de Durbin–Watson, observou-se a presença de autocorrelação em nove dos 19 parques para a base S1, e em oito parques tanto para as bases S4 como S7. Este resultado sugere a existência de padrões temporais residuais que não são completamente captados pelo modelo linear, indicando uma limitação potencial da abordagem nos casos afetados.

No que se refere à homocedasticidade, avaliada pelo teste de Levene, os resultados foram, em geral, satisfatórios. Este pressuposto foi cumprido por todos os parques analisados com base S4, enquanto que foram encontradas algumas exceções noutras bases: os parques PV01 e PV05 com base S7 e PV08 com base S1.

Apesar de se terem verificado algumas violações pontuais dos pressupostos estatísticos nomeadamente no que respeita à independência dos resíduos, é importante referir que, segundo Wilks (2011), a presença de autocorrelação é um fenómeno comum em séries temporais de natureza energética e climática. Tal limitação não invalida a abordagem, mas reforça a necessidade de recorrer, em fases posteriores da análise, a modelos de previsão mais sofisticados, capazes de captar estruturas temporais mais complexas. Neste sentido, a secção seguinte explora alternativas metodológicas com o objetivo de melhorar a qualidade dos ajustamentos e aumentar a robustez estatística da estimativa de produção energética em parques fotovoltaicos.

4.4 Modelos alternativos ao MCP

A análise comparativa entre a metodologia tradicional Measure–Correlate–Predict, baseada em regressão linear simples, e modelos alternativos como regressão linear múltipla e algoritmos de machine learning (XGBoost e Random Forest) permitiu aferir o desempenho relativo de cada abordagem na estimativa da produção mensal de energia em parques fotovoltaicos em operação.

A motivação para explorar modelos alternativos surgiu da constatação, anteriormente documentada, de que em alguns parques não se verificavam integralmente os pressupostos da regressão linear (em especial a independência dos resíduos), apesar do elevado R^2 global superior a 0,97. A regressão linear múltipla, embora tenha revelado ligeiras melhorias nas métricas estatísticas globais, apresentou limitações metodológicas semelhantes à regressão linear simples, sobretudo pela ausência de validação cruzada e a dependência de pressupostos estatísticos. Assim, apesar do seu potencial para incorporar variáveis meteorológicas adicionais, o modelo não garantiu ganhos substanciais de robustez estatística, mostrando-se insuficiente para colmatar as limitações detetadas nos pressupostos da regressão tradicional.

Neste contexto, optou-se por explorar abordagens não paramétricas baseadas em *Machine Learning*: Random Forest e XGBoost. Algoritmos reconhecidos pela sua eficácia em dados tabulares e capacidade de modelar relações não lineares complexas (Breiman, 2001; Chen and Guestrin, 2016).

Ao contrário dos modelos lineares, os algoritmos de machine learning foram treinados com validação cruzada temporal (TimeSeriesSplit com cinco divisões). Esta técnica é mais adequada a séries cronológicas, já que respeita a ordem temporal e evita fugas de informação, ao contrário do k-fold tradicional que mistura pontos de treino e teste (Hyndman and Athanasopoulos, 2018). Os dados foram previamente sujeitos a pré-processamento rigoroso, incluindo remoção de valores ausentes, exclusão de outliers (identificados na etapa anterior) e eliminação de parques com menos de 35 observações garantindo qualidade mínima das séries temporais.

Mesmo com esta limitação de dados, os modelos XGBoost e Random Forest apresentaram desempenho robusto, com R^2 de teste médio superior a 0.89 e erros normalizados (RMSE e MBE) geralmente dentro dos limites aceites. Em três parques observou-se sobreajuste (diferença R^2 treino–teste > 0.15), não totalmente mitigado mesmo após tentativas de

regularização. No entanto, em 11 parques, os modelos mantiveram estabilidade entre treino e teste, demonstrando boa capacidade de generalização.

Este resultado ganha ainda maior significado ao considerar que as normas técnicas e científicas geralmente aceitam erros relativos até 10–15% como aceitáveis para estudos de previsão de produção (IEA-PVPS, 2020; IEC 61724-1:2021). O modelo XGBoost manteve, em média, RMSE inferior a 10% em mais de metade dos parques e MBE dentro de ±10%, posicionando-se claramente dentro desses intervalos.

A Tabela 7 mostra a superioridade do XGBoost face ao modelo de machine learning random forest (ver o apêndice D para resultados completos), e aproxima-se dos resultados obtidos com a regressão linear, apesar de não beneficiar da ausência de validação cruzada, como os modelos lineares. Esta diferença metodológica impede comparações diretas absolutas. No entanto, é precisamente a aplicação de uma abordagem mais exigente aos modelos ML que torna os seus resultados ainda mais significativos.

Tabela 7 - Comparação das médias resumida dos modelos.

Modelos	R^2	RMSE (%)	MBE (%)
XGBoost	89,4	10,1	7,6
Random Forest	76,3	23,715	16,45
Regressão Linear Simples	98,1	0,019	0,007
Regressão Linear Múltipla	98,3	0,017	0,005

A regressão linear, embora mais ajustada à amostra, beneficia da ausência de separação entre treino e teste, o que pode levar a uma sobrestimação do desempenho. Já os modelos de ML, treinados com validação temporal realista, aproximam-se do desempenho dos modelos lineares mesmo sob este cenário exigente, evidenciando o seu potencial prático.

Para além das métricas, foram gerados gráficos de comparação entre a produção mensal medida e as previsões dos diferentes modelos. Nestes, o modelo XGBoost destacou-se pela capacidade de acompanhar os padrões sazonais e captar flutuações extremas, especialmente em parques como PV17 e PV13 (representados nas Figuras 11 e 12), onde a regressão linear falhou em acompanhar adequadamente os picos de produção.

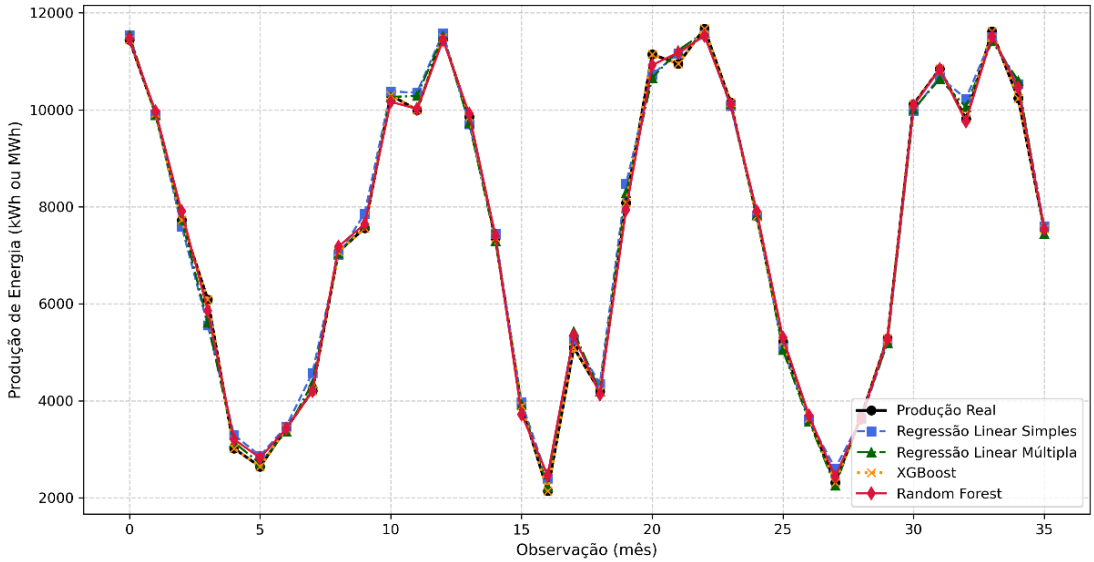


Figura 10- Comparação produção real vs previsões (PV13).

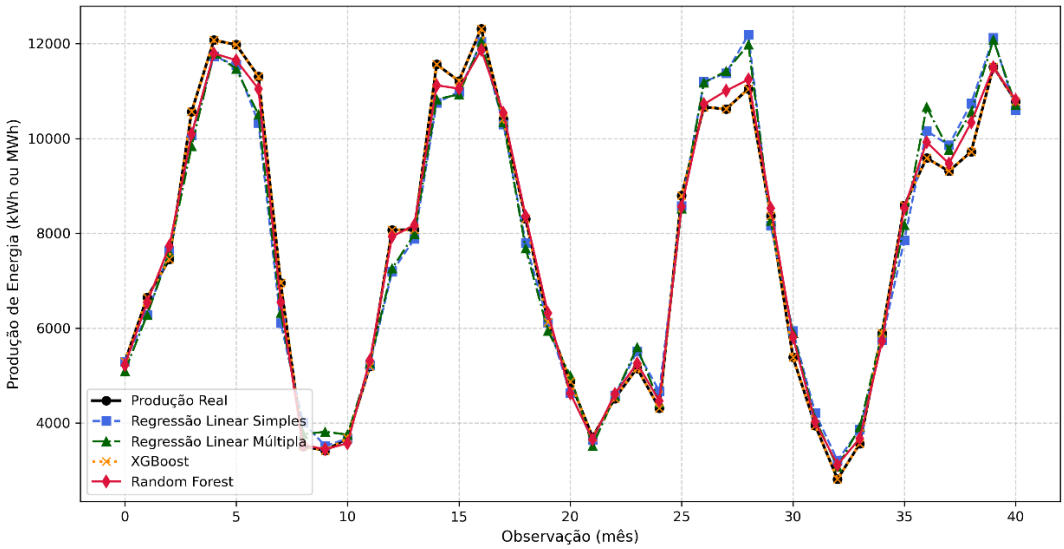


Figura 11 - Comparação produção real vs previsões (PV17).

Nos parques com comportamento mais regular como PV07 e PV08 (Figuras 13 e 14), a regressão linear manteve boa capacidade de previsão, mas o XGBoost mostrou maior adaptabilidade a variações subtis. Estes resultados reforçam a utilidade do XGBoost como modelo preditivo robusto mesmo em séries mensais com ruído e variabilidade residual, ou seja, este comportamento sugere maior adaptabilidade do modelo às variações reais do recurso solar.

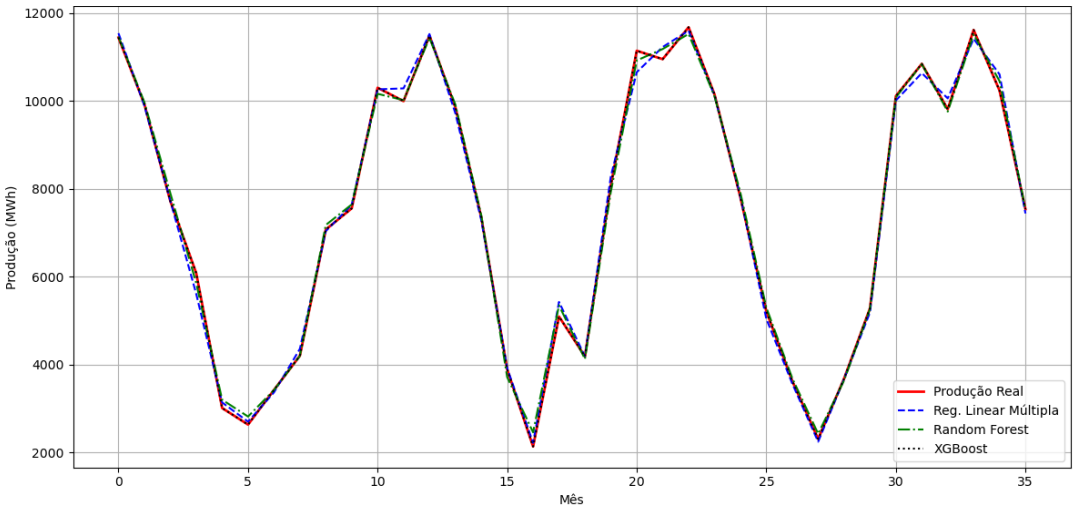


Figura 12 - Comparação produção real vs previsões (PV07).

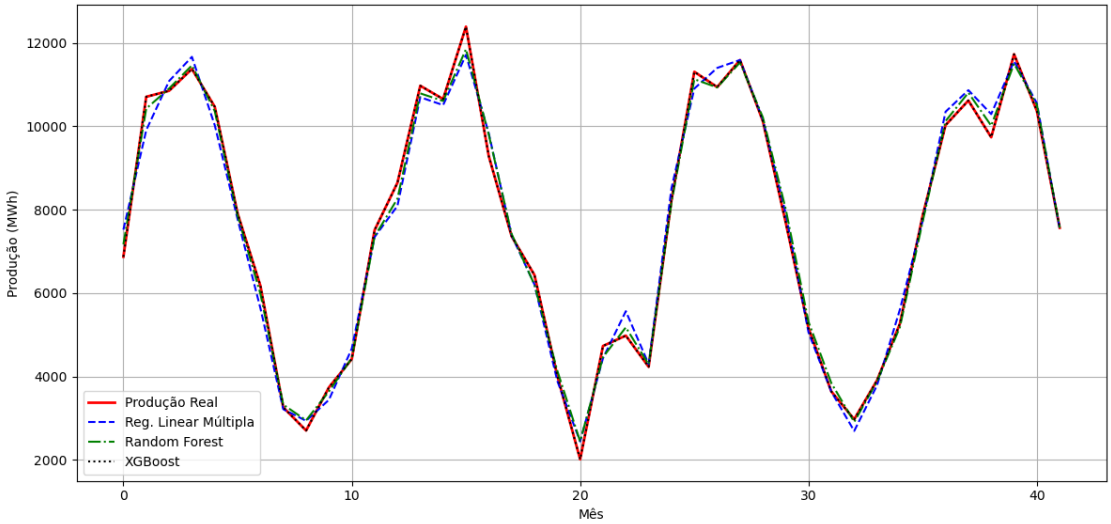


Figura 13 - Comparação produção real vs previsões (PV08).

Apesar dos bons resultados, é importante reconhecer os trade-offs[9]. Algoritmos de ML requerem maior volume de dados e estão mais expostos ao risco de sobreajuste. A regressão [10] linear, mesmo com pressupostos violados, mantém a vantagem da transparência e replicabilidade em ambiente industrial, fatores valorizados por operadores e investidores. Esta discussão reforça que a proposta não pretende substituir o MCP, mas sim avaliar alternativas viáveis em contextos específicos.

A principal contribuição desta análise reside na demonstração empírica de que é possível aplicar com sucesso algoritmos de ML, nomeadamente o XGBoost, a séries mensais de produção fotovoltaica, preenchendo assim uma importante lacuna na literatura científica. Ao contrário da maioria dos estudos existentes, focados em granularidades horárias ou diárias, este trabalho mostra que, mesmo com dados limitados, é possível atingir desempenhos estatisticamente robustos, desde que aplicadas técnicas adequadas de validação e controlo de sobreajuste.

Assim, o XGBoost não só se mostra como uma alternativa válida à metodologia MCP em contextos operacionais reais, como também pode ser considerado um modelo preditivo independente, com aplicações potenciais em estudos de previsão P50/P90, análise de perdas e planeamento energético de longo prazo em sistemas já instalados.

[9] Expressão *trade-offs* refere-se à necessidade de equilibrar vantagens e desvantagens, ou ganhos e perdas, em decisões técnicas, económicas ou metodológicas.

5. Conclusão

O objetivo primário desta investigação foi validar uma aplicação da metodologia MCP para estimativas de previsão de energia elétrica anual em parques fotovoltaicos após a construção.

Para esse efeito, foram utilizados dados mensais de 19 parques fotovoltaicos localizados em Espanha, bem como dados de radiação de longo prazo provenientes de várias bases de dados, nomeadamente Solargis, ERA5 e PVGIS. A metodologia foi primeiro adaptada ao contexto fotovoltaico, testada e submetida a diversas análises estatísticas rigorosas incluindo deteção de valores atípicos, tratamento e aplicação de modelos preditivos alternativos. O intuito, acima de tudo, foi avaliar a robustez estatística, a praticabilidade e a preditividade do MCP face aos cenários práticos reais.

Com base nos resultados obtidos, foram tiradas as seguintes conclusões:

Desempenho da regressão linear no MCP

A aplicação da regressão linear simples na base da metodologia MCP demonstrou um elevado desempenho estatístico na maioria dos parques, com um valor de R^2 superior a 97%. A metodologia foi validada globalmente, apesar de terem sido encontradas algumas limitações no incumprimento de certos pressupostos da regressão linear devido à presença de ligeiras autocorrelações em alguns parques.

Influência da base de radiação na qualidade do modelo

A escolha da base de radiação revelou-se como um fator importante para o desempenho do modelo. O PVGIS, no sentido geral, mostrou maior capacidade de estudar e caracterizar as variações dos parques, com a melhor correlação entre GHI e produção elétrica em 12 dos 19 parques, tendo o melhor R^2. Por outro lado, a Solargis, embora tenha tido um desempenho ligeiramente inferior ao da base de dados PVGIS na caracterização da variabilidade dos parques, foi a base escolhida, dado um viés sistemático (MBE) mais reduzido, sendo que a diferença no R^2 médio foi apenas de 0,01%.

Impacto da análise de valores atípicos

A análise e remoção de valores atípicos mostrou ser uma etapa fundamental para a melhoria da qualidade dos modelos. A utilização de critérios estatísticos, operacionais e sazonais permitiu eliminar dados anómalos que estavam a comprometer os ajustes. Em particular, verificaram-se melhorias substanciais nas métricas de regressão em parques com instabilidade operacional. Estes resultados sublinham a importância de integrar rotinas sistemáticas de controlo de qualidade nas análises de previsão energética.

Desempenho dos modelos alternativos à MCP

Na comparação com modelos preditivos alternativos, o algoritmo de machine learning XGBoost demonstrou um desempenho competitivo face à regressão linear, mesmo quando sujeito a validação cruzada temporal (TimeSeriesSplit), com dados mensais limitados. O modelo atingiu um R^2 médio de 89% com valores de RMSE e MBE dentro dos limites considerados aceitáveis pelas normas técnicas. Os resultados confirmaram a sua elevada capacidade de generalização e adaptabilidade a variações sazonais e extremas, posicionando-o como uma alternativa viável ao MCP para contextos reais.

Contributo científico do estudo

Esta investigação representa uma contribuição relevante e original para o setor solar, ao demonstrar que a metodologia MCP, originalmente desenvolvida para o setor eólico, pode ser adaptada com sucesso ao contexto fotovoltaico. A integração de validações estatísticas formais, análise de valores atípicos e aplicação de modelos de machine learning a dados mensais reais permitiu colmatar lacunas existentes na literatura e fornecer ferramentas práticas para auditorias técnicas, avaliação de riscos e previsão de produção energética em projetos pós-construção.

5.1 Recomendações e Trabalhos Futuros

A investigação realizada permitiu validar, com elevado grau de confiança, a aplicabilidade da metodologia MCP para previsão de produção anual em parques fotovoltaicos operacionais. Ainda assim, os resultados obtidos e as limitações identificadas apontam para várias oportunidades de aprofundamento e melhoria futura, quer em termos metodológicos, quer na aplicação prática do modelo em contextos reais. Com base nisso, apresentam-se as seguintes recomendações:

- **Aprofundar a análise da autocorrelação residual**

  Apesar da robustez global dos modelos de regressão linear simples, a presença de autocorrelação em alguns parques sugere a possibilidade de incluir mecanismos de correção ou adaptação temporal. Recomenda-se explorar modelos autorregressivos ou técnicas de regressão com estrutura de erros autocorrelacionados, que possam melhorar a capacidade explicativa em séries com dependência temporal.

- **Avaliar o desempenho em outras localizações geográficas**

  Os resultados apresentados referem-se exclusivamente a parques localizados em Espanha. A replicação desta abordagem noutras regiões com características climáticas e operacionais distintas permitiria validar a generalização do modelo, testando a sua robustez em contextos diversos.

- **Explorar abordagens híbridas e não paramétricas**

  O bom desempenho do XGBoost abre espaço para o desenvolvimento de modelos híbridos que combinem a interpretabilidade dos modelos lineares com a flexibilidade dos algoritmos de machine learning. Adicionalmente, a aplicação de redes neurais recorrentes (RNN) ou modelos baseados em *long short-term memory* (LSTM) poderia ser testada em séries mensais mais longas, desde que a quantidade de dados o permita.

- **Automatizar processos de deteção e classificação de outliers**

Tendo em conta a diminuição das incertezas nas previsões seria benéfico desenvolver ferramentas automatizadas para deteção, classificação e interpretação de outliers, incorporando critérios sazonais, operacionais e estatísticos, com integração direta em plataformas de monitorização de performance.

Bibliografia

Alcañiz, A. et al. (2023) 'Trends and gaps in photovoltaic power forecasting with machine learning', *Energy Reports*, 9, pp. 447–471. Available at: https://doi.org/10.1016/j.egyr.2022.11.208.

Asiedu, S.T. et al. (2024) 'Machine learning forecasting of solar PV production using single and hybrid models over different time horizons', *Heliyon*, 10(7), e28898. Available at: https://doi.org/10.1016/j.heliyon.2024.e28898.

Bergmeir, C. and Benítez, J.M. (2012) 'On the use of cross-validation for time series predictor evaluation', *Information Sciences*, 191, pp. 192–213. Available at: https://doi.org/10.1016/j.ins.2011.12.028.

Breiman, L. (2001) 'Random forests', *Machine Learning*, 45(1), pp. 5–32. Available at: https://doi.org/10.1023/A:1010933404324.

Carta, J.A., Velázquez, S. and Cabrera, P. (2013) 'A review of measure-correlate-predict (MCP) methods used to estimate long-term wind characteristics at a target site', *Renewable and Sustainable Energy Reviews*, 27, pp. 362–400. Available at: https://doi.org/10.1016/j.rser.2013.07.004.

Chatterjee, S. and Hadi, A.S. (2006) *Regression analysis by example*. 4th edn. Hoboken, NJ: John Wiley & Sons. Available at: https://doi.org/10.1002/0470055464.

Chen, T. and Guestrin, C. (2016) 'XGBoost: A scalable tree boosting system', *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. Available at: https://doi.org/10.1145/2939672.2939785.

Cleveland, R.B., Cleveland, W.S., McRae, J.E. and Terpenning, I. (1990) 'STL: A seasonal-trend decomposition procedure based on Loess', *Journal of Official Statistics*, 6(1), pp. 3–73. Available at: https://www.math.unm.edu/~lil/Stat581/STL.pdf.

Deceglie, M.G., Anderson, K., Fregosi, D., Hobbs, W.B., Mikofski, M.A., Theristis, M., Meyers, B.E. et al. (2023) 'Perspective: Performance loss rate in photovoltaic systems', *Solar RRL*, 7(10),

2300196. Available at: https://doi.org/10.1002/solr.202300196.

Department of Statistics and Data Science – Yale University (1998) *ANOVA for regression*. Available at: http://www.stat.yale.edu/Courses/1997-98/101/anovareg.htm (Accessed: 5 July 2025).

Duffie, J.A. and Beckman, W.A. (2013) *Solar engineering of thermal processes*. 4th edn. Hoboken, NJ: Wiley. Available at: https://doi.org/10.1002/9781118671603.

Eletrónica PT (n.d.) *Efeito fotovoltaico*. Available at: https://www.electronica-pt.com/efeito-fotovoltaico (Accessed: 6 June 2025).

Friedman, J.H. (2001) 'Greedy function approximation: A gradient boosting machine', *Annals of Statistics*, 29(5), pp. 1189–1232. Available at: https://doi.org/10.1214/aos/1013203451.

Gastwirth, J.L. and Gel, Y.R. (2009) 'The impact of Levene's test of equality of variances on statistical theory and practice', *Statistical Science*, 24(3), pp. 343–360. Available at: https://doi.org/10.1214/09-STS301.

Gören, D. (2021) *Long-term energy yield estimation of a solar photovoltaic power plant in METU NCC*. MSc thesis. Middle East Technical University. Available at: https://hdl.handle.net/11511/95214.

Huang, L., Kang, J., Wan, M., Fang, L., Zhang, C. and Zeng, Z. (2021) 'Solar radiation prediction using different machine learning algorithms and implications for extreme climate events', *Frontiers in Earth Science*, 9, 596860. Available at: https://doi.org/10.3389/feart.2021.596860.

Hyndman, R.J. and Athanasopoulos, G. (2018) *Forecasting: Principles and practice*. 2nd edn. OTexts. Available at: https://otexts.org/fpp2/.

International Energy Agency (IEA) (2024) *Solar PV – Renewables 2025*. Paris: International Energy Agency. Available at: https://www.iea.org/energy-system/renewables/solar-pv (Accessed: 6 August 2025).

International Renewable Energy Agency (IRENA) (2024) *Highest annual growth of renewables jobs in 2023, reaching 16.2 million* [Press release, 1 October]. Abu Dhabi: IRENA. Available at: https://www.irena.org/News/pressreleases/2024/Oct/Highest-Annual-Growth-of-

Renewables-Jobs-in-2023-Reaching-16-point-2-Million-PT (Accessed: 6 August 2025).

IEA Photovoltaic Power Systems Programme (IEA-PVPS) (2020) *Task 16 – Technical guidelines for PV long-term yield prediction*. Paris: International Energy Agency Photovoltaic Power Systems Programme. Available at: https://iea-pvps.org/research-tasks/technical-guidelines-for-pv-long-term-yield-prediction/ (Accessed: 6 August 2025).

IEA PVPS (2022) *Guidelines for operation and maintenance of photovoltaic power plants in different climates*. IEA-PVPS Report T13-25:2022. Paris: International Energy Agency Photovoltaic Power Systems Programme. Available at: https://iea-pvps.org/wp-content/uploads/2022/11/IEA-PVPS-Report-T13-25-2022-OandM-Guidelines.pdf.

IEA PVPS (2023) *Assessment of photovoltaic system performance in real conditions*. IEA PVPS Task 13 report. Paris: International Energy Agency Photovoltaic Power Systems Programme. Available at: https://iea-pvps.org/.

IEC (2021) *IEC 61724-1:2021 – Photovoltaic system performance – Part 1: Monitoring*. Geneva: International Electrotechnical Commission. Available at: https://cdn.standards.iteh.ai/samples/103495/100a1ea212914dd38ba72a9138af4ce1/IEC-61724-1-2021.pdf.

IEC (2022) *prEN IEC 61400-15-1:2022 – Wind energy generation systems – Part 15-1: Site suitability input conditions*. Brussels: European Committee for Electrotechnical Standardization (CENELEC). Available at: https://standards.iteh.ai/catalog/standards/clc/de31ee96-f284-434a-b678-2c224fdb0a81/pren-iec-61400-15-1-2022.

Jadidi, H. et al. (2020) 'Bayesian updating of solar resource data for risk mitigation in project finance', *Solar Energy*, 207, pp. 1390–1403. Available at: https://doi.org/10.1016/j.solener.2020.07.096.

Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R. and Wu, A.Y. (2002) 'An efficient k-means clustering algorithm: Analysis and implementation', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7), pp. 881–892. Available at: https://doi.org/10.1109/TPAMI.2002.1017616.

Kenton, W. (2023) 'Durbin Watson test: What it is in statistics, with examples', *Investopedia*.

Available at: https://www.investopedia.com/terms/d/durbin-watson-statistic.asp (Accessed: 5 July 2025).

Kruskal, W.H. and Wallis, W.A. (1952) 'Use of ranks in one-criterion variance analysis', *Journal of the American Statistical Association*, 47(260), pp. 583–621. Available at: https://doi.org/10.2307/2280779.

Lindvall, J. (2016) *Post-construction production assessment of wind farms: Assessment and optimization of the energy production of operational wind farms. Part 1*. Stockholm: Energiforsk. Available at: https://energiforsk.se/en.

Markovics, D. and Mayer, M.J. (2022) 'Comparison of machine learning methods for photovoltaic power forecasting based on numerical weather prediction', *Renewable and Sustainable Energy Reviews*, 161, 112364. Available at: https://doi.org/10.1016/j.rser.2022.112364.

MEASNET (2016) *Procedure: Evaluation of site-specific wind conditions. Version 2.0 (April 2016)*. MEASNET. Available at: https://www.measnet.com/wp-content/uploads/2016/05/Measnet_SiteAssessment_V2.0.pdf.

Meng, M. and Song, C. (2020) 'Daily photovoltaic power generation forecasting model based on random forest algorithm for North China in winter', *Sustainability*, 12(6), 2247. Available at: https://doi.org/10.3390/su12062247.

Mohanasundaram, V. and Rangaswamy, B. (2025) 'Photovoltaic solar energy prediction using the seasonal trend decomposition layer and ASOA optimized LSTM neural network model', *Scientific Reports*, 15, 4032. Available at: https://doi.org/10.1038/s41598-025-87625-0.

Montgomery, D.C. and Runger, G.C. (2014) *Applied statistics and probability for engineers*. 6th edn. Hoboken, NJ: Wiley. Available at: https://www.wiley.com/en-us/Applied+Statistics+and+Probability+for+Engineers%2C+6th+Edition-p-9781118802267.

Mubari, L. and Ramahana, T. (2024) 'Comparison of operational energy yield assessments to pre-construction energy yield assessments for solar photovoltaic plants', in *2024 IEEE PES/IAS PowerAfrica*. Johannesburg, South Africa: IEEE, pp. 1–3. Available at: https://doi.org/10.1109/PowerAfrica61624.2024.10759322.

Narvaez, G. et al. (2021) 'Machine learning for site-adaptation and solar radiation forecasting', *Renewable Energy*, 167, pp. 333–342. Available at: https://doi.org/10.1016/j.renene.2020.11.089.

Nguyen, T.N. and Müsgens, F. (2022) 'What drives the accuracy of PV output forecasts?', *Applied Energy*, 323, 119603. Available at: https://doi.org/10.1016/j.apenergy.2022.119603.

Okorieimoh, C.C., Norton, B. and Conlon, M. (2024) 'Disaggregating longer-term trends from seasonal variations in measured PV system performance', *Electricity*, 5(1), pp. 1–23. Available at: https://doi.org/10.3390/electricity5010001.

Polo, J. et al. (2016) 'Preliminary survey on site-adaptation techniques for satellite-derived and reanalysis solar radiation datasets', *Solar Energy*, 132, pp. 25–37. Available at: https://doi.org/10.1016/j.solener.2016.03.001.

Probst, P., Wright, M.N. and Boulesteix, A.L. (2019) 'Hyperparameters and tuning strategies for random forest', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3), e1301. Available at: https://doi.org/10.1002/widm.1301.

Rbiedermann (n.d.) *Map of Spain vector image [Vector ID 5241147]*. VectorStock. Available at: https://www.vectorstock.com/royalty-free-vector/map-of-spain-vector-5241147 (Accessed: 6 June 2025).

Reich, N. et al. (2015) 'On-site performance verification to reduce yield prediction uncertainties', in *2015 IEEE 42nd Photovoltaic Specialist Conference (PVSC)*. New Orleans, LA: IEEE, pp. 1–6. Available at: https://doi.org/10.1109/PVSC.2015.7355614.

Renné, D.S. (2016) 'Resource assessment and site selection for solar heating and cooling systems', in *Advances in solar heating and cooling*. Elsevier, pp. 13–41. Available at: https://doi.org/10.1016/B978-0-08-100301-5.00002-3.

Rystad Energy (2023) *Iberia: A new European energy powerhouse emerges*. Oslo: Rystad Energy. Available at: https://www.rystadenergy.com/news/iberia-a-new-european-energy-powerhouse-emerges (Accessed: 6 August 2025).

Shapiro, S.S. and Wilk, M.B. (1965) 'An analysis of variance test for normality (complete samples)', *Biometrika*, 52(3/4), pp. 591–611. Available at: https://doi.org/10.2307/2333709.

Swami, G., Sheth, K. and Patel, D. (2024) ‘PV capacity evaluation using ASTM E2848: Techniques for accuracy and reliability in bifacial systems’, *Smart Grid and Renewable Energy*, 15(9), pp. 201–216. Available at: https://doi.org/10.4236/sgre.2024.159012.

Tavares, A.M. et al. (2024) ‘Effect of solar irradiation inter-annual variability on PV and CSP power plants production capacity: Portugal case-study’, *Energies*, 17(21), 5490. Available at: https://doi.org/10.3390/en17215490.

Tukey, J.W. (1977) *Exploratory data analysis*. Reading, MA: Addison-Wesley. Available at: https://doi.org/10.1002/bimj.4710230408.

U.S. Department of Energy (DOE) (2022) *Understanding solar photovoltaic system performance*. Washington, DC: U.S. Department of Energy. Available at: https://www.energy.gov/eere/solar.

Ward, J.H. (1963) ‘Hierarchical grouping to optimize an objective function’, *Journal of the American Statistical Association*, 58(301), pp. 236–244. Available at: https://doi.org/10.1080/01621459.1963.10500845.

Wild, M. (2012) ‘Enlightening global dimming and brightening’, *Bulletin of the American Meteorological Society*, 93(1), pp. 27–37. Available at: https://doi.org/10.1175/BAMS-D-11-00074.1.

Wilks, D.S. (2006) *Statistical methods in the atmospheric sciences*. 2nd edn. Amsterdam: Academic Press (Elsevier). Available at: https://sunandclimate.wordpress.com/wp-content/uploads/2009/05/statistical-methods-in-the-atmospheric-sciences-0127519661.pdf.

Apêndice A – Escolha da melhor base

Tabela A.1- Resultados do desempenho das métricas estatísticas (RMSE e MBE) para todas as bases de dados de radiação.

Parque	S1		S2		S3		S4		S5		S6		S7	
	RMSE (%)	MBE (%)	RMSE (%)	MBE (%)	RMSE (%)	MBE (%)	RMSE (%)	MBE (%)	RMSE (%)	MBE (%)	RMSE (%)	MBE (%)	RMSE (%)	MBE (%)
PV01	2,24	0,65	2,56	0,54	2,53	0,51	2,52	0,53	2,58	0,52	2,52	0,66	2,11	0,72
PV02	2,14	0,57	2,34	0,46	2,29	0,43	2,42	0,48	2,38	0,45	2,35	0,62	1,97	0,65
PV03	1,81	0,38	1,61	0,44	1,86	0,47	1,66	0,41	1,75	0,41	1,43	0,61	1,48	0,50
PV04	2,89	0,92	3,67	1,44	3,31	1,29	3,40	1,36	3,21	1,27	3,69	1,59	3,38	1,28
PV05	1,81	0,62	2,56	1,14	2,52	1,02	2,51	1,08	2,60	1,03	2,72	1,39	2,17	0,94
PV06	3,08	0,29	3,72	0,48	4,04	0,50	3,68	0,42	3,83	0,41	3,83	0,72	3,50	0,55
PV07	0,89	0,63	1,66	1,05	1,75	1,02	1,64	1,01	1,80	1,15	1,55	1,28	1,00	0,86
PV08	1,49	0,49	1,77	0,84	1,98	0,82	1,76	0,81	1,92	0,93	1,61	1,04	1,37	0,68
PV09	1,18	0,63	2,10	1,25	2,12	1,22	2,20	1,15	2,14	1,15	2,07	1,53	1,14	0,92
PV10	1,36	0,71	1,42	1,18	1,46	1,19	1,40	1,07	1,41	1,09	1,53	1,64	1,23	1,11
PV11	1,42	0,77	2,17	1,43	2,17	1,39	2,23	1,32	2,23	1,32	2,26	1,82	1,46	1,15
PV12	0,94	0,76	1,76	1,33	1,75	1,31	1,66	1,23	1,53	1,19	1,88	1,68	0,99	1,15
PV13	2,73	1,09	3,45	1,63	3,51	1,70	3,62	1,62	3,65	1,65	3,46	1,80	2,95	1,39
PV14	1,14	0,50	1,48	1,03	1,58	1,12	1,45	0,95	1,47	0,99	1,09	1,16	0,83	0,73
PV15	1,74	0,82	2,57	1,44	2,76	1,56	2,60	1,36	2,66	1,42	2,39	1,59	1,55	1,03
PV16	1,17	0,47	1,73	0,81	1,77	0,78	1,67	0,77	1,95	0,93	1,50	0,98	1,08	0,64
PV17	2,85	079	2,92	0,65	2,90	0,63	2,91	0,65			2,91	0,78	2,51	0,82

PV18	1,81	0,69	2,39	1,18	2,54	1,28	2,40	1,12	2,45	1,15	2,01	1,26	1,55	0,83
PV19	3,07	0,91	3,22	0,76	3,22	0,74	3,26	0,78	3,31	0,76	3,24	0,91	2,66	0,90

Tabela A.2- Desempenho estatístico (R^2, RMSE e MBE) das três bases de dados de radiação selecionadas.

Parque	S1			S4			S7		
	RMSE (%)	MBE (%)	R^2 (%)	MBE (%)	MBE (%)	R^2 (%)	RMSE (%)	MBE (%)	R^2 (%)
PV01	2,24	0,65	97,76	2,52	0,53	97,48	2,11	0,72	97,89
PV02	2,14	0,57	97,86	2,42	0,48	97,58	1,97	0,65	98,03
PV03	1,81	0,38	98,19	1,66	0,41	98,34	1,48	0,50	97,83
PV04	2,89	0,92	97,11	3,40	1,36	96,60	3,38	1,28	96,62
PV05	1,81	0,62	98,19	2,51	1,08	97,49	2,17	0,94	97,83
PV06	3,08	0,29	96,92	3,68	0,42	96,32	3,50	0,55	96,50
PV07	089	0,63	99,11	1,64	1,01	98,36	1,00	0,86	99,00
PV08	1,49	0,49	98,51	1,76	0,81	98,24	1,37	0,68	98,63
PV09	1,18	0,63	98,82	2,20	1,15	97,80	1,14	0,92	98,86
PV10	1,36	0,71	98,64	1,40	1,07	98,60	1,23	1,11	98,77
PV11	1,42	0,77	98,58	2,23	1,32	97,77	1,46	1,15	98,54
PV12	0,94	0,76	99,06	1,66	1,23	98,34	0,99	1,15	99,01
PV13	2,73	1,09	97,27	3,62	1,62	96,38	2,95	1,39	97,05
PV14	1,14	0,50	98,86	1,45	0,95	98,55	0,83	0,73	99,18
PV15	1,74	0,82	98,26	2,60	1,36	97,40	1,55	1,03	98,45
PV16	1,17	0,47	98,83	1,67	0,77	98,33	1,08	0,64	98,92
PV17	2,85	0,79	97,15	2,91	0,65	97,09	2,51	0,82	97,49
PV18	1,81	0,69	98.19	2,40	1,12	97,60	1,55	0,83	98,45
PV19	3,07	0,91	96,93	3,26	0,78	96,74	2,66	0,90	97,34
MEDIA	1,88	0,67	98,12	2,37	0,95	97,63	1,84	0,89	98,13

Apêndice B – ANÁLISES DE OUTLIERS

Tabela B.1- Resultados dos outliers removidos.

Parque	Mes	Ano	Metodo	Disponibilidade	PR	Gross EP	Estimada	ΔEP	Outliers hibridos
PV01	5	21	Z-score	99,84	81	102,708	96,37749	6,330546	
PV01	7	22	Z-score	99,7	64	95,77698	109,5952	-13,8182	X
PV01	8	22	Z-score	99,44	63	78,38972	94,05008	-15,6604	X
PV02	6	22	Z-score	99,71	68	94,73066	102,0706	-7,33993	
PV02	8	22	Z-score	99,45	68	80,11434	94,25412	-14,1398	X
PV03	3	20	IQR	97,16	90	19,43032	61	-41,6586	X
PV03	10	20	IQR	52,08	87	73,08506	56,69478	16,39028	X
PV03	3	21	IQR	18,8	73	94,06702	71	23,27869	X
PV03	6	21	IQR + Z-score	78,79	0	19,4017	101,6748	-82,2731	X
PV03	3	22	IQR	100	76	1,8076	44,86655	-43,0589	X
PV03	7	22	IQR + Z-score	71,65	74	159,3108	112,4113	46,89948	X
PV04	3	20	IQR + Z-score	92,92	87	17,69264	55,29575	-37,6031	X
PV04	10	20	IQR	52,81	86	65,67998	51,56177	14,11822	X
PV04	3	21	IQR + Z-score	17,41	76	107,525	64,10318	43,42181	X
PV05	3	20	IQR + Z-score	90,05	84.81	14,90372	47,9949	-33,0912	X
PV05	3	21	IQR	33,3	0	69,34024	56,17024	13,17	X
PV05	8	23	IQR + Z-score	9,09	4.62	55,23762	83,07684	-27,8392	X
PV05	9	23	IQR + Z-score	17,17	8.84	33,73733	59,35786	-25,6205	X

PV06	3	20	IQR	98,91	85	18,1	58,40451	-40,3045	X
PV06	3	21	IQR + Z-score	7,85	73	225,2815	68,39272	156,8888	X
PV06	6	21	IQR + Z-score	79,92	0	19,13	99,575	-80,4478	X
PV06	3	22	IQR	100	70	2,23	42,18405	-39,954	X
PV07	10	21	Z-score	99,65	108	65,59518	59,76266	5,832521	
PV07	1	22	IQR + Z-score	98,05	106	48,78878	40,45984	8,328939	
PV07	3	23	Z-score	99,33	83	79,82493	73,68604	6,138885	
PV08	3	20	IQR + Z-score	89,93	91	29,41666	61,79513	-32,3785	X
PV09	3	20	IQR + Z-score	77,89	80	31,02902	61,80096	-30,7719	X
PV09	4	20	IQR	79,65	83	90,56949	78,19274	12,37676	X
PV09	3	21	IQR + Z-score	19,28	0	112,1551	72,35343	39,80166	
PV09	2	23	IQR + Z-score	59,64	89	87,50671	53,24425	34,26246	X
PV10	3	20	IQR	99,26	88.06	25,37266	59,96367	-34,591	X
PV10	3	21	IQR + Z-score	6,11	0	285,7185	70,48456	215,2339	X
PV10	6	21	IQR + Z-score	95,98	0	23,54563	104,1161	-80,5705	X
PV10	10	21	IQR	79,9	76.8	3,630788	59,26946	-55,6387	X
PV10	3	22	IQR	100	83.6	22,26	42,37777	-20,1178	X
PV11	3	20	IQR + Z-score	95,21	88	21,00116	47,45122	-26,4501	X
PV11	3	21	IQR + Z-score	20,59	129	126,593	55,82497	70,76803	X
PV11	10	22	IQR	100	88	52,8364	39,08793	13,74847	X
PV12	3	21	IQR + Z-score	16,31	48	141,1281	71,38785	69,7403	X
PV12	6	21	IQR + Z-score	94,03	0	23,73955	103,6911	-79,9516	X
PV12	10	21	IQR + Z-score	92,49	103	4,134069	60,30753	-56,1735	X
PV12	3	22	IQR	99,88	93	22,68172	44,26009	-21,5784	X
PV13	3	20	IQR	99,88	81	33,27279	57,75973	-24,4869	X
PV13	3	21	IQR + Z-score	99,88	0	38,91193	69,18101	-30,2691	X
PV13	5	22	IQR + Z-score	98,78	75	66,16452	102,669	-36,5045	X
PV13	1	23	IQR	99,26	80	10,77499	35,41488	-24,6399	X

PV13	2	23	IQR	97,42	80	26,075	50,1272	-24,0522	X
PV13	6	23	IQR	98,93	76	74,20897	97,67896	-23,47	X
PV14	3	20	IQR + Z-score	99,39	84	28,24751	56,65854	-28,411	X
PV14	1	23	IQR + Z-score	89,53	81	10,73858	33,37812	-22,6395	X
PV14	5	23	IQR	99,42	71	86,39861	103,4979	-17,0992	X
PV14	6	23	IQR + Z-score	99,57	72	73,82797	98,24917	-24,4212	X
PV15	3	20	IQR + Z-score	99,98	84	30,97414	56,22898	-25,2548	X
PV15	1	23	IQR + Z-score	82,37	90	10,76153	33,02656	-22,265	X
PV15	4	23	IQR	100	70	83,34459	96,40685	-13,0623	X
PV15	5	23	IQR	99,22	69	86,15446	102,9114	-16,7569	X
PV15	6	23	IQR + Z-score	99,57	73	73,42627	97,68026	-24,254	X
PV16	3	20	IQR + Z-score	99,89	90	29,24224	62,21602	-32,9738	X
PV16	10	20	IQR	97,62	84	66,0336	57,05843	8,975169	
PV16	7	21	IQR + Z-score	66,56	105	172,1208	117,5544	54,56643	X
PV16	1	22	IQR	83,28	83	53,7464	41,8177	11,9287	X
PV17	8	22	Z-score	99,44	62	87,33	104,6168	-17,291	X
PV18	3	20	IQR + Z-score	89,76	86	30,11976	57,85883	-27,7391	X
PV18	1	23	IQR + Z-score	95,63	79	8,0412	35,14168	-27,1005	X
PV18	4	23	IQR	93,1	69	79,08872	97,44554	-18,3568	X
PV18	5	23	IQR	99,34	71	85,03473	103,8765	-18,8417	X
PV18	6	23	IQR + Z-score	99,84	61	74,66536	98,67093	-24,0056	X
PV18	9	23	IQR + Z-score	99,93	77	47,93535	73,36739	-25,432	X
PV19	8	22	IQR + Z-score	99,95	58	80,90084	104,756	-23,8551	X

Figura B.1- Dendrograma do agrupamento hierárquico (método de Ward) aplicado aos padrões mensais.

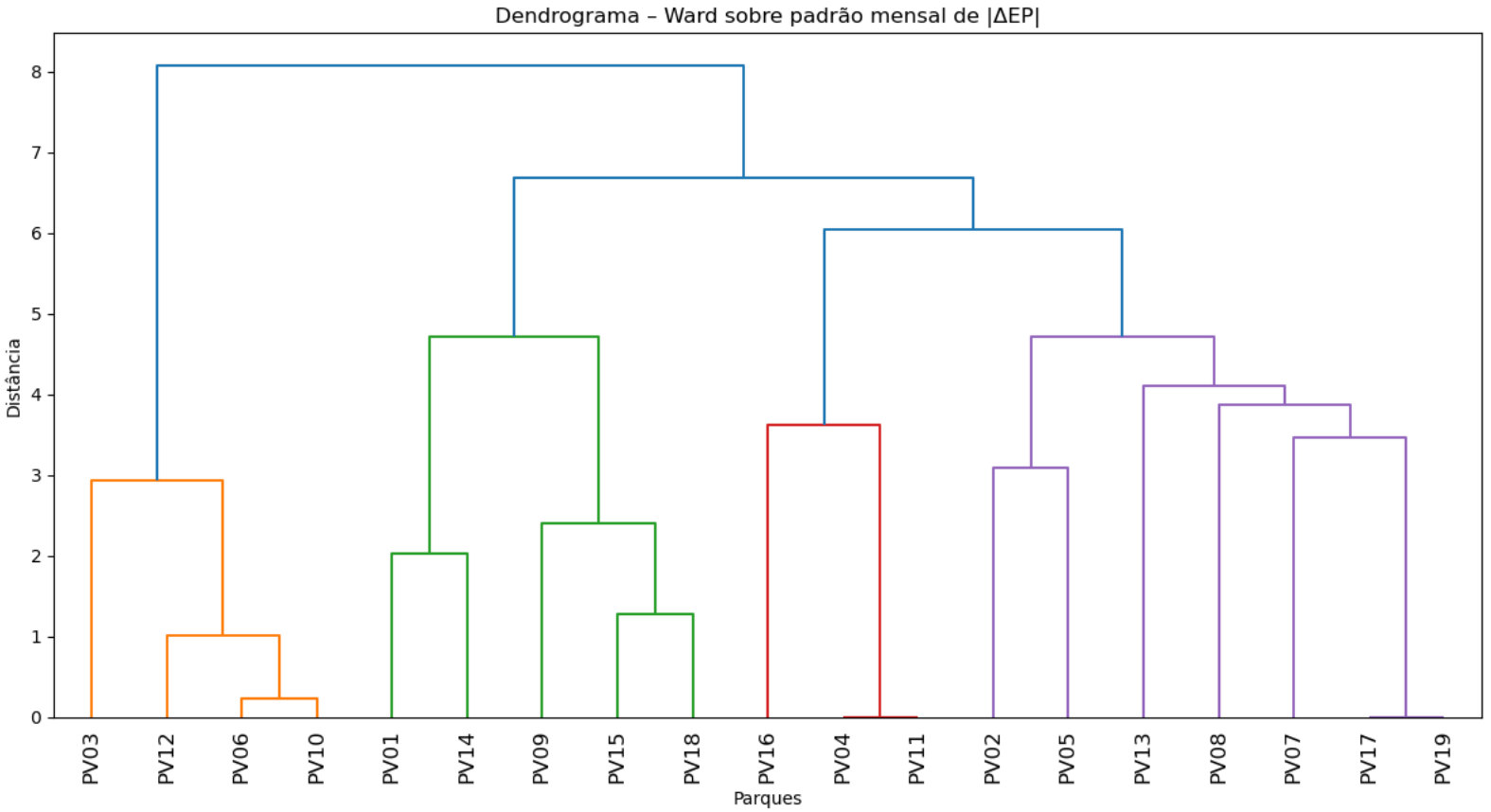


Figura B.2- Representação bidimensional (PCA) da clusterização K-means (k=3), evidenciando a separação entre parques em três grupos distintos.

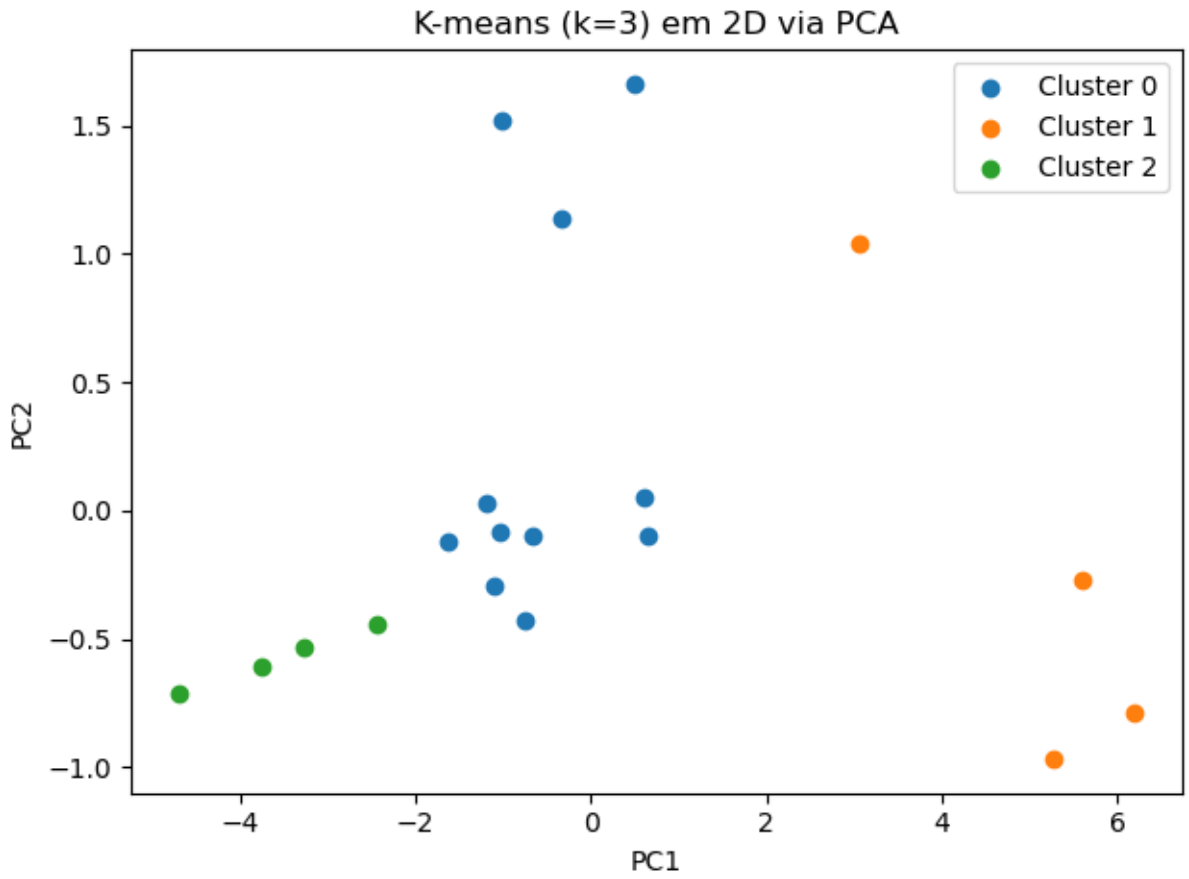
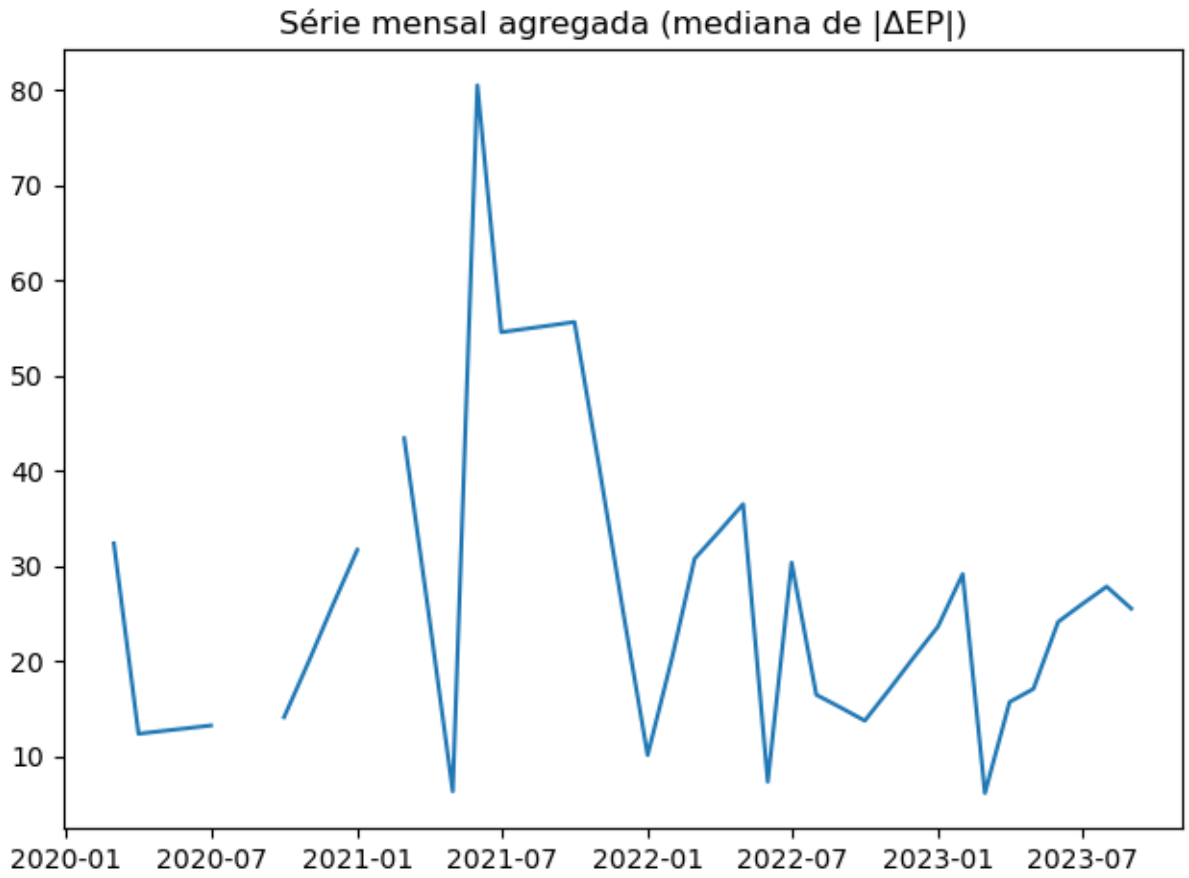


Figura B.3- Série temporal da mediana mensal de |ΔEP|, base para a decomposição STL.



Apêndice C – VALIDAÇÃO DA METODOLOGIA

Tabela C.1- Resultados dos testes de pressupostos da regressão linear para a base S1 (Solargis)

Parque	p_ANOVA	Significância	p_Shapiro	Normalidade	Durbin–Watson	Independência	p_Levene	Homocedasticidade
PV01	3,92659E-32	✔	0,939783793	✔	1,610012415	✔	0,345658	✔
PV02	1,17441E-31	✔	0,253673626	✔	1,397488472	✘	0,181506	✔
PV03	6,75449E-19	✔	0,055262243	✔	2,104078284	✔	0,097152	✔
PV04	4,28758E-23	✔	0,24209538	✔	1,042952083	✘	0,149017	✔
PV05	1,02918E-27	✔	0,475390933	✔	1,240319487	✘	0,136636	✔
PV06	2,37735E-17	✔	0,13480916	✔	0,9025975	✘	0,06324	✔
PV07	1,36808E-39	✔	0,008913466	⚠	2,342222686	✔	0,652279	✔
PV08	3,61479E-38	✔	0,344461172	✔	2,211476096	✔	0,04631	✘
PV09	1,64179E-29	✔	0,960873713	✔	2,457570016	✔	0,459721	✔
PV10	7,32579E-25	✔	0,450441524	✔	1,597127979	✔	0,206175	✔
PV11	2,8974E-29	✔	0,108438963	✔	2,034976198	✔	0,08471	✔
PV12	7,26704E-27	✔	0,537351448	✔	2,568382751	✘	0,171194	✔
PV13	5,72053E-29	✔	0,334173636	✔	1,142143362	✘	0,979709	✔
PV14	1,28227E-35	✔	0,454193845	✔	2,111441314	✔	0,541091	✔
PV15	2,80992E-33	✔	0,471405674	✔	1,142216127	✘	0,129639	✔
PV16	2,76821E-38	✔	0,035039185	⚠	2,246513068	✔	0,266954	✔
PV17	9,45113E-32	✔	0,770543825	✔	1,234705451	✘	0,508792	✔
PV18	4,44166E-32	✔	0,154614962	✔	2,173796956	✔	0,470744	✔
PV19	2,39432E-30	✔	0,773359856	✔	1,053604314	✘	0,510452	✔

Tabela C.2- Resultados dos testes de pressupostos da regressão linear para a base S2 (ERA5)

Parque	p_ANOVA	Significância	p_Shapiro	Normalidade	Durbin–Watson	Independência	p_Levene	Homocedasticidade
PV01	4,69269E-31	✔	0,485658407	✔	1,845755972	✔	0,081164907	✔
PV02	5,854E-31	✔	0,522177706	✔	1,409439499	✘	0,126149658	✔
PV03	2,07997E-19	✔	0,259915499	✔	2,140696961	✔	0,40204656	✔
PV04	1,22853E-21	✔	0,059724642	✔	1,383707685	✘	0,172729516	✔
PV05	1,9628E-25	✔	0,087035836	✔	1,588132093	✔	0,11378125	✔
PV06	1,71969E-16	✔	0,141397947	✔	0,986368906	✘	0,219413748	✔
PV07	1,57857E-34	✔	0,82934861	✔	2,213898082	✔	0,095125523	✔
PV08	1,13852E-36	✔	0,572810053	✔	2,67779454	✘	0,026745166	✘
PV09	6,86093E-26	✔	0,635759518	✔	2,406659156	✔	0,133796393	✔
PV10	1,27493E-24	✔	0,995728854	✔	2,143919191	✔	0,066753756	✔
PV11	1,6557E-26	✔	0,037783435	⚠	1,911781772	✔	0,149179706	✔
PV12	1,92689E-23	✔	0,819812914	✔	2,355158996	✔	0,077654663	✔
PV13	3,47292E-27	✔	0,146678591	✔	1,389351705	✘	0,688343335	✔

PV14	1,2957E-33	✔	0,371793896	✔	1,908256327	✔	0,386082425	✔
PV15	3,16848E-30	✔	0,486920673	✔	1,251500008	✘	0,845645466	✔
PV16	4,1695E-35	✔	0,952541773	✔	2,169931222	✔	0,123800867	✔
PV17	1,53613E-31	✔	0,639207138	✔	1,34603158	✘	0,08336577	✔
PV18	5,72834E-30	✔	0,110337709	✔	2,151047569	✔	0,170328317	✔
PV19	5,65696E-30	✔	0,545943726	✔	1,251605414	✘	0,121916534	✔

Tabela C.3- Resultados dos testes de pressupostos da regressão linear para a base S3 (ERA5)

Parque	p_ANOVA	Significância	p_Shapiro	Normalidade	Durbin–Watson	Independência	p_Levene	Homocedasticidade
PV01-S3	3,81847E-31	✔	0,443432412	✔	1,890908843	✔	0,043611274	✘
PV02-S3	3,94837E-31	✔	0,524239328	✔	1,46463007	✘	0,124050896	✔
PV03-S3	9,01682E-19	✔	0,415656977	✔	1,878526911	✔	0,730837362	✔
PV04-S3	2,88777E-22	✔	0,070142846	✔	1,398460434	✘	0,20581554	✔
PV05-S3	1,56923E-25	✔	0,045904117	⚠	1,659076265	✔	0,220076471	✔

PV06-S3	4,10233E-16	✔	0,420913499	✔	0,851928397	✘	0,336082323	✔
PV07-S3	4,22604E-34	✔	0,728884008	✔	2,157933218	✔	0,336446066	✔
PV08-S3	1,14108E-35	✔	0,579309508	✔	2,521826208	✘	0,168168153	✔
PV09-S3	8,26098E-26	✔	0,366753556	✔	2,268110986	✔	0,40291143	✔
PV10-S3	1,81407E-24	✔	0,930081593	✔	2,212991575	✔	0,104846296	✔
PV11-S3	1,57388E-26	✔	0,05525889	✔	1,905238989	✔	0,405549661	✔
PV12-S3	1,70046E-23	✔	0,617530062	✔	2,372252333	✔	0,087221438	✔
PV13-S3	4,69163E-27	✔	0,076774149	✔	1,425285998	✘	0,737315671	✔
PV14-S3	3,83721E-33	✔	0,476300792	✔	1,992284621	✔	0,802669496	✔
PV15-S3	1,13283E-29	✔	0,520184335	✔	1,318458806	✘	0,565915662	✔
PV16-S3	6,62376E-35	✔	0,96666315	✔	1,992557016	✔	0,419692588	✔
PV17-S3	1,37194E-31	✔	0,630847409	✔	1,360169217	✘	0,085715889	✔
PV18-S3	1,6228E-29	✔	0,136415477	✔	2,184040987	✔	0,421261838	✔
PV19-S3	5,91611E-30	✔	0,426255226	✔	1,289438047	✘	0,115905336	✔

Tabela C.4- Resultados dos testes de pressupostos da regressão linear para a base S4 (ERA5)

Parque	p_ANOVA	Significância	p_Shapiro	Normalidade	Durbin–Watson	Independência	p_Levene	Homocedasticidade
PV01	3,54174E-31	✔	0,613975988	✔	1,858723638	✔	0,123984785	✔
PV02	1,0851E-30	✔	0,523161335	✔	1,444388599	✘	0,188280924	✔
PV03	2,76675E-19	✔	0,235200194	✔	1,980918875	✔	0,868505551	✔
PV04	1,0955E-08	✔	0,151019905	✔	0,943242362	✘	0,750504704	✔
PV05	1,43378E-25	✔	0,374410922	✔	1,544879507	✔	0,239025505	✔
PV06	1,51229E-16	✔	0,653315642	✔	0,944702229	✘	0,282514179	✔
PV07	1,28636E-34	✔	0,666731835	✔	2,093315959	✔	0,206414238	✔
PV08	1,00377E-36	✔	0,492227196	✔	2,5825359	✘	0,089298709	✔
PV09	1,37738E-25	✔	0,153856388	✔	2,183491203	✔	0,550959196	✔
PV10	1,09722E-24	✔	0,904406711	✔	2,00486394	✔	0,124278476	✔
PV11	2,50388E-26	✔	0,039625475	⚠	1,814522392	✔	0,437209559	✔
PV12	9,28746E-24	✔	0,258122203	✔	2,397592621	✔	0,090609263	✔
PV13	8,35506E-27	✔	0,559802519	✔	1,357791404	✘	0,609813617	✔
PV14	9,00685E-34	✔	0,926889378	✔	1,819383069	✔	0,266282786	✔
PV15	4,05114E-30	✔	0,777940346	✔	1,133260682	✘	0,935184217	✔
PV16	2,09457E-35	✔	0,62685383	✔	2,078772432	✔	0,489894319	✔
PV17	1,44754E-31	✔	0,571130342	✔	1,359433024	✘	0,118947466	✔
PV18	6,26543E-30	✔	0,243148097	✔	2,04828034	✔	0,116012705	✔
PV19	7,22933E-30	✔	0,528088281	✔	1,243988822	✘	0,14838428	✔

Tabela C.5- Resultados dos testes de pressupostos da regressão linear para a base S5 (ERA5)

Parque	p_ANOVA	Significância	p_Shapiro	Normalidade	Durbin–Watson	Independência	p_Levene	Homocedasticidade
PV01-S5	5,61776E-31	✔	0,675138956	✔	1,911761224	✔	0,077184664	✔
PV02-S5	8,00678E-31	✔	0,636986684	✔	1,518074297	✔	0,206486227	✔
PV03-S5	4,74516E-19	✔	0,499677533	✔	1,807252157	✔	0,869736582	✔
PV04-S5	9,11702E-09	✔	0,165944139	✔	0,971963201	✘	0,754722493	✔
PV05-S5	2,51312E-25	✔	0,26694628	✔	1,585249892	✔	0,29449597	✔
PV06-S5	2,30343E-16	✔	0,778862177	✔	0,888460862	✘	0,427396331	✔
PV07-S5	6,64862E-34	✔	0,903683516	✔	2,27071336	✔	0,064063603	✔
PV08-S5	5,58943E-36	✔	0,621123276	✔	2,664165397	✘	0,036912817	✘
PV09-S5	9,01341E-26	✔	0,200329744	✔	2,073399821	✔	0,812657503	✔
PV10-S5	1,13022E-24	✔	0,916840776	✔	2,036572785	✔	0,149017826	✔
PV11-S5	2,46373E-26	✔	0,027980368	⚠	1,750200841	✔	0,788873994	✔
PV12-S5	3,18364E-24	✔	0,237524749	✔	2,460873823	✔	0,0763176	✔
PV13-S5	9,54893E-27	✔	0,414978557	✔	1,428787103	✘	0,994586009	✔

PV14-S5	1,07761E-33	✔	0,501174113	✔	1,938005475	✔	0,543807232	✔
PV15-S5	6,03778E-30	✔	0,881010149	✔	1,19708303	✘	0,747188111	✔
PV16-S5	4,06674E-34	✔	0,742683841	✔	2,162936976	✔	0,079605493	✔
PV18-S5	9,0104E-30	✔	0,258298099	✔	2,138820474	✔	0,247324353	✔
PV19-S5	9,99559E-30	✔	0,462632369	✔	1,303124388	✘	0,148366438	✔

Tabela C.6- Resultados dos testes de pressupostos da regressão linear para a base S6 (PVGIS)

Parque	p_ANOVA	Significância	p_Shapiro	Normalidade	Durbin–Watson	Independência	p_Levene	Homocedasticidade
alcazar1-S6	3,4724E-31	✔	0,105183875	✔	2,018980768	✔	0,011845622	✘
alcazar2-S6	6,21769E-31	✔	0,30338357	✔	1,408244338	✘	0,073674759	✔
PV03-S6	6,26675E-20	✔	0,569874758	✔	1,851059973	✔	0,934786784	✔
PV04-S6	1,30576E-21	✔	0,025212739	⚠	1,323875139	✘	0,12228055	✔
PV05-S6	4,93192E-25	✔	0,125392288	✔	1,577901831	✔	0,030782923	✘
PV06-S6	2,29842E-16	✔	0,01004242	⚠	1,039261726	✘	0,171220918	✔
PV07-S6	4,62518E-35	✔	0,915644305	✔	1,888991188	✔	0,346515178	✔
PV08-S6	1,74622E-37	✔	0,145144104	✔	2,431328874	✔	0,05564529	✔

PV09-S6	5,69428E-26	✔	0,947578023	✔	2,27486273	✔	0,247910403	✔
PV10-S6	3,27471E-24	✔	0,410870546	✔	2,094221767	✔	0,178087079	✔
PV11-S6	2,97522E-26	✔	0,064897907	✔	1,871515757	✔	0,209143454	✔
PV12-S6	4,24964E-23	✔	0,521662631	✔	2,126697738	✔	0,188621761	✔
PV13-S6	3,61917E-27	✔	0,116765542	✔	1,47808519	✘	0,86498804	✔
PV14-S6	5,7517E-36	✔	0,445690469	✔	1,824543814	✔	0,23592782	✔
PV15-S6	8,35779E-31	✔	0,452776763	✔	1,262865833	✘	0,844407398	✔
PV16-S6	2,80419E-36	✔	0,823643742	✔	2,079984569	✔	0,30648906	✔
PV17-S6	1,43442E-31	✔	0,353677596	✔	1,529859791	✔	0,040765061	✘
PV18-S6	2,6507E-31	✔	0,187111492	✔	1,915061011	✔	0,137410099	✔
PV19-S6	6,59158E-30	✔	0,07233632	✔	1,287887597	✘	0,029593875	✘

Tabela C.6- Resultados dos testes de pressupostos da regressão linear para a base S7 (PVGIS)

Parque	p_ANOVA	Significância	p_Shapiro	Normalidade	Durbin–Watson	Independência	p_Levene	Homocedasticidade
PV01-S7	1,27845E-32	✔	0,058177643	✔	1,75375418	✔	0,018964033	✘
PV02-S7	2,56917E-32	✔	0,128574658	✔	1,424037593	✘	0,057424764	✔
PV03-S7	8,99365E-20	✔	0,248726115	✔	1,668949797	✔	0,216814524	✔
PV04-S7	5,47207E-08	✔	0,15543141	✔	1,000317741	✘	0,780490636	✔
PV05-S7	1,57443E-26	✔	0,119332937	✔	1,138142134	✘	0,020742666	✘
PV06-S7	8,89649E-17	✔	0,013303019	⚠	0,80411438	✘	0,077214315	✔
PV07-S7	1,25466E-38	✔	0,495237884	✔	1,738757205	✔	0,563389674	✔
PV08-S7	6,91014E-39	✔	0,349947977	✔	2,133294711	✔	0,067047024	✔
PV09-S7	9,79619E-30	✔	0,353920796	✔	2,153720127	✔	0,300163316	✔
PV10-S7	2,08299E-25	✔	0,906246021	✔	1,765313327	✔	0,29046051	✔
PV11-S7	4,27405E-29	✔	0,01558644	⚠	1,727966307	✔	0,449632338	✔
PV12-S7	1,45737E-26	✔	0,351772133	✔	1,997201537	✔	0,334107322	✔
PV13-S7	2,22551E-28	✔	0,084336959	✔	1,295830624	✘	0,890251872	✔

PV14-S7	3,90404E-38	✔	0,373811464	✔	1,843824845	✔	0,924400684	✔
PV15-S7	3,46101E-34	✔	0,772327459	✔	1,019979989	✘	0,2681042	✔
PV16-S7	5,92387E-39	✔	0,328907751	✔	1,722936378	✔	0,275945914	✔
PV17-S7	7,8021E-33	✔	0,141091453	✔	1,391829239	✘	0,100561072	✔
PV18-S7	3,03328E-33	✔	0,251828359	✔	1,848030562	✔	0,30737878	✔
PV19-S7	1,56023E-31	✔	0,077690875	✔	1,143102548	✘	0,134363383	✔

Figura C.1 -Gráficos de resíduos para validação dos pressupostos da regressão linear PV01.

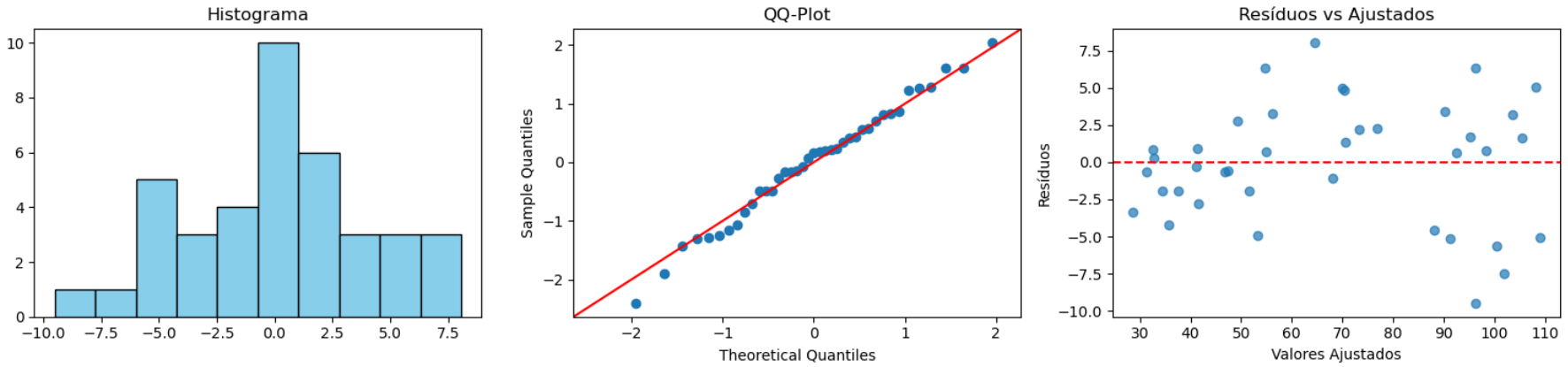


Figura C.2 -Gráficos de resíduos para validação dos pressupostos da regressão linear PV02.

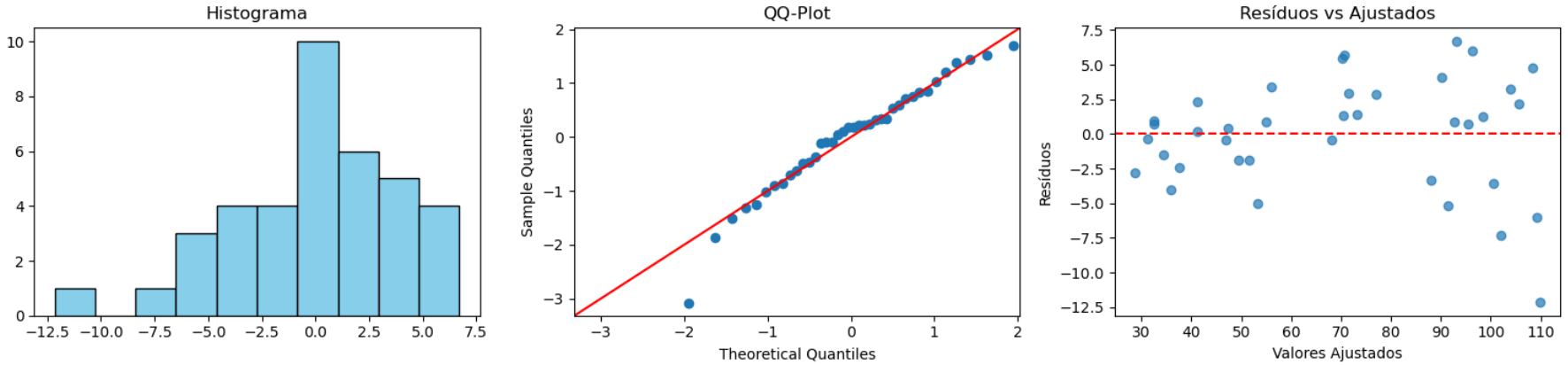


Figura C.3 -Gráficos de resíduos para validação dos pressupostos da regressão linear PV03.

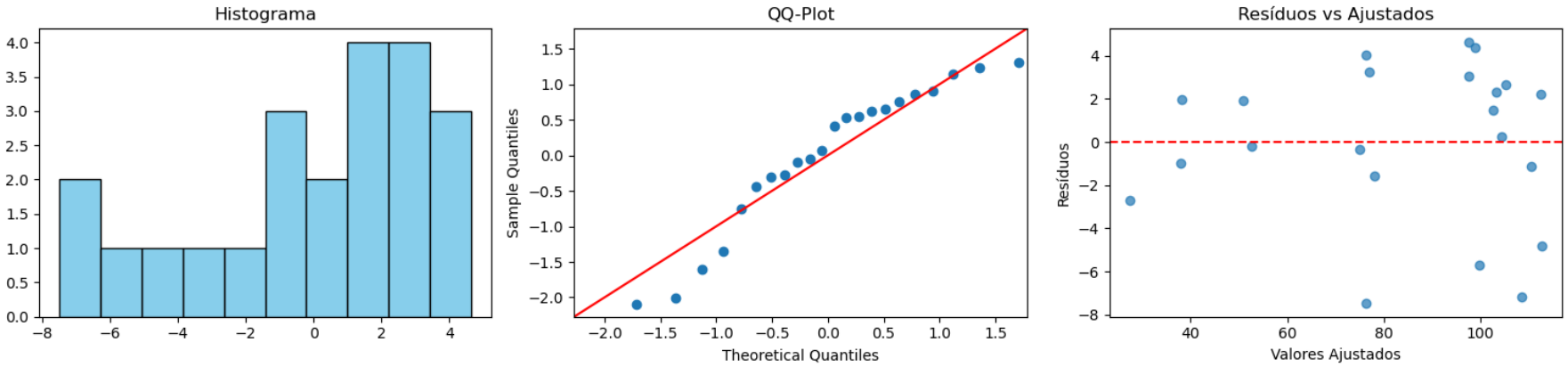


Figura C.4 -Gráficos de resíduos para validação dos pressupostos da regressão linear PV04.

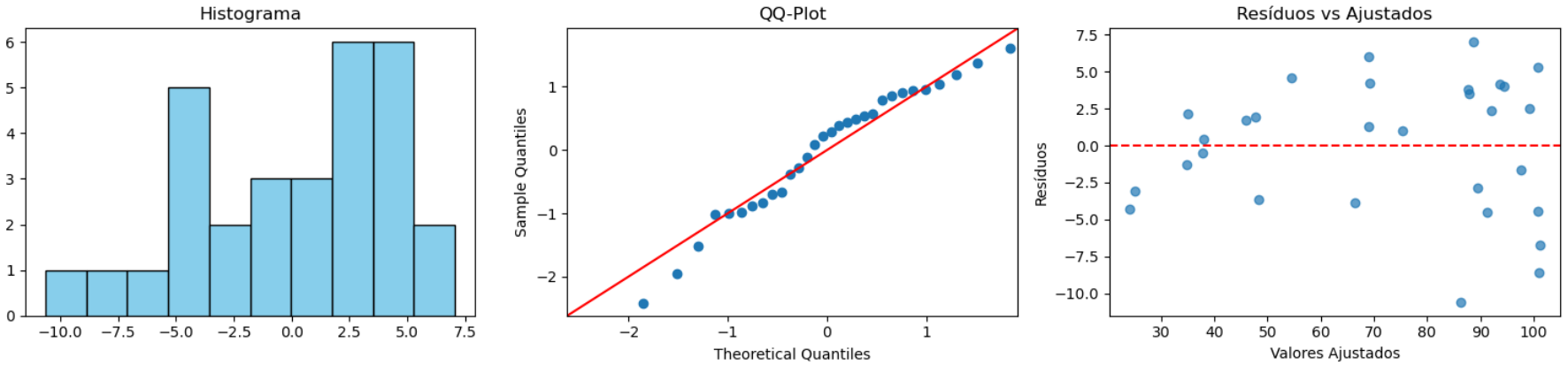


Figura C.5 -Gráficos de resíduos para validação dos pressupostos da regressão linear PV10.

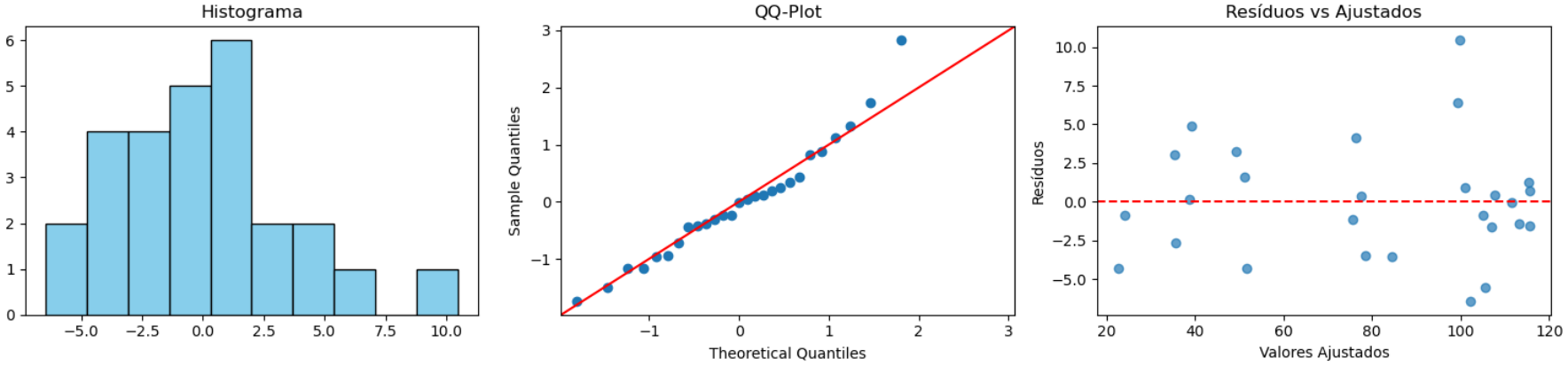


Figura C.6 -Gráficos de resíduos para validação dos pressupostos da regressão linear PV09.

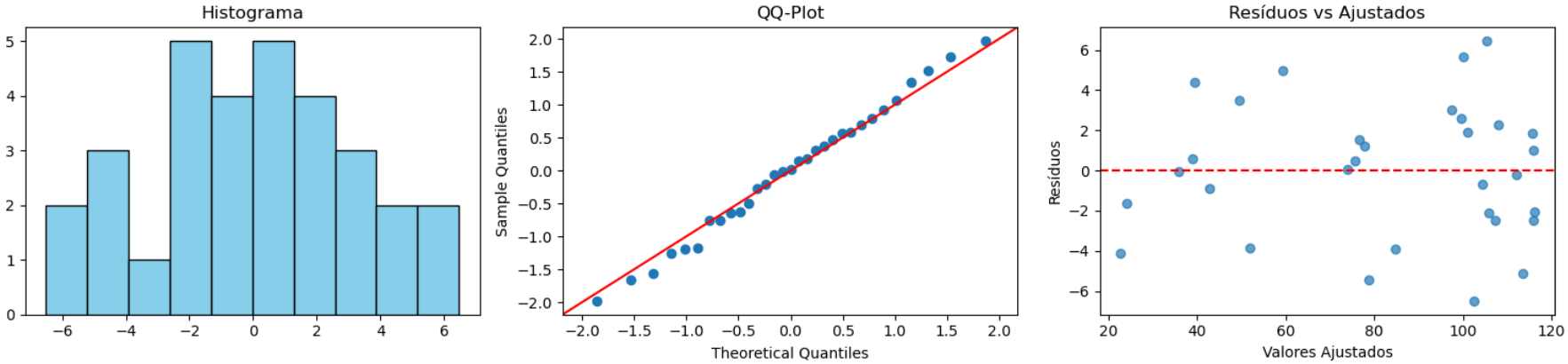


Figura C.7 -Gráficos de resíduos para validação dos pressupostos da regressão linear PV08.

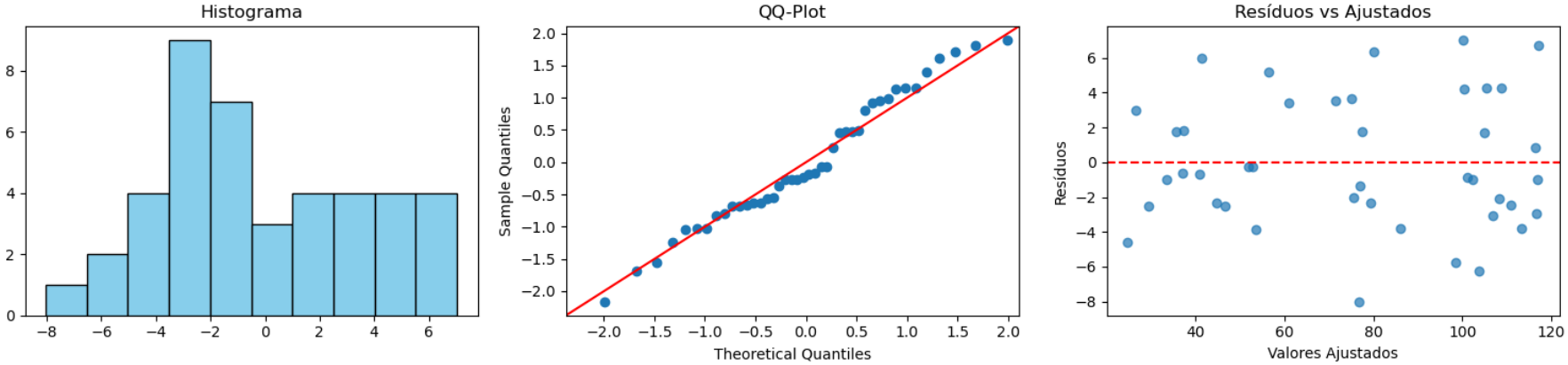


Figura C.8 -Gráficos de resíduos para validação dos pressupostos da regressão linear PV07.

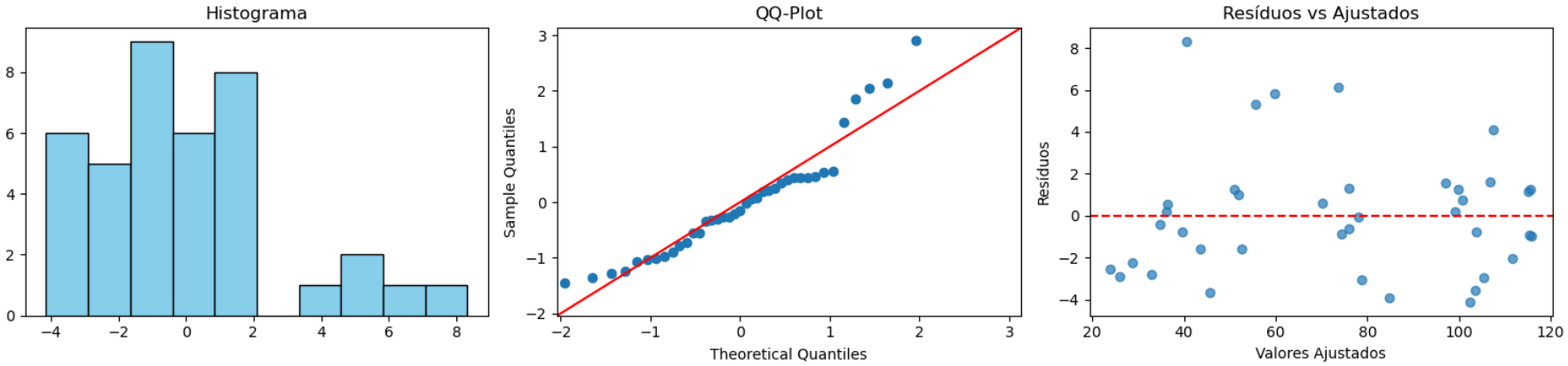


Figura C.9 -Gráficos de resíduos para validação dos pressupostos da regressão linear PV06.

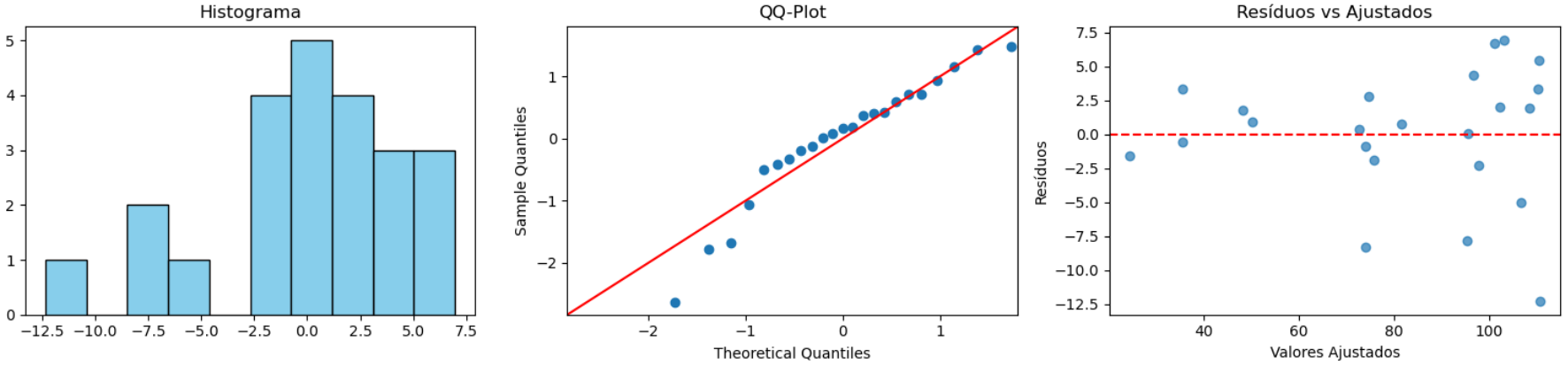


Figura C.10 -Gráficos de resíduos para validação dos pressupostos da regressão linear PV18.

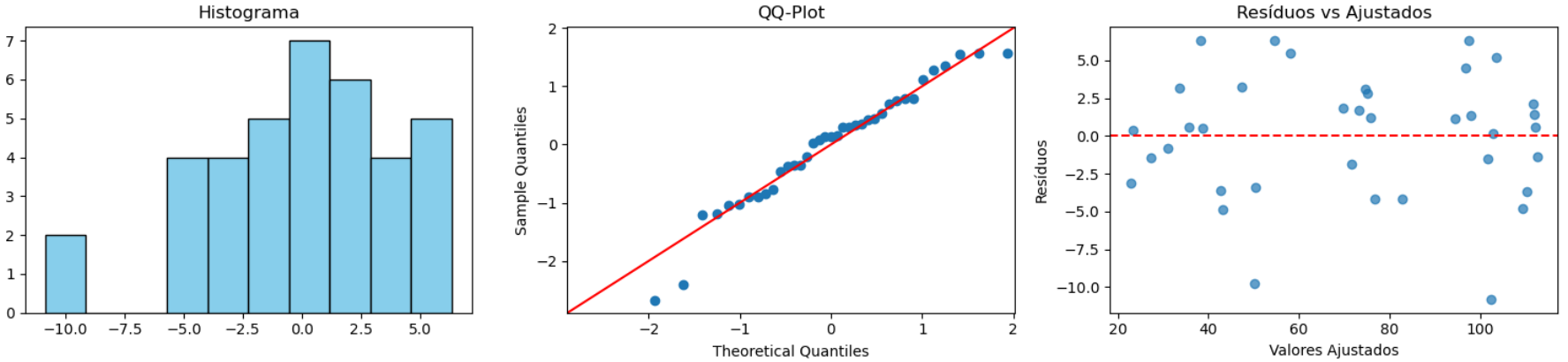


Figura C.11 -Gráficos de resíduos para validação dos pressupostos da regressão linear PV17.

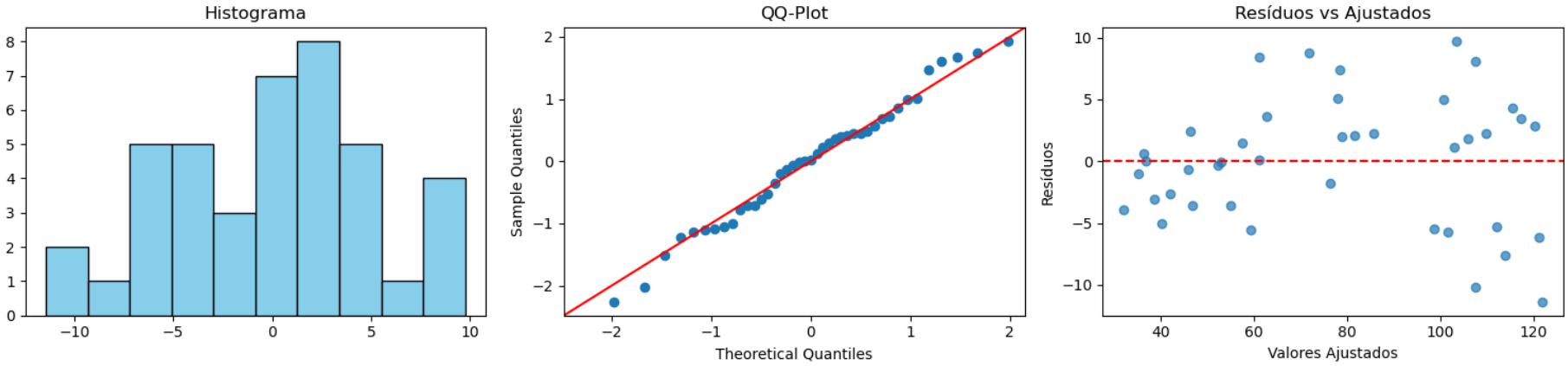


Figura C.12 -Gráficos de resíduos para validação dos pressupostos da regressão linear PV16.

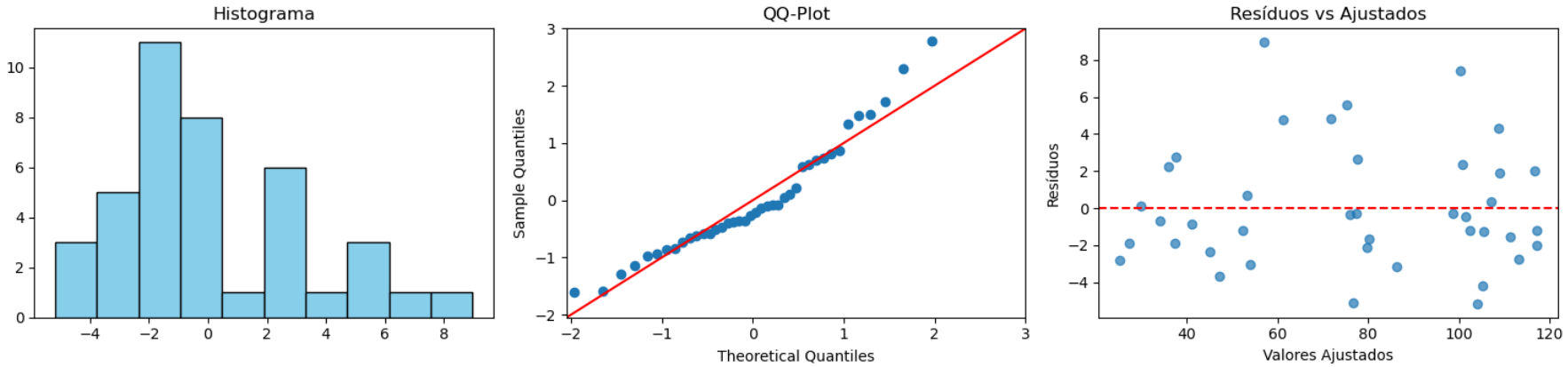


Figura C.13 -Gráficos de resíduos para validação dos pressupostos da regressão linear PV15.

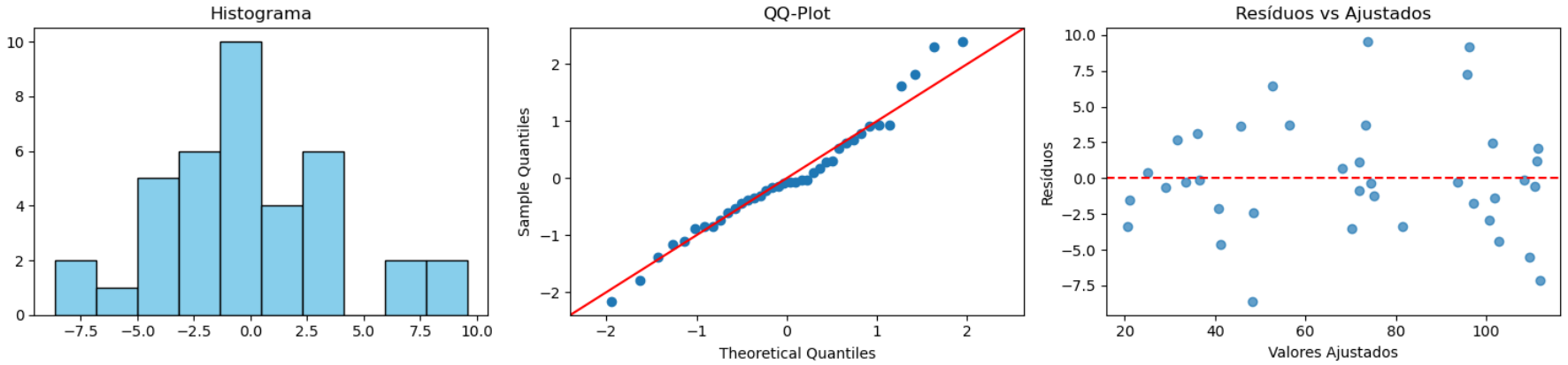


Figura C.14 -Gráficos de resíduos para validação dos pressupostos da regressão linear PV14.

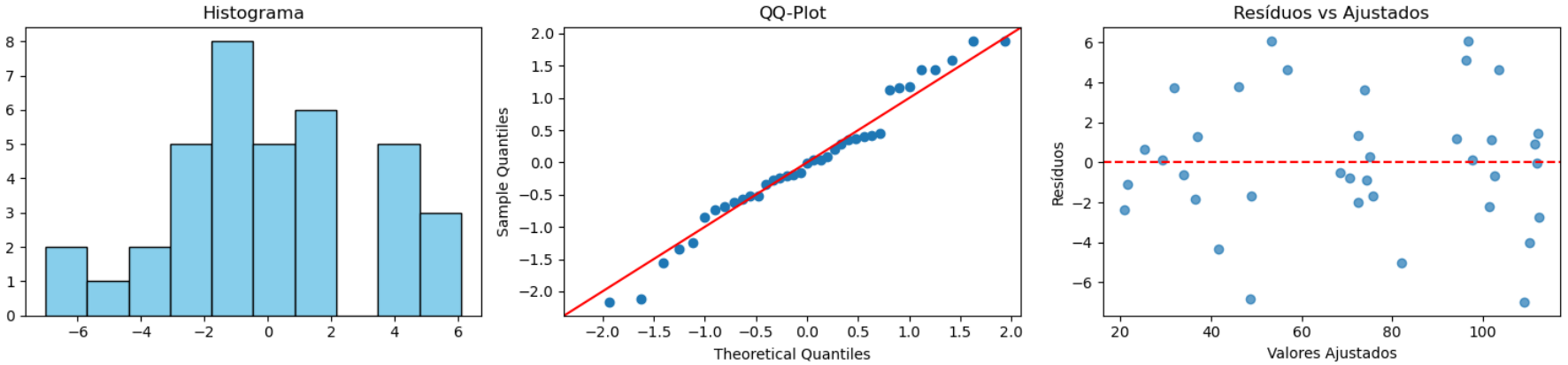


Figura C.15 -Gráficos de resíduos para validação dos pressupostos da regressão linear PV13.

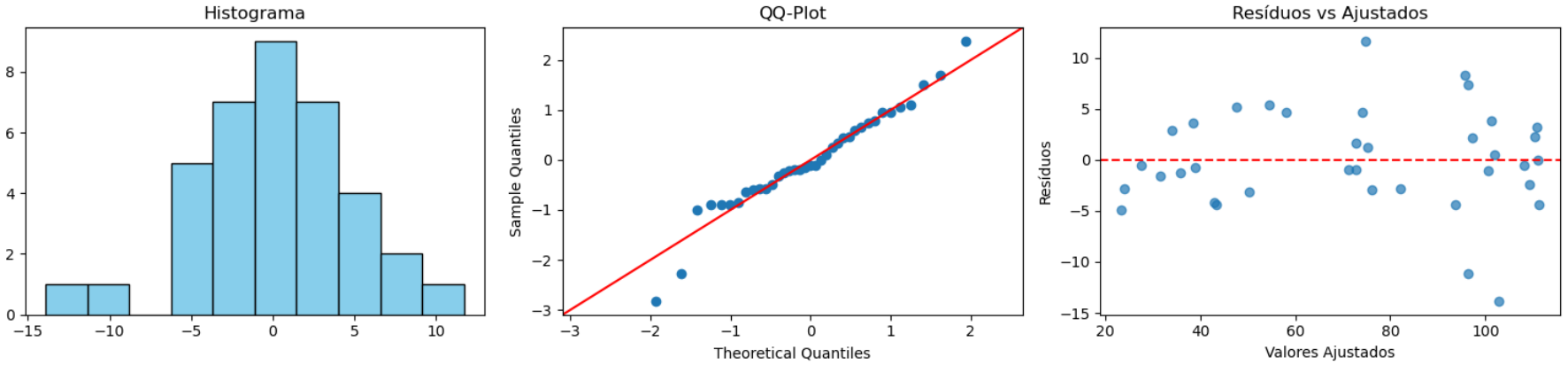


Figura C.16 -Gráficos de resíduos para validação dos pressupostos da regressão linear PV12.

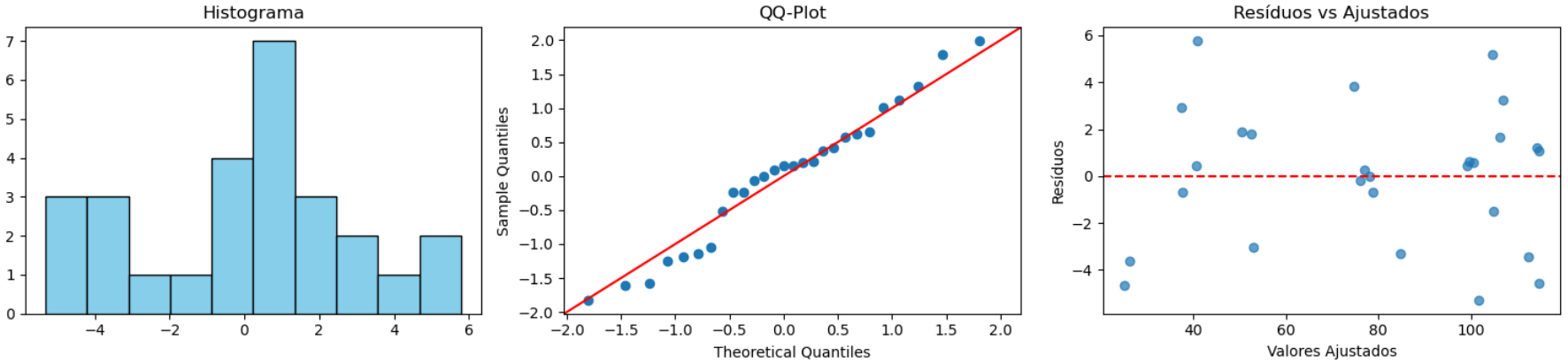


Figura C.17 -Gráficos de resíduos para validação dos pressupostos da regressão linear PV19.

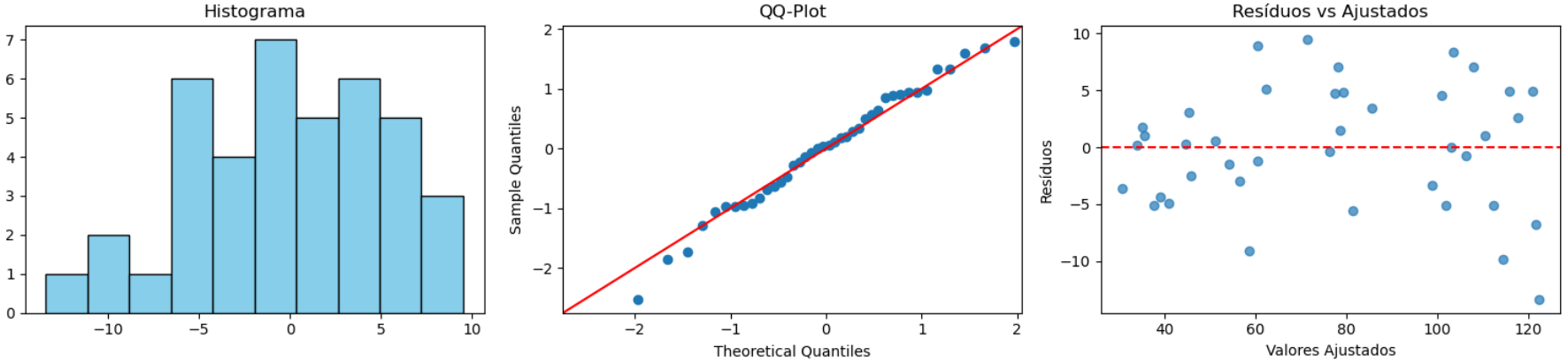


Figura C.18 -Gráficos de resíduos para validação dos pressupostos da regressão linear PV05.

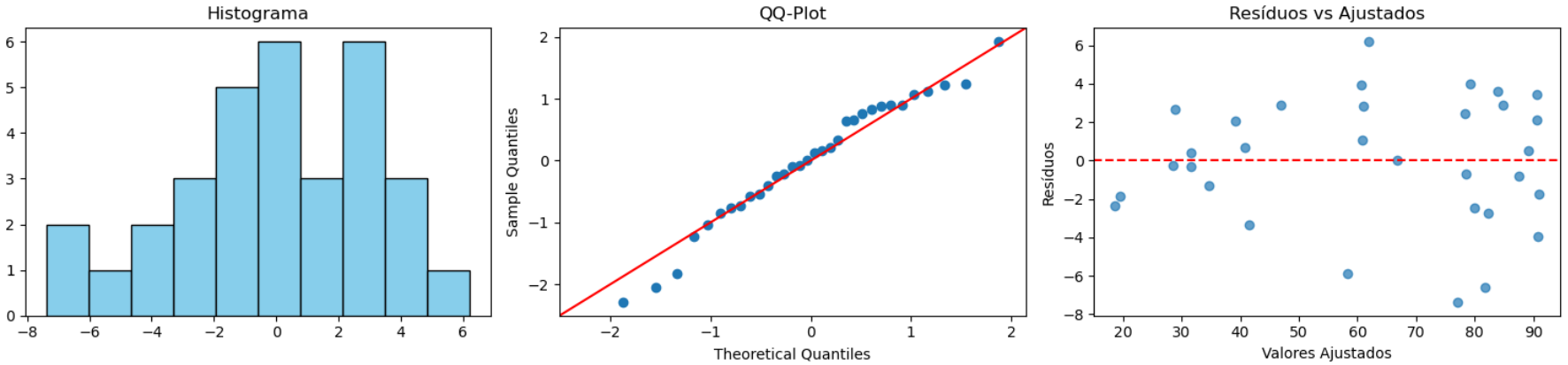
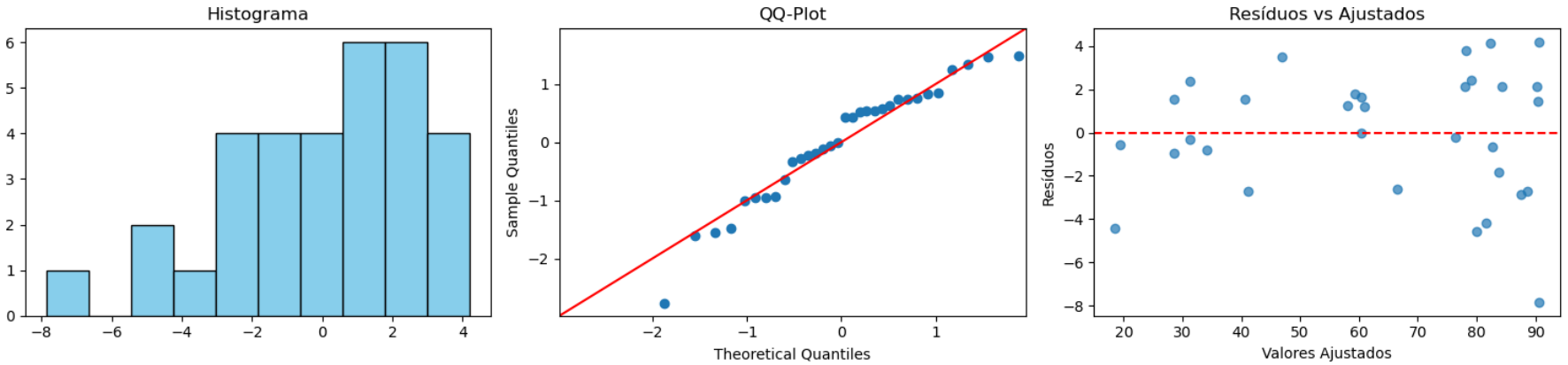


Figura C.19 -Gráficos de resíduos para validação dos pressupostos da regressão linear PV11.

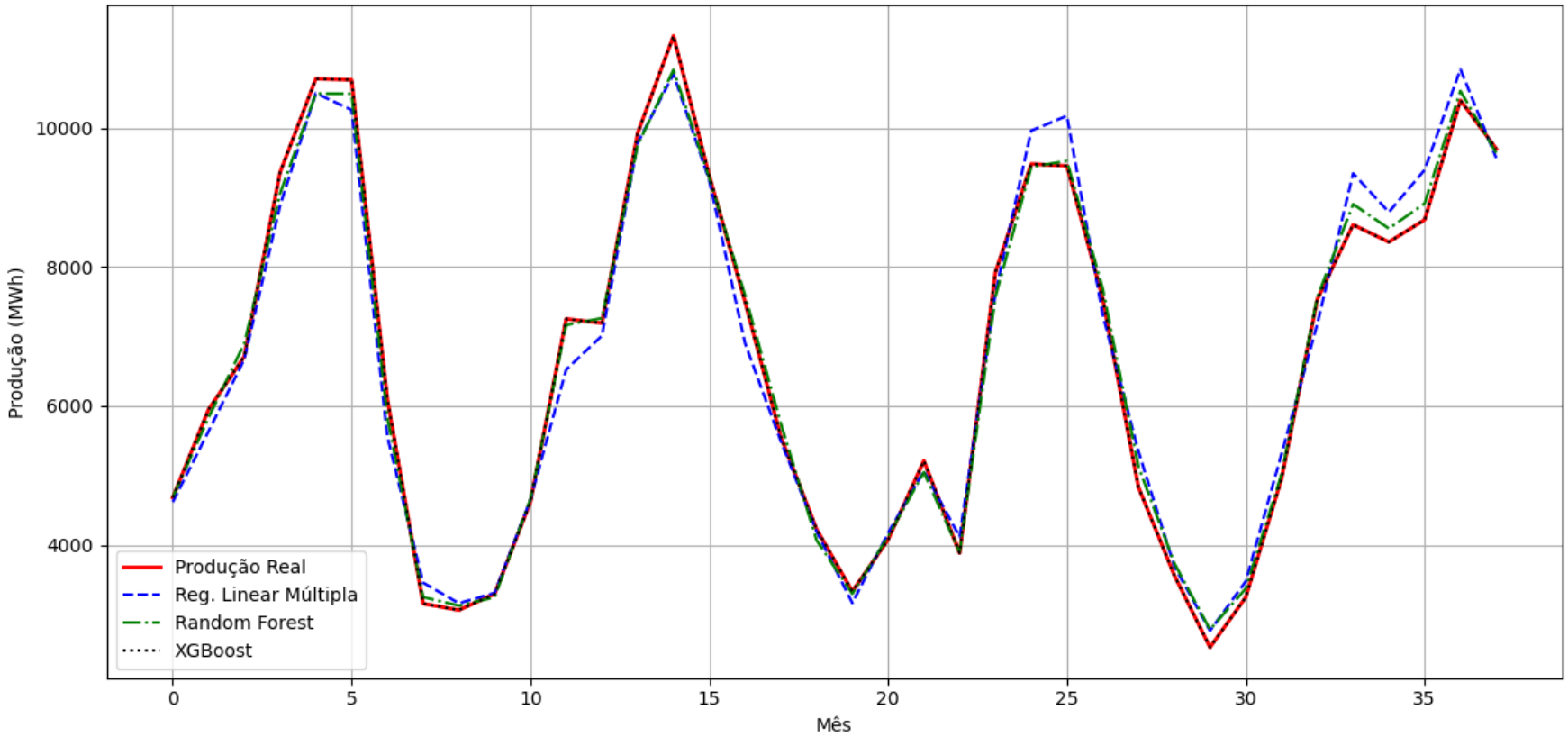


Apêndice D – MODELOS ALTERNATIVOS AO MCP

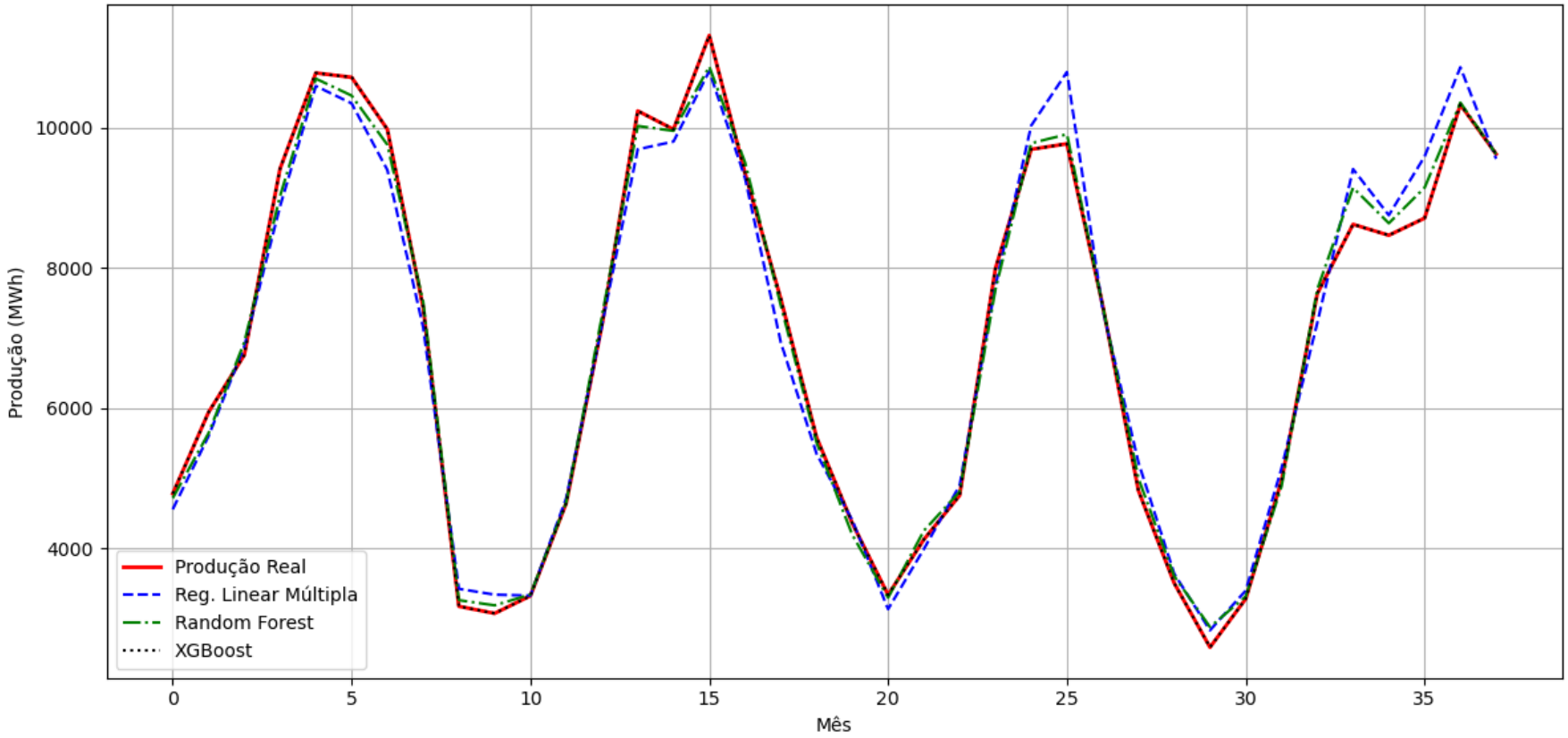
Tabela D.1- Resultados do desempenho dos modelos alternativos.

Parque	XGBoost			Random forest			Regressao linear simples (MCP)			Regressao linear multipla		
	R^2 (%)	RMSE (%)	MBE (%)	R^2 (%)	RMSE (%)	MBE (%)	R^2 (%)	RMSE (%)	MBE (%)	R^2 (%)	RMSE (%)	MBE (%)
PV01	95,63	4,37	3,64	86,32	13,68	10,85	97,76	2,24	0,65	97,94	2,06	0,57
PV02	91,34	8,66	8,44	76,91	23,09	17,21	97,86	2,14	0,57	97,90	2,10	0,50
PV17	91,11	8,89	7,93	71,03	28,97	18,5	97,15	2,85	0,79	97,40	2,60	0,68
PV19	90,15	9,85	9,95	68,32	31,68	21,03	96,93	3,07	0,91	97,06	2,94	0,74
PV18	78,08	21,92	17,94	60,53	39,47	27,06	98,19	1,81	0,69	98,36	1,64	0,48
PV08	84,29	15,71	12,74	70,66	29,34	18,7	98,51	1,49	0,49	98,69	1,31	0,37
PV14	93,56	6,44	1,8	90,72	9,28	9,05	98,86	1,14	0,50	98,35	1,65	0,54
PV15	94,33	5,67	7,57	82,3	17,7	19,48	98,26	1,74	0,82	98,41	1,59	0,73
PV16	90,59	9,41	1,48	0.953	4,7	4,44	98,83	1,17	0,47	99,21	0,79	0,20
PV07	96,12	3,88	-2,16	84,49	15,51	4,92	99,11	0,89	0,63	99,65	0,35	0,13
PV13	77,65	22,35	18,64	52,56	47,44	29,76	97,27	2,73	1,09	97,82	2,18	0,83

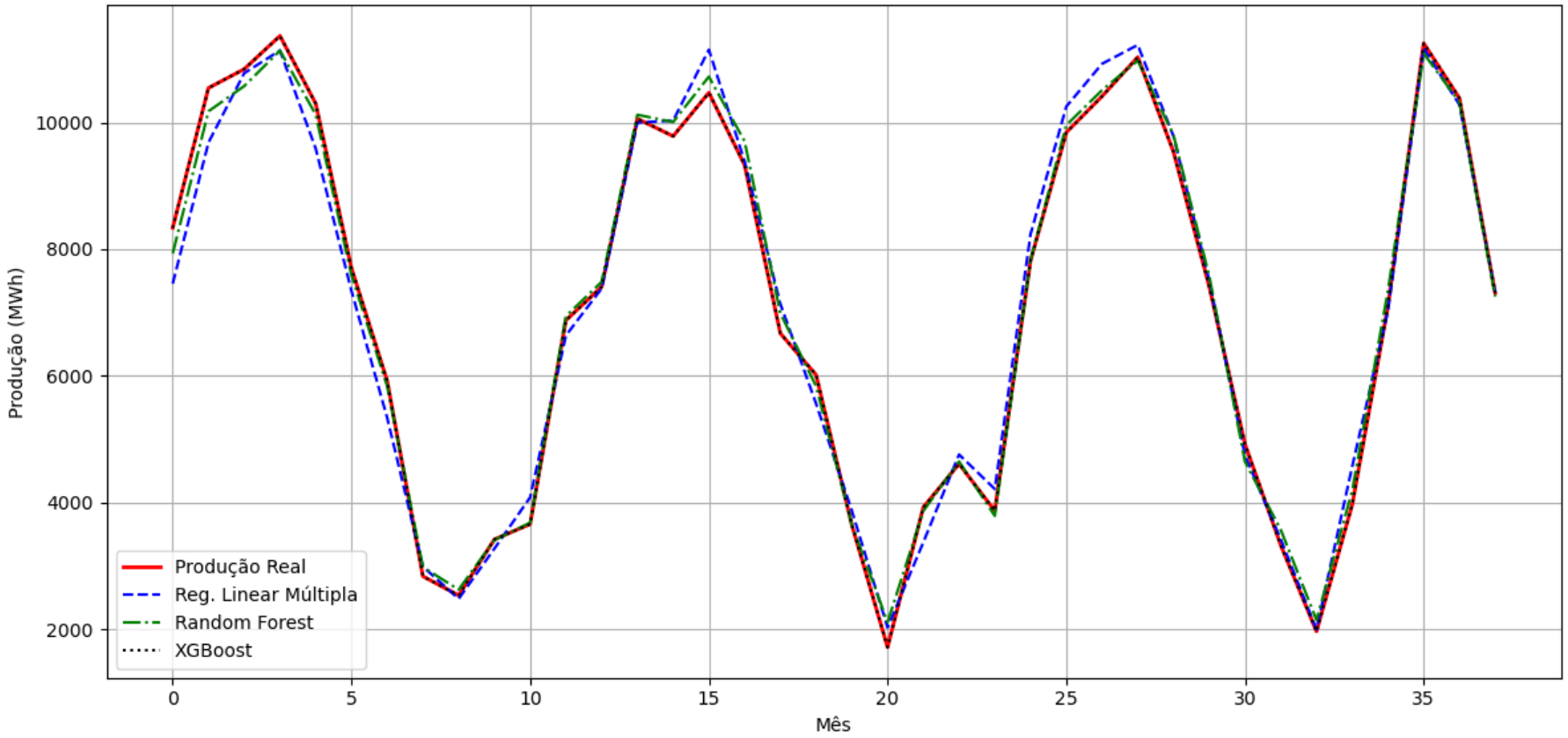
Figuras D.1- Comparação entre a produção mensal medida e as previsões obtidas com os modelos alternativos (PV01).



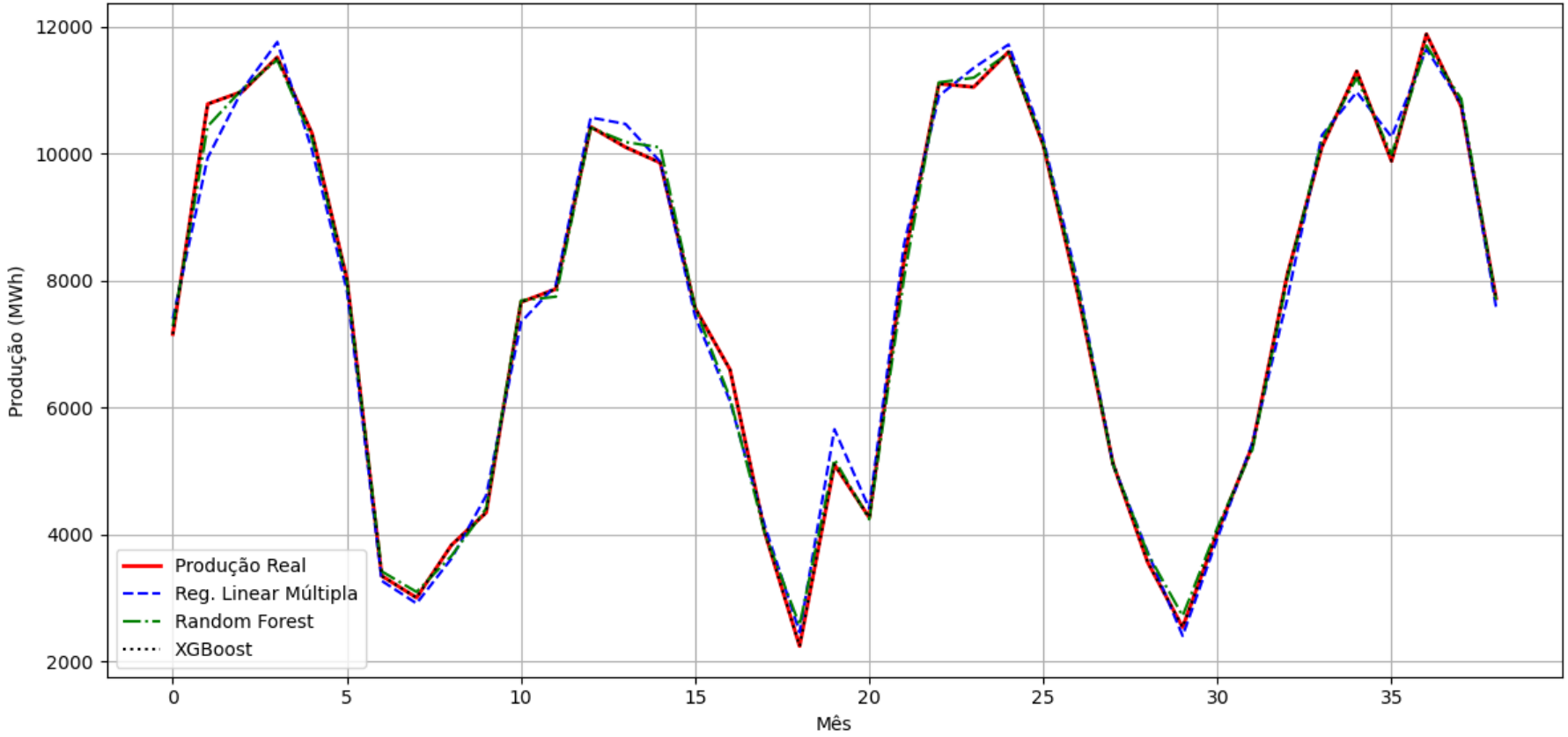
Figuras D.2- Comparação entre a produção mensal medida e as previsões obtidas com os modelos alternativos (PV02).



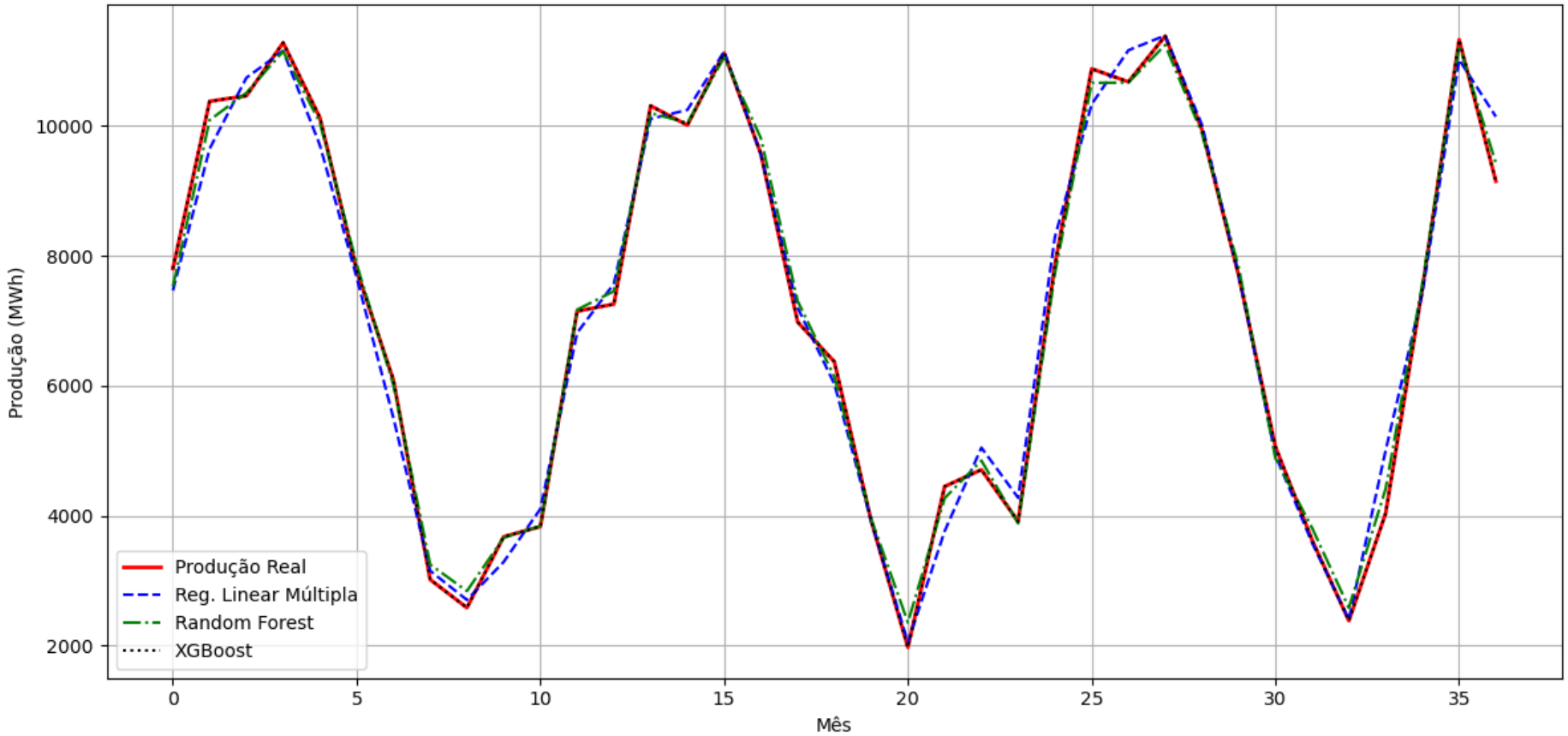
Figuras D.3- Comparação entre a produção mensal medida e as previsões obtidas com os modelos alternativos (PV15).



Figuras D.4- Comparação entre a produção mensal medida e as previsões obtidas com os modelos alternativos (PV16).



Figuras D.5- Comparação entre a produção mensal medida e as previsões obtidas com os modelos alternativos (PV18).



Apêndice E – SCRIPT EM PHYTON PARA TREINO E VALIDAÇÃO DO MODELO XGBOOST EM SERIES MENSAIS

```
import pandas as pd
import numpy as np
from sklearn.model_selection import TimeSeriesSplit
from sklearn.metrics import r2_score
from xgboost import XGBRegressor


#  Caminho para o ficheiro Excel
file_path = r"C:\Users\pc\OneDrive\Desktop\TESE\energia fotovoltaica\ML\ML data.xlsx"
xls = pd.ExcelFile(file_path)
sheet_names = xls.sheet_names


# Lista para armazenar resultados
resultados = []
#  Função com fórmula estilo MCP
def calcular_rmse_mcp(y_true, y_pred):
    numerador = np.sum((y_true - y_pred) ** 2)
    media_y = np.mean(y_true)
    denominador = np.sum((y_true - media_y) ** 2)
    rmse_pct = (numerador / denominador) * 100
    return rmse_pct
```

```
def calcular_mbe_mcp(y_true, y_pred):
    mbe_pct = np.mean((y_pred - y_true) / y_true) * 100
    return mbe_pct

# Loop por parque
for parque in sheet_names:
    try:
        df = pd.read_excel(xls, sheet_name=parque).dropna()
        df = df.select_dtypes(include=[np.number])

        if "Prod" not in df.columns or df.shape[0] < 35:
            continue

        X = df.drop(columns=["Prod"])
        y = df["Prod"]

        tscv = TimeSeriesSplit(n_splits=5)
        r2_train, r2_test, rmse_pct_list, mbe_pct_list = [], [], [], []

        for train_idx, test_idx in tscv.split(X):
            X_train, X_test = X.iloc[train_idx], X.iloc[test_idx]
            y_train, y_test = y.iloc[train_idx], y.iloc[test_idx]

            model = XGBRegressor(
                n_estimators=100,
                learning_rate=0.05,
```

```
            max_depth=3,
            subsample=0.8,
            colsample_bytree=0.8,
            reg_alpha=1.0,
            reg_lambda=1.0,
            random_state=42
        )
        model.fit(X_train, y_train)
        y_pred_train = model.predict(X_train)
        y_pred_test = model.predict(X_test)
        # R²
        r2_train.append(r2_score(y_train, y_pred_train))
        r2_test.append(r2_score(y_test, y_pred_test))
        # RMSE (%) e MBE (%) - estilo MCP
        rmse_pct = calcular_rmse_mcp(y_test, y_pred_test)
        mbe_pct = calcular_mbe_mcp(y_test, y_pred_test)
        rmse_pct_list.append(rmse_pct)
        mbe_pct_list.append(mbe_pct)
    resultados.append({
        "Parque": parque,
        "R² Treino": round(np.mean(r2_train), 4),
        "R² Teste": round(np.mean(r2_test), 4),
        "RMSE (%)": round(np.mean(rmse_pct_list), 2),
        "MBE (%)": round(np.mean(mbe_pct_list), 2),
        "Overfitting": "Sim" if (np.mean(r2_train) - np.mean(r2_test)) > 0.2 else "Não"
    })
```

```
    except Exception as e:
        print(f"Erro em {parque}: {e}")


print(df_resultados)
```