MODELOS DE PREVISÃO DA CONCENTRAÇÃO DE GÁS CARBÔNICO NA ATMOSFERA

TACIANA ARAÚJO-SILVA; PAULO RENATO ALVES FIRMINO; FRANK GOMES-SILVA

RESUMO

Uma previsão eficiente da emissão de dióxido de carbono pode contribuir para fomentar políticas públicas de redução da emissão desse gás na atmosfera. Nesse artigo são formulados quatro modelos otimizados baseados nas metodologias ARIMA, ETS, ANN e SVR, para prever a emissão de CO^2. Adotou-se uma série temporal da concentração de dióxido de carbono atmosférico obtida através da Administração nacional oceânica e atmosférica. O estudo foi conduzido com um subconjunto, no total de 80%, da série na etapa treinamento dos modelos e outro para teste. Empregou-se um conjunto de medidas de desempenho para a avaliação dos modelos. O modelo que apresentou melhores resultados foi SVR, enquanto o modelo menos promissor foi o ETS, que não se destacou em nenhuma das métricas quando comparado aos demais.

Palavras-chave: Modelagem e previsão de séries temporais, Regressão de vetores de suporte, Redes neurais artificiais.

1 INTRODUÇÃO

As mudanças climáticas se tornaram motivo de preocupação essencial para a sociedade e têm despertado a atenção de cientistas ao longo do tempo. O aquecimento global tem impacto em diversos setores como imigração, agricultura e trouxe efeitos negativos para a sociedade e vida humana. Eventos extremos têm ocorrido devido à alta emissão de gases do efeito estufa na atmosfera (ZHOU *et al*, 2021; FANG *et al*, 2018).

O gás carbônico, ou dióxido de carbono, (CO^2) é um dos principais gases do efeito estufa e sua emissão tem aumentado continuamente nos últimos anos, mesmo com inúmeras políticas públicas para redução. Com o crescimento da população mundial é natural que os níveis de gases do efeito estufa sejam elevados, uma vez que há aumento de demanda de energia entre outros. A principal fonte de emissão de CO^2 é a combustão de combustíveis fósseis, como petróleo e gás natural (LI *et al*, 2018).

A previsão da emissão de CO^2 é uma preocupação mundial e uma ferramenta de conscientização pública na tentativa de solucionar os problemas climáticos, portanto é de suma importância uma previsão confiável de suas emissões futuras (NYONI; BONGA, 2019).

O estudo estatístico-computacional de séries temporais relacionadas a esse tema pode contribuir para o desenvolvimento de ações necessárias de prevenção e mitigação de efeitos danosos ao meio ambiente provenientes do acúmulo desse gás na atmosfera. Uma série temporal é uma sequência de observações cronológicas amostradas de um fenômeno (MONTGOMERY; JENNINGS; KULAHCI, 2015).

DOI: 10.51189/ii-econline/13709

Na literatura é possível encontrar modelos ARIMA (LI *et al*, 2018) e de ANN (NGUYEN; HALEM, 2018) para previsão de CO^2. Em 2018, Fang *et al* realizou a previsão da emissão de dióxido de carbono com base na regressão de processos gaussianos aprimorados com base em PSO modificado obtendo bons resultados para EUA, China e Japão.

Mais recentemente, Kallio *et al* (2021) sugeriu modelos de aprendizados de máquina baseados em regressão de Ridge, árvore de decisão, floresta aleatória e Perceptron multicamadas, para modelar a concentração futura de CO^2 em residências e Zhou *et al* (2021) propôs um novo mecanismo de rolagem cinza baseado no princípio de prioridade da informação e obteve resultados satisfatórios para prever e analisar a tendência das emissões de dióxido de carbono na China.

Neste trabalho objetiva-se avaliar estatisticamente o desempenho de quatro modelos preditivos otimizadas, Modelo Autorregressivo Integrado de Médias Móveis (ARIMA), Modelos de Suavização Exponencial (ETS), Redes Neurais Artificiais (RNA) e Regressão de Vetores de Suporte (SVR), para previsão da concentração de CO^2 na atmosfera. Na busca pelo modelo mais promissor, será avaliada a eficiência de cada um junto ao um rigoroso conjunto de métricas.

2 MATERIAIS E MÉTODOS

2.1 Formalismos de Modelagem Individual

Modelos Autorregressivos Integrados de Médias Móveis - ARIMA

A classe de modelos ARIMA é uma importante ferramenta de previsão e é a base de muitas ideias fundamentais na análise de séries temporais (CHATFIELD, 2000). Podemos escrever o modelo ARIMA como uma equação linear na forma

$$z_t^{(d)} = \theta_0 + \sum_{i \in I_u} \theta_i^{ar} z_{t-t_i}^{(d)} + \sum_{j \in I_{\hat{u}}} \theta_j^{ma} e_{t-t_j}^{(d)} + e_t^{(d)}, \quad (1)$$

na qual θ_0 é o intercepto do modelo, *d* é o número de diferenciações necessárias para estabilizar a série z_t, que indica, por sua vez, a ordem de integração *I* do modelo, o primeiro e o segundo somatórios indicam, respectivamente, as partes autorregressiva (AR) e de médias móveis (MA) do modelo. Enquanto $\{\theta_i^{ar}\}_{i \in I_u}$ representa os coeficientes dos termos autorregressivos e $\{\theta_j^{ma}\}_{j \in I_{\hat{u}}}$ são os coeficientes dos resíduos do modelo e $e_t^{(d)}$ é o resíduo correspondente ao ajuste do modelo a z_t. A primeira diferença é dada por $z_t^{(1)} = z_t - z_{t-1}$ e a *d*- ésima diferença é $z_t^{(d)} = z_t^{(d-1)} - z_{t-1}^{(d-1)}$.

t t $t-1$

Modelos de Suavização Exponencial - ETS

Os modelos de suavização exponencial são bastante populares devido à sua simplicidade, eficiência computacional e razoável precisão (MORETTIN e TOLOI, 2018). Essa técnica consiste em atribuir pesos a observações passadas da série que decaiam exponencialmente, da mais recente a mais distante, ao longo do tempo, fazendo com que observações mais recentes tenham pesos maiores.

Este formalismo propõe a decomposição da série em suas componentes erro (*e*), tendência (T) e sazonalidade (S), podendo ser classificados de acordo com o comportamento dessas componentes. A tendência pode-se apresentar como nenhuma, aditiva, multiplicativa, aditiva amortecida ou multiplicativa amortecida, já a sazonalidade pode ser nenhuma, aditiva ou multiplicativa, considerando ainda o erro como aditivo ou multiplicativo, formando um total de 30 modelos distintos. Matematicamente o modelo aditivo pode ser expresso como na Equação 2 enquanto o multiplicativo está representado na Equação 3, vale salientar que outras variações podem ocorrer.

$$z_t = T + S + e\ , \tag{2}$$

$$z_t = T.S.e\ . \tag{3}$$

Redes Neurais Artificiais - RNA

Inspirada no funcionamento do cérebro humano, uma Rede Neural Artificial (RNA) é um processador paralelo distribuído, constituído de unidades de processamento simples (neurônios ou nodos), que são naturalmente propensos a armazenarem conhecimento experimental e torná- lo disponível para o uso (HAYKIN, 2001).

O processo de uma RNA consiste e 3 ações realizadas pelos seus elementos básicos: 1) o conjunto de dentritos (*I*) recebe os estímulos externos ao corpo celular (S); 2) S pondera esses estímulos através de operações agregativas simples e 3) o axônio (*A*) processa a informação recebida de S, por meio de operações mais sofisticadas, gerando as respostas dos neurônios aos estímulos da camada de entrada. Para séries temporais usa-se uma rede com uma camada intermediária, levando a uma função de soma simples, onde θ_i é o coeficiente que pondera as observações z_{t-t_i} e θ_0 é o intercepto,

$$S_h(I_h) = \sum_{i=1}^{I} \theta_{hi} \qquad + \theta_{h0}, \tag{6}$$
$$z_{t-t_i}$$

DOI: 10.51189/ii-econline/13709

com respectivas funções de ativação $A_h(S_h(I_h))$, que são agora operadas por uma função de ativação final que dá uma estimativa para z_t,

$$\hat{z_t} = A_0(\sum_{h=1}^{H} \theta_{hi} A_h(S_h(I_h))). \quad (7)$$

***Support Vector Regression* - SVR**

A *Support Vector Regression* - SVR busca uma função capaz de prever valores futuros da série. Seja $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^n \times \mathbb{R}$ uma amostra de treinamento, em que $\boldsymbol{x_i}$ é um vetor de entradas *n*-dimensional, y_i são as saídas e n é o número de observações desse conjunto, a relação entre a entrada e saída é definida pela fórmula (KANG; LI, 2016):

$$f(x) = \langle \boldsymbol{w}\boldsymbol{x} \rangle + b \ , \quad (8)$$

onde $f(x)$ é o valor predito com, no máximo, um desvio ε de y_i, b é o intercepto, $\boldsymbol{w}$ é um vetor de pesos, $\langle . \rangle$ indica o produto interno na dimensão $\mathbb{R}^n$.

Em alguns casos é necessário estender a regressão a um caso de regressão não linear, para tal utiliza-se uma função chamada Kernel que mapeia os dados do espaço de entrada para uma dimensão superior na qual a regressão torna-se possível. Após alguma álgebra, a Equação 8 pode ser reescrita na forma

$$f(x) = \sum_{i=1}^{n} (\alpha - \alpha^*) K(x_i, x_j) + b \ , \quad (9)$$

sendo α e α^*os multiplicadores de Lagrange e $K(.)$ uma função Kernel.

Estudo de Caso

Os dados utilizados neste trabalho tratam-se de dados reais Concentração de Dióxido de Carbono na Atmosfera (CO^2). A série possui periodicidade mensal, conta com 726 observações, sendo que destas, 580 foram destinadas para o treinamento dos modelos e 145 para o teste. A série estudada foi analisada sob a ótica da teoria de séries temporais em um procedimento divido em 3 etapas:

i) inicialmente, a série foi dividida em dois subconjuntos, um para treinamento, com 80% dos dados, e outro para teste dos modelos com 20% dos dados, a série foi normalizada no intervalo [0,4; 0,6], dada a necessidade das entradas das RNA's estarem no intervalo [0,0; 1,0];

ii) implementação dos modelos individuais ARIMA, ANN, ETS e SVR;

Os melhores modelos ARIMA e ETS foram selecionados através do BIC, por sua vez os modelos RNA e SVR foram otimizados pelo Algoritmo *Sumulated Annealing* (SA), que também busca a minimização do BIC. Para variações do AL na RNA foram adotados: BPROP, RPROP+, RPROP-, SAG e SLR e para a FA tomou-se FL, FS e FTH. No caso do SVR, o tipo adotado foi *eps-regression* e as variações para função Kernel foram: *Linear* (LK), *Polynomial* (PK), *Sigmoid* (SK) e *Radial Basis* (RBK);

iii) avaliação do desempenho dos modelos através de um conjunto das métricas de desempenho, a saber, MSE, MAPE, ARV, ID, Theil e W-POCID.

Para efeito de comparação dos métodos a nível de desempenho, serão aplicados a séries temporais reais, destinando-se 80% dos valores observados para o treinamento e o restante para o teste do modelo. Considerando $\hat{z_t}$ o valor previsto para z_t no instante t e n o total de observações regressas, o desempenho dos modelos será avaliado diante das seguintes métricas:

$$MSE = \sum_{t=1}^{n} (z_t - \hat{z_t})^2/n \quad (11)$$

$$MAPE = 100/n \sum_{t=1}^{n} |z_t - \hat{z_t}/z_t| \quad (12)$$

$$Theil = \sum_{t=1}^{n} (z_t - \hat{z_t}$$

DOI: 10.51189/ii-econline/13709

$)^2 / \sum^{n} \ (z_t - z_{t-1})^2$

$$\sum_{t=1}^{n} \quad \sum_{t=1}^{n} \quad (13)$$

$$ARV = \sum_{t=1}^{n} (z_t - \hat{z_t})^2 / \sum_{t=1}^{n} (z_t - \bar{z_{t-1}})^2 \quad (14)$$

$$POCID = 100 \times \sum_{t=1}^{n} D_t / n \quad (15)$$

O MSE avalia a precisão e eficiência do modelo, assim como o MAPE, que o faz em valores relativos. O Theil compara a previsão feita pelo modelo com um *random walker* e o ARV o faz diante da média dos valores regressos da série. Por sua vez, o POCID indica a taxa de acerto do modelo quanto a previsão de tendência. Neste estudo foi adotado o complementar do POCID, WPOCID = 1 – POCID, para buscar o modelo capaz de minimizar todas as métricas.

Todas as análises, gráficos e tabelas foram geradas no *software* gratuito R (COMPUTING, 2015), através da interface RStudio.

3 RESULTADOS E DISCUSSÃO

Na Tabela 1 estão descritos os modelos obtidos para a série CO^2. O modelo RNA utilizou 14 entradas autorregressivas e apenas um nó na camada intermediária, sendo o que mais acertou a previsão de tendência da série (WPOCID = 0.035). O ARIMA, por sua vez, utilizou dois termos autorregressivos, um de média móvel e uma diferenciação para estabilizar a série. No modelo ETS, adota-se um erro aditivo e nenhuma sazonalidade ou tendência, se aproximando de um *random walk*. O modelo SVR utilizou a função kernel polinomial, com apenas 5 vetores de suporte.

Tabela 1 – Descrição dos modelos individuais para a série CO^2.

Formalismo	Modelo	Descrição
ANN	ANN(14, 1, 1)	PAR = 13, PARS = 1, S = 24, LA = SLR, F A = LF

DOI: 10.51189/ii-econline/13709

ARIMA	ARIMA(2, 1, 1)	-
ETS	ETS(A, N, N)	$\alpha = 1.0$
SVR	-	PAR = 13, PARS = 5, S = 24, cost = 20.4725, ε = 0.0877, kernel = P K, SV = 5, degree = 1.7707, coef 0 = 0.4813 γ = 1589.2588

A Tabela 2 mostra o desempenho dos modelos individuais para a série CO^2, da qual pode-se observar o desempenho superior do SVR, que foi o melhor modelo entre todos. Com destaque para o pequeno número de erros cometidos pelo modelo (ver MSE e MAPE). O modelo ETS teve baixo desempenho perante todas as métricas, quando comparado aos demais modelos.

Tabela 2 – Performance dos modelos de previsão (ANN, ARIMA, ETS, SVR) para a série CO^2 (fase de teste). Os melhores valores encontram-se em negrito.

Métrica	ANN	ARIMA	ETS	SVR
MSE	1.830	0.605	1.764	**0.539**
MAPE	0.003	0.002	0.003	**0.001**
ARV	0.037	0.009	0.026	**0.009**
Theil	1.054	0.348	1.000	**0.310**
WPOCID	**0.035**	0.167	0.167	0.049

A representação gráfica das previsões dos modelos individuais para a série CO^2 pode ser observada na Figura 1.

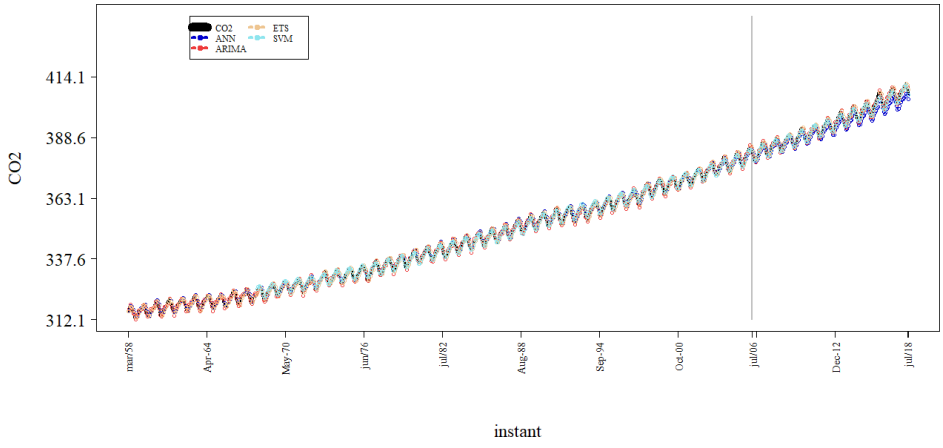

Figura 1: Modelos Individuais (ANN, ARIMA, ETS e SVR) para a série CO^2 (Fase de teste).

Ao calcularmos as médias agregadas normalizadas dos valores de todas as medidas de desempenho utilizadas, observa-se que o SVR foi o melhor modelo, cujo valor da média foi de 0.0239.

4 CONCLUSÃO

DOI: 10.51189/ii-econline/13709

Neste trabalho foram analisadas quatro metodologias preditivas com o objetivo de avaliar estatisticamente o desempenho de cada uma ao prever a concentração de dióxido de carbono na atmosfera, os modelos foram adotados ARIMA, ETS, RNA e SVR.

Os resultados mostraram, sob a óptica do conjunto de métricas de desempenho aqui utilizadas, que o modelo SVR teve desempenho superior na previsão de CO^2. O modelo com pior desempenho foi o ETS, que não se destacou em nenhuma das métricas analisadas.

A previsão e os resultados aqui mostrados podem contribuir na tomada de decisão e direcionamento de políticas de redução da concentração de dióxido de carbono na atmosfera.

REFERÊNCIAS

CHATFIELD, C. **Time-series forecasting**. Washington, USA: Chapman & Hall/CRC, 2000.

COMPUTING, R. F. for S. A language and environment for statistical computing. 2015. Disponível em: https://www.R-project.org/.

FANG, D. et al. A novel method for carbon dioxide emission forecasting based onimproved gaussian processes regression. **Journal of cleaner production**, **Elsevier**, v. 173, p. 143–150, 2018.

HAYKIN, S. **Redes Neurais princípios e prática.** Porto Alegre: Boobman, 2001.

KALLIO, J. et al. Forecasting office indoor co2 concentration using machine learning with a one-year dataset. **Building and Environment, Elsevier**, v. 187, p. 107409, 2021.

KANG, F.; LI, J. Artificial bee colony algorithm optimized support vector regression for system reliability analysis of slopes. **Journal of Computing in Civil Engineering,** v. 30, p. 3–14, 2016.

LI, F. et al. Modelling of a post-combustion co2 capture process using deep belief network. **Applied Thermal Engineering, Elsevier**, v. 130, p. 997–1003, 2018.

MONTGOMERY, D. C.; JENNINGS, C. L.; KULAHCI, M. **Introduction to time series analysis and forecsating**. New Jersey, USA: John Wiley & Sons, 2015.

MORETTIN, P. A.; TOLOI, C. M. **Análise de Séries Temporais**. São Paulo: Blucher, 2018.

NGUYEN, P.; HALEM, M. Prediction of co2 flux using long short term memory (lstm) recurrent neural networks with data from flux towers and oco-2 remote sensing. In: **AGU Fall Meeting Abstracts**. [S.l.: s.n.], 2018.

NOOA. Administração nacional oceânica e atmosférica: tendências do dióxido de carbono atmosférico. 2018.

NYONI, T.; BONGA, W. G. Prediction of co2 emissions in india using arima models. DRJ-**Journal of Economics & Finance,** v. 4, n. 2, p. 01–10, 2019.

ZHOU, W. et al. Forecasting chinese carbon emissions using a novel grey rolling prediction model. **Chaos, Solitons & Fractals, Elsevier**, v. 147, p. 110968, 2021.

DOI: 10.51189/ii-econline/13709