

METODOLOGIA ESTRUTURADA BASEADA EM DESIGN OF EXPERIMENTS (DOE) PARA AVALIAÇÃO COMPARATIVA DE MÉTODOS ESTATÍSTICOS EM DADOS SINTÉTICOS

Francielle da Silva Leandro – franciellesilvaleandro@unifei.edu.br
Pedro Paulo Balestrassi – pedro@unifei.edu.br
Simone Carneiro Streitenberger – simone.p4@gmail.com

Palavras-chave: Planejamento de experimentos (DOE), análise estatística, comparação de métodos, dados sintéticos.

1. INTRODUÇÃO

Séries temporais lineares e não lineares estão presentes em diferentes contextos de empresas e indústrias sendo vital a análise correta dos seus dados para tomadas de decisão e compreensão de fenômenos dinâmicos ao longo do tempo. A escolha do método estatístico mais adequado impacta diretamente a qualidade das inferências realizadas. No entanto, com a vasta gama de técnicas e métodos estatísticos disponíveis, a identificação de qual método é mais adequado para resolver um problema específico, dificulta o processo de decisão e pode levar o pesquisador a conclusões incorretas.

Os estudos comparativos frequentemente aplicam múltiplos métodos - como redes neurais, aprendizado de máquina e/ou métodos estatísticos tradicionais - a um único banco de dados, utilizando uma métrica estatística para avaliação dos resultados e considera-se como melhor técnica aquela que apresenta o melhor desempenho na métrica analisada. Contudo, esse tipo de comparação apresenta riscos de conclusões equivocadas ou superficiais, uma vez que não são consideradas análises estatísticas abrangentes, generalização de cenários e a identificação dos fatores mais relevantes para o problema em questão.

Com base nos estudos avaliados, identifica-se uma lacuna na literatura quanto à utilização de uma ferramenta para uma análise sistemática de comparação entre diferentes métodos estatísticos utilizando dados sintéticos para problemas de séries temporais.

A presente pesquisa busca superar essas limitações ao propor uma metodologia abrangente de avaliação baseada no design de experimentos (DOE), fundamentada em princípios estatísticos rigorosos, para avaliação sistemática e reproduzível, identificação de padrões de desempenho sob diversas condições experimentais, avaliação de métodos estatísticos por meio de simulação computacional de dados sintéticos, com obtenção de resultados escaláveis, reproduzíveis e interpretáveis, com eficiência significativamente superior às avaliações *ad hoc* ou baseadas em tentativa e erro, métodos comumente adotados por pesquisadores.

2. DESENVOLVIMENTO

Não há um método estatístico que possa ser utilizado universalmente para a resolução de problemas de séries temporais, sendo assim, a análise do desempenho dos diferentes métodos, depende dos parâmetros escolhidos e a natureza dos dados analisados.

2.1 Referencial teórico

De acordo com Morettin e Toloi (2018), uma série temporal consiste em valores observados sequencialmente ao longo do tempo. São caracterizadas por: dependência temporal (correlação), padrões sazonais, tendências de longo prazo; e podem ser decompostas em quatro componentes, sendo eles: tendência (T), ciclo, sazonalidade (S) e erro (ε) (Hewamalage *et al.*, 2021).

Nos trabalhos analisados de análise e comparação de métodos estatísticos, os mesmos são aplicados a padrões lineares ou não lineares, e análise do seu é realizado por meio de métricas estatísticas dentre as quais destacam-se: MAE/MSE/RMSE, para noção absoluta do erro; métricas MAPE/sMAPE, para comparar modelos em escalas diferentes e a função de desejabilidade (*Desirability*) que atribui uma pontuação a um conjunto de respostas. Dentre os trabalhos destacam-se: Bao *et al.*, 2024; Milenković *et al.*, 2024; Alves *et al.*, 2023; Iftikhar *et al.* 2024, Hewamalage *et al.*, 2021; entre outros.

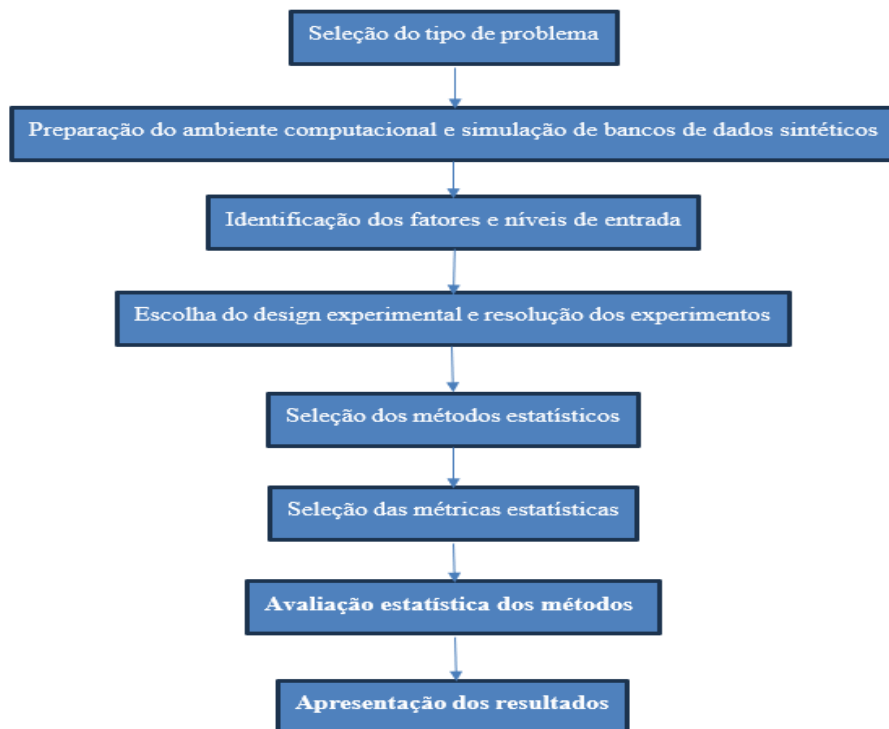
2.2 Metodologia

A metodologia, observada na Figura 1, foi desenvolvida por meio de um código em Python, que contou com o auxílio da inteligência artificial *DeepSeek Chat* (2024), para otimização de funções, sugestões de implementação e resolução de erros. Os bancos de dados sintéticos foram gerados por meio da biblioteca *TimeSynth*, desenvolvida por Maat *et al.* (2017), que gera séries temporais regulares e irregulares, combinando diferentes sinais. O código final foi revisado e validado pela autora, por meio de testes estatísticos.

Para esta análise, foram selecionados dez métodos estatísticos com características distintas: Modelos clássicos ARIMA e SARIMA que são métodos estatísticos lineares baseados em autocorrelação; modelos ARIMA *Boost* e SARIMA *Boost* que combinam o poder do ARIMA/SARIMA para capturar a dependência temporal linear com o *boosting* para modelar os resíduos não lineares; modelos *LightGBM* e *XGBoost* utilizados em competições de *machine learning*, que podem capturar relações não lineares e padrões complexos; métodos de redes neurais LSTM e RNN que capturam dependências temporais e padrões complexos; método *Prophet* projetado para séries com forte sazonalidade, tendências não lineares e outliers; e método SVR usado para modelar problemas não lineares e alta dimensionalidade.

O estudo comparativo de diferentes métodos de previsão de séries temporais foi realizado por meio de um experimento fatorial, baseado nos fatores e níveis apresentados na Figura 2.

Figura 1 – Metodologia estruturada de comparação de métodos estatísticos



Fonte: A autora

Figura 2 – Fatores e níveis analisados no estudo de séries temporais

Fator	Nível alto (+)	Nível baixo (-)
Tamanho do conjunto de dados (<i>n-samples</i>)	500	100
Correlação (<i>correlation</i>)	<i>high</i>	<i>low</i>
Número de variáveis (<i>n_features</i>)	5	3
Autocorrelação (<i>autocorrelation</i>)	<i>strong</i>	<i>weak</i>
Relação (<i>relation</i>)	<i>non-linear</i>	<i>linear</i>

Fonte: A autora

3. RESULTADOS E DISCUSSÃO

Para o teste do código desenvolvido, foram realizados experimentos com cinco fatores e dois níveis (2^5), com resolução completa. O design experimental pode ser visualizado na Figura 3.

A análise dos melhores resultados dos métodos por cada métrica pode ser observada na Figura 4.

Figura 3 – Design experimental completo para 5 fatores– Série temporal

Plano de experimentos DOE gerado:

	n_samples	correlation	n_features	autocorrelation	relation	Execução
0	100	low	3	weak	linear	3
1	100	low	3	weak	non-linear	6
2	100	low	3	strong	linear	4
3	100	low	3	strong	non-linear	8
4	100	low	5	weak	linear	3
5	100	low	5	weak	non-linear	4
6	100	low	5	strong	linear	3
7	100	low	5	strong	non-linear	8
8	100	high	3	weak	linear	7
9	100	high	3	weak	non-linear	5
10	100	high	3	strong	linear	3
11	100	high	3	strong	non-linear	4
12	100	high	5	weak	linear	8
13	100	high	5	weak	non-linear	3
14	100	high	5	strong	linear	5
15	100	high	5	strong	non-linear	5
16	500	low	3	weak	linear	3
17	500	low	3	weak	non-linear	4
18	500	low	3	strong	linear	3
19	500	low	3	strong	non-linear	4
20	500	low	5	weak	linear	5
21	500	low	5	weak	non-linear	3
22	500	low	5	strong	linear	8
23	500	low	5	strong	non-linear	5
24	500	high	3	weak	linear	7
25	500	high	3	weak	non-linear	9
26	500	high	3	strong	linear	3
27	500	high	3	strong	non-linear	9
28	500	high	5	weak	linear	4
29	500	high	5	weak	non-linear	6
30	500	high	5	strong	linear	2
31	500	high	5	strong	non-linear	5

Fonte: A autora.

Figura 4 – Melhores resultados das métricas estatísticas por experimento

Exp.	Melhor Modelo	MAE	MSE	RMSE	MAPE	sMAPE	Desirability
1	SARIMA_Boost	0,07	-0,63	0,4	0,32	0,51	0,18 (Prophet)
2	SARIMA	-0	0,6	0,9	0,49	0,38	0,01 (LightGBM)
3	SARIMA_Boost	0,27	1,19	0,48	0,11	0,1	0,32 (Prophet)
4	SVR	0,18	0,05	0,22	1,91	75,91	0,01 (ARIMA_Boost)
5	LightGBM	1,96	4,56	2,14	0,21	24,18	0,15 (ARIMA_Boost)
6	LightGBM	0,39	0,23	0,48	1,46	62,23	0,02 (LightGBM)
7	SARIMA_Boost	3,12	19,2	0	0	0,93	0,37 (SARIMA_Boost)
8	SVR	0,34	0,16	0,4	4,94	67,37	0,04 (SARIMA_Boost)
9	Prophet	0,09	0,01	0,11	0,01	0,98	0,51 (Prophet)
10	SVR	0,14	0,03	0,18	1,2	39,35	0,02 (SVR)
11	Prophet	0,02	0	0,02	0	0,2	0,83 (Prophet)
12	SVR	0,13	0,03	0,17	1,97	42,73	0,02 (SVR)
13	Prophet	0,13	0,02	0,14	0,01	1,39	0,42 (Prophet)
14	LightGBM	0,35	0,21	0,46	4,4	64,14	0,01 (LightGBM)
15	Prophet	0,01	0	0,02	0	0,15	0,87 (Prophet)
16	SVR	0,1	0,02	0,13	1,62	32,47	0,03 (SVR)
17	LightGBM	1,27	2,17	1,47	0,13	14,78	0,06 (LightGBM)
18	LightGBM	0,49	0,39	0,62	2,57	77,29	0,01 (LightGBM)
19	LightGBM	1,33	2,09	1,44	0,14	15,63	0,06 (LightGBM)
20	SVR	0,1	0,01	0,12	0,69	25,13	0,04 (SVR)
21	LightGBM	1,29	2,13	1,46	0,14	14,92	0,06 (LightGBM)
22	LightGBM	0,68	0,66	0,82	8,23	99,61	0,01 (LightGBM)
23	Prophet	0,06	0,01	0,08	0,01	0,71	0,58 (Prophet)
24	SVR	0,1	0,02	0,13	0,8	26,39	0,04 (SVR)
25	Prophet	0,09	0,01	0,11	0,01	0,96	0,51 (Prophet)
26	SVR	0,11	0,02	0,14	0,73	30,66	0,03 (SVR)
27	Prophet	0,04	0	0,04	0	0,4	0,71 (Prophet)
28	SVR	0,1	0,01	0,11	4,42	31,2	0,031 (SVR)
29	Prophet	0,07	0,01	0,09	0,01	0,8	0,56 (Prophet)
30	LightGBM	0,46	0,3	0,55	4,02	71,14	0,01 (LightGBM)
31	Prophet	0,02	0	0,02	0	0,2	0,83 (Prophet)
32	SVR	0,1	0,02	0,13	0,62	29,07	0,03 (SVR)

Fonte: A autora

Os métodos com melhores resultados nas métricas em estudo foram SVR (10 experimentos), *Prophet* (9 experimentos), *LightGBM* (9 experimentos), *SARIMA Boost* (3 experimentos) e *SARIMA* (1 experimento). Na métrica de *desirability* os melhores métodos foram *Prophet*, *SARIMA* e *ARIMA Boost*.

O design de experimentos (DOE) implementado permitiu avaliar sistematicamente a influência dos fatores no desempenho dos modelos de previsão, e a análise de variância (ANOVA) permitiu quantificar os efeitos principais e interações entre fatores.

Os testes t pareado, teste Wilcoxon e teste Friedman foram aplicados para identificar diferenças estatísticas entre os resultados das métricas de cada método, com nível de significância $p\text{-value} < 0,05$. Os resultados da ANOVA e testes estatísticos são observados na Figura 5.

Figura 5 – Resultados da influência e interação dos fatores por métrica e testes estatísticos

Fator/Métrica	MAE	MSE	RMSE	MAPE	sMAPE	Desirability
Relation				F = 20,46, p = 0,00	F = 41,09, p = 0,000	F = 34,28, p = 0,00
Correlation	F = 7,58, p = 0,01	F = 4,62, p = 0,05	F = 11,50, p = 0,00			F = 5,05, p = 0,03
Autocorrelation			F = 5,55, p = 0,02			F = 4,72, p = 0,04
Correlation x Relation	F = 7,02, p = 0,01		F = 4,36, p = 0,05			F = 22,55, p = 0,00
Correlation x autocorrelation			F = 4,36, p = 0,05			
N_samples x autocorrelation					F = 4,54, p = 0,04	
N_samples x n_features						F = 10,81, p = 0,00
N_features x autocorrelation			F = 5,38, p = 0,03		F = 4,45, p = 0,05	
Autocorrelation x relation					F = 17,19, p = 0,00	F = 13,23, p = 0,00
N_samples x correlation x n_features	F = 18,16, p = 0,00				F = 5,56, p = 0,03	
N_samples x autocorrelation x relation					F = 7,04, p = 0,02	
N_samples x n_features x relation	F = 9,29, p = 0,01					
Correlation x n_features x autocorrelation x relation			F = 7,68, p = 0,01			
N_samples x correlation x autocorrelation x relation					F = 5,42, p = 0,03	
Teste T pareado	p-value: 0.00	p-value: 0.00	p-value: 0.00	p-value: 0.00	p-value: 0.00	p-value: 0.00
Teste Wilcoxon	p-value: 0.00	p-value: 0.00	p-value: 0.00	p-value: 0.00	p-value: 0.00	p-value: 0.00
Teste Friedman	p-value: 0.00	p-value: 0.00	p-value: 0.00	p-value: 0.00	p-value: 0.00	p-value: 0.00

Fonte: A autora.

Os fatores *relation*, *correlation* e *autocorrelation* foram os fatores de maior impacto nas métricas. A autocorrelation destacou-se nas interações de dois fatores. O fator *n_samples* se destacou nas interações de três fatores, enquanto *relation*, *autocorrelation* e *correlation* foram destaque nas interações de quatro fatores.

Nos testes paramétricos e não paramétricos foram identificadas diferenças estatisticamente significativas entre todos os modelos.

4. CONSIDERAÇÕES FINAIS

A metodologia estruturada desenvolvida neste trabalho permitiu uma avaliação sistemática e reproduzível entre dez diferentes métodos estatísticos por meio de diferentes métricas estatísticas, possibilitando identificar padrões de desempenho sob diversas condições experimentais em uma simulação computacional de dados sintéticos, servindo como suporte à tomada de decisão. O design experimental (DOE) se mostrou eficiente para: gerar bancos de dados sintéticos, permitir análises sistemáticas de métodos estatísticos, generalizar os cenários, quantificar a influência dos fatores analisados e auxiliar na identificação do melhor método a ser utilizado em cada experimento.

A contribuição científica desta pesquisa está centrada na redução do tempo de processamento e análise de dados, a minimização de erros na comparação e escolha de métodos estatísticos e a disponibilização de uma estrutura adaptável para uso de estudantes e pesquisadores, com ampla aplicabilidade, podendo ser utilizada em diversas áreas do conhecimento e adaptada a diferentes tipos de problemas estatísticos.

Para pesquisas futuras sugere-se utilizar ferramentas de IA para automatizar a estrutura, explorar outros designs de DOE - como Taguchi e Plackett-Burman - e técnicas de otimização, incluir métricas adicionais à análise e características externas ao banco de dados analisados.

AGRADECIMENTOS

Agradeço ao apoio financeiro fornecido pela CAPES e a Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG).

REFERÊNCIAS

Alves, E. P.; Oliveira, J. F. L. de; Madeiro, F.; Marinho, M. H. d. N. *A nonlinear optimized hybrid system for energy consumption forecasting from smart meters. Learning and Nonlinear Models*, v. 20, p. 17-30, 2022.

Bao, X.; Wang, D.; Wu, T.; Yu, Z. *Probabilistic Time Series Forecasting Based on the Principle of Justifiable Granularity. 9th International Conference on Intelligent Computing and Signal Processing (ICSP)*, Xian, China, 2024, p. 517-522. DOI: 10.1109/ICSP62122.2024.10743812.

DEEPSEEK. *DeepSeek Chat (versão DeepSeek-V3)*, 2024. Disponível em: <https://www.deepseek.com>. Acesso em: jan. 2025.

Hewamalage, H.; Bergmeir, C.; Bandara, K. *Recurrent neural networks for time series forecasting: Current status and future directions. International Journal Forecasting*, v. 37, p. 388-427, 2021. DOI: 10.1016/j.ijforecast.2020.06.008. Acesso em: mai. 2025.

Iftikhar, H.; S., Mancha Gonzales; Zywiólek, J.; López-Gonzales, J. L. *Electricity Demand Forecasting Using a Novel Time Series Ensemble Technique. IEEE Access*, v. 12, p. 88963-88975, 2024. DOI: 10.1109/ACCESS.2024.3419551.

Maat, J. R.; Malali, A.; Protopapas, P. *TimeSynth: A Multipurpose Library for Synthetic Time Series in Python*. 2017. Disponível em: <http://github.com/TimeSynth/TimeSynth>. Acesso em: jan. 2025.

Milenković, M., Gligorić, M., Bojović, N., & Gligorić, Z. *A comparison between ARIMA, LSTM, ARIMA-LSTM and SSA for cross-border rail freight traffic forecasting: the case of Alpine-Western Balkan Rail Freight Corridor. Transportation Planning and Technology*, v. 47, n. 1, p. 89-112, 2023. DOI: 10.1080/03081060.2023.2245389.

Morettin, P.; Toloi, C. *Análise de séries temporais*. 2. ed. São Paulo: E. Blucher, 2006. 564 p.