UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE TECNOLOGIA E GEOCIÊNCIAS
DEPARTAMENTO DE ENERGIA NUCLEAR
PROGRAMA DE PÓS-GRADUAÇÃO EM TECNOLOGIAS ENERGÉTICAS E NUCLEARES

DIEGO RODRIGUES DE MIRANDA

PROCEDIMENTO DE CLASSIFICAÇÃO E REGRESSÃO APLICADO AO SITE ADAPTATION DA RADIAÇÃO SOLAR

Recife
2023

DIEGO RODRIGUES DE MIRANDA

PROCEDIMENTO DE CLASSIFICAÇÃO E REGRESSÃO APLICADO AO SITE ADAPTATION DA RADIAÇÃO SOLAR

Dissertação apresentada ao Programa de Pós-Graduação em Tecnologias Energéticas e Nucleares da Universidade Federal de Pernambuco, como requisito para a obtenção do título de Mestre em Tecnologias Energéticas e Nucleares.
Área de concentração: Fontes Renováveis de Energia.

Orientadora: Profa. Dra. Olga de Castro Vilela

Coorientador: Prof. Dr. Alexandre Carlos Araújo da Costa

Recife

2023

Catalogação na fonte
Bibliotecário Gabriel Luz CRB-4 / 2222

M672p Miranda, Diego Rodrigues de.
Procedimento de classificação e regressão aplicado ao site adaptation da radiação solar / Diego Rodrigues de Miranda. 2023.
108 f: il.

Orientadora: Profa. Dra. Olga de Castro Vilela.
Coorientador: Prof. Dr. Alexandre Carlos Araújo da Costa.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CTG. Programa de Pós-Graduação em Tecnologias Energéticas e Nucleares, Recife, 2023.
Inclui referências.

1. Tecnologias energéticas e nucleares. 2. Site adaptation. 3. Classificação não supervisionada. 4. Classificação supervisionada. 5. Randomização dos dados. I. Vilela, Olga de Castro (Orientadora). II. Costa, Alexandre Carlos Araújo da (Coorientador). III. Título.

UFPE

621.042 CDD (22. ed.) BCTG / 2023 - 73

DIEGO RODRIGUES DE MIRANDA

PROCEDIMENTO DE CLASSIFICAÇÃO E REGRESSÃO APLICADO À TÉCNICAS DE SITE ADAPTATION DA RADIAÇÃO SOLAR

Dissertação apresentada ao Programa de Pós-Graduação em Tecnologias Energéticas e Nucleares da Universidade Federal de Pernambuco, como requisito para a obtenção do título de Mestre em Tecnologias Energéticas e Nucleares.
Área de concentração: Fontes Renováveis de Energia.

Aprovada em: 28/02/2023

BANCA EXAMINADORA

_____
Prof. Dr. Tsang Ing Ren (Examinador Externo)
Universidade Federal de Pernambuco

_____
Prof. Dr. Rodrigo Alonso Suárez (Examinador Externo)
Universidad de la República Uruguay

_____
Prof. Dr. Germán Ariel Salazar (Examinador Externo)
Universidad Nacional de Salta

_____
Prof. Dr. Emmanuel Damilano Dutra (Examinador Interno)
Universidade Federal de Pernambuco

Dedico esse trabalho aos pesquisadores do Centro de Energias Renováveis (CER-UFPE) que tanto me inspiraram e auxiliaram na solução do problema proposto.

AGRADECIMENTOS

Agradeço imensamente aos pesquisadores do Centro de Energias Renováveis da UFPE, onde fui acolhido de braços abertos por pessoas maravilhosas. Janis, Renan, João, Valentin, Leonardo Petribú, Gabriel, Lucas, Rodrigo, e tantos outros amigos pesquisadores que fizeram o ato de pesquisar ser também um ato de compartilhar. Compartilhar ideias, compartilhar sorrisos, compartilhar lágrimas, pulsações. Todos os projetos que fizemos juntos, todo o aprendizado, as reuniões, compras de equipamentos, discussões para relatórios, viagens, apresentações dos nossos trabalhos de pesquisa e as trocas de ideias fazem do CER-UFPE um espaço inspirador e que faz qualquer pesquisador brilhar o olhar.

Agradeço imensamente aos meus orientadores Olga Vilela e Alexandre Costa, vocês são professores que fazem qualquer um querer seguir a pesquisa. Obrigado pelas orientações, obrigado pelo olhar perpiscaz, pelo apoio quando a gente se desespera, pelos conselhos. Acredito que reencontrei uma veia de querer descobrir as coisas com a experiência do mestrado. Agradeço à professora Elielza por todo o apoio durante o mestrado e a Evelyn por todo o apoio administrativo nos projetos.

Agradeço também a todo o apoio da minha família, que sempre deram suporte nessa caminhada pelo Recife. Sobretudo à minha mãe, Maria de Lourdes, que sempre me apoiou nas decisões e deu suporte em todas as viagens para Petrolina. Em especial, à Larissa e à Diego por todos os conselhos, revisões e suporte durante o tempo do mestrado.

Por fim, agradeço ao apoio financeiro da Companhia Hidro Elétrica do São Francisco – CHESF por meio do projeto Plataforma Solar de Petrolina CVI 80.2020.0010.00 (23076.009704/2020-56), do projeto IBITU.INTELIPREV, ambos Programas de P&D ANEEL, como também à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio aos bolsistas no âmbito do Programa de Pós-Graduação em Tecnologias Energéticas e Nucleares – PROTEN.

RESUMO

A estimativa da radiação solar em diferentes localidades durante vários anos é importante para o desenvolvimento de projetos fotovoltaicos e heliotérmicos de grande porte. Nesse contexto, modelos de *site adaptation* são utilizados para ajustar séries de radiação provenientes de bases de dados históricas com medições em terra. Neste trabalho, modelos estatísticos são aplicados de forma global e local, bem como de forma combinada, para realização do *site adaptation* na resolução temporal de 15 minutos da irradiância global horizontal (GHI) e da irradiância direta normal (DNI). Quatro estações meteorológicas são utilizadas para testar o *site adaptation*: duas na Argentina (El Rosal e Salta), uma no Brasil (Petrolina) e uma na Namíbia (Gobabeb). As variáveis regressoras são formadas por dados provenientes da *Copernicus Atmosphere Monitoring Service* (CAMS) e do *European Centre for Medium-Range Weather Forecasts* (ECMWF). Nos modelos globais são aplicados modelos regressivos para toda a série temporal, enquanto nos modelos locais, as variáveis regressoras são divididas em grupos de acordo com uma classificação das condições do céu no local. A regressão linear múltipla e a rede neural do tipo *multilayer perceptron* são os modelos regressivos utilizados neste trabalho. Os subconjuntos locais das séries temporais são definidos através de uma classificação não supervisionada, sendo utilizado o algoritmo de *k-means*. Em seguida é feita uma classificação supervisionada com o modelo *Random Forest*, onde as variáveis regressoras devem aprender o comportamento das classes estabelecidas na etapa anterior. Os dados são divididos em três conjuntos para aplicação dos modelos: calibração, validação e teste. No conjunto de validação, são treinados os modelos regressivos de combinação, utilizando todos os modelos globais e locais cujos parâmetros foram inferidos no conjunto de calibração. Este trabalho também avalia a robustez dos modelos para 62 diferentes divisões dos dados nos três conjuntos citados. Algumas dessas divisões mantém a sequência cronológica das séries temporais enquanto outras fazem uma randomização dos *timestamps* de 15 minutos antes das divisões nos três conjuntos. Os resultados do *site adaptation* da GHI nas estações de El Rosal e Salta mostram que os modelos de combinação e locais apresentam maior acurácia que os modelos globais e da CAMS, enquanto nas estações de Petrolina e Gobabeb, os resultados de todos os modelos são similares aos da CAMS. É provável que os resultados tenham sido melhores nas estações de

El Rosal e Salta, pois as mesmas estão localizadas na borda do campo de visão do satélite METEOSAT, o que pode induzir a mais erros nas séries de radiação estimadas pela CAMS. Já para o *site adaptation* da DNI nas estações de Petrolina e Gobabeb, os modelos de combinação e locais conseguem maior acurácia que os modelos globais e da CAMS. Os resultados também mostram que modelos aplicados com os dados divididos com a estratégia de randomização dos *timestamps* de 15 minutos apresentam melhores resultados estatísticos que as divisões que mantêm a sequência cronológica das séries temporais.

Palavras-chave: *site adaptation*; classificação não supervisionada; classificação supervisionada; randomização dos dados.

ABSTRACT

Estimating the solar radiation accurately during many years in different regions is necessary to develop photovoltaic and heliothermic solar power plant projects. In this work, statistical models are employed using ground measurements to correct long-term time series of radiation, a procedure known as site adaptation. These statistical models are applied globally, locally and in combination to make the site adaptation of global horizontal irradiance (GHI) and direct normal irradiance (DNI) in 15 minutes temporal resolution. Four meteorological stations located in Argentina (El Rosal e Salta), Brazil (Petrolina) and Namib (Gobabeb) are used to test the models. The variables used in the models are from Copernicus Atmosphere Monitoring Service (CAMS) and European Centre for Medium-Range Weather Forecasts (ECMWF). The global models are applied to all-time series, while in the local ones, the regression variables are split into clusters according to a classification based on the sky conditions. Multilinear regression and multilayer perceptron neural networks are the two main regression algorithms used in this work. The subsets of local modes are defined using a non-supervised classification to create a vector of classes made by the k-means algorithm. Then a supervised classification is applied using the random forest algorithm, where the variables will learn the classes of the non-supervised step. The data are split into three sets to apply and evaluate the models: calibration, validation and test sets. In the validation set, the combination models are trained using all results from global and local models whose parameters were defined in the calibration set. This work also evaluates the robustness of the models for 62 different divisions in the three sets mentioned. Some of these divisions maintain the chronological sequence of time series while others divisions shuffled the 15 minutes timestamps before the split into the three sets. The results of GHI site adaptation in El Rosal and Salta stations show better accuracy for the combination and local models than global models and CAMS, while in Petrolina and Gobabeb stations, the results of all models are similar to CAMS. The results were probably better in El Rosal and Salta stations because these stations are located at the edge of the METEOSAT satellite view, which can induce errors in CAMS estimations. The results of DNI site adaptation in the Petrolina and Gobabeb stations show that the local and combination models achieve better accuracy than the global models and CAMS. The overall results also suggest that the models applied with the data divided by shuffling the 15 minutes timestamps present better

statistical results than the divisions that maintain the chronological sequence of time series.

Keywords: site adaptation; non-supervised classification; supervised classification; data shuffling

LISTA DE ABREVIATURAS E SIGLAS

AIC	*Akaike information criterion*
A_r	Índice de razão da área
CAMS	*Copernicus Atmosphere Monitoring Service*
CLD	*Cloud Cover*
CM-SAF	*Satellite Application Facility on Climate Monitoring*
D	Dimensão Fractal
DLR-Solemi	*German Aerospace Center - Solar Energy Mining*
ECMWF	*European Centre for Medium-Range Weather Forecasts*
EHF	*Energy and Semiconductor Research Laboratory, University of Oldenburg*
F_m	Índice de fração da manhã
glm	*Generalized Linear Models*
gsd	*Granulometric Size Distribution*
GMM	Gaussian Mixture Models
H	Umidade
HelioClim-3	Banco de dados de radiação solar derivado de imagens de satélite e modelos físicos desenvolvido pela empresa SODA (Solar Radiation Data)
I	Intermitência
I_{bn} ou DNI	Irradiância direta normal
I_d ou DHI	Irradiância difusa horizontal
I_g ou GHI	Irradiância global horizontal
$I_{g,ClearSky}$	Irradiância global horizontal de céu claro
I_{oeff}	Irradiância extraterrestre efetiva
I_{oh}	Irradiância extraterrestre horizontal
JRA-55	Conjunto de dados de reanálise de segunda geração iniciado em 2013 pela Agência Meteorológica do Japão (*Japan Meteorological Agency*)
k_b	Transmitância normal
k_c	Índice de céu claro
k_d	Razão difusa

$k_{h,d}^{*}$	Índice de claridade horário sem tendência sazonal
k_t	Índice de claridade
k_{tn}	Índice de claridade normalizado
m_a	Massa de ar
MBE	*Mean Bias Error*
MOS	*Model Output Statistics*
MLP	Redes neurais do tipo *Multilayer Perceptron*
NASA	*National Aeronautics and Space Administration*
NCEP-DOE	*National Centre for Environmental Prediction, USA - Department of Energy, USA*
NCEP-FNL	*National Centre for Environmental Prediction, USA - Final*
NCEP-GFS	*National Centre for Environmental Prediction, USA - Global Forecast System*
NCEP-NCAR	*National Centre for Environmental Prediction, USA - National Centre for Atmospheric Research, USA*
NSRDB	*National Solar Radiation Database*
NWP	*Numerical Weather Prediction*
P	Pressão
PAM	*Partition Around Medoids*
PCA	*Principal Component Analysis*
POP_D	Probabilidade diária da persistência
PVGIS	*Photovoltaic Geographical Information System*
QM	*Quantile Mapping*
r	Correlação
RMSE	*Root Mean Square Error*
SH	*Sunshine Hour*
SVM-C	*Support Vector Machines for Classification*
SVM-R	*Support Vector Machines for Regression*
T	Temperatura
VI	Índice de variabilidade
3TIER	Banco de dados de radiação solar derivado de imagens de satélite e modelos físicos desenvolvido pela empresa Vaisala

SUMÁRIO

1 INTRODUÇÃO

No desenvolvimento de projetos fotovoltaicos ou heliotérmicos de grande porte, uma boa estimativa do recurso solar da região de interesse é necessária para obter estimativas de produção de energia mais acuradas. Para tanto, campanhas de medições de radiação solar, temperatura, umidade e velocidade do vento, entre outras variáveis meteorológicas, são realizadas nos locais específicos onde pretende-se conhecer o recurso solar, com vistas à implementação de empreendimentos de energia. Atualmente no Brasil, as campanhas de medições para projetos fotovoltaicos de grande porte devem ter duração mínima de 1 ano, sendo a principal medição a da irradiância global horizontal (EPE, 2017). Já para projetos heliotérmicos ou projetos fotovoltaicos com concentração, ambos também de grande porte (mais de 5 MW), a principal medição é a da irradiância direta normal por, no mínimo, três anos. As instruções direcionadas aos empreendimentos de energia solar de grande porte, incluindo as instruções associadas às campanhas de medições, são publicadas pela Empresa de Pesquisa Energética para possibilitar às empresas a participação nos leilões de energia promovidos pela Agência Nacional de Energia Elétrica (ANEEL).

Contudo, para estimar o recurso solar em dada região, não basta somente realizar a campanha de medições no local específico, mas estimar o comportamento climatológico das variáveis de interesse. Como a fonte solar é intermitente, com variações estocásticas ao longo do tempo, a caracterização do recurso deve considerar uma janela temporal de, no mínimo, 10 anos, período muitas vezes denominado como climatologicamente significativo (HABTE ET AL., 2017). Para isso, bases de dados históricas que fornecem estimativas de variáveis ambientais durante vários anos podem ser utilizadas, já que é inviável medir a radiação solar por mais de 10 anos em projetos de energia solar. Como as séries históricas de radiação solar obtidas a partir de modelos de reanálise ou modelos baseados em informações de satélites podem não reproduzir com acurácia a radiação solar no local de interesse, modelos estatísticos podem ser empregados para ajuste dessas séries. O *site adaptation* é um modelo estatístico que utiliza as medições realizadas em terra em um período menor de tempo (geralmente, de 1 a 3 anos) para corrigir as séries climatológicas fornecidas pelas bases de dados históricas (POLO ET AL., 2016).

As séries históricas de radiação podem ser provenientes tanto de modelos de previsão numérica do tempo (*Numerical Weather Prediction* - NWP) ou reanálise, a

exemplo da base de dados do *European Centre for Medium-Range Weather Forecasts* (ECMWF, e.g. modelos ERA-5 e ERA-5 Land) e da *National Aeronautics and Space Administration* (NASA, e.g. modelo MERRA-2), quanto a partir de modelos que utilizem imagens de satélite (e.g. PVGIS, CERES, NSDBR, CAMS). Atualmente, os modelos baseados em imagens de satélites são mais utilizados para avaliação do recurso solar, pois apresentam estimativas mais acuradas que modelos de NWP. Apesar disso, diferenças significativas ainda são encontradas entre as séries fornecidas por modelos baseados em informações de satélites e as séries medidas por estações solarimétricas, principalmente para radiação direta normal. Fernández-Peruchena et al. (2020) citam quatro principais fontes de erros de modelos baseados em imagens de satélites: o efeito das nuvens na atenuação da radiação solar; o modelo de céu claro utilizado; a área do pixel do satélite que pode levar a erros principalmente durante dias nublados (nuvens intermitentes) e devido a mudanças na quantidade de aerossóis presentes na atmosfera; e os erros devido a terrenos elevados e altos albedos (radiação refletida pela superfície do solo) provenientes de desertos ou neve. Os efeitos desses potenciais erros podem resultar em séries com um viés (BIAS) e desvio (amplitude de variação) diferentes do esperado. Uma possível solução para superar esses potenciais erros é fazer o *site adaptation* das séries fornecidas pelas bases históricas.

Para realizar o *site adaptation*, as séries de longo prazo devem ter uma cobertura temporal que inclua o período da campanha de medições realizada em terra. Assim, calibram-se e validam-se os modelos estatísticos no período concomitante entre as medições realizadas pela estação solarimétrica no local de interesse e as bases de dados históricas. Para gerar a série histórica adaptada ao local é necessário aplicar o modelo estatístico calibrado e validado à toda a série histórica. Essa série de longo prazo ajustada pode ser utilizada em diferentes aplicações. Em termos de empreendimentos solares, a série histórica é utilizada para o cálculo da da variabilidade interanual do recurso (KARIUKI E SATO, 2018), obtenção das médias de longo prazo de radiação solar, geração de um ano meteorológico típico para a região de interesse (WILCOX E MARION, 2008) e cálculos da energia produzida P50 e P95 (EPE, 2017), fatores importantes para a viabilidade econômica de projetos. Além disso, a série pode ser utilizada para o preenchimento das lacunas de campanhas de medições públicas ou privadas (SCHWANDT ET AL., 2014), bem como em previsões da potência de saída de um empreendimento solar.

Nesse sentido, modelos regressivos ou modelos baseados em *quantile mapping* (QM) são largamente utilizados para adaptar as séries no período simultâneo entre as medições em terra e as saídas das bases de dados históricas. As séries utilizadas nesses métodos estatísticos podem incluir saídas de modelos de satélite (irradiâncias global horizontal, difusa horizontal e direta normal), saídas de NWP e outras variáveis como ângulo zenital, hora solar aparente, entre outras. As técnicas estatísticas podem ser aplicadas globalmente, sobre a série temporal como um todo, ou localmente, sobre subconjuntos das séries. Miranda et al. (2020) mostram que modelos regressivos aplicados globalmente para irradiância global horizontal (em inglês *Global Horizontal Irradiance*, GHI ou I_g) podem não apresentar melhoras significativas nos estatísticos quando comparados ao modelo de satélite para irradiância global horizontal a depender da região de estudo. De fato, existem situações em que a acurácia do modelo de satélite é tão alta para irradiância global horizontal que é difícil conseguir melhorar ainda mais essa acurácia a partir do uso de medições em terra (POLO ET AL., 2020). Contudo, pode ser possível obter melhoras significativas nos estatísticos aplicando modelos locais separados de acordo com uma classificação previamente estabelecida. Como não é possível separar esses subconjuntos no período em que não se tem as medições, pode-se treinar modelos para que as variáveis provenientes das bases históricas consigam acertar determinada classificação adotada.

Uma boa divisão das séries temporais de radiação pode ser feita buscando separar os momentos de céu claro, céu variável e céu nublado. Quando as séries fornecidas por modelos de satélite são utilizadas para fazer essa divisão, erros podem ser gerados, já que os modelos que utilizam imagens de satélite não estimam tão bem a atenuação da radiação pela cobertura de nuvens nos momentos nublados (HUANG ET AL., 2019). De fato, a cobertura de nuvens é um dos fatores que mais afetam a estimativa da radiação de modelos de satélite, já que os modelos de céu claro mais avançados (SUN ET AL., 2021) conseguem estimar bem o comportamento da radiação em condições de céu sem nuvens. Como a série medida na estação solarimétrica está sujeita a todas as variações que ocorrem, de fato, com a radiação solar em certa região, ela é mais confiável em termos de identificar as diferentes condições do céu no local. Por exemplo, métodos de classificação não supervisionada podem ser utilizados para definir as condições do céu em certa região com as medições feitas na estação solarimétrica. Em seguida, técnicas supervisionadas de

classificação podem ser empregadas para que as variáveis das bases de dados históricas aprendam a classificação estabelecida. A combinação de técnica não supervisionada e supervisionada de classificação já foi aplicada para resolver problemas de previsão de curto prazo da radiação solar (JIMÉNEZ-PÉREZ E MORA-LÓPEZ., 2016).

Este trabalho busca solucionar o problema de *site adaptation* a partir da utilização de modelos globais, locais (em subconjuntos das séries temporais) e combinação de modelos. Primeiro, será proposta uma classificação não supervisionada das séries de radiação para identificar as diferentes condições do céu na região, bem como o treinamento das variáveis provenientes das bases de dados históricas para acertar as classes previamente estabelecidas (classificação supervisionada). Após obter o modelo que irá classificar e, portanto, dividir os dados em seus respectivos subconjuntos, modelos estatísticos são utilizados de forma a adaptar a série de radiação proveniente de uma base de dados histórica ao local. Os modelos para correção das séries de radiação da base de dados histórica serão calibrados e validados no período simultâneo entre a campanha de medições e a base histórica. Procedimentos similares podem ser aplicados para outras variáveis importantes para a estimativa da potência gerada por sistemas heliotérmicos e fotovoltaicos tais como a velocidade do vento, temperatura, entre outras.

Assim, os objetivos deste trabalho são: realizar o *site adaptation* utilizando modelos globais, locais e de combinação na resolução temporal de 15 minutos; nos modelos locais, agrupar os dados de acordo com as condições do céu na região utilizando os dados da estação solarimétrica minuto a minuto (classificação não supervisionada); treinar um algoritmo de *machine learning* para que as variáveis regressoras provenientes das bases de dados históricas aprendam a classificação feita na etapa anterior, de forma que seja possível obter as diferentes condições do céu na região para os períodos no passado em que não existem dados da estação solarimétrica; aplicar os modelos regressivos para realização do *site adaptation* utilizando diferentes estratégias de pré-processamento dos dados de entrada e pós processamento das saídas de modelos; e, por fim, avaliar os resultados da metodologia utilizando diferentes estrátegias de divisão dos dados no conjuntos de calibração, validação e teste.

2 CONCEITOS PRELIMINARES

Nesta seção serão apresentados os conceitos preliminares das técnicas utilizadas neste trabalho. São apresentados os conceitos dos dois principais modelos estatísticos utilizados para o *site adaptation*: a regressão linear múltipla e as redes neurais do tipo *multilayer perceptron*. Os modelos de *site adaptation* podem ser pós processados com os modelos de *quantile mapping* e correção de BIAS e desvio padrão, também apresentados nesta seção. Em relação aos modelos de classificação, são apresentados os conceitos associados ao modelo de *k-means* (não supervisionado), árvores de decisão (supervisionado) e *random forest* (supervisionado). Por fim, é apresentada uma técnica de pré-processamento das variáveis regressoras aplicada nos modelos globais, a análise de componentes principais.

2.1 MODELOS REGRESSIVOS UTILIZADOS PARA O SITE ADAPTATION

Para o *site adaptation*, os modelos de regressão linear múltipla e redes neurais do tipo *multilayer perceptron* serão utilizados. A seguir são descritos esses dois grupos de modelos.

2.1.1 Regressão Linear Múltipla

Um dos métodos de regressão linear mais utilizados na relação entre variáveis meteorológicas é a regressão linear múltipla (*Multiple Linear Regression* – MLR) (WILKS, 2013). No MLR, preditores (e.g., dados provenientes de modelos de satélites ou dados atmosféricos de reanálise) são utilizados em uma estimativa linear de um determinado preditando (e.g., dados observacionais de radiação solar ou temperatura). A Eq. (1) descreve a regressão linear múltipla, sendo y_i o preditando, x_{ij} os preditores, α_j os coeficientes de regressão associados às variáveis regressoras (preditores) j e ε_i, o erro do modelo regressivo.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1j} \\ 1 & x_{21} & x_{22} & & x_{2j} \\ & \vdots & & \ddots & \vdots \\ 1 & x_{i1} & x_{i2} & \cdots & x_{ij} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_j \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \end{bmatrix} \quad (1)$$

Para os casos de regressão linear simples, o conjunto de preditores é composto somente por uma variável regressora (*j = 1*). Neste trabalho, para estimativas dos coeficientes α_j da regressão, é empregado o método dos mínimos quadrados, que busca minimizar o erro quadrático médio (*Mean Square Error* – MSE) entre a variável observada (y_i) e a variável estimada pelo modelo regressivo ($\hat{y}_i$). Na Eq. (2), é apresentado o MSE.

$$MSE = \sum_{n=1}^{T} (y_i - \hat{y}_i)^2 \tag{2}$$

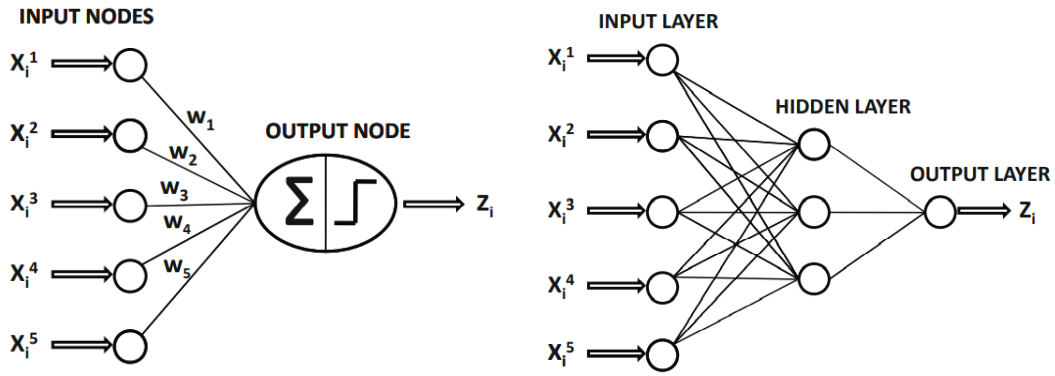
2.1.2 Redes Neurais do tipo Multilayer Perceptron

Uma rede neural pode ser descrita como uma rede de nós ou "neurônios" interconectados entre si utilizados para solucionar problemas de classificação e regressão. A rede neural mais simples é aquela composta por duas camadas: uma de entrada e uma de saída, o que é uma arquitetura conhecida como *perceptron* (Figura 1.a). Na rede do tipo *perceptron* há apenas um neurônio (o da camada de saída) que processa as informações da camada de entrada para fornecer o valor previsto na saída. Já a Figura 1.b apresenta um tipo de rede composta por uma camada de entrada (*input layer*) que se comunica com camadas ocultas (*hidden layers*), que, por sua vez, se conectam a uma camada de saída (*output layer*). Esse tipo de rede é denominado *multilayer perceptron*, já que pode possuir várias camadas ocultas formadas por diferentes quantidades de neurônios, sendo que as camadas posteriores não possuem uma saída ligada às camadas anteriores, ou seja, a interconexão entre os neurônios flui em uma direção sendo denominadas redes de fluxo unidirecional ou *feed-forward networks* (AGGARWAL ET AL, 2015).

Os dados da camada de entrada provenientes das variáveis regressoras são repassados para os neurônios da primeira camada oculta para que uma rede *multiplayer perceptron* inicie suas iterações. É no primeiro neurônio da primeira camada oculta que os cálculos matemáticos começam: um vetor de pesos W_j é inicializado randomicamente, onde *j* corresponde ao número de variáveis utilizadas na rede (variáveis $X_1, X_2, \dots, X_j$). O neurônio calcula, então, uma soma ponderada das entradas de acordo com os pesos estabelecidos, aplicando uma função de ativação à soma ponderada antes de repassar a informação para o neurônio de uma próxima

camada. Este processo é feito até chegar no neurônio da camada de saída, onde a estimativa final da rede deve ser comparada com o *target*, de forma a ajustar todos os pesos da rede com os dados de treinamento a partir de algoritmos de minimização de erro. Os pesos da rede são ajustados para cada vetor de valores *i* do conjunto de variáveis X_j, iterativamente. Após percorrer todo o conjunto de treinamento, o algoritmo pode começar novamente o processo iterativo naquele conjunto, só que, dessa vez, o primeiro vetor de valores *i* será computado não mais com os pesos W_j inicializados de forma randômica, mas com os pesos W_j estimados pela rede após percorrer uma vez o conjunto de treinamento. Quando a rede percorre todo o conjunto de treinamento uma vez significa que essa rede passou por uma época. Uma rede neural pode ser treinada, portanto, com várias épocas, onde em cada uma delas os dados do conjunto de treinamento devem passar pela rede para realizar um ajuste contínuo dos pesos W_j. Além disso, outro parâmetro que pode ser ajustado é a quantidade de inicializações randômicas dos pesos W_j feitas no algoritmo, de forma que se possa testar a convergência da rede para diferentes inicializações (HAN ET AL., 2012; DAWSON E WILBY, 2001; PERRUCI, 2018).

Figura 1 – Ilustração do modelo de redes neurais *feed-forward*.



(a) Redes neurais do tipo *perceptron*. (b) Redes do tipo *multiplayer perceptron*.
Fonte: Aggarwal et al. (2015).

2.2 MODELOS UTILIZADOS NO PÓS-PROCESSAMENTO DOS MODELOS REGRESSIVOS

Os modelos regressivos apresentados na seção anterior podem ser pós-processados utilizando o modelo de *quantile mapping* ou um modelo para correção

do BIAS (diferença entre a média da série estimada pela série observada) e desvio padrão. A combinação sequencial de modelo regressivo seguido por *quantile mapping* foi proposta por Fernández-Peruchena et al. (2020) na solução de problemas de *site adaptation*. Sendo assim, esses dois modelos para pós-processamento serão descritos a seguir.

2.2.1 Quantile Mapping (QM)

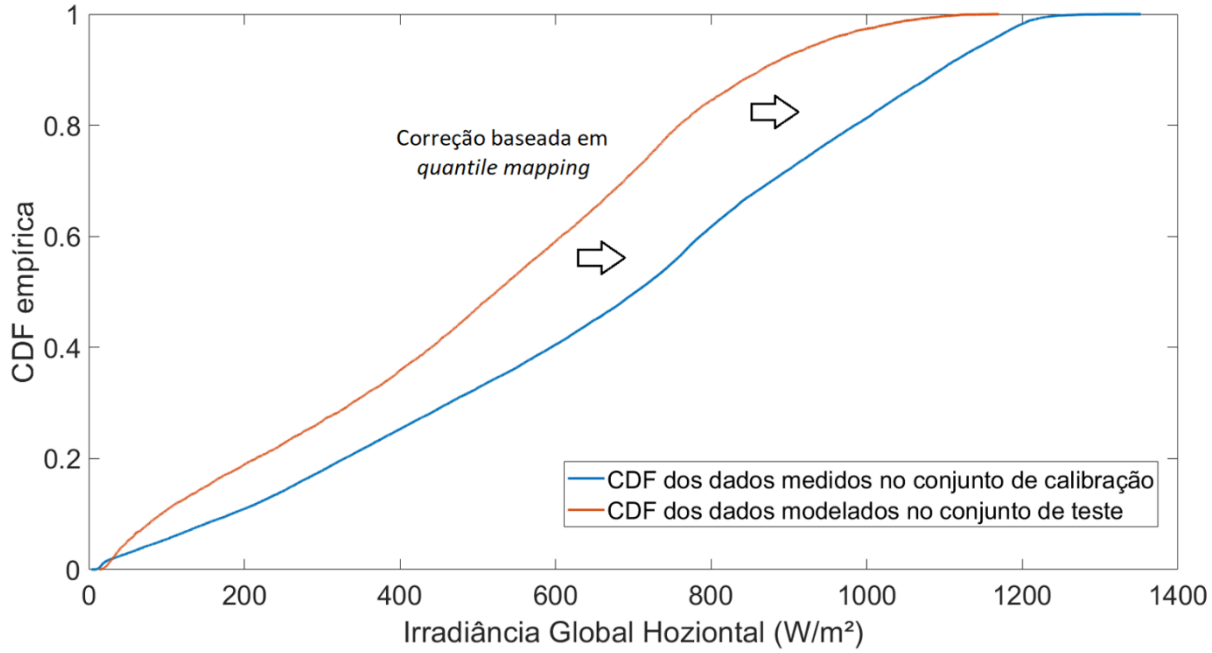
A técnica de QM aplica uma transformação inversa de forma a utilizar a curva de distribuição acumulada dos dados observados como referência para corrigir a curva dos dados modelados. A curva de distribuição acumulada (*Cumulative Distribution Functions* – CDFs) fornece a probabilidade de certa variável assumir valores iguais ou menores que determinado valor. A correção dos dados pode ser feita tanto utilizando funções paramétricas (CDFs) quanto funções empíricas (*Empirical Cumulative Density Distribution – ecdf*). Neste trabalho, a abordagem empírica é utilizada.

Para gerar a *ecdf*, as séries das variáveis envolvidas devem ser ordenadas do menor valor para o maior e deve ser estimada a densidade de probabilidade acumulada da série temporal. Considerando ***n*** o número de amostras da série temporal e ordenando os valores da variável de interesse do menor para o maior, a densidade de probabilidade acumulada pode ser obtida pela Eq. (3), onde ***p*** é um vetor de 1 até ***n***. Nas séries de radiação, a *ecdf* é formada ao ordenar os valores da irradiância global horizontal do menor para o maior (eixo *x* da Figura 2) em função da densidade de probabilidade obtida com a Eq. (3), conforme mostrado no eixo *y* da Figura 2. Para corrigir a distribuição dos dados modelados no conjunto de teste, basta então aplicar a transformação inversa usando a distribuição dos dados observados no conjunto de calibração (***obs***). Na Eq. (4), $\hat{y}_{Corr}$ é a série da variável modelada corrigida por QM no conjunto de teste. Na Figura 2, a CDF dos dados medidos e modelados é estimada empiricamente (*ecdf*) e a correção baseada em QM busca aproximar a curva em vermelho da distribuição de probabilidade acumulada da curva dos dados medidos, em azul.

$$ecdf = \frac{1}{n}p \qquad p = 1,2,3,\dots,n \tag{3}$$

$$\hat{y}_{Corr} = ecfd_{obs}^{-1}(ecfd_{mod}(\hat{y}_c)) \quad (4)$$

Figura 2 – Ilustração do modelo de *Quantile Mapping*.



Fonte: própria.

2.2.2 Correção de BIAS e desvio padrão

Outro modelo utilizado como pós-processamento dos modelos regressivos é a correção de BIAS (diferença entre as médias das séries estimada e observada) e desvio padrão entre o modelo e a observação, conforme Eq. (5). A média aritmética e o desvio padrão das séries temporais no conjunto de calibração são utilizadas para corrigir a série estimada no conjunto de teste. Na Eq. (5), $\hat{y}_{test}$ é a série estimada no conjunto de teste pelo modelo regressivo, $\overline{\hat{y}_{cal}}$ a média da série estimada pelo modelo regressivo no conjunto de calibração, $\sigma_{y,cal}$ e $\sigma_{\hat{y},cal}$ são os desvios das séries observada e estimada no conjunto de calibração, $\bar{y}_{cal}$ é a média da série de medições no conjunto calibração e, por fim, $\hat{y}_{cor}$ é a série observacional corrigida (MIRANDA et al., 2020). Essas correções são importantes para adequar as saídas dos modelos estatísticos ao desvio padrão e à média (BIAS) da série de dados observacionais.

$$\hat{y}_{cor} = (\hat{y}_{test} - \overline{\hat{y}_{cal}}) \left(\frac{\sigma_{y,cal}}{\sigma_{\hat{y},cal}} \right) + \overline{y_{cal}} \quad (5)$$

Onde:
$\hat{y}_{test}$: **dados modelados** no conjunto de teste (a série que vai receber a correção de BIAS e desvio)

$\overline{\hat{y}_{cal}}$: média dos **dados modelados** no conjunto de calibração
$\sigma_{y,cal}$: desvio padrão da série dos **dados observados** no conjunto de calibração
$\sigma_{\hat{y},cal}$: desvio padrão da série dos **dados modelados** no conjunto de calibração
$\overline{y_{cal}}$: média dos **dados observados** no conjunto de calibração
$\hat{y}_{cor}$: **dados modelados** no conjunto de teste corrigidos por BIAS e desvio

2.3 MODELOS DE CLASSIFICAÇÃO

A metodologia proposta neste trabalho para separação dos dados em subconjuntos envolve uma combinação de técnica não supervisionada de classificação seguida de uma técnica supervisionada. A seguir são descritos os modelos de classificação não supervisionada e supervisionada.

2.3.1 Algoritmo de k-means

O *clustering* é um procedimento utilizado para separar um conjunto de dados em subconjuntos, de acordo com funções objetivas que estimam a similaridade ou dissimilaridade entre os grupos ou *clusters* (HAN ET AL., 2012). É um método de aprendizado não supervisionado, em que o algoritmo aprende por observações ao invés de exemplos como nas técnicas supervisionadas. Uma das abordagens mais utilizadas em *clustering* envolve o particionamento dos dados, separando o conjunto inicial de dados em grupos específicos. Nessa abordagem, o número de *clusters* que se pretende atingir pode ser ou não definido previamente. Neste trabalho, o número de *clusters* (K) utilizado foi definido previamente como 5 classes, sendo esse parâmetro o ponto de partida do algoritmo de *k-means*.

Para exemplificar o algoritmo de *k-means* (MACQUEEN, 1967), pode-se partir de um conjunto de pontos em um espaço euclidiano de duas dimensões representado por $D = \{p_1, p_2, \dots, p_n\}$, onde $p = (x, y)$. O algoritmo de particionamento organiza cada ponto p em K *clusters* $C_1, C_2, \dots, C_K$. Uma função objetiva é utilizada para avaliar a similaridade entre os pontos de um *cluster*, de tal maneira que os pontos dentro de um mesmo *cluster* devem ser similares entre si e dissimilares dos pontos pertencentes a outros *clusters*. Cada agrupamento pode ser representado pelo seu centróide, isto é, um ponto central daquele grupo, como, por exemplo, a média dos pontos de determinado *cluster*. A distância de um ponto $p_j \in C_j$ ao seu centróide c_j, pode ser medida pela distância Euclidiana. Uma das funções mais utilizadas no *clustering*

baseado em particionamento dos dados para avaliar a similaridade entre os pontos é a soma dos erros quadráticos dos pontos em relação a certo centróide, conforme Eq. (6).

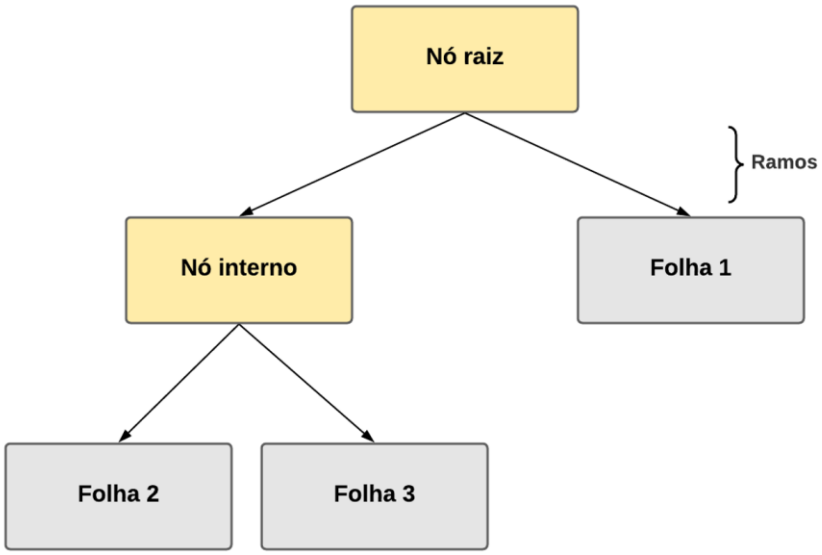
$$e^2(C, D) = \sum_{j=1}^{K} \sum_{i=1}^{n_j} \left\| p_i^{(j)} - c_j \right\|^2 \quad (6)$$

O *k-means* inicia, geralmente, com uma separação aleatória do conjunto de dados em K centróides. Tanto a quantidade K de centróides quanto os centróides c_j onde o algoritmo deve iniciar as suas iterações podem ser definidos a priori. O algoritmo vai reatribuindo iterativamente os agrupamentos, baseado na similaridade dos pontos de certo *cluster* ao seu centróide, até que um critério de convergência é atingido, como, por exemplo, quando a soma dos erros quadráticos não diminui consideravelmente após certo número de iterações. Vale salientar que um dos problemas dos algoritmos de *clustering* baseado no particionamento é que muitas vezes eles atingem um mínimo local da função utilizada como critério de avaliação ao invés do mínimo global, devido à sensibilidade do algoritmo às inicializações dos agrupamentos (JAIN ET AL., 2000).

2.3.2 Árvore de decisão

Modelos baseados em árvores de decisão são metodologias tipicamente utilizadas para solução de problemas de classificação. Em uma árvore de decisão, um conjunto de decisões hierárquicas é tomada, de tal forma que os *nós* de decisão formam uma estrutura similar a uma árvore, crescendo de cima para baixo, conforme ilustrado na Figura 3. A decisão feita em cada nó é baseada em um critério de divisão tomado conforme as variáveis disponíveis no conjunto de treinamento (AGGARWAL ET AL., 2015).

Figura 3 – Estrutura básica de uma árvore de decisão.



Fonte: Própria.

De forma geral, o algoritmo necessita iniciar em um nó raiz, definindo qual variável será utilizada, dentre todas as disponíveis, para dividir os dados. Em seguida, o processo de divisão continua em nós internos, dividindo os dados de acordo com as classes estabelecidas no *target* até que um critério de parada seja atingido. O critério de parada cria um nó final, chamado de folha, e a classe dominante naquele nó será a classe que o algoritmo irá estimar caso algum dado do conjunto de teste atinja essa determinada folha. Um critério de parada simples é, por exemplo, quando todos os dados do conjunto de treinamento de uma certa folha pertencerem a somente uma classe. Contudo, deixar a árvore atingir esse nível de especificação não é uma boa estratégia, já que o modelo vai estar muito especializado em estimar todas as classes do conjunto de treinamento, de tal forma que o modelo poderá não ter um bom desempenho para estimar as classes do conjunto de teste, situação conhecida como *overfitting*. Para evitar o *overfitting*, técnicas de podagem que definem um ponto onde a árvore deve parar o seu crescimento ou que elimine nós que acrescentem pouca informação adicional para divisão dos dados podem ser utilizadas (HAN ET AL., 2012).

O critério de divisão dos dados em determinado nó deve considerar tanto a seleção das diferentes variáveis disponíveis, quanto, nos casos em que as variáveis tenham valores contínuos, em qual ponto determinada variável será dividida para formar os ramos da árvore. Para definir a variável que será utilizada em um nó e como ela será dividida são utilizadas medidas que buscam selecionar o melhor critério de

divisão para um certo nó da árvore. O objetivo é obter o critério de divisão que melhor separe os dados de acordo com as classes estabelecidas no *target*. As duas métricas mais utilizadas são o índice gini e a entropia. O índice gini (*G*) quantifica a distribuição das classes em determinado conjunto de dados S, candidato a certo nó, e é calculado utilizando as frequências relativas $p_1, p_2, \dots, p_k$ associadas à ocorrência de cada classe no conjunto S (Eq. 7). Para quantificar a divisão de um certo conjunto S em *r* ramos (Eq. 8), é utilizada uma soma ponderada de acordo com as frequências relativas das divisões de classes ($|S|$ indica a quantidade de elementos no conjunto S e $|S_i|$ a quantidade de elementos no subconjunto S_i). De forma similar, a entropia (*E*) é definida para um certo conjunto S, conforme a Eq. (9), e para levar em consideração a divisão do conjunto S em *r* ramos, utiliza-se a Eq. (10). Tanto para entropia quanto para o índice gini, os menores valores indicam as melhores alternativas para dividir os dados.

$$G(S) = 1 - \sum_{j=1}^{k} p_j^2 \tag{7}$$

$$G(S \to S_1, \dots, S_r) = \sum_{i=1}^{r} \frac{|S_i|}{|S|} G(S_i) \tag{8}$$

$$E(S) = -\sum_{j=1}^{k} p_j log_2(p_j) \tag{9}$$

$$E(S \to S_1, \dots, S_r) = \sum_{i=1}^{r} \frac{|S_i|}{|S|} E(S_i) \tag{10}$$

A Tabela 1 apresenta um subconjunto dos dados utilizados neste trabalho para exemplificar a construção de uma árvore de decisão. O *target* é composto pelas classes 1, 2 e 3 que correspondem, respectivamente, às condições de céu claro, céu parcialmente claro e céu nublado. As variáveis utilizadas para classificar são o índice de cobertura de nuvens (*CloudCoverage*), a coluna total de vapor d'água (*tcwv*) e a irradiância difusa horizontal (*DHI*). Se o critério utilizado é o gini e a árvore construída é binária, ou seja, cada nó é subdividido em dois ramos, deve-se, primeiro, calcular os diferentes índices gini para uma certa variável de forma a obter o nó raiz. Como as variáveis são contínuas, para decidir o nó raiz deve-se, também, levar em consideração os dois ramos obtidos ao dividir a variável em valores maiores e menores que determinado valor estabelecido.

Tabela 1 – Parte do conjunto de dados utilizado neste trabalho para a localidade de El Rosal na Argentina.

Time	Classes	CloudCoverage (%)	tcwv (kg/m²)	DHI (W/m²)
05/01/2014 12:30	1	0,00	17,98	90,6
09/07/2014 14:30	1	0,00	4,52	61,2
11/02/2014 18:30	2	100,00	19,66	70,8
25/08/2015 11:45	1	0,00	3,88	64,2
24/04/2016 14:00	2	100,00	8,76	380,4
26/09/2016 09:15	1	0,00	4,46	58,8
20/12/2015 16:15	1	75,87	11,04	396,6
19/01/2016 17:45	1	54,87	14,75	292,2
20/05/2016 14:15	1	65,13	6,38	280,2
22/08/2016 17:00	1	0,00	3,52	37,8
31/07/2016 14:00	1	0,00	3,56	64,2
05/09/2016 14:15	1	0,00	1,77	52,8
01/06/2015 16:30	1	0,00	3,71	46,2
12/07/2016 12:00	1	0,00	2,18	43,8
19/12/2014 10:15	1	0,00	12,61	79,8
30/08/2014 17:00	3	100,00	6,00	96,6
03/02/2015 17:45	3	100,00	20,20	167,4
12/05/2016 09:30	1	47,47	8,98	143,4
18/02/2015 09:45	2	0,00	9,53	78
17/06/2015 10:15	2	100,00	5,70	172,8

Fonte: Própria.

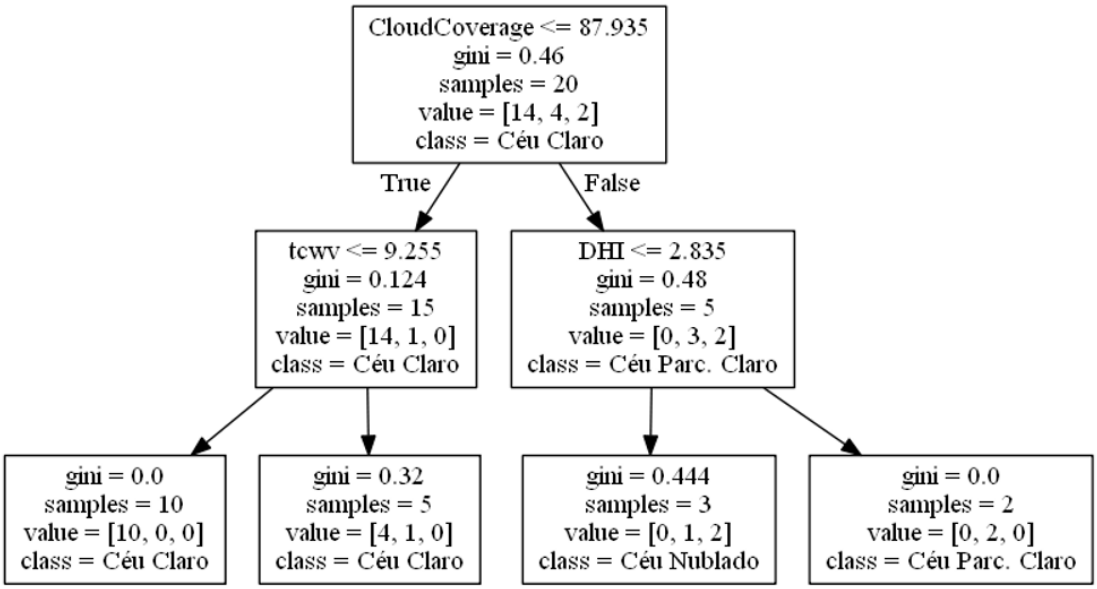
A Figura 4 apresenta uma árvore de decisão construída com os dados da Tabela 1. Note que o nó raiz é definido para valores da variável *CloudCoverage* menores que 87,9%, após o algoritmo testar diferentes critérios de divisão para todas as variáveis. Para o nó raiz, o índice gini do critério de divisão *CloudCoverage ≤ 87,9% e CloudCoverage > 87,9%* (Eq. 13) é obtido considerando a soma ponderada dos índices gini do ramo direito e do ramo esquerdo (Eqs. 11 e 12). Para cada divisão das diferentes variáveis testadas, o menor índice gini indica a melhor divisão dos dados. Após definir a variável e o ponto em que ela será dividida no nó raiz, deve-se fazer o mesmo para os nós internos, até que critérios de paradas específicos sejam atingidos. Após a árvore estar formada utilizando os dados do conjunto de treinamento, o modelo poderá estimar a classe de uma amostra do conjunto de teste ao percorrer essa nova amostra pelos nós da árvore até que chegue à determinada folha, na qual a classe dominante dela será a classe estimada para a nova amostra de dados. Vale salientar que o índice gini igual a 0,46 no nó raiz é calculado utilizando a Eq. (7) no *target* de classes (que possui 14 amostras para classe 1, 4 para classe 2 e 2 amostras para classe 3).

$$G(CloudCoverage \leq 87{,}9\%) = 1 - \left(\frac{14}{15}\right)^2 - \left(\frac{1}{15}\right)^2 - \left(\frac{0}{15}\right)^2 = 0{,}124 \quad (11)$$

$$G(CloudCoverage > 87{,}9\%) = 1 - \left(\frac{0}{5}\right)^2 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0{,}48 \quad (12)$$

$$G(CloudCoverage \rightarrow \leq 87{,}9\% \text{ e} > 87{,}9\%) = \left(\frac{15}{20}\right)0{,}124 + \left(\frac{5}{20}\right)0{,}48 = 0{,}213 \quad (13)$$

Figura 4 – Exemplo da construção de uma árvore de decisão.



Fonte: própria.

2.3.3 Random Forest

A estratégia de combinar diferentes estimativas de modelos de classificação previamente treinados é denominada como um modelo de *ensemble*. Modelos de *ensemble* utilizam diferentes classificadores ($M_1, M_2, \ldots, M_k$) treinados em diferentes subconjuntos do conjunto de treinamento ($T_1, T_2, \ldots, T_k$) com objetivo de obter um modelo de classificação combinado. Assim, dada uma certa amostra do conjunto de teste, a estimativa da classe que aquela amostra deve pertencer é feita para cada um dos M_k classificadores, e a classificação final é baseada na classe majoritária resultante dos modelos de classificação aplicados. Como o *ensemble* combina diferentes modelos de classificação, ele é menos propenso a causar *overfitting* do que a utilização de somente um modelo de classificação, como, por exemplo, uma árvore de decisão (HAN ET AL., 2012).

O *random forest* é um modelo de *ensemble* que considera a solução do problema de classificação por uma combinação de diferentes árvores de decisão, formando uma espécie de “floresta”. Cada árvore de decisão pode ser treinada em um subconjunto aleatório dos dados de treinamento e as variáveis selecionadas para formar a estrutura da árvore também são selecionadas de forma aleatória. Por exemplo, considere-se que sejam utilizadas 13 variáveis ou *features* em determinado problema de classificação. Considerando o subconjunto 1 dos dados de treinamento, pode-se selecionar de forma aleatória apenas 3 das 13 variáveis originais para a decisão de como dividir o nó raiz da árvore de decisão treinada nesse subconjunto, enquanto outras 3 variáveis, também escolhidas de forma aleatória dentre as 13 disponíveis, poderiam ser utilizadas para dividir o nó interno logo após o nó raiz, e assim sucessivamente. Geralmente, as árvores criadas nesses subconjuntos podem crescer até o máximo possível, sem a utilização de técnicas de podagem. A estratégia descrita é conhecida como seleção de entradas aleatória (*random input selection*) e é bastante utilizada nos modelos de *random forest*.

Os subconjuntos de treinamento para cada árvore de decisão também podem ser montados de forma aleatória a partir do conjunto de dados originais. Por exemplo, supondo hipoteticamente que o conjunto de treinamento possua 38000 amostras e que serão treinadas 100 árvores de decisão. Então, serão necessários 100 subconjuntos do conjunto de treinamento, um para cada árvore. Pode-se criar os 100 subconjuntos utilizando uma estratégia que considere reposição de amostras, ou seja, o subconjunto 1 pode conter 38000 amostras, que é a mesma quantidade do conjunto original, sendo que várias dessas amostras estão, na verdade, repetidas, o que exclui uma parte das amostras originais. Essa estratégia utilizada para criar os subconjuntos com reposição é conhecida como *bootstrap*. Ao combinar o *bootstrap* com a estratégia que considera a classe majoritária dentre todas as resultantes dos 100 modelos de árvores de decisão para estimar a classe de determinada amostra, tem-se a metodologia conhecida como *bootstrap agreggation* ou *bagging* (BREIMAN, 2001). O modelo de *random forest* pode utilizar todas as estratégias descritas acima. Vale salientar que os modelos de *random forest* são eficientes em reduzir a variância nos resultados finais e, também, são modelos que conseguem ter bom desempenho mesmo na presença de *outliers* ou ruídos provenientes das séries temporais (AGGARWAL ET AL., 2015).

2.4 PRÉ-PROCESSAMENTO DE VARIÁVEIS POR ANÁLISE DE COMPONENTES PRINCIPAIS

A Análise de Componentes Principais (*Principal Component Analysis* – PCA) é uma técnica de redução da dimensionalidade que pode ser aplicada a conjuntos de dados com muitas variáveis redundantes, ou seja, correlacionadas entre si. Para tanto, uma transformação linear é aplicada ao conjunto original dos dados, obtendo um novo conjunto, o das componentes principais, menor que o original e que explica boa parte da variância do conjunto original.

Tomando o conjunto original dos dados $X_{m\times n}$, composto por n variáveis de m valores ou medições cada, a PCA é uma rotação específica que transforma os m vetores de tamanho n em vetores rotacionados de uma base ordenada específica do espaço vetorial $\mathbb{R}^n$, obtendo uma matriz de componentes principais, $PC_{m\times n}$, composta por n componentes principais ordenadas, sendo as primeiras componentes as que explicam a maior parte da variância dos dados originais. Para encontrar essa base ótima da rotação, deve-se maximizar a variância associada aos eixos da base de rotação, de tal forma que o primeiro vetor da base vai estar associado à direção de máxima variância do conjunto de dados originais, o segundo vetor à segunda direção de maior variância e, assim, sucessivamente. Jollife (1986) demonstra como essa variância pode ser maximizada, obtendo, no final da demonstração, uma base ótima de rotação composta pelos autovetores associadas à matriz de covariância do conjunto dos dados, $\Sigma_{n\times n}$. Outro resultado interessante é que os autovalores (λ) de Σ estão associados à variância do conjunto original de dados, de tal forma que o autovetor associado ao maior λ é o vetor da base que está na direção de maior variância do conjunto, o segundo autovetor vai estar na direção da segunda maior variância da nuvem de pontos formada pelos dados e, assim, sucessivamente. A matriz de rotação $Rot_{n\times n}$ será composta, então, pelos autovetores da matriz de covariância ordenados de acordo com os respectivos autovalores, do maior para o menor. A Eq. (14) mostra a transformação do conjunto de dados originais para o conjunto das componentes principais pela matriz de rotação.

$$PC_{m\times n} = X_{m\times n} Rot_{n\times n} \tag{14}$$

A redução da dimensionalidade está justamente na escolha das k primeiras componentes principais (primeiras colunas da matriz $PC_{m\times n}$) que representem a maior variância do conjunto original de dados.

2.5 DIAGRAMA DE TAYLOR E AVALIAÇÃO ESTATÍSTICA DOS RESULTADOS

O diagrama de Taylor (TAYLOR, 2001) é uma ferramenta visual de avaliação estatística de resultados bastante útil. Os principais estatísticos associados ao diagrama são a correlação (Eq. 16), o desvio padrão (Eq. 15) e um estatístico proposto por Taylor, o *skill score* (SS4). A Figura 5 apresenta um dos diagramas de Taylor relacionados com este trabalho, como um exemplo. Os pontos coloridos no diagrama representam os modelos, enquanto o ponto em magenta localizado no eixo x representa os dados observacionais. A linha em magenta representa o desvio padrão das observações, que pode ser visto no eixo y como sendo 322,77 W/m². Quanto mais próximo os modelos (pontos coloridos) estiverem da linha em magenta, melhor eles representam o desvio padrão da observação. O coeficiente de correlação pode ser visto na posição azimutal; assim, quanto mais próximo os modelos estiverem do eixo x, maior a correlação.

As regiões R1 (cinza), R2 (amarelo) e R3 (magenta) são indicadores de desempenho dos modelos. Modelos dentro da região R1 (em cinza) apresentam uma correlação superior a 0,5 e estão mais próximos da linha em magenta, sendo, portanto, uma região em que os modelos podem apresentar bons resultados. Já os modelos dentro das regiões R2 (em amarelo) e R3 (em magenta) apresentam um desvio consideravelmente distante da observação ou correlações muito baixas, respectivamente, sendo, portanto, regiões em que os modelos não reproduzem o sinal observado. A razão entre os desvios é a razão entre o desvio padrão do modelo pelo desvio da observação, conforme mostrado na Eq. (17). O SS4 (Eq. 18) é representado pelas linhas em vermelho, partindo-se do ponto dos dados observados, em magenta; é um estatístico que indica o desempenho geral do modelo já que leva em consideração tanto o STDRatio, que está associada à amplitude de variação do modelo em relação à observação, quanto a correlação, que está associada à estrutura de fase e frequência entre os dois sinais comparados. Nas Eqs. de (15) a (18), x_i são os dados observados e $\bar{x}$ sua média, enquanto m_i são os dados estimados e $\overline{m}$ sua média.

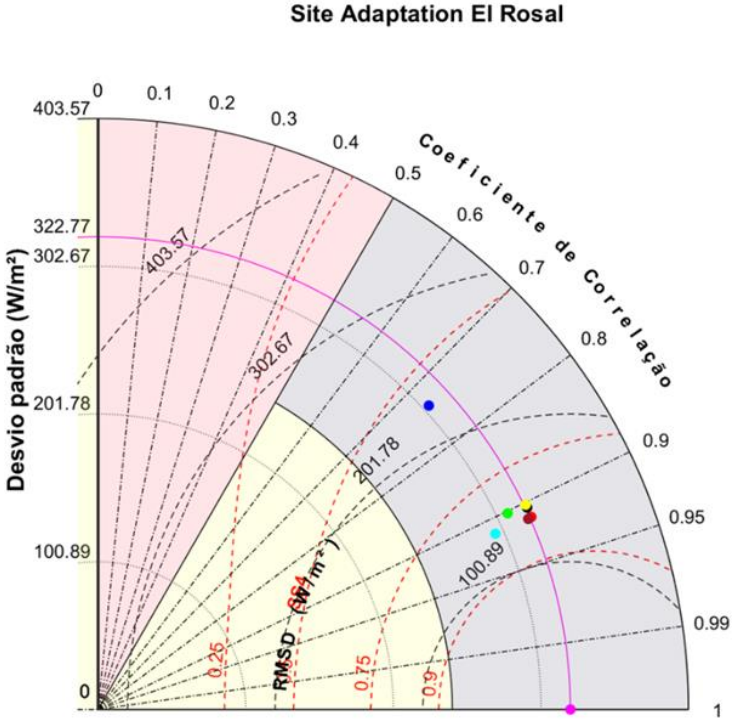
$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}} \tag{15}$$

$$\rho = \frac{\frac{1}{N}\sum_{i=1}^{N}(m_i - \bar{m})\,(x_i - \bar{x})}{\sigma_{mod}\sigma_{obs}} \tag{16}$$

$$STDRatio = \frac{\sigma_{mod}}{\sigma_{obs}} \tag{17}$$

$$SS4 = \frac{(1+\rho)^4}{4(STDRatio + 1/STDRatio)^2} \tag{18}$$

Figura 5 – Exemplo de um diagrama de Taylor.



Fonte: própria.

Os modelos são avaliados, também, com o erro médio (em inglês *Mean Bias Error* – MBE) ou BIAS e com a raiz do erro quadrático médio (em inglês, *Root Mean Square Error* – RMSE), bem como com ambos os estatísticos normalizados pela média da série observada (MBEn e RMSEn). As Eqs. (19) e (20) apresentam o MBE e o RMSE.

$$MBE = \bar{m} - \bar{x} \tag{19}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{N}^{i=1}(x_i - m_i)^2} \tag{20}$$

3 REVISÃO DE LITERATURA

Diversos modelos estatísticos podem ser utilizados para o *site adaptation* da radiação. Polo et al. (2020) testaram diferentes modelos de *site adaptation* com objetivo de obter um recurso solar de longo prazo mais acurado. As séries de longo prazo, são, no geral, tratadas na base horária. As técnicas analisadas incluem os modelos de regressão linear simples, *quantile mapping* (THEMEßL ET AL., 2012; MICHELANGELI ET AL., 2009), *Kernel density distribution mapping* (KDM), regressão linear múltipla, *sequential regressive-quantile mapping procedure* (SIM) e regressões utilizando subconjuntos dos dados. Alguns dos modelos, como KDM e SIM, foram aplicados a subconjuntos separados em dias de céu claro e dias de céu nublado. Dentre todos os modelos testados, aqueles com abordagem local, ou seja, que fazem a separação entre dias de céu claro e céu nublado, foram os que apresentaram os melhores resultados. Portanto, modelos de classificação para definição dos subconjuntos sobre os quais serão aplicados os modelos regressivos são uma possibilidade para aumentar a acurácia do *site adaptation*.

3.1 MODELOS DE CLASSIFICAÇÃO EM ENERGIA SOLAR

Os métodos de classificação são muito utilizados em diversas áreas do conhecimento para agrupamento de dados (*clusters*). Os métodos podem ser supervisionados, em que já há informações sobre os agrupamentos ou classes das variáveis e, portanto, treina-se um algoritmo para que os dados consigam acertar determinada classificação; ou não supervisionados, em que técnicas estatísticas são aplicadas ao conjunto de dados para definir os *clusters* ao qual cada dado deve pertencer (PÉREZ-ORTIS ET AL., 2016).

No âmbito da energia solar, as duas técnicas são largamente utilizadas. Muselli et al. (2000) classificam os dias da série de irradiância global horizontal (I_g) em claros, nublados e parcialmente nublados baseados em parâmetros obtidos a partir do perfil diário do índice de claridade (k_t – razão entre I_g e a irradiância extraterrestre horizontal, I_{oh}) horário utilizando o método de agregação de Ward. Moreno et al. (2017) utilizam a técnica de *k-medoids* para classificar os diferentes dias presentes em uma série da irradiância direta normal (em inglês, *Direct Normal Irradiance,* DNI

ou I_{bn}). Para tanto utilizam diferentes índices: o índice de variabilidade (VI), proposto por Stein et al. (2012), que busca quantificar a variabilidade da radiação solar em determinada janela temporal; a probabilidade diária da persistência da transmitância normal (k_b – razão entre I_{bn} pela irradiância extraterrestre efetiva, I_{oeff}), chamado índice POP_D; e o índice de fração da manhã, F_m, que determina se a radiação está concentrada na primeira ou segunda parte do dia. Eles estabelecem 7 tipos de dias, e validam os *clusters* utilizando a avaliação de especialistas da área.

Soubdhan et al. (2009) empregam dados de irradiância global horizontal medidos na ilha de Guadaloupe (região tropical) para classificar os diferentes dias presentes na série. Os autores utilizam distribuições diárias de probabilidade do índice de claridade medido segundo a segundo para classificar os diferentes tipos de dias, baseado nas associações das diferentes distribuições diárias de k_t estimadas por uma mistura finita de distribuições Dirichlet (EMILION, 2002); o algoritmo de *Stochastic Approximation Expectation-Maximization* (SAEM) é utilizado para fazer a classificação das distribuições diárias de k_t. Gastón-Romeo et al. (2011) utilizam uma ideia similar à de Soubdhan et al. (2009), mas empregam técnicas de matemática morfológica para calcular distribuições granulométricas (gsd – *granulometric size distribution*) dos dados de I_{bn} para certo dia ao invés de utilizar a distribuição diária de k_b ou k_t. Os autores aplicam então o modelo de *Partitioning Around Medoids* (PAM) para classificar as distribuições granulométricas calculadas; os dados utilizados de radiação direta normal, difusa e global horizontal integralizados em 10 minutos são provenientes de uma planta heliotérmica do tipo torre solar, situada no sul da Espanha. Fortuna et al. (2016) aplicam a técnica de *fuzzy-c means* em dois índices diários estimados a partir das séries de radiação, o índice de razão de área (A_r), definido como a razão entre a área do dia solar medido e a área de um dia típico da região; e a intermitência (I), calculado a partir das diferenças das potências de densidade espectral (domínio da frequência) do dia medido e do dia típico. Todas as técnicas de classificação descritas nos artigos acima fazem a classificação não supervisionada de séries de radiação; ou seja, classificam diferentes momentos da série em, por exemplo, céu claro, céu variável, céu nublado, entre outras classes utilizadas.

Dentre as técnicas de classificação supervisionada, Calbó et al. (2001) usam diferentes parâmetros obtidos a partir de dados horários de radiação global e difusa horizontais e observações visuais da nebulosidade para classificar as diferentes horas

em 9 ou 5 classes. O método da máxima verossimilhança (*maximum-likelihood*) é aplicado para classificar os dados, sendo que o mesmo conjunto de dados foi utilizado para treinar e validar o método. Jiménez-Pérez e Mora-López (2016) utilizam uma combinação de técnica não supervisionada e técnica supervisionada para fazer a previsão da irradiância global horizontal de todas as horas de um próximo dia. Os autores utilizam primeiro o modelo de *k-means* para agrupar os dados em 4 *clusters*: dia claro, dia nublado, dia nublado sobretudo durante a manhã e dia nublado sobretudo durante a tarde. Após agrupar os dados, os autores utilizam técnicas supervisionadas de classificação (árvores de decisão e *Support Vector Machines for Classification* – SVM-C) para estimar qual *cluster*, dentre os 4 citados, atribuir ao próximo dia. Alimohammadi e He (2016) também empregam uma combinação de técnica não supervisionada para classificar os dados de irradiância global horizontal baseado na aplicação de *Gaussian Mixture Models* (GMM); seguida de uma técnica supervisionada, o método do *k nearest neighbor* (k-NN), em que as variáveis exógenas e saídas de um *Numerical Weather Prediction* são treinados para acertar a classificação previamente estabelecida.

Por fim, há autores que fazem a classificação baseado em valores limiares de parâmetros específicos calculados (*threshold values*), ou seja, os autores classificam sem utilizar nenhuma técnica estatística específica, mas apenas o conhecimento empírico dos parâmetros calculados (DUCHON E O'MALLEY, 1999; TRUEBLOOD ET AL., 2013; HARROUNI E MAAFI, 2004; LI E LAM, 2001; RENO E HANSEN, 2016; ASSUNÇÃO ET AL., 2006; KANG E TAM, 2013).

A Tabela 2 apresenta um resumo das classificações feitas em energia solar. Os modelos de classificação que utilizam valores *threshold* são inclusos no grupo de técnicas não supervisionadas como modelos determinísticos, enquanto que os modelos estatísticos utilizados para separação das classes, como por exemplo o *k-means*, são modelos não determinísticos. Já as técnicas supervisionadas empregam modelos de aprendizado de máquina (*machine learning*) para que as variáveis regressoras utilizadas aprendam certo *target*. As classes que compõe o *target* podem ser obtidas por observações das condições do céu no local, imagens do estado do céu ou técnicas de classificação não supervisionada previamente aplicadas ao conjunto de dados.

Tabela 2 – Revisão bibliográfica sobre métodos de classificação em energia solar.

Classificação Não Supervisionada				
Autores	Resolução Temporal	Parâmetros utilizados	Número de classes	Método de Agrupamento
Muselli et al. (2000)*	1 hora	k_t	3	Método de Agregação de Ward
Moreno et al. (2017)*	1 min	VI, F_m, POP_D	7	*k-medoids*
Soubdhan et al. (2009)*	1 s	Distribuições de Dirichlet	4	*Stochastic Approximation Expectation-Maximization*
Gastón-Romeo et al. (2011)*	10 min	gsd	4	*Partition Around Medoids*
Fortuna et al. (2016)*	5 min	A_r, I	4	*Fuzzy c-means*
Alimohammadi e He (2016)**	10 min	GMM	3	*Expectation-Maximization*
Jiménez Pérez e Mora-Lopéz (2016)*	1 min	$k^*_{h,d}$	4	*k-means*
Chicco et al. (2014)*	1 min	$I_g, I_{g,ClearSky}$ (normalizadas)	4	*k-means*
Duchon e O'Malley (1999)***	1 min	Desvios de $I_g, I_{g,ClearSky}$	7	Empírico
Harrouni e Maafi (2004)*	1 min, 10 min	D, k_t	3	Empírico
Reno e Hansen (2016)***	1 min	5 parâmetros derivados de I_g	2	Empírico
Trueblood et al. (2013)*	1 min	k_c, VI	5	Empírico
Li e Lam (2001)**	1 hora	k_t, k_d, CLD, SH	3	Empírico
Assunção et al. (2006)*;**	5 min	k_t, m_a	4	Empírico
Kang e Tam (2013)*	1 min	k_t, POP_D	10	Empírico
Classificação Supervisionada				
Autores	Resolução Temporal	Parâmetros utilizados	Número de classes	Modelos de Classificação
Calbó et al. (2001)**	1h	k_t, k_d, k_{tn} + 10 parâmetros derivados	5 e 9	*Maximum-likelihood*
Pagès et al. (2003)**	30 min	I_g, T, H	4	*Maximum-likelihood*
Alimohammadi e He (2016)**	1 h	GMM	3	k-NN
Jiménez-Pérez e Mora-Lopéz (2016)*	1 h	k_t, T, H, P	4	Árvores de Decisão, SVM-C

Escala das classificações: diária (*), horária (**), menor que horária (***). Fonte: própria.

3.2 MODELOS DE SITE ADAPTATION

Segundo Polo et al. (2020), o *site adaptatio*n é um nome genérico para descrever o procedimento de correção de séries de longo prazo de radiação (no geral, mais que 10 anos), geralmente oriundas de modelos baseados em imagens de satélite, utilizando, para tal, séries medidas de boa qualidade em um curto período de tempo (no geral, 2 a 3 anos). Diversos modelos estatísticos já foram e estão sendo propostos para a adaptação da radiação ao local, destacando-se duas principais abordagens: modelos regressivos e modelos baseados em *quantile mapping* (QM). Há autores que propõe também a adaptação das séries a partir de um ajuste do BIAS (diferença entre a média da série medida e modelada), metodologia mais simples conhecida como método da razão e baseada no cálculo da razão entre as radiações medida e do modelo de satélite (ou reanálise) no período concomitante entre elas. Essa razão é, então, aplicada à série completa da radiação modelada (CEBECAUER E SURI, 2016; GUEYMARD ET AL., 2009; RUSCHEL E PONTE, 2018). Suri et al. (2010) sugerem, também, aplicar o método da razão seguido de uma correção da curva de distribuição acumulada da série de radiação do satélite para que ela se adeque à curva de distribuição das frequências acumuladas da série medida. O método da correção da curva de frequência acumulada está dentro dos modelos baseados em *quantile mapping* (QM). A abordagem dos modelos baseados em QM utiliza as probabilidades acumuladas (quantis) das séries de radiação medida e modelada e aplica uma transformação inversa usando a curva de distribuição acumulada dos dados observados para corrigir a curva dos dados modelados; esse processo pode ser feito tanto utilizando as distribuições empíricas dos dados observados e modelados, quanto a partir de um ajuste dessas curvas à distribuições paramétricas (SCHUMANN ET AL., 2011; INES ET AL., 2006; THEMEßL ET AL., 2012; CANNON ET AL., 2015).

Já dentre os modelos regressivos, o mais simples é a regressão linear no período simultâneo entre as observações e o modelo, aplicando essa regressão a toda a série modelada para adaptá-la (AGUIAR ET AL., 2019). Gueymard et al. (2012) e Bender et al. (2011) utilizam uma regressão linear múltipla para remover o BIAS e ajustar a variância dos dados de radiação modelados; as variáveis regressoras utilizadas pelos autores incluem as saídas dos modelos baseados em imagens de satélite ou reanálise, bem como outras variáveis meteorológicas oriundas de NWP.

Os autores classificam a técnica utilizada como um *Model Output Statistics* (MOS). Vernay et al. (2013) propuseram utilizar uma transformada de Fourier para determinar as frequências dominantes do erro entre o índice de céu claro diário do modelo baseado em imagens de satélite e da estação solarimétrica. Essas frequências dominantes são usadas, então, como variáveis regressoras de uma regressão linear múltipla. Mieslinger et al. (2014) concluíram, após comparar os dados provenientes de modelos de satélite com as observações, que baixas irradiâncias são, geralmente, sobre-estimadas, enquanto altas irradiâncias são subestimadas pelos modelos de satélite. Por isso, os autores propõem utilizar uma função polinomial do terceiro grau ao invés de uma função linear para fazer o *site adaptation*.

Fernández-Peruchena et al. (2020) aplicam duas abordagens de forma sequencial: regressões lineares múltiplas seguidas pelo QM. Em uma etapa de pré-processamento, uma regressão linear múltipla é feita utilizando o melhor conjunto de variáveis regressoras para o local em estudo. Para encontrar esse melhor conjunto, são comparados e ranqueados uma exaustiva lista de modelos lineares (*generalized linear models* – glm) utilizando o critério de informação de Akaike (*Akaike information criterion* – AIC). Em seguida, uma correção baseada no *quantile mapping* é aplicada à variável estimada pela regressão. É importante aplicar o procedimento nessa ordem, já que, de outra maneira, os resultados podem não ser bons, especialmente para irradiância direta normal. Finalmente, um pós-processamento é feito para evitar inconsistências nas séries adaptadas. Miranda et al. (2022) propõem utilizar as variáveis regressoras fornecidas pela CAMS, o que inclui séries de irradiância, irradiância de céu claro, profundidade ótica de aerossóis, profundidade ótica da nuvem, bem como variáveis meteorológicas do modelo ERA5-Land do ECMWF para realizar o *site adaptation* na resolução temporal de 15 minutos. Os autores propõem uma etapa de pré-processamento das variáveis empregando análise de componentes principais (PCA) e utilizam a regressão linear múltipla e redes neuras do tipo *multilayer perceptron* como modelos regressivos para o *site adaptation*, que podem ser pós-processados ou não com *quantile mapping*. Uma metodologia similar a de Miranda et al. (2022) é empregada neste trabalho para os modelos estatísticos globais.

Salamalikis et al. (2022), Narvaez et al. (2021) e Tiba et al. (2019) propõem utilizar técnicas de *machine learning* (ML) para realizar o *site adaptation*. Os autores testam modelos regressivos não lineares que, até então, haviam sido pouco citados na literatura na área de *site adaptation*. Tiba et al. (2019) testam os modelos de redes

neurais e *Support Vector Machine for Regression* (SVM-R) para realizar o *site adaptation*, comparando os resultados com modelos regressivos lineares; a técnica de SVM-R apresentou os melhores resultados estatísticos. Narvaez et al. (2021) implementam os modelos de regressão linear, redes neurais, *random forest* e AdaBoost e comparam os seus resultados com o *Quantile Mapping*. Os modelos não lineares de *random forest* e redes neurais apresentaram resultados mais acurados que o modelo de QM. Salamalikis et al. (2022) aplicam modelos de ML em subconjuntos de céu claro, céu nublado e céu intermediário. Esses subconjuntos foram definidos utilizando uma metodologia baseada em valores limiares do índice de claridade modificado (INEICHEN ET AL. 2009) definidos ao aplicar *Hidden Markov Models* (HMM). Para cada subconjunto local, os autores testam os modelos de ML de redes neurais, XGBoost, *random forests*, regressão linear do tipo *elastic net*, *Multivariate Adaptive Regression Spline* (MARS) e *Support Vector Regression* (SVR). A Tabela 3 lista algumas das publicações que abordam diferentes técnicas para fazer o *site adaptation*.

Tabela 3 – Revisão bibliográfica sobre técnicas de *site adaptation*.

Continua

Site Adaptation			
Autores	Variável Adaptada	Base de dados histórica utilizada	Modelos para o *Site Adaptation*
Polo et al. (2015)*	I_{bn}	*Meteosat Indian Ocean Data Coverage*	Regressão Linear
Tahir et al. (2020)**,***	I_g	NCEP-NCAR, NCEP-DOE, JRA-55 (reanálise) e NCEP-FNL, NCEP-GFS (análise)	Regressão Linear
Aguiar et al. (2019)*	I_g	CM-SAF (SARAH *database*)	Regressão Linear
Bender et al. (2011)*	I_{bn}	3TIER	MOS (regressão linear múltipla)
Gueymard et al. (2012)*	I_{bn}, I_g	3TIER	MOS (regressão linear múltipla)
Vernay et al. (2013)+	I_g	HelioClim-3	Transformada de Fourier + MLR

Tabela 3 – Revisão bibliográfica sobre técnicas de *site adaptation*.

Conclusão.

Site Adaptation			
Autores	Variável Adaptada	Base de dados histórica utilizada	Modelos para o S*ite Adaptation*
Schumann et al. (2011)*	I_{bn}, I_g	DLR-Solemi e EHF	*Quantile Mapping*
Polo et al. (2020)*	I_{bn}, I_g	Satélite/Reanálise (não especificado)	*Ensemble* de regressões lineares múltiplas, *Quantile Mapping, Kernel Density Distribution Mapping*
Mieslinger et al. (2014)*	I_{bn}	DLR-Solemi, EHF, HelioClim-3	Regressão com polinômio de terceiro grau
Miranda et al. (2022)'	I_g	CAMS, ERA5-Land (ECMWF)	MLR, Redes Neurais do tipo MLP, QM, PCA
Fernández-Peruchena et al. (2020)*	I_d, I_g	CAMS, NSRDB	*Generalized Linear Models* (GLM) + *Quantile Mapping*
Schumann et al. (2011)*	I_{bn}, I_g	DLR-Solemi e EHF	*Quantile Mapping*
Tiba et al. (2019)*	I_{bn}	Solargis	Regressão Linear, MLR, Redes Neurais e SVM-R
Narvaez et al. (2021)*	I_g	NSRDB	Regressão Linear, Redes Neurais, *Random Forests*, AdaBoost
Salamalikis et al. (2022)*	I_g	CAMS	Modelos de ML aplicados em subconjuntos locais: redes neurais, XGBoost, R*andom Forests*, regressão linear do tipo *elastic net*, MARS e SVR.

Site adaptation aplicado em dados com resolução temporal de 15 minutos ('), horária (*), de três horas (**), de seis horas (***), diária ($^+$). Fonte: própria.

3.3 ANÁLISE DAS METODOLOGIAS UTILIZADAS

Em síntese, há na literatura uma vasta gama de metodologias de classificação não supervisionada que visam definir subconjuntos de uma série de radiação medida, o que geralmente é feito em escala diária. Assim, a maioria dos métodos expostos na

Tabela 1 para classificação supervisionada empregam dados medidos de radiação em uma resolução temporal que varia de 1 segundo a 1 hora (sendo o intervalo mais utilizado o de 1 minuto), para classificar os dias da série temporal de radiação. Além disso, os dias são classificados na maioria dos trabalhos em 3 ou 4 classes, conforme indicado na Tabela 2. É possível perceber que poucas publicações abordam a classificação supervisionada para definição de subconjuntos das séries temporais de radiação. No geral, a classificação supervisionada pressupõe a utilização de observação humana ou imagens do céu para validação dos resultados e pode ser por isso que ela seja pouca empregada, já que a maioria das medições feitas em estações meteorológicas não possuem esse tipo de observação. Outra abordagem descrita em trabalhos de previsão da radiação solar é utilizar a classificação supervisionada para que as variáveis medidas aprendam o conjunto de classes definido previamente por uma classificação não supervisionada.

Para o *site adaptation*, as principais técnicas utilizadas são baseadas em modelos regressivos lineares, não lineares e *quantile mapping*. A realização do *site adaptation* utilizando modelos locais, baseado em subconjuntos das séries temporais, é citado, principalmente, nos artigos de Polo et al. (2020), Fernández-Peruchena et al. (2020) e Salamalikis et al. (2022). Alguns desses autores não explicitam, contudo, a metodologia utilizada para classificação, somente citando que o modelo de céu claro foi utilizado para dividir a série temporal em duas classes, dias de céu claro e dias de céu nublado, o que caracteriza uma classificação não supervisionada.

As metodologias que combinam classificação não supervisionada com classificação supervisionada para, em seguida, realizar o *site adaptation* aplicando modelos regressivos ou modelos baseados em *quantile mapping* para cada subconjunto ainda foram pouco exploradas. Tal metodologia pode ser encontrada em trabalhos de previsão da radiação solar, mas não aplicada para realização do *site adaptation*. Assim, a metodologia proposta neste trabalho pretende testar diferentes combinações de classificações não supervisionada e supervisionada para realização de um *site adaptation*. Vale salientar que os resultados encontrados poderão ser aplicados também como metodologias para previsão da radiação solar.

4 METODOLOGIA

Considerando o estado da arte no que se refere ao *site adaptation* das variáveis associadas à energia solar, a metodologia deste trabalho propõe a realização de um *site adaptation* em 8 etapas:

i) classificação não supervisionada dos dados disponíveis da estação solarimétrica, identificando as condições do céu na região;
ii) divisão dos dados provenientes das bases históricas nos conjuntos de treinamento, validação e teste, utilizando diferentes estratégias para dividir os dados nos três conjuntos;
iii) aplicação de modelos estatísticos para solução do problema de classificação supervisionada que utiliza como *target* as classes definidas no item *i)*;
iv) seleção e/ou extração de variáveis das bases de dados históricas utilizando somente os dados pertencentes ao conjunto de treinamento;
v) aplicação dos modelos regressivos para realização do *site adaptation* de forma global e de forma local;
vi) avaliação dos resultados por classes no conjunto de validação para escolha dos melhores modelos e treinamento dos modelos de combinação, também no conjunto de validação;
vii) aplicação dos modelos no conjunto de teste;
viii) avaliação dos resultados.

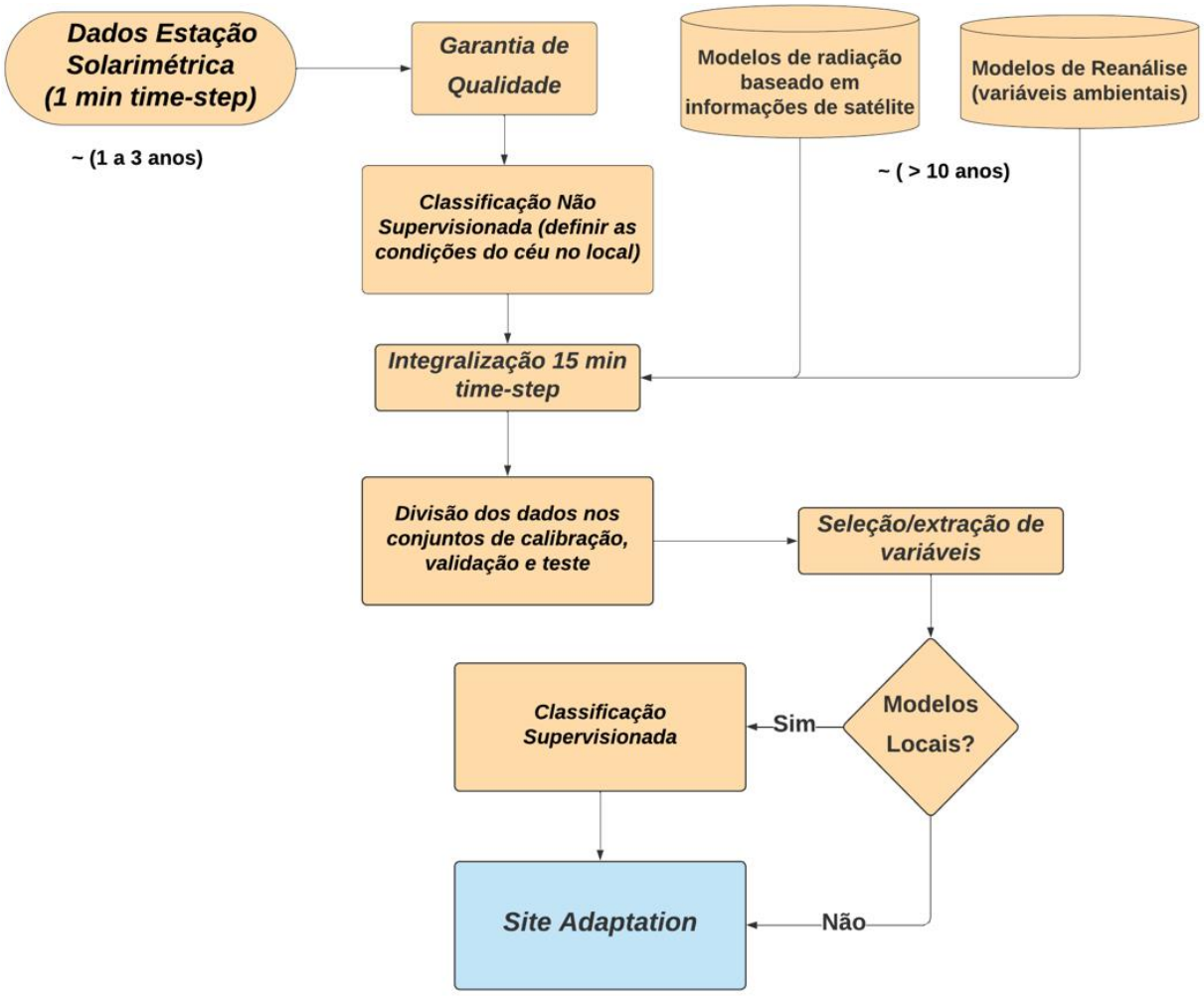
Neste trabalho, a base de dados histórica escolhida para radiação foi a da *Copernicus Atmosphere Monitoring Service (CAMS Radiation Service)*. Salazar et al. (2019) valida a base de dados da CAMS para a cidade de Petrolina, Pernambuco, e afirma que, dentre as 11 bases de dados de radiação solar analisadas, a da CAMS é a mais consistente para utilização na região. Como os dados empregados neste trabalho englobam uma estação solarimétrica no município de Petrolina – PE, escolheu-se a base de dados da CAMS para realização do *site adaptation*. A CAMS fornece séries temporais de irradiância global horizontal (em inglês, *Global Horizontal Irradiance* – GHI), difusa horizontal (em inglês, *Diffuse Horizontal Irradiance* – DHI ou I_d), direta normal (em inglês, *Direct Normal Irradiance* – DNI) e as respectivas séries de irradiância em condições de céu claro no período de fevereiro de 2004 até dois

dias anteriores à data atual, em escalas temporais de minutos, 15 minutos e horárias. A resolução espacial é interpolada para o ponto de interesse. A base de dados da CAMS é fornecida pelo serviço atmosférico *Copernicus*, um programa europeu para desenvolver a capacidade de observação europeia do planeta com respeito ao monitoramento terrestre, marinho e atmosférico, planejamentos de emergência, segurança e mudanças climáticas. É utilizado um modelo físico que utiliza aproximações do modelo de transferência radiativa, o método do Heliosat-4 (QU ET AL., 2017). O método é baseado no modelo *McClear* (LEFÈVRE ET AL., 2013) para cálculo da radiação em condições de céu claro e no modelo *McCloud* (SCHROEDTER-HOMSCHEIDT ET AL., 2019) para calcular a atenuação da radiação devido à presença de nuvens. As principais entradas do método Heliosat-4 são propriedades dos aerossóis, coluna total de vapor d'água e coluna total de ozônio, que são fornecidos pelo serviço global da CAMS com uma resolução temporal de 3 horas e resolução espacial de 0,8° (até 20 de Junho de 2016) e 0,4° (desde 21 de Junho de 2016). Propriedades das nuvens são obtidas a partir da segunda geração dos satélites METEOSAT, que fornece imagens a cada 15 minutos em uma resolução espacial de 3 km a 10 km.

Os dados medidos de radiação utilizados nesse trabalho são inicialmente qualificados por um procedimento de garantia de qualidade, conforme descrito em Petribú et al. (2017), o que inclui os filtros de qualidade recomendados pela *Baseline Surface Radiation Network* (BSRN). Dados considerados como anômalos nos testes de qualidade não são utilizados nos modelos, sendo substituídos por *NaN* (*not a number*). Em seguida, as medições de GHI da estação solarimétrica na resolução temporal de minutos são utilizadas para classificar as condições do céu no local, utilizando uma técnica não supervisionada. Os dados são, então, integralizados para uma escala de 15 min, de acordo com os procedimentos descritos por Salazar et al. (2019) e Roesch et al. (2011). Com o conjunto de dados na resolução de 15 min, são feitas as diferentes divisões nos conjuntos de calibração, validação e teste antes da aplicação dos modelos regressivos. Os modelos de extração/seleção de variáveis são aplicados para definir os conjuntos de variáveis regressoras que serão utilizados nos problemas de regressão. Para aplicação dos modelos locais, uma técnica supervisionada é utilizada para que as variáveis dos bancos de dados históricos acertem a classificação previamente estabelecida na etapa não supervisionada. Por fim, os modelos regressivos são treinados tanto no caso dos modelos locais (nos

subconjuntos de saída da classificação supervisionada) quanto dos modelos globais utilizando o conjunto de calibração. Modelos de combinação são também treinados no conjunto de validação. Os modelos são validados por avaliação estatística. A Figura 6 resume a metodologia utilizada neste trabalho.

Figura 6 – Diagrama metodológico para realização do *site adaptation*.



Fonte: própria.

4.1 CLASSIFICAÇÃO NÃO SUPERVISIONADA

A metodologia proposta para classificação dos dados envolve uma combinação de técnica não supervisionada de classificação seguida de uma técnica supervisionada. Na técnica não supervisionada serão utilizados somente os dados medidos na estação solarimétrica na resolução de minutos de forma a classificar as condições do céu no local durante todo o período da série temporal. Para a

classificação não supervisionada, este trabalho utiliza o índice de céu claro (k_c) e o índice de variabilidade (*variability index* – VI) como principais parâmetros para classificação do céu na região em 5 condições: céu claro, céu parcialmente claro, céu nublado, céu parcialmente variável e céu variável. A Tabela 4 apresenta as respectivas classes para cada condição do céu.

Em seguida, é aplicado o algoritmo de *k-means* para separar o conjunto de dados nos 5 *clusters* especificados. O índice de céu claro (k_c) é calculado como a razão entre a GHI (I_g) e a GHI de céu claro ($I_{gClearSky}$). Já o índice de variabilidade (VI) é calculado conforme proposto por Stein et al. (2012) com adaptações (Eq. 21).

Tabela 4 – Classes e respectivas condições do céu.

Número da classe	Condição do céu
Classe 1	Céu Claro
Classe 2	Céu Parcialmente Claro
Classe 3	Céu Nublado
Classe 4	Céu Variável
Classe 5	Céu Parcialmente Variável

Fonte: própria.

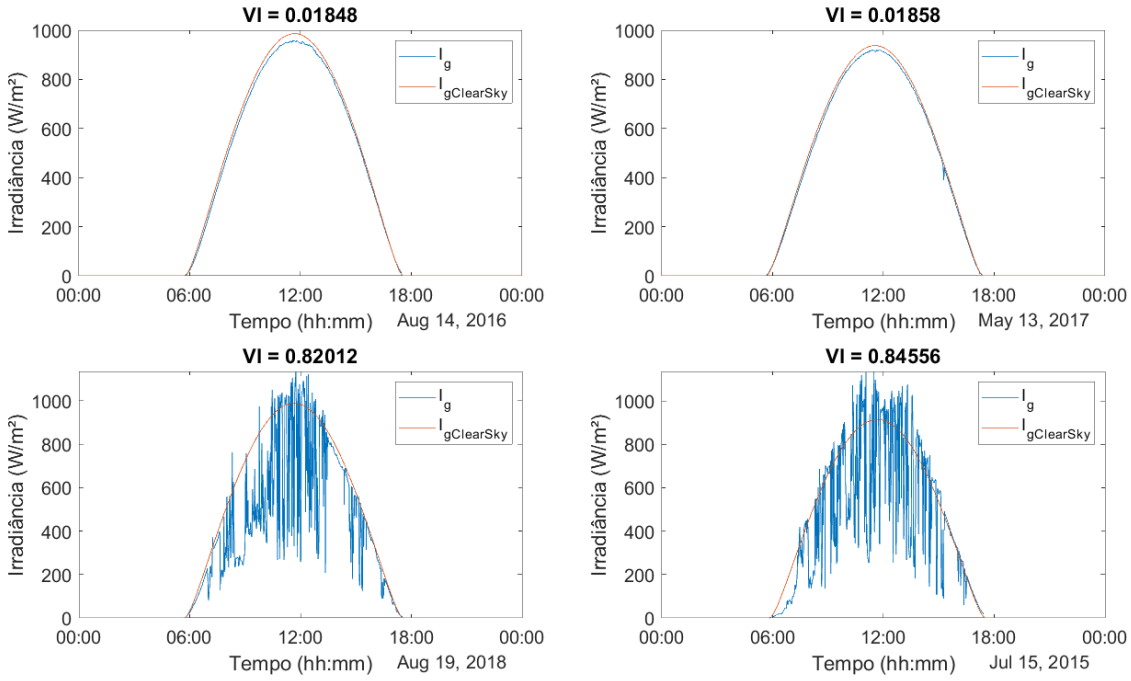
$$VI = \frac{\sum_{k=2}^{n} \sqrt{(I_{g,k} - I_{g,k-1})^2 + 1}}{\sum_{k=2}^{n} \sqrt{(I_{gClearSky,k} - I_{gClearSky,k-1})^2 + 1}} \tag{21}$$

O índice de variabilidade é o somatório das distâncias entre um ponto e outro de I_g dividido pelo mesmo somatório das distâncias para $I_{gClearSky}$. Vale salientar que a Eq. 21 deve ser aplicada com dados na resolução temporal de minutos. Por exemplo, para a irradiância global horizontal medida em minutos, caso a classificação seja feita em base diária, $n = 1440$ (1 dia); já no caso onde a classificação é feita em janelas temporais de 15 min, $n = 15$. O modelo de céu claro *McClear*, proposto por Lefèvre et al. (2013) e disponível na base de dados da CAMS, foi utilizado neste trabalho. Para facilitar a análise, o VI foi normalizado pelo seu valor máximo, assumindo valores entre 0 e 1. Em um dia de céu claro, assumindo que o modelo de céu claro esteja de acordo com as medições, e também em dias muito nublados, é esperado que VI seja próximo a zero. As duas condições, céu claro ou céu nublado durante todo o dia, são características de dias com baixa variabilidade. Já valores de VI próximo a 1 indicam dias com variabilidade elevada, como ilustrado na Figura 7.

Como o *site adaptation* será aplicado na escala de 15 min neste trabalho, as medições de irradiância global horizontal feitas a cada 1 minuto da estação

solarimétrica e o modelo de céu claro, também na escala dos minutos, são empregados para obter classificações de janelas de 15 minutos da série temporal. Assim, por exemplo, se às 08:45 de um certo dia, a classe resultante é a classe 2, significa que na janela de 08:31 até 08:45 o céu se manteve parcialmente claro na região analisada.

Figura 7 – Valores do índice de variabilidade diário e respectivos dias.



Fonte: própria.

De forma a considerar os momentos da série temporal em que as medições são mais confiáveis, não foram utilizados dados com elevação solar menor que 7º (períodos de amanhecer e entardecer), bem como, não foram utilizados dados cujo $k_c > 1{,}3$ (o que também corresponde a períodos de início e fim de dia). Além disso, como a dispersão $VI\ x\ k_c$ possui muitos pontos, já que é uma dispersão na escala de 15 minutos, alguns pontos podem ser previamente classificados, de acordo com o conhecimento empírico em relação às classes e às variáveis analisadas. Assim, os pontos do índice de variabilidade acima do percentil P95 são previamente classificados como céu variável (classe 4). Foram classificados previamente, também, alguns pontos como pertencentes à classe céu claro (classe 1), caso tais pontos satisfaçam as duas equações mostradas na Tabela 5 (*i* e *ii*) que consideram a média e o desvio padrão na janela de 15 min das séries de I_g e $I_{g,ClearSky}$.

Tabela 5 – Equações para classificação prévia de alguns pontos na classe céu claro (classe 1).

$i) \left\lvert \overline{I_g} - \overline{I_{g,ClearSky}} \right\rvert < 20\ W/m^2$	A diferença entre as médias das irradiâncias na janela temporal de 15 min deve ser menor que 20 W/m²
$ii) \left\lvert \sigma_{I_g} - \sigma_{I_{g,ClearSky}} \right\rvert < 2\ W/m^2$	A diferença entre os desvios padrões das irradiâncias na janela temporal de 15 min deve ser menor que 2 W/m²

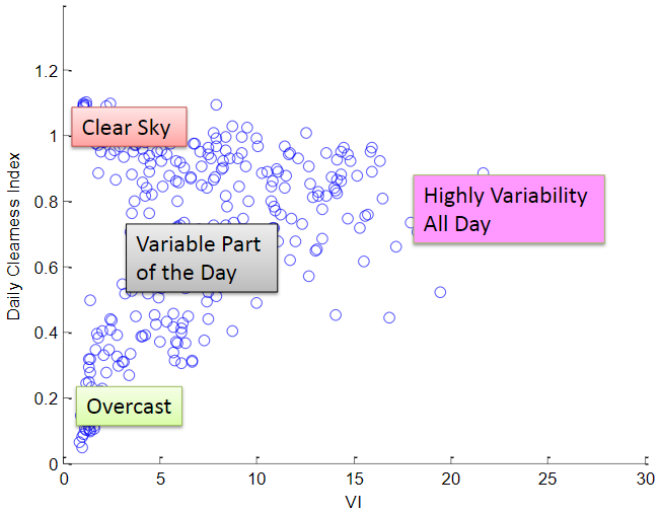
Fonte: própria.

Após classificar previamente alguns pontos da dispersão $VI\ x\ k_c$, pode-se utilizar o algoritmo de *k-means* para classificação objetiva do conjunto de dados restante. Para tanto, foi estimada a função de densidade de probabilidade conjunta de ocorrer um valor de k_c e um valor de VI, utilizando o método do *kernel density estimator* (KDE). Ao aplicar o KDE para estimar a probabilidade conjunta de duas variáveis ocorrerem, a curva resultante é uma superfície que descreve a função densidade de probabilidade para aquela dispersão (VARGAS, 2015; MAGENUKA ET AL., 2020). Estimar a função de densidade de probabilidade é útil para definir melhor os pontos de inicialização do algoritmo de *k-means*, já que um pico é formado na região associada às condições de céu claro. No modelo de *Kernel Density Estimation* (KDE), o *kernel* utilizado foi o gaussiano, sendo estimado o melhor desvio padrão para as distribuições normais (*bandwidth*).

Stein et al. (2012) sugerem a classificação subjetiva proposta na Figura 8, utilizando somente a dispersão $VI\ x\ k_c$. Os pontos com k_c baixo e VI baixo indicam uma condição de céu nublado, em que a radiação é baixa e não possui variações grandes. Já os pontos com k_c alto e VI baixo indicam céu claro, já que o VI baixo indica que a radiação não varia e o k_c alto indica que, naquela janela temporal, o modelo está próximo do modelo de céu claro. Uma condição intermediária de céu variável parcialmente ocorre no centro da dispersão, enquanto que os pontos mais à direita do gráfico, com VI elevado indicam condições de céu variável. Uma classificação similar será utilizada neste trabalho, com a diferença de que serão classificadas janelas temporais de 15 minutos e não diárias e será utilizado para classificação não somente os dois parâmetros citados, mas também a densidade de probabilidade conjunta de ocorrer os valores de VI e k_c. A Figura 9 mostra os 5 pontos iniciais do algoritmo de *k-means* (centróides iniciais), que foram escolhidos de forma subjetiva de acordo com o conhecimento prévio sobre a relação entre a dispersão $VI\ x\ k_c\ x\ KDE$ e as condições do céu no local. Os valores de VI normalizado próximos a zero e os valores de k_c elevados estão em uma região que apresenta uma alta probabilidade de ocorrência, ou seja, com maior KDE, o que forma um pico

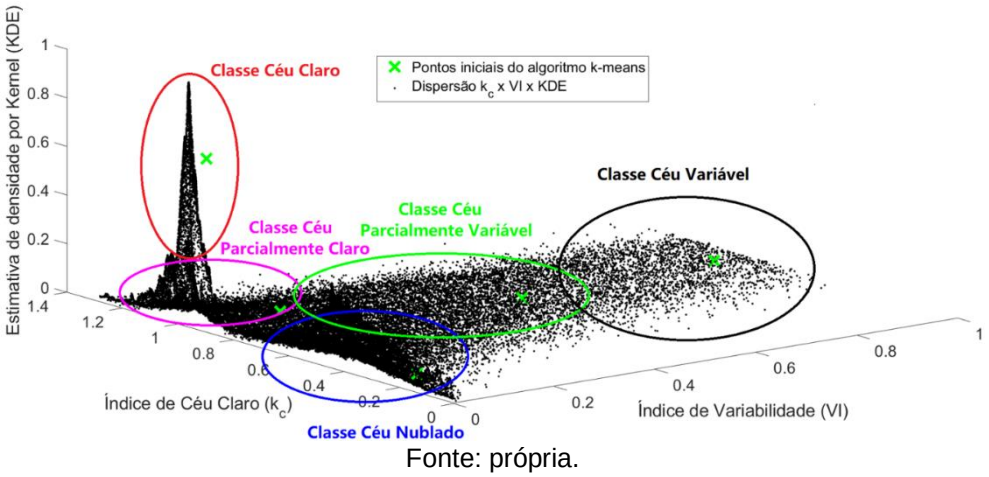
característico da condição de céu claro, conforme destacado em vermelho na Figura 9. Logo abaixo da classe céu claro, tem-se k_c elevado, VI próximo a zero e KDE baixo, o que foi considerado como uma classe de céu parcialmente claro (em magenta). O círculo azul indica as características da classe céu nublado (k_c baixo e VI próximo a zero). A região mais à direita é caracterizada por VI mais elevado, devendo pertencer à classe céu variável (destacado em preto). Por fim, uma condição intermediária entre as classes céu parcialmente claro, nublado e variável ocorre no meio da dispersão, o que foi considerado como a classe céu parcialmente variável (destacado em verde). Para os centróides iniciais, foi selecionado um ponto em cada uma das regiões da Figura 9, conforme destacado com um 'X' verde. Os centróides escolhidos são apresentados na Tabela 6. Foram utilizados os mesmos centróides iniciais para todas as estações solarimátricas utilizadas neste trabalho.

Figura 8 – Classes propostas por Stein et al. (2012) para a dispersão $VI\ x\ k_c$.



Fonte: Stein et al. (2012).

Figura 9 – Classes propostas neste trabalho para a dispersão $VI\ x\ k_c\ x\ KDE$.



Fonte: própria.

Tabela 6 – Centróides iniciais da dispersão $VI\ x\ k_c$ para inicialização de *k-means*.

Classe	Índice de céu claro (k_c)	Índice de variabilidade (VI)	*Kernel Density Estimation* (KDE)
Céu Claro	0,94	0,025	0,69
Céu Parcialmente Claro	0,912	0,156	0,025
Céu nublado	0,15	0,01	0,07
Céu variável	0,766	0,937	0,007
Céu parcialmente variável	0,735	0,539	0,005

Fonte: própria.

4.2 DIVISÃO DOS DADOS NOS CONJUNTOS DE CALIBRAÇÃO, VALIDAÇÃO E TESTE

Após a etapa de classificação não supervisionada, obtém-se um vetor de classes para cada *time-step* de 15 minutos da série temporal da estação solarimétrica. Para não depender das medições feitas na estação, que, geralmente, compreendem um período de medições curto (e.g. de 1 a 2 anos), propõem-se empregar variáveis provenientes de modelos de radiação baseados em imagens e informações de satélite e modelos de reanálise, de tal forma que essas variáveis possam aprender o comportamento das classes previamente definidas, permitindo fazer classificações nos vários anos em que a estação solarimétrica não fez medições na região. Assim, os dados da estação, o vetor de classes construído na etapa não supervisionada e os dados provenientes das bases de dados históricas, todos considerados no *time-step* de 15 minutos, devem ser divididos nos conjuntos de calibração, validação e teste para aplicação dos modelos de classificação supervisionada e *site adaptation*. Vale salientar que os mesmos conjuntos serão utilizados tanto na etapa de classificação quanto na de regressão.

De forma a avaliar a robustez dos modelos, foram definidas 62 configurações diferentes para dividir os dados nos conjuntos de calibração, validação e teste. Essas 62 diferentes divisões podem ser agrupadas em 4 principais categorias, conforme descrito na Tabela 7.

A categoria 1 (divisões de 1 a 10) mantém a sequência temporal dos dados, a categoria 2 (divisões de 11 a 22) faz uma intercalação entre os dias da série temporal, a categoria 3 (divisões de 23 a 42) faz a randomização em base diária dos dados antes das divisões nos três conjuntos, enquanto a categoria 4 (divisões de 43 a 62) faz a randomização ou embaralhamento dos *timestamps* de 15 min antes de dividir nos três conjuntos. A Tabela 8 apresenta os percentuais dos dados pertencentes aos

conjuntos de calibração, validação e teste para cada divisão. Por fim, a Tabela 9 apresenta as principais operações feitas em cada um dos três conjuntos.

Tabela 7 – Distintas divisões dos dados e a metodologia de cada grupo de divisão.

Diferentes divisões dos dados	Descrição das estratégias utilizadas para dividir os dados
1) Divisões de 1 a 10 (10 configurações)	Diferentes posições de calibração, validação e teste respeitando a sequência cronológica das séries temporais, definidas de acordo com uma janela móvel utilizando os percentuais de 70%, 10% e 20%, respectivamente
2) Divisões de 11 a 22 (22 configurações)	Intercalação dos dias para definição das amostras de calibração, validação e teste (por exemplo, os dois primeiros dias da série para calibração, o terceiro para validação e o quarto para teste e, assim, sucessivamente);
3) Divisões de 23 a 42 (20 configurações)	Embaralhamento randômico em base diária dos dados na resolução temporal de 15 minutos, com diferentes tamanhos para cada um dos três subconjuntos
4) Divisões de 43 a 62 (20 configurações)	Embaralhamento randômico dos timestamps de 15 minutos, com diferentes tamanhos para cada um dos três subconjuntos

Fonte: própria.

Tabela 8 – Percentuais dos dados para cada divisão nos conjuntos de calibração, validação e teste.

Div.	Cal.	Val.	Teste	Div.	Cal.	Val.	Teste	Div.	Cal.	Val.	Teste
1	70%	10%	20%	**21**	~50%	~25%	~25%	**41**	~20%	~40%	~40%
2	70%	10%	20%	**22**	~50%	~25%	~25%	**42**	~10%	~50%	~40%
3	70%	10%	20%	**23**	~80%	~10%	~10%	**43**	~80%	~10%	~10%
4	70%	10%	20%	**24**	~70%	~20%	~10%	**44**	~70%	~20%	~10%
5	70%	10%	20%	**25**	~60%	~30%	~10%	**45**	~60%	~30%	~10%
6	70%	10%	20%	**26**	~50%	~40%	~10%	**46**	~50%	~40%	~10%
7	70%	10%	20%	**27**	~40%	~50%	~10%	**47**	~40%	~50%	~10%
8	70%	10%	20%	**28**	~70%	~10%	~20%	**48**	~70%	~10%	~20%
9	70%	10%	20%	**29**	~60%	~20%	~20%	**49**	~60%	~20%	~20%
10	70%	10%	20%	**30**	~50%	~30%	~20%	**50**	~50%	~30%	~20%
11	~50%	~25%	~25%	**31**	~40%	~40%	~20%	**51**	~40%	~40%	~20%
12	~50%	~25%	~25%	**32**	~30%	~50%	~20%	**52**	~30%	~50%	~20%
13	~50%	~25%	~25%	**33**	~60%	~10%	~30%	**53**	~60%	~10%	~30%
14	~50%	~25%	~25%	**34**	~50%	~20%	~30%	**54**	~50%	~20%	~30%
15	~50%	~25%	~25%	**35**	~40%	~30%	~30%	**55**	~40%	~30%	~30%
16	~50%	~25%	~25%	**36**	~30%	~40%	~30%	**56**	~30%	~40%	~30%
17	~50%	~25%	~25%	**37**	~20%	~50%	~30%	**57**	~20%	~50%	~30%
18	~50%	~25%	~25%	**38**	~50%	~10%	~40%	**58**	~50%	~10%	~40%
19	~50%	~25%	~25%	**39**	~40%	~20%	~40%	**59**	~40%	~20%	~40%
20	~50%	~25%	~25%	**40**	~30%	~30%	~40%	**60**	~30%	~30%	~40%
-	-	-	-	-	-	-	-	**61**	~20%	~40%	~40%
-	-	-	-	-	-	-	-	**62**	~10%	~50%	~40%

Fonte: própria.

Tabela 9 – Descrição dos procedimentos adotados nas três divisões dos dados.

Conjunto Estatístico	Finalidade
Conjunto de calibração	Treinamento dos modelos regressivos e da classificação supervisionada
Conjunto de validação	Seleção dos melhores modelos locais e treinamento das combinações de modelos
Conjunto de teste	Aplicação de todos os modelos treinados e avaliação dos resultados

Fonte: própria.

4.3 CLASSIFICAÇÃO SUPERVISIONADA

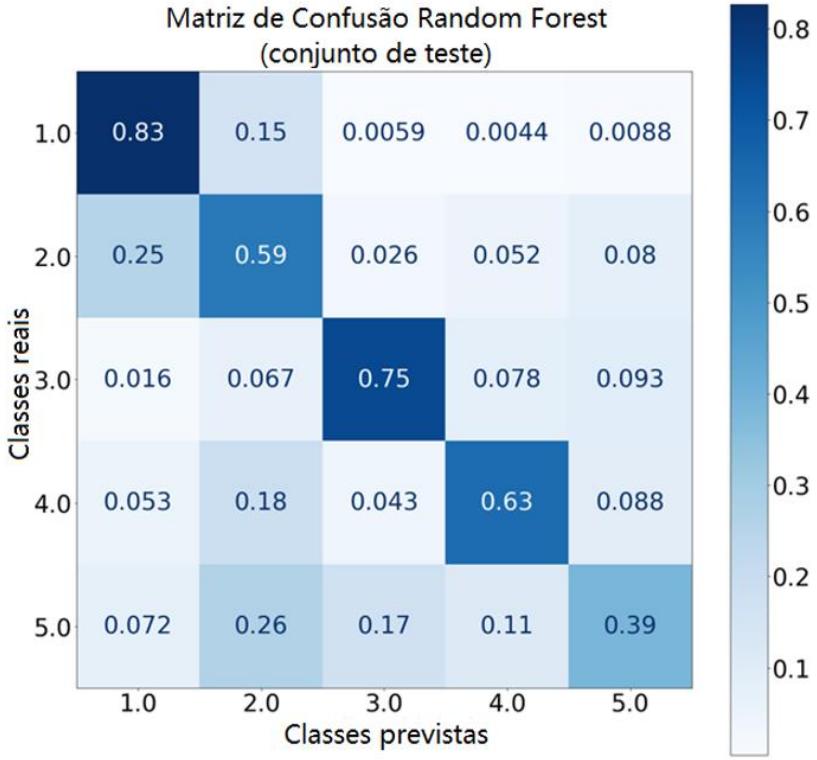
O problema de classificação supervisionada pode ser resolvido utilizando modelos de aprendizagem de máquina, desde os mais simples, como, por exemplo, regressão logística e árvores de decisão, até modelos mais complexos, como *Support Vector Machine* (para classificação) e *Random Forest*. Neste trabalho, o algoritmo empregado para classificação supervisionada é o *Random Forest*. São utilizadas as séries temporais oriundas de modelos de radiação baseado em informações de satélite ou NWP como entradas de um modelo de classificação, onde essas variáveis poderão aprender as classes definidas previamente. Para avaliar o modelo de classificação supervisionada serão analisadas a acurácia e a matriz de confusão. A acurácia (em inglês, *accuracy*) é calculada conforme a Eq. (22) e indica o desempenho geral do modelo, com valores entre 0 (pior acurácia) e 1 (melhor acurácia). Já a matriz de confusão é uma tabela que indica no eixo x o resultado do modelo (classe prevista) e no eixo y as classes observadas (classe real). Por exemplo, na Figura 10, analisando os resultados da classe 1 (classe real), é possível observar que o modelo acertou 83% dos casos dessa classe, contudo 15% das amostras foram consideradas erroneamente como pertencentes à classe 2, 0,59% à classe 3, 0,44% à classe 4 e 0,88% como pertencentes à classe 5. O modelo apresentou um bom desempenho para estimar a classe 1 nesse caso. Já no caso da classe 5, o modelo acerta apenas 39% dos casos, classificando erroneamente 26% das amostras como classe 2, apresentando, portanto, um desempenho ruim para essa classe.

$$Acc = \frac{\#Número\ de\ classificações\ corretas}{\#Número\ de\ amostras} \tag{22}$$

Para definir o melhor conjunto de hiperparâmetros no algoritmo de *Random Forest*, uma busca em grade (*grid search*) é realizada no conjunto de treinamento,

considerando os seguintes hiperparâmetros: a utilização ou não do *bootstrap* para criação dos subconjuntos com reposição, a profundidade máxima das árvores de decisão estimadas em cada subconjunto (‘*max_depth*’), a quantidade de *features* que deve ser testada em cada árvore (‘*max_features*’), o número mínimo de amostras que uma folha deve conter (‘*min_samples_split*’), o mínimo número de divisões de um nó interno (‘*min_samples_split*’) e a quantidade de árvores de decisão utilizadas (‘*n_estimators*’). Os valores dos hiperparâmetros utilizados no *grid search* para avaliar a melhor combinação são apresentados na Tabela 10.

Figura 10 – Exemplo de uma matriz de confusão.



Fonte: própria.

Tabela 10 – Hiperparâmetros de busca para os modelos de classificação supervisionada.

Modelo de Classificação	Busca pelos melhores hiperparâmetros
Random Forest	"bootstrap": [True, False], "max_depth": [10, 20, 30], "max_features": ["auto", "sqrt"], "min_samples_leaf": [2, 4], "min_samples_split": [2, 5], "n_estimators": [200, 500, 700]

Fonte: própria.

As variáveis regressoras utilizadas para aprender o comportamento das classes definidas na etapa supervisionada incluem todas as saídas de modelo

fornecidas pela CAMS Radiation Service (e.g. aerossóis, coluna total de vapor de água, irradiâncias, cobertura de nuvens, etc) e variáveis meteorológicas do modelo ERA-5 Land do ECMWF (temperatura de ponto de orvalho, velocidade do vento, pressão atmosférica, etc). As variáveis são utilizadas no modelo *random* forest sem uma etapa de pré-processamento. Todas as variáveis regressoras são utilizadas na resolução de 15 minutos, sendo, então, o *site adaptation* feito nessa escala temporal.

4.4 SELEÇÃO E EXTRAÇÃO DE VARIÁVEIS

A seleção e extração de variáveis é feita como uma etapa de pré-processamento para os modelos de *site adaptation*. Para selecionar/extrair as variáveis mais relevantes do domínio estudado podem ser utilizados tanto algoritmos de seleção de variáveis (*feature selection*), como filtros baseados em coeficientes estatísticos, quanto algoritmos de extração de variáveis, a exemplo de técnicas de redução de dimensionalidade como a análise de componentes principais.

Nos modelos de *site adaptation*, é importante selecionar as variáveis (ou *features*) que mais expliquem o *target*. Por isso, a seleção das variáveis é um dos primeiros passos para resolver de forma mais eficiente problemas de classificação/regressão. Os modelos de seleção podem ser divididos em três principais categorias (AGGARWAL ET AL., 2015):

- ***Filters models***: é utilizado algum estatístico ou parâmetro matemático para avaliar a qualidade das variáveis de forma individual ou de grupos de variáveis. Um critério ou valor *threshold* é adotado para filtrar variáveis irrelevantes na solução do problema;
- ***Wrapper models***: é utilizado algum algoritmo de classificação/regressão para avaliar os resultados de determinado modelo em diferentes grupos de variáveis. Um algoritmo de busca pelo melhor grupo de variáveis é, então, implementado, para determinar o melhor conjunto de variáveis que soluciona o problema. Esta metodologia pode demandar um alto esforço computacional, a depender do conjunto de *features* analisado;
- ***Embedded models***: a solução de alguns modelos de classificação/regressão (e.g. *lasso regression*, regressão logística utilizando regularizações L1 e L2) podem conter informação sobre as variáveis mais relevantes do problema

proposto. As variáveis mais relevantes são, então, selecionadas, de tal forma que o modelo é novamente treinado no novo conjunto de variáveis encontrado.

A técnica aqui utilizada para seleção de variáveis é um filtro baseado na correlação de *Pearson*. No geral, os algoritmos de seleção buscam os grupos de variáveis que possuam a mínima redundância entre elas e a máxima relevância das mesmas com o *target* (HALL, 1999; YU E LIU, 2004). A seleção de variáveis é feita neste trabalho em duas partes. Primeiro, para eliminar a redundância dos grupos de variáveis no conjunto de treinamento, o que inclui as *features* que tenham uma alta correlação entre si, é calculada a matriz de correlação das variáveis regressoras utilizadas. A matriz de correlação é uma matriz quadrada, simétrica e com todos os elementos da diagonal principal iguais a 1, já que a diagonal principal representa a correlação das variáveis com elas mesmas. A dimensão da matriz de correlação é a mesma que a quantidade de variáveis regressoras. Por exemplo, em um caso hipotético com 3 variáveis regressoras, a matriz de correlação deve ser montada conforme a Eq. (23), onde ρ faz referência à correlação de Pearson, enquanto v_j faz referência à j-ésima variável regressora (no exemplo, j = 1,2,3). O algoritmo analisa as correlações linha a linha, buscando por correlações superiores a 0,99. Caso em determinada linha seja encontrada uma correlação superior a 0,99, a respectiva coluna e, consequentemente, variável regressora, deve ser eliminada do conjunto de variáveis. Esse método simples retira variáveis redundantes entre si do conjunto de variáveis regressoras.

$$M_{corr} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} 1 & \rho(v_1, v_2) & \rho(v_1, v_3) \\ \rho(v_1, v_2) & 1 & \rho(v_2, v_3) \\ \rho(v_1, v_3) & \rho(v_2, v_3) & 1 \end{bmatrix} \quad (23)$$

Na segunda parte é avaliada a correlação entre as variáveis regressoras que ficaram após a primeira filtragem e o *target*. As variáveis regressoras que apresentarem correlação inferior a 0,2 são eliminadas do conjunto de dados. Vale salientar que caso todas as variáveis regressoras estejam com uma correlação com o *target* inferior a 0,2, o critério de correlação é diminuído para 0,1. Um estudo de sensibilidade foi feito para escolha dos valores adotados como critérios de seleção

nas duas partes. Por fim, as variáveis selecionadas são, então, utilizadas nos modelos regressivos.

Para a extração de variáveis é adotado o modelo de análise de componentes principais (PCA), sendo selecionadas as primeiras componentes principais que explicam mais do que 95% do conjunto original dos dados. Assim, os modelos regressivos são pré-processados neste trabalho tanto utilizando seleção com filtros pela correlação quanto análise de componentes principais.

4.5 SITE ADAPTATION

Para o *site adaptation*, os dois modelos regressivos principais utilizados são a regressão linear múltipla e a rede neural do tipo *multilayer perceptron*, podendo as saídas desses modelos serem pós-processadas através do modelo de *quantile mapping* e correção por BIAS e desvio. Nos modelos que envolvem redes neurais, a função de ativação utilizada é a tangente hiperbólica com algoritmo de minimização de erro de Levenberg-Marquardt (GAVIN ET AL. 2020). A arquitetura da rede *multilayer perceptron* considera três camadas ocultas (com 5 neurônios na primeira camada, 4 na segunda e 3 na terceira) e uma camada de saída. Por fim, a quantidade de épocas e inicializações utilizadas foram de 50 e 100, respectivamente.

A aplicação dos modelos regressivos para *site adaptation* podem ser agrupados em três principais categorias: os modelos globais, locais e as combinações de modelos. Os modelos globais são treinados utilizando as séries temporais completas das variáveis regressoras no conjunto de calibração. No caso dos modelos locais, inicialmente a classificação não supervisionada define as condições do céu (classes) para cada janela de 15 minutos da série temporal. Essa etapa inicial do trabalho é feita antes das divisões dos dados em calibração, validação e teste. Assim que o vetor de classes é obtido, os dados são divididos nos três conjuntos. Então, o modelo *Random Forest* (classificação supervisionada) é treinado usando os dados do conjunto de treinamento. Para cada classe de saída do *Random Forest* no conjunto de calibração, são treinados os modelos regressivos locais. Assim, haverá, por exemplo, um MLR específico para a classe céu claro ou um MLP específico para a classe céu nublado. Vale salientar que todos os modelos globais utilizam a análise de componentes principais (PCA) para extração das variáveis regressoras. Já os

modelos locais utilizam tanto a seleção de variáveis por filtros de correlação quanto a extração de variáveis por PCA para definir as variáveis regressoras.

No conjunto de validação são selecionados os melhores modelos para cada subconjunto da série, bem como são treinados os modelos de combinação. Para escolher os melhores modelos dentre os globais e os locais, dois critérios estáticos são utilizados, formando dois novos modelos (‘ModSelec’ e ‘ModSelec2’) compostos pelos melhores modelos locais em cada subconjunto. Os dois critérios analisam os estatísticos correlação, razão entre os desvios (STDRatio), erro quadrático médio normalizado (RMSEn) e SS4 para escolha do melhor modelo e são apresentados na Tabela 11.

Tabela 11 – Critérios estáticos para escolha dos melhores modelos para cada subconjunto local no conjunto de validação.

Modelos	**Critérios utilizados para seleção dos melhores modelos**
ModSelec	1) No conjunto de validação, selecionar, dentre todos os modelos globais e locais treinados no conjunto de treinamento, os 5 modelos com maior correlação; 2) Dentre os cinco modelos selecionados, avaliar os três que possuem a razão entre os desvios mais próxima a 1, em valor absoluto; 3) Dentre os três modelos selecionados, escolher o que possui o menor nRMSE.
ModSelec2	1) No conjunto de validação, selecionar, dentre todos os modelos globais e locais treinados no conjunto de treinamento, os 5 modelos com maior SS4; 2) Dentre os cinco modelos selecionados, avaliar os três que possuem o menor nRMSE; 3) Dentre os três modelos selecionados, escolher o que possui o STDRatio mais próximo a 1, em valor absoluto.

Fonte: própria.

Por fim, os modelos de combinação são treinados, também no conjunto de validação, levando-se em consideração todos os modelos globais e locais treinados no conjunto de treinamento, bem como os dois novos modelos ‘ModSelec’ e ‘ModSelec2’. A definição das variáveis regressoras para os modelos de combinação é feita a partir de seleção de variáveis por filtros de correlação no conjunto de validação. Contudo, no caso dos modelos de combinação, a seleção é feita primeiro avaliando a correlação das variáveis (modelos globais e locais no conjunto de validação) com o *target*, excluindo-se as variáveis que tenham uma correlação inferior a 0,85 do conjunto. Caso todas as variáveis tenham correlação inferior a 0,85 com o *target*, esse critério é diminuído para 0,5. Em seguida, em um segundo passo, a matriz de correlação é calculada sobre as variáveis remanescentes e as que apresentarem uma correlação superior a 0,999 entre si, analisando linha a linha, são eliminadas do

conjunto. Os valores de correlação para seleção das variáveis regressoras dos modelos de combinação foram escolhidos após testar diferentes possibilidades de valores para escolha das variáveis.

A Tabela 12 descreve todos os modelos empregados para o *site adaptation*. O modelo de referência é o produto fornecido pela CAMS de radiação (GHI ou DNI). Vale salientar que as séries finais de radiação fornecidas pela CAMS consideram, na maioria dos casos, um ajuste estatístico na saída do modelo feita pela própria CAMS. A CAMS fornece, também, as séries de irradiância sem considerar essa correção estatística (CAMSnocorr). Os modelos globais, locais e de combinação serão, portanto, comparados com o modelo da CAMS na avaliação dos resultados.

Tabela 12 – Descrição de todos os modelos utilizados para o *site adaptation*.

Continua

MODELOS DE REFERÊNCIA	
CAMS	GHI ou DNI resultante do modelo da CAMS
CAMSnocorr	GHI ou DNI resultante do modelo da CAMS sem aplicação de modelos estatísticos para ajuste do modelo (*BIAS correction*)
MODELOS GLOBAIS	
MLR	MLR com variáveis regressoras extraídas por PCA
MLR_QM	MLR com variáveis regressoras extraídas por PCA e pós-processamento utilizando QM
MLR_BD	MLR com variáveis regressoras extraídas por PCA e pós-processamento utilizando correção por BIAS e desvio
MLP	MLP com variáveis regressoras extraídas por PCA
MLP_QM	MLP com variáveis regressoras extraídas por PCA e pós-processamento utilizando QM
MLP_BD	MLP com variáveis regressoras extraídas por PCA e pós-processamento utilizando correção por BIAS e desvio

Fonte: própria.

Tabela 12 – Descrição de todos os modelos utilizados para o *site adaptation*.
Conclusão.

MODELOS LOCAIS	
MLR_SEL_Loc	5 MLRs com variáveis regressoras selecionadas utilizando filtros de correlação (SEL) para cada subconjunto
MLR_SEL_BD_Loc	5 MLRs com variáveis regressoras selecionadas utilizando filtros de correlação (SEL) e pós-processadas com correção por BIAS e desvio para cada subconjunto
MLR_PCA_Loc	5 MLRs com variáveis regressoras extraídas com PCA para cada subconjunto
MLR_PCA_BD _Loc	5 MLRs com variáveis regressoras extraídas com PCA e pós-processadas com correção por BIAS e desvio para cada subconjunto
MLP_SEL_Loc	5 MLPs com variáveis regressoras selecionadas utilizando filtros de correlação (SEL) para cada subconjunto
MLP_SEL_BD_Loc	5 MLPs com variáveis regressoras selecionadas utilizando filtros de correlação (SEL) e pós-processadas com correção por BIAS e desvio para cada subconjunto
MLP_PCA_Loc	5 MLPs com variáveis regressoras extraídas com PCA para cada subconjunto
MLP_PCA_BD _Loc	5 MLPs com variáveis regressoras extraídas com PCA e pós-processadas com correção por BIAS e desvio para cada subconjunto
ModSelec	Modelo composto pelos melhores modelos para cada subconjunto (classe) selecionados no conjunto de validação utilizando critérios estatísticos específicos (Tabela 11)
ModSelec2	Modelo composto pelos melhores modelos para cada subconjunto (classe) selecionados no conjunto de validação utilizando critérios estatísticos específicos (Tabela 11)
MODELOS DE COMBINAÇÃO	
MLRComb	MLR treinado no conjunto de validação com variáveis regressoras selecionadas utilizando filtros de correlação sobre todos os modelos globais e locais
MLRComb_BD	MLR treinado no conjunto de validação com variáveis regressoras selecionadas utilizando filtros de correlação sobre todos os modelos globais e locais, pós-processados por correção de BIAS e desvio
MLPComb	MLP treinado no conjunto de validação com variáveis regressoras selecionadas utilizando filtros de correlação sobre todos os modelos globais e locais
MLPComb_BD	MLP treinado no conjunto de validação com variáveis regressoras selecionadas utilizando filtros de correlação sobre todos os modelos globais e locais, pós-processados por correção de BIAS e desvio

Fonte: própria.

5 RESULTADOS

Inicialmente, são descritos os dados observacionais utilizados para testar a metodologia proposta, bem como as bases de dados históricas utilizadas (seção 5.1). Em seguida, para cada estação solarimétrica, são apresentados na seção 5.2 os resultados quanto à classificação não supervisionada dos dados. A seção 5.3 apresenta os resultados da classificação supervisionada e a seção 5.4 apresenta os resultados do *site adaptation* empregado em cada estação, descrevendo os resultados dos modelos globais, locais e de combinação.

5.1 BASE DE DADOS UTILIZADAS

Quatro estações solarimétricas serão utilizadas para aplicação do *site adaptation*. As características dessas estações são apresentadas na Tabela 12. As estações de El Rosal e Salta medem irradiância global horizontal, enquanto as estações de Petrolina e Gobabeb medem as três componentes da radiação. Os dados passaram por um algoritmo para garantia e controle da qualidade, tal como proposto em Petribú et al. (2017). Todos os dados considerados como anômalos (*outliers*) pelos testes de qualidade não foram considerados para a realização do *site adaptation*.

Tabela 12 – Informações sobre as estações solarimétricas utilizadas.

Localização/Cidade	**El Rosal**	**Salta**	**Petrolina**	**Gobabeb**
Província/Estado	Salta	Salta	Pernambuco	Orongo
País	Argentina	Argentina	Brasil	Namíbia
Latitude (°)	-24,393	-24,72872	-9,107	-23.5614
Longitude (°)	-65,7683	-65,40958	-40,442	15.04198
Altitude (m)	3355	1233	404	409
Piranômetro	Kipp&Zonen CMP3	Eppley PSP	EKO MS-80A (GHI) e EKO MS-57 (DNI)	Kipp&Zonen CMP22 (GHI) e Kipp&Zonen CHP1 (DNI)
Intervalo de integração	1 min	1 min	1 min	1 min
Período de medição	2013 - 2018	2013 - 2015	2018 – 2022	2015 - 2020

Fonte: própria.

Em relação às as bases de dados históricas, foram utilizadas variáveis do modelo da CAMS *Radiation Service* que fornece séries temporais de radiação de acordo com o campo de visão do satélite METEOSAT de -66° a 66° em ambas latitude e longitude (CAMS radiation service - www.soda-pro.com) e interpolações das saídas

do ERA5-Land do ECMWF (ERA5-Land hourly data from 1981 to present (copernicus.eu). A CAMS fornece as séries temporais já interpoladas para a região de interesse na escala temporal de minutos até meses; neste trabalho, foram utilizadas as saídas da CAMS na escala temporal dos minutos integralizadas para a escala de 15 min. Já para variáveis do ECMWF, foi utilizado uma interpolação bilinear simples para obter as séries temporais na coordenada de interesse e uma interpolação temporal simples para obter as séries de 15 min a partir das séries horárias. A Tabela 13 mostra todas as variáveis das bases de dados utilizadas neste trabalho.

Além destas variáveis, no *site adaptation* da irradiância direta normal utiliza-se também o índice de claro (k_t – razão entre a GHI e a irradiância extraterrestre efetiva horizontal), a fração difusa (k_d – razão entre a DHI e a GHI), a transmitância normal (k_n – razão entre a DNI e a irradiância extraterrestre efetiva) e os índices de céu claro para cada uma das componentes da radiação (as componentes de radiação divididas pelas respectivas séries de céu claro).

Tabela 13 – Variáveis utilizadas para realização do site adaptation.

Continua

#	Nome da variável	Descrição da variável	Modelo
1	TOA	Irradiação horizontal no topo da atmosfera (W/m²)	CAMS
2	Clear sky GHI	Irradiação global horizontal de céu claro (W/m²)	CAMS
3	Clear sky BHI	Irradiação direta horizontal de céu claro (W/m²)	CAMS
4	Clear sky DHI	Irradiação difusa horizontal de céu claro (W/m²)	CAMS
5	Clear sky BNI	Irradiação direta normal de céu claro (W/m²)	CAMS
6	GHI	Irradiação global horizontal (W/m²)	CAMS
7	BHI	Irradiação direta horizontal (W/m²)	CAMS
8	DHI	Irradiação difusa horizontal (W/m²)	CAMS
9	BNI	Irradiação direta normal (W/m²)	CAMS
10	sza	Ângulo zenital (°)	CAMS
11	summer/winter split	Divisão inverno/verão (1 - verão e 2 - inverno)	CAMS
12	tco3	Coluna total de ozônio (Dobson unit)	CAMS
13	tcwv	Coluna total do vapor de água (kg/m²)	CAMS
14	AOD BC	Profundidade ótica do aerossol a 550 nm para carbono negro	CAMS
15	AOD DU	Profundidade ótica do aerossol a 550 nm para poeira	CAMS
16	AOD SS	Profundidade ótica do aerossol a 550 nm para sal marinho	CAMS
17	AOD OR	Profundidade ótica do aerossol a 550 nm para matéria orgânica	CAMS
18	AOD SU	Profundidade óptica do aerossol a 550 nm para sulfato	CAMS
19	fiso	MODIS BRDF parâmetro fiso	CAMS
20	fvol	MODIS BRDF parâmetro fvol	CAMS
21	fgeo	MODIS BRDF parâmetro fgeo	CAMS
22	albedo	Albedo	CAMS
23	Cloud optical depth	Profundidade ótica da nuvem	CAMS
24	Cloud coverage of the pixel	Cobertura de nuvem do pixel	CAMS

Fonte: própria.

Tabela 13 – Variáveis utilizadas para realização do site adaptation.

Conclusão

#	Nome da variável	Descrição da variável	Modelo
25	CloudType	0=no clouds; 5=low-level cloud; 6=medium-level cloud; 7=high-level cloud and 8=thin cloud	CAMS
26	GHInocorr	Irradiação global horizontal (W/m²) sem correção estatística	CAMS
27	BHInocorr	Irradiação direta horizontal (W/m²) sem correção estatística	CAMS
28	DHInocorr	Irradiação difusa horizontal (W/m²) sem correção estatística	CAMS
29	BHInocorr	Irradiação direta normal (Wh/m²) sem correção estatística	CAMS
25	Vel	Velocidade do Vento (m/s) a 10 m	ERA5-Land
26	d2m	Temperatura do ponto de orvalho a 2 m (K)	ERA5-Land
27	t2m	Temperatura ambiente a 2 m (K)	ERA5-Land
28	skt	Temperatura de uma fina camada da superfície em equilíbrio radiativo (K)	ERA5-Land
29	sp	Pressão atmosférica na superfície (Pas)	ERA5-Land
30	ssrd	Irradiância global horizontal (W/m²)	ERA5-Land
31	strd	Irradiância térmica da superfície (W/m²)	ERA5-Land
32	tp	Precipitação total (m)	ERA5-Land
33	e	Evapotranspiração total (m de água equivalente)	ERA5-Land

Fonte: própria.

´

As variáveis da Tabela 13 são a entrada da extração/seleção de variáveis. Assim, o conjunto utilizado como entrada da análise de componentes principais (PCA) e da seleção de variáveis são as 33 variáveis listadas acima, mais as variáveis adimensionais como o índice de céu claro, índice de claridade, fração difusa, entre outras, sendo todos os índices obtidos a partir das séries de radiação da CAMS.

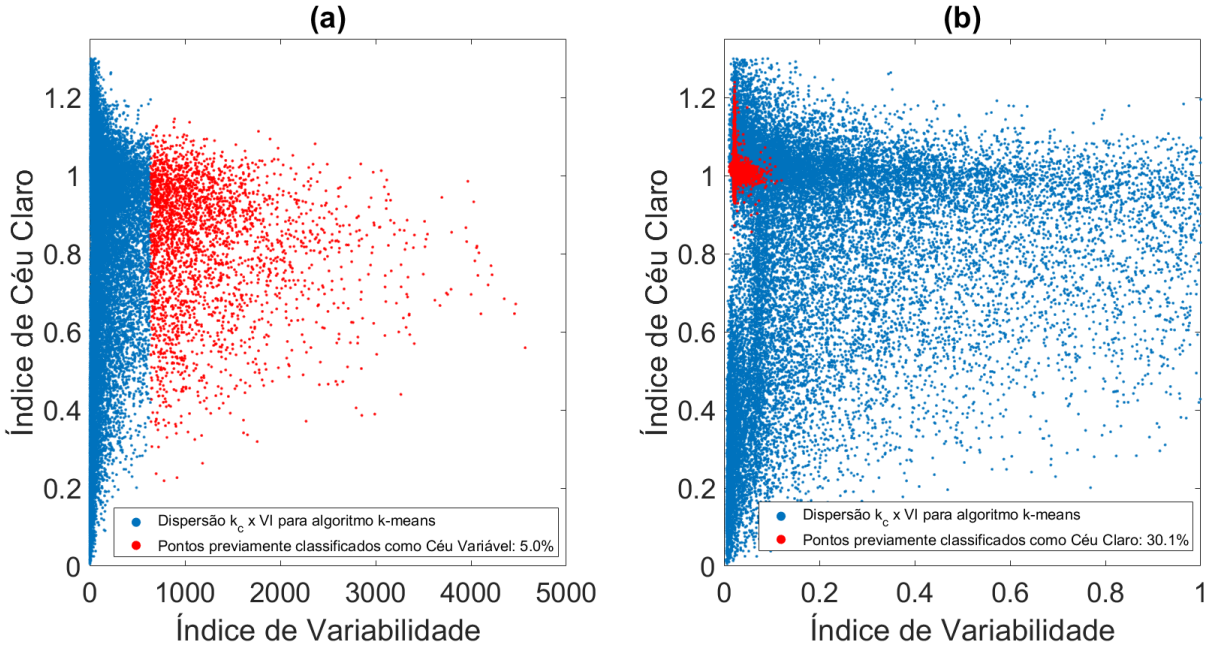
5.2 CLASSIFICAÇÃO NÃO SUPERVISIONADA

Utilizando a série temporal de irradiância global horizontal medida e a série resultante do modelo de céu claro *McClear* para I_g, ambas na resolução de minutos, é possível calcular o índice de céu claro (k_c) e o índice de variabilidade (VI) em janelas temporais de 15 min. São feitas, então, as classificações prévias nas classes céu variável (para valores de VI acima do percentil P95) e céu claro (Tabela 5), conforme apresentado na Figura 11 para a estação de El Rosal.

Em seguida, é estimada a densidade de probabilidade conjunta por *Kernel Density Estimation* (KDE) e são escolhidos os pontos iniciais de *k-means* (Figura 12). Esses pontos iniciais para o algoritmo de *k-means* são os mesmos para todos os casos analisados. Pode-se notar que, para todas as estações solarimétricas, há um pico da densidade de probabilidade na região de céu claro (com k_c elevado e VI pequeno). Nessa região, o ponto escolhido mais acima representa a classe céu claro

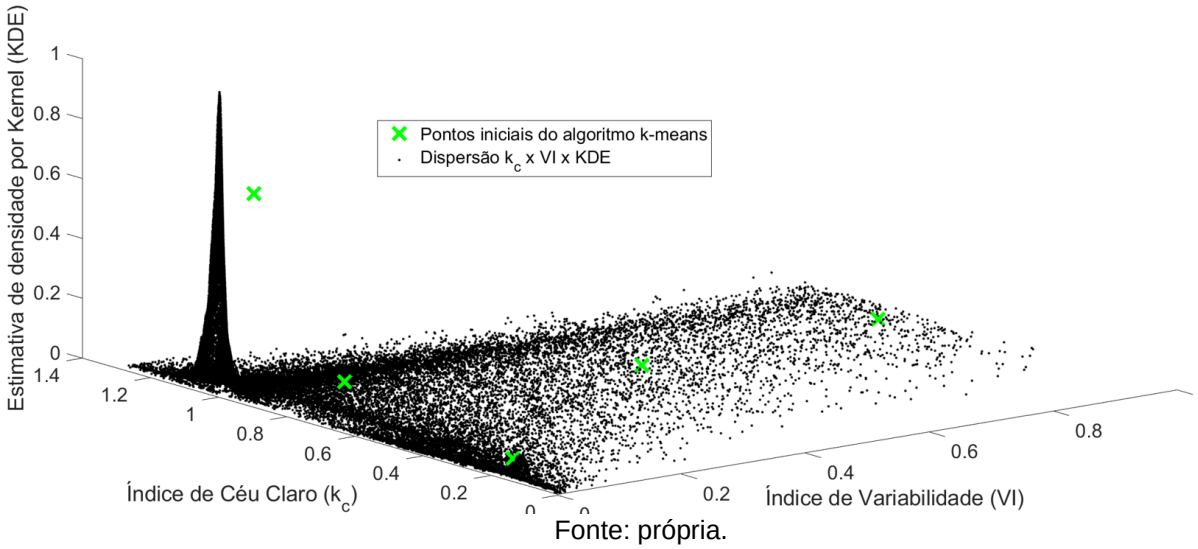
(classe 1), enquanto o ponto localizado na base do pico representa a classe céu parcialmente claro (classe 2). O ponto mais abaixo, na região de k_c pequeno, representa a classe céu nublado (classe 3), enquanto o ponto mais à direita da dispersão, com o VI elevado, representa a classe céu variável (classe 4). Por fim, os pontos que ficam entre as classes céu parcialmente claro, céu variável e céu nublado, abrangem uma região intermediária da dispersão e representam a classe céu variável parcialmente (classe 5).

Figura 11 – Classificação prévia da dispersão $VI\ x\ k_c$ para a estação de El Rosal. Em (a), 5% dos dados da estação El Rosal foram previamente classificados como céu variável (classe 4); enquanto em (b), 30,1% dos dados foram previamente classificados na classe 1 (céu claro).



Fonte: própria.

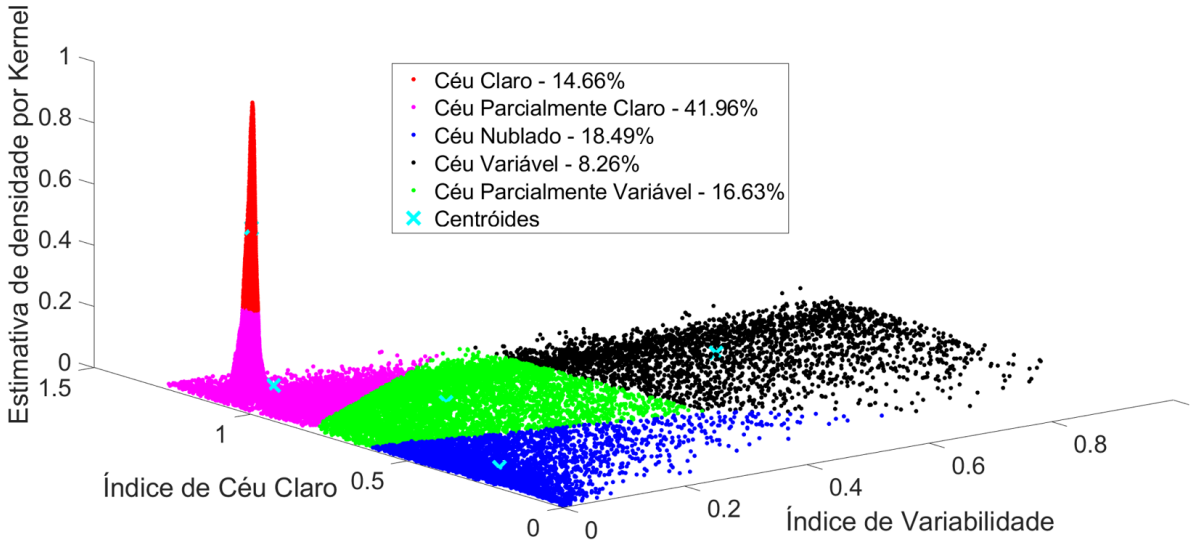
Figura 12 – Dispersão $k_c\ x\ VI\ x\ KDE$ com pontos iniciais do algoritmo de *k-means*.



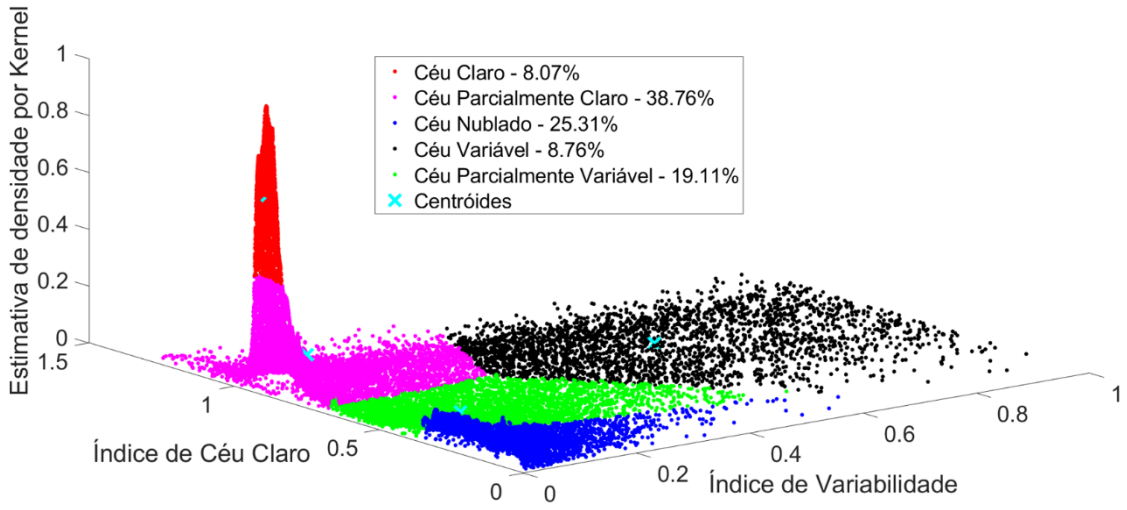
Fonte: própria.

Definidos então a quantidade e a localização dos centróides iniciais, o algoritmo de *k-means* pode ser utilizado para definir os 5 *clusters* de cada estação, conforme apresentado na Figura 13. O padrão de classificação é similar para os 5 casos (mostrando que os pontos iniciais escolhidos apresentam convergência) e está de acordo com as classes definidas anteriormente de forma empírica (Figura 9). Vale salientar que os percentuais de cada classe mostrados nas legendas dos gráficos da Figura 13 não levam em consideração as amostras pré-classificadas nas classes céu variável e céu claro.

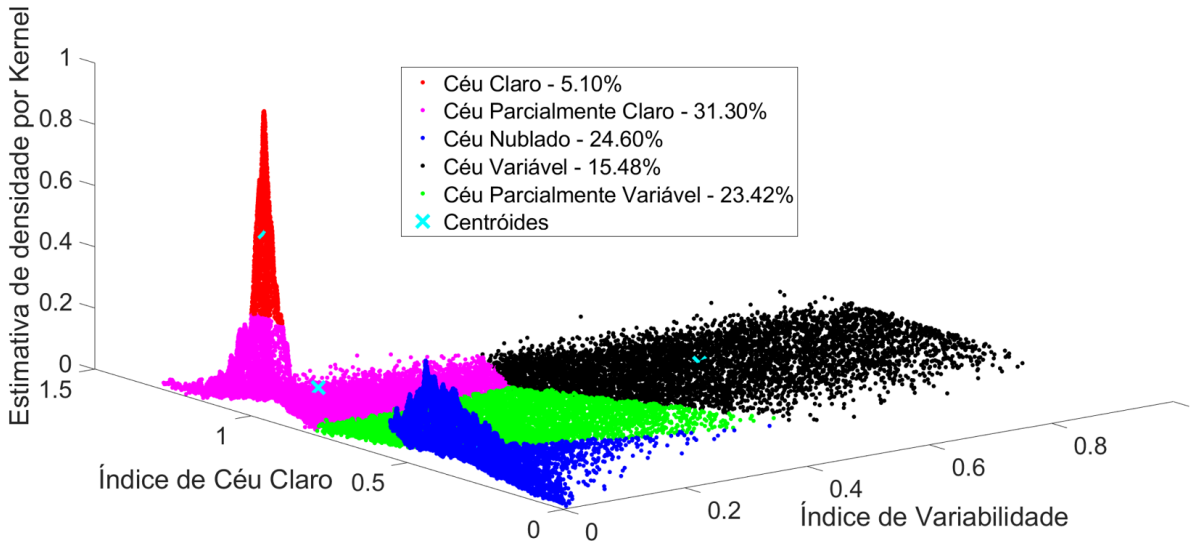
Figura 13 – Resultados do algoritmo de *k-means* para cada estação analisada.



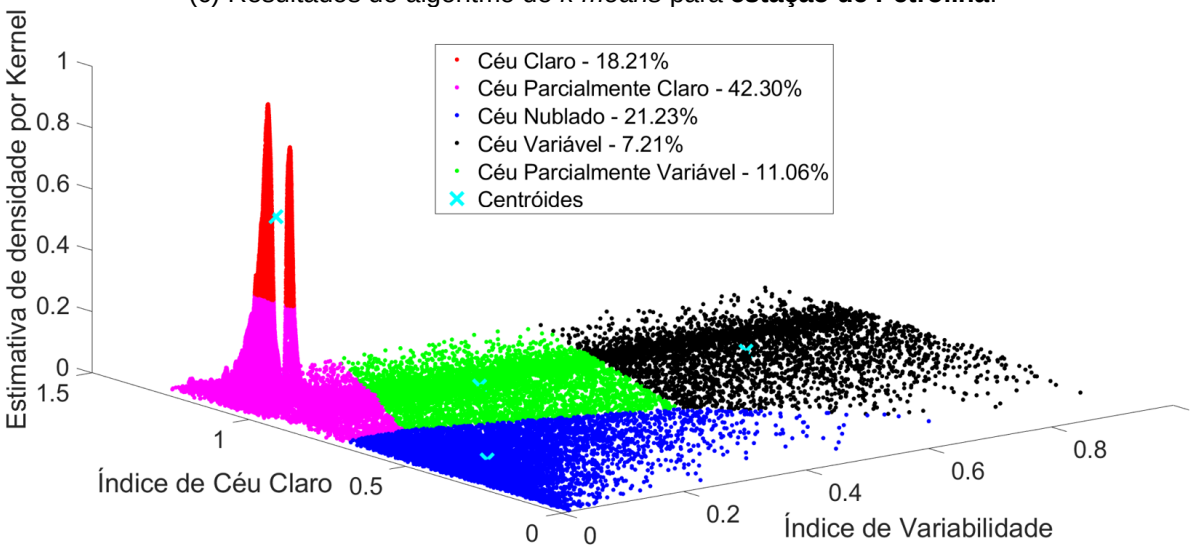
(a) Resultados do algoritmo de *k-means* para **estação de El Rosal**.



(b) Resultados do algoritmo de *k-means* para **estação de Salta**.



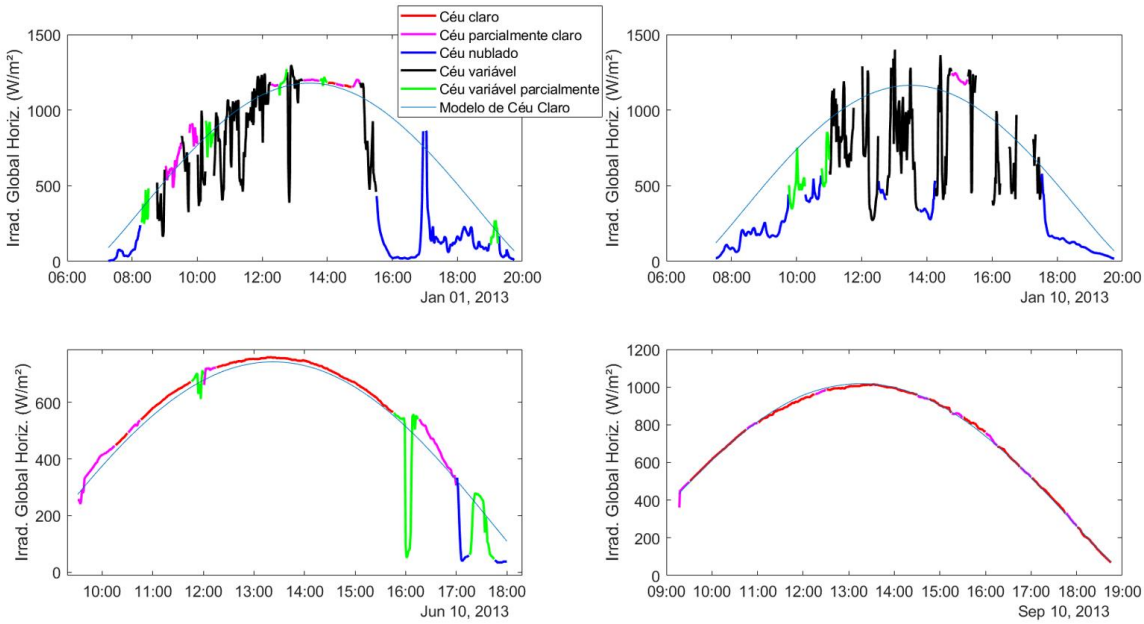
(c) Resultados do algoritmo de *k-means* para **estação de Petrolina**.



(d) Resultados do algoritmo de *k-means* para **estação de Gobabeb**.
Fonte: própria.

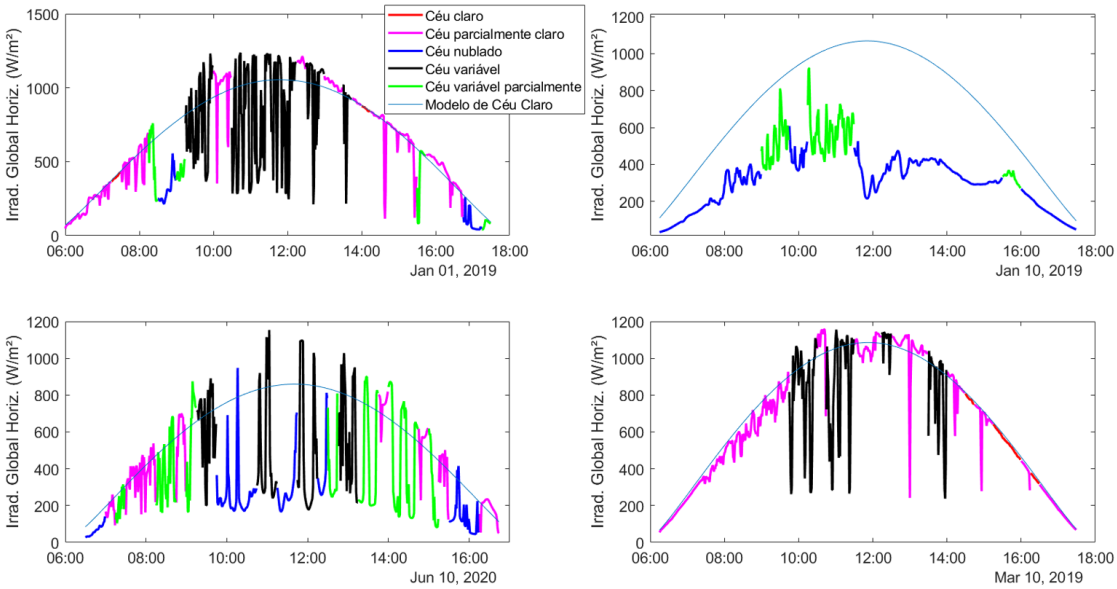
As Figuras 14 e 15 apresentam os dados minuto a minuto com as classificações resultantes das janelas temporais de 15 min para alguns dias das séries de GHI de El Rosal e Petrolina. Pode-se observar que, no geral, as classes correspondem às condições estabelecidas. Contudo, algumas classificações podem representar regiões de fronteiras entre as classes, onde os comportamentos são intermediários e mais difíceis de classificar. Por exemplo, no dia 01/jan/2019 para a estação de Petrolina, algumas janelas temporais classificadas como céu parcialmente claro (em magenta na Figura 15) apresentam um comportamento que pode estar mais próximo da classe céu variável parcialmente (em verde).

Figura 14 – Algumas classes das janelas temporais de 15 min obtidas para a estação de El Rosal.



Fonte: própria.

Figura 15 – Classes das janelas temporais de 15 min para a estação de Petrolina.

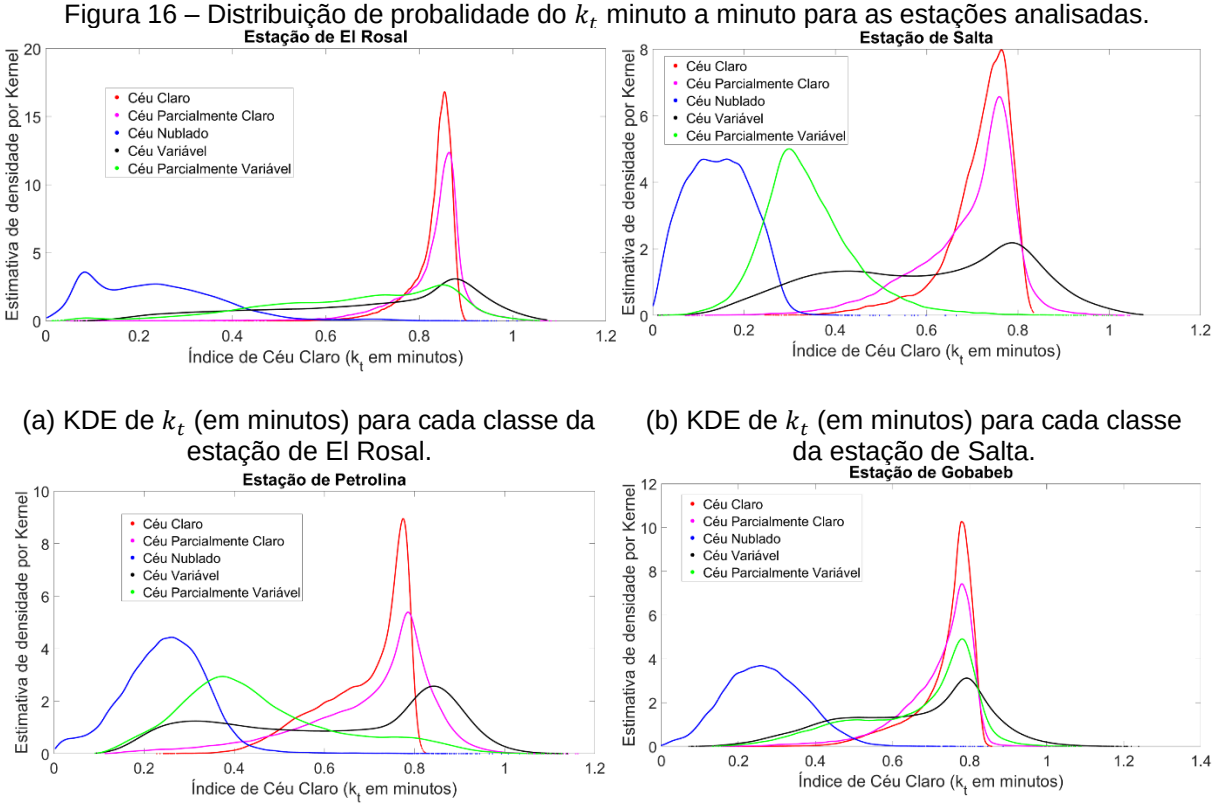


Fonte: própria.

Uma outra forma de avaliar a classificação dos dados feita é plotando as curvas de densidade de probabilidade (estimadas por KDE) do índice de claridade (k_t) na resolução de minutos para cada classe. No modelo de *Kernel Density Estimation* (KDE), o *kernel* utilizado foi o gaussiano, sendo estimado o melhor desvio padrão para a PDF de Gauss (*bandwidth*). Para a classe céu claro, por exemplo, é esperado um pico em valores de k_t elevado, enquanto que para a classe céu nublado é esperado um pico para valores de k_t baixos. A Figura 16 apresenta as curvas de densidade de

k_t para cada classe. As classes céu claro e céu parcialmente claro apresentam um pico na região de k_t mais elevado, enquanto a classe céu nublado apresenta um pico na região de k_t baixo. A classe céu variável apresenta uma distribuição bimodal, sendo uma das modas em k_t baixo e a outra em k_t elevado, com exceção da estação de El Rosal que apresenta somente uma moda (na região de k_t elevado). Por fim, a classe céu parcialmente variável (em verde), que representa uma classe intermediária nas dispersões da Figura 13, possui uma PDF cujo formato varia de acordo com os diferentes climas onde estão as estações.

Figura 16 – Distribuição de probalidade do k_t minuto a minuto para as estações analisadas.



(a) KDE de k_t (em minutos) para cada classe da estação de El Rosal.

(b) KDE de k_t (em minutos) para cada classe da estação de Salta.

(c) KDE de k_t (em minutos) para cada classe da estação de Petrolina.

(d) KDE de k_t (em minutos) para cada classe da estação de Gobabeb.

Fonte: própria.

Os percentuais finais de cada *cluster* unindo a classificação prévia e a classificação resultante do algoritmo de *k-means* são apresentados na Tabela 15. Observa-se que as estações de El Rosal e Gobabeb possuem uma predominância de condições de céu claro, já que mais de 66% e 80% das classes, respectivamente, estão entre as classes 1 (céu claro) ou 2 (céu parcialmente claro). Esse percentual mais elevado de céu claro nessas estações pode estar relacionado à altitude no caso de El Rosal e à região desértica no caso de Gobabeb. Já as estações de Petrolina e

Salta possuem condições do céu mais distribuídas entre todas as classes no período analisado.

Tabela 15 – Percentuais dos dados pertencentes a cada classe para cada estação avaliada.

Classe	El Rosal	Salta	Petrolina	Gobabeb
Céu Claro	38,75%	10,31%	15,43%	65,29%
Céu Parcialmente Claro	27,75%	35,79%	26,25%	15,37%
Céu nublado	11,82%	23,31%	20,48%	7,71%
Céu variável	10,67%	13,05%	18,10%	7,62%
Céu parcialmente variável	11,01%	17,55%	19,66%	4,01%

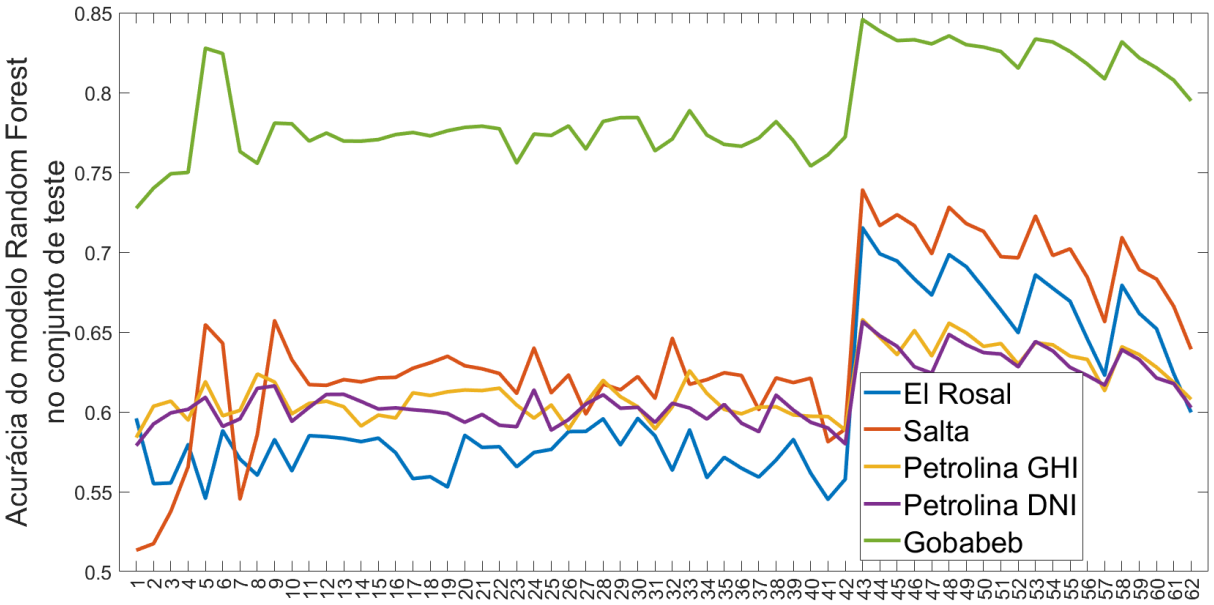
Fonte: própria.

5.3 CLASSIFICAÇÃO SUPERVISIONADA

O *target* de classes construído na etapa anterior será utilizado nos modelos de classificação supervisionada para que os dados provenientes das bases de dados da CAMS e do ECMWF possam aprender as condições do céu na região. Após testar diferentes modelos de classificação, como a regressão logística, redes neurais e árvores de decisão, o modelo *Random Forest* foi o que apresentou os melhores resultados para o problema de classificação proposto. As *features* de entrada para o *random forest* são todas as variáveis da CAMS e do ECMWF.

A Figura 17 apresenta a acurácia do modelo *Random Forest* no conjunto de teste para cada uma das divisões dos dados da Tabela 7 (seção 4.2). Pode-se notar que a acurácia do modelo aumenta significativamente a partir da divisão 43, onde a estratégia de randomização de todos os *timestamps* de 15 min antes de dividir os dados nos conjuntos de calibração, validação e teste é utilizada; em particular, a divisão 43 faz uma proporção de 80%, 10% e 10%, respectivamente, para os três conjuntos citados, e apresenta a maior acurácia da matriz de confusão dentre todas as divisões em todas as estações analisadas. Vale salientar que a estação de Petrolina apresenta resultados diferentes da classificação supervisionada para GHI e DNI, pois devido ao percentual de dados considerados como anômalos pela garantia de qualidade ser mais elevado na DNI que na GHI, decidiu-se utilizar conjuntos de dados distintos para ambas as componentes. Isso não ocorre com a estação de Gobabeb, já que os percentuais de *outliers* são similares para GHI e DNI, podendo-se manter o mesmo conjunto de dados para cada componente.

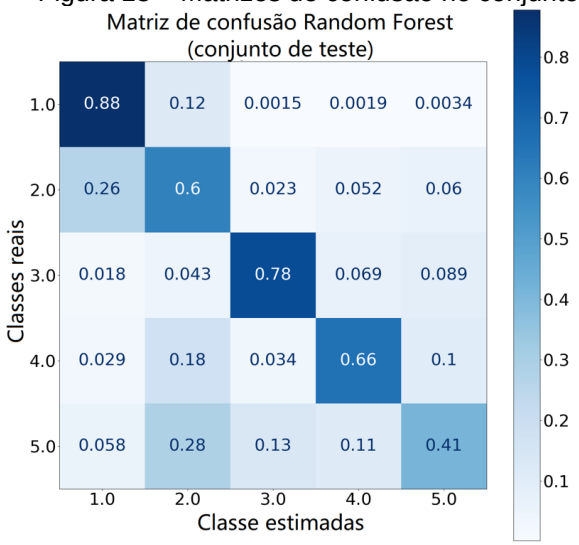
Figura 17 – Acurácia do modelo *random forest* para as estações analisadas.



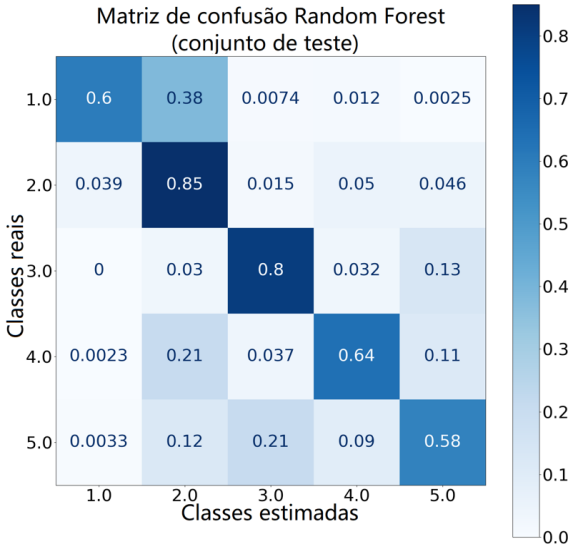
Fonte: própria.

A Figura 18 apresenta as matrizes de confusão no conjunto de teste para as melhores divisões (no caso, a divisão 43) em cada estação estudada. As estações de El Rosal e Gobabeb apresentam os maiores percentuais para céu claro dentre todas as estações (38,75% e 65,29%), o que justifica o maior acerto do modelo na classe 1 para essas estações. Pode-se notar também que o modelo confunde a classificação entre as classes 1 e 2 em todas as estações, atribuindo à classe céu parcialmente claro (classe 2) dados que pertencem à classe céu claro (classe 1) e vice-versa. Por exemplo, para a classe 1 na estação El Rosal, o modelo acertou 88% dos dados classificando-os corretamente como pertencentes à classe céu claro, mas confundiu 12% desses dados classificando-os como céu parcialmente claro. Essa confusão é aceitável já que as duas classes apresentam comportamento similar de céu claro. Para a classe 3 (céu nublado) os resultados mostram uma acurácia superior a 76% para todos os casos, enquanto para a classe 4 (céu variável) os resultados são superiores a 64% de acurácia para todos os casos. Por fim, a classe 5 (céu parcialmente variável) apresenta acurácias mais baixas em todas as estações por se tratar de uma classe intermediária entre as classes céu variável, céu nublado e céu parcialmente claro, o que dificulta que o modelo classifique corretamente essa classe.

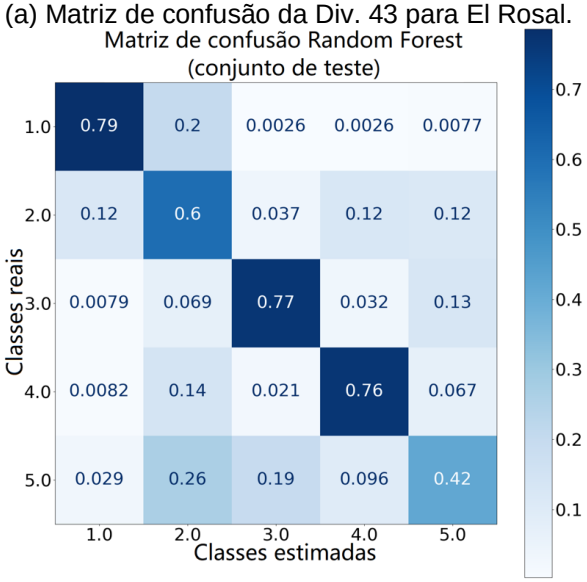
Figura 18 – Matrizes de confusão no conjunto de teste para as quatro estações na divisão 43.



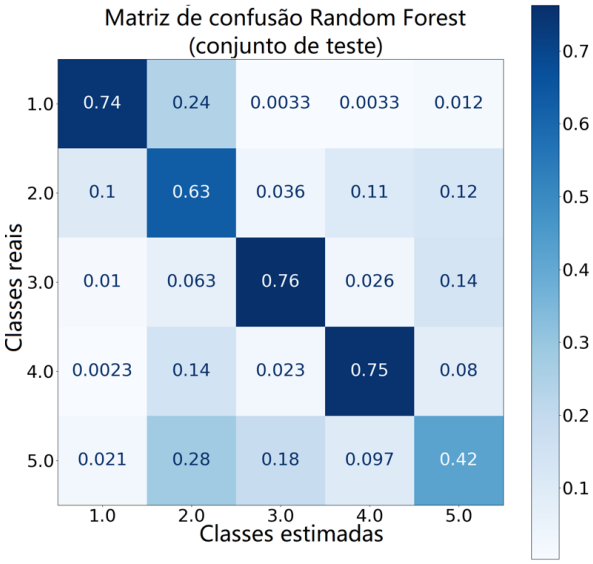
(a) Matriz de confusão da Div. 43 para El Rosal.



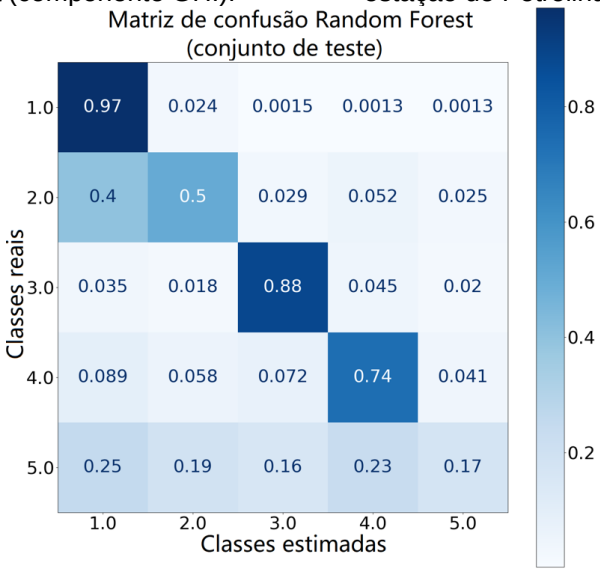
(b) Matriz de confusão da Div. 43 para Salta.



(c) Matriz de confusão da Divisão 43 para estação de Petrolina (componente GHI).



(d) Matriz de confusão da Divisão 43 para estação de Petrolina (componente DNI).



(e) Matriz de confusão da Divisão 43 para estação de Gobabeb.
Fonte: própria.

5.4 SITE ADAPTATION

Nesta etapa são estimados os parâmetros dos modelos globais e locais (utilizando a saída do modelo *Random Forest* para separar os dados) no conjunto de treinamento. No conjunto de validação, são escolhidos os melhores modelos para cada uma das 5 classes dentre todos os modelos globais, locais e o próprio modelo da CAMS, formando os modelos 'ModSelec' e 'ModSelec2', de acordo com os critérios da Tabela 11 (seção 4.5). O treinamento dos modelos de combinação também é feito no conjunto de validação. Por fim, todos os modelos são aplicados e seus resultados avaliados no conjunto de teste, para cada uma das 62 divisões dos dados. Vale salientar que o modelo da CAMS é o modelo de referência com o qual todos os outros resultados serão sempre comparados. As próximas seções apresentam os resultados do *site adaptation* por estação solarimétrica.

Para avaliar os modelos nas diferentes divisões dos dados, são selecionados os melhores modelos das três principais categorias (globais, locais e de combinação) para cada divisão. Essa seleção é feita de acordo com os critérios estabelecidos na Tabela 16.

Tabela 16 – Critérios para escolha dos melhores modelos considerando as 62 divisões dos dados.

Categoria de modelo	**Avaliação do modelo para cada uma das 62 divisões**
Modelos globais	1) São selecionados os quatro modelos globais com maior SS4; 2) Dentre esses quatro modelos são escolhidos os dois que possuem a razão entre os desvios mais próxima a 1, em valor absoluto; 3) O melhor modelo dentre os dois do passo anterior é o que possui o menor RMSEn.
Modelos locais	1) São selecionados os seis modelos locais com maior SS4; 2) Dentre esses seis modelos são escolhidos os três que possuem a razão entre os desvios mais próxima a 1, em valor absoluto; 3) O melhor modelo dentre os três do passo anterior é o que possui o menor RMSEn.
Modelos de combinação	1) São selecionados os três modelos de combinação com maior SS4; 2) Dentre esses três modelos são escolhidos os dois que possuem a razão entre os desvios mais próxima a 1, em valor absoluto; 3) O melhor modelo dentre os dois do passo anterior é o que possui o menor RMSEn.
Escolha da melhor divisão (após escolher o melhor modelo por categoria)	1) Das 62 divisões são selecionadas as dez divisões cujos resultados do modelo escolhido apresentem maior SS4; 2) Dentre essas dez divisões são escolhidas as cinco que possuem a razão entre os desvios mais próxima a 1, em valor absoluto; 3) Dentre essas cinco divisões são escolhidas as três que possuírem o menor RMSEn. 4) A melhor divisão dos dados dentre as três do passo anterior é a cujo modelo possui o menor MBEn.

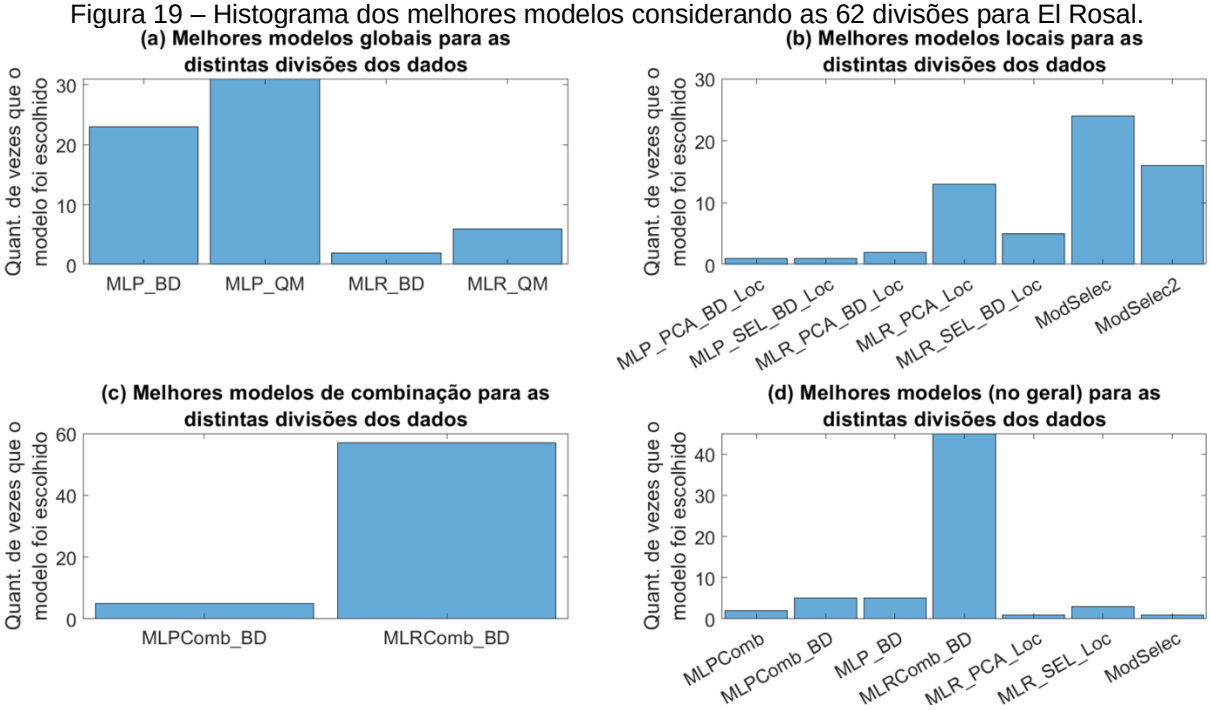
Fonte: própria.

Em seguida, são avaliados histogramas que apresentam os melhores modelos de cada categoria (global, local ou de combinação) e a quantidade de vezes que os mesmos foram selecionados como melhores modelos para decidir qual dos modelos de cada categoria escolher. Após a escolha dos melhores modelos dentre as três principais categorias, é preciso selecionar qual a melhor divisão de dados. De forma similar aos critérios estabelecidos anteriormente, é apresentado na Tabela 16 o critério utilizado para escolha da melhor divisão de dados.

5.4.1 Estação El Rosal

Considerando todos os modelos para o *site adaptation* da GHI de El Rosal nas distintas divisões dos dados, foi escolhido para cada uma das 62 divisões, o melhor modelo de acordo com os critérios estabelecidos na Tabela 16 para cada categoria. A Figura 19 apresenta o histograma com a quantidade de vezes que os modelos foram selecionados como melhores nas distintas divisões. Pode-se notar que, na estação El Rosal, os melhores modelos global, local e de combinação são, respectivamente, MLP_QM, ModSelec e MLRComb_BD. Considerando todos os modelos, a Figura 19.d mostra que MLRComb_BD apresenta os melhores resultados para as distintas divisões dos dados.

Figura 19 – Histograma dos melhores modelos considerando as 62 divisões para El Rosal.

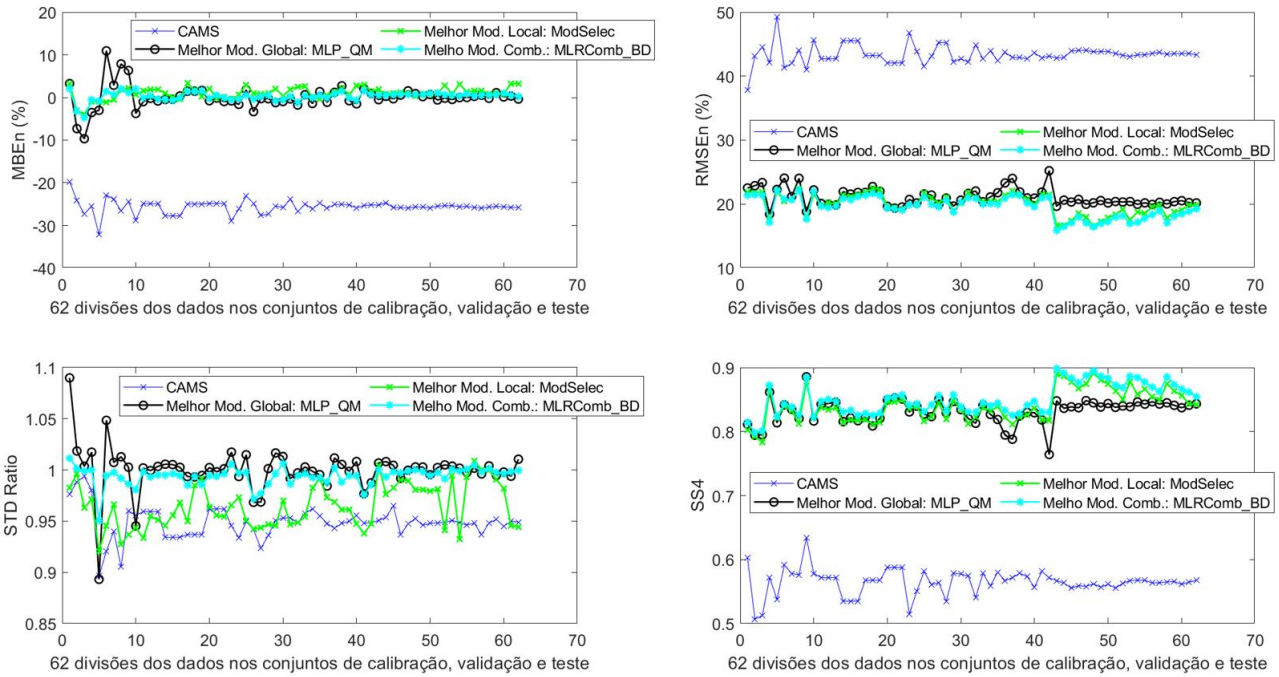


Fonte: própria.

A Figura 20 apresenta o MBEn, RMSEn, STDRatio e SS4 para cada uma das divisões dos dados dos melhores modelos global, local e de combinação em comparação com o modelo da CAMS. Pode-se notar que os modelos locais e de combinação apresentam resultados similares em termos de MBEn, RMSEn e SS4, contudo divergem quanto ao STDRatio, com o modelo local ModSelec apresentando uma maior variação nos resultados para o STDRatio na faixa de 0,95 a 1. No geral, os modelos locais e de combinação apresentam melhores resultados que a CAMS e que os modelos globais. As melhores divisões dos dados para Estação de El Rosal foram as divisões 21, 53 e 53 para o melhor modelo global, local e de combinação, respectivamente.

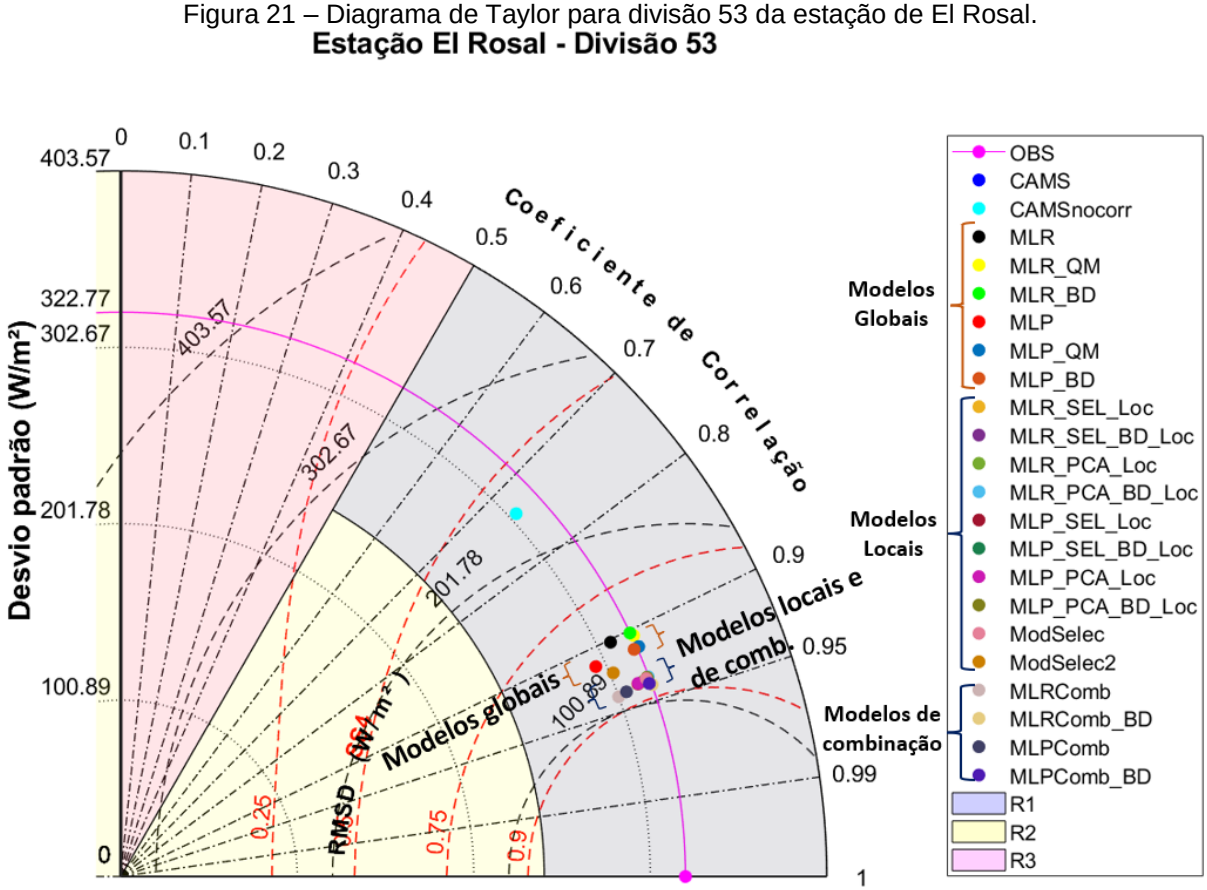
A Figura 21 apresenta o diagrama de Taylor no conjunto de teste para a divisão 53, já que o modelo de combinação MLRComb_BD foi o que apresentou o melhor resultado. O modelo da CAMS não representa bem a GHI para a localidade da estação de El Rosal, já que possui uma correlação de 0,74 e um RMSEn de 43,1%. Já os modelos globais ajustam a correlação para valores próximos a 0,9, enquanto os locais e de combinação ajustam para valores próximos de 0,94.

Figura 20 – Resultados dos melhores modelos por categoria para cada uma das 62 divisões na estação de El Rosal.



Fonte: própria.

Figura 21 – Diagrama de Taylor para divisão 53 da estação de El Rosal.



Fonte: própria.

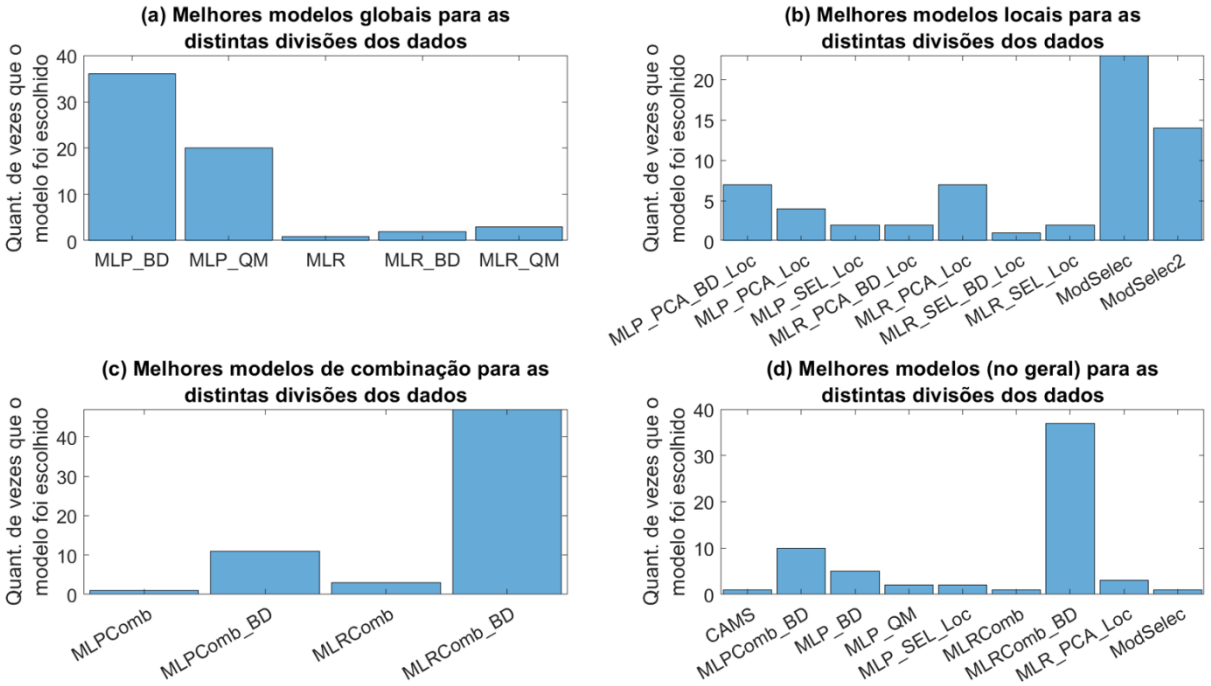
5.4.2 Estação Salta

A Figura 22 apresenta o histograma com a quantidade de vezes que os modelos foram selecionados como melhores nas distintas divisões. Na estação de Salta os melhores modelos global, local e de combinação são, respectivamente, MLP_BD, ModSelec e MLRComb_BD. Considerando todos os modelos, a Figura 22.d mostra que MLRComb_BD apresenta os melhores resultados para as distintas divisões dos dados.

Na Figura 23, os modelos apresentam variações consideráveis em todos os estatísticos para as 10 primeiras divisões dos dados que mantêm a sequência cronológica das séries temporais. Já considerando as divisões de 43 a 62, com randomização dos *timestamps* de 15 minutos, os modelos de combinação apresentam menores RMSEn e maiores SS4 que os outros modelos. Novamente, os modelos locais e de combinação apresentam melhores resultados que a CAMS e que os modelos globais no caso da estação de Salta. As melhores divisões dos dados para

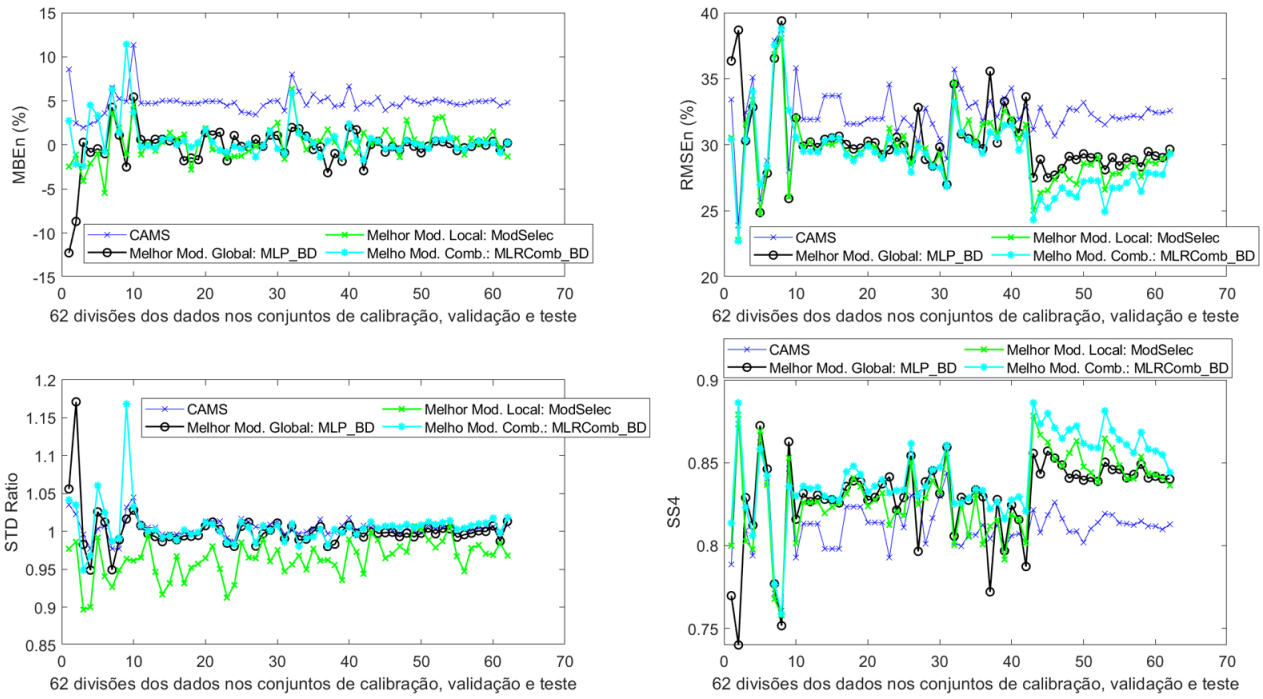
estação de Salta foram as divisões 43, 43 e 46 para o melhor modelo global, local e de combinação, respectivamente.

Figura 22 – Histograma dos melhores modelos considerando as 62 divisões para Salta.



Fonte: própria.

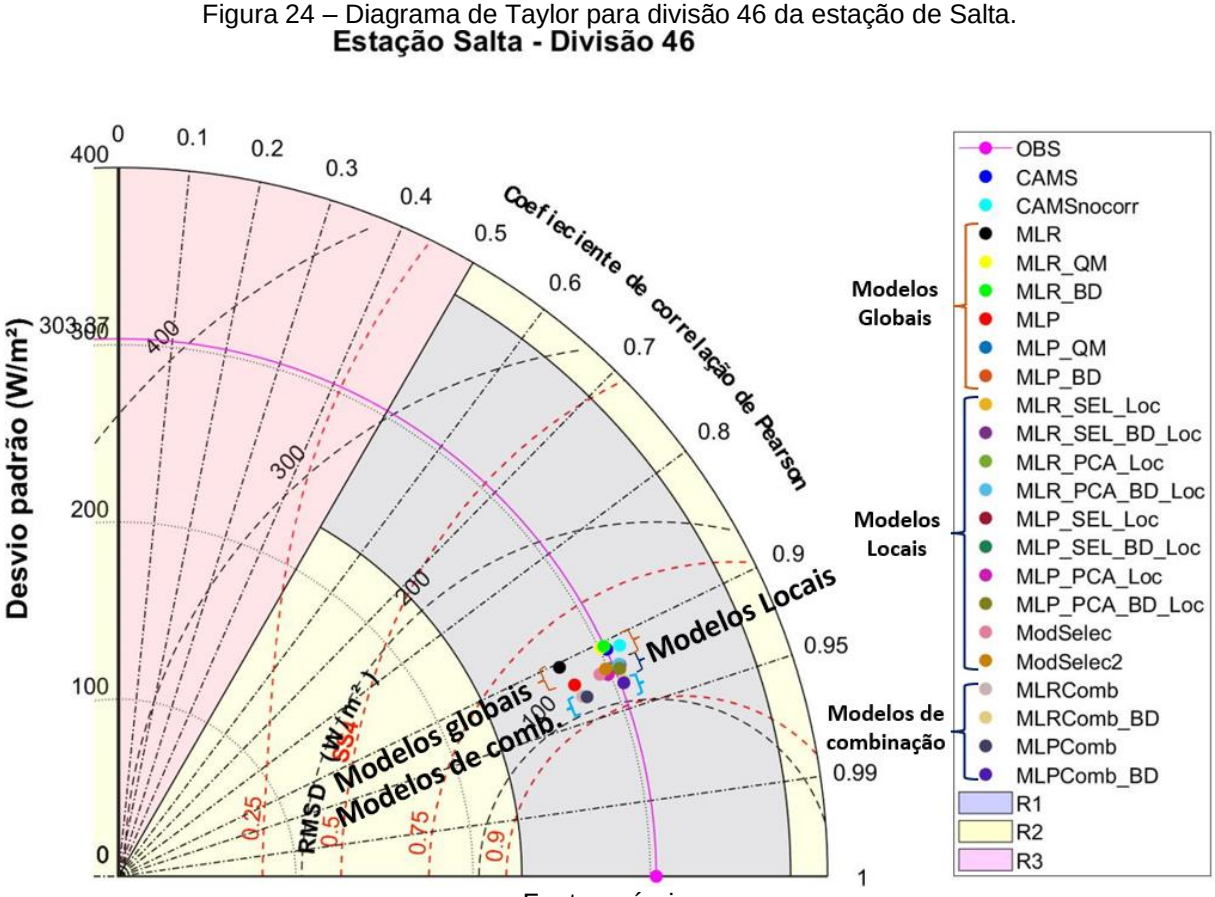
Figura 23 – Resultados dos melhores modelos por categoria para cada uma das 62 divisões para a estação de Salta.



Fonte: própria.

A Figura 24 apresenta o diagrama de Taylor no conjunto de teste para a divisão 46, já que o modelo de combinação MLRComb_BD foi o que apresentou o melhor resultado. O modelo da CAMS e os modelos globais apresentam resultados similares, com correlações próximas a 0,91. Já os modelos locais e de combinação apresentam uma leve melhora na correlação que fica próximo a 0,94.

Figura 24 – Diagrama de Taylor para divisão 46 da estação de Salta.

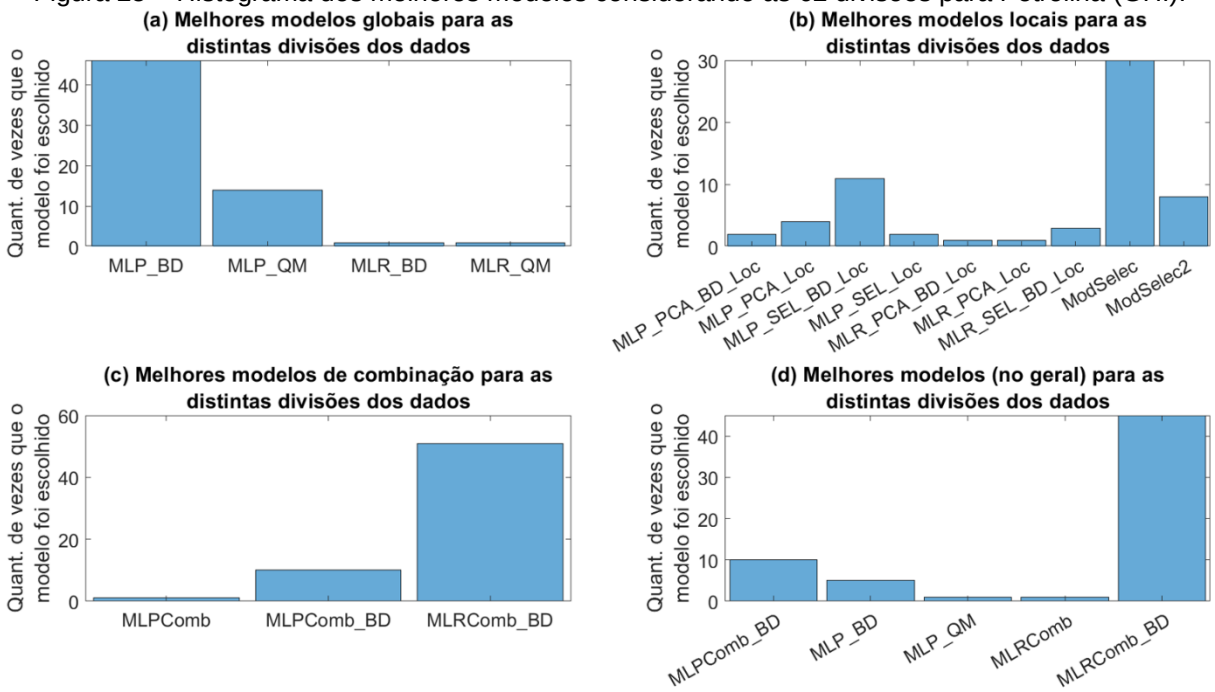


Fonte: própria.

5.4.3 Estação Petrolina (GHI)

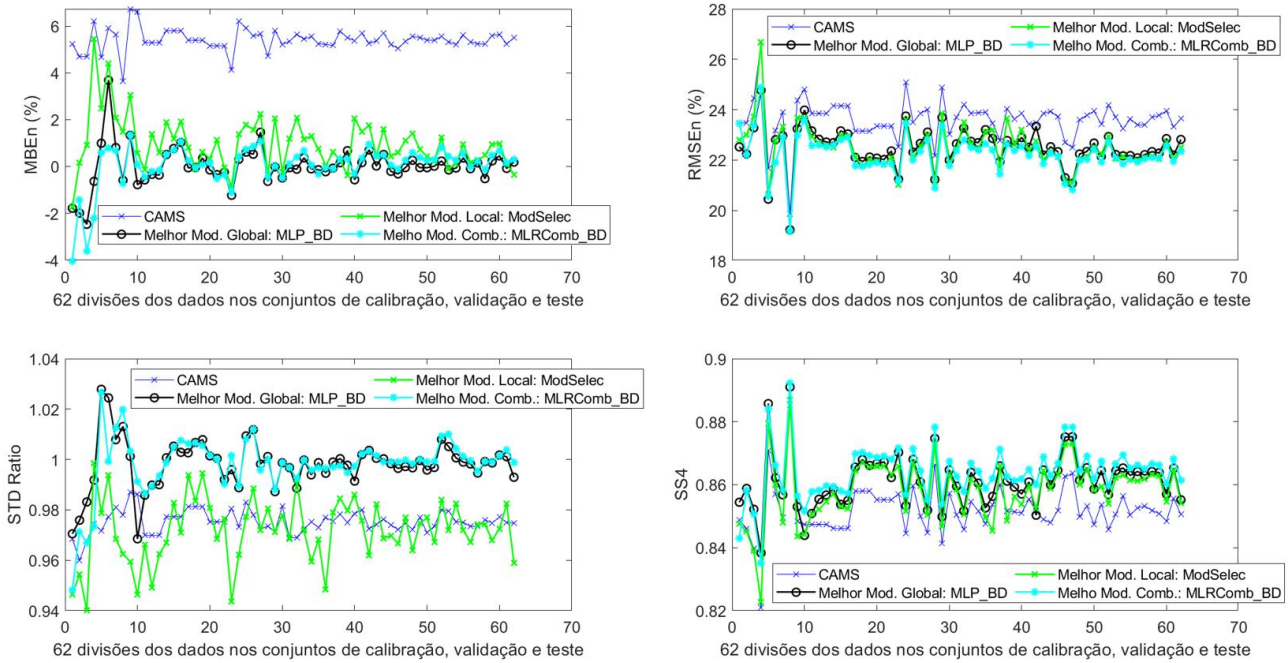
A Figura 25 apresenta o histograma com a quantidade de vezes que os modelos foram selecionados como melhores nas distintas divisões. Pode-se notar que, na estação de Petrolina para GHI, os melhores modelos global, local e de combinação são, respectivamente, MLP_BD, ModSelec e MLRComb_BD. Considerando todos os modelos, a Figura 25.d mostra que MLRComb_BD apresenta os melhores resultados para as distintas divisões dos dados.

Figura 25 – Histograma dos melhores modelos considerando as 62 divisões para Petrolina (GHI).



Fonte: própria.

Figura 26 – Resultados dos melhores modelos por categoria para cada uma das 62 divisões para a estação de Petrolina (GHI).

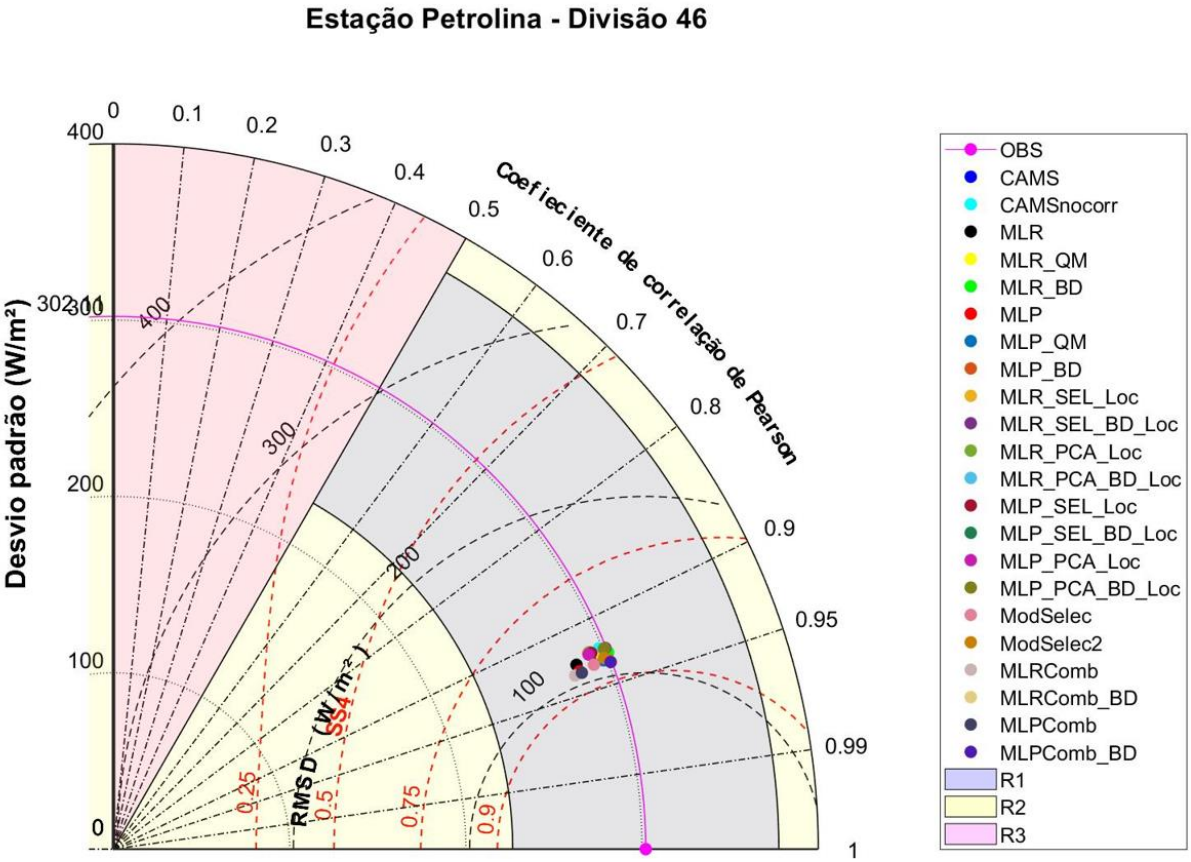


Fonte: própria.

Da Figura 26, pode-se notar que o BIAS (MBEn) apresentado pelo modelo da CAMS é ajustado por todos os modelos, sem considerar as divisões de dados que mantém a sequência cronológica das séries temporais (divisões 1 a 10), já que essas divisões apresentam uma variação maior no BIAS, bem como nos outros estatísticos.

Contudo, os resultados do RMSEn e do SS4 são similares para todos os modelos, com uma melhora pouco expressiva quando comparado com o modelo de referência da CAMS. Os modelos globais e de combinação apresentam resultados ligeiramente melhores que a CAMS. As melhores divisões dos dados para estação de Petrolina, componente GHI, foram as divisões 47, 28 e 46 para o melhor modelo global, local e de combinação, respectivamente. A Figura 27 apresenta o diagrama de Taylor no conjunto de teste para a divisão 46. Pode-se notar que todos os modelos apresentam resultados similares em termos de correlação.

Figura 27 – Diagrama de Taylor para divisão 46 da estação de Petrolina (GHI).

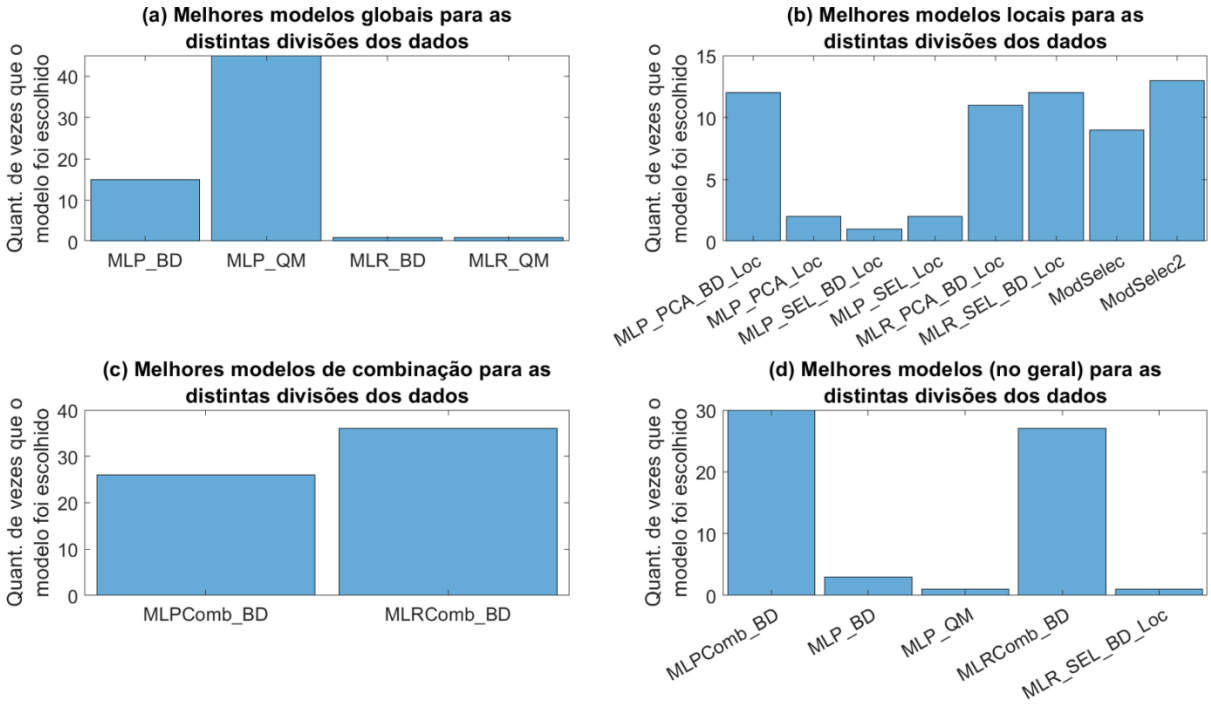


Fonte: própria.

5.4.4 Estação Petrolina (DNI)

Para a estação de Petrolina (componente DNI), os melhores modelos global, local e de combinação são, respectivamente, MLP_QM, ModSelec2 e MLRComb_BD, de acordo com a Figura 28. Considerando todos os modelos, a Figura 28.d mostra que MLPComb_BD apresenta os melhores resultados para as distintas divisões dos dados.

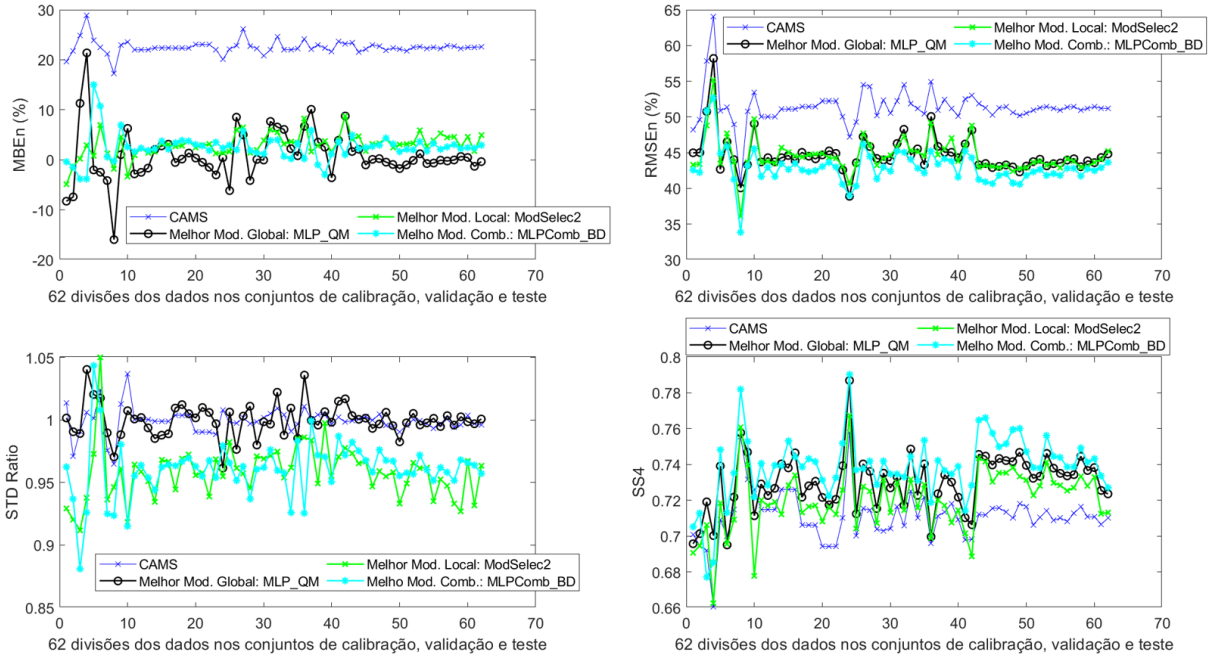
Figura 28 – Histograma dos melhores modelos considerando as 62 divisões para Petrolina (DNI).



Fonte: própria.

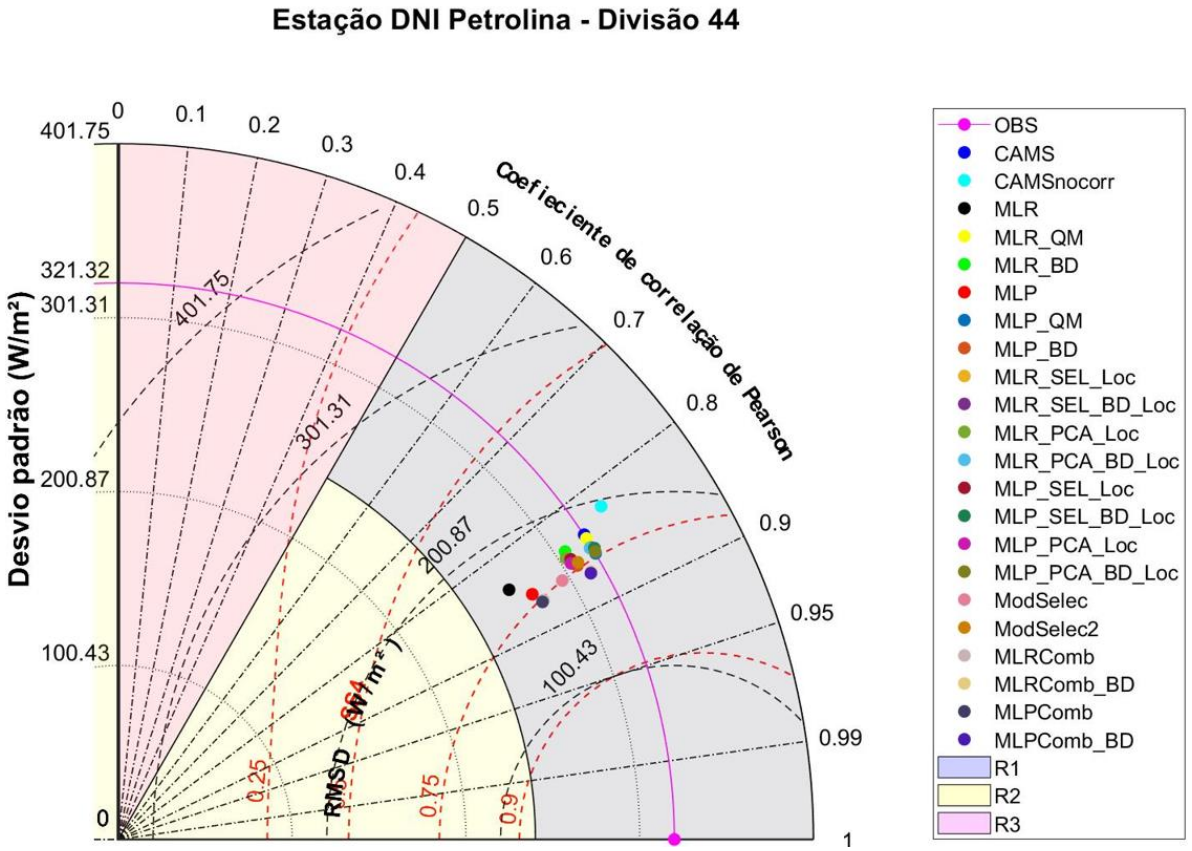
Da Figura 29, pode-se notar que o BIAS (MBEn) apresentado pelo modelo da CAMS da ordem de 20% é bem ajustado por todos os modelos considerando as divisões de 11 a 62. Novamente, as divisões que mantém a sequência cronológica (de 1 a 10) apresentam maior variação nos estatísticos finais que as outras divisões. Os modelos globais, locais e de combinação apresentam resultados melhores que a CAMS em termos de RMSEn e SS4. Já em relação ao STDRatio, os modelos globais e da CAMS apresentam variações nesse estatístico próximo a 1, o que significa que ambos estão representando o desvio padrão da observação. Já o STDRatio dos modelos locais e de combinação apresentam uma variação em torno de 0,95, o que é um valor ainda elevado, porém não tão bom quanto os resultados dos modelos globais e da CAMS. As melhores divisões dos dados para estação de Petrolina, componente DNI, foram as divisões 58, 43 e 44 para o melhor modelo global, local e de combinação, respectivamente. A Figura 30 apresenta o diagrama de Taylor no conjunto de teste para a divisão 44. Pode-se notar que os modelos locais e de combinação apresentam melhores resultados em termos de correlação quando comparados ao da CAMS.

Figura 29 – Resultados dos melhores modelos por categoria para cada uma das 62 divisões para a estação de Petrolina (DNI).



Fonte: própria.

Figura 30 – Diagrama de Taylor para divisão 44 da estação de Petrolina (DNI).



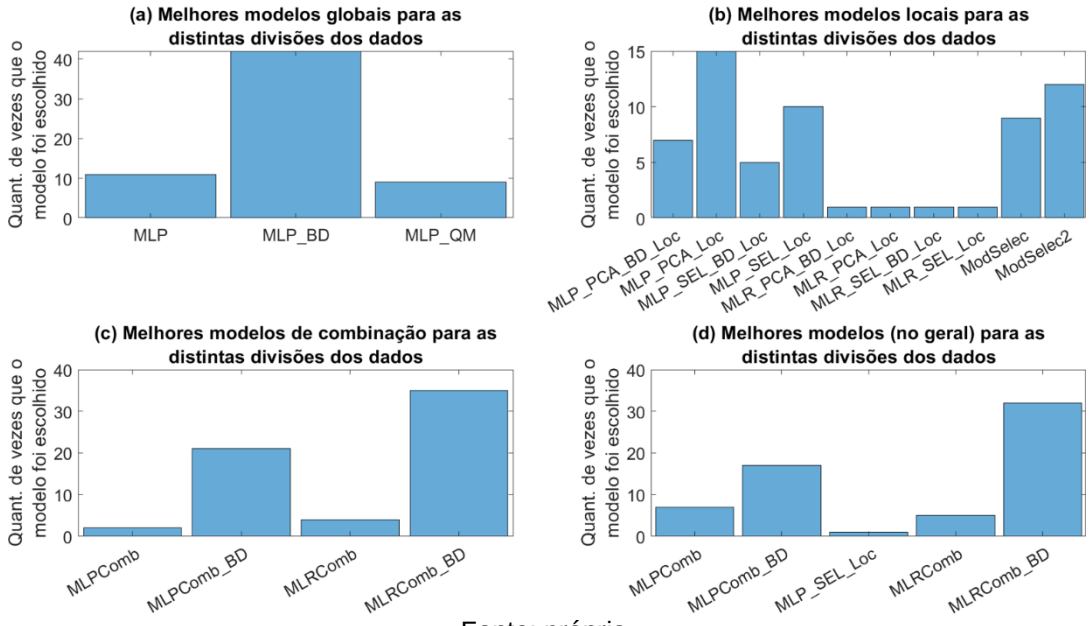
Fonte: própria.

5.4.5 Estação Gobabeb (GHI)

Para a estação de Gobabeb (GHI), os melhores modelos global, local e de combinação foram, respectivamente, MLP_BD, MLP_PCA_Loc e MLRComb_BD, conforme Figura 31. Considerando todos os modelos, a Figura 31.d mostra que MLRComb_BD apresenta os melhores resultados nas distintas divisões dos dados.

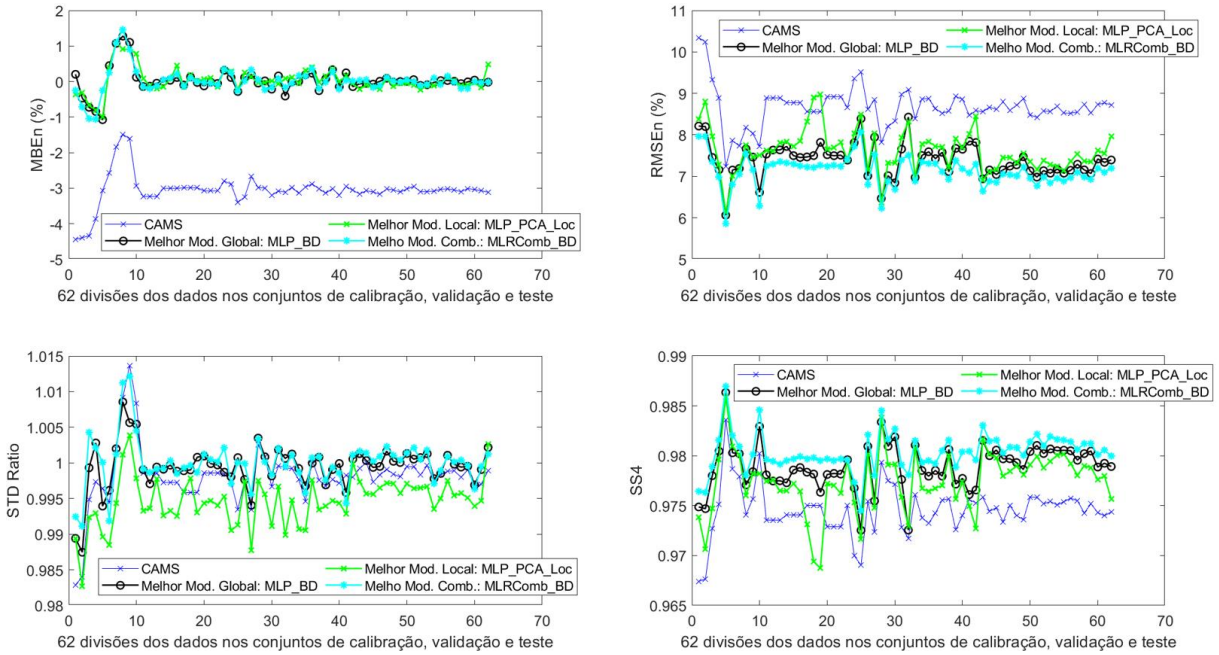
Conforme a Figura 32, os modelos apresentam resultados similares entre si, com RMSEn abaixo de 10% e SS4 com pequenas variações em torno de 0,98. Os modelos globais, locais e de combinação apresentam resultados ligeiramente melhores que a CAMS em termos de MBEn, RMSEn e SS4. Já em relação ao STDRatio, todos os modelos apresentam variações próximas a 1. As melhores divisões dos dados para estação de Gobabeb, componente GHI, foram as divisões 33, 28 e 43 para o melhor modelo global, local e de combinação, respectivamente. A Figura 33 apresenta o diagrama de Taylor no conjunto de teste para a divisão 43. Todos os modelos apresentam resultados similares em termos de correlação com valores próximos a 0,99. Vale salientar que os modelos apresentam bons resultados para essa estação provavelmente devido ao percentual de mais de 80% dos dados pertencerem às classes céu claro ou céu parcialmente claro (região desértica), o que faz o resultado dos modelos apresentarem menos variações.

Figura 31 – Histograma dos melhores modelos considerando as 62 divisões para Gobabeb (GHI).



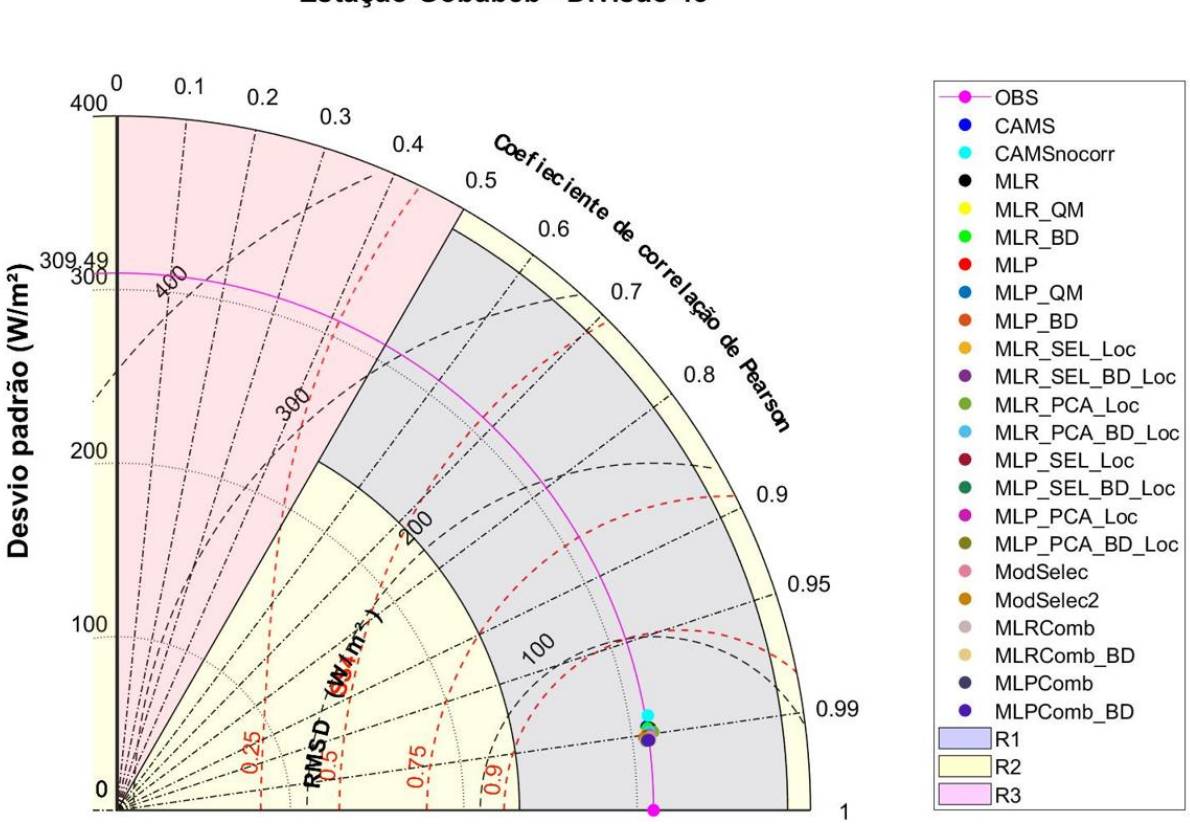
Fonte: própria.

Figura 32 – Resultados dos melhores modelos por categoria para cada uma das 62 divisões para a estação de Gobabeb (GHI).



Fonte: própria.

Figura 33 – Diagrama de Taylor para divisão 43 da estação de Gobabeb (GHI).

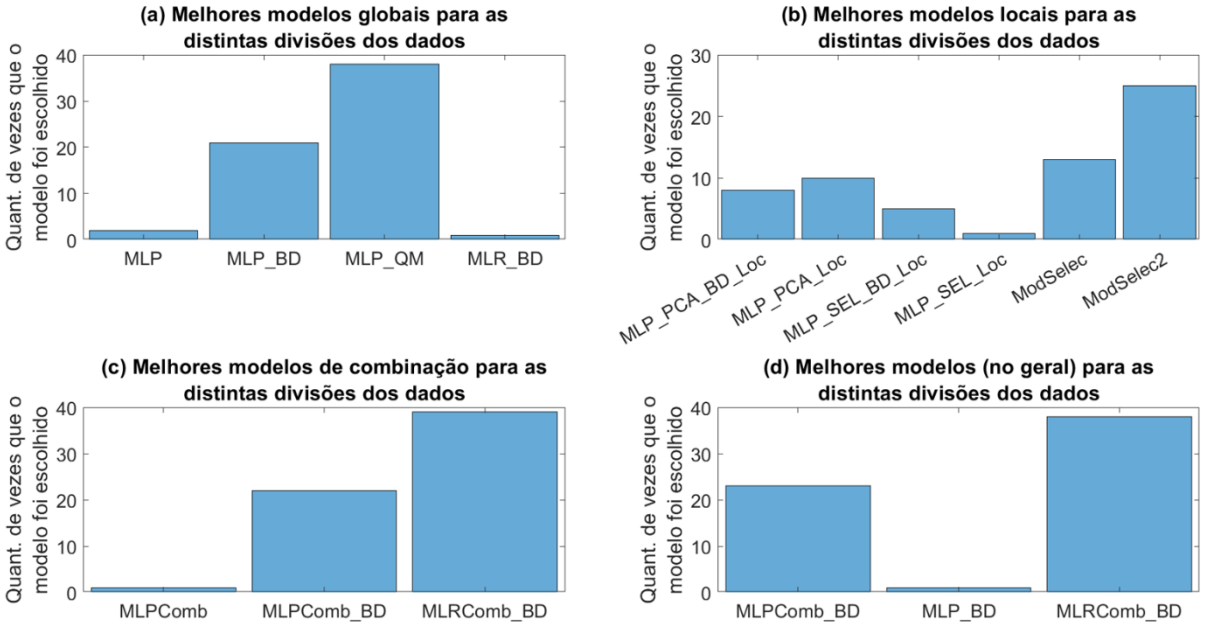


Fonte: própria.

5.4.6 Estação Gobabeb (DNI)

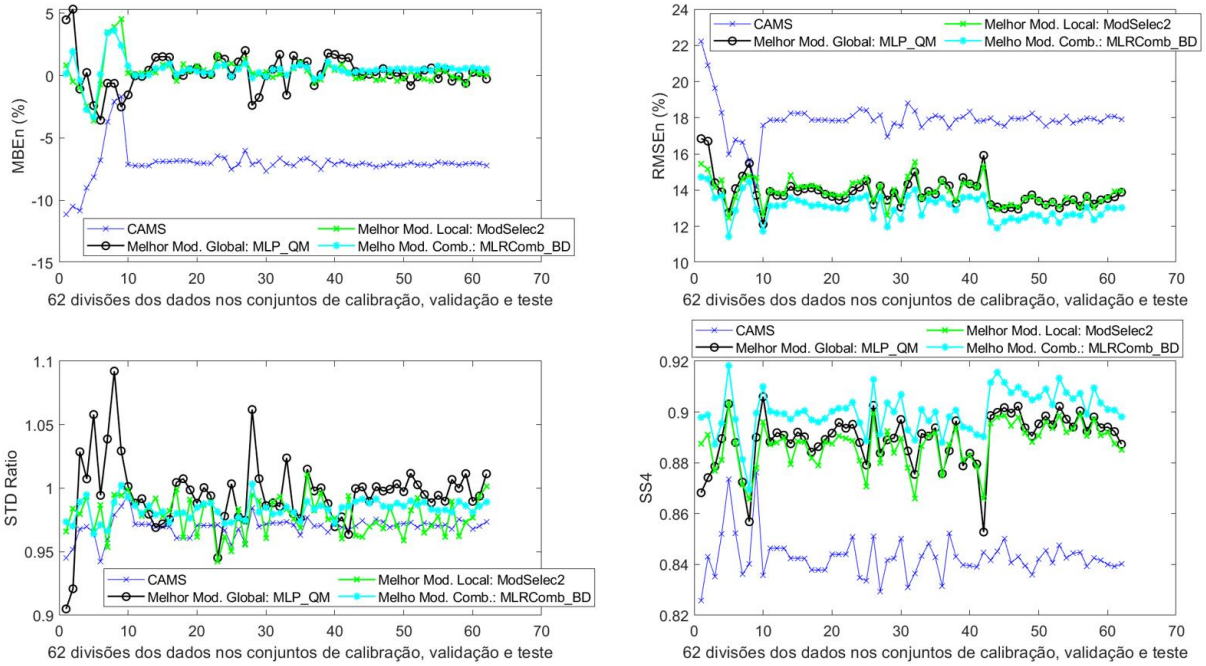
Para a estação de Gobabeb (componente DNI), os melhores modelos global, local e de combinação foram, respectivamente, MLP_QM, ModSelec2 e MLRComb_BD, de acordo com a Figura 34. Considerando todos os modelos, a Figura 34.d mostra que MLRComb_BD apresenta os melhores resultados para as distintas divisões dos dados. Os modelos globais, locais e de combinação ajustam o MBEn, apresentam baixo RMSEn e alto SS4 quando comparados ao modelo de referência da CAMS, como mostra a Figura 35. Os modelos apresentam resultados similares em relação ao STDRatio, com variações entre 0,95 e 1. Os modelos de combinação se destacam por apresentar resultados melhores que os modelos globais, locais e da CAMS. As melhores divisões dos dados para estação de Gobabeb, componente DNI, foram as divisões 46, 10 e 43 para o melhor modelo global, local e de combinação, respectivamente. A Figura 36 apresenta o diagrama de Taylor no conjunto de teste para a divisão 43. O modelo da CAMS tem uma correlação próxima a 0,92, enquanto os modelos locais e de combinação estão próximos a 0,95 de correlação com os dados observados. Já os modelos globais que utilizam o MLR apresentam resultados próximos ao da CAMS, enquanto os que utilizam MLP apresentam resultados melhores.

Figura 34 – Histograma dos melhores modelos considerando as 62 divisões para Gobabeb (DNI).



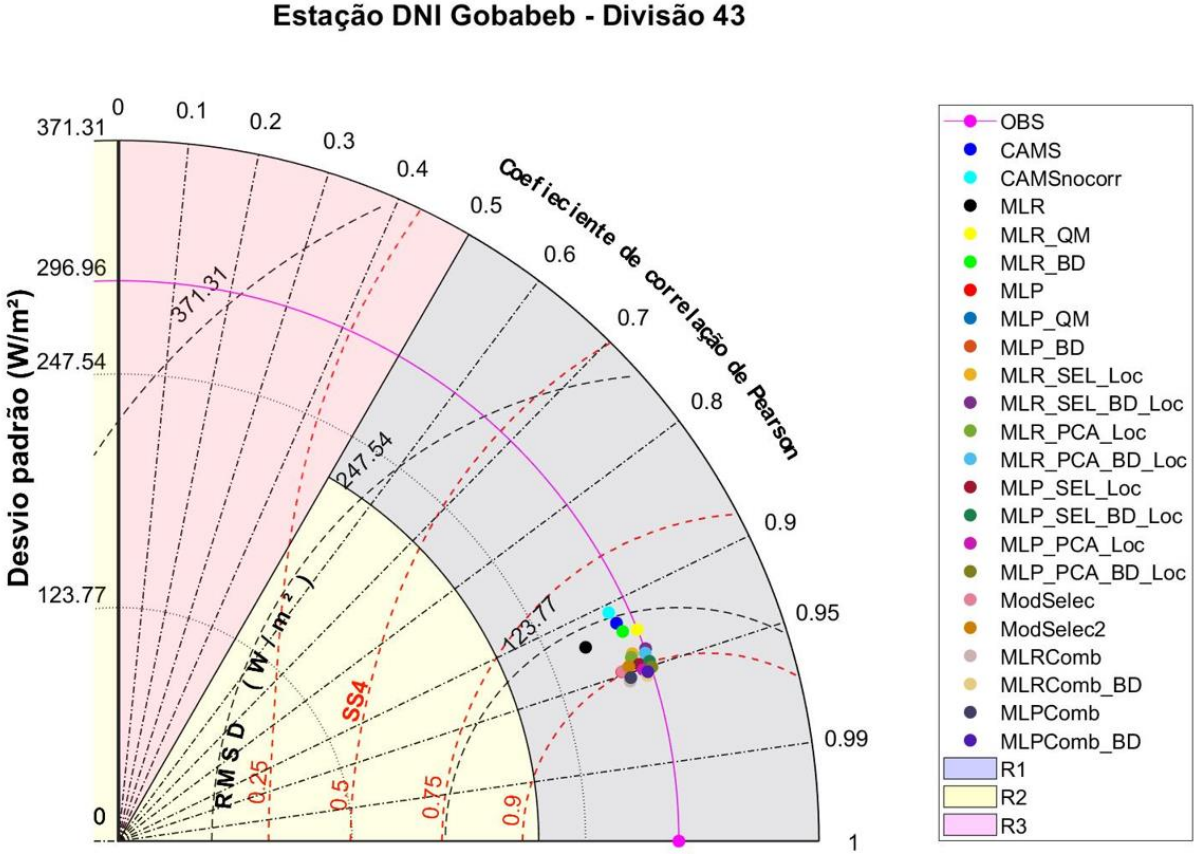
Fonte: própria.

Figura 35 – Resultados dos melhores modelos por categoria para cada uma das 62 divisões para a estação de Gobabeb (DNI).



Fonte: própria.

Figura 36 – Diagrama de Taylor para divisão 43 da estação de Gobabeb (DNI).



Fonte: própria.

5.4.7 Avaliação dos resultados

A Tabela 17 apresenta os resultados para as melhores divisões dos dados considerando o modelo de combinação, conforme apresentado nas seções anteriores. Para El Rosal, o *site adaptation* consegue um resultado muito bom, com uma diminuição no RMSEn de 26,2% e um aumento na correlação de 0,57 para 0,89 quando comparado o modelo de combinação com o da CAMS. Já no *site adaptation* para estação de Salta, o modelo de combinação tem melhores resultados que o modelo da CAMS, contudo, a melhora não é tão expressiva quanto no caso da estação de El Rosal. Para as estações de Petrolina e Gobabeb, componente GHI, os resultados dos 4 modelos são similares entre si. Já no caso da DNI para as estações Petrolina e Gobabeb, os modelos de combinação apresentam os melhores resultados (maiores SS4 e menores RMSEn).

A Tabela 17 também mostra os resultados para irradiância difusa horizontal (DHI) no caso das estações de Petrolina e Gobabeb, que foi obtida por diferença utilizando os resultados do *site adaptation* das outras duas componentes da radiação. Lembrado que a irradiância global horizontal (GHI) é igual a soma da DHI com a irradiância direta horizontal, esta última obtida ao multiplicar a DNI pelo cosseno do ângulo zenital. Por exemplo, o melhor modelo global para DHI da estação de Petrolina foi o *Multilayer Perceptron* (MLP), sendo essa DHI obtida pela diferença entre o modelo MLP da GHI pelo modelo MLP da DNI. Assim, a difusa foi calculada por diferença considerando todos os modelos globais, locais e de combinação de GHI e DNI, sendo apresentados na Tabela 17 os modelos com melhores resultados para cada categoria. Vale salientar que a divisão dos dados considerada para calcular a radiação difusa foi a divisão que apresentou os melhores resultados para DNI. Para ambas as estações, os melhores resultados para DHI estão nos modelos de combinação que apresentam menores dispersões e maiores correlações, sobretudo para a estação de Gobabeb, em que foi possível diminuir o RMSEn em -11,8% em relação ao modelo da CAMS.

Tabela 17 – Resultados das melhores divisões (de acordo com os modelos de combinação).

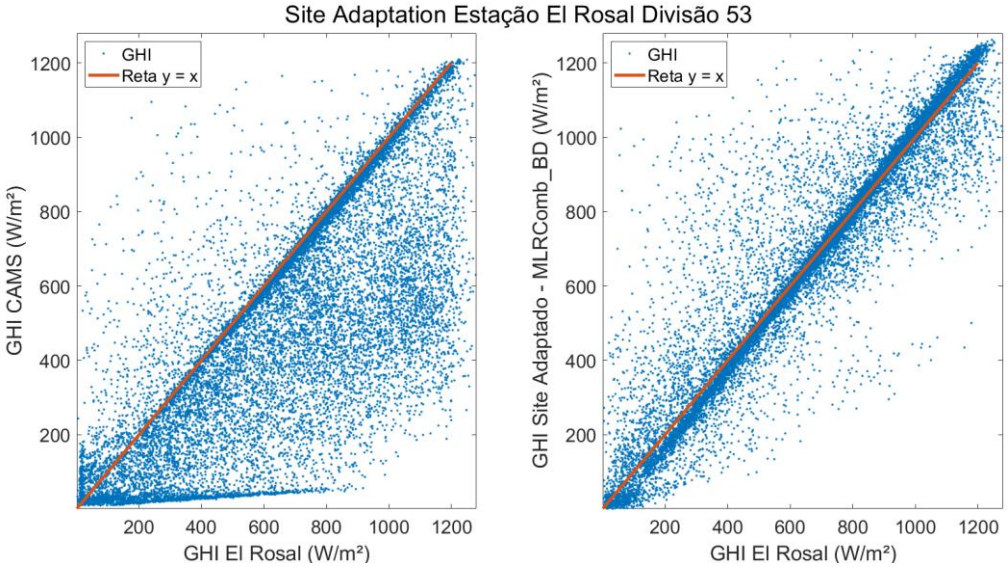
Estação	Comp. da radiação	Modelos	Divisão	MBEn	STDRatio	RMSEn	ρ	SS4
El Rosal	GHI	CAMS	53	-25,4%	0,95	43,1%	0,74	0,57
		Global (MLP_QM)		-0,5%	1,00	20,4%	0,91	0,84
		Local (ModSelec)		0,8%	0,99	17,6%	0,94	0,88
		Comb. (MLRComb_BD)		0,5%	1,00	16,9%	0,94	0,89
Salta	GHI	CAMS	46	4,6%	1,00	30,7%	0,91	0,83
		Global (MLP_BD)		-0,2%	1,00	27,7%	0,92	0,85
		Local (ModSelec)		0,6%	0,97	27,4%	0,92	0,85
		Comb. (MLRComb_BD)		-0,2%	1,01	25,9%	0,93	0,87
Petrolina	GHI	CAMS	46	5,0%	0,97	22,7%	0,93	0,86
		Global (MLP_BD)		-0,3%	1,00	21,3%	0,93	0,88
		Local (ModSelec)		0,6%	0,97	21,2%	0,93	0,87
		Comb. (MLRComb_BD)		-0,1%	1,00	21,1%	0,94	0,88
Petrolina	DNI	CAMS	44	21,5%	1,00	51,3%	0,84	0,71
		Global (MLP_QM)		2,1%	1,00	43,5%	0,86	0,74
		Local (ModSelec2)		4,7%	0,97	43,2%	0,86	0,74
		Comb. (MLPComb_BD)		2,2%	0,98	40,9%	0,87	0,77
Petrolina	DHI (obtida por diferença)	CAMS	44	-12,8%	0,93	34,5%	0,85	0,73
		Global (MLP)		5,7%	0,97	29,9%	0,87	0,77
		Local (ModSelec2)		-0,9%	1,08	31,8%	0,88	0,77
		Comb. (MLRComb)		6,5%	0,99	29,8%	0,89	0,79
Gobabeb	GHI	CAMS	43	0,0%	1,00	7,8%	0,99	0,98
		Global (MLP_BD)		-0,1%	1,00	6,9%	0,99	0,98
		Local (MLP_PCA_Loc)		-0,2%	1,00	6,9%	0,99	0,98
		Comb. (MLRComb_BD)		0,0%	1,00	6,6%	0,99	0,98
Gobabeb	DNI	CAMS	43	-7,2%	0,97	18,0%	0,92	0,84
		Global (MLP_QM)		0,3%	1,00	13,2%	0,95	0,90
		Local (ModSelec2)		-0,2%	0,96	13,2%	0,95	0,90
		Comb. (MLRComb_BD)		0,3%	0,99	12,2%	0,95	0,91
Gobabeb	DHI (obtida por diferença)	CAMS	43	8,6%	0,79	40,5%	0,85	0,70
		Global (MLP_BD)		1,0%	1,00	31,9%	0,91	0,83
		Local (MLP_PCA_Loc)		1,7%	1,00	30,2%	0,92	0,85
		Comb. (MLRComb_BD)		0,6%	1,01	28,7%	0,93	0,86

Fonte: própria.

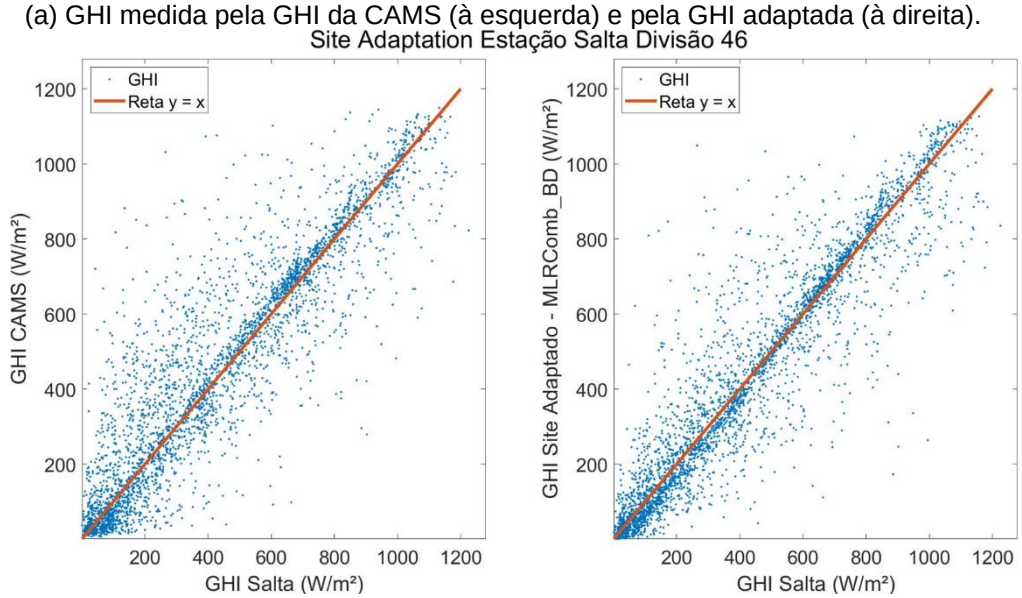
As Figuras de 37 a 39 apresentam as dispersões finais entre os dados medidos de radiação *versus* o modelo da CAMS (gráficos à esquerda) e *versus* o melhor modelo para o *site adaptation* (gráficos à direita), considerando as divisões da Tabela 17. A Figura 37 apresenta essas dispersões para GHI das estações de El Rosal e Salta, enquanto as Figuras 38 e 39 apresentam as dispersões de GHI, DNI e DHI para as estações de Petrolina e Gobabeb, respectivamente. Pode-se notar que, no geral, o *site adaptation* diminui a dispersão e alinha os dados melhor em torno da reta x = y,

sobretudo para a estação de El Rosal e as componentes DNI e DHI das estações de Gobabeb e Petrolina.

Figura 37 – Dispersões dos melhores modelos e divisões dos dados para GHI das estações de Salta e El Rosal.



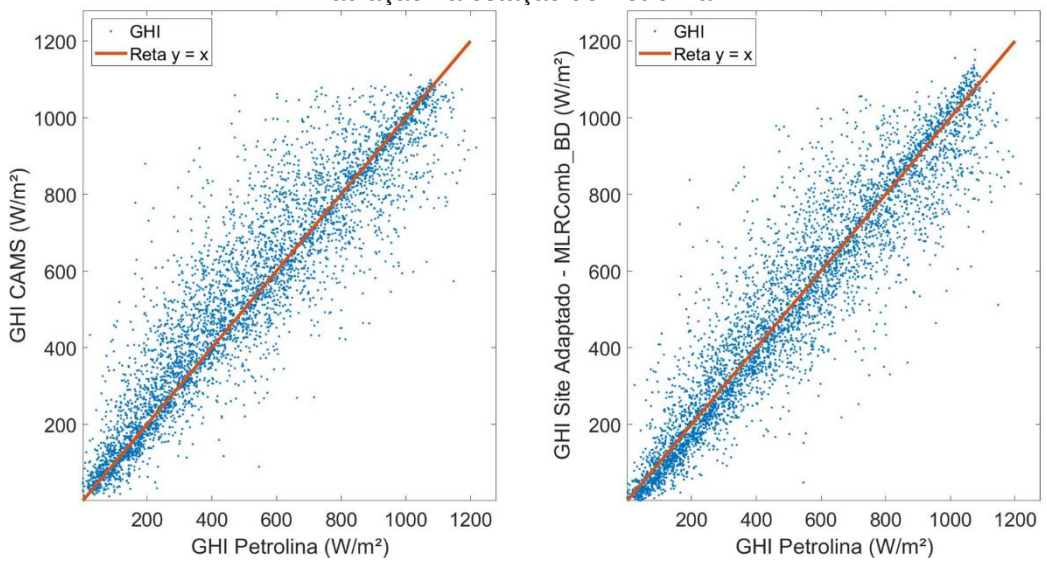
(a) GHI medida pela GHI da CAMS (à esquerda) e pela GHI adaptada (à direita).



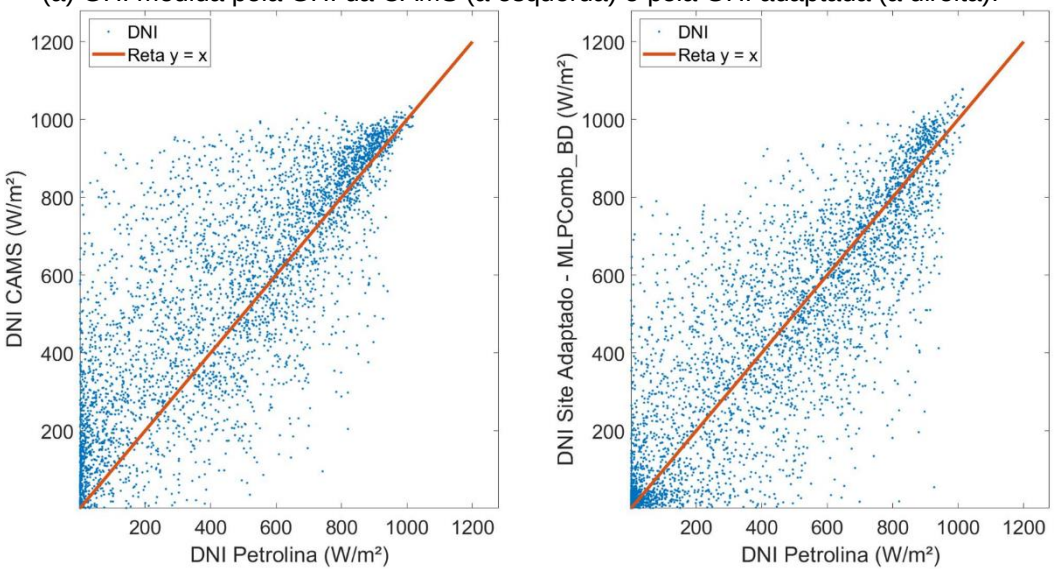
(b) GHI medida pela GHI da CAMS (à esquerda) e pela GHI adaptada (à direita).

Fonte: própria.

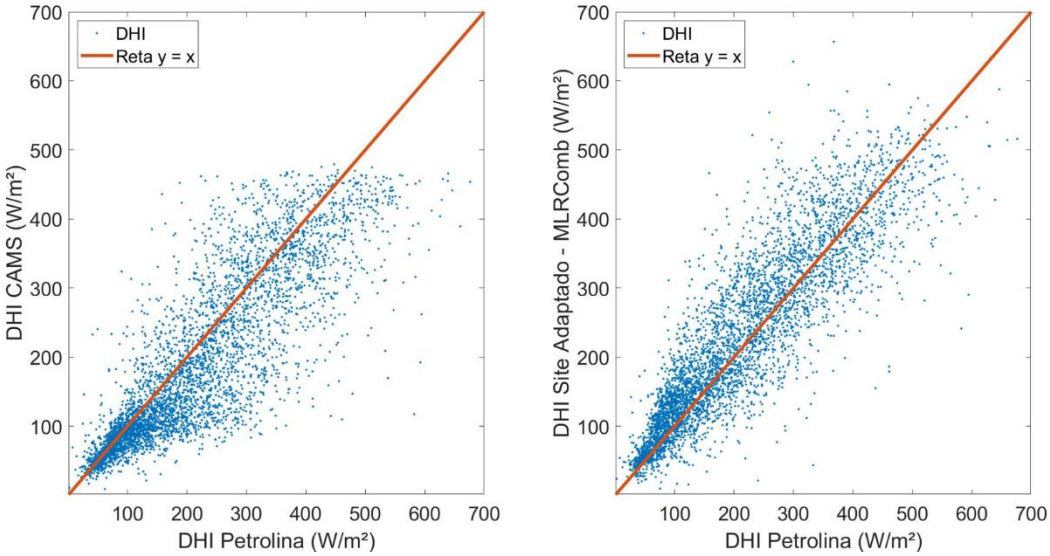
Figura 38 – Dispersões dos melhores modelos e divisões dos dados para as três componentes da radiação na estação de Petrolina.



(a) GHI medida pela GHI da CAMS (à esquerda) e pela GHI adaptada (à direita).

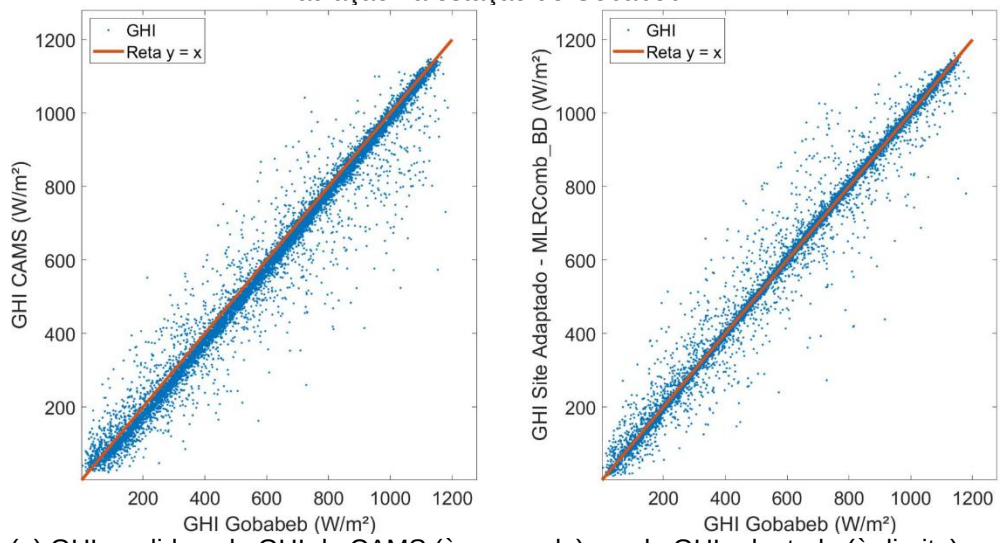


(b) DNI medida pela DNI da CAMS (à esquerda) e *versus* pela DNI adaptada (à direita).

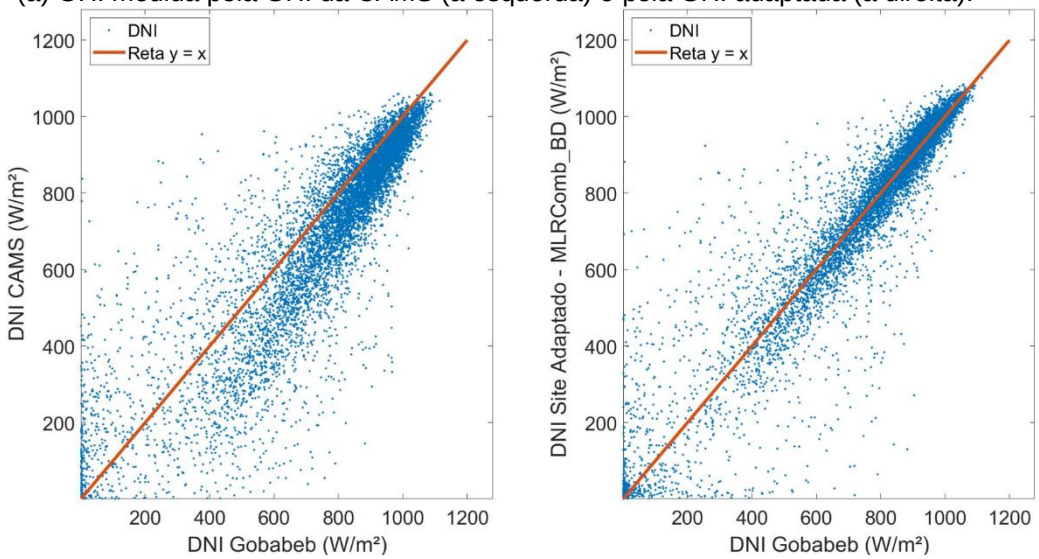


(c) DHI medida pela DHI da CAMS (à esquerda) e pela DHI adaptada (à direita).
Fonte: própria.

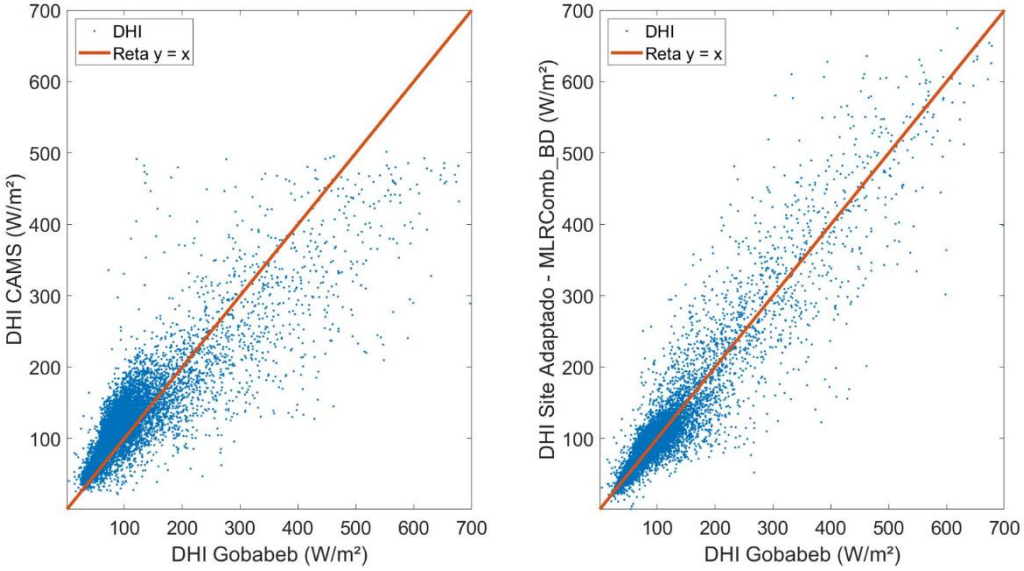
Figura 39 – Dispersões dos melhores modelos e divisões dos dados para as três componentes da radiação na estação de Gobabeb.



(a) GHI medida pela GHI da CAMS (à esquerda) e pela GHI adaptada (à direita).



(b) DNI medida pela DNI da CAMS (à esquerda) e pela DNI adaptada (à direita).



(c) DHI medida pela DHI da CAMS (à esquerda) e pela DHI adaptada (à direita).
Fonte: própria.

Os resultados da Tabela 17 também podem ser observados nos subconjuntos locais das séries temporais, obtidos de acordo com as condições do céu em casa região. A Tabela 18 apresenta os resultados da divisão 53 para cada subconjunto dos dados da estação de El Rosal, ordenados pelo SS4. O objetivo é avaliar a acurácia dos modelos nos subconjuntos das séries. Os resultados com maiores SS4 estão destacados em negrito, enquanto o modelo de combinação MLRComb_BD é destacado em laranja para comparação. Para a classe céu claro, o modelo mais acurado é o MLP_PCA_BD_Loc, pois apresenta maior correlação e menor RMSEn. Nos casos das classes 2, 3 e 5, MLRComb_BD está entre os modelos mais acurados. Por fim, para a classe 4, o modelo mais acurado é o MLR_PCA_BD_Loc, enquanto o modelo MLRComb_BD apresenta um STDRatio baixo (0,54).

Tabela 18 – Resultados de GHI para as melhores divisões por classe na estação de El Rosal no conjunto de teste.

Continua

El Rosal - Divisão 53 - Subconjunto Céu Claro (classe 1)							
Modelos	MBE (W/m²)	MBEn	STDRatio	RMSE (W/m²)	RMSEn	ρ	SS4
CAMS	-129,7	-17,8%	1,09	230,2	31,6%	0,73	0,56
MLP_PCA_BD_Loc	**4,1**	**0,6%**	**0,98**	**34,2**	**4,7%**	**0,99**	**0,98**
MLP_PCA_Loc	4,1	0,6%	0,98	34,2	4,7%	0,99	0,98
MLP_SEL_BD_Loc	4,4	0,6%	0,99	34,7	4,8%	0,99	0,98
MLP_SEL_Loc	4,2	0,6%	0,99	34,7	4,8%	0,99	0,98
MLPComb	3,3	0,4%	0,99	35,0	4,8%	0,99	0,98
MLRComb	-2,4	-0,3%	0,98	34,9	4,8%	0,99	0,98
MLR_SEL_Loc	4,3	0,6%	0,99	35,8	4,9%	0,99	0,98
MLR_SEL_BD_Loc	4,3	0,6%	0,99	35,8	4,9%	0,99	0,98
MLPComb_BD	2,4	0,3%	1,03	36,6	5,0%	0,99	0,98
MLR_PCA_Loc	4,1	0,6%	1,00	36,2	5,0%	0,99	0,98
MLRComb_BD	**1,8**	**0,2%**	**1,05**	**37,3**	**5,1%**	**0,99**	**0,98**
El Rosal - Divisão 53 - Subconjunto Céu Parcialmente Claro (classe 2)							
Modelos	MBE (W/m²)	MBEn	STDRatio	RMSE (W/m²)	RMSEn	ρ	SS4
CAMS	-241,2	-33,3%	1,02	331,9	45,9%	0,74	0,57
MLRComb_BD	**12,9**	**1,8%**	**1,03**	**85,8**	**11,9%**	**0,96**	**0,93**
MLPComb	13,7	1,9%	0,98	84,5	11,7%	0,96	0,93
MLRComb	8,3	1,1%	0,96	83,0	11,5%	0,96	0,93
MLPComb_BD	13,1	1,8%	1,02	86,2	11,9%	0,96	0,93
MLR_PCA_Loc	28,9	4,0%	0,98	89,0	12,3%	0,96	0,93
MLP_PCA_BD_Loc	28,4	3,9%	0,97	88,5	12,2%	0,96	0,93
MLP_PCA_Loc - MS - MS2	27,8	3,8%	0,97	88,3	12,2%	0,96	0,93
MLR_PCA_BD_Loc	28,8	4,0%	0,97	88,8	12,3%	0,96	0,93
MLP_SEL_BD_Loc	28,3	3,9%	0,97	88,8	12,3%	0,96	0,93
MLP_SEL_Loc	27,8	3,8%	0,96	88,6	12,3%	0,96	0,93

Tabela 18 – Resultados de GHI para as melhores divisões por classe na estação de El Rosal no conjunto de teste.

Conclusão

El Rosal - Divisão 53 - Subconjunto Céu Nublado (classe 3)							
Modelos	MBE (W/m²)	MBEn	STDRatio	RMSE (W/m²)	RMSEn	ρ	SS4
CAMS	-44,4	-22,7%	0,89	202,1	103,3%	0,41	0,24
MLRComb_BD	**-25,1**	**-12,8%**	**0,77**	**147,7**	**75,5%**	**0,66**	**0,44**
MLP	87,7	44,8%	1,24	206,0	105,4%	0,64	0,43
MLRComb	5,5	2,8%	0,73	144,5	73,9%	0,66	0,43
MLP_QM	62,1	31,7%	1,26	201,2	102,9%	0,63	0,42
MLPComb_BD	-21,0	-10,7%	0,73	148,1	75,7%	0,65	0,42
MLR	129,6	66,3%	1,23	229,5	117,3%	0,62	0,42
MLP_BD	64,2	32,8%	1,30	206,1	105,4%	0,63	0,41
MLR_BD - MS2	115,9	59,3%	1,28	227,1	116,1%	0,62	0,41
MLPComb	3,8	2,0%	0,69	146,3	74,8%	0,65	0,40
MLR_QM	96,2	49,2%	1,38	227,6	116,4%	0,63	0,40
El Rosal - Divisão 53 - Subconjunto Céu Variável (classe 4)							
Modelos	MBE (W/m²)	MBEn	STDRatio	RMSE (W/m²)	RMSEn	ρ	SS4
CAMS	-243,6	-29,1%	0,89	390,4	46,6%	0,27	0,16
MLR_PCA_BD_Loc	**19,1**	**2,3%**	**0,73**	**242,6**	**28,9%**	**0,48**	**0,27**
MLP_SEL_BD_Loc	18,4	2,2%	0,72	242,2	28,9%	0,48	0,27
MLR_SEL_BD_Loc	19,6	2,3%	0,73	243,2	29,0%	0,48	0,27
MLP_PCA_BD_Loc	15,5	1,8%	0,74	246,7	29,4%	0,47	0,26
MLR_BD	76,3	9,1%	0,69	253,4	30,2%	0,47	0,25
MLR	67,0	8,0%	0,66	249,3	29,7%	0,47	0,25
MLR_QM	72,9	8,7%	0,65	250,0	29,8%	0,47	0,24
MLP_QM	43,4	5,2%	0,73	256,2	30,6%	0,43	0,24
MLP_BD	45,8	5,5%	0,70	253,9	30,3%	0,43	0,23
MLPComb_BD	19,0	2,3%	0,57	236,1	28,2%	0,48	0,22
MLRComb_BD	**19,4**	**2,3%**	**0,54**	**232,7**	**27,8%**	**0,49**	**0,22**
El Rosal - Divisão 53 - Subconjunto Céu Parcialmente Variável (classe 5)							
Modelos	MBE (W/m²)	MBEn	STDRatio	RMSE (W/m²)	RMSEn	ρ	SS4
CAMS	-189,5	-35,6%	0,92	287,6	54,1%	0,65	0,46
MLPComb_BD	**4,0**	**0,8%**	**0,99**	**124,7**	**23,4%**	**0,89**	**0,80**
MLRComb_BD	**2,6**	**0,5%**	**0,95**	**123,2**	**23,2%**	**0,89**	**0,80**
MLPComb	13,3	2,5%	0,94	123,5	23,2%	0,89	0,80
MLR_SEL_BD_Loc	6,2	1,2%	0,97	125,6	23,6%	0,89	0,79
MLP_PCA_BD_Loc	4,9	0,9%	0,96	125,2	23,5%	0,89	0,79
MLP_SEL_BD_Loc	3,9	0,7%	0,96	126,1	23,7%	0,89	0,79
MLRComb	10,9	2,0%	0,89	122,6	23,1%	0,89	0,79
MLR_SEL_Loc	6,4	1,2%	0,92	124,0	23,3%	0,89	0,79
MLR_PCA_BD_Loc - MS	5,1	1,0%	0,96	127,3	23,9%	0,88	0,79
MLP_PCA_Loc - MS2	-2,1	-0,4%	0,90	123,6	23,2%	0,89	0,79

Fonte: própria.

A Tabela 19 apresenta os resultados da divisão 46 para cada subconjunto dos dados da estação de Salta. Os resultados com maior SS4 estão destacados em negrito. Para as classes 1, 2, 3 e 4 os modelos com maior SS4 são, respectivamente, MLP_PCA_BD_Loc, MLPComb, MLP_BD e CAMS. Para a classe 5, o modelo com maior SS4 e com RMSEn baixo em comparação com os outros modelos foi o MLRComb_BD. Na classe 3, apesar do modelo MLRComb_BD não apresentar o maior SS4, possui um RMSEn 7,6% mais baixo que o modelo com maior SS4 (MLP_BD).

Tabela 19 – Resultados de GHI para as melhores divisões por classe na estação de Salta no conjunto de teste.

Continua

Salta - Divisão 46 - Subconjunto Céu Claro (classe 1)							
Modelos	MBE (W/m²)	MBEn	STDRatio	RMSE (W/m²)	RMSEn	ρ	SS4
CAMS	4,6	0,7%	1,02	38,5	6,3%	0,98	0,96
MLP_PCA_BD_Loc	**2,3**	**0,4%**	**1,00**	**27,2**	**4,4%**	**0,99**	**0,98**
MLR_SEL_Loc	2,4	0,4%	1,00	27,2	4,4%	0,99	0,98
MLR_SEL_BD_Loc	2,4	0,4%	1,00	27,2	4,4%	0,99	0,98
MLP_PCA_Loc - MS - MS2	2,4	0,4%	1,00	27,2	4,4%	0,99	0,98
MLP_SEL_BD_Loc	2,1	0,3%	1,00	27,6	4,5%	0,99	0,98
MLP_SEL_Loc	2,4	0,4%	0,99	27,6	4,5%	0,99	0,98
MLR_PCA_Loc	2,2	0,4%	1,01	28,3	4,6%	0,99	0,98
MLR_PCA_BD_Loc	1,9	0,3%	0,99	28,1	4,6%	0,99	0,98
MLRComb	-13,9	-2,3%	0,98	31,4	5,1%	0,99	0,98
MLPComb	-11,9	-1,9%	1,00	31,8	5,2%	0,99	0,98
MLRComb_BD	**2,6**	**0,4%**	**1,07**	**32,3**	**5,3%**	**0,99**	**0,97**
Salta - Divisão 46 - Subconjunto Céu Parcialmente Claro (classe 2)							
Modelos	MBE (W/m²)	MBEn	STDRatio	RMSE (W/m²)	RMSEn	ρ	SS4
CAMS	5,2	0,9%	1,04	81,8	14,7%	0,96	0,92
MLPComb	**-0,3**	**-0,1%**	**0,97**	**71,8**	**12,9%**	**0,97**	**0,93**
MLPComb_BD	11,1	2,0%	1,05	76,4	13,7%	0,97	0,93
MLRComb_BD	**12,9**	**2,3%**	**1,04**	**76,6**	**13,7%**	**0,97**	**0,93**
MLRComb	0,3	0,1%	0,95	72,7	13,0%	0,97	0,93
MLP_PCA_BD_Loc	18,4	3,3%	0,98	79,4	14,2%	0,96	0,92
MLP_PCA_Loc - MS - MS2	19,2	3,4%	0,97	79,5	14,3%	0,96	0,92
MLP_BD	0,2	0,0%	1,04	80,2	14,4%	0,96	0,92
MLP	-10,0	-1,8%	0,96	77,5	13,9%	0,96	0,92
MLP_SEL_BD_Loc	19,0	3,4%	0,98	80,4	14,4%	0,96	0,92
MLP_SEL_Loc	19,6	3,5%	0,97	80,4	14,4%	0,96	0,92

Tabela 19 – Resultados de GHI para as melhores divisões por classe na estação de Salta no conjunto de teste.

Conclusão

Salta - Divisão 46 - Subconjunto Céu Nublado (classe 3)							
Modelos	MBE (W/m²)	MBEn	STDRatio	RMSE (W/m²)	RMSEn	ρ	SS4
CAMS	35,8	27,4%	1,32	148,1	113,5%	0,53	0,32
MLP_BD	**-4,0**	**-3,1%**	**0,94**	**109,1**	**83,5%**	**0,59**	**0,40**
MLP - MS	19,1	14,7%	0,88	107,4	82,2%	0,60	0,40
MLP_QM - MS2	-0,4	-0,3%	0,86	105,8	81,1%	0,59	0,39
MLPComb_BD	-21,9	-16,8%	0,64	96,1	73,6%	0,66	0,39
MLRComb_BD	**-25,1**	**-19,2%**	**0,67**	**99,1**	**75,9%**	**0,64**	**0,39**
MLR	56,8	43,5%	1,03	131,8	100,9%	0,56	0,37
MLPComb	0,1	0,0%	0,60	93,8	71,9%	0,66	0,37
MLRComb	-0,5	-0,4%	0,63	95,6	73,2%	0,64	0,37
MLR_BD	33,0	25,3%	1,13	129,6	99,3%	0,56	0,36
MLR_QM	26,9	20,6%	1,06	124,7	95,5%	0,55	0,36
Salta - Divisão 46 - Subconjunto Céu Variável (classe 4)							
Modelos	MBE (W/m²)	MBEn	STDRatio	RMSE (W/m²)	RMSEn	ρ	SS4
CAMS	**49,4**	**7,4%**	**0,74**	**216,0**	**32,5%**	**0,58**	**0,35**
CAMSnocorr	30,8	4,6%	0,79	218,7	32,9%	0,56	0,35
MLP_QM	-5,6	-0,8%	0,79	218,3	32,9%	0,55	0,34
MLP_BD	6,0	0,9%	0,74	215,2	32,4%	0,56	0,33
MLR_QM - MS - MS2	-16,1	-2,4%	0,79	221,4	33,3%	0,54	0,33
MLR_BD	-7,9	-1,2%	0,78	221,6	33,4%	0,54	0,33
MLP_SEL_BD_Loc	52,9	8,0%	0,72	222,7	33,5%	0,54	0,32
MLP	-13,2	-2,0%	0,68	212,8	32,0%	0,56	0,32
MLPComb_BD	26,0	3,9%	0,61	206,1	31,0%	0,59	0,31
MLR	-28,6	-4,3%	0,71	218,9	33,0%	0,54	0,31
MLRComb_BD	**27,7**	**4,2%**	**0,59**	**208,6**	**31,4%**	**0,57**	**0,29**
Salta - Divisão 46 - Subconjunto Céu Parcialmente Variável (classe 5)							
Modelos	MBE (W/m²)	MBEn	STDRatio	RMSE (W/m²)	RMSEn	ρ	SS4
MLRComb_BD	**-22,8**	**-7,5%**	**0,73**	**125,1**	**41,0%**	**0,73**	**0,51**
MLPComb_BD	-15,0	-4,9%	0,71	124,2	40,7%	0,73	0,50
MLP_BD	-2,1	-0,7%	0,97	142,1	46,6%	0,68	0,50
MLP - MS2	8,1	2,7%	0,90	137,7	45,2%	0,68	0,49
MLP_QM	-14,0	-4,6%	0,92	140,5	46,1%	0,67	0,49
MLRComb	-11,5	-3,8%	0,67	124,1	40,7%	0,73	0,48
MLR_BD	-22,2	-7,3%	0,92	144,6	47,4%	0,66	0,47
MLPComb	-6,0	-2,0%	0,66	124,0	40,7%	0,73	0,47
CAMS	16,3	5,3%	1,05	154,9	50,8%	0,65	0,46
MLR	-8,2	-2,7%	0,84	138,7	45,5%	0,66	0,46

Fonte: própria.

A Tabela 20 apresenta os resultados da divisão 46 para cada subconjunto dos dados da estação de Petrolina (GHI). Os resultados com maior SS4 estão destacados em negrito. Todos os modelos apresentam resultados próximos para a classe 1, com

ótima acurácia. Na classe 3, o modelo global MLP_BD apresenta SS4 similar ao modelo MLRComb_BD, porém com um melhor STDRatio. Já na classe 4, o modelo MLRComb_BD apresenta o menor SS4 e o STDRatio mais distante de 1.

Tabela 20 – Resultados de GHI para as melhores divisões por classe na estação de Petrolina no conjunto de teste.

Continua

Petrolina (GHI) - Divisão 46 - Subconjunto Céu Claro (classe 1)							
Modelos	MBE (W/m²)	MBEn	STDRatio	RMSE (W/m²)	RMSEn	ρ	SS4
CAMS	**2,6**	**0,5%**	**0,99**	**33,3**	**6,0%**	**0,99**	**0,99**
MLPComb	-0,2	0,0%	0,99	34,4	6,2%	0,99	0,99
MLR_PCA_Loc	8,9	1,6%	0,98	35,5	6,3%	0,99	0,99
CAMSnocorr	16,7	3,0%	0,99	38,4	6,9%	0,99	0,99
MLP_PCA_BD_Loc	8,9	1,6%	0,98	35,7	6,4%	0,99	0,99
MLR_PCA_BD_Loc	8,9	1,6%	0,98	35,7	6,4%	0,99	0,99
MLP_PCA_Loc	8,7	1,6%	0,98	35,7	6,4%	0,99	0,99
MLR_SEL_BD_Loc	8,7	1,6%	0,98	35,7	6,4%	0,99	0,99
MLR_SEL_Loc	8,7	1,6%	0,98	35,7	6,4%	0,99	0,99
MLP_SEL_BD_Loc	8,7	1,6%	0,98	35,9	6,4%	0,99	0,99
MLRComb_BD	**2,0**	**0,4%**	**1,05**	**38,6**	**6,9%**	**0,99**	**0,98**
Petrolina (GHI) - Divisão 46 - Subconjunto Céu Parcialmente Claro (classe 2)							
Modelos	MBE (W/m²)	MBEn	STDRatio	RMSE (W/m²)	RMSEn	ρ	SS4
MLRComb_BD	**7,9**	**1,4%**	**1,01**	**97,1**	**17,3%**	**0,94**	**0,88**
MLPComb_BD	6,2	1,1%	1,01	97,2	17,3%	0,94	0,88
MLP_BD	-3,3	-0,6%	1,00	96,5	17,2%	0,94	0,88
MLPComb	-2,0	-0,4%	0,95	95,0	16,9%	0,94	0,88
CAMS	25,3	4,5%	0,98	99,6	17,7%	0,94	0,88
MLR_BD	-4,8	-0,8%	1,02	99,3	17,7%	0,94	0,88
MLP_QM - MS	-8,6	-1,5%	1,00	98,2	17,5%	0,94	0,88
MLRComb	2,1	0,4%	0,94	94,9	16,9%	0,94	0,88
MLP - MS2	-4,1	-0,7%	0,94	95,3	17,0%	0,94	0,88
MLR	-7,6	-1,3%	0,96	96,7	17,2%	0,94	0,88
Petrolina (GHI) - Divisão 46 - Subconjunto Céu Nublado (classe 3)							
Modelos	MBE (W/m²)	MBEn	STDRatio	RMSE (W/m²)	RMSEn	ρ	SS4
MLP_BD	**-16,4**	**-7,5%**	**0,92**	**94,8**	**43,2%**	**0,80**	**0,65**
MLRComb_BD	**-23,4**	**-10,7%**	**0,86**	**93,8**	**42,8%**	**0,80**	**0,65**
MLP	3,8	1,7%	0,87	92,1	42,0%	0,80	0,64
MLPComb_BD	-19,8	-9,0%	0,82	92,4	42,1%	0,80	0,64
MLP_QM - MS2	-10,1	-4,6%	0,83	92,1	42,0%	0,80	0,63
MLRComb	-2,3	-1,1%	0,81	90,4	41,2%	0,80	0,63
CAMS	26,8	12,2%	1,03	105,4	48,1%	0,78	0,63
MLPComb	-6,7	-3,1%	0,77	90,6	41,3%	0,80	0,62
MLR_BD - MS	1,0	0,5%	0,98	102,0	46,5%	0,77	0,61
MLR	19,5	8,9%	0,92	101,6	46,3%	0,77	0,61

Tabela 20 – Resultados de GHI para as melhores divisões por classe na estação de Petrolina no conjunto de teste.

Conclusão

Petrolina (GHI) - Divisão 46 - Subconjunto Céu Variável (classe 4)							
Modelos	MBE (W/m²)	MBEn	STDRatio	RMSE (W/m²)	RMSEn	ρ	SS4
MLP_QM	**6,7**	**0,9%**	**0,77**	**164,6**	**21,1%**	**0,60**	**0,39**
MLP_PCA_BD_Loc	-4,8	-0,6%	0,79	167,9	21,5%	0,59	0,38
MLP_SEL_BD_Loc	-2,3	-0,3%	0,79	167,9	21,5%	0,59	0,38
MLR_SEL_BD_Loc	1,3	0,2%	0,78	167,1	21,4%	0,59	0,38
MLR_BD	8,8	1,1%	0,79	168,9	21,6%	0,59	0,38
MLR_QM - MS2	5,0	0,6%	0,78	168,3	21,6%	0,59	0,37
CAMSnocorr	19,6	2,5%	0,85	176,4	22,6%	0,57	0,37
MLR_PCA_BD_Loc	0,4	0,1%	0,75	166,2	21,3%	0,59	0,37
MLP_BD	12,7	1,6%	0,70	162,8	20,9%	0,60	0,37
MLPComb_BD	10,6	1,4%	0,68	161,0	20,6%	0,61	0,37
MLRComb_BD	**9,9**	**1,3%**	**0,66**	**160,5**	**20,6%**	**0,61**	**0,36**
Petrolina (GHI) - Divisão 46 - Subconjunto Céu Parcialmente Variável (classe 5)							
Modelos	MBE (W/m²)	MBEn	STDRatio	RMSE (W/m²)	RMSEn	ρ	SS4
MLRComb_BD	**-0,4**	**-0,1%**	**0,87**	**115,2**	**26,2%**	**0,75**	**0,58**
MLP_BD	5,1	1,2%	0,91	118,9	27,1%	0,75	0,57
MLPComb_BD	-4,1	-0,9%	0,85	115,4	26,3%	0,75	0,57
MLP	11,1	2,5%	0,86	117,4	26,7%	0,75	0,57
MLRComb	3,4	0,8%	0,80	114,0	25,9%	0,75	0,56
MLP_QM - MS2	-8,4	-1,9%	0,85	117,7	26,8%	0,74	0,56
MLPComb	-4,7	-1,1%	0,80	114,4	26,0%	0,75	0,56
MLR_BD - MS	-3,7	-0,8%	0,84	120,9	27,5%	0,73	0,54
CAMS	41,3	9,4%	0,91	133,2	30,3%	0,71	0,53
MLR_QM	-13,4	-3,1%	0,83	122,1	27,8%	0,72	0,53

Fonte: própria.

A Tabela 21 apresenta os resultados da divisão 44 para cada subconjunto dos dados da estação de Petrolina (DNI). Os resultados com maior SS4 estão destacados em negrito, enquanto o modelo de combinação MLPComb_BD é destacado em laranja para comparação por ter sido o melhor modelo de combinação. Na classe 1, os três primeiros modelos (MLRComb_BD, MLPComb_BD e CAMS) apresentam resultados similares com RMSEn baixo e SS4 próximo a 0,85. Na classe 2, o modelo MLP_QM – MS (o subescrito MS indica que esse modelo foi selecionado no conjunto de validação como melhor modelo local) apresenta o melhor resultado com maior SS4, STDRatio mais próximo a 1 e RMSEn similar aos outros modelos de maior SS4. A classe 3 (céu nublado) não apresenta bons resultados para nenhum modelo com RMSEn superior a 240% em todos os casos. Na classe 4, o MLRComb_BD possui maior SS4 e um RMSEn baixo quando comparado com os demais, porém apresenta

um STDRatio de 0,8. Por fim, na classe 5, o modelo com maior SS4 é o MLP_QM, embora apresente um RMSEn maior que o modelo MLPComb_BD.

Tabela 21 – Resultados de DNI para as melhores divisões por classe na estação de Petrolina no conjunto de teste.

Continua

Petrolina (DNI) - Divisão 44 - Subconjunto Céu Claro (classe 1)							
Modelos	MBE (W/m²)	MBEn	STDRatio	RMSE (W/m²)	RMSEn	ρ	SS4
MLRComb_BD	**35,2**	**5,1%**	**0,96**	**94,5**	**13,6%**	**0,92**	**0,85**
MLPComb_BD	**38,5**	**5,6%**	**1,01**	**98,3**	**14,2%**	**0,92**	**0,84**
CAMS	23,7	3,4%	1,00	95,4	13,8%	0,91	0,84
MLPComb	-12,2	-1,8%	0,88	89,2	12,9%	0,92	0,83
MLP_PCA_BD_Loc	24,9	3,6%	0,84	92,7	13,4%	0,92	0,82
MLRComb	-14,2	-2,0%	0,83	90,3	13,0%	0,92	0,82
MLP_SEL_BD_Loc	24,5	3,5%	0,85	93,3	13,5%	0,92	0,82
MLP_QM	10,2	1,5%	0,97	97,2	14,0%	0,90	0,82
MLR_QM	-21,0	-3,0%	0,96	98,8	14,3%	0,90	0,82
MLR_SEL_BD_Loc	23,8	3,4%	0,85	94,2	13,6%	0,91	0,82
Petrolina (DNI) - Divisão 44 - Subconjunto Céu Parcialmente Claro (classe 2)							
Modelos	MBE (W/m²)	MBEn	STDRatio	RMSE (W/m²)	RMSEn	ρ	SS4
MLP_QM - MS	**26,4**	**4,6%**	**0,90**	**189,2**	**32,7%**	**0,70**	**0,52**
MLP_BD	12,5	2,2%	0,87	185,0	32,0%	0,71	0,52
MLPComb_BD	**39,1**	**6,8%**	**0,81**	**183,5**	**31,8%**	**0,72**	**0,52**
MLRComb_BD	48,2	8,4%	0,80	185,8	32,2%	0,71	0,52
MLR_BD	14,1	2,4%	0,87	190,4	33,0%	0,69	0,50
MLR_QM	31,4	5,4%	0,87	192,8	33,4%	0,69	0,50
CAMS	107,6	18,6%	0,85	218,1	37,8%	0,69	0,49
MLP - MS2	-13,1	-2,3%	0,75	180,6	31,3%	0,71	0,49
MLPComb	3,9	0,7%	0,70	177,7	30,8%	0,72	0,48
MLRComb	12,6	2,2%	0,70	178,4	30,9%	0,71	0,48
Petrolina (DNI) - Divisão 44 - Subconjunto Céu Nublado (classe 3)							
Modelos	MBE (W/m²)	MBEn	STDRatio	RMSE (W/m²)	RMSEn	ρ	SS4
MLR	**32,3**	**84,7%**	**0,76**	**104,9**	**274,8%**	**0,35**	**0,19**
CAMS	25,8	67,6%	0,99	118,4	310,3%	0,30	0,18
MLP	11,7	30,7%	0,66	97,3	255,0%	0,35	0,17
MLR_BD - MS	10,7	28,1%	0,66	97,8	256,1%	0,34	0,17
MLR_QM	-12,4	-32,4%	0,65	98,3	257,6%	0,33	0,16
CAMSnocorr - MS2	15,7	41,2%	0,89	113,4	297,1%	0,27	0,16
MLP_QM	-15,9	-41,5%	0,62	98,2	257,4%	0,32	0,15
MLP_BD	1,0	2,6%	0,60	95,8	251,0%	0,33	0,15
MLRComb	-1,3	-3,4%	0,44	91,8	240,6%	0,35	0,11
MLPComb	-3,8	-10,0%	0,34	91,6	239,9%	0,35	0,08
MLRComb_BD	-10,1	-26,6%	0,37	93,9	245,9%	0,30	0,08

Tabela 21 – Resultados de DNI para as melhores divisões por classe na estação de Petrolina no conjunto de teste.

Conclusão

Petrolina (DNI) - Divisão 44 - Subconjunto Céu Variável (classe 4)							
Modelos	MBE (W/m²)	MBEn	STDRatio	RMSE (W/m²)	RMSEn	ρ	SS4
MLRComb_BD	**-7,6**	**-1,6%**	**0,80**	**206,1**	**44,7%**	**0,62**	**0,41**
MLP_PCA_BD_Loc	-11,1	-2,4%	0,82	209,1	45,4%	0,62	0,41
MLP_QM - MS2	21,6	4,7%	0,99	229,8	49,9%	0,60	0,41
MLP_BD	11,1	2,4%	0,93	222,9	48,4%	0,60	0,41
MLPComb_BD	**3,1**	**0,7%**	**0,79**	**206,2**	**44,7%**	**0,62**	**0,41**
MLP_SEL_BD_Loc	-8,6	-1,9%	0,81	211,4	45,9%	0,61	0,40
MLP	1,3	0,3%	0,81	212,2	46,0%	0,60	0,39
MLRComb	-20,1	-4,3%	0,69	202,6	43,9%	0,62	0,38
CAMS	129,3	28,0%	0,96	268,2	58,2%	0,56	0,37
MLPComb	-11,4	-2,5%	0,68	202,5	43,9%	0,62	0,37
Petrolina (DNI) - Divisão 44 - Subconjunto Céu Parcialmente Variável (classe 5)							
Modelos	MBE (W/m²)	MBEn	STDRatio	RMSE (W/m²)	RMSEn	ρ	SS4
MLP_QM	**-8,6**	**-4,0%**	**0,87**	**181,0**	**83,8%**	**0,60**	**0,41**
MLP_BD	-3,7	-1,7%	0,84	178,0	82,4%	0,60	0,40
MLR_QM	-1,6	-0,8%	0,90	187,5	86,8%	0,58	0,39
MLRComb_BD	-38,8	-17,9%	0,69	173,0	80,1%	0,62	0,38
MLP - MS2	19,7	9,1%	0,74	174,5	80,8%	0,60	0,38
MLPComb_BD	**-33,0**	**-15,3%**	**0,67**	**170,3**	**78,8%**	**0,63**	**0,38**
MLR_BD - MS	10,8	5,0%	0,79	180,0	83,3%	0,58	0,37
CAMS	121,4	56,2%	0,91	231,7	107,3%	0,54	0,35
MLRComb	-16,2	-7,5%	0,61	168,7	78,1%	0,62	0,34
CAMSnocorr	87,4	40,5%	1,00	225,5	104,4%	0,53	0,34

Fonte: própria.

A Tabela 22 apresenta os resultados da divisão 43 para cada subconjunto dos dados da estação de Gobabeb (GHI). Os resultados com maior SS4 estão destacados em negrito. Na classe 1, todos os modelos apresentam resultados muito próximos da observação (SS4 próximo a 1 e RMSEn menor que 5%). Na classe 2, os resultados seguem a mesma tendência que na classe 1. Em relação à classe 3, o modelo MLRComb_BD apresenta o maior SS4 e um RMSEn 4,1% menor que o da CAMS; contudo possui um STDRatio mais distante de 1 e um aumento no MBEn (em valor absoluto) quando comparado com os resultados da DNI fornecida pela CAMS. Na classe 4, o maior SS4 é o apresentado pelo modelo MLPComb_BD que possui resultados similares ao modelo MLRComb_BD. Na classe 5, todos os modelos apresentam resultados estatísticos levemente melhores que os da CAMS, sendo que o modelo MLP – MS – MS2 apresenta o menor RMSEn e maior SS4. Os subescritos

MS e MS2 indicam que esse modelo foi selecionado no conjunto de validação como melhor modelo do subconjunto céu parcialmente variável (Tabela 11 da seção 4.5).

Tabela 22 – Resultados de GHI para as melhores divisões por classe na estação de Gobabeb no conjunto de teste.

Continua

Gobabeb (GHI) - Divisão 43 - Subconjunto Céu Claro (classe 1)							
Modelos	MBE (W/m²)	MBEn	STDRatio	RMSE (W/m²)	RMSEn	ρ	SS4
CAMS	-18,0	-2,8%	1,01	26,4	4,1%	1,00	1,00
MLP_PCA_BD_Loc	**0,5**	**0,1%**	**1,00**	**13,6**	**2,1%**	**1,00**	**1,00**
MLP_PCA_Loc - MS - MS2	0,6	0,1%	1,00	13,6	2,1%	1,00	1,00
MLRComb	0,5	0,1%	1,00	13,7	2,1%	1,00	1,00
MLRComb_BD	**0,0**	**0,0%**	**1,00**	**13,7**	**2,1%**	**1,00**	**1,00**
MLPComb	0,8	0,1%	1,00	13,8	2,2%	1,00	1,00
MLPComb_BD	0,0	0,0%	1,00	13,9	2,2%	1,00	1,00
MLP_SEL_BD_Loc	0,6	0,1%	1,00	14,2	2,2%	1,00	1,00
MLP_SEL_Loc	0,2	0,0%	1,00	14,2	2,2%	1,00	1,00
MLR_SEL_Loc	0,5	0,1%	1,00	14,7	2,3%	1,00	1,00
MLR_SEL_BD_Loc	0,5	0,1%	1,00	14,7	2,3%	1,00	1,00
Gobabeb (GHI) - Divisão 43 - Subconjunto Céu Parcialmente Claro (classe 2)							
Modelos	MBE (W/m²)	MBEn	STDRatio	RMSE (W/m²)	RMSEn	ρ	SS4
CAMS	-15,8	-2,5%	1,02	46,9	7,4%	0,99	0,98
MLPComb_BD	**3,7**	**0,6%**	**1,00**	**31,5**	**4,9%**	**0,99**	**0,99**
MLPComb	4,5	0,7%	1,00	31,6	5,0%	0,99	0,99
MLRComb_BD	**4,7**	**0,7%**	**1,00**	**31,8**	**5,0%**	**0,99**	**0,99**
MLRComb	5,2	0,8%	0,99	31,9	5,0%	0,99	0,99
MLP_PCA_BD_Loc	8,7	1,4%	0,99	33,9	5,3%	0,99	0,99
MLP_SEL_BD_Loc	8,3	1,3%	0,99	33,8	5,3%	0,99	0,99
MLP_SEL_Loc	8,0	1,3%	0,98	33,8	5,3%	0,99	0,99
MLP_PCA_Loc	6,4	1,0%	0,98	33,5	5,3%	0,99	0,99
MLP	2,9	0,5%	1,00	33,7	5,3%	0,99	0,99
MLP_BD	3,3	0,5%	1,01	34,0	5,3%	0,99	0,99
Gobabeb (GHI) - Divisão 43 - Subconjunto Céu Nublado (classe 3)							
Modelos	MBE (W/m²)	MBEn	STDRatio	RMSE (W/m²)	RMSEn	ρ	SS4
MLRComb_BD	**-10,1**	**-5,5%**	**0,86**	**68,6**	**37,4%**	**0,87**	**0,74**
MLRComb	-5,8	-3,2%	0,86	68,2	37,1%	0,87	0,74
MLPComb_BD	-5,3	-2,9%	0,85	68,3	37,2%	0,87	0,74
MLPComb	-2,7	-1,5%	0,85	68,1	37,1%	0,87	0,74
MLP_QM	5,6	3,0%	0,94	72,6	39,5%	0,85	0,73
MLP_BD	3,9	2,1%	0,91	71,9	39,2%	0,85	0,73
MLP - MS2	7,7	4,2%	0,90	72,1	39,3%	0,85	0,73
MLP_PCA_BD_Loc	-26,3	-14,3%	0,79	75,8	41,3%	0,86	0,70
CAMS	-7,1	-3,9%	0,88	76,3	41,5%	0,83	0,69
MLP_SEL_BD_Loc	-25,8	-14,0%	0,78	77,1	42,0%	0,85	0,69

Tabela 22 – Resultados de GHI para as melhores divisões por classe na estação de Gobabeb no conjunto de teste.

Conclusão

Gobabeb (GHI) - Divisão 43 - Subconjunto Céu Variável (classe 4)							
Modelos	MBE (W/m²)	MBEn	STDRatio	RMSE (W/m²)	RMSEn	ρ	SS4
CAMS	-47,4	-6,4%	0,89	131,0	17,8%	0,83	0,69
MLPComb_BD	**2,1**	**0,3%**	**0,92**	**106,5**	**14,5%**	**0,87**	**0,77**
MLPComb	2,5	0,3%	0,92	106,4	14,4%	0,87	0,77
MLRComb_BD	**4,9**	**0,7%**	**0,91**	**106,4**	**14,4%**	**0,87**	**0,76**
MLRComb	4,6	0,6%	0,90	106,2	14,4%	0,87	0,76
MLP_PCA_BD_Loc	-5,0	-0,7%	0,96	110,0	14,9%	0,87	0,76
MLP_SEL_BD_Loc	-3,5	-0,5%	0,96	111,3	15,1%	0,87	0,76
MLP_BD	-1,9	-0,3%	0,94	110,4	15,0%	0,87	0,75
MLP - MS2	-3,1	-0,4%	0,93	110,2	15,0%	0,87	0,75
MLP_QM - MS	2,8	0,4%	0,93	110,1	14,9%	0,87	0,75
MLP_PCA_Loc	-4,4	-0,6%	0,87	108,2	14,7%	0,87	0,75
Gobabeb (GHI) - Divisão 43 - Subconjunto Céu Parcialmente Variável (classe 5)							
Modelos	MBE (W/m²)	MBEn	STDRatio	RMSE (W/m²)	RMSEn	ρ	SS4
CAMS	-20,9	-3,9%	1,02	84,4	15,8%	0,95	0,91
MLP - MS - MS2	**2,8**	**0,5%**	**1,00**	**70,2**	**13,1%**	**0,96**	**0,93**
MLP_BD	2,2	0,4%	1,00	70,5	13,2%	0,96	0,93
MLP_QM	7,8	1,5%	1,00	71,3	13,3%	0,96	0,93
MLRComb_BD	**12,6**	**2,4%**	**0,99**	**71,8**	**13,4%**	**0,96**	**0,93**
MLRComb	13,9	2,6%	0,98	71,8	13,4%	0,96	0,93
MLP_PCA_BD_Loc	14,9	2,8%	0,99	72,6	13,6%	0,96	0,93
MLP_PCA_Loc	13,4	2,5%	0,98	72,1	13,5%	0,96	0,92
MLPComb	11,0	2,0%	1,00	72,6	13,6%	0,96	0,92
MLPComb_BD	9,8	1,8%	1,00	72,6	13,6%	0,96	0,92
MLR_SEL_BD_Loc	18,0	3,4%	1,00	75,1	14,0%	0,96	0,92

Fonte: própria.

Por fim, a Tabela 23 apresenta os resultados da divisão 43 para cada subconjunto dos dados da estação de Gobabeb (DNI). Os resultados com maior SS4 estão destacados em negrito. Na classe 1, todos os modelos da Tabela 23 apresentam resultados melhores que o da CAMS, com destaque para o maior SS4 do modelo de combinação MLRComb_BD. Na classe 2, de forma similar à classe céu claro, todos os modelos listados apresentam resultados melhores que o da CAMS, com destaque para o modelo MLPComb_BD. Na classe 3, todos os modelos listados não ajustam bem a DNI em condições de céu nublado e apresentam RMSEn superior a 250%. O modelo MLRComb_BD apresenta o menor SS4 da lista, com um STDRatio de 0,41 e um MBEn de -31,8%, apesar de ter o segundo menor RMSEn da lista (258,3%). Para as classes 4 e 5, os modelos com maiores SS4 são os modelos globais MLP_QM e

MLP_BD, respectivamente, que possuem também STDRatio mais próximo a 1 que os modelos de combinação, o que justifica o maior SS4 destes. Apesar dessa diferença no STDRatio, o modelo de combinação MLRComb_BD apresenta menores RMSEn em ambas as classes 4 e 5.

Tabela 23 – Resultados de DNI para as melhores divisões por classe na estação de Gobabeb no conjunto de teste.

Continua

Gobabeb (DNI) - Divisão 43 - Subconjunto Céu Claro (classe 1)							
Modelos	MBE (W/m²)	MBEn	STDRatio	RMSE (W/m²)	RMSEn	ρ	SS4
CAMS	-69,8	-8,3%	1,18	108,8	13,0%	0,91	0,81
MLRComb_BD	**2,1**	**0,3%**	**0,99**	**55,8**	**6,7%**	**0,95**	**0,90**
MLRComb	-5,5	-0,7%	0,95	55,5	6,6%	0,95	0,89
MLPComb_BD	3,6	0,4%	0,99	56,5	6,7%	0,94	0,89
MLPComb	-4,2	-0,5%	0,95	56,0	6,7%	0,94	0,89
MLP_PCA_BD_Loc	5,0	0,6%	0,94	57,1	6,8%	0,94	0,89
MLP_PCA_Loc	5,3	0,6%	0,92	57,2	6,8%	0,94	0,88
MLP_BD	-3,8	-0,5%	1,01	61,0	7,3%	0,94	0,88
MLP	-8,9	-1,1%	0,96	60,3	7,2%	0,94	0,88
MLP_QM	0,1	0,0%	0,99	60,6	7,2%	0,94	0,88
MLP_SEL_BD_Loc	5,4	0,6%	0,94	61,2	7,3%	0,93	0,87
Gobabeb (DNI) - Divisão 43 - Subconjunto Céu Parcialmente Claro (classe 2)							
Modelos	MBE (W/m²)	MBEn	STDRatio	RMSE (W/m²)	RMSEn	ρ	SS4
CAMS	-51,4	-6,9%	1,13	155,8	21,0%	0,79	0,63
MLPComb_BD	**15,6**	**2,1%**	**0,91**	**101,1**	**13,6%**	**0,88**	**0,77**
MLPComb	11,2	1,5%	0,88	100,3	13,5%	0,88	0,77
MLRComb_BD	**24,7**	**3,3%**	**0,90**	**104,6**	**14,1%**	**0,87**	**0,76**
MLRComb	20,5	2,8%	0,86	103,5	13,9%	0,87	0,75
MLP_BD	15,4	2,1%	0,97	110,1	14,8%	0,86	0,75
MLP_QM - MS2	21,6	2,9%	0,96	111,6	15,0%	0,86	0,74
MLP	14,2	1,9%	0,92	108,4	14,6%	0,86	0,74
MLP_PCA_BD_Loc	31,3	4,2%	0,86	113,8	15,3%	0,85	0,72
MLP_SEL_BD_Loc	31,2	4,2%	0,86	116,2	15,7%	0,85	0,71
MLP_PCA_Loc	29,8	4,0%	0,81	113,7	15,3%	0,85	0,70

Tabela 23 – Resultados de DNI para as melhores divisões por classe na estação de Gobabeb no conjunto de teste.

Conclusão

Gobabeb (DNI) - Divisão 43 - Subconjunto Céu Nublado (classe 3)							
Modelos	MBE (W/m²)	MBEn	STDRatio	RMSE (W/m²)	RMSEn	ρ	SS4
MLP	**30,0**	**63,5%**	**0,81**	**131,4**	**277,8%**	**0,38**	**0,22**
MLP_BD	14,3	30,2%	0,76	127,4	269,4%	0,37	0,20
MLP_QM	-6,7	-14,1%	0,79	130,0	274,8%	0,35	0,20
CAMS	16,1	34,1%	0,91	140,6	297,4%	0,33	0,19
MLR	95,4	201,8%	1,08	182,4	385,6%	0,29	0,17
MLR_BD - MS	58,8	124,4%	1,02	162,7	344,1%	0,29	0,17
CAMSnocorr	28,0	59,3%	1,02	156,4	330,6%	0,27	0,16
MLR_QM	10,7	22,7%	1,05	156,6	331,1%	0,26	0,16
MLRComb	-2,5	-5,4%	0,46	120,3	254,3%	0,32	0,11
MLPComb	4,9	10,4%	0,58	131,8	278,8%	0,21	0,10
MLRComb_BD	**-15,0**	**-31,8%**	**0,41**	**122,2**	**258,3%**	**0,29**	**0,09**
Gobabeb (DNI) - Divisão 43 - Subconjunto Céu Variável (classe 4)							
Modelos	MBE (W/m²)	MBEn	STDRatio	RMSE (W/m²)	RMSEn	ρ	SS4
CAMS	17,5	3,4%	0,86	216,7	41,9%	0,70	0,51
MLP_QM	**8,7**	**1,7%**	**1,04**	**203,5**	**39,3%**	**0,78**	**0,62**
MLPComb_BD	-16,0	-3,1%	0,89	189,4	36,6%	0,78	0,62
MLP_BD	9,7	1,9%	1,00	199,7	38,6%	0,77	0,62
MLP	19,7	3,8%	0,95	196,0	37,9%	0,77	0,62
MLRComb_BD	**-8,0**	**-1,6%**	**0,86**	**186,1**	**36,0%**	**0,78**	**0,62**
MLPComb	-9,8	-1,9%	0,86	187,4	36,2%	0,78	0,61
MLP_PCA_BD_Loc	-24,0	-4,6%	0,90	193,6	37,4%	0,77	0,61
MLRComb	-0,6	-0,1%	0,82	184,9	35,7%	0,78	0,61
MLP_SEL_BD_Loc	-28,2	-5,4%	0,88	200,4	38,7%	0,75	0,58
MLP_PCA_Loc	-27,6	-5,3%	0,75	190,2	36,8%	0,77	0,57
Gobabeb (DNI) - Divisão 43 - Subconjunto Céu Parcialmente Variável (classe 5)							
Modelos	MBE (W/m²)	MBEn	STDRatio	RMSE (W/m²)	RMSEn	ρ	SS4
CAMS	-17,6	-3,2%	0,96	229,4	42,3%	0,72	0,54
MLP_BD	**3,3**	**0,6%**	**0,99**	**202,5**	**37,3%**	**0,78**	**0,63**
MLP_QM - MS - MS2	1,1	0,2%	1,02	206,9	38,1%	0,78	0,63
MLPComb_BD	-4,6	-0,9%	0,93	197,3	36,4%	0,78	0,63
MLP	12,8	2,4%	0,94	198,7	36,6%	0,78	0,63
MLRComb_BD	**9,1**	**1,7%**	**0,92**	**196,4**	**36,2%**	**0,78**	**0,63**
MLPComb	0,0	0,0%	0,89	194,9	35,9%	0,78	0,62
MLRComb	14,6	2,7%	0,88	194,5	35,8%	0,78	0,62
MLP_PCA_BD_Loc	16,4	3,0%	0,94	206,9	38,1%	0,76	0,60
MLP_SEL_BD_Loc	9,0	1,7%	0,93	206,2	38,0%	0,76	0,60
MLP_SEL_Loc	5,1	0,9%	0,86	201,7	37,2%	0,76	0,59

Fonte: própria.

6 CONCLUSÕES E PERSPECTIVAS FUTURAS

As diferentes condições do céu definidas na etapa de classificação não supervisionada foram validadas de forma empírica, mostrando que a classificação adotada das janelas de 15 minutos apresenta resultados satisfatórios, já que segue as regiões pré-definidas da dispersão $VI\ x\ k_c\ x\ KDE$. Os resultados mostram que as estações de El Rosal e Gobabeb apresentam um percentual considerável de céu claro (66,5% e 80,7%, respectivamente, somando os percentuais das classes céu claro e céu parcialmente claro), o que pode estar relacionado à altitude da estação de El Rosal e ao fato da estação Gobabeb estar localizada em uma região desértica. Já os resultados da classificação supervisionada mostram que as maiores acurácias no conjunto de teste foram obtidas para as divisões de 43 a 62, que fazem a randomização dos *timestamps* de 15 min.

O *site adaptation* das estações localizadas na Argentina (El Rosal e Salta) apresentam melhores resultados estatísticos do que o modelo da CAMS. Como essas estações estão localizadas em uma região na borda do campo de visão do satélite METEOSAT, o modelo da CAMS pode não estimar bem a atenuação da radiação pelas nuvens nessa região. Em particular, a estação de El Rosal apresenta excelentes resultados estatísticos (26,2% de RMSEn normalizado a menos que a CAMS na comparação com o modelo MLRComb_BD), o que pode estar relacionado com a altitude da estação solarimétrica de 3335 m. A estação de Salta se encontra a 1233 m de altitude e os resultados estatísticos do *site adaptation* são melhores que os da CAMS, contudo, não são tão expressivos como os da estação El Rosal (4,8% de RMSEn normalizado a menos que a CAMS na comparação com o modelo MLRComb_BD).

Por outro lado, os resultados para irradiância global horizontal nas estações de Petrolina e Gobabeb mostram estatísticos similares para os modelos globais, locais, de combinação e da CAMS. Como ambas as estações são melhor visualizadas pelo satélite METEOSAT, sobretudo para a estação de Gobabeb, o modelo da CAMS consegue estimar melhor a radiação nessas estações. Além disso, como a estação de Petrolina fica próxima a uma estação que faz parte da *Baseline Surface Radiation Network* (BSRN) e a estação de Gobabeb é uma estação da BSRN, o modelo da CAMS pode ter feito ajustes estatísticos nas regiões de Petrolina e Gobabeb utilizando os dados públicos dessas estações. Finalmente, para irradiância direta normal e

irradiândia difusa horizontal (obtida por diferença) das estações de Petrolina e Gobabeb, os resultados dos modelos globais e locais são similares e melhores que os da CAMS, enquanto os resultados dos modelos de combinação são ainda melhores do que os modelos globais e locais.

De maneira geral, os resultados do *site adaptation* são heterogêneos, isto é, dependem fortemente da região analisada e da componente da radiação. Do ponto de vista dos modelos globais, a rede neural do tipo *multilayer perceptron* (MLP) apresenta melhores resultados que a regressão linear múltipla (MLR). Como a MLP utiliza uma técnica não linear para resolver o problema de forma global, ela consegue atingir resultados que a regressão linear não alcança. Em relação aos modelos locais, por se tratarem de modelos aplicados em subconjuntos específicos da série temporal divididos de acordo com as condições do céu, eles conseguem melhores resultados estatísticos, no geral, que os modelos globais. Por fim, dado que os modelos globais e locais não são redundantes, a combinação dos mesmos cujos parâmetros foram treinados no conjunto de validação, obtém resultados ainda melhores do que os modelos individualmente.

A avaliação feita das diferentes divisões dos dados mostra que, em sua maioria, as divisões dos dados que consideram randomização apresentam resultados com maior acurácia do que as divisões que mantém a sequência cronológica das séries temporais ou as divisões com dias intercalados. Além disso, para todos os modelos analisados, os melhores resultados estão nos modelos de combinação que podem utilizar como variáveis de entrada as saídas dos modelos globais, locais e da CAMS, a depender do resultado da seleção de variáveis.

A análise dos resultados nos subconjuntos locais das melhores divisões dos dados mostra que em alguns subconjuntos o modelo de combinação pode não estimar de forma acurada a radiação naquela condição de céu específica, apesar de apresentar melhores resultados ao avaliar os estatísticos das séries temporais de forma global. Para buscar melhorar as estimativas também nos subconjuntos das séries temporais, poderiam ser analisados em trabalhos futuros os resultados de modelos de combinação treinados para cada subconjunto local no conjunto de validação, de forma a obter resultados combinados para cada subconjunto da série. Além disso, pode-se testar a metodologia utilizando um número diferentes de classes (como, por exemplo, apenas 3 ou 4 classes) de forma a avaliar se os resultados finais são sensíveis à quantidade de classes escolhidas.

REFERÊNCIAS

AGGARWAL, C. C. Data Mining: The Textbook. 1ª. ed., Elsevier, 2015.

AGUIAR, L. M.; POLO, J.; VINDEL, J. M.; OLIVER, A. Analysis of Satellite Derived Solar Irradiance in Islands with Site Adaptation Techniques for Improving the Uncertainty. Renewable Energy, 135: 98–107, 2019

ALIMOHAMMADI, S.; HE, D. Multi-Stage Algorithm for Uncertainty Analysis of Solar Power Forecasting. IEEE Power and Energy Society General Meeting, 2016.

ASSUNÇÃO, H. F.; ESCOBEDO, J. F.; OLIVEIRA, A.P. A New Algorithm to Estimate Sky Condition Based on 5 Minutes-Averaged Values of Clearness Index and Relative Optical Air Mass. Theoretical and Applied Climatology, 90(3–4): 235–48, 2007.

BENDER, G.; DAVIDSON, F.; EICHELBERGER, S.; GUEYMARD, C. The road to bankability: Improving assessments for more accurate financial planning. Solar Conference, Raleigh, Estados Unidos, 2011.

BREIMAN, L. Random Forests. *Machine Learning* 45, 5–32, 2001.

CALBÓ, J.; GONZÁLEZ, J. A.; PEGÈS, D. 2001. A Method for Sky-Condition Classification from Ground-Based Solar Radiation Measurements. Journal of Applied Meteorology, 40(12): 2193–99, 2001.

CANNON, A. J.; SOBIE, S. R.; MURDOCK, T. Q. Bias Correction of GCM Precipitation by Quantile Mapping: How Well Do Methods Preserve Changes in Quantiles and Extremes? Journal of Climate, 28(17): 6938–59, 2015.

CEBECAUER, T.; SURI, M. Site-Adaptation of Satellite-Based DNI and GHI Time Series: Overview and SolarGIS Approach. AIP Conference Proceedings, 1734, 2016.

CHICCO, G.; COCINA, V.; SPERTINO, F. Characterization of Solar Irradiance Profiles for Photovoltaic System Studies through Data Rescaling in Time and Amplitude. Proceedings of the Universities Power Engineering Conference, 2014

DAWSON, C. W.; WILBY, R. L. Hydrological modelling using artificial neural networks. Progress in physical Geography, v. 25, n. 1, p. 80-108, 2001.

DUCHON, C. E.; O'MALLEY, M. S. Estimating Cloud Type from Pyranometer Observations. Journal of Applied Meteorology, 38(1): 132–41, 1999.

EMILION, R. Classification and mixture process. Academic Science Report. Paris, 335, series I, 189-193, 2002.

EPE, EMPRESA DE PESQUISA ENERGÉTICA. Instruções para Solicitação de Cadastramento e Habilitação Técnica de Empreendimentos Fotovoltaicos com Vistas à Participação nos Leilões de Energia Elétrica. 2017. Disponível em

<http://www.epe.gov.br/sites-pt/leiloes-de-energia/Documents/EPE-DEE-065_2013_R5_2017_UFV.pdf> Acesso em 03 de dezembro de 2019.

FERNÁNDEZ-PERUCHENA, CARLOS M., JESÚS POLO, LUIS MARTÍN, AND LUIS MAZORRA. Site-Adaptation of Modeled Solar Radiation Data: The SiteAdapt Procedure. *Remote Sensing* 12(13): 1–17, 2020.

FORTUNA, LUIGI, GIUSEPPE NUNNARI, AND SILVIA NUNNARI. A New Fine-Grained Classification Strategy for Solar Daily Radiation Patterns. *Pattern Recognition Letters* 81: 110–17, 2016.

FRANCESCA, GWEN BENDER, DAVIDSON SCOTT, AND CHRISTIAN A. GUEYMARD. The Road to Bankability: Improving Assessments for More Accurate Financial Planning. *40th ASES National Solar Conference 2011, SOLAR 2011* 1(June): 733–38, 2011.

GASTÓN-ROMEO, MARTÍN, TERESA LEON, FERMÍN MALLOR, AND LOURDES RAMÍREZ-SANTIGOSA. 2011. A Morphological Clustering Method for Daily Solar Radiation Curves. *Solar Energy* 85(9): 1824–36, 2011.

GAVIN, H. P. The Levenberg-Marquardt algorithm for nonlinear least squares curve-fitting problems. Department of Civil and Environmental Engineering Duke University, 2020.

GUEYMARD, C. A.; WILCOX, S. M. Spatial and Temporal Variability in the Solar Resource: Assessing the Value of Short-Term Measurements at Potential Solar Power Plant Sites. *38th ASES National Solar Conference 2009, SOLAR 2009* 5(January 2009): 3026–53, 2009.

GUEYMARD, C. A.; GUSTAFSON, W. T.; ETRINGER, A.; STORCK, P. Evaluation of procedures to improve solar resource assessments: optimum use of short-term data from a local weather station to correct bias in long-term satellite derived solar radiation time series. World Renewable Energy Forum, Denver, CO, v. 3, p. 2092–99, 2012.

HABTE, A.; STOFFEL, T.; PEREZ, R.; MYERS, D.; GUEYMARD, C.; BLANC, P.; WILBERT, S. Overview of Solar Radiation Resource Concepts. Best Practices Handbook for the Collection and Use of Solar Resource Data for Solar Energy Applications: Second Edition, NREL, 2.1-2.22, 2017.

HALL, M. A. Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. Conference: Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, 2000.

HAN, J.; KAMBER, M.; PEI, J. Data Mining: Concepts and Techniques. 3ª. ed., Elsevier, 2012.

HUANG, G.; LI, Z.; LI, X.; LIANG, S.; YANG, K.; WANG, D.; ZHANG, Y. Estimating Surface Solar Irradiance from Satellites: Past, Present, and Future Perspectives. Remote Sensing of Environment, 233, 111371, 2019.

HARROUNI, S., A. GUESSOUM, AND A. MAAFI. 2005. Classification of Daily Solar Irradiation by Fractional Analysis of 10-Min-Means of Solar Irradiance. *Theoretical and Applied Climatology* 80(1): 27–36, 2005.

INEICHEN, P.; BARROSO, C.S.; GEIGER, B.; HOLLMANN, R.; MARSOUIN, A.; MUELLER, R. Satellite application facilities irradiance products: hourly time step comparison and validation over Europe. Int. J. Rem. Sens. 30 (21), p. 332-344, 2009.

INES, A.V.M.; HANSEN, J. W. Bias Correction of Daily GCM Rainfall for Crop Simulation Studies. Agricultural and Forest Meteorology, 138(1–4): 44–53, 2006.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data Clustering: A Review. 31(3), 2000.

JIMÉNEZ-PÉREZ, PEDRO F., AND LLANOS MORA-LÓPEZ. Modeling and Forecasting Hourly Global Solar Radiation Using Clustering and Classification Techniques. *Solar Energy* 135: 682–91, 2016.

JOLLIFFE, I. T. Principal Component Analysis. 2ª. ed.: Elsevier, 1986.

KANG, BYUNG O., AND KWA SUR TAM.A New Characterization and Classification Method for Daily Sky Conditions Based on Ground-Based Solar Irradiance Measurement Data. *Solar Energy* 94: 102–18, 2013.

KARIUKI, B. W.; SATO, T. Interannual and Spatial Variability of Solar Radiation Energy Potential in Kenya Using Meteosat Satellite. Renewable Energy 116:88–96, 2018.

LEFÈVRE, M.; OUMBE, A.; BLANC, P.; ESPINAR, B.; GSCHWIND, B.; QU, Z.; WALD, L.; SCHROEDTER-HOMSCHEIDT, M.; HOYER-KLICK, C.; AROLA, A.; BENEDETTI, A.; KAISER, J. W.; MORCRETTE, E J. J. McClear: A New Model Estimating Downwelling Solar Radiation at Ground Level in Clear-Sky Conditions. Atmospheric Measurement Techniques, 6(9):2403–18, 2013.

LI, DANNY H.W., AND JOSEPH C. LAM. 2001. An Analysis of Climatic Parameters and Sky Condition Classification. *Building and Environment* 36(4): 435–45, 2001.

MACQUEEN, J. SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS. 233(233): 281–97, 1967.

MAGENUKA, T. K.; MUSASA, K.; AKINDEJI, K. T. Kernel Density Estimation of Solar Radiation and Wind Speed for South Africa. Conference: 5th North American International Conference on Industrial Engineering and Operations Management, Detroit, Michigan, USA, 2020.

MICHELANGELI, P. A., M. VRAC, AND H. LOUKOS. Probabilistic Downscaling Approaches: Application to Wind Cumulative Distribution Functions. Geophysical Research Letters 36(11):2–7, 2009.

MIESLINGER, THERESA, FELIX AMENT, KAUSHAL CHHATBAR, AND RICHARD MEYER. A New Method for Fusion of Measured and Model-Derived Solar Radiation Time-Series. *Energy Procedia* 48: 1617–26, 2014.

MIRANDA, D. R.; ARAÚJO, E. V.; VILELA, O. C. Análise do recurso solar de longo prazo na cidade de São João do Rio do Peixe, sertão da Paraíba. VIII Congresso Brasileiro de Energia Solar – CBENS, 2020, Fortaleza-CE. Anais do VIII Congresso Brasileiro de Energia Solar, 2020.

MIRANDA, D. R.; COSTA, R. S. S.; VILELA, O. C.; SALAZAR, G. A.; PEREIRA, A. C.; JATOBA, E. B.; NETO, A. C.; FILHO, J. B. M. *Site Adaptation* da radiação global horizontal para estações solarimétricas na Argentina e no Brasil. IX Congresso Brasileiro de Energia Solar – CBENS, Florianópolis-SC. Anais do IX Congresso Brasileiro de Energia Solar, 2022.

MORENO-TEJERA, S.; SILVA, M. A; SANTIGOSA, L.; LILLO, I. Classification of days according to DNI profiles using clustering techniques. Solar Energy, v. 146, p. 319-333, 2017.

MUSELLI, M., POGGI, P., NOTTON, G., & LOUCHE, A. Classification of typical meteorological days from global irradiation records and comparison between two Mediterranean coastal sites in Corsica Island. *Energy Conversion and Management*, *41*(10), 1043–1063, 2000.

NARVAEZ, GABRIEL, LUIS FELIPE GIRALDO, MICHAEL BRESSAN, AND ANDRES PANTOJA. Machine Learning for Site-Adaptation and Solar Radiation Forecasting. *Renewable Energy* 167(December): 333–42, 2021

PAGÈS, D., JOSEP CALBÓ, E J. A. GONZÁLEZ. Using Routine Meteorological Data to Derive Sky Conditions. *Annales Geophysicae* 21(3): 649–54. 2003

PÉREZ-ORTIZ, M.; JIMÉNEZ-FERNÁNDEZ, S.; GUTIÉRREZ, P. A.; ALEXANDRE, H.; HERVÁS-MARTÍNEZ, C.; SALCEDO-SANZ, S. A Review of Classification Problems and Algorithms in Renewable Energy Applications. *Energies* 9(8): 1–27, 2016.

PERRUCI, V. P. Análise de complementaridade entre diferentes técnicas estatísticas para aumento na resolução espacial do comportamento do vento local. Dissertação de Mestrado, Universidade Federal de Pernambuco, 2018.

PETRIBÚ, L. B.; SABINO, E.; BARROS, H.; COSTA, A.; BARBOSA, E.; VILELA, O.C. Procedimento objetivo para a garantia de qualidade de dados de radiação solar. Presentado em XL Reunión de Trabajo de la Asociación Argentina de Energías Renovables y Ambiente (ASADES), San Juan, Argentina, 2017.

POLO, J., MARTÍN, L.; VINDEL, J. M. Correcting Satellite Derived DNI with Systematic and Seasonal Deviations: Application to India. *Renewable Energy* 80: 238–43, 2015.

POLO J., WILBERT S., RUIZ-ARIAS J. A., MEYER R., GUEYMARD C., SÚRI M., ET AL. Preliminary survey on site-adaptation techniques for satellite-derived and reanalysis solar radiation datasets. Solar Energy, 132:25–37, 2016.

POLO, J.; FERNÁNDEZ-PERUCHENA, C.; SALAMALIKIS,V.; AGUIAR, L. M.; TURPIN, M.; MARTÍN-POMARES, L.; KAZANTZIDIS, A.; BLANC, P.; REMUND, J.

Benchmarking on Improvement and Site-Adaptation Techniques for Modeled Solar Radiation Datasets. Solar Energy 201(March):469–79, 2020.

QU, ZHIPENG, ARMEL OUMBE, PHILIPPE BLANC, BELLA ESPINAR, GERHARD GESELL, BENOIT GSCHWIND, LARS KLÜSER, MIREILLE LEFÈVRE, LAURENT SABORET, MARION SCHROEDTER-HOMSCHEIDT, LUCIEN WALD. Fast Radiative Transfer Parameterisation for Assessing the Surface Solar Irradiance: The Heliosat-4 Method. Meteorologische Zeitschrift 26(1):33–57, 2017.

RENO, M. E HANSEN, C. Identification of Periods of Clear Sky Irradiance in Time Series of GHI Measurements. Renewable Energy, v. 90, p. 520-531, 2016.

ROESCH, A.; WILD, M.; OHMURA, A.; DUTTON, E.G.; LONG, C.N.; ZHANG T. Assessment of BSRN radiation records for the computation of monthly means. Atmos. Meas. Tech., v. 4, p. 339–54, 2011.

RUSCHEL, C. G E PONTE, G. P. Metodologias de ajuste de dados solarimétricos visando a estimativa de produção de energia de longo prazo. VII Congresso Brasileiro de Energia Solar, Gramado, 2018.

SALAMALIKIS, V.; TZOUMANIKAS, P.; ARGIRIOU, A. A.; KAZANTZIDIS, A. Site adaptation of global horizontal irradiance from the Copernicus Atmospheric Monitoring Service for radiation using supervised machine learning techniques. Renewable Energy 195: 92-106, 2022.

SALAZAR, G; GUEYMARD, C.; GALDINO, J.; VILELA, O. C.; FRAIDENRAICH, N. Solar irradiance time series derived from high-quality measurements, satellite-based models, and reanalyses at a near-equatorial site in Brazil. Renewable and Sustainable Energy Reviews, 117, 109478, 2020.

SCHROEDTER-HOMSCHEIDT, MARION, CARSTEN HOYER-KLICK, NIELS KILLIUS, JETHRO BETCKE, MIREILLE LEFÈVRE, LUCIEN WALD, ETIENNE WEY, LAURENT SABORET. User's Guide to the CAMS Radiation Service (CRS) Status December 2019. (December):1–74, 2019.

SCHUMANN, K.; BEYER, H. G.; CHHATBAR, K.; MEYER, R. Improving Satellite-Derived Solar Resource Analysis with Parallel Ground-Based Measurements. Proceedings of Proceeding from ISES Solar World Congress, Kasel, Germany, 2011.

SCHWANDT, MARKO, KAUSHAL CHHATBAR, RICHARD MEYER, KATHARINA FROSS, INDRADIP MITRA, RAMADHAN VASHISTHA, GODUGUNUR GIRIDHAR, S. GOMATHINAYAGAM, AND ASHVINI KUMAR. Development and Test of Gap Filling Procedures for Solar Radiation Data of the Indian SRRA Measurement Network. Energy Procedia 57:1100–1109, 2014.

SOUBDHAN, T.; EMILION, R.; CALIF, R. Classification of Daily Solar Radiation Distributions Using a Mixture of Dirichlet Distributions. *Solar Energy* 83(7): 1056–63, 2009.

STEIN, J. S.; HANSEN, C. W.; RENO, M. J. The Variability Index: A New and Novel Metric for Quantifying Irradiance and Pv Output Variability. *World Renewable Energy Forum, WREF 2012, Including World Renewable Energy Congress XII and Colorado Renewable Energy Society (CRES) Annual Conferen* 4(May): 2764–70, 2012.

SUN, XIXI, JAMIE M. BRIGHT, CHRISTIAN A. GUEYMARD, XINYU BAI, BRENDAN ACORD, AND PENG WANG. Worldwide Performance Assessment of 95 Direct and Diffuse Clear-Sky Irradiance Models Using Principal Component Analysis. Renewable and Sustainable Energy Reviews 135(March 2020):110087, 2021.

SÚRI, M.; CEBECAUER, T., PEREZ., R. Quality Procedures of SolarGIS for Provision Site-Specific Solar Resource Information. *Conference SolarPACES*: 1–5, 2010.

TAHIR, Z. R; AZHAR, M.; BLAN, P.; ASIM, M.; IMRAN, S.; HAYAT, N.; SHAHID, H.; ALI, H. The Evaluation of Reanalysis and Analysis Products of Solar Radiation for Sindh Province, Pakistan. *Renewable Energy* 145: 347–62, 2020.

TAYLOR, K.E. Summarizing multiple aspects of model performance in a single diagram. Journal of Geophysical Research, v. 106, p. 7183-7192, 2011.

THEMEßL, MATTHIAS JAKOB, ANDREAS GOBIET, AND GEORG HEINRICH. Empirical-Statistical Downscaling and Error Correction of Regional Climate Models and Its Impact on the Climate Change Signal. Climatic Change 112(2):449–68, 2012.

TIBA, C.; SÁ, M. H; LACERDA; L. F. Methods for Site-Adaptation of Satellite-Based DNI Time Series: Application to Brazilian Northeast. *Proceedings of the ISES Solar World Congress 2019 and IEA SHC International Conference on Solar Heating and Cooling for Buildings and Industry 2019* (Ml): 2275–82. 2019

TRUEBLOOD, C.; COLEY, S.; KEY, T.; ROGERS, L.; ELLIS, A.; HANSEN, C.; PLILPOT, E. PV Measures Up for Fleet Duty. *IEEE Power & Energy Magazine* (March/April): 33–44, 2016.

VARGAS, S. A. Previsão da distribuição da densidade de probabilidade da geração de energia eólica usando técnicas não paramétricas. Tese de Doutorado, Pontifícia Universidade Católica do Rio de Janeiro, 2015.

VERNAY, C.; BLANC, P.; PITAVAL, S. Characterizing measurements campaigns for an innovative calibration approach of the global horizontal irradiation estimated by HelioClim-3. Renewable Energy, v. 57, p. 339–347, 2013.

WILCOX, S. E MARION, W. Users manual for TMY3 data sets. Technical Report NREL/TP-581-43156, 2008.

WILKS, D. S. Statistical methods in the atmospheric sciences. 2ª. ed. [S.l.]: Elsevier, 2013.

YU, L.; LIU, H. Efficient Feature Selection via Analysis of Relevance and Redundancy. Journal of Machine Learning Research 5, p. 1205–1224, 2004.