Predição de séries temporais de radiação solar utilizando modelos estatísticos e aprendizagem profunda na região de Quixeramobim - CE

Raul Victor de O. Paiva * **Tarcisio F. Maciel** * **Wilker de O. Feitosa** * **Nícolas de A. Moreira** *

* *Departamento de Engenharia de Teleinformática, Universidade Federal do Ceará, CE, (e-mail: raul.paiva@gtel.ufc.br, maciel@gtel.ufc.br, wilker@gtel.ufc.br, nicolas@ufc.br).*

Abstract: Weather forecasting is essential in the renewable energy sector, and it is indispensable to use climate information such as air humidity, atmospheric pressure, temperature, wind speed and solar irradiation, which can be considered variables for forecasts in a certain region. In particular, there is a notable potential for the use of photovoltaic solar energy in the Brazilian Northeast due to the high solar irradiation levels in the region and, from the climatic time series, it is possible to train deep learning models that seek to predict it at short and long term. This work aims to statistically analyze and predict a time series of total daily solar irradiation in the municipality of Quixeramobim, Ceará, through machine learning methods. The results obtained indicate that the models predict solar irradiance with low prediction error when compared to existing results in the literature.

Resumo: A previsão do tempo é fundamental no setor de energias renováveis, sendo indispensável o uso de informações climáticas como umidade do ar, pressão atmosférica, temperatura, velocidade do vento e incidência de radiação solar, as quais podem ser consideradas variáveis para predições em uma determinada região. Em particular, há um notório potencial para o emprego de energia solar fotovoltaica no Nordeste brasileiro devido à grande incidência de radiação solar na região e, a partir das séries temporais climáticas, é possível treinar modelos de aprendizagem profunda que busquem prever a curto e longo prazo a irradiação solar. Este trabalho visa analisar estatisticamente e prever uma série temporal de incidência de radiação solar total diária no município de Quixeramobim, no Ceará, através de métodos de aprendizagem de máquina. Os resultados obtidos indicam que os modelos predizem a irradiação solar com erro de predição baixo quando comparado a resultados existentes na literatura.

Keywords: solar irradiation; time series; data analysis; machine learning
Palavras-chaves: radiação solar; séries temporais; análise de dados; aprendizado de máquina

1. INTRODUÇÃO

Atualmente, o avanço tecnológico fornece recursos essenciais no auxílio à tomada de decisões em áreas diversas, como na Economia, Finanças e Gestão Ambiental.

Em particular, para modelos de predição de séries temporais de dados climáticos, ainda existem limitações de processamento computacional pelo fato dos modelos numéricos clássicos, como os Modelos Numéricos de Circulação Geral (MCGs), serem muito complexos. De fato, os MCGs usam equações matemáticas representativas das leis físicas que regem os movimentos da atmosfera e as interações com os componentes do sistema climático e cuja solução depende de métodos numéricos de altíssima demanda computacional (Escobar, 2007; Sampaio and Dias, 2014). A partir daí surge a necessidade de investigar novos métodos de predição de séries temporais climáticas, mais simples, com menor custo computacional, e que forneçam bons resultados, sejam eles estatísticos ou baseados em Redes Neurais Artificiais (RNAs).

Neste contexto, a metodologia clássica usando o método *Autoregressive Integrated Moving Average* (ARIMA), também conhecida como metodologia de Box-Jenkins, é uma abordagem comumente utilizada em predição de séries temporais, buscando expressar o comportamento futuro das séries baseado na variação estatística de seus dados no passado. O método ARIMA é considerado um método clássico devido à sua generalidade, podendo lidar com séries estacionárias ou não-estacionárias, sem ou com elementos sazonais (Babai et al., 2013; Maddala, 2003).

Além dos modelos clássicos, RNAs vêm sendo empregadas com sucesso na predição de séries temporais por sua capacidade generalizada de aproximar funções não-lineares (Fernandes et al., 1996; Calôba et al., 2002; Torres-Jr et al., 2005).

De fato, conforme Mourao (2019); Teixeira et al. (2019); Pereira (2017); Ghaderi et al. (2017); Grover et al. (2015), é viável treinar modelos preditivos de Aprendizagem Profunda (AP) a partir de séries temporais. A mesma conclusão é tomada em Santos and Costa (2013) em que se afirma

que os instrumentos utilizados pelos meteorologistas foram se desenvolvendo e, com eles, a precisão das previsões do tempo foi melhorada substancialmente. Assim, o uso de AP neste trabalho para predição de séries temporais climáticas encontra parte de sua justificativa.

Por outro lado, no contexto socioeconômico a participação das fontes fósseis na matriz energética mundial é de 79,5% e, em 2017, 179 países tinham metas para aumentar essa participação em suas matrizes, um número que vem crescendo ao longo dos anos seguindo a introdução de novas políticas regulatórias para energias renováveis (REN21, 2018). No caso dos grandes empreendimentos para geração elétrica a partir da energia solar fotovoltaica no Brasil, uma expansão, iniciada em 2014 com os primeiros projetos de Usinas Fotovoltaicas (UFVs) vencedores de leilões de energia, é atribuída à redução dos custos de investimento, ao aumento da capacidade das usinas e à estimativa convidativa sobre a redução dos custos do empreendimento no horizonte de entrega da energia (EPE, 2018).

Em Mendes et al. (2017), aponta-se que o aumento das UFVs demandará dos responsáveis pela operação de sistemas elétricos a aplicação de ferramentas capazes de predizer a disponibilidade de recursos solares em curto prazo, em particular com o uso de RNAs na predição da radiação solar global. Lá destaca-se ainda que métodos de predição com RNAs podem ser aplicáveis a quaisquer regiões do Brasil, até mesmo àquelas em que não há estações de monitoramento suficientes, dada a capacidade de generalização das RNAs.

Como pode ser visualizado na Figura 1, o nordeste brasileiro possui grande potencial para ampliação da energia fotovoltaica como componente da matriz energética. No contexto regional, o Governo do Estado do Ceará tem favorecido e atraído investidores no setor das energias renováveis. Conforme EPE (2018), no leilão A-4/2018 para contratação de projetos de energia solar, dentre 29 empreendimentos que foram contratados no Brasil, 14 deles são no Ceará, totalizando 390 MW de potência a ser instalada no estado. Deste modo, a justificativa e relevância de estudos como o conduzido neste trabalho são fortalecidas e o trabalho atual encontra-se portanto alinhado às iniciativas do Governo do Estado.

Através de técnicas de aprendizagem de máquina, é possível realizar um amplo estudo referente a séries temporais climáticas nas regiões do sertão central cearense, fornecendo predições de temperatura e incidência de radiação solar, por exemplo, em uma área específica do Ceará. Logo, este trabalho colabora com o desenvolvimento do setor de energia solar no Estado.

Neste trabalho será abordado o treinamento de modelos preditivos para séries temporais de incidência de radiação solar a fim de mapear parte do potencial fotovoltaico da região de Quixeramobim-CE, utilizando o método estatístico ARIMA, bem como a AP com o *Long Short-Term Memory* (LSTM).

O restante deste artigo está organizado como segue. A Seção 2 revisita brevemente as séries temporais, o método estatístico ARIMA e o LSTM de AP. Na Seção 3, são apresentados 4 trabalhos relacionados ao presente tema proposto, os descrevendo brevemente. Já na Seção 4 apresenta a metodologia utilizada, os valores aplicados na parametrização dos estudos realizados obtidos através da análise do histograma, da densidade de probabilidade, das funções de autocorrelação e de autocorrelação parcial dos dados das séries temporais, além da aplicação de testes de estacionariedade e de tendência. Na Seção 5 a discussão dos resultados é realizada, além de apresentar a comparação entre os métodos investigados, resultados das previsões. A Seção 6 contém as principais conclusões obtidas a partir dos resultados atingidos e aponta algumas perspectivas de trabalhos futuros.

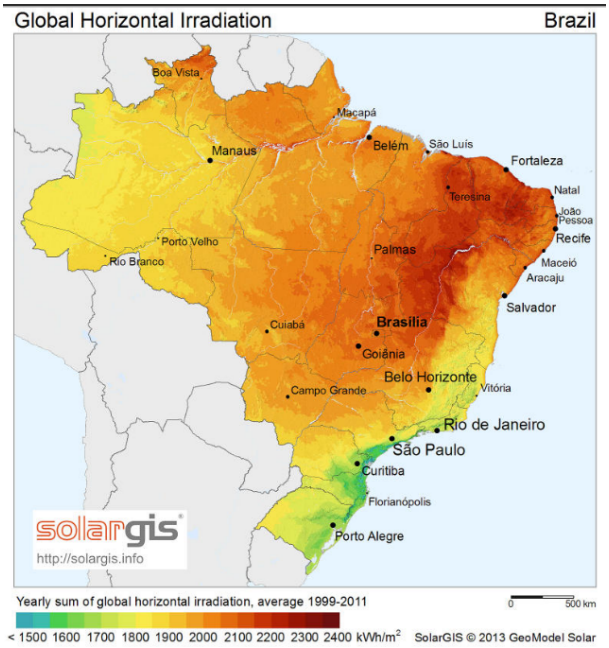

Figura 1. Irradiação solar no território Brasileiro. Média dos anos 1999 a 2011. Fonte: `http://solargis.info`.

2. MODELAGEM DE SÉRIES TEMPORAIS DE RADIAÇÃO SOLAR

2.1 Séries temporais climáticas e o modelo ARIMA

Séries de dados colhidos ao longo do tempo são chamadas de séries temporais e, quando se referem a dados climáticos, são chamadas de séries temporais climáticas. Séries temporais climáticas, como temperatura, pressão atmosférica, velocidade do vento e incidência de radiação solar, podem ser encontradas em bases de dados de órgãos competentes como o Instituto Nacional de Meteorologia (INMET) e a Fundação Cearense de Meteorologia e Recursos Hídricos (FUNCEME), entre outros, subsidiando estudos que abrangem a temática de predições de fenômenos/eventos climáticos.

Uma série temporal $y_{t=1}^{T}, t = 1, \ldots, T$, é normalmente uma realização particular de um processo estocástico que pode ser representada como:

$$y_{t=1}^{T} = \{y_1, y_2, \ldots, y_t, \ldots, y_{T-1}, y_T\}. \quad (1)$$

Através da análise dos dados contidos na série, diversas propriedades de interesse sobre os dados, como média, variância, periodicidade e estacionariedade, entre outras, podem ser extraídas a partir da aplicação de métodos

estatísticos. Estas informações podem fornecer informações importantes sobre o fenômeno subjacente aos dados e auxiliar na tomada de decisões.

Como demonstrado em Sampaio and Dias (2014); Teixeira et al. (2019); Grover et al. (2015); Mendes et al. (2017), a partir de séries temporais climáticas é possível construir modelos preditivos utilizando a modelagem clássica ARIMA e suas variantes, assim como as RNAs de AP. Além disso, é de conhecimento geral que modelos de AP permitem avaliar uma quantidade maior de parâmetros, capturando os padrões ocultos nos dados das séries temporais tornando-se, portanto, uma ferramenta de grande potencial para a área.

Passando à breve discussão sobre os modelos ARIMA, os mesmos são normalmente caracterizados por três parâmetros:

- p: o número de observações passadas inclusas no modelo, também conhecido como "ordem de *lag*", associadas ao componente Autoregressivo (AR) do método;
- d: o número de diferenças tomadas ou "grau de diferenciação", associadas à componente integrativa do método e que busca estabelecer estacionariedade;
- q: o tamanho da janela de média móvel ou "ordem de média móvel", associada a componente de Média Móvel (MA) do método.

Um modelo ARIMA é usualmente representado pela notação ARIMA(p, d, q).

No contexto dos modelos ARIMA, a metodologia Box-Jenkins investiga a autocorrelação entre valores da série em diferentes instantes sucessivos de tempo. Os padrões de autocorrelação, em geral, possibilitam identificar um ou vários modelos possíveis para a série temporal (Mehdi and Mehdi, 2010). Por este motivo, o modelo se torna bastante usual para séries temporais. Dito isto, é indispensável utilizar outros modelos para fins de comparação. As Redes Neurais Recorrentes (RNRs) fornecem alguns modelos para tratar do mesmo problema, essencialmente em AP.

2.2 RNR e LSTM

Conforme resume Teixeira et al. (2019), as RNRs são representadas por uma realimentação na arquitetura padrão. O alcance dos valores utilizados para correção dos pesos é restrito, pois a influência da saída nas camadas ocultas decai exponencialmente à medida em que passa pelas conexões recorrentes da rede. Várias tentativas para resolver esta questão, conhecida como problema de dissipação de gradiente, foram abordadas durante os anos 90, e uma das soluções adotadas foi a criação das redes LSTM.

A ideia central por trás da arquitetura LSTM é uma atualização da memória C_n. O bloco de memória *LSTM* está ilustrado na Figura 2.

Este bloco possui a capacidade de remover ou adicionar informações a esta memória em cada passo de tempo em uma sequência, controlada cuidadosamente por um *forget gate* f_n e um *input gate* i_n, que empregam a mesma estrutura de uma rede neural de única camada com a função de ativação sigmoide (Rasmus et al., 2019), e se relacionam com os demais de elementos (pesos, entradas, saídas, vieses, etc.) segundo a equação

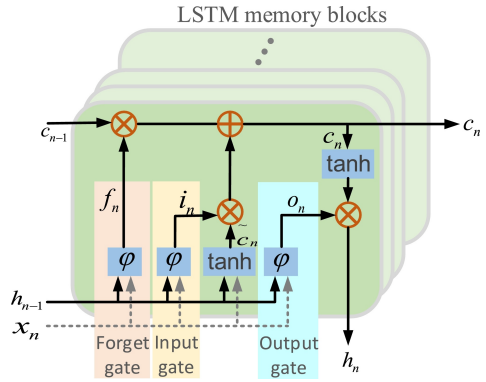

Figura 2. Representação de uma célula LSTM Rasmus et al. (2019).

$$f = \varphi(b_f + \mathbf{u}_f^T \mathbf{x}_n + \mathbf{w}_f^T \mathbf{h}_{n-1}), \tag{2a}$$
$$i_t = \varphi(b_i + \mathbf{u}_i^T \mathbf{x}_n + \mathbf{w}_i^T \mathbf{h}_{n-1}). \tag{2b}$$

Os autores em Rasmus et al. (2019) ressalvam que $\mathbf{x}_n$ é a sequência de entrada no passo de tempo n, e $\mathbf{h}_{n-1}$ é o vetor de saída da LSTM no passo de tempo passado. Os parâmetros $\mathbf{u}_i$, $\mathbf{w}_i$, $\mathbf{u}_f$ e $\mathbf{w}_f$ são as entradas e os vetores de peso recorrentes do *input* e *forget gates*, respectivamente, e os b's são todos os termos dos vieses. Os autores ainda ressaltam que a função de ativação sigmoide é responsável por controlar o quanto de cada componente deve passar. Ainda segundo Rasmus et al. (2019), a memória C_n é atualizada esquecendo parcialmente a memória existente e adicionando um novo conteúdo de memória $C_{\tilde{n}}$ dado por

$$C_n = f_n C_{n-1} + i_n \tilde{C}_n, \tag{3a}$$
$$\tilde{C}_n = \tanh(b_i + \mathbf{u}_c^T \mathbf{x}_n + \mathbf{w}_c^T \mathbf{h}_{n-1}). \tag{3b}$$

O *output gate* o_n possui estrutura familiar à do *input* e *forget gate*. A saída da LSTM é dada por

$$h_n = o_n \tanh(C_n), \tag{4a}$$
$$o_n = \tanh(b_o + \mathbf{w}_o^T \mathbf{x}_n + \mathbf{u}_o^T \mathbf{h}_{n-1}), \tag{4b}$$

em que os termos $\mathbf{u}_o$ e $\mathbf{w}_o$ são os vetores de peso de entrada e recorrente do *output gate*, respectivamente (Rasmus et al., 2019). Uma análise mais detalhada de RNRs, LSTM, e suas aplicações, foge ao escopo deste trabalho, sendo o leitor interessado em descrições mais aprofundadas direcionado às referências (Rasmus et al., 2019; Ghaderi et al., 2017; Pereira, 2017).

Dada esta breve revisão sobre de séries temporais, o contexto em que está inserido e as técnicas de modelagem, podemos montar os modelos preditivos com tais métodos após realizar uma análise prévia dos dados em questão.

3. TRABALHOS CORRELATOS

A presente seção discorre sobre alguns trabalhos que abordaram a mesma temática que o presente artigo, os descrevendo brevemente.

O primeiro a ser citado tem como título "A utilização das redes neurais artificiais na previsão de radiação solar global" por Mendes et al. (2017) que propõem utilizar RNA, mais especificamente com o *Multilayer Perceptron* (MLP), para previsão da radiação de energia solar global a partir dos dados disponibilizadas pelo Sistema de Organização

Nacional de Dados Ambientais (SONDA). Os autores incluíram a pesquisa no contexto de energias renováveis, mesmo assunto alvo do presente artigo. Os autores concluíram que o método pode ser aplicável a qualquer região do Brasil, embora tenham apresentado uma metodologia curta, o que abre espaço para experimentação de outras metodologias, objetivo do presente trabalho.

Já o artigo de Brahma and Wadhvani (2020), intitulado "Solar Irradiance Forecasting Based on Deep Learning Methodologies and Multi-Site Data", foram desenvolvidas metodologias baseadas em AP e utilizaram dados de 36 anos (1983-2019) de irradiação solar diário em dois locais da India, obtidos no repositório do projeto *Prediction Of Worldwide Energy Resources* (POWER) da *NASA*. Os resultados indicaram que o LSTM bidirecional e modelos LSTM baseados em *attention* podem ser usados para predição de irradiação solar diária. Os estudos dos autores revelaram que dados multivariados de irradiação solar melhoram os resultados sobre a performance da predição de uma localidade com dados univariados.

Um outro trabalho com mesma temática é o de Mourao (2019), "Predição de séries temporais climáticas com aprendizagem profunda", que propõe a aplicação de modelos de AP para construir preditores sobre dados climáticos. O mesmo aborda inicialmente a predição de séries de temperaturas mínima e máxima para o município de Crateús-CE, utilizando diferentes tipos de normalização de dados e depois abrange para Maceió-AL, Campos Sales-CE, Turiacu-MA, Belem-PA e Monte Alegre-PA. A proposta é interessante, mas o autor pecou em não fazer uma análise prévia dos dados em questão, ponto crucial para compreensão da problemática.

Cadenas and Rivera (2010) abordam um interessante estudo sobre predição de velocidade do vento em três localidades distintas, em seu trabalho intitulado "Wind speed forecasting in three different regions of Mexico, using a hybrid ARIMA-ANN model", em que mesmo sendo um trabalho de 12 anos atrás, a pesquisa ainda é relevante. No estudo, os autores optaram por comparar o desempenho dos modelos ARIMA (estatístico), RNA e um modelo híbrido ARIMA-RNA. Os autores concluíram que os três modelos se saíram bem, porém o melhor avaliado foi o modelo híbrido, mostrando ser uma boa alternativa para predições de velocidades do vento, em que tendências lineares e não-lineares são encontradas. É válido ressaltar que poucos trabalhos relatam a respeito do custo computacional utilizados por eles, ponto que poderia ser dado mais atenção e, por este motivo, o assunto é abordado no presente artigo.

Dessa maneira, trabalhos como esses servem de estímulo para pesquisar a respeito de desafios pertinentes, tais como meios para analisar e tratar os dados, contornar problemas de aprendizado de máquina e assim aprimorar as previsões e até mesmo meios para diminuir o custo computacional. Portanto, tais itens são parte da motivação do presente trabalho.

4. METODOLOGIA

Os estudos realizados neste trabalho utilizaram a linguagem de programação `Python 3.10` para a análise e modelagem dos dados. O computador utilizado possui um SSD de 256 GB de armazenamento, processador Intel Core i5 4ª geração *single thread*, uma placa gráfica *Nvidia Quadro K2200*, e 16 GB de RAM.

Em particular, o método ARIMA foi acessado a partir da biblioteca `statsmodels` (Seabold and Perktold, 2010). Para a otimização dos parâmetros, o método *walk-forward* também foi utilizado, o qual encontra aplicação em finanças (Kirkpatrick and Dahlquist, 2010) e lá permite determinar parâmetros ótimos para uma estratégia de *trading*. O mesmo método é aplicado aqui otimizando a amostra de dados para treinamento por meio de uma janela temporal aplicada nas séries. Os dados remanescentes são reservados para testes ficando fora da amostra de treinamento. Uma pequena porção dos dados reservados seguindo os dados na amostra é testada e os resultados são guardados. Os dados na amostra da janela temporal são deslocados para frente pelo período abordado pela saída das amostras de teste, e o processo é repetido. Maiores detalhes sobre o método de Kirkpatrick and Dahlquist (2010) podem ser vistos em seu trabalho.

Os parâmetros p e q dos modelos ARIMA foram determinados através de uma busca direta nos intervalos $0 \leq p \leq 15$ e $0 \leq q \leq 4$, respectivamente, enquanto que $d = 0$ foi adotado não tendo sido portanto diferenciada a série. O LSTM foi acessado aplicando a biblioteca `Keras`, onde também foi realizada uma busca por valores de parâmetros adequados.

4.1 Análise da série temporal

Antes de iniciar a modelagem, o objeto de estudo em questão é analisado aqui, i.e., a série temporal é analisada a fim de melhorar a compreensão sobre a mesma. Inicia-se aqui com a decomposição da série, passando para o histograma e distribuição de densidade de probabilidade, *Autocorrelation Function* (ACF) e *Partial ACF* (PACF), e finaliza-se análise com os testes *Augmented Dickey-Fuller* (ADF) e *Mann-Kendall* (MK).

A série de radiação solar coletada na região de Quixeramobim foi acessada na base de dados públicos da página oficial do INMET. A Figura 3 apresenta a série temporal completa, com 1827 observações, juntamente com a decomposição da série nas componentes de tendência, sazonalidade e resíduo.

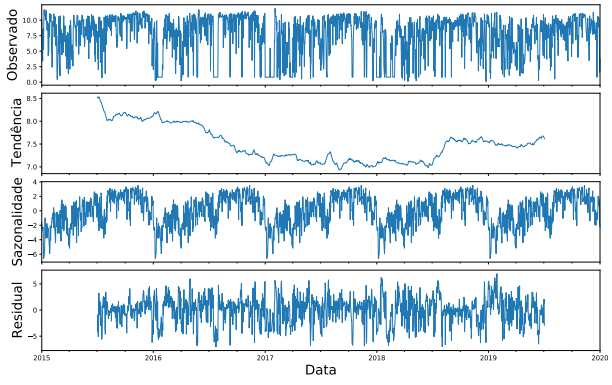

Figura 3. Decomposição da série completa nas componentes de tendência, sazonalidade e resíduo.

A decomposição foi realizada para verificar o comportamento da série do tempo em suas componentes tendência, sazonalidade e resíduos contidos nesta. É possível observar uma tendência crescente em determinadas épocas e decrescente em outras. A sazonalidade, com frequência equivalente a 365 pontos, apresenta uma série que possui um padrão intrínseco. A respeito do erro, nota-se a partir do último gráfico, grandes níveis de resíduo na série.

Na Figura 4, obtida com o método `histplot` e `distplot` da biblioteca `seaborn` com o *Kernel Density Estimation* (KDE) ativado, vê-se no histograma e na distribuição de densidade de probabilidade que a maior incidência de radiação está entre 9 a 11 kJ/m^2.

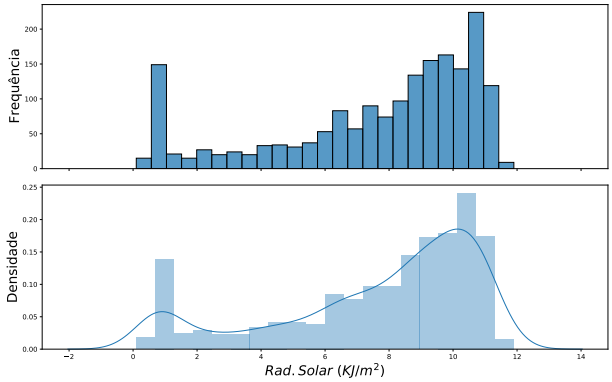

Figura 4. Histograma e densidade de probabilidade da série de radiação solar.

As análises indicadas acima são importantes para melhor compreensão dos dados em questão. Todavia, é necessário realizar uma avaliação prévia dos parâmetros (p, d, q) para iniciar a modelagem com o método ARIMA e as análises da ACF e PACF podem ajudar nesta.

Na Figura 5, pode-se notar que a ACF revela um nível de correlação significante (com nível equivalente a 5%) até por volta de 20 *lags* tendendo a zero à medida que o *lag* aumenta. Analogamente, verifica-se na PACF relevância até 3 *lags*. Estas avaliações são importantes para a determinação inicial dos parâmetros do modelo ARIMA, uma vez que os *lags* na ACF ajudam a definir o parâmetro q da média móvel, enquanto os da PACF definem o p da autorregressão.

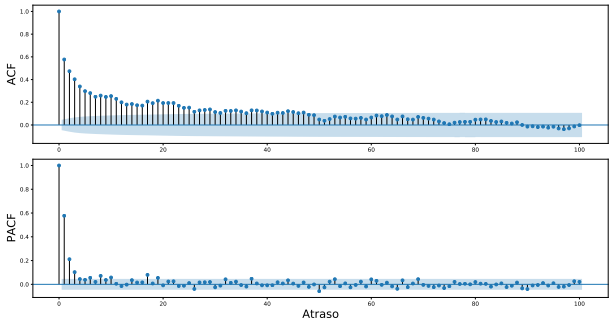

Figura 5. Funções de autocorrelação e autocorrelação parcial.

Ainda sobre os parâmetros do modelo ARIMA, também é possível verificar inicialmente se é necessário utilizar algum grau de diferenciação (d) na série. Para isto, o teste ADF fornece as respostas necessárias, pois segundo Vasconcellos (2000), ao realizar a análise de séries temporais, é necessário verificar a estacionariedade. Portanto, o teste ADF foi executado, testando a hipótese nula para a investigação da estacionariedade, com nível de significância equivalente a 5%, em que a rejeição da hipótese nula foi apontada, como indicado na Tabela 1, tanto para a série original quanto para a diferenciada.

Tabela 1. Resultados dos testes ADF e MK do conjunto de dados de Quixeramobim.

Série	ADF	Valor p	MK	Valor p
Original	-5,70	$7{,}60\times10^{-7}$	$-1{,}95\times10^{-4}$	$1{,}73\times10^{-2}$
Diferenciada	-15,90	$7{,}80\times10^{-29}$	0,00	$9{,}93\times10^{-1}$

Adicionalmente a análise de tendência, culminando em um estudo mais apurado, foi realizada através do teste MK, pois segundo Goossens and Berger (1986) o teste de MK é o método mais apropriado para analisar mudanças climáticas, além de permitir a detecção e localização aproximada do ponto inicial de determinada tendência. Os resultados dos testes estão indicados na Tabela 1. O teste MK retornou tendência decrescente para a série original, como pôde ser observado na Figura 3, e nula para a série diferenciada de primeira ordem. Diferente do que foi realizado em Mourao (2019), a breve análise foi importante para compreender a natureza da série temporal em questão.

Assim, tendo em vista a análise prévia da série de radiação solar no município de Quixeramobim, foi possível criar modelos preditivos com o modelo ARIMA e com o LSTM. Para isto, a base de dados foi dividida em 80% para treino (janeiro de 2015 a dezembro de 2018) e 20% para teste (janeiro de 2019 a janeiro de 2020), cf. Figura 6.

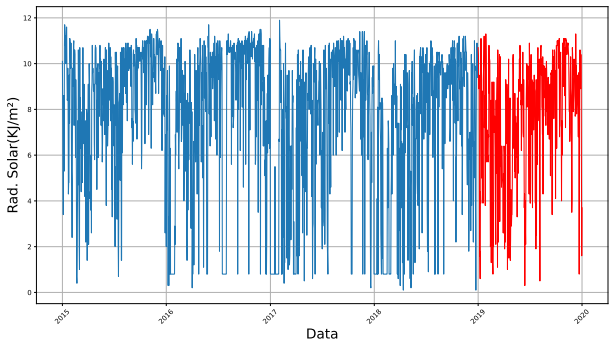

Figura 6. Divisão da base de treino e teste.

A metodologia aplicada para fazer as previsões está separada em quatro partes e está organizada como segue. A primeira parte consta em uma busca de parâmetros ótimos para ambos modelos (ARIMA e LSTM), e então a configuração de cada modelo é montada. O segundo momento revela a comparação entre as previsões realizadas, sendo que a avaliação é composta levando em conta a base de teste completa. A terceira parte revela a segunda avaliação, utilizando de Janelas de Previsão (JP) para verificar a qualidade das previsões em determinadas faixas temporais e, consequentemente, o acumulo de erro ao longo do tempo. Além das avaliações de previsões, os tempos de processamento de cada modelo também são comparados na

ultima parte da metodologia. Os resultados são mostrados na seção subsequente.

5. RESULTADOS

A presente seção tem como objetivo evidenciar os resultados obtidos a partir da metodologia aplicada e está organizada como segue. A primeira subseção destaca as previsões com o ARIMA e a segunda com o LSTM, a terceira revela as avaliações das janelas de previsão para os dois modelos e a ultima mostra os tempos de processamento de cada.

5.1 Previsões com ARIMA

A busca anterior pelos parâmetros do modelo ARIMA apontou $(7, 0, 3)$ como bons parâmetros (p, d, q) sob o ponto de vista da métrica *Root Mean Square Error* (RMSE), para a previsão do ano de 2019, cf. Figura 7. No entanto, tornou-se necessário avaliar o modelo modificando o parâmetro autorregressivo (p) com d e q fixados para fins de comparação nas métricas de avaliação *Mean Absolute Percentage Error* (MAPE), *Mean Absolute Error* (MAE) e RMSE, tornando viável estudar a influência desses na previsão, em que p varia entre 2, 7 e 15. Estas variantes são avaliadas conforme indicado na Tabela 2.

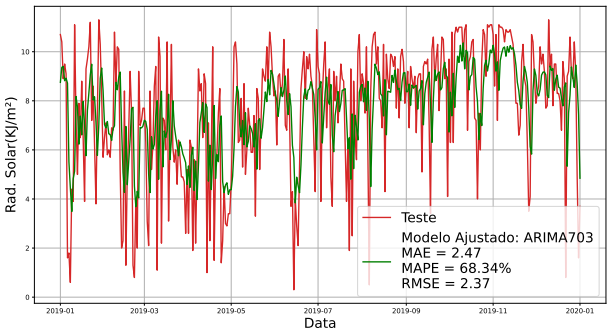

Figura 7. Predição com o modelo ARIMA(7, 0, 3).

5.2 Previsões com LSTM

Adicionalmente, foi realizado um estudo prévio na busca dos parâmetros adequados para modelagem com LSTM e julgada apropriada a rede com *dropout* equivalente a 35%, 4 camadas contendo o conjunto de unidades de células (36, 18, 18, 6, 1) do início ao fim da rede, utilizando o `Leaky Relu` como função de ativação na camada de saída e otimizador `adam` da biblioteca `Keras` da *Application Programming Interfacet* (API) do `Tensorflow`. O *Look Back* (LB) considerado inicialmente equivale a 14, ou seja, 14 pontos no passado são usados para prever estes no futuro. O treinamento foi realizado em 250 épocas. As simulações foram rodadas na placa gráfica *Nvidia Quadro K2200*. Os resultados são mostrados na próxima seção. Sob o ponto de vista da métrica RMSE, o modelo descrito acima previu o ano de 2019 errando cerca de $\pm 2{,}34$ kJ/m^2. A Figura 8 revela a comparação entre a base de teste e a previsão.

A Tabela 2 compara as métricas de avaliação da previsão do ano de 2019 com o modelo ARIMA, variando o parâmetro autorregressivo p em 2, 7 e 15, e com o LSTM, variando o LB em 2, 7, 14 e 30.

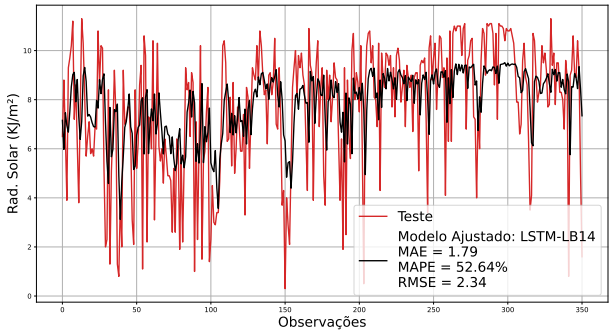

Figura 8. Predição com o modelo LSTM.

Tabela 2. Métricas de avaliação dos modelos.

	ARIMA (p)			**LSTM** (LB)			
	2	7	15	2	7	14	30
MAPE	68,03	68,34	68,35	54,15	52,03	52,64	56,49
MAE	2,46	2,47	2,47	1,84	1,85	1,79	1,83
RMSE	2,39	2,37	2,37	2,39	2,37	2,34	2,46

A partir desses resultados prévios, é possível verificar que as métricas MAPE e MAE para os modelos LSTM são bem menores que os do ARIMA. No entanto, a análise acima ainda é muito superficial, uma vez que os modelos são avaliados na base de teste completa, assim acumulando cerca de 365 pontos de erro. É importante avaliar os erros a curto, médio e longo prazo, sobretudo para as previsões com LSTM, pois é utilizado o LB, que define basicamente um horizonte de previsão. Na sequência, a avaliação das JP é visualizada. Vale ressaltar que a variação com o parâmetro p equivalente a 30 no modelo ARIMA não foi avaliado devido ao custo computacional na etapa de treinamento utilizando o método *walk-forward*.

5.3 Avaliação das JP

As janelas de predição nos intervalos de 2 a 30 dias também foram analisadas seguindo as métricas MAE e RMSE, uma vez que a base de teste possui muitas observações tornando difícil indicar um prazo de previsão ideal desses modelos, pois é mais vantajoso avaliar por janelas de previsões realizadas que um ano completo, visto que as métricas se tratam de médias. A avaliação é realizada por JP, que é a quantidade de pontos, equivalentes a 2, 7, 14 e 30 (correspondente aos dias), a serem avaliados respeitando a evolução temporal. A avaliação dessas previsões por modelo é mostrada na Tabela 3.

As avaliações presentes na tabela acima revelam como cada configuração de modelo se comporta conforme as JP mudam. Em linhas gerais, os resultados são adequados para previsões a um prazo de até 30 dias. As avaliações destacadas em cinza servem para termos um parâmetro de qual configuração melhor previu a série temporal de radiação solar total diária em distintas janelas de previsão, para as avaliações MAE e RMSE. No modelo ARIMA, é possível notar que o parâmetro p aparenta ter uma influência nas previsões, tanto que para os JP presentes, há praticamente uma equivalência entre a mudança do parâmetro p e a avaliação da JP, basta verificar as metricas destacadas, note que para p igual a 2, a melhor avaliação é para JP equivalente a 2, da mesma forma ocorre para p

Tabela 3. Avaliação das JP de curto alcance para diferentes configurações de cada modelo.

	ARIMA (p,d,q)	**MAE**	**RMSE**
(JP = 2)	$(\mathbf{2},\mathbf{0},\mathbf{3})$	1,200	1,260
	$(\mathbf{7},\mathbf{0},\mathbf{3})$	1,500	1,560
	$(\mathbf{15},\mathbf{0},\mathbf{3})$	1,726	1,767
(JP = 7)	$(\mathbf{2},\mathbf{0},\mathbf{3})$	2,110	3,216
	$(\mathbf{7},\mathbf{0},\mathbf{3})$	2,170	3,165
	$(\mathbf{15},\mathbf{0},\mathbf{3})$	2,150	3,320
(JP = 14)	$(\mathbf{2},\mathbf{0},\mathbf{3})$	2,220	3,140
	$(\mathbf{7},\mathbf{0},\mathbf{3})$	2,301	3,100
	$(\mathbf{15},\mathbf{0},\mathbf{3})$	2,180	3,100
(JP = 30)	$(\mathbf{2},\mathbf{0},\mathbf{3})$	2,150	2,740
	$(\mathbf{7},\mathbf{0},\mathbf{3})$	2,210	2,715
	$(\mathbf{15},\mathbf{0},\mathbf{3})$	2,100	2,670
	LSTM (lb!)	**MAE**	**RMSE**
(JP = 2)	$\mathbf{lb = 2}$	0,459	0,464
	$\mathbf{lb = 7}$	0,228	0,295
	$\mathbf{lb = 14}$	1,084	1,220
	$\mathbf{lb = 30}$	1,230	1,240
(JP = 7)	$\mathbf{lb = 2}$	3,100	4,265
	$\mathbf{lb = 7}$	1,632	2,556
	$\mathbf{lb = 14}$	2,352	2,527
	$\mathbf{lb = 30}$	0,900	1,082
(JP = 14)	$\mathbf{lb = 2}$	2,600	3,700
	$\mathbf{lb = 7}$	1,980	2,620
	$\mathbf{lb = 14}$	2,423	2,813
	$\mathbf{lb = 30}$	1,980	2,771
(JP = 30)	$\mathbf{lb = 2}$	2,460	3,240
	$\mathbf{lb = 7}$	1,880	2,480
	$\mathbf{lb = 14}$	2,260	2,805
	$\mathbf{lb = 30}$	2,173	3,025

$= 7$, ainda que o melhor MAE tenha sido atribuído ao de 2, para $p = 15$ segue a equivalência e com JP equivalente 30, a melhor ainda é para $p = 15$.

Já a respeito da modelagem com LSTM, ao mudar o LB também é possível notar uma determinada influência deste na previsão. Nas métricas de avaliação mostradas ainda na Tabela 3, note pelos resultados destacados que as previsões para LBs equivalentes a 2 e 7 são muito bem avaliadas dentro da JP = 2, já para JP = 7 o melhor resultado foi para LB igual a 30, enquanto para JP = 30 o mais acertado foi com LB correspondente a 7.

Os resultados poderiam ser bem superiores se não fosse a grande quantidade de ruído presente na série temporal original. Esse tipo de problema é muito recorrente no eixo de pesquisa desse ramo, uma vez que estamos trabalhando com variáveis aleatórias. Ainda assim, as soluções propostas aqui foram interessantes.

A partir dos resultados obtidos com a metodologia aplicada, podemos inferir que o LSTM entrega melhores resultados em comparação com o ARIMA sob o ponto de vista das métricas avaliadas. No geral, os resultados foram inferiores aos de Brahma and Wadhvani (2020) para previsão de séries temporais de irradiação solar, em que os melhores resultados obtidos pelos autores foram para o modelo *Bidir* no cenário *Multi-Location* com *Mean Square Error* (MSE) e RMSE equivalente a 9.094×10^{-3} e 9.536×10^{-2}, respectivamente, para a primeira localidade e 7.610×10^{-3} e 7.610×10^{-3} para a segunda localidade, ambos para horizonte de previsão equivalente a 1. Entretanto, os resultados apresentados aqui foram bem melhores que os de Mendes et al. (2017), em que o melhor resultado obtido pelos autores na etapa de validação foi de 19.88 da métrica MAE, enquanto nós obtivemos 2.47 e 1.79 com os modelos ARIMA e LSTM, respectivamente.

Mesmo com os resultados das métricas de avaliação obtidos, por outro lado ainda é necessário avaliar o tempo de processamento de cada variação dos modelos propostos e verificar os tempos de treinamento e predição.

5.4 Tempo de processamento dos modelos

Para estimar o tempo de processamento dos métodos, para cada configuração dos modelos descritos foram realizadas 30 repetições, extraindo-se delas os tempos mínimo, máximo, médio μ e o desvio padrão σ dos mesmos. Os resultados obtidos estão mostrados na Tabela 4.

Tabela 4. Tempos de processamento (s).

		$\boldsymbol{\mu \pm \sigma}$	**[min; máx]**
ARIMA (Treinamento)	$(\mathbf{2},\mathbf{0},\mathbf{3})$	$377{,}60 \pm 7{,}50$	[367,50; 397,15]
	$(\mathbf{7},\mathbf{0},\mathbf{3})$	$1275{,}40 \pm 15{,}90$	[1254,8; 1324,8]
	$(\mathbf{15},\mathbf{0},\mathbf{3})$	$3811{,}80 \pm 52{,}70$	[3760,80; 3958,20]
LSTM (Treinamento)	$\mathbf{lb = 2}$	$93{,}50 \pm 2{,}70$	[89,20; 98,50]
	$\mathbf{lb = 7}$	$92{,}30 \pm 10{,}7$	[85,80; 148,7]
	$\mathbf{lb = 14}$	$91{,}90 \pm 3{,}10$	[86,50; 99,90]
	$\mathbf{lb = 30}$	$94{,}20 \pm 10{,}3$	[89,85; 148,9]
ARIMA (Predição)	$(\mathbf{2},\mathbf{0},\mathbf{3})$	$1{,}02 \pm 3{,}00 \times 10^{-2}$	[0,99; 1,10]
	$(\mathbf{7},\mathbf{0},\mathbf{3})$	$1{,}03 \pm 1{,}60 \times 10^{-2}$	[1,00; 1,07]
	$(\mathbf{15},\mathbf{0},\mathbf{3})$	$1{,}01 \pm 1{,}90 \times 10^{-2}$	[0,98; 1,06]
LSTM (Predição)	$\mathbf{lb = 2}$	$1{,}30 \pm 2{,}65 \times 10^{-1}$	[1,10; 2,30]
	$\mathbf{lb = 7}$	$1{,}50 \pm 3{,}45 \times 10^{-1}$	[1,15; 2,40]
	$\mathbf{lb = 14}$	$1{,}60 \pm 3{,}10 \times 10^{-1}$	[1,10; 2,50]
	$\mathbf{lb = 30}$	$1{,}55 \pm 1{,}90 \times 10^{-1}$	[1,30; 2,00]

Conforme os dados da Tabela 4, constata-se que o LSTM tem melhor desempenho que o ARIMA, pois este utiliza o método *walk-forward* na etapa de treinamento que inclui a validação da previsão por etapas dentro da base de treino e eleva o custo computacional de cada realização. Isto pode ser verificado pelo aumento do tempo médio com o aumento de p.

6. CONCLUSÃO

Neste trabalho foi realizada uma breve motivação o potencial fotovoltaico do território nordestino, bem como uma rápida visita aos modelos ARIMA e LSTM, contextualizando sua aplicação na predição de irradiação solar na região de Quixeramobim.

Adicionalmente, a avaliação das janelas de previsão revelou que os modelos de AP possuem bons resultados quando submetidos a curtos prazos e que a influência do *look back* é a causa disso, ao menos com os parâmetros utilizados. Entretanto, ainda é necessário avaliar uma quatidade maior de parâmetros na tentativa de fazer previsões mais acertadas a médio prazo, uma vez que predições a longo prazo

são um grande desafio no contexto de séries temporais climáticas.

A análise dos erros de predição e a breve comparação com resultados disponíveis na literatura, juntamente com a verificação dos tempos de processamento que destacou o impacto do custo computacional necessário para realização dos modelos nas etapas de treino e teste, foram satistórios a proposta do presente artigo. A respeito dos resultados dos tempos de processamento, o método do ARIMA foram mais custosos, pois foram executados apenas no processador *single thread* sem a possibilidade de uso da *Graphics Processing Units* (GPU) para as bibliotecas aqui consideradas e teve um desempenho inferior ao LSTM. Como o `Tensorflow` facilita o processamento de modelos de AP com suporte a GPUs, principalmente da *Nvidia*, o LSTM foi favorecido.

Logo, a aplicação desses modelos de AP parece adequada ao estudo de séries temporais climáticas. A previsão acertada da série original é um significativo desafio, e buscar diferentes maneiras de lidar com este fato, dependendo da aplicação, pode ser um caminho viável. Dessa forma, fica como uma perspectiva para futuros trabalhos aplicar séries de irradiação filtradas no LSTM, ou seja, destinar médias móveis para realização das previsões e justificar o uso em contextos distintos.

REFERÊNCIAS

Babai, M.Z., Ali, M.M., Boylan, J.E., and Syntetos, A.A. (2013). Forecasting and inventory performance in a two-stage supply chain with arima (0, 1, 1) demand: Theory and empirical analysis. *International Journal of Production Economics*, 143(2), 463–471.

Brahma, B. and Wadhvani, R. (2020). Solar irradiance forecasting based on deep learning methodologies and multi-site data. *Symmetry*, 12(11), 1–20.

Cadenas, E. and Rivera, W. (2010). Wind speed forecasting in three different regions of mexico, using a hybrid arima-ann model. *Renewable Energy - Elsevier*, 35, 2732–2738.

Calôba, G.M., Calôba, L.P., and Saliby, E. (2002). Cooperação entre redes neurais artificiais e técnicas clássicas para previsão de demanda de uma série de vendas de cerveja na austrália. *Pesquisa Operacional*, 22, 345–358.

EPE (2018). *Projetos fotovoltaicos nos leilões de energia: Características dos empreendimentos participantes nos leilões de 2013 a 2018*. Empresa de pesquisa energética.

Escobar, G.C.J. (2007). Padrões sinóticos associados a ondas de frio na cidade de são paulo. *Revista Brasileira de Meteorologia*, 22, 241–254.

Fernandes, L.G.L., Portugal, M.S., and Navaux, P.O.A. (1996). Previsão de séries de tempo: redes neurais artificiais e modelos estruturais. *Pesquisa e Planejamento Econômico*, 26(2), 253–276.

Ghaderi, A., Sanandaji, B.M., and Ghaderi, F. (2017). Deep forecast: Deep learning-based spatio-temporal forecasting. In *The 34th International Conference on Machine Learning (ICML), Time series Workshop*.

Goossens, C. and Berger, A. (1986). Annual and seasonal climatic variations over the northern hemisphere and europe during the last century. In *Annales Geophysicae*, volume 4, 385–400.

Grover, A., Kapoor, A., and Horvitz, E. (2015). A deep hybrid model for weather forecasting. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 379–386.

Kirkpatrick, C.D. and Dahlquist, J. (2010). *Technical analysis: the complete resource for financial market technicians*. Pearson Education, Inc., second edition.

Maddala, G.S. (2003). *Introdução à econometria*. Rio de Janeiro: LTC, third edition.

Mehdi, K. and Mehdi, B. (2010). An artificial neural network (p, d, q) model for timeseries forecasting. *Expert Systems with applications*, 37(1), 479–489.

Mendes, I.A., Rezende, R.A.D., Ferreira, T.H., e Silva Nascimento, J.S.F., and Silva, O.F. (2017). A utilização das redes neurais artificiais na previsão de radiação solar global. In *Congresso Técnico Científico da Engenharia e da Agronomia. Belém-PA*.

Mourao, I.S. (2019). Predição de séries temporais climáticas com aprendizagem profunda. Monografia (Bacharel em Ciência da Computação).

Pereira, M.M. (2017). Aprendizado profundo: Redes lstm. Monografia (Bacharel em Sistemas de Informação).

Rasmus, A.S., Peimankar, A., and Puthusserypady, S. (2019). A deep learning approach for real-time detection of atrial fibrillation. *Expert Syst. Appl.*, 115, 465–473.

REN21 (2018). *Renewables 2018 global status report. A comprehensiveannual overview of the state of renewable energy*. Renewable energy policy network for the 21st century.

Sampaio, G. and Dias, P.S. (2014). Evolução dos modelos climáticos e de previsão de tempo e clima. *Revista USP*, (103), 41–54.

Santos, A. and Costa, O.A. (2013). Sistema de recepção de dados do satélite meteosat-9 na secretaria de meio ambiente e recursos hídricos–sergipe: Implementação e aplicações. In *Anais XVI Simpósio Brasileiro de Sensoriamento Remoto - SBSR*.

Seabold, S. and Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*. URL `https://www.statsmodels.org/stable/index.html`.

Teixeira, R., Silva, D., Mello-Junior, H., Forero, L., Lima, A., and Figueiredo, K. (2019). Previsão de séries temporais de velocidade do vento utilizando redes neurais artificiais e métodos estatísticos na região de arraial do cabo-rj. In *Anais do 14 Congresso Brasileiro de Inteligência Computacional*, 1–7.

Torres-Jr, R.G., Machado, M.A.S., and Souza, R.C. (2005). Previsão de séries temporais de falhas em manutenção industrial usando redes neurais. *Engevista*.

Vasconcellos, M.A.S. (2000). *Manual de econometria*. São Paulo: Editora Atlas, first edition.