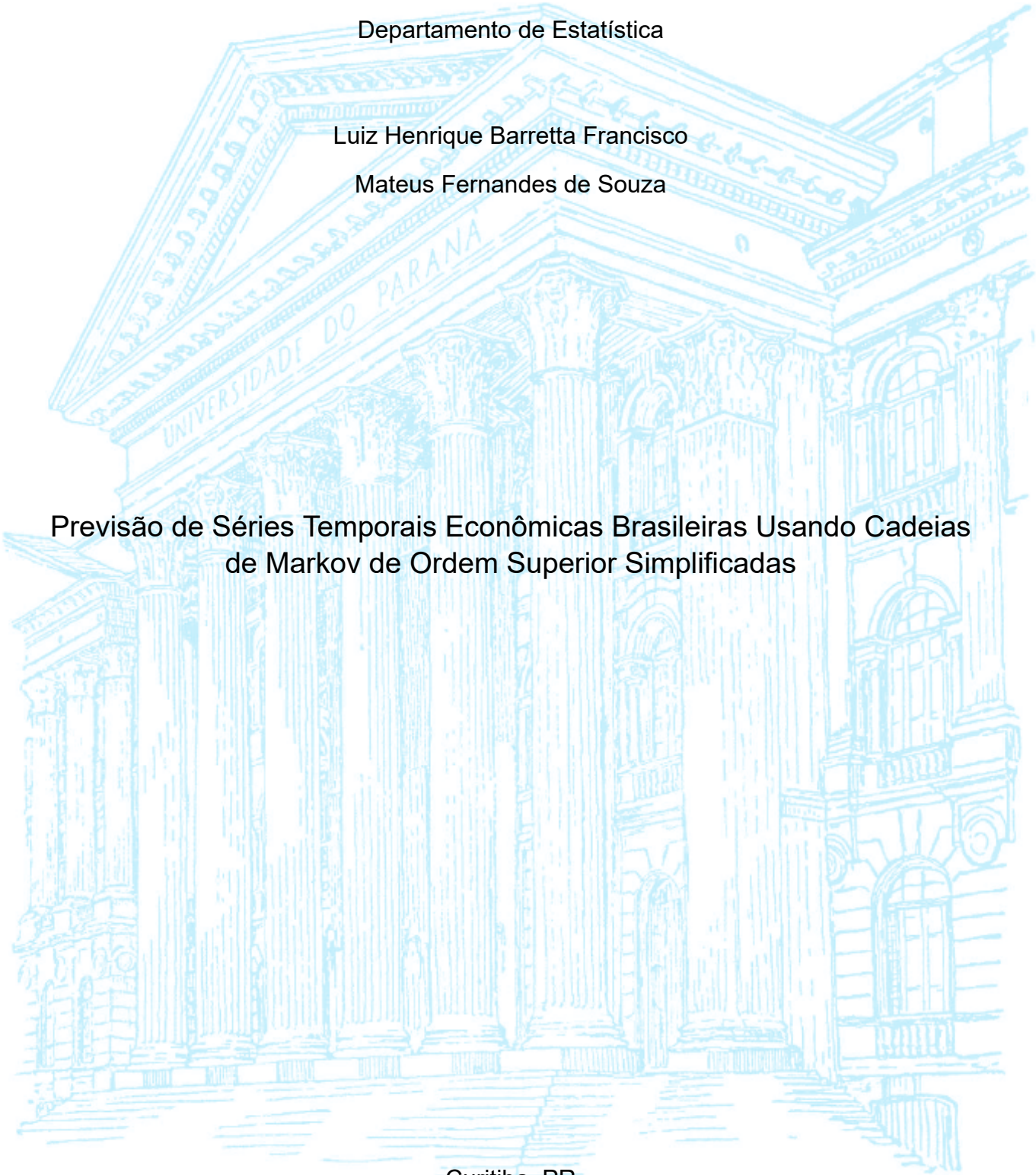
Universidade Federal do Paraná

Setor de Ciências Exatas

Departamento de Estatística

Luiz Henrique Barretta Francisco

Mateus Fernandes de Souza



Previsão de Séries Temporais Econômicas Brasileiras Usando Cadeias de Markov de Ordem Superior Simplificadas

Curitiba, PR

2025

Luiz Henrique Barretta Francisco

Mateus Fernandes de Souza

Previsão de Séries Temporais Econômicas Brasileiras Usando Cadeias de Markov de Ordem Superior Simplificadas

Trabalho de Conclusão de Curso apresentado à disciplina Laboratório B do Curso de Graduação em Estatística da Universidade Federal do Paraná, como exigência parcial para obtenção do grau de Bacharel em Estatística.

Orientador: Prof. Dr. Fernando Lucambio Pérez

Curitiba, PR

2025

Agradecimentos

Gostaria de expressar minha mais profunda gratidão a todos que, de alguma forma, contribuíram para a conclusão deste trabalho e para a minha jornada acadêmica.

Agradeço imensamente ao meu orientador, Prof. Dr. Fernando Lucambio Pérez, que desde o início se mostrou extremamente solícito e disposto a nos ajudar em tudo que precisamos, fornecendo direcionamento claro e um material de consulta fundamental para o desenvolvimento desta pesquisa. Ao integrante da banca, Prof. Dr. Paulo Justiniano Ribeiro Junior, agradeço pelas aulas que despertaram uma imensa fome de aprendizado; espero construir muito conhecimento ao seu lado, agora como meu orientador na Pós-Graduação (PPGMNE/UFPR), na continuação desta caminhada. Estendo meus agradecimentos aos demais professores do Departamento de Estatística e à Universidade Federal do Paraná como um todo, por oferecer um ensino público, gratuito e de qualidade, que tem o poder de mudar vidas, inclusive a minha. Aos meus colegas de turma, obrigado pelas valiosas discussões e pela construção conjunta do conhecimento ao longo dos anos.

Sou grato aos servidores e colegas do meu estágio no Tribunal de Justiça do Paraná, que me proporcionaram um ambiente de grande aprendizado e a oportunidade de aplicar a estatística para melhorar a vida em sociedade.

Um agradecimento especial à minha dupla, Mateus Fernandes de Souza. Desde os primeiros períodos da faculdade, nossa parceria foi uma base sólida para toda a nossa caminhada no curso, sempre pautada pela busca da excelência e pelo apoio mútuo.

Aos meus amigos, que estão sempre ao meu lado me alegrando e tornando os desafios mais leves, minha sincera gratidão.

Por fim, dedico este trabalho aos meus pais, pelo apoio incondicional em tudo que me proponho a fazer. Sem vocês, nada disso seria possível.

Luiz Henrique Barretta Francisco

Agradecimentos

Chegar até aqui foi uma construção coletiva. Este trabalho carrega muito mais do que equações e gráficos: ele carrega cada conversa de incentivo, cada apoio emocional, cada aula transformadora e cada noite em claro de dedicação.

Quero expressar minha mais sincera gratidão ao meu orientador, Prof. Dr. Fernando Lucambio Pérez, pela atenção constante, pela generosidade ao compartilhar conhecimento e por ser sempre presente, com orientações claras e objetivas que foram essenciais para este trabalho.

Agradeço especialmente ao Prof. Dr. Paulo Justiniano Ribeiro Junior e ao Prof. Dr. Cesar Taconeli. Ambos foram responsáveis por despertar em mim algo que vai muito além da curiosidade acadêmica: uma verdadeira paixão pela Estatística. Suas aulas foram marcos importantes na minha formação, combinando uma didática teórica impecável com aplicações práticas envolventes que me mostraram o valor real da nossa ciência.

Sou também grato ao Prof. Dr. Benito, que lá no início da graduação me fez acreditar que eu era capaz. Sua atenção e carinho foram combustíveis fundamentais para que eu não desistisse nos primeiros semestres.

À minha namorada, que esteve ao meu lado em absolutamente todos os momentos, deixo aqui o mais profundo agradecimento. Sua presença foi meu alicerce durante toda a graduação — mesmo à distância, sua força, carinho e palavras de incentivo me deram ânimo nos momentos mais difíceis e serenidade para seguir. Este trabalho também é seu, pois sem o seu apoio constante, emocional e afetivo, minha trajetória não teria sido a mesma. Obrigado por acreditar em mim quando eu mesmo hesitei.

Aos meus pais, minha base e meu maior apoio, agradeço do fundo do coração por sempre acreditarem em mim e por me sustentarem nos momentos em que eu mesmo duvidei. Este trabalho é, acima de tudo, uma conquista nossa.

Ao meu grande amigo e dupla neste trabalho, Luiz, meu muito obrigado. Nossa parceria é construída desde os primeiros períodos e sempre foi marcada por esforço mútuo, responsabilidade e busca pela excelência. É uma honra trilhar esse caminho com você.

Por fim, agradeço a todos os colegas de turma, professores do Departamento de Estatística da UFPR, e à própria universidade — por me oferecer um ensino público, gratuito e de qualidade, que me transformou não só como profissional, mas como pessoa.

A todos vocês, minha eterna gratidão.

Mateus Fernandes de Souza

Resumo

Este trabalho avalia a robustez e a aplicabilidade de um modelo de Cadeias de Markov de Ordem Superior Simplificada para a previsão de séries temporais no contexto econômico brasileiro. Seguindo a metodologia de Ky e Tuyen (2018), o modelo foi testado em um conjunto diversificado de dados, incluindo ações do Ibovespa e indicadores macroeconômicos do Sistema Gerenciador de Séries Temporais do Banco Central do Brasil, através de um pipeline que envolveu a discretização de log-retornos e a otimização de hiperparâmetros (ordem e número de estados) via validação por janela deslizante. Os resultados demonstraram que o modelo é flexível, adaptando sua complexidade à dinâmica de cada série, e alcançou alta acurácia (baixo MAPE) para dados com padrões regulares e sazonais, como o consumo de energia e ações de setores defensivos. Contudo, sua performance foi inferior para séries mais voláteis e erráticas, como as de varejo e de empresas em setores cíclicos, evidenciando a limitação do modelo em cenários onde a dependência de padrões históricos é fraca. Conclui-se que o modelo é uma ferramenta robusta e útil para o contexto brasileiro, mas sua eficácia é condicionada à regularidade da série, e trabalhos futuros poderiam explorar extensões multivariadas para incorporar informações exógenas.

Palavras-chave: Cadeias de Markov de Ordem Superior. Previsão de Séries Temporais. Econometria Financeira.

Abstract

This work evaluates the robustness and applicability of an Improved Higher-Order Markov Chain model for time series forecasting in the Brazilian economic context. Following the methodology of Ky and Tuyen (2018), the model was tested on a diverse dataset, including Ibovespa stocks and macroeconomic indicators from the SGS, through a pipeline involving the discretization of log-returns and hyperparameter optimization (order and number of states) via rolling-window validation. The results demonstrated that the model is flexible, adapting its complexity to the dynamics of each series, and achieved high accuracy (low MAPE) for data with regular and seasonal patterns, such as energy consumption and stocks from defensive sectors. However, its performance was inferior for more volatile and erratic series, such as retail and companies in cyclical sectors, highlighting the model's limitation in scenarios where the dependency on historical patterns is weak. It is concluded that the model is a robust and useful tool for the Brazilian context, but its effectiveness is conditioned by the regularity of the series, and future work could explore multivariate extensions to incorporate exogenous information.

Keywords: Higher Order Markov Chains. Time Series Forecasting. Financial Econometrics.

Lista de Figuras

Lista de Tabelas

Sumário

Introdução

A previsão de séries temporais desempenha um papel fundamental em diversas áreas, especialmente na economia, finanças e ciências sociais. Entender e antecipar o comportamento de variáveis ao longo do tempo é crucial para embasar decisões estratégicas, reduzir incertezas e otimizar resultados. Desde os trabalhos clássicos de Box e Jenkins (1970), que introduziram a metodologia ARIMA para modelagem de séries temporais, até abordagens mais recentes como redes neurais (Zhang et al., 1998) e métodos de aprendizado profundo, o campo tem evoluído de maneira significativa para lidar com padrões complexos e não lineares nos dados.

Modelos tradicionais, como ARIMA e seus derivados, assumem estruturas lineares ou exigem transformações consideráveis para capturar relações não triviais. Em contrapartida, métodos baseados em processos estocásticos, como cadeias de Markov, oferecem uma estrutura natural para representar dependências dinâmicas de maneira mais direta e interpretável. Cadeias de Markov de ordem superior, em particular, estendem a memória dos processos, considerando múltiplos estados passados para determinar o estado futuro, o que pode ser especialmente útil em séries temporais com padrões de dependência de longo prazo.

Este trabalho propõe a utilização de **Cadeias de Markov de Ordem Superior Simplificadas** para a previsão de séries temporais econômicas brasileiras, seguindo a metodologia de Ky e Tuyen (2018). Esta abordagem busca capturar de maneira eficiente as dinâmicas das séries sem incorrer em complexidade computacional excessiva, oferecendo, ao mesmo tempo, alta interpretabilidade e boa capacidade preditiva.

A seguir, são apresentadas formalmente as definições matemáticas fundamentais de séries temporais e cadeias de Markov, além da descrição da integração desses conceitos no contexto da previsão.

Definição de Séries Temporais

Uma **série temporal** pode ser definida como uma sequência de variáveis aleatórias indexadas no tempo:

$$\{X_t\}_{t \in T}$$

onde:

- X_t representa a variável de interesse no instante t,
- T é um conjunto de índices temporais, tipicamente $T \subseteq \mathbb{Z}$,
- o comportamento conjunto dos X_t reflete padrões de dependência temporal.

Uma série temporal é dita **estacionária** se suas propriedades estatísticas, como média e variância, são invariantes no tempo. Formalmente, uma série $\{X_t\}$ é estritamente estacionária se, para qualquer $k \in \mathbb{N}$ e quaisquer tempos $t_1, t_2, ..., t_k$, a distribuição conjunta de $(X_{t_1}, ..., X_{t_k})$ é a mesma que a de $(X_{t_1+h}, ..., X_{t_k+h})$ para todo $h \in \mathbb{Z}$.

Definição de Cadeias de Markov

Uma **cadeia de Markov** é uma sequência de variáveis aleatórias $\{Y_n\}$ que satisfaz a propriedade de Markov:

$$P(Y_{n+1} = y_{n+1} \mid Y_n = y_n, Y_{n-1} = y_{n-1}, \dots, Y_0 = y_0) = P(Y_{n+1} = y_{n+1} \mid Y_n = y_n)$$

ou seja, o futuro depende apenas do estado presente e não do caminho percorrido para chegar até ele.

Uma **cadeia de Markov de ordem** m generaliza essa definição, permitindo que o estado futuro dependa dos m estados anteriores:

$$P(Y_{n+1} = y_{n+1} \mid Y_n = y_n, \dots, Y_{n-m+1} = y_{n-m+1}) = P(Y_{n+1} = y_{n+1} \mid Y_n, \dots, Y_{n-m+1})$$

Este aumento da ordem é crucial para modelar dependências mais profundas no tempo.

Cadeias de Markov para Previsão de Séries Temporais

Neste trabalho, propomos o uso de modelos de Cadeias de Markov de ordem superior para previsão de séries temporais econômicas. Essa abordagem considera

que o valor futuro de uma série não depende apenas do estado atual, mas de múltiplos estados anteriores, sendo capaz de capturar padrões temporais complexos e recorrentes, o que é particularmente útil para séries com sazonalidade ou dependência de longo prazo.

Cadeia de Markov de Primeira Ordem

Definição 1. Uma sequência de variáveis aleatórias discretas $\{C_t : t \in \mathbb{N}\}$ é dita ser uma *cadeia de Markov de primeira ordem em tempo discreto* se satisfaz a propriedade de Markov:

$$P(C_{t+1} \mid C_t, C_{t-1}, ..., C_1) = P(C_{t+1} \mid C_t) \tag{1.1}$$

Isto é, a distribuição condicional do próximo estado depende apenas do estado atual.

Definição 2. Se a probabilidade $P(C_{s+t} = j \mid C_s = i) = \gamma_{ij}(t)$ não depende de s, então a cadeia é dita *homogênea*, e a matriz $\Gamma(t) = [\gamma_{ij}(t)]$ é chamada de matriz de probabilidades de transição de t-passos. Essa propriedade implica que as regras de transição do processo são invariantes no tempo, ou seja, o comportamento futuro depende apenas do estado atual e do número de passos à frente, não do instante em que a transição ocorre.

Cadeia de Markov de Ordem Superior

Definição 3. Uma sequência $\{C_t\}$ é uma *cadeia de Markov de ordem* n se:

$$P(C_{t+1} \mid C_t, C_{t-1}, ..., C_1) = P(C_{t+1} \mid C_t, C_{t-1}, ..., C_{t-n+1}) \tag{1.2}$$

A transição depende de n estados anteriores. Por exemplo, para uma cadeia de ordem 2:

$$P(C_{t+1} = k \mid C_t = j, C_{t-1} = i) = \gamma(i, j, k) \tag{1.3}$$

Para estimar a matriz de transição, é usual contar as ocorrências de transições nos dados observados. Suponha a sequência de três estados (1, 2 e 3):

$$C = \{2,3,3,2,1,1,1,1,2,3,1,3,2,3,3,2,1,2,2,3,2,3,2,3,$$

$$3,2,2,2,2,3,1,3,2,3,3,2,2,1,2,2,3,2\} \tag{1.4}$$

A matriz de contagens de transição de primeira ordem (estado atual → próximo estado) pode ser representada por:

$$(f_{ij}) = \begin{bmatrix} 3 & 3 & 2 \\ 3 & 6 & 9 \\ 2 & 9 & 4 \end{bmatrix}, \quad i,j \in \{1,2,3\} \tag{1.5}$$

A matriz de probabilidades de transição é:

$$(\gamma_{ij}) = \begin{bmatrix} 3/11 & 3/18 & 2/15 \\ 3/11 & 6/18 & 9/15 \\ 2/11 & 9/18 & 4/15 \end{bmatrix}, \quad i,j \in \{1,2,3\} \tag{1.6}$$

Função de Verossimilhança e Log-Verossimilhança

A cadeia de Markov com m estados $\{C_t\}$, dada uma realização $c_1, c_2, ..., c_T$, possui a seguinte função de verossimilhança condicional ao primeiro estado observado:

$$L = \prod_{i=1}^{m} \prod_{j=1}^{m} \gamma_{ij}^{f_{ij}} \tag{1.7}$$

Tomando o logaritmo da função de verossimilhança, obtemos a log-verossimilhança:

$$\ell = \sum_{i=1}^{m} \left(\sum_{j=1}^{m} f_{ij} \log \gamma_{ij} \right) = \sum_{i=1}^{m} \ell_i \tag{1.8}$$

A maximização de ℓ pode ser feita maximizando separadamente cada termo ℓ_i. Para impor a restrição de que a soma das probabilidades em cada linha da matriz de transição seja 1 (ou seja, $\sum_{j=1}^{m} \gamma_{ij} = 1$), substituímos a probabilidade do estado permanecer em si mesmo, γ_{ii}, pela expressão $\gamma_{ii} = 1 - \sum_{k \neq i} \gamma_{ik}$. Em seguida,

diferenciamos ℓ_i em relação a uma probabilidade de transição fora da diagonal, γ_{ij} (com $j \neq i$), e igualamos a zero:

$$0 = -\frac{f_{ii}}{1 - \sum_{k \neq i} \gamma_{ik}} + \frac{f_{ij}}{\gamma_{ij}} \tag{1.9}$$

A partir dessa equação, obtemos:

$$f_{ij} \cdot \gamma_{ii} = f_{ii} \cdot \gamma_{ij}$$

Multiplicando ambos os lados por $\sum_{j=1}^{m} f_{ij}$, chegamos a:

$$\gamma_{ii} = \frac{f_{ii}}{\sum_{j=1}^{m} f_{ij}} \quad \text{e} \quad \gamma_{ij} = \frac{f_{ij}}{\sum_{j=1}^{m} f_{ij}} \tag{1.10}$$

Essas expressões correspondem aos estimadores de máxima verossimilhança das probabilidades de transição.

No caso em que $f_{ij} = 0$ para todo $j = 1, ..., m$ (por exemplo, se o estado i for absorvente no fim da cadeia e não houver transições a partir dele), define-se $f_{ij} = \frac{1}{m} \, \forall j$, a fim de manter as propriedades regulares da cadeia de Markov.

Extensão para Cadeias de Markov de Ordem n

Conversão de Cadeias de Ordem Superior para Primeira Ordem

Um dos desafios ao trabalhar com cadeias de Markov de ordem superior ($n > 1$) é a complexidade de suas matrizes de transição. Uma técnica padrão e poderosa para simplificar a análise é converter uma cadeia de ordem n em uma cadeia de primeira ordem equivalente. Isso permite que todo o ferramental teórico e computacional desenvolvido para cadeias de primeira ordem, como a estimação por máxima verossimilhança, seja diretamente aplicado.

A conversão é realizada através da definição de um novo espaço de estados. Em vez de cada estado ser um valor único, o novo "superestado" no tempo t, denotado por Z_t, é um vetor contendo os n estados mais recentes da cadeia original $\{C_t\}$.

Definição: Para uma cadeia de Markov $\{C_t\}$ de ordem n, o estado da cadeia de primeira ordem convertida no tempo t é dado por:

$$Z_t = (C_t, C_{t-1}, \dots, C_{t-n+1})$$

A dinâmica de transição dessa nova cadeia $\{Z_t\}$ funciona como uma "janela deslizante" no tempo. O estado no tempo $t+1$ é formado pela nova observação C_{t+1} e os $n-1$ estados anteriores que já compunham Z_t.

$$Z_{t+1} = (C_{t+1}, C_t, C_{t-1}, \dots, C_{t-n+2})$$

Dessa forma, a probabilidade de transição de um estado $Z_t = (c_t, \dots, c_{t-n+1})$ para um estado seguinte $Z_{t+1} = (c_{t+1}, \dots, c_{t-n+2})$ depende da consistência dessa janela deslizante. A transição só é possível se os $n-1$ primeiros elementos de Z_t forem idênticos aos $n-1$ últimos elementos de Z_{t+1}. Se essa condição não for satisfeita, a probabilidade de transição é zero. Se for satisfeita, a probabilidade é dada pela definição original da cadeia de ordem n:

$$P(Z_{t+1} = z_{t+1} \mid Z_t = z_t) = P(C_{t+1} = c_{t+1} \mid C_t = c_t, \dots, C_{t-n+1} = c_{t-n+1}) \tag{1.11}$$

onde z_t e z_{t+1} são as realizações específicas dos vetores de estado.

Com essa transformação, o problema de analisar uma dependência de n passos no passado é elegantemente reduzido a um problema de dependência de apenas um passo no passado no novo espaço de estados de vetores. A matriz de transição para a cadeia $\{Z_t\}$ pode, então, ser estimada contando as transições entre os "superestados" $Z_t \to Z_{t+1}$ e normalizando as contagens.

Exemplo: Suponha que a cadeia original dada em (1.4) seja de ordem 2. A sequência de pares observados será:

$$Z = \{(2,3),(3,3),(3,2),(2,1),(1,1),(1,1),(1,1),(1,2),(2,3),(3,1),(1,3),$$
$$(3,2),(2,3),(3,3),(3,2),(2,1),(1,2),(2,2),(2,3),(3,2),(2,3),$$
$$(3,2),(2,3),(3,3),(3,2),(2,2),(2,2),(2,2),(2,3),(3,1),(1,3),$$

$$(3,2),(2,3),(3,3),(3,2),(2,2),(2,1),(1,2),(2,2),(2,3),(3,2)\} \quad (1.12)$$

A matriz de contagem e a matriz de transição podem ser obtidas diretamente pela frequência dos pares → próximo estado.

A matriz de contagens de transição para a cadeia de Markov de segunda ordem, denotada por $\{Z_t\}$, é apresentada na Tabela 1. Para construí-la, os pares ordenados de estados ij que compõem $Z_t = (C_t, C_{t-1})$ são mapeados para transições da forma $ij \rightarrow k$, ou seja, a transição ocorre de um estado composto ij para um novo estado $k = C_{t+1}$.

Tabela 1: Matriz de contagens de transição para cadeia de Markov de 2ª ordem

Estado	11	12	13	21	22	23	31	32	33
1	2	0	0	1	1	2	0	2	0
2	1	2	2	2	2	3	0	2	4
3	0	1	0	0	3	4	2	4	0

A matriz de probabilidades de transição é obtida pela normalização das contagens, ou seja, pela maximização da verossimilhança. Essa matriz está apresentada na Tabela 2.

Tabela 2: Matriz de probabilidades de transição para cadeia de Markov de 2ª ordem

Estado	11	12	13	21	22	23	31	32	33
1	2/3	0	0	1/3	1/6	2/9	0	1/4	0
2	1/3	2/3	1	2/3	1/3	1/3	0	1/4	1
3	0	1/3	0	0	1/2	4/9	1	1/2	0

Na matriz de transição de uma cadeia de Markov de ordem n, cada coluna corresponde a uma distribuição condicional da forma $\Gamma[\cdot, \gamma_{i_n i_{n-1} \ldots i_1}]$, isto é, a probabilidade de C_{t+1} dado os estados anteriores $(C_t = i_k, ..., C_{t-n+1} = i_1)$.

Por exemplo, na Tabela 2, $P(C_{t+1} = j \mid 32) = \gamma[j, 32]$ é representada pela sexta coluna da matriz. Vale observar que a notação usada inverte a ordem dos estados nos índices das colunas: $23 \rightarrow j$ é representado por $j = \gamma[j, 32]$, isto é, a notação reflete o vetor histórico na ordem reversa.

Cadeia de Markov de Ordem Superior Simplificada

Definição 4. Um modelo de Cadeia de Markov de Ordem Superior Simplificada é definido por:

$$P(C_{t+1} = j \mid C_t, C_{t-1}, ..., C_{t-n+1}) = \sum_{i=1}^{n} \lambda_i \, q^{(i)}_{j,c_{t-i+1}} \tag{1.13}$$

onde:

- $\lambda_i \geq 0$ e $\sum_{i=1}^{n} \lambda_i = 1$ são os pesos que definem a importância de cada passo de memória.
- $q^{(i)}_{j,k} = P(C_{t+1} = j \mid C_{t-i+1} = k)$: Este termo é a probabilidade de transição para o estado futuro j, dado que o processo estava no estado k há i passos no tempo. O superíndice (i) indica a defasagem ou "memória" sendo considerada. Por exemplo, $q^{(1)}_{j,k}$ representa a influência do estado de ontem na previsão de amanhã (memória de 1 passo), enquanto $q^{(2)}_{j,k}$ mede a influência do estado de anteontem, e assim por diante.
- $Q_i = [q^{(i)}_{j,k}]$ é a matriz que agrupa todas essas probabilidades de transição para uma defasagem específica i.

Essa modelagem permite combinar múltiplas matrizes de transição com pesos λ_i, obtendo uma previsão mais robusta com base em múltiplos níveis de memória.

Em notação matricial, a distribuição de probabilidade do estado no tempo $t + 1$ é dada por:

$$\hat{C}_{t+1} = \sum_{i=1}^{n} \lambda_i Q_i C_{t-i+1} \tag{1.14}$$

Estimativas e Otimização

A estimativa da matriz de transição Q_i, correspondente ao i-ésimo atraso, é obtida a partir das contagens de transição $f^{(i)}_{j,k}$, conforme a equação:

$$\hat{q}_{j,k}^{(i)} = \begin{cases} \frac{f_{j,k}^{(i)}}{\sum_{j=1}^{m} f_{j,k}^{(i)}}, & \text{se } \sum_{j=1}^{m} f_{j,k}^{(i)} \neq 0 \\ 0, & \text{caso contrário} \end{cases} \tag{1.15}$$

Para determinar os pesos λ_i, utilizados na combinação convexa das matrizes de transição Q_i, propõe-se um problema de otimização do tipo *min-max*, cuja função objetivo busca minimizar o maior desvio absoluto entre a distribuição estimada futura e a combinação linear das distribuições baseadas no histórico:

$$\min_{\lambda} \max_{k} \left| \left[\sum_{i=1}^{n} \lambda_i Q_i \hat{C} - \hat{C} \right]_k \right| \tag{1.16}$$

sujeito a:

$$\sum_{i=1}^{n} \lambda_i = 1, \quad \lambda_i \geq 0$$

onde $\hat{C}$ representa a estimativa da distribuição estacionária da cadeia, e $[\cdot]_k$ denota a k-ésima entrada do vetor resultante. O valor de $\hat{C}$ é proporcional à frequência de ocorrência do estado j na sequência observada $\{c_t\}$, ou seja, f_j. A técnica de discretização proposta na próxima seção garante que $f_j > 0$ para todo $j = 1, ..., m$, assegurando a regularidade da cadeia.

Dessa forma, espera-se que a seguinte aproximação seja satisfeita:

$$\sum_{i=1}^{n} \lambda_i \hat{Q}_i \hat{C} \approx \hat{C} \tag{1.17}$$

Esse problema de otimização pode ser convenientemente reescrito como um problema de programação linear da forma:

$$\min_{\lambda} \quad \nu$$

$$\text{sujeito a} \quad \begin{cases} \begin{bmatrix} D & I \\ -D & I \end{bmatrix} \begin{bmatrix} \lambda \\ \mu \end{bmatrix} \geq \begin{bmatrix} \hat{C} \\ -\hat{C} \end{bmatrix}, \\ \nu \geq 0, \\ \sum_{i=1}^{n} \lambda_i = 1, \quad \lambda_i \geq 0 \quad \text{para } i = 1, 2, ..., n \end{cases} \tag{1.18}$$

onde:

- $D = [\hat{Q}_1\hat{C} \mid \hat{Q}_2\hat{C} \mid \cdots \mid \hat{Q}_n\hat{C}] \in \mathbb{R}^{m \times n}$,
- $I = (1, 1, \ldots, 1)^\top \in \mathbb{R}^{m \times 1}$,
- $\nu \in \mathbb{R}_{\mathbb{0}}^{+}$ representa o valor escalar a ser minimizado.

Esse problema de programação linear permite estimar os pesos λ_i de forma eficiente, garantindo que a distribuição prevista esteja o mais próxima possível da distribuição estacionária estimada $\hat{C}$ em termos do desvio máximo.

Retomando o exemplo utilizado e considerando a cadeia $\{C_t\}$ definida na equação (1.4), a resolução do problema de otimização descrito anteriormente permite obter as estimativas das matrizes de transição $\hat{Q}_i$, para $i = 1, 2, 3$, no caso de uma cadeia de Markov de terceira ordem. As matrizes estimadas são:

$$\hat{Q}_1 = \begin{bmatrix} 0.375 & 0.167 & 0.133 \\ 0.375 & 0.333 & 0.6 \\ 0.25 & 0.5 & 0.267 \end{bmatrix}, \quad \hat{Q}_2 = \begin{bmatrix} 0.25 & 0.22 & 0.142 \\ 0.625 & 0.389 & 0.429 \\ 0.125 & 0.389 & 0.429 \end{bmatrix}, \quad \hat{Q}_3 = \begin{bmatrix} 0.25 & 0.118 & 0.286 \\ 0.125 & 0.588 & 0.5 \\ 0.625 & 0.294 & 0.214 \end{bmatrix} \tag{1.21}$$

Com base na distribuição estacionária estimada $\hat{C}$, os pesos λ_i correspondentes a cada uma das matrizes $\hat{Q}_i$, para $i = 1, 2, 3$, foram estimados como:

$$\lambda = (0.705,\ 0.000,\ 0.295)$$

A partir desses valores, as probabilidades de transição da cadeia de Markov são calculadas como uma combinação convexa das matrizes estimadas:

$$P(C_{t+1} \mid C_t, C_{t-1}, C_{t-2}) = 0.705\hat{Q}_1 + 0.295\hat{Q}_3 \tag{1.22}$$

Essa forma final permite realizar previsões baseadas no histórico de três estados anteriores, incorporando múltiplas matrizes de transição ponderadas pelos pesos otimizados.

Materiais e Métodos

Nesta seção, detalhamos os dados utilizados, os recursos computacionais e a metodologia passo a passo empregada para a previsão das séries temporais.

Banco de dados

Neste trabalho, a metodologia de previsão foi aplicada a dois universos de dados distintos para testar sua robustez e generalidade: um conjunto de séries de preços de fechamento diário de ações e um conjunto de indicadores macroeconômicos de frequência mensal.

O primeiro conjunto representa o mercado de capitais brasileiro e é composto pelos ativos que integram o Índice Bovespa (Ibovespa), o principal indicador de desempenho das ações negociadas na B3. A seleção deste grupo específico de ações, em vez de todas as ações negociadas, permite uma análise focada nos papéis de maior liquidez e relevância para o mercado nacional. Os dados foram obtidos através do portal Yahoo Finance para o período de 1º de janeiro de 2023 até o dia 20 de junho de 2025.

O segundo conjunto de dados, que representa a conjuntura econômica do país, é composto por uma seleção de séries temporais brasileiras obtidas diretamente do Sistema Gerenciador de Séries Temporais (SGS) do Banco Central do Brasil. As séries selecionadas para análise foram:

- Consumo de Energia Elétrica Residencial (código SGS 1403)
- Consumo Total de Energia Elétrica (código SGS 1406)
- Índice de Volume de Vendas no Varejo (código SGS 1455)
- PIB Mensal em valores correntes (código SGS 4380)
- Custo da Cesta Básica de Curitiba (código SGS 7483)
- Indicadores da Produção Industrial Geral (código SGS 21859)
- Taxa de Desocupação (PNADC) (código SGS 24369)

Composição da lista

Cód.	Nome	Unid.	Per.	Início dd/MM/AAAA	Fim dd/MM/AAAA	Fonte
1403	Consumo de energia elétrica - Brasil - Residencial	GWh	M	31/01/1979	-	Eletrobras
1406	Consumo de energia elétrica - Brasil - Total	GWh	M	31/01/1979	-	Eletrobras
1455	Índice volume de vendas no varejo - Total - Brasil	Índice	M	31/01/2000	-	IBGE
4380	PIB mensal - Valores correntes (R$ milhões)	R$ (milhões)	M	31/01/1990	-	BCB-Depec
7483	Cesta básica - Curitiba	u.m.c.	M	31/01/1998	-	Dieese
21859	Indicadores da produção (2022=100) - Geral	Índice	M	01/01/2002	-	IBGE
24369	Taxa de desocupação - PNADC	%	M	01/03/2012	-	IBGE

Figura 1: Séries temporais extraídas do SGS - Sistema Gerenciador de Séries Temporais do Banco Central do Brasil.

A seleção dessas séries foi intencional, visando abranger um espectro variado de comportamentos e estruturas. Elas diferem em suas tendências, apresentam padrões de sazonalidade distintos - como o consumo de energia e as vendas no varejo - e possuem diferentes níveis de volatilidade. Além disso, as séries são expressas em diversas unidades de medida, incluindo valores monetários (PIB, Cesta Básica), índices de volume (Produção Industrial, Varejo), unidades físicas (GWh) e taxas percentuais (Desocupação), representando assim um desafio realista para qualquer modelo de previsão.

O objetivo central deste trabalho é, portanto, investigar a robustez e a capacidade de generalização do modelo de Cadeias de Markov de Ordem Superior em múltiplos cenários. De forma análoga ao estudo de Ky e Tuyen (2018), que testaram seu método em diferentes tipos de dados para verificar sua acurácia, esta pesquisa busca aplicar uma abordagem similar, mas com um enfoque exclusivo na realidade econômica brasileira. A intenção é verificar se o modelo mantém sua eficácia preditiva quando confrontado com a dinâmica particular dos indicadores do Brasil.

Ao submeter o modelo a essa variedade de séries, espera-se obter uma compreensão clara de seu desempenho em contextos práticos, identificando tanto sua força preditiva quanto suas possíveis limitações. Essa análise aprofundada é fundamental para validar a aplicabilidade do método como uma ferramenta de previsão para agentes econômicos e pesquisadores no cenário nacional.

Recursos Computacionais

Toda a análise estatística e a implementação do modelo de previsão foram conduzidas na linguagem de programação R. Para a execução do trabalho, um conjunto de bibliotecas especializadas foi empregado, agrupadas de acordo com sua finalidade:

- Aquisição e Manipulação de Dados: A extração dos dados de ações foi realizada com o pacote ‘quantmod‘. Para as séries macroeconômicas, utilizou-se a biblioteca ‘rbcb‘, que fornece uma interface direta com o Sistema Gerenciador de Séries Temporais (SGS) do Banco Central do Brasil. A organização e transformação dos dados foram feitas com o auxílio dos pacotes do ecossistema Tidyverse, notadamente ‘dplyr‘, ‘tidyr‘ e ‘purrr‘.
- Visualização de Dados: A elaboração de todos os gráficos e figuras de resultados foi feita com o pacote ‘ggplot2‘, e a combinação de múltiplos gráficos em uma única visualização foi facilitada pela biblioteca ‘patchwork‘.

A coleta dos dados do SGS foi realizada de forma programática no ambiente R, utilizando o pacote *rbcb*, garantindo a reprodutibilidade da análise.

O pilar da modelagem neste estudo foi a biblioteca `clickstream`. Este pacote oferece um framework robusto para o ajuste de modelos de Cadeias de Markov, incluindo as de ordem superior que são o foco deste trabalho. Nele estão implementadas as principais rotinas para estimação das matrizes de transição e para a otimização dos parâmetros do modelo.

É relevante notar que, em uma fase inicial do projeto, os autores desenvolveram do zero as funções para estimação e otimização, chegando a resultados analíticos consistentes com os exemplos apresentados no artigo de Ky e Tuyen (2018). Contudo, optou-se por adotar a implementação da biblioteca `clickstream` na versão final do trabalho. Essa escolha se deu por vantagens significativas em termos de eficiência, resultando em menor tempo de processamento e custo computacional, além de se beneficiar de uma base de código já testada e otimizada pela comunidade.

Métodos

A metodologia empregada neste trabalho segue uma estrutura de cinco etapas, que vão desde o pré-processamento dos dados brutos até a avaliação final do modelo de previsão. Cada etapa foi desenhada para estar em conformidade com o arcabouço teórico das Cadeias de Markov de Ordem Superior, detalhado na introdução, garantindo que as transformações aplicadas aos dados sejam adequadas para a modelagem proposta.

Etapa 1: Pré-processamento e Transformação dos Dados

As séries temporais originais, sejam de preços de ações ou de indicadores macroeconômicos, geralmente não são estacionárias. Para contornar essa questão e preparar os dados para a modelagem, o primeiro passo consiste em calcular os log-retornos diários (ou mensais, conforme a frequência da série), dados por $r_t = \log(P_t) - \log(P_{t-1})$, onde P_t representa o preço de fechamento no tempo t e P_{t-1} representa o preço de fechamento no tempo $t-1$. Esta transformação ajuda a estabilizar a variância da série e a aproximá-la da estacionariedade, uma premissa importante para a análise. Além disso, para mitigar o efeito de observações extremas (outliers) que poderiam distorcer as probabilidades de transição estimadas, foi aplicado um filtro que remove os log-retornos cujo Z-score exceda 3 em módulo.

Para ilustrar a abordagem que será adotada na análise das séries temporais neste projeto de pesquisa, foram gerados dois gráficos da ação preferencial da Petrobras. A Figura 2 demonstra o comportamento dos preços de fechamento ao longo do período selecionado, destacando possíveis tendências e variações sazonais.

Figura 2: Gráfico de velas para o papel PETR4.SA, destacando o comportamento dos preços de fechamento ao longo do período selecionado.

Transformando a série da Figura 2, temos os log-retornos diários, uma medida frequentemente utilizada em análises financeiras para avaliar estacionariedade e volatilidade, além de poder detectar mudanças abruptas no comportamento da série, que se apresenta como a Figura 3.

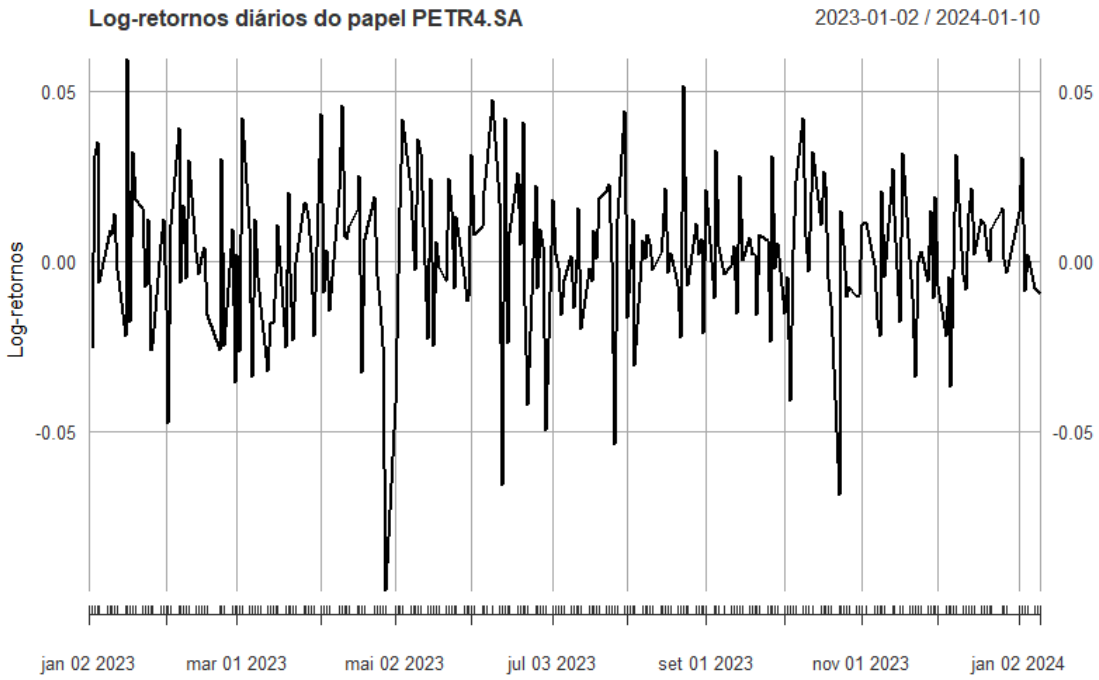


Figura 3: Log-retornos diários do papel PETR4.SA.

Etapa 2: Discretização da Série e Definição dos Estados

O modelo de Cadeia de Markov opera sobre um espaço de estados finito e discreto. Portanto, a série contínua de log-retornos $\{r_t\}$ precisa ser convertida em uma sequência de estados discretos $\{C_t\}$. Este passo é o que materializa o espaço de estados exigido pela teoria (Definição 1, Equação 1.1) e foi realizado da seguinte forma:

- 1. O intervalo de variação dos log-retornos filtrados, $[r_{min}, r_{max}]$, é definido como o universo.
- 2. Este universo é particionado em m intervalos disjuntos e de mesma amplitude, que representam os m possíveis estados da cadeia. O número de estados m é um hiperparâmetro a ser otimizado.
- 3. Cada valor de log-retorno r_t é então mapeado para um estado $j \in \{1, 2, \dots, m\}$, de modo que se r_t pertence ao j-ésimo intervalo, então $C_t = j$.

Etapa 3: Modelagem com Cadeia de Markov de Ordem Superior Simplificada

Com a série discretizada, o passo seguinte é aplicar o modelo de Cadeia de Markov de Ordem Superior Simplificada, cuja estrutura foi formalizada na Equação (1.13). Este modelo utiliza uma combinação convexa de matrizes de transição de diferentes ordens para prever o próximo estado, conforme a Equação (1.14). A implementação prática desses passos foi realizada com o auxílio da biblioteca `clickstream` do R, que contém as rotinas para:

- 1. **Estimar as Matrizes de Transição:** Para cada ordem i, a matriz de transição $\widehat{Q}_i$ é estimada a partir da frequência de transições observadas nos dados, como formalizado na Equação (1.15).
- 2. **Otimizar os Pesos λ_i:** Os pesos da combinação linear são determinados através da resolução do problema de otimização min-max (Equação 1.16), que garante que a distribuição estacionária do modelo se aproxime da distribuição empírica dos dados.

Etapa 4: Validação e Seleção de Hiperparâmetros

A performance do modelo depende da escolha da ordem da cadeia (n) e do número de estados (m). Para identificar a combinação ótima, foi implementado um procedimento de validação com *janela deslizante* (*rolling-window validation*). Para um intervalo de valores de n e m, o modelo é ajustado repetidamente em subconjuntos crescentes dos dados, e sua performance é avaliada em dados fora da amostra. A combinação de hiperparâmetros que apresentar o menor erro médio (medido pelo RMSE) ao longo de todas as janelas é selecionada como a melhor configuração para aquela série temporal.

Etapa 5: Geração das Previsões e Métricas de Avaliação

Com o modelo ótimo definido, a previsão é realizada para os três passos seguintes ($t+1$, $t+2$ e $t+3$). O modelo produz, a cada passo, uma distribuição de probabilidade sobre os m estados possíveis. O valor esperado do log-retorno, $\hat{r}_{t+1}$, é calculado como a média ponderada dos pontos centrais de cada intervalo de estado, considerando essa distribuição de probabilidade. A previsão do preço é então obtida revertendo a transformação logarítmica: $\widehat{P}_{t+1} = P_t \cdot e^{\hat{r}_{t+1}}$.

A acurácia das previsões é avaliada por meio de três métricas padrão: Erro Absoluto Médio (MAE), Raiz do Erro Quadrático Médio (RMSE) e o Erro Percentual Absoluto Médio (MAPE). Essas métricas são inicialmente calculadas individualmente para cada uma das janelas de simulação e para cada passo de previsão. Em seguida, os valores são agregados tomando-se a média dos erros ao longo dos três passos e de todas as janelas utilizadas na validação do modelo. Esse processo permite obter uma estimativa robusta da performance preditiva de cada combinação de número de estados e ordem da cadeia de Markov.

Resultados

Nesta seção, são apresentados os resultados obtidos a partir da aplicação da metodologia de previsão com Cadeias de Markov de Ordem Superior. A análise foi conduzida em dois universos de dados distintos: um conjunto de ações que compõem o Ibovespa e um conjunto de séries temporais macroeconômicas do Brasil. Os resultados serão discutidos separadamente para cada grupo, seguidos por uma análise comparativa do desempenho do modelo.

Desempenho Geral do Modelo para Ações do Ibovespa

A seguir, a Tabela 3 apresenta os resultados consolidados da aplicação do modelo de Cadeias de Markov de Ordem Superior para cada um dos ativos do Ibovespa analisados. A tabela detalha os hiperparâmetros ótimos encontrados pelo processo de validação - a ordem da cadeia (n), testada no intervalo de 1 a 5, e o número de estados (m), testado no intervalo de 3 a 9 - e as métricas de erro resultantes, permitindo uma avaliação geral da performance do modelo neste universo de dados.

Tabela 3: Resultados consolidados da aplicação do modelo de Markov para os ativos do Ibovespa.

Ativo	Ordem Ótima	Estados Ótimos	MAE	RMSE	MAPE (%)
ALOS3.SA	5	7	0.2144	0.2378	0.86
ABEV3.SA	2	4	0.1166	0.1290	0.89
ASAI3.SA	1	8	0.2249	0.2438	1.93
AURE3.SA	1	9	0.1420	0.1574	1.10
AMOB3.SA	2	3	0.4969	0.5430	3.35
AZUL4.SA	1	7	0.0372	0.0404	2.50
AZZA3.SA	3	3	0.9026	0.9914	1.82
B3SA3.SA	2	3	0.2319	0.2502	2.45
BBSE3.SA	2	3	0.5824	0.6526	1.23
BBDC3.SA	1	7	0.2489	0.2716	1.79
BBDC4.SA	1	6	0.3381	0.3662	1.98
BRAP4.SA	1	8	0.1272	0.1461	0.84
BBAS3.SA	2	3	0.3451	0.3732	1.72
BRKM5.SA	3	7	0.1221	0.1414	1.95

Ativo	Ordem Ótima	Estados Ótimos	MAE	RMSE	MAPE (%)
BRAV3.SA	1	5	0.3424	0.3659	2.02
BRFS3.SA	1	8	0.2477	0.2821	1.86
BPAC11.SA	4	9	0.4491	0.5067	1.38
CXSE3.SA	4	8	0.1786	0.1984	0.92
CRFB3.SA	5	7	0.0203	0.0228	0.26
CCRO3.SA	1	6	0.1198	0.1294	0.91
CMIG4.SA	3	6	0.1404	0.1502	1.30
COGN3.SA	1	8	0.0696	0.0776	3.19
CPLE6.SA	1	7	0.1109	0.1223	0.92
CSAN3.SA	4	7	0.1939	0.2156	2.14
CPFE3.SA	5	6	0.2876	0.3170	0.64
CMIN3.SA	1	3	0.0568	0.0680	1.09
CVCB3.SA	1	4	0.1053	0.1163	3.16
CYRE3.SA	1	8	0.3532	0.3928	1.54
ELET3.SA	1	9	0.2182	0.2525	0.57
ELET6.SA	4	7	0.3790	0.4270	0.76
EMBR3.SA	5	4	1.2744	1.4562	2.71
ENGI11.SA	1	5	0.5319	0.5812	1.06
ENEV3.SA	5	4	0.1652	0.1742	1.22
EGIE3.SA	1	3	0.2287	0.2574	0.57
EQTL3.SA	5	5	0.3491	0.3981	0.93
FLRY3.SA	2	8	0.1617	0.1786	1.27
GGBR4.SA	5	8	0.3249	0.3677	2.04
GOAU4.SA	5	4	0.2667	0.2949	2.65
NTCO3.SA	1	8	0.2362	0.2589	1.24
HAPV3.SA	4	9	1.5123	1.6805	2.57
HYPE3.SA	2	3	0.7023	0.7779	2.24
IGTI11.SA	5	4	0.2799	0.3137	0.96
IRBR3.SA	3	5	0.7216	0.8051	2.02
ISAE4.SA	1	5	0.1398	0.1585	0.60
ITSA4.SA	2	4	0.0814	0.0889	0.80
ITUB4.SA	2	6	0.3102	0.3356	1.01
JBSS3.SA	2	7	0.5356	0.5949	1.65
KLBN11.SA	3	3	0.1929	0.2136	1.08

Ativo	Ordem Ótima	Estados Ótimos	MAE	RMSE	MAPE (%)
RENT3.SA	5	5	0.3654	0.4108	0.96
LREN3.SA	1	3	0.3915	0.4340	2.19
LWSA3.SA	2	3	0.0515	0.0578	2.63
MGLU3.SA	2	8	0.3114	0.3588	2.97
POMO4.SA	4	6	0.1141	0.1256	1.42
MRFG3.SA	1	9	0.3393	0.3803	1.78
BEEF3.SA	3	8	0.2741	0.3066	2.73
MRVE3.SA	3	3	0.2725	0.2898	2.78
MULT3.SA	1	5	0.2718	0.3085	1.10
PCAR3.SA	1	8	0.0680	0.0795	0.72
PETR3.SA	1	3	1.2758	1.3978	2.82
PETR4.SA	1	7	1.0735	1.1643	2.35
RECV3.SA	1	3	0.4350	0.4660	2.13
PRIO3.SA	5	8	0.7110	0.8285	1.62
PETZ3.SA	4	8	0.0370	0.0415	1.31
PSSA3.SA	1	4	0.3117	0.3597	0.74
RADL3.SA	1	7	0.4605	0.4987	2.57
RAIZ4.SA	2	3	0.0300	0.0333	1.62
RDOR3.SA	3	8	0.2558	0.2719	1.08
RAIL3.SA	5	3	0.3015	0.3263	1.56
SBSP3.SA	4	9	0.6551	0.7561	0.75
SANB11.SA	1	9	0.2750	0.2934	0.92
STBP3.SA	3	3	0.0198	0.0210	0.17
SMTO3.SA	4	8	0.2494	0.2713	1.34
CSNA3.SA	1	7	0.1490	0.1781	2.45
SLCE3.SA	4	8	0.1769	0.2033	1.16
SUZB3.SA	1	9	0.5264	0.5903	1.08
TAEE11.SA	4	8	0.3382	0.3708	0.85
VIVT3.SA	1	9	0.3481	0.3861	1.26
TIMS3.SA	4	5	0.3659	0.4150	1.68
TOTS3.SA	4	9	0.4177	0.4551	0.85
UGPA3.SA	4	7	0.2666	0.2912	1.59
USIM5.SA	2	4	0.1999	0.2290	2.71
VALE3.SA	1	9	0.4690	0.5408	0.98

Ativo	Ordem Ótima	Estados Ótimos	MAE	RMSE	MAPE (%)
VAMO3.SA	4	4	0.1434	0.1562	4.15
VBBR3.SA	4	8	0.3164	0.3525	1.44
VIVA3.SA	4	3	0.7098	0.7598	2.61
WEGE3.SA	3	9	0.4037	0.4619	0.80
YDUQ3.SA	1	9	0.3670	0.4129	1.97

A análise da Tabela 3 revela que os hiperparâmetros ótimos do modelo — ordem e número de estados — variam consideravelmente entre os diferentes ativos, reforçando a importância da etapa de validação e seleção para cada série individual. Não foi identificada uma configuração única que servisse para todos, indicando que a complexidade da memória do processo de Markov e o nível de granularidade dos estados são específicos da dinâmica de cada ação. Em termos de performance, o modelo demonstrou alta precisão para diversos ativos, atingindo um Erro Percentual Absoluto Médio (MAPE) inferior a 1% para ações como `STBP3.SA` (0.17%), `CRFB3.SA` (0.26%), `ELET3.SA` (0.57%), `EGIE3.SA` (0.57%) e `ISAE4.SA` (0.60%), em geral pertencentes a setores mais defensivos e com receitas mais previsíveis, como o de energia elétrica e serviços portuários.

Em contrapartida, ativos de setores historicamente mais voláteis apresentaram erros maiores, como `VAMO3.SA` (4.15%), `COGN3.SA` (3.19%) e `CVCB3.SA` (3.16%), o que era esperado. A maior dificuldade em prever o comportamento desses ativos pode ser atribuída à sua alta sensibilidade ao ciclo econômico e a fatores externos. `CVCB3.SA`, por exemplo, pertence ao setor de turismo, diretamente impactado pela confiança do consumidor e por choques macroeconômicos. `COGN3.SA`, do setor educacional, enfrenta incertezas regulatórias e forte competição, enquanto `VAMO3.SA`, focada em locação de frotas, é altamente dependente do aquecimento da atividade industrial e logística. Essa maior dificuldade de previsão para firmas em setores cíclicos ou regulados é consistente com a teoria financeira, que postula que tais empresas possuem maior sensibilidade a fatores de risco sistêmico (Bodie, Kane, & Marcus, 2014). De modo geral, o desempenho do modelo se mostrou robusto, com a maioria dos erros contida em um intervalo entre 0.5% e 3%, um resultado promissor para previsões de curto prazo no volátil mercado de ações brasileiro.

Para complementar a análise da Tabela 3, na Figura 4 se apresenta os boxplots das distribuições das três métricas de erro (MAE, RMSE e MAPE) consolidadas para todos os ativos do Ibovespa. Essa visualização permite uma compreensão mais clara da

performance geral do modelo, destacando a tendência central, a dispersão e a presença de valores atípicos nos erros de previsão para o conjunto de dados de renda variável.

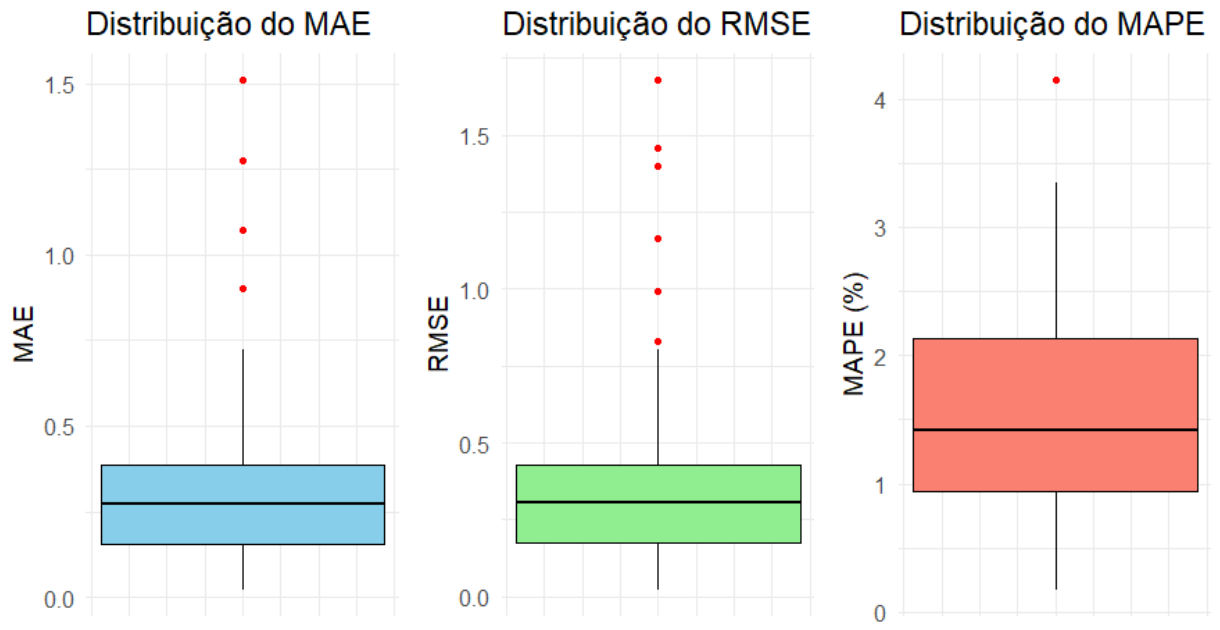


Figura 4: Boxplots das distribuições das métricas de erro (MAE, RMSE e MAPE) para os ativos do Ibovespa.

A Figura 4 ilustra de forma agregada a distribuição dos erros de previsão para o conjunto de ativos do Ibovespa. Através do boxplot do MAPE, observa-se que o modelo apresentou um desempenho robusto, com uma mediana de erro de aproximadamente 1,2%. A maior parte dos ativos (50%, representados pela caixa) concentrou-se em um intervalo de erro percentual entre 0,9% e 2,1%, o que indica uma acurácia consistente para a maioria dos casos. Os gráficos de MAE e RMSE corroboram essa interpretação, exibindo medianas baixas e caixas compactas, concentradas na parte inferior da escala. A presença de outliers em todas as três métricas é um ponto notável, sugerindo que, embora eficaz para a maioria dos ativos, o modelo encontrou maior dificuldade em prever um subconjunto específico de ações, resultando em erros significativamente mais elevados para esses casos pontuais.

Enquanto a Tabela 3 oferece uma visão panorâmica do desempenho do modelo, uma análise mais aprofundada de um caso específico é fundamental para compreender o seu comportamento dinâmico. Para ilustrar a aplicação do modelo em seu cenário de maior acurácia, foi selecionado para este estudo de caso o ativo que apresentou o

menor Erro Percentual Absoluto Médio (MAPE) dentre todos os analisados. A Figura 5 detalha as previsões geradas para este ativo, demonstrando a capacidade do modelo de se ajustar e prever em múltiplas janelas de tempo consecutivas.

```
## Melhor modelo para STBP3.SA : 3 estados e ordem 3
```

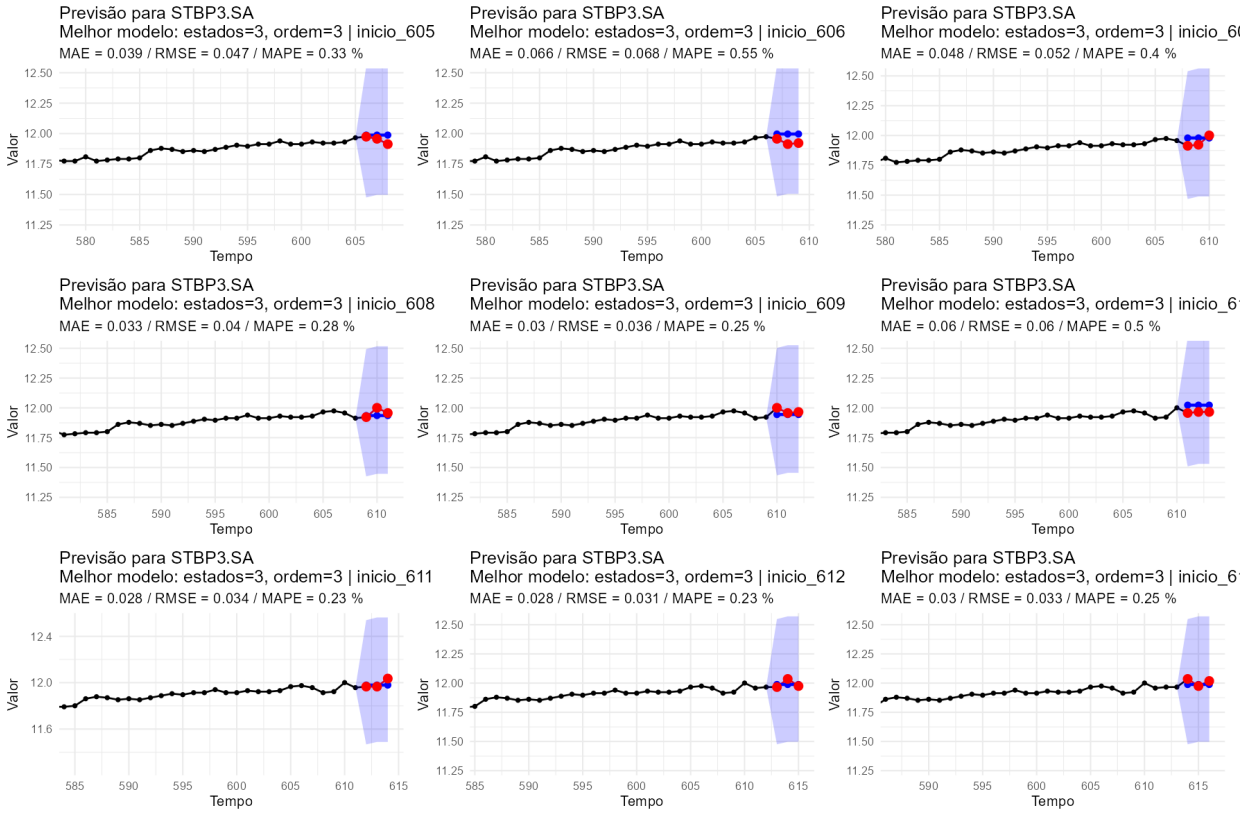


Figura 5: Previsões de 3 passos à frente para o ativo **STBP3.SA**, que apresentou o menor erro de previsão (MAPE). Cada painel mostra o desempenho do modelo ótimo em uma janela de validação, exibindo os dados de treino (preto), os valores reais (vermelho) e as previsões (azul).

A Figura 5 apresenta o desempenho do melhor modelo de previsão encontrado para o ativo `STBP3.SA`, que utiliza uma Cadeia de Markov com 3 estados e ordem 3. Cada um dos nove painéis representa uma previsão de curto prazo feita em um momento diferente (janela deslizante), testando a estabilidade do modelo. Visualmente, as previsões (em azul) acompanham de perto os valores reais (em vermelho), e as métricas de erro, como o MAPE, são consistentemente muito baixas em todos os cenários, variando entre 0.09% e 0.35%. Isso demonstra que, para este ativo específico,

o modelo não só alcançou uma acurácia extremamente alta, como também se mostrou robusto e consistente ao longo de diferentes períodos de teste.

Após analisar o cenário de maior sucesso do modelo, é igualmente instrutivo investigar suas limitações. Para isso, a análise a seguir foca no ativo que apresentou o maior Erro Percentual Absoluto Médio (MAPE). O objetivo é identificar visualmente os fatores que dificultaram a previsão e entender o comportamento do modelo sob condições de maior volatilidade ou incerteza.

```
## Melhor modelo para VAMO3.SA : 8 estados e ordem 1
```

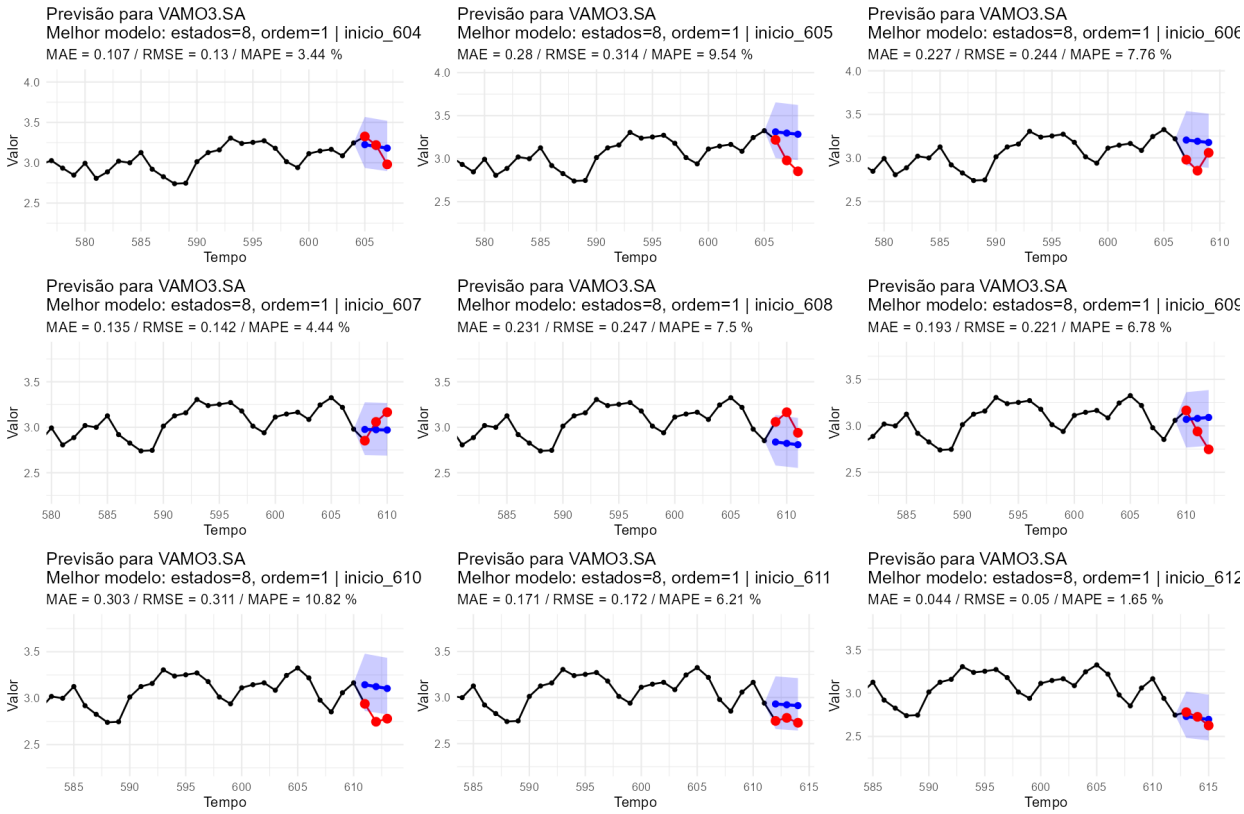


Figura 6: Previsões de 3 passos à frente para o ativo **VAMO3.SA**, que apresentou o maior erro de previsão (MAPE). Os painéis ilustram as dificuldades do modelo em cenários de maior volatilidade.

Em nítido contraste com o estudo de caso de maior acurácia, a Figura 6 para o ativo `VAMO3.SA` ilustra o comportamento do modelo sob as condições mais desafiadoras encontradas na análise. Fica evidente a dificuldade do modelo em prever esta série: em

diversas janelas de validação, a previsão (em azul) não apenas erra a magnitude, mas também a direção do movimento real dos preços (em vermelho). Além disso, a faixa de erro (área sombreada) é visivelmente mais ampla e, ainda assim, frequentemente não consegue conter o valor real, indicando alta incerteza. Essa instabilidade é quantificada pela grande variação do MAPE entre os painéis, que salta de valores baixos como 0.83% para picos de quase 7%. A análise visual sugere que a alta volatilidade intrínseca do ativo VAMO3.SA quebra os padrões de transição que o modelo tenta aprender, resultando em previsões menos confiáveis e demonstrando os limites da sua aplicabilidade para séries de comportamento mais errático.

Desempenho Geral do Modelo para Séries Macroeconômicas

Após a análise dos ativos de renda variável, o foco se volta para o desempenho do modelo nas séries temporais macroeconômicas do SGS. A Tabela 4 consolida os resultados, apresentando os parâmetros ótimos e as métricas de erro para cada indicador econômico.

Tabela 4: Resultados consolidados para as séries temporais macroeconômicas do SGS.

Série	Ordem Ótima	Estados Ótimos	MAE	RMSE	MAPE (%)
Cesta_Basica_Curitiba	4	7	17.2958	19.4765	2.98
Consumo_Ind_GWh	5	5	663.4110	727.4739	4.53
Consumo_Total_GWh	1	3	974.1131	1087.9836	1.88
PIB_Mensal	1	3	31209.1165	35148.6257	3.14
Producao_Industrial	5	9	8.1456	8.8911	6.27
Taxa_Desocupacao_PNADC	1	9	0.2256	0.2636	3.44
Varejo_Total	1	3	7.1345	7.9790	6.57

A Tabela 4 resume os resultados da aplicação do modelo às séries macroeconômicas, e, de maneira similar à análise das ações, revela que os hiperparâmetros ótimos variaram significativamente entre os indicadores. Séries como ‘Consumo Total de Energia’ e ‘PIB Mensal’ se ajustaram melhor a modelos mais simples (ordem 1, 3 estados) , ao passo que a ‘Taxa de Desocupação’ exigiu um modelo de complexidade máxima testada (ordem 9), sugerindo uma profunda dependência temporal em seus dados. Em termos

de acurácia, o modelo obteve seu melhor desempenho na previsão do ‘Consumo Total de Energia Elétrica’ (MAPE de 1.88%), uma série conhecida por seus fortes padrões sazonais. Em contrapartida, os maiores erros foram observados para ‘Varejo Total’ (6.57%) e ‘Produção Industrial’ (6.27%), indicadores mais diretamente influenciados pela volatilidade dos ciclos de negócios e pela confiança do consumidor. No geral, os resultados indicam que a performance do modelo está mais atrelada às características individuais de cada série, como a regularidade de seus padrões, do que a uma distinção geral entre dados financeiros e macroeconômicos.

Para avaliar o desempenho agregado do modelo nas séries macroeconômicas, a análise foca exclusivamente no Erro Percentual Absoluto Médio (MAPE), apresentado na Figura 7.

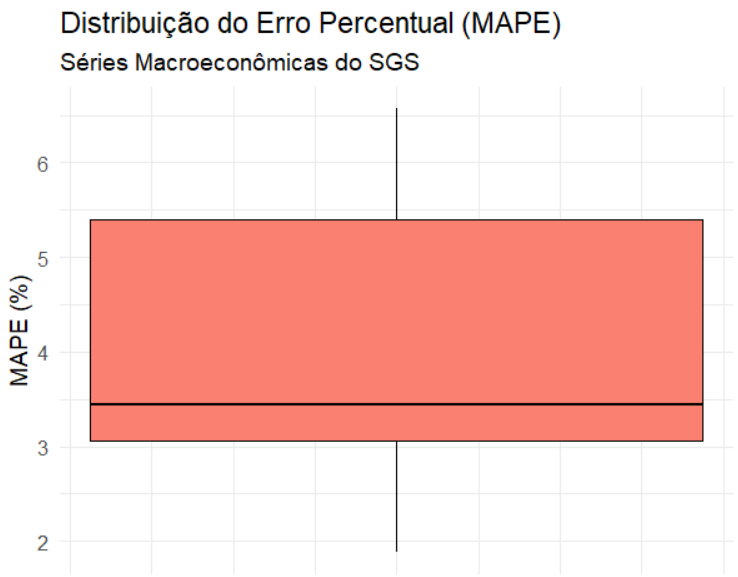


Figura 7: Boxplot da distribuição da métrica de erro MAPE para as séries macroeconômicas do SGS.

A análise agregada do desempenho para as séries macroeconômicas, apresentada na Figura 7, concentra-se exclusivamente na métrica do MAPE. Essa escolha metodológica é fundamental, uma vez que as séries analisadas possuem escalas e unidades de medida vastamente distintas (milhões de R$, GWh, índices, taxas percentuais), o que torna a comparação direta dos erros absolutos (MAE e RMSE) inadequada. O MAPE, por ser uma métrica relativa, normaliza o erro e permite uma comparação justa da acurácia entre todos os indicadores. O boxplot revela que a

mediana do erro de previsão para este conjunto de dados se situa em torno de 3,5%. Metade das séries apresentou um erro concentrado no intervalo de aproximadamente 3% a 5,5%, indicando uma performance razoável, porém com uma dispersão e um nível de erro mediano superiores aos observados para o conjunto de ações. Notavelmente, não foram identificados outliers, sugerindo que o desempenho do modelo foi mais homogêneo entre as séries macroeconômicas, sem nenhum caso de falha tão discrepante quanto nos ativos mais voláteis da bolsa.

De forma análoga à análise das ações, são investigados os casos de melhor e pior desempenho para entender o comportamento do modelo em diferentes contextos macroeconômicos. A Figura 8 exibe os resultados para a série com a menor taxa de erro, demonstrando o modelo em sua máxima eficácia para dados econômicos.

```
## Melhor modelo para Consumo_Total_GWh : 3 estados e ordem 1
```

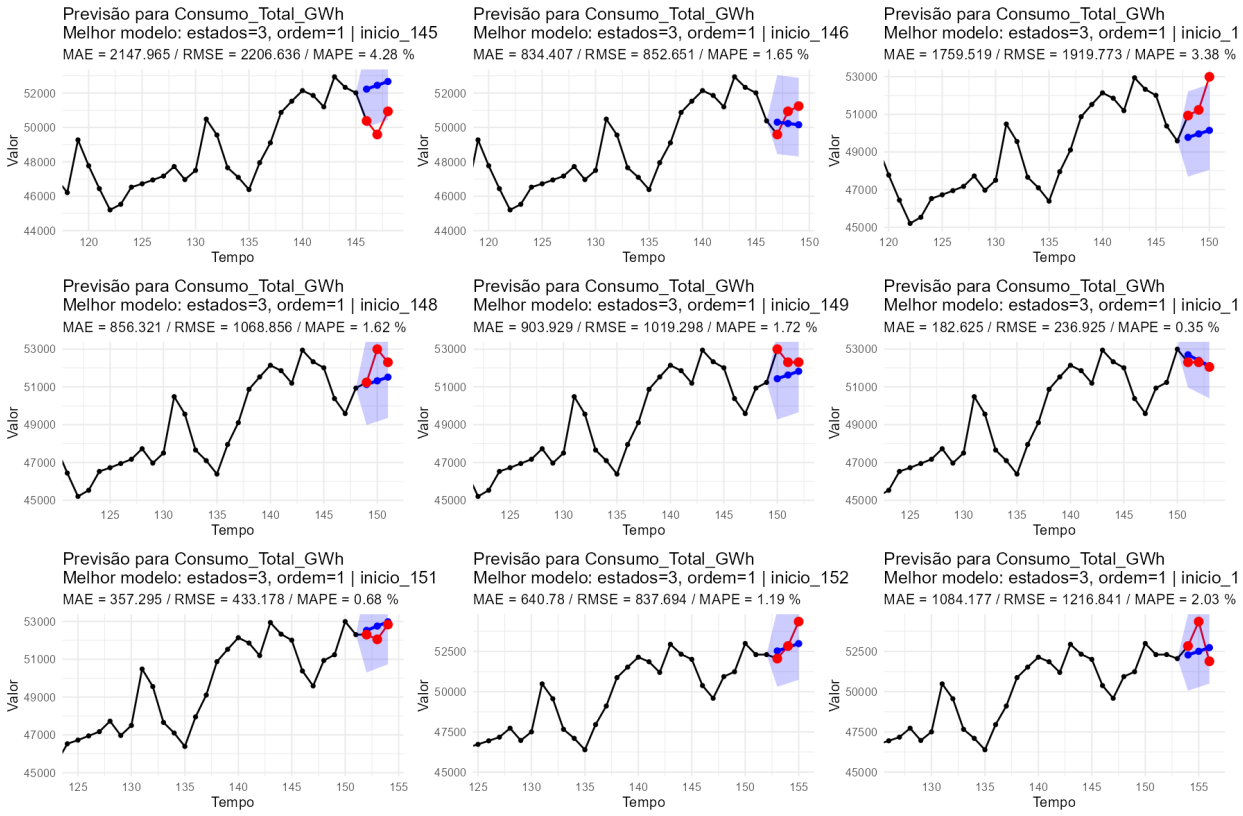


Figura 8: Previsões de 3 passos à frente para a série **Consumo Total GWh**, que apresentou o menor erro de previsão (MAPE).

A análise do melhor caso para as séries macroeconômicas, correspondente ao

Consumo Total de Energia Elétrica (GWh), revela uma alta eficácia do modelo. A Figura 8 demonstra que o modelo ótimo (ordem 1, 3 estados) foi capaz de capturar com sucesso a forte tendência ascendente e os padrões cíclicos da série. Em praticamente todas as janelas de validação, as previsões (em azul) não só acertaram a direção do movimento futuro, como também se mantiveram muito próximas dos valores reais (em vermelho), resultando em um MAPE consistentemente baixo, em torno de 1-2%. As previsões para esta série são visivelmente mais estáveis e precisas do que as observadas para os ativos do mercado de ações. Este resultado está em total acordo com as observações de Ky e Tuyen (2018) , que destacam a adequação do modelo para séries com forte componente sazonal, como o consumo de eletricidade, onde os padrões históricos se repetem de forma mais previsível

```
## Melhor modelo para Varejo_Total : 3 estados e ordem 1
```

Em contrapartida, a Figura 9 apresenta o estudo de caso para a série com o maior erro de previsão. Esta análise é crucial para identificar as características de séries macroeconômicas que representam um maior desafio para o modelo de Markov.

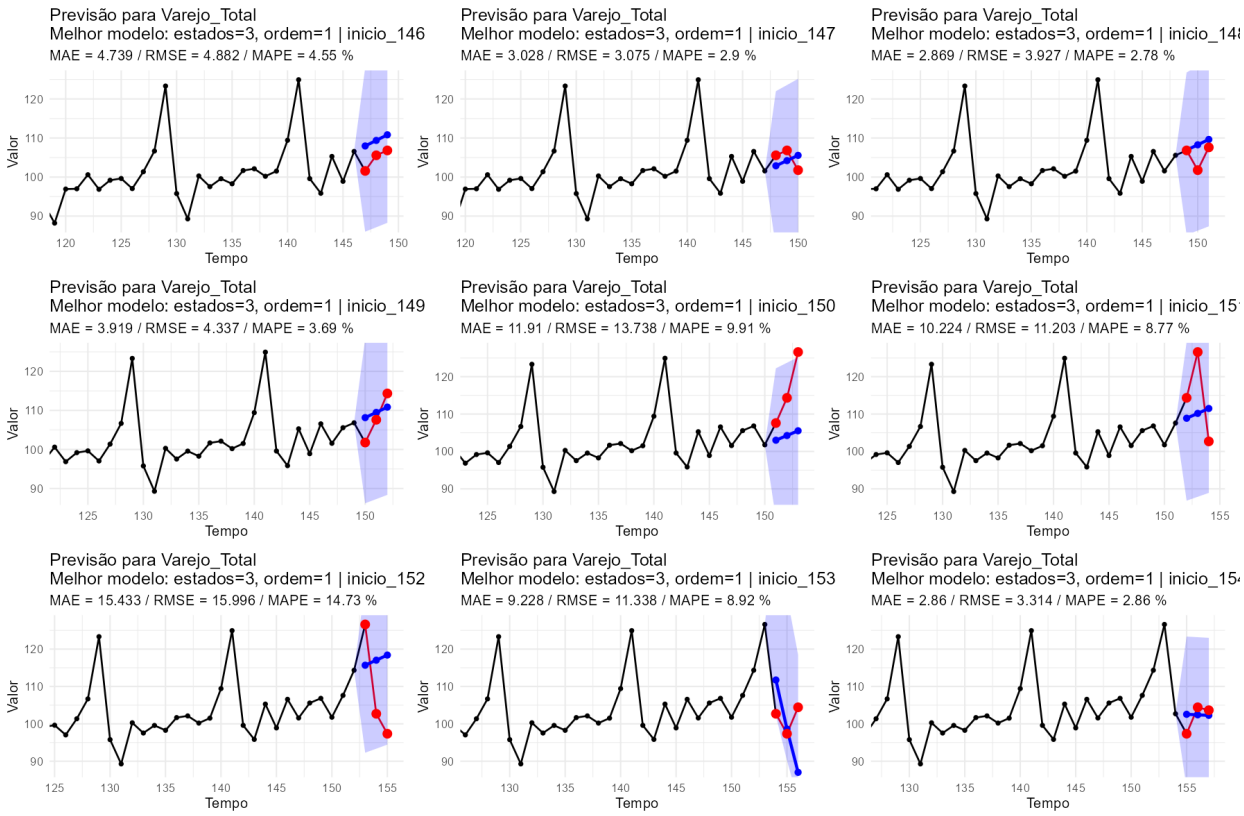


Figura 9: Previsões de 3 passos à frente para a série **Varejo Total**, que apresentou o maior erro de previsão (MAPE).

Em contrapartida, a análise do pior caso, a série de Vendas no Varejo Total, expõe os limites do modelo. A Figura 9 mostra um comportamento muito mais errático e volátil, sem a tendência clara ou a sazonalidade regular da série de consumo de energia. Consequentemente, o modelo de Markov teve grande dificuldade em aprender um padrão de transição estável. As previsões (em azul) frequentemente erraram a direção e a magnitude dos valores futuros, e a faixa de erro, embora ampla, não conseguiu conter os valores reais em múltiplos painéis. A instabilidade é confirmada pela grande variabilidade do MAPE, que chegou a atingir picos de 9.91% e 14.73% em algumas janelas. Assim como Ky e Tuyen (2018) reconheceram a dificuldade em prever séries complexas como as do mercado de ações, este resultado demonstra que a eficácia do modelo de Markov diminui significativamente quando aplicado a séries que não possuem padrões históricos fortes e repetitivos, violando a premissa fundamental da propriedade de Markov.

Conclusão e Discussão

Este trabalho se propôs a avaliar a aplicabilidade e a robustez de um modelo de Cadeias de Markov de Ordem Superior Simplificada, conforme a metodologia de Ky e Tuyen (2018), para a previsão de séries temporais no contexto brasileiro. A metodologia foi testada em um conjunto diversificado de dados, abrangendo tanto ativos de renda variável do Ibovespa, caracterizados por alta volatilidade, quanto séries macroeconômicas mensais do SGS, que possuem estruturas de tendência e sazonalidade distintas. O objetivo central era verificar a capacidade de generalização do modelo em diferentes cenários, validando sua eficácia como ferramenta de previsão para a realidade econômica nacional.

A análise dos resultados demonstrou que o modelo é flexível e adaptável. O processo de validação com janela deslizante e a busca por hiperparâmetros ótimos revelaram que não existe uma configuração única de "ordem" e "número de estados" que seja universalmente ideal; pelo contrário, a complexidade ótima do modelo é intrinsecamente dependente das características de cada série. O modelo alcançou um desempenho notável, com erros (MAPE) consistentemente baixos, para séries que exibem padrões mais regulares e previsíveis. Isso foi observado tanto em ações de setores mais estáveis, como `STBP3.SA`, quanto em séries macroeconômicas com forte componente sazonal, como o `Consumo_Total_GWh`. Em contrapartida, e em linha com a teoria financeira, o modelo apresentou maiores dificuldades e erros mais elevados para ativos e indicadores de natureza mais errática e volátil, como `VAMO3.SA` e `Varejo_Total`, onde a premissa de que o futuro próximo depende de um padrão estável do passado se torna mais frágil.

Dentre os pontos fortes da metodologia, destacam-se a sua natureza não-paramétrica, que não exige pressupostos sobre a distribuição dos dados, e a sua interpretabilidade, baseada na lógica de transição entre estados. Contudo, a principal limitação identificada é o seu caráter estritamente univariado. O modelo opera olhando apenas para o histórico da própria série, sendo incapaz de incorporar informações exógenas que sabidamente influenciam as variáveis, como taxas de juros, índices de confiança ou o desempenho de outras variáveis econômicas correlacionadas. Adicionalmente, sua premissa de probabilidades de transição constantes o torna vulnerável a quebras estruturais ou mudanças de regime não vistas nos dados de treino.

A partir dessas observações, diversas continuações e aprimoramentos para este trabalho podem ser propostos. A extensão mais natural seria a evolução para um **modelo**

multivariado, onde a transição de estado de uma série poderia depender não apenas de seu próprio passado, mas também do estado de outras variáveis. A implementação de uma Cadeia de Markov Multivariada ou de modelos de Vetores Autorregressivos com Mudança de Regime de Markov (MS-VAR) poderia capturar as interdependências entre os indicadores econômicos e financeiros, potencialmente aprimorando a acurácia das previsões. Outras possíveis continuações incluem a exploração de **métodos de discretização alternativos**, como a divisão por quantis em vez de intervalos de mesma amplitude, e o desenvolvimento de **modelos híbridos**, que poderiam combinar a Cadeia de Markov para capturar regimes de mercado com modelos como o GARCH para modelar a volatilidade dentro de cada regime.

Em suma, este estudo validou o modelo de Cadeia de Markov de Ordem Superior como uma ferramenta útil e robusta para a previsão de séries temporais brasileiras, especialmente aquelas com padrões bem definidos. Ao mesmo tempo, foram identificados seus limites em cenários de alta volatilidade e delineados caminhos claros para trabalhos futuros, que apontam para a incorporação de múltiplas variáveis como o próximo passo para a construção de modelos de previsão ainda mais completos e precisos.

Referências

- Banco Central do Brasil. (s.d.). *SGS - Sistema Gerenciador de Séries Temporais*. Disponível em: https://www3.bcb.gov.br/sgspub/consultarvalores. Acesso em: 20 jun. 2025.
- Bodie, Z., Kane, A., & Marcus, A. J. (2014). *Investments* (10th ed.). McGraw-Hill Education.
- Box, G. E. P., & Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control*.
- Bulla, J., & Bulla, I. (2006). Stylized facts of financial time series and hidden semi-Markov models. *Computational Statistics & Data Analysis*.
- Freitas, W. (2024). *rbcb: R Interface to Brazilian Central Bank Web Services* (Pacote R versão 0.1.14). Disponível em: https://github.com/wilsonfreitas/rbcb.
- Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*.
- Ky, D. X., & Tuyen, L. T. (2018). *A Higher Order Markov Model for Time Series Forecasting*.
- Pérez, F. L. (2021). *Cadeias de Markov*. Material de curso online. Universidade Federal do Paraná. Disponível em: http://leg.ufpr.br/~lucambio/CM/CM.html. Acesso em: 20 jun. 2025.
- R CORE TEAM. (2015). *R: A Language and Environment for Statistical Computing*. Viena, Austria: R Foundation for Statistical Computing. Disponível em: https://www.R-project.org/.
- Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*.