Trabalho de Conclusão de Curso apresentado para obtenção do título
de especialista em Data Science e Analytics – 2025

Modelagem e previsão do IBOVESPA com variáveis macroeconômicas

João Paulo Fumes Beneton; Daniel Alvarez Firmino

Modelagem e previsão do IBOVESPA com variáveis macroeconômicas

Resumo (ou Sumário Executivo)

Este estudo examinou a previsibilidade mensal do Ibovespa para apoiar gestão de risco e alocação tática. Comparou-se a capacidade preditiva de um ARIMA univariado (apenas o preço do índice) e de ARIMAX com Selic, IPCA e USD/BRL como variáveis exógenas, no período 2000–2025. Seguiu-se um protocolo reprodutível e sem vazamento: diagnosticou-se estacionariedade por ADF, diferenciou-se o Ibovespa (ΔIBOV), padronizou-se apenas no treino, selecionaram-se ordens (p, d, q) por AICc e definiram-se defasagens das exógenas por validação interna no treino (lag comum e por variável). Posteriormente, avaliou-se fora da amostra por rolling 1-step em hold-out de 36 meses, com re-estimação a cada passo, comparando modelos por MAE/RMSE e pelo teste de Diebold–Mariano; conduziram-se ainda análises de sensibilidade via remoção de variáveis (FULL/DROP/ONLY) e permutation importance em blocos. Os resultados mostraram que o ARIMA (3,0,4) univariado permaneceu como benchmark robusto, enquanto especificações ARIMAX, mesmo com defasagens selecionadas, apresentaram ganhos apenas marginais e não estatisticamente significativos no hold-out; sinais in-sample, como o do IPCA defasado, não se sustentaram fora da amostra. Concluiu-se que, no horizonte mensal, a dinâmica própria do índice capturou a maior parte do sinal útil e que a principal contribuição do estudo residiu no protocolo de validação transparente e replicável para comparação de modelos preditivos do Ibovespa.
Palavras-chave: ARIMA; ARIMAX; variáveis exógenas; validação fora da amostra; Diebold–Mariano.

Introdução

O interesse em prever valores futuros de ativos financeiros acompanha os mercados acionários desde sua origem. Por algum tempo, prevaleceu a ideia de que tais previsões seriam relativamente simples, sobretudo a partir de características da própria ação. A literatura, contudo, demonstra que a previsão de preços é um problema complexo (Satchell, 2007). Frequentemente, essa complexidade decorre da escolha das variáveis utilizadas para antecipar o prêmio de risco acionário. Nesse contexto, diversos estudos buscaram identificar variáveis com elevada capacidade preditiva. Todavia, há trabalhos que documentam algum poder preditivo — sobretudo de variáveis macroeconômicas — na formulação de modelos; por outro lado, existem evidências de que as previsões baseadas nessas variáveis são instáveis ao longo do tempo e sensíveis à janela amostral, tanto dentro quanto fora da amostra (Goyal e Welch, 2008).

No contexto brasileiro, a heterogeneidade de resultados também é observada em estudos que tomam o Ibovespa como variável a ser prevista. O índice reflete a variação de preços de uma carteira selecionada de ativos listados na B3, influenciada por fatores internos e externos (Ross; Westerfield; Jaffe; Lamb, 2015). Nessa perspectiva, parte da literatura não detecta causalidade estatisticamente significativa entre indicadores macroeconômicos — taxa de juros (Selic), taxa de câmbio (PTAX) e inflação (IPCA) — e o principal índice brasileiro (REAd, 2006). Em contraste, outros trabalhos, via ARDL/VEC em janelas específicas, encontram relações de longo prazo e efeitos assimétricos/heterogêneos — em geral, choques de juros e câmbio reduzem o Ibovespa, ao passo que o IPCA depende do horizonte (curto vs. longo prazo) e do regime (OJBM, 2024). Em síntese, a evidência permanece mista e fortemente sensível à janela amostral e ao período considerado — padrão enfatizado tanto em estudos de mercados emergentes (Çakıcı, Yan & Zaremba, 2024) quanto no próprio Ibovespa (Arévalo, Gómez & Villamil, 2023).

Considerando esse contexto, este trabalho propõe uma investigação aplicada, com dados mensais (2000–2025), sobre o Ibovespa e três variáveis macroeconômicas clássicas — Selic, IPCA e câmbio (USD/BRL) —, frequentemente empregadas na literatura. Buscando transparência e comparabilidade, adota-se uma metodologia quantitativa que compreende: (i) diagnóstico de estacionariedade (Augmented Dickey–Fuller [ADF]; análise da função de autocorrelação [ACF)] e diferenciação quando necessário; (ii) estimação de ARIMA (univariado, benchmark principal) e ARIMAX (com preditores exógenos); (iii) seleção de ordens (p,d,q) por Akaike Information Criterion corrected [AICc] e verificação de autocorrelação residual (Ljung–Box); (iv) definição empírica de defasagens para as exógenas (lag comum e lags por variável) via validação interna no treino; e (v) avaliação fora da amostra

por rolling 1-step em hold-out de 36 meses com janela de treino expansiva (expanding window), com métricas MAE/RMSE e teste de Diebold–Mariano para comparação formal de acurácia. Como extensões diagnósticas, realizam-se ablation tests (FULL/DROP/ONLY) e permutation importance em blocos no teste. Reconhece-se, como limitação, a heterocedasticidade típica de séries financeiras (não modelada aqui via GARCH) e a possibilidade de mudanças de regime; sazonalidade mensal explícita não foi modelada por parcimônia (Clark; West, 2007; Goyal; Welch, 2008).

Com base em dados mensais, os resultados não indicam ganhos robustos de previsão ao incluir exógenas defasadas no ARIMAX frente ao ARIMA univariado (benchmark). Em particular, o ARIMA (3,0,4) mostrou-se muito competitivo na avaliação rolling 1-step com janela de treino expansiva no hold-out de 36 meses, enquanto especificações ARIMAX (1,0,1) com lags por variável (5,3,5) e (4,3,3) apresentaram, fora da amostra, ganhos marginais não significativos segundo o teste de Diebold–Mariano. Em contraste, análises in-sample sugerem sinal informativo para alguns preditores (por exemplo, IPCA defasado), mas a generalização fora da amostra é fraca. No agregado, ablation e permutation importance apontam pequena contribuição da SELIC defasada e impacto limitado ou adverso de IPCA e USD/BRL na janela estudada, em linha com a literatura que documenta previsibilidade mensal modesta e instável.

Diante do exposto, o objetivo deste Trabalho de Conclusão de Curso é mensurar e comparar a capacidade preditiva de modelos ARIMA e ARIMAX para o Ibovespa em horizonte mensal de 1 passo à frente, utilizando Selic, IPCA e USD/BRL como candidatas a regressoras exógenas. A contribuição central é metodológica e empírica: fornecer um pipeline reprodutível e sem vazamento (seleção de ordens e defasagens por validação interna, avaliação out of sample [OOS] por rolling 1-step e comparação formal por Diebold–Mariano) aplicado ao período 2000–2025, cobrindo choques relevantes (2008, 2015–2016, 2020 e 2021–2023). Os achados reforçam que, para previsões mensais de curto prazo, ARIMAX parcimonioso não supera consistentemente o ARIMA univariado, embora possa empatar ou melhorar marginalmente em janelas específicas — resultado coerente com a evidência internacional recente para mercados emergentes.

Metodologia ou Material e Métodos

1) Variáveis

O período 2000–2025 no Brasil foi marcado por mudanças estruturais em seu cenário — como a adoção do regime de metas para a inflação (1999), o câmbio flutuante (1999) e a Lei de Responsabilidade Fiscal (2000) — e por choques relevantes — como a crise global de 2008, a recessão de 2015–2016 e a COVID-19 em 2020. Nesse contexto, a B3 consolidou-se como a principal bolsa da América Latina, com expansão do valor de mercado e do volume negociado. Esses marcos motivam a investigação, neste estudo, de como variáveis macroeconômicas se relacionam com o comportamento do Ibovespa.

Neste contexto, o Ibovespa — criado em 1968 — é o principal índice de referência do mercado acionário brasileiro. Calculado a partir de uma carteira teórica das ações mais negociadas e líquidas da B3, o índice reflete tanto variações de preços quanto o reinvestimento de proventos das empresas componentes. Diante disso, o alvo preditivo deste trabalho é a variação mensal do Ibovespa em primeira diferença (ΔIBOVESPA), que constitui uma proxy prática do comportamento do mercado acionário brasileiro em horizonte de curto prazo, dada a relevância do índice. Tal objetivo já foi explorado por diversos autores como elencado no decorrer do presente estudo.

Para prever o comportamento do Ibovespa, este estudo considera três preditores macroeconômicos recorrentes na literatura: inflação (IPCA), taxa de juros (SELIC) e taxa de câmbio (USD/BRL – PTAX). O objetivo é avaliar seu poder preditivo sobre ΔIBOVESPA, reconhecendo que, na evidência para o Brasil, choques em juros e câmbio costumam pressionar negativamente o índice e que os efeitos da inflação podem ser heterogêneos no tempo. Exemplos incluem com VECM (resposta negativa a câmbio e juros e positiva a IPCA); Montes e Tiberto (2011), via OLS para 2000–2010 (associação negativa para câmbio e juros); Franzen (2009) (desvalorizações cambiais elevam a aversão ao risco, pressionando o índice); e Nunes et al. (2005) (relação positiva entre inflação e atividade, compatível com a Curva de Phillips). Apesar das divergências entre trabalhos, a recorrência dessas variáveis justifica sua inclusão e reavaliação no período analisado.

Em contraste com os resultados acima, parte da literatura não encontra relação estável de longo prazo entre o Ibovespa e o câmbio. Estudos como The Dynamic Relationship between Stock Prices and Exchange Rates: Evidence for Brazil apontam que não há cointegração entre o Ibovespa e a taxa de câmbio (dólar). De modo semelhante, (REAd, 2006) (1994–2005) concluiu que as variáveis macroeconômicas clássicas — inflação (IPCA), juros (SELIC) e câmbio (PTAX) — não apresentaram previsibilidade robusta sobre o Ibovespa. A

análise em voga demonstrou que os choques próprios do índice explicaram a maior parte do erro de previsão dentro de uma janela de até 24 meses, indicando que o comportamento do Ibovespa é, em grande parte, autocontido. Esses achados convergem com o ceticismo de Goyal e Welch (2008) quanto ao baixo desempenho fora da amostra de preditores amplamente utilizados. À luz dessa divergência, este estudo reexamina essas relações no recorte 2000–2025, período que inclui choques e possíveis mudanças de regime capazes de afetar a evidência empírica. Para operacionalizar a análise, utilizam-se séries mensais alinhadas no calendário: o Ibovespa em fechamento do último dia útil (obtido via base pública, p.ex., Yahoo Finance) e, como variáveis explicativas, USD/BRL – PTAX (fim de mês), IPCA (variação mensal) e SELIC (SGS/BCB). As exógenas serão defasadas (lag comum e/ou lags por variável) e essas defasagens são escolhidas empiricamente por validação interna no treino; quando aplicável, as séries são padronizadas (z-score) com ajuste do scaler exclusivamente no conjunto de treino (antivazamentos).

A figura apresentada abaixo (Figura 1) representa a evolução histórica mensal, no período de janeiro de 2000 a julho de 2025, das quatro variáveis analisadas neste estudo — Ibovespa (fechamento do último dia útil), SELIC, IPCA (m/m) e USD/BRL (PTAX, fim de mês). No Ibovespa, observa-se uma trajetória ascendente de longo prazo, com episódios de volatilidade em momentos de crise — global, como a pandemia de COVID-19 em 2020, e nacional, como a recessão de 2015–2016. Na SELIC, notam-se ciclos de aperto e afrouxamento, com destaque para a elevação de 2021–2023; no IPCA, picos em anos de choques de preços e período de estabilidade inflacionária; na taxa de câmbio, fases de apreciação e depreciação do real. Em conjunto, esses padrões sugerem não-estacionariedade em nível e possíveis quebras de regime, motivando as transformações e os testes econométricos apresentados no tópico a seguir.

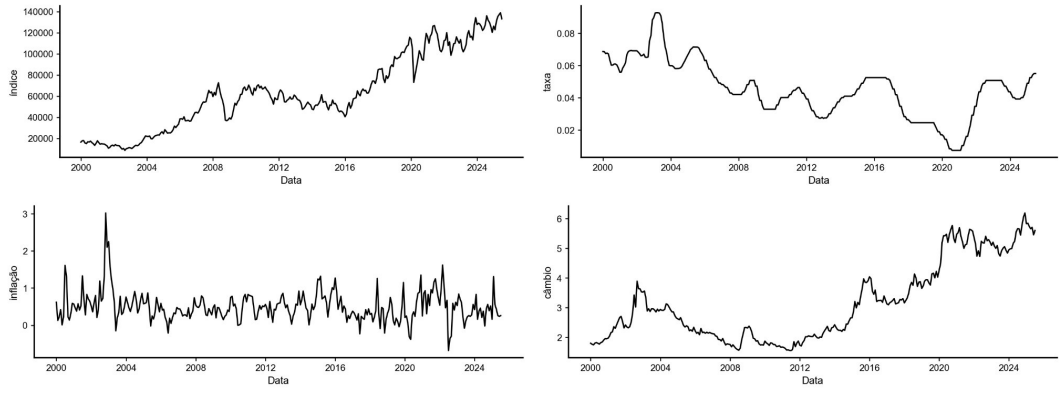

Figura 1. Evolução histórica mensal, no período de janeiro de 2000 a julho de 2025
Fonte: Resultados originais da pesquisa

Por inspeção visual, não se observa padrão sazonal mensal robusto; essa hipótese também será testada formalmente nas seções subsequentes. Na sequência, gerou-se a tabela descritiva (Tabela 1) das variáveis utilizadas. Em nível, IBOVESPA e USD/BRL exibem maior amplitude/volatilidade (desvios-padrão elevados), compatível com os choques de 2008, 2015–2016 e 2020; SELIC e IPCA são relativamente mais estáveis.

Tabela 1. Análise Descritiva

	IBOVESPA	SELIC	IPCA	DÓLAR
count	307.0	307.0	307.0	307.0
mean	62529.91	0.05	0.5	3.15
std	35639.1	0.02	0.39	1.32
mix	8623.0	0.01	-0.68	1.56
max	138855.0	0.09	3.02	6.19

Fonte: Resultados originais da pesquisa

2) Estacionariedade, Sazonalidade e Padronização

À luz da evolução histórica das quatro variáveis apresentadas na figura do tópico anterior, observa-se que o IPCA (m/m) demonstra relativa estabilidade, enquanto o Ibovespa e o USD/BRL exibem tendência e sinais de quebras, ao passo que a SELIC é marcadamente cíclica. Tais comportamentos configuram não-estacionariedade, pois as médias, variâncias e autocorrelações variam ao longo do tempo. Séries com tendência e/ou sazonalidade são exemplos típicos de séries não estacionárias, dado que ocorre alteração no nível esperado desses componentes ao longo do período analisado. Esse padrão é problemático, pois prejudica o ajuste/diagnóstico e a estabilidade preditiva em modelos ARIMA/ARIMAX. Nesta seção, verifica-se formalmente a estacionariedade e, quando necessário, aplica-se transformações, tomando como referência a definição usual do conceito (propriedades constantes no tempo; Hyndman & Athanasopoulos (2021)). Na pipeline deste presente estudo, o alvo é a variação mensal do Ibovespa (ΔIBOVESPA), e as exógenas entram defasadas; quando padronizadas (z-score), o ajuste do scaler é feito apenas no conjunto de treino (evitando vazamento). Por inspeção visual, não se observou sazonalidade mensal robusta; por parcimônia, não houve a modelagem componente sazonal.

Nesta perspectiva, como primeira verificação, inspeciona-se, como apontado no parágrafo anterior, os gráficos das séries (bem como ACF/PACF): séries estacionárias tendem a oscilar em torno de um nível constante; no nosso caso, o IPCA (m/m) sugere esse padrão, enquanto Ibovespa, SELIC e USD/BRL não. Complementarmente, aplica-se o teste Augmented Dickey–Fuller [ADF], cuja hipótese nula é a presença de raiz unitária (não-estacionariedade), e reportam-se os p-valores na Tabela 2 abaixo.

Tabela 1. Teste Augmented Dickey–Fuller

VARIÁVEL	ADF-ESTATÍSTICA	P-VALOR	REJEITA H0
IPCA	-8.64	0.00	SIM
SELIC	-2.32	0.17	NÃO
DÓLAR	-0.65	0.86	NÃO
IBOVESPA	-0.14	0.94	NÃO

Fonte: Resultados originais da pesquisa

De acordo com os resultados do teste ADF aplicado às séries em nível (com constante), apenas o IPCA (m/m) rejeitou a hipótese nula ao nível de 5%, indicando ausência de raiz unitária nessa série. Para as demais variáveis, a hipótese nula não foi rejeitada, caracterizando não-estacionariedade em nível nas especificações adotadas. A função de autocorrelação (ACF), em nível, reforça esse diagnóstico: no IPCA, as autocorrelações caem rapidamente para zero; nas demais séries, o decréscimo é lento, padrão típico de processos com raiz unitária. A Figura 2 apresenta as ACFs correspondentes.

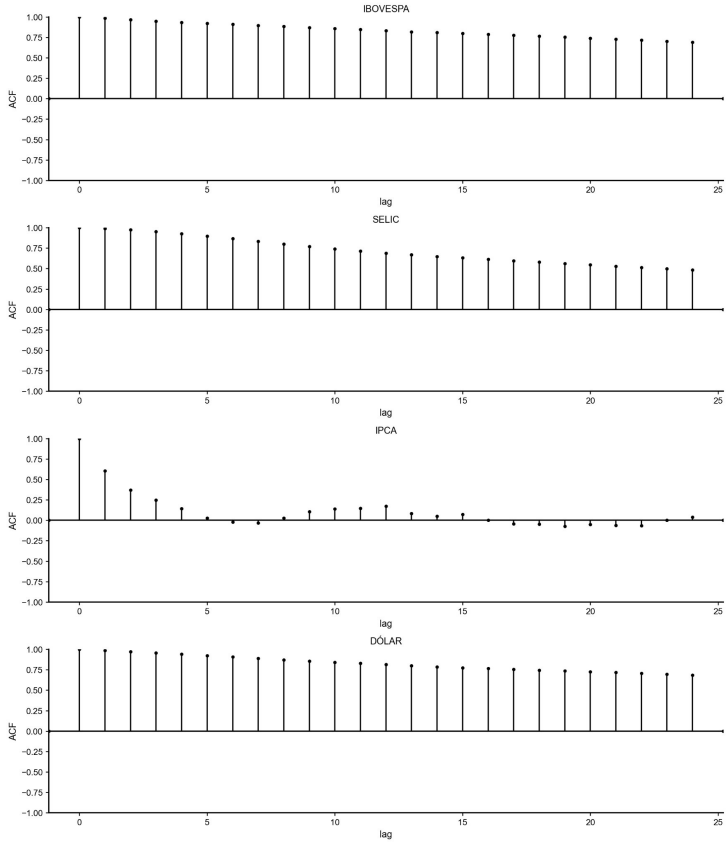

Figura 2. Função de autocorrelação para as variáveis exógenas
Fonte: Resultados originais da pesquisa

A análise da função de autocorrelação (ACF) em nível corrobora o teste Augmented Dickey–Fuller (ADF): Ibovespa, SELIC e USD/BRL exibem autocorrelações que decaem lentamente e não rejeitam a hipótese de raiz unitária nas especificações usuais (com constante e tendência), caracterizando não-estacionariedade em nível. Já o IPCA (m/m) oscila em torno de um patamar relativamente estável, compatível com estacionariedade nas especificações adotadas.

Diante desse diagnóstico, e seguindo a recomendação de Hyndman & Athanasopoulos (2021), adotamos a primeira diferença para Ibovespa, SELIC e USD/BRL, de modo a estabilizar a média e reduzir a autocorrelação de longo prazo. Em termos práticos, cada série é transformada em Delta [$Delta\ y_t = y_t - y_{\{t-1\}}$]. O IPCA é mantido como variação mensal (m/m), pois já se apresenta estável. A partir deste ponto, o alvo de previsão do estudo é ΔIBOV, e as exógenas usadas no ARIMAX são ΔSELIC, IPCA (m/m) e ΔUSD/BRL, possivelmente defasadas.

Após a diferenciação, aplica-se, novamente, o ADF (Tabela 2) às séries IBOVESPA, ΔSELIC e ΔUSD/BRL e obtivemos p-valores < 5%, indicando ausência de raiz unitária pós-transformação; no IPCA, o resultado permanece consistente com estacionariedade. As ACFs das séries em diferença também passam a decair rapidamente para zero, reforçando o diagnóstico.

Tabela 2. Teste Augmented Dickey–Fuller após diferenciação

VARIÁVEL	ADF-ESTATÍSTICA	P-VALOR	REJEITA H0
SELIC	-11.25	0.00	SIM
DÓLAR	-5.59	0.00	SIM
IBOVESPA	-11.19	0.00	SIM

Fonte: Resultados originais da pesquisa

Analisada a estacionariedade e, de acordo com os gráficos de ACF (Figura 2) em nível e nas diferenças (lags 12 e 24), não se observam picos pronunciados ou padrões repetitivos que caracterizem sazonalidade mensal nas séries. Em Ibovespa e USD/BRL, as autocorrelações decaem lenta e gradualmente — sem estrutura periódica; a SELIC exibe comportamento cíclico, mas não estritamente mensal; e, no IPCA (m/m), os lags sazonais não se destacam. Diante disso, opta-se por não incluir componente sazonal (SARIMA), mantendo especificações não sazonais. Como checagem, os testes de Ljung–Box aplicados aos resíduos (Q=12) não indicaram autocorrelação sazonal remanescente.

Posteriormente, realizados os testes que viabilizam a aplicação de modelos de série temporal, implementa-se a padronização (z-score) das variáveis exógenas — SELIC, IPCA e USD/BRL —, pela qual cada x é reescalada segundo $z=(x-\mu)/\sigma$. Os parâmetros μ e σ são estimados

apenas no conjunto de treino a cada passo do rolling 1-step (janela expansiva) e, em seguida, aplicados à observação do mês previsto, evitando vazamento de informação. Esse procedimento coloca os preditores na mesma escala, melhora a estabilidade numérica da estimação e facilita a leitura relativa dos coeficientes (em unidades padronizadas), sem padronizar o alvo ΔIBOV. Ademais, séries financeiras podem apresentar heterocedasticidade. Embora este estudo não modele GARCH, essa limitação é reconhecida.

Por fim, estima-se a matriz de correlação entre as exógenas (SELIC, IPCA, USD/BRL), calculada sobre as séries na forma usada na modelagem — isto é, após as transformações e defasagens (lag comum ou por variável) e na amostra comum alinhada. A Figura 3 apresenta o heatmap correspondente.

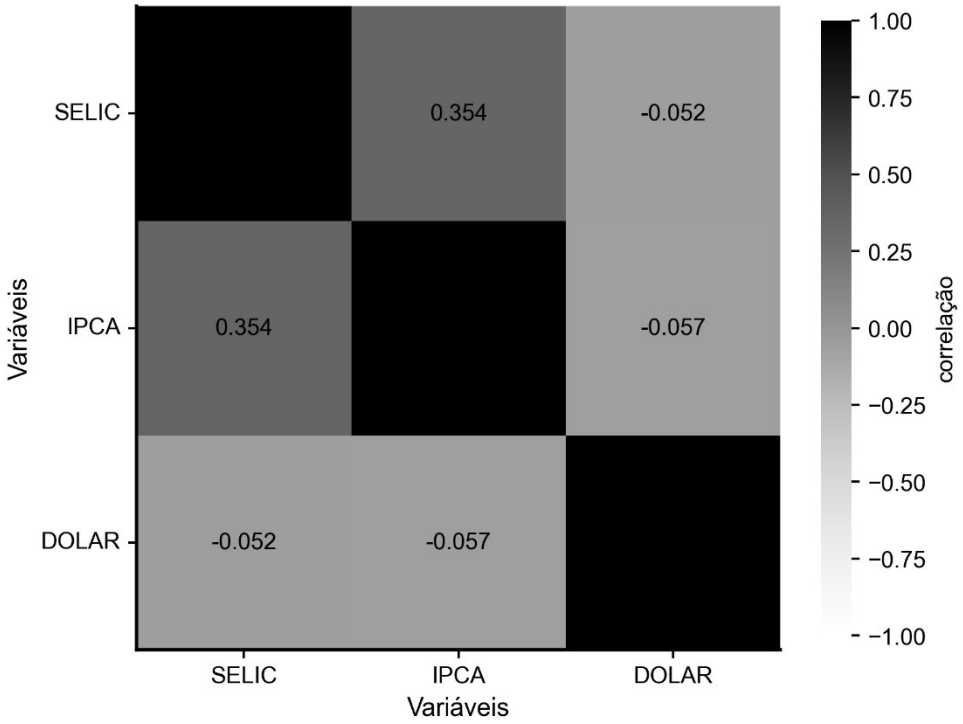

Figura 3. Heatmap (correlações entre as variáveis exógenas)

Fonte: Resultados originais da pesquisa

A partir da matriz de correlação, não se identificam evidências de multicolinearidade entre os preditores, é benéfico para a estabilidade do ARIMAX, uma vez que há correlações fracas a, no máximo, moderadas entre os preditores — ρ (SELIC, IPCA) ≈ 0,35; ρ (SELIC, USD/BRL) ≈ −0,05; ρ (IPCA, USD/BRL) ≈ 0,06.

3) Especificação e lags das exógenas: AICc (ordens) + rolling 1-step (validação)

Uma vez concluídos os diagnósticos, adotam-se como modelos de previsão o Autoregressive Integrated Moving Average [ARIMA] e o Auto- Regressive Moving Average Model including Exogenous covariates [ARIMAX] — SELIC, IPCA e USD/BRL, como

exógenas — para a série-alvo (Ibovespa). No ARIMA, os componentes são: (a) AutoRegressive (AR), que modela o valor atual como combinação linear de valores passados da própria série; (b) Integrated (I), que aplica diferenciações de ordem *d* para induzir estacionariedade; e (c) Moving Average (MA), que modela o erro atual como combinação linear de erros de previsão passados (resíduos defasados). No ARIMAX, além desses componentes, incluem-se regressoras exógenas (possivelmente defasadas) para capturar efeitos contemporâneos ou dinâmicos sobre o Ibovespa. As respectivas fórmulas do modelo ARIMA e ARIMAX, encontram-se abaixo (Hyndman & Athanasopoulos, 2021).

a) $$\Phi(B)\ \Delta_{x_t}^{d} y = c + \Theta(B)\ \varepsilon_{x,t} \qquad (1)$$

b) $$y_{x_t} = \beta_0 + \sum_{j=0}^{k} \gamma_j\, x_{t-j} + n_{x,t}, n_t \sim ARIMA(p,d,q) \quad (2)$$

Em que: y_{x_t} é a série-alvo no tempo t; x_t é a variável exógena; Δ^d é o operador de diferenciação de ordem d; B é o operador de defasagem $B\,y_t = y_{t-1}$; $\Phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ e $\Theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$ são, respectivamente, os polinômios AR (ordem p) e MA (ordem q); c é a constante; $\varepsilon_{x,t}$ é a inovação (ruído branco); β_0 são o intercepto; γ_j é o coeficiente da defasagem j de x e $\sum_{j=0}^{k} \gamma_j\, x_{t-j}$ representa o componente exógeno até a defasagem k (estende-se para múltiplas exógenas somando termos análogos); $n_{x,t} \sim ARIMA(p,d,q)$ é o erro que segue ARIMA (p, d, q) onde p é a ordem AR, d a ordem de diferenciação e q a ordem MA.

À luz desse arcabouço, modelam-se as variações do Ibovespa (ΔIBOV); assim, na estimação, *d* = 0, pois a diferenciação é aplicada previamente. A presença de intercepto (constante) foi avaliada e optou-se por estimar sem constante (conforme validação interna). As defasagens das variáveis exógenas foram definidas por validação (tanto lag comum quanto lags específicos por variável). As condições de estacionariedade e invertibilidade foram impostas na estimação (raízes dos polinômios AR/MA fora do círculo unitário) e a adequação do ajuste foi avaliada por diagnóstico dos resíduos (e.g., autocorrelação remanescente).

Nessa perspectiva e ainda segundo Hyndman & Athanasopoulos (2021), o método ARIMA/ARIMAX segue a abordagem clássica de Box–Jenkins: (i) identificação (verificação/transformação para estacionariedade — tratadas no tópico anterior); (ii) estimação dos parâmetros; e (iii) validação do modelo. Neste estudo, as ordens (p, q) não são impostas a priori: segue-se a lógica de Hyndman–Khandakar, testando combinações parcimoniosas e escolhendo por AICc no treino. Para as variáveis exógenas, avaliam-se defasagens $L \in \{1,\ldots,6\}$ (lag comum e, quando indicado, lags por variável) e seleciona-se,

também por AICc no treino, a estrutura que segue para a validação fora da amostra em janela rolling 1-step (hold-out de 36 meses, janela expansiva), com re-estimação a cada passo e checagem de resíduos (Ljung–Box, e.g., Q (12)).

No que tange à estimação de (p, q) — isto é, a definição das ordens do AR/MA — adotou-se um grid search limitado, ou seja, optou-se por restringir o conjunto de combinações possíveis para (p, q) — isto se deve à três principais fatores: eficiência, parcimônia (vinculada à sobreparametrização) e, principalmente, alinhamento com os sinais dos diagnósticos ACF/PACF e diferenciação prévia, as quais já indicavam faixas plausíveis. A seleção do conjunto otimizado é através do $AIC = -2\,log\,log\,L\,\left(\hat{\theta}\right) + 2k$, em que $L\left(\hat{\theta}\right)$ é a verossimilhança maximizada e k o número de parâmetros (AR, MA), constante quando presente; no ARIMAX, somam-se os coeficientes das exógenas). O AICc corrige o viés do AIC em amostras finitas e é o critério preferido. Além disso, também, impõe-se ausência de termo determinístico na média do modelo (constante), opção usual para séries em diferença e confirmada pela validação empírica. A partir desse cenário e dado os conjuntos candidatos, o benchmark univariado selecionado foi o ARIMA (3,0,4), cujo valor da AICc alcançou 5872.73 — fixou-se d=0, como apontado anteriormente; a partir dele e para isolar o ganho das exógenas, comparou-se as variantes ARIMAX com lags exógenos escolhidos anteriormente, e então procedeu-se com a validação fora da amostra (OOS).

Com o arcabouço definido no parágrafo anterior, delimitou-se o procedimento de comparação e validação fora da amostra. As variáveis exógenas (SELIC, IPCA e USD/BRL) tiveram defasagens pré-fixadas via validação interna no treino, utilizando rolling 1-step com re-estimação a cada passo: testou-se um lag comum $(L \in \{0,\ldots,6\})$ e lags específicos por variável (cada exógena podendo ter uma defasagem distinta). A configuração que melhor atendeu aos critérios (AICc/erro no treino) foi congelada antes de qualquer uso no teste, prevenindo vazamento de informação (defasagens aplicadas antes do split e padronização ajustada apenas no treino de cada passo). Para a comparação, considera-se, portanto: (i) o baseline ARIMA (3,0,4); (ii) um ARIMAX (3,0,4), preservando a mesma ordem para isolar o efeito das exógenas; (iii) um ARIMAX (3,0,4) com lags por variável, obtidos em grade $(L\_SELIC, L\_IPCA, L\_DÓLAR) \in \{1,\ldots,6\}^3$ e levados à validação interna (finalistas como (4, 3, 3) e (5, 3, 5)), (iv) uma alternativa parcimoniosa ARIMAX (1,0,1) — escolhida por ser um padrão mínimo usual para séries em diferenças e por diagnósticos internos (ACF/PACF sugerindo memória curta; AICc no treino não indicando ordens altas). A avaliação fora da amostra utilizou um hold-out de 36 meses em rolling 1-step com janela expansiva e re-estimação a cada passo; as comparações consideraram MAE e RMSE e o teste de Diebold–Mariano entre pares de modelos. No Tópico 4, detalha-se o protocolo OOS e os critérios de comparação adotados.

4) Benchmarks e Métricas (ARIMA univariado + RW/Drift; MAE/RMSE; Diebold–Mariano)

À luz do parágrafo anterior, a avaliação do desempenho preditivo é realizada com base em métricas fora da amostra e na comparação com modelos de referência (*benchmarks*). Adota-se como benchmark principal o ARIMA univariado estimado sobre a série-alvo em diferenças (ΔIBOV) e selecionado por AICc no conjunto de treino. Complementarmente, utilizam-se benchmarks ingênuos que servem de piso de desempenho e ajudam a qualificar eventuais ganhos: ruído branco — ARIMA (0,0,0), random walk — ARIMA (0,1,0) sem constante, e random walk com *drift* — ARIMA (0,1,0) com constante. Como aqui a variável-alvo é ΔIBOV, o *no-change* apropriado para a meta é ARIMA (0,0,0) aplicado à série em diferenças (previsão igual a zero) (Hyndman & Athanasopoulos, 2021).

O ARIMAX proposto será considerado superior apenas se apresentar ΔMAE e/ou ΔRMSE negativos em relação ao benchmark principal e se a diferença for estatisticamente significativa pelo teste de Diebold–Mariano. Todas as previsões são geradas em rolling 1-step sobre um hold-out de 36 meses, com re-estimação a cada passo e controle de vazamento (defasagens fixadas antes do corte e z-score ajustado apenas no treino).

Nessa perspectiva e sobre as métricas utilizadas, Manu Joseph e Jeffrey Tackes (2024) apontam que, o Erro Absoluto é um erro dependente de escala, ou seja, sua magnitude está vinculada à escala do conjunto de dados e ele mede o erro absoluto médio das previsões realizadas. Sua equação encontra-se abaixo.

$$MAE = (1 \div H) \times (|e_1| + |e_2| + |e_h|) \qquad (3)$$

A métrica raiz do erro quadrático médio (RMSE) calcula o erro médio absoluto, utilizando-se de uma penalização maior para erros maiores. Sua fórmula consiste na raiz quadrada do erro quadrático médio.

$$RMSE = \sqrt{\frac{1}{H}\left(e_1^2 + e_2^2 + e_H^2\right)} \qquad (4)$$

Onde H é o número de pontos avaliados (tamanho do teste), $e_h = y_h - \widehat{y_h}$ é o erro de previsão no mês h.

De acordo com Hyndman e Athanasopoulos (2021), ambas as métricas são amplamente usadas em séries temporais; valores menores indicam melhor desempenho. Em conjunto, o MAE favorece interpretações diretas (mediana/robustez), enquanto o RMSE destaca diferenças quando há erros grandes. Para complementar MAE e RMSE, utiliza-se, também, do teste de Diebold–Mariano, o qual, de acordo com Diebold & Mariano, compara diretamente a acurácia preditiva de dois modelos a partir das diferenças de erro ao longo do

tempo. Ainda segundo os autores, o teste calcula a média dessas diferenças e a padroniza por uma estimativa de variância robusta a autocorrelação e heterocedasticidade, permitindo decidir se um modelo erra menos que o outro de forma estatisticamente significativa. Para este presente estudo, como o horizonte é de um passo à frente (janela rolling 1-step), aplica-se a versão padrão do DM com a correção de pequena amostra de Harvey–Leybourne–Newbold. (Diebold & Mariano, 1995; Harvey, Leybourne & Newbold, 1997).

Por fim, além das métricas fora da amostra (OOS) — usadas para eleger o melhor modelo —, avalia-se a contribuição das exógenas por dois diferentes meios: (i) experimento de inclusão/remoção (“ablation”) nas variantes FULL / DROP / ONLY (Hurvich & Tsai, 1989) – mensurando o respectivo desempenho OOS; e (ii) Importância por Permutação à la Breiman (2001), adaptada a séries temporais por embaralhamento em blocos (b = 3 meses) apenas no teste, para preservar a dependência temporal e evitar vazamento (Künsch, 1989; Politis & Romano, 1994).

Resultados e Discussão

Dado todo o contexto e pipeline abordados e definidos na seção anterior — diagnósticos de estacionariedade, diferenciação prévia do Ibovespa (ΔIBOV), padronização aplicada apenas no treino, prevenção de vazamento temporal e escolha de ordens por AICc — o benchmark adotado é o ARIMA (3,0,4) estimado sem constante sobre ΔIBOVESPA. Para medir o "poder preditivo" das variáveis macroeconômicas, compara-se esse ARIMA univariado a um ARIMAX com a mesma ordem (3,0,4) e os três preditores (SELIC, IPCA, USD/BRL) defasados em 1 mês. Iniciar com lag 1 é um ponto de partida parcimonioso e economicamente plausível: os indicadores são mensais e choques macro costumam repercutir nos preços com defasagem curta; além disso, essa configuração "mínima" evita super parametrização antes de explorar defasagens maiores ou específicas por variável (etapa tratada adiante). Como checagem de desempenho, confronta-se, também, os dois candidatos com três benchmarks ingênuos na mesma amostra: (i) ruído branco nos níveis (ARIMA (0,0,0)), (ii) random walk com drift (ARIMA (0,0,0) com constante) e (iii) OLS com exógenas defasadas em 1 mês (ARIMA (0,0,0)). Em termos de AICc e de acordo com a tabela abreviada abaixo (Tabela 3), todos esses benchmarks possuem um desempenho pior do que o ARIMA (ΔAICc ≈ +105 contra o ARIMA) puro e o ARIMAX, confirmando que modelos mais estruturados capturam a dinâmica básica de ΔIBOVESPA.

Tabela 3. Δ AICc vs ARIMA univariado

MODELO	ORDEM	AICc	ΔAICc vs ARIMA
ARIMA uni variado	(3, 0, 4)	5872.73	0.00
ARIMAX (L=1)	(3, 0, 4)	5868.47	-4.25
WHITE NOISE	(0, 0, 0)	5978.26	+105.53
RW com drift	(0, 0 ,0)	5977.97	+105.24
OLS exógenas	(0, 0, 0)	5978.41	+105.68

Fonte: Resultados originais da pesquisa

No ARIMAX, manter a mesma estrutura ARMA isola o efeito das exógenas, ou seja, qualquer ganho de ajuste é atribuível às variáveis macro, e não a mudanças na dinâmica autorregressivo/média-móvel. Na comparação in-sample — tabela acima —, os resultados indicam empate técnico entre ARIMA (3,0,4) e ARIMAX (3,0,4) com lag 1: ΔAICc ≈ −4,25 em favor do ARIMAX — diferença pequena demais para ser interpretada como ganho material. Em ambos os casos, os testes de Ljung–Box sobre os resíduos (Q (12)) retornam *p-values* elevados (≈0,83 no ARIMA e ≈0,79 no ARIMAX), indicando ausência de autocorrelação remanescente e, portanto, ajustes bem-comportados. Esse resultado, a priori, dialoga com o

estudo proposto por Pimenta Júnior & Higuchi (2008), no qual, através de estimação dos VARs para o período de 1994–2005, não se encontrou causalidade de Granger de SELIC, IPCA e PTAX para o Ibovespa, ou seja, choques no próprio índice explicam a maior parcela do erro de previsão.

No entanto, para que a comparação com a literatura seja válida, é necessário analisar o desempenho com dados fora da amostra (OOS). No hold-out de 36 meses — ou seja, no espaço amostral de testes (3 anos) —, com janela expansiva e re-estimação a cada passo (previsão de um novo mês, por vez), mantendo a ordem fixa (3, 0, 4) — para isolar o efeito das exógenas —, o ARIMA puro alcançou MAE=4513,45 e RMSE=5432,58, ao passo que o ARIMAX(3,0,4) com defasagem comum de 1 mês obteve MAE=4601,87 e RMSE=5481,06 — piora de ~1,9% no MAE e ~0,9% no RMSE. Com isso, para essa configuração inicial, as variáveis macroeconômicas não acrescentaram poder preditivo relevante em um cenário de curto prazo; a pequena vantagem *in-sample* do ARIMAX (ΔAICc ≈ −4,25) não se materializou OOS. Essa conclusão, novamente, é coerente com achados de previsibilidade frágil e instável em frequência mensal (e.g., Goyal & Welch, 2008 e Rapach & Zhou (2013). Economicamente, uma explicação plausível para tal observado é o fato de que, em horizonte mensal, a dinâmica autorregressiva do próprio índice captura a maior parte do sinal útil, deixando pouco espaço para ganho incremental das exógenas, ou seja, o histórico recente do próprio índice já é um previsor tão bom para o próximo mês que outras variáveis externas (como juros, câmbio etc.) acabam não adicionando informações novas e relevantes para melhorar a previsão.

Nesta perspectiva, para verificar se o insucesso da defasagem de um mês para as variáveis exógenas (L = 1) não decorre de um mau alinhamento temporal — e para testar outras defasagens —, realizou-se — somente no treino, mantendo a estrutura ARMA (3, 4), fazendo o shift das exógenas antes do corte e aplicando z-score ajustado — seleção interna de defasagem comum $L \in \{1,\ldots,6\}$. A grade por AICc (Tabela 4) indica L=4 como o menor AICc (5035,87), à frente de L=1 (5037,84) e L=6 (5039,17). Mesmo havendo uma diferença irrisória (ΔAICc ≈ 2 p.p em relação à L = 1) — sinalizando preferência por defasagem de 4 meses — é uma aplicação plausível do ponto de vista econômico, uma vez que, a transmissão de choques de juros e inflação aos preços pode levar alguns meses. Afirmação, essa, que corrobora com as evidências de que, em ARDL/VECM, o Ibovespa documenta relações de longo prazo e ajustes de curto prazo defasados, sugerindo que os efeitos de choques de juros e inflação não são contemporâneos (da Silva et al., 2014; OJBM, 2024). Portanto, levou-se L=4 como candidata para a etapa OOS, repetindo o rolling 1-step com re-estimação — como feito para a defasagem de um mês, exposto no parágrafo anterior.

Tabela 4. Δ AICc vs ARIMA univariado

LAG	AICc
4	5035.87
1	5037.84
6	5039.17
5	5043.00
3	5048.05
2	5049.48

Fonte: Resultados originais da pesquisa

Repetindo o protocolo OOS com a defasagem estipulada e sendo igual à quatro (L=4) para todas as variáveis exógenas e mantendo a estrutura ARIMA fixada em (3, 0, 4), o ARIMA pura obtive MAE=4444,24 e RMSE=5334,03, enquanto o ARIMAX (3, 0, 4; L=4) registrou MAE=4643,70 e RMSE=5563,83 — uma piora de ≈4,5% no MAE e ≈4,3% no RMSE. Percebe-se que os valores do ARIMA univariado (puro) diferem dos valores do parágrafo anterior e isto se deve ao fato de (i) testar L=4, a amostra efetiva é realinhada e, (ii) no rolling 1-step com re-estimação, cada passo refita os parâmetros com um histórico ligeiramente distinto. No que tange o desempenho inferior do ARIMAX com L=4, explica-se que, a vantagem in-sample era tênue (ΔAICc ~ 2 pontos face a L=1) e, portanto, facilmente não generalizável. Além disso, escolher L [lag] no treino pode introduzir viés de seleção; com sinal macro mensal fraco/instável, os três coeficientes adicionais tendem a elevar a variância das estimativas sem compensação em viés, sobretudo num hold-out curto. Na prática, o que se viu foi exatamente o recomendado pela literatura: benchmark univariado forte e avaliação OOS rigorosa expõem ganhos marginais das exógenas que não se materializam fora da amostra (Goyal & Welch, 2008).

Como verificação adicional de robustez, realizou-se um estudo de performance fora da amostra (rolling 1-step com janela expansiva e re-estimação a cada passo — in-sample) — avaliando, Tabela 5, lags comuns $L \in \{1,\ldots,6\}$ e, Tabela 6, tuplas por variável (isto é, cada variável exógena possuindo uma defasagem diferente). Nessa exploração na base de testes, algumas configurações superaram o benchmark em métricas: p.ex., lag comum L=6 obteve MAE≈4429 e RMSE≈5349 (vs. 4505/5452 do ARIMA puro na mesma amostra) e a tupla (SELIC, IPCA, DÓLAR) = (2,1,6) atingiu MAE≈4366 e RMSE≈5240 (ΔMAE≈−139; ΔRMSE≈−212). Entretanto, por utilizarem o mesmo hold-out para busca e avaliação, tais ganhos são otimistas (risco de *data snooping*).

Tabela 5. Métricas envolvendo diferentes defasagens no ambiente de teste

LAG	MAE ARIMAX	RMSE ARIMAX	MAE PURO	RMSE PURO
6	4429.11	5349.93	4505.21	5452.48

1	4585.98	5497.02	4505.21	5452.48
5	4575.90	5590.70	4505.21	5452.48
4	4712.84	5637.29	4505.21	5452.48
3	4812.14	5780.10	4505.21	5452.48
2	4863.40	5784.78	4505.21	5452.48

Fonte: Resultados originais da pesquisa

Tabela 6. Métricas envolvendo tuplas por variáveis em um cenário de teste

LAG SELIC	LAG IPCA	LAG DÓLAR	MAE	RMSE
2	1	6	4366.23	5240.43
5	1	4	4259.60	5263.32
3	1	6	4287.95	5268.09
6	2	6	4397.60	5275.68
3	4	6	4291.16	5316.27
6	1	4	4378.00	5326.10
6	4	6	4358.31	5331.62
3	6	6	4366.08	5344.70
6	6	6	4429.12	5350.00
6	1	6	4401.70	5360.74

Fonte: Resultados originais da pesquisa

Para mitigar o otimismo indicado no parágrafo anterior — decorrente de buscar e avaliar no mesmo hold-out (ambiente de teste) — adota-se uma abordagem em formato de funil só no treino em duas etapas: (i) pré-seleção por AICc in-sample, aplicando as defasagens antes do corte e fazendo z-score apenas no treino, mantendo ARMA (3, 0, 4) fixo; e (ii) validação interna com rolling 1-step (~24 meses) apenas nos finalistas, com re-estimação a cada passo. Nessa validação, o lag comum vencedor foi L=3 (superando L=4 e L=6). Importa notar que o fato de L=3 superar L=4 decorre do método de seleção: a escolha prévia de L=4 veio de uma busca in-sample por AICc, sem o rolling 1-step; já o funil com rolling penaliza configurações que não generalizam. Em sequência, Tabela 7, nota-se que a melhor combinação por variável foi (SELIC, IPCA, DÓLAR) = (4, 3, 3). Com base nisso, pré registra-se para o hold-out final de 36 meses a comparação contra o benchmark usando o lag comum L=3 e, como alternativa, a tupla (4,3,3), preservando todo o protocolo de antivazamento.

Tabela 7. Métricas envolvendo diferentes defasagens no ambiente de treino

LAG SELIC	LAG IPCA	LAG DÓLAR	MAE	RMSE
4	3	3	5043.44	6274.97
5	3	5	4917.00	6348.73

4	5	6	5076.05	6375.80
4	5	5	5227.96	6641.37
4	2	5	5538.30	6762.29
4	5	3	5500.80	6811.94
4	5	1	5280.10	6811.94
4	4	5	5635.54	6911.13
1	3	5	5750.82	7231.81
4	3	5	5678.12	7322.54

Fonte: Resultados originais da pesquisa

Posteriormente, dado o ARIMAX com as tuplas pré-selecionadas (envolvendo o parágrafo anterior), avalia-se esse novo modelo contra o benchmark ARIMA (3, 0, 4), para averiguar se, com diferentes defasagens para as variáveis exógenas, há a obtenção de menores valores para as métricas escolhidas frente ao modelo univariado. Com isso, observa-se na tabela acima que, nenhuma delas (comparou-se, também, a tupla (5, 3, 5) - segunda mais bem colocada) superou o univariado: para (SELIC, IPCA, DÓLAR) = (5,3,5) obtivemos MAE≈4917,0 e RMSE≈6348,7 (piora em relação ao MAE e RMSE do ARIMA puro — MAE≈4513,5 / RMSE≈5432,6). Para (4,3,3), MAE≈4702,5 e RMSE≈5750,3 (resultado, também, inferior ao benchmark selecionado). Neste contexto, os testes de Diebold–Mariano não rejeitam igualdade de acurácia: para (5,3,5), DM(MAE)≈−0,75, p≈0,45 e DM(MSE)≈−1,34, p≈0,18; para (4,3,3), DM(MAE)≈−1,03, p≈0,30 e DM(MSE)≈−1,85, p≈0,065. Em suma, as tuplas exógenas avaliadas não acrescentaram poder preditivo estatisticamente significativo no curto prazo e, em nível de erro, ficaram levemente piores que o benchmark. Esse padrão é consistente com evidência brasileira e latino-americana de baixa causalidade/instabilidade de preditores macro (Pimenta Júnior & Higuchi, 2008).

Por fim, envolvendo a etapa no hold-out (H=36), inclui-se o ARIMAX (1,0,1) como alternativa parcimoniosa — essa nova adição está em linha com Box–Jenkins/Hyndman–Khandakar, uma vez que, os diagnósticos ACF/PACF de ΔIBOV sugerem memória curta e, ao adicionar exógenas, manter o ARMA baixo ajuda a isolar o efeito das regressoras. Percebe-se, com isso e com as tuplas (5,3,5) e (4,3,3) para as variáveis exógenas, que o ARIMAX (1,0,1) - (5,3,5) apresentou leve melhora sobre o univariado (MAE≈4468,7 e RMSE≈5400,1 vs. 4513,5/5432,6; ΔMAE≈−44,9; ΔRMSE≈−19,0), enquanto o ARIMAX (1,0,1) - (4,3,3) melhorou o RMSE (≈5380,2) mas piorou o MAE (≈4585,3). Todavia, em ambos os casos, o teste de Diebold–Mariano não rejeita igualdade de acurácia (p(MAE)≈0,90/0,81; p(MSE)≈0,94/0,86). Assim, mesmo com a alternativa parcimoniosa, os ganhos frente ao benchmark selecionado foram marginais e estatisticamente não significativos.

Tabela 8. Métricas envolvendo ARIMAX (1, 0, 1) com diferentes defasagens

MODELO	MAE ARIMAX	RMSE ARIMAX	MAE PURO	RMSE PURO
ARIMAX (1, 0, 1) – (4, 3, 3)	4585.28	5380.17	4513.54	5419.09
ARIMAX (1, 0, 1) – (5, 3, 5)	4468.68	5400.08	4513.54	5419.09
ARIMAX (1, 0, 1) – (L=3)	4679.98	5699.69	4513.54	5419.09

Fonte: Resultados originais da pesquisa

Como checagem adicional — uma vez obtido resultados promissores, no que tange às métricas utilizadas — reestima-se o ARIMAX (1, 0, 1), in-sample, para as tuplas, ou seja, analise-se os resultados do ARIMAX para as defasagens selecionadas (5,3,5) e (4,3,3). No caso da tupla (5, 3, 5), o IPCA foi a única variável macroeconômica que apresentou um sinal negativo, enquanto SELIC e dólar não são significantes para a estimação; os resíduos passaram no diagnóstico Ljung-Box e há indícios de heterocedasticidade. Neste contexto, afirma-se que os sinais se alinham parcialmente à literatura vigente: o efeito adverso da inflação sobre retornos (Fama, 1981; Nunes et al., 2005); impactos negativos de depreciações cambiais sobre o Ibovespa (Franzen et al., 2009; Montes & Tiberto, 2011) — todavia, para este estudo, o coeficiente do dólar seja pequeno/não significativo, assim como, ocorre com a SELIC. Em conjunto, a baixa significância e a ausência de ganho OOS ajudam a explicar por que as tuplas não superaram o benchmark no hold-out — como visto no nono parágrafo deste tópico. Ressalta-se, portanto, que, entre o ARIMAX (1, 0, 1) com diferentes defasagens para as variáveis exógenas e o ARIMA univariado, este último deve seguir como referência principal.

Como extensão exploratória para este presente estudo, realizou-se o teste de exclusão de variáveis (FULL/DROP/ONLY) — análise de sensibilidade das variáveis (Tabela 9) — no ARIMAX (1, 0, 1) com a tupla (5, 3, 5). No treino (AICc), o modelo FULL (SELIC+IPCA+DÓLAR) foi o melhor (AICc = 5122.28); retirar DÓLAR penalizou pouco (+0.44), usar só IPCA (+0.72) ou dropar SELIC (+0.78) piorou moderadamente, enquanto só DÓLAR (+3.17), só SELIC (+4.09) e dropar IPCA (+4.74) foram claramente inferiores— sinal de que o IPCA carrega boa parte do ajuste in-sample. Já fora da amostra, Tabela 10, o ranking muda: ONLY_SELIC obteve MAE=4387 e RMSE≈5267, superando levemente o PURO (sem exógenas, ARIMA (1,0,1)) (MAE=4422; RMSE=5281); DROP_DÓLAR ficou próximo (RMSE≈5297), ao passo que o FULL piorou (RMSE≈5400). Os ganhos são pequenos (<3%) e instáveis across specs, e não alteram a conclusão central de que o ARIMA (3,0,4) univariado permanece o benchmark mais sólido no período analisado. Em suma, o experimento de remoção/adição de variáveis sugere que a SELIC pode ajudar marginalmente no curtíssimo

prazo nesta janela, enquanto o IPCA explica mais o ajuste in-sample— uma observação que corrobora com os estudos apontados no parágrafo anterior.

Tabela 9. Teste de exclusão de variáveis (FULL/DROP/ONLY)

MODELO	EXOG	AICc	AICc vs FULL
FULL	SELIC + IPCA + DÓLAR	5122.27	0.00
DROP DOLAR	SELIC + IPCA	5122.72	0.44
ONLY IPCA	IPCA	5123.00	0.72
DROP SELIC	IPCA + DÓLAR	5123.06	0.78
ONLY DÓLAR	DÓLAR	5125.44	3.16
ONLY SELIC	SELIC	5126.37	4.09
DROP IPCA	SELIC + DÓLAR	5127.01	4.73

Fonte: Resultados originais da pesquisa

Tabela 10. Teste de exclusão de variáveis (FULL/DROP/ONLY) - OOS

MODELO	MAE	RMSE
ONLY SELIC	4387.15	5266.97
PURO	4421.61	5281.42
DROP DOLAR	4396.74	5296.99
ONLY IPCA	4477.21	5321.22
DROP IPCA	4448.23	5337.60
ONLY DOLAR	4482.02	5350.30
FULL	4468.68	5400.08
DROP SELIC	4546.80	5414.07

Fonte: Resultados originais da pesquisa

Ademais, com o intuito de complementar graficamente este presente estudo, apresenta-se as Figuras 4 e 5, as quais apontam para os achados quantitativos e comparação de desempenho entre os melhores modelos encontrados. A Figura 4 representa a previsão em nível para H=2, ou seja, ilustra o ARIMAX (1, 0, 1) com a tupla (4, 3, 3). Os dois pontos previstos (em vermelho) apontam queda, mas envoltos por um leque de incerteza amplo (faixa cinza), o que é natural quando se projeta em diferenças e depois se acumula para o nível — a variância cresce com o horizonte e torna o ponto uma indicação direcional, não um valor preciso. A Figura 5 representa a comparação OOS em ΔIBOV, ou seja, ela mostra que tanto o ARIMA (3,0,4) quanto o ARIMAX suavizam a volatilidade observada (linha preta), porém o ARIMA (cinza) acompanha melhor as viradas e a magnitude dos movimentos, enquanto o ARIMAX (vermelho) tende a encolher para a média (under-reaction), ficando atrás nos picos

e vales. Visualmente, portanto, os gráficos corroboram o que as métricas e os testes indicaram ao longo da seção: no horizonte mensal e sob o protocolo de validação adotado, o ARIMAX com (4,3,3) tente a apresentar um desempenho inferior frente ao benchmark.

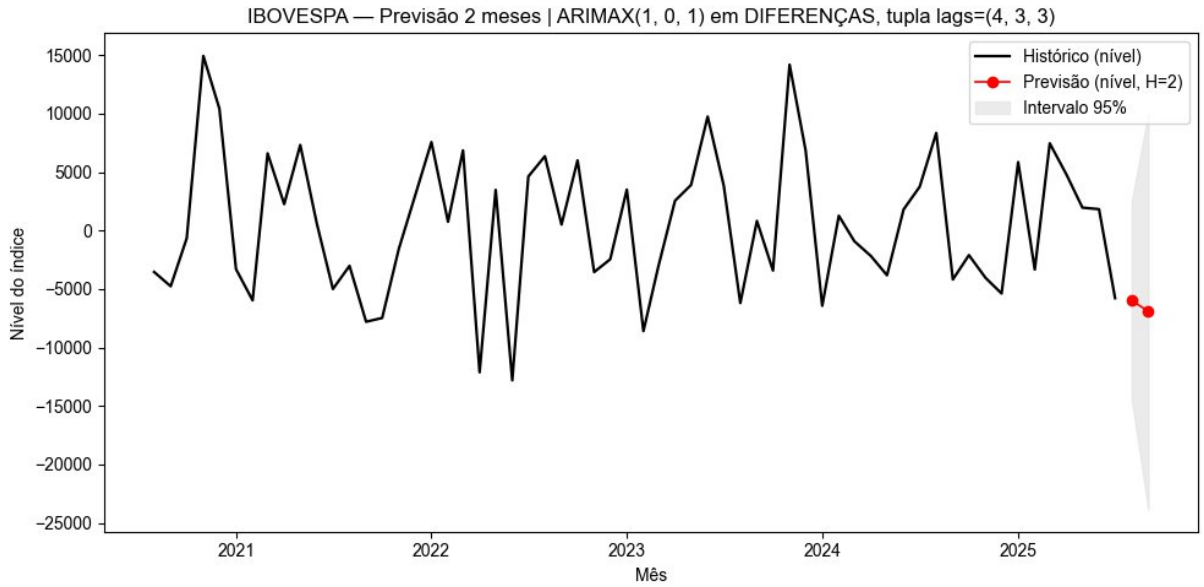

Figura 4. a previsão em nível para H=2 para ARIMAX (1, 0, 1) – (4, 3, 3)

Fonte: Resultados originais da pesquisa

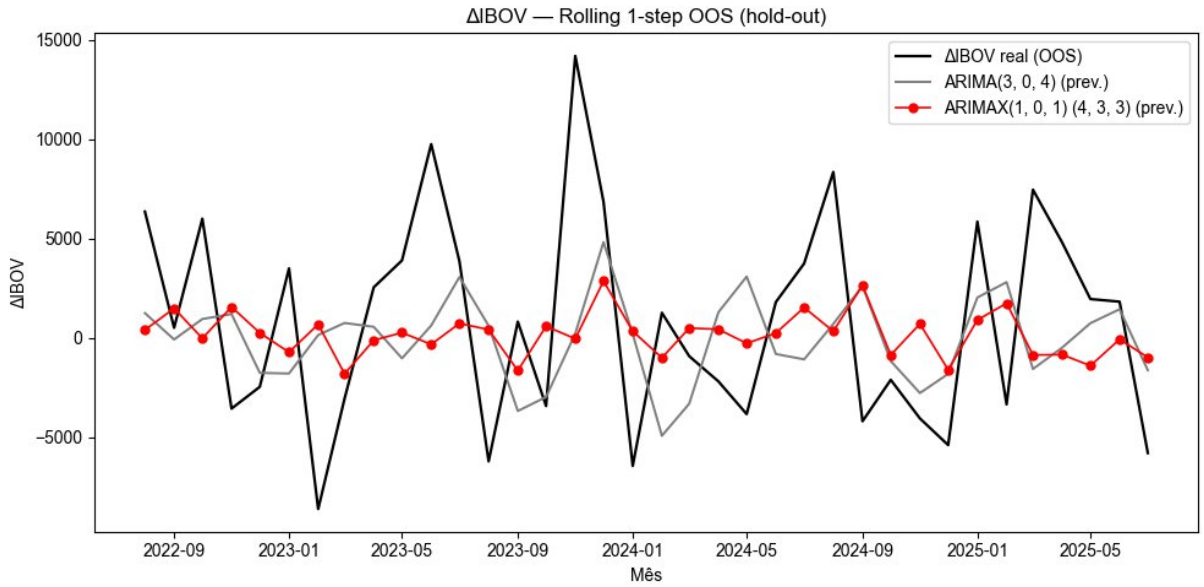

Figura 5. comparação OOS em ΔIBOV

Fonte: Resultados originais da pesquisa

Por fim, confrontou-se os achados deste estudo com a dissertação de Holanda Filho (2024). Para 2012–2022, o autor testou ADF/KPSS, um Logit com dummies de calendário e

controles, e variantes ARIMAX; em seu estudo estabeleceu-se que, dummies semanais/mensais não foram relevantes e as variáveis macro usuais (SELIC, IPCA e câmbio) exibiram baixo poder explicativo, ao passo que medidas endógenas do próprio índice (médias móveis) foram mais informativas. Essa conclusão obtida pelo autor em questão reflete com os resultados obtidos até então — em frequência mensal, o ARIMA univariado segue um benchmark difícil de superar e ganhos com ARIMAX, quando surgem, são pontuais e não persistentes fora da amostra, ou seja, a previsibilidade mensal de curto prazo do Ibovespa é dominada pela própria dinâmica ARMA do índice, e choques macro raramente se traduzem em melhorias robustas de acurácia sob avaliação OOS rigorosa. Ademais, entre as exógenas testadas neste trabalho, a que mais se destacou foi o IPCA — com sinal adverso e significância apenas in-sample em algumas especificações —, mas sem sustentação no hold-out; SELIC e USD/BRL mostraram efeitos ainda mais tímidos.

Conclusão(ões) ou Considerações Finais

Este estudo comparou, em horizonte mensal de um passo à frente, a capacidade preditiva do ARIMA univariado e do ARIMAX com SELIC, IPCA e USD/BRL como exógenas. Para 2000–2025, o ARIMA (3,0,4) manteve-se como referência robusta. O ARIMAX, construído de forma parcimoniosa e com controle de vazamento, chegou a empatar ou apresentar pequenos ganhos, porém não persistentes. Assim, recomenda-se o ARIMA (3,0,4) como baseline, com o ARIMAX servindo como complemento exploratório.

As inferências indicam que, no curto prazo, a maior parte do sinal útil é capturada pela própria dinâmica autorregressiva e de média móvel do índice. As variáveis macroeconômicas, mesmo defasadas e padronizadas, acrescentaram pouco ganho incremental. Houve melhorias pontuais ao ajustar defasagens por variável, mas não se sustentaram fora da amostra. Operacionalmente, um modelo enxuto, estável e reestimado ao longo do tempo mostrou-se mais vantajoso do que ampliar o conjunto de preditores — em linha com o desempenho observado do ARIMAX (1,0,1).

A principal contribuição é metodológica. Apresentamos um pipeline reprodutível que combina seleção por AICc (parcimônia), validação interna sem vazamento e avaliação fora da amostra por rolling 1-step, com clara separação entre escolhas de treino e verificação em teste. Esse protocolo reduz sobre ajuste, formaliza a comparação entre modelos e pode ser replicado em outros contextos e bases.

Por fim, aponta-se que, por este trabalho ter adotado (i) frequência mensal, (ii) três preditores macro e (iii) avaliação fora da amostra 1-passo à frente com hold-out de 36 meses, algumas extensões relevantes ficaram fora do escopo. Entre elas: (i) ampliar o conjunto de variáveis — commodities, curva a termo de juros, spreads de crédito, risco-país e fatores globais — e incluir medidas de sentimento (notícias, comunicados oficiais, textos de balanço) obtidas por NLP, para verificar se tais adições elevam a previsibilidade do ARIMAX; (ii) tratar heterocedasticidade e mudanças de regime (p.ex., ARIMAX-GARCH e parâmetros variantes no tempo); (iii) empregar seleção regularizada e modelos de aprendizado de máquina/redes neurais (LASSO/Elastic Net, Random Forest/GBM, XGBoost/LightGBM) com validação em blocos e validação aninhada para controlar sobre ajuste; e (iv) realizar testes de estabilidade (quebras estruturais), avaliar janelas alternativas e considerar métricas adicionais de perda. Uma agenda natural inclui ampliar preditores, incorporar sentimento textual e comparar modelos estatísticos e de ML sob o mesmo protocolo antivazamentos. Finalmente, analisar o Ibovespa em frequência diária (ou semanal) exigirá nowcasting das macro mensais e tratamento explícito da volatilidade intramês, mas pode revelar padrões de curtíssimo prazo que não emergem na frequência mensal.

Referências

BREIMAN, Leo. Random forests. *Machine Learning*, v. 45, n. 1, 2001.

Diebold, F.X.; Mariano, R.S. 1995. Comparing Predictive Accuracy. *Journal of Business & Economic Statistics* **13(3): 253–263**. Disponível em: (Taylor & Francis/JSTOR). Acesso em: 26 set. 2025

FRANZEN, André; MEURER, Roberto; GONÇALVES, Carlos Eduardo Soares; SEABRA, Fernando. Determinantes do fluxo de investimentos de portfólio para o mercado acionário brasileiro. *Estudos Econômicos (São Paulo)*, v. 39, n. 2, p. 301–328, 2009.

Goyal, A.; Welch, I. 2008. A comprehensive look at the empirical performance of equity premium prediction. Review of Financial Studies 21(4): 1455-1508.

Harvey, D.I.; Leybourne, S.J.; Newbold, P. 1997. Testing the Equality of Prediction Mean Squared Errors. *International Journal of Forecasting* **13(2): 281–291**. Disponível em: (Elsevier/EconPapers). Acesso em: 28 set. 2025.

Holanda Filho, I. 2024. Impacto de variáveis macroeconômicas no retorno de ações brasileiras: evidências para o mercado acionário brasileiro (Ibovespa). Dissertação de Mestrado em [Programa]. Universidade Federal do Ceará, Fortaleza, CE, Brasil. Disponível em: <URL>. Acesso em: 29 set. 2025

Hurvich, C.M.; Tsai, C.-L. 1989. Regression and time series model selection in small samples. Biometrika 76(2): 297-307.

HYNDMAN, Rob J.; ATHANASOPOULOS, George. 2021. *Forecasting: Principles and Practice*. 3. ed. OTexts. Disponível em: https://otexts.com/fpp3/ (ver tb. cap. “1.7 The statistical forecasting perspective”: https://otexts.com/fpp3/perspective.html). Acesso em: 21 set. 2025.

KNIGHT, John; SATCHELL, Stephen (eds.). 2007. *Forecasting Volatility in the Financial Markets*. 3. ed. Oxford: Butterworth-Heinemann/Elsevier. ISBN 978-0-7506-6942-9. Disponível em: https://shop.elsevier.com/books/forecasting-volatility-in-the-financial-markets/satchell/978-0-7506-6942-9. Acesso em: 30 set. 2025.

Künsch, H.R. 1989. The jackknife and the bootstrap for general stationary observations. The Annals of Statistics 17(3): 1217-1241.

NUNES, Maurício S.; COSTA JR., Newton C. A. da; MEURER, Roberto. A relação entre o mercado de ações e as variáveis macroeconômicas: uma análise econométrica para o Brasil. *Revista Brasileira de Economia*, v. 59, n. 4, p. 585–607, 2005.

OLIVEIRA, Luma de; ABRITA, Mateus Boldrine. Interaction between Macroeconomics Variables and IBOVESPA, the Brazilian Stock Market’s Index. *Transnational Corporations Review*, v. 5, n. 4, p. 81–95, 2013.

Pimenta Júnior, T.; Higuchi, R.H. 2008. Variáveis macroeconômicas e o Ibovespa: um estudo da relação de causalidade. *REAd – Revista Eletrônica de Administração* **14(2): 296–315**. Disponível em: (UFRGS/REAd). Acesso em: 30 set. 2025.

Politis, D.N.; Romano, J.P. 1994. The stationary bootstrap. Journal of the American Statistical Association 89(428): 1303-1313.

Rapach, D.; Zhou, G. 2013. Forecasting Stock Returns. *Handbook of Economic Forecasting*, **Vol. 2A**, Elsevier, **pp. 328–383**. Disponível em: (SSRN/working paper; RePEc ficha). Acesso em: 10 set. 2025.

ROSS, Stephen A.; WESTERFIELD, Randolph W.; JAFFE, Jeffrey; LAMB, Roberto. 2015. *Administração Financeira*. 10. ed. Porto Alegre: AMGH. ISBN 978-85-8055-432-8. Disponível em: https://books.google.com/books?id=N3sTBwAAQBAJ. Acesso em: 25 set. 2025.

Santos, A.S.; Rondina Neto, A.; Araújo, E.C.; Oliveira, L.; Abrita, M.B. 2013. Interaction between Macroeconomics Variables and IBOVESPA, the Brazilian Stock Market's Index. *Transnational Corporations Review* **5(4): 81–95**. Disponível em: (ScienceDirect). Acesso em: 28 set. 2025.

SANTOS, Allan Silveira dos; RONDINA NETO, Angelo; ARAÚJO, Eliane Cristina; SILVA, Fabiano Mello da; CORONEL, Daniel Arruda; VIEIRA, Kelmara Mendes. Causality and Cointegration Analysis between Macroeconomic Variables and the Bovespa. *PLOS ONE*, v. 9, n. 2, e89765, 2014.

Tabak, B.M. 2006. The Dynamic Relationship between Stock Prices and Exchange Rates: Evidence for Brazil. *Banco Central do Brasil*, **Working Paper 124**. Disponível em: (BCB WPS 124; versão IJTAF também disponível). Acesso em: 25 set. 2025.

Vartanian, P.R.; Farias, H.; Fronzaglia, M.L. 2022. A comparative analysis of macroeconomic and financial variables' influence on Brazilian stock and real estate markets. In: Encontro da ANPAD (ENANPAD), 46., 2022, Rio de Janeiro, RJ, Brasil. Anais. Disponível em: <URL>. Acesso em: 29 set. 2025.

VIEIRA, Edson Roberto; FERRANDO, Gabriel Silva. Macroeconomic Determinants of the Brazilian Stock Market: An Autoregressive Distributed Lag Approach. *Open Journal of Business and Management*, v. 12, n. 6, p. 4055–4072, 2024.

ÇAKICI, Nusret; ZAREMBA, Adam. 2024. What drives stock returns across countries? Insights from machine learning models. *International Review of Financial Analysis*, v. 96 (Part A), 103569. DOI: 10.1016/j.irfa.2024.103569. Disponível em: https://ideas.repec.org/a/eee/finana/v96y2024ipas1057521924005015.html. Acesso em: 25 set. 2025.