UNIVERSIDADE FEDERAL DE JUIZ DE FORA
FACULDADE DE ECONOMIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ECONOMIA

Lucas Figueira Mesquita Ribeiro

Uma análise de *nowcasting* do crescimento do PIB brasileiro: uma abordagem por meio de métricas de florestas aleatórias

Juiz de Fora

2026

Lucas Figueira Mesquita Ribeiro

Uma análise de *nowcasting* do crescimento do PIB brasileiro: uma abordagem por meio de métricas de florestas aleatórias

Dissertação apresentada ao Programa de Pós-Graduação em Economia da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Economia Aplicada. Área de concentração: Economia Regional e Macroeconomia.

Orientador: Prof. Dr. Rafael Morais de Souza

Coorientador: Prof. Dr. Wilson Luiz Rotatori Corrêa

Juiz de Fora

2026

Ficha catalográfica elaborada através do Modelo Latex do CDC da UFJF
com os dados fornecidos pelo(a) autor(a)

Ribeiro, Lucas Figueira Mesquita.
Uma análise de *nowcasting* do crescimento do PIB brasileiro : uma abordagem por meio de métricas de florestas aleatórias / Lucas Figueira Mesquita Ribeiro. – 2026.
122 f. : il.

Orientador: Rafael Morais de Souza
Coorientador: Wilson Luiz Rotatori Corrêa
Dissertação (Mestrado) – Universidade Federal de Juiz de Fora, Faculdade de Economia. Programa de Pós-Graduação em Economia, 2026.

1. *Nowcasting*. 2. Produto Interno Bruto. 3. *Dynamic Factor Model*. 4. LASSO. 5. *Random Forest* I. Souza, Rafael Morais de, orient. II. Corrêa, Wilson Luiz Rotatori, coorient. III. Título.

Lucas Figueira Mesquita Ribeiro

Uma análise de nowcasting do crescimento do PIB brasileiro: uma abordagem por meio de métricas de florestas aleatórias

Dissertação apresentada ao Programa de Pós-graduação em Economia da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Economia Aplicada. Área de concentração: Economia

Aprovada em 19 de fevereiro de 2026.

BANCA EXAMINADORA

Dr. Rafael Morais de Souza - Orientador

Universidade Federal de Juiz de Fora

Dr. Wilson Luiz Rotatori Corrêa - Coorientador

Universidade Federal de Juiz de Fora

Dr. Douglas Sad Silveira

Universidade Federal de Juiz de Fora

Dr. Bruno Pérez Ferreira

Universidade Federal de Minas Gerais

Juiz de Fora, 22/01/2026.

Documento assinado eletronicamente por **Rafael Morais de Souza**, **Professor(a)**, em 19/02/2026, às 16:47, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Wilson Luiz Rotatori Correa**, **Professor(a)**, em 20/02/2026, às 08:03, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Bruno Pérez Ferreira**, **Usuário Externo**, em 20/02/2026, às 08:06, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Douglas Sad Silveira**, **Professor(a)**, em 20/02/2026, às 08:22, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no Portal do SEI-Ufjf (www2.ufjf.br/SEI) através do ícone Conferência de Documentos, informando o código verificador **2836723** e o código CRC **082524FB**.

Para Ricardo e Lucia.

AGRADECIMENTOS

Escrever os agradecimentos é sempre difícil, mas necessário. Uma quantidade incontável de pessoas contribuiu, direta ou indiretamente, para que este trabalho fosse concluído. Agradeço primeiramente aos meus pais, Ricardo e Lucia, que foram fundamentais em toda a minha vida, incluindo nesses dois anos de Mestrado. Agradeço à Andressa, minha namorada, pelo constante apoio e carinho. Agradeço à nossa pequena Kira, que alegra os nossos dias desde que entrou em nossa vida. Agradeço a toda minha família: Rafaela, Bernardo, Bia, Zenir, Mere, Dani, Monique, Camila, Roger, Bruno(s), Carol, Naldo, Kyle, Rogério, Paulinho(s), Felipe, Nathália, Cleyton, Kíssila, Rômulo e tantos outros, pelo carinho e preocupação. Agradeço aos meus amigos de Juiz de Fora: João, Lucas, Vinícius, Samuel, Luângela, Angélica, Elenton, Murillo, César, Juninho, Gabriel, Márcio, Léo, Juliano, Felipe(s), Fraga, Gustavo(s) e tantos outros. Agradeço aos meus amigos de Campos: Paulo(s), Marlon, Ronald, Davi, Jamal, Helena, Lais, Gabriel, Allana, Raynara, Kevin, Thauffic, Álvaro, Rodrigo, Maraca, Luiz, Stella, Thamyres, Murilo, Vinicin, Mayconn, Jéssica e tantos outros. Seria impossível escrever o nome de todos, infelizmente. Agradeço ao Cezar, que me recebeu, com muita gentileza e preocupação, em Juiz de Fora. Agradeço imensamente ao Rafael, que, sempre muito gentil e atencioso, me ajudou desde o início do Mestrado, sendo meu professor, mestre, amigo e conselheiro; sem Rafael eu jamais teria concluído essa penosa jornada, sem dúvida. Agradeço ao Wilson, que sempre foi muito atencioso, gentil e disposto a tirar minhas dúvidas triviais. Agradeço aos amigos da Computação, Cristiano e Stênio, que me acompanharam durante grande parte dessa jornada. Agradeço também a Bruno, Cristiano, Douglas e Sidney, por seus valiosos comentários. Por fim, não posso esquecer de agradecer à CAPES pelo financiamento. Como toda contribuição científica, aliás, como toda contribuição de qualquer tipo, esta Dissertação não é só minha, mas de todos que me ajudaram, de uma forma ou de outra. Evidentemente, os erros que por ventura permanecerem são de minha única responsabilidade.

"O mundo não pode ser compreendido sem números, mas também não pode ser compreendido apenas com números."

Hans Rosling

ABSTRACT

The present study aims to conduct *nowcasting* exercises of Brazilian GDP growth rates. To this end, dimensionality reduction and/or shrinkage techniques are employed in order to select the most relevant variables; once this selection is performed, GDP forecasts are generated using several specifications of a Dynamic Factor Model (DFM), which are compared to an Autoregressive Model (AR). The contribution of this study lies in the use of block permutation importance (BPI) from the *Random Forest* (RF), implemented with both a *Moving Block Bootstrap* (MBB) and a *Circular Block Bootstrap* (CBB), to select the most relevant independent variables. The central hypothesis is that the application of these modern dimensionality reduction techniques may contribute to reducing the Mean Squared Forecast Error (MSFE) and, consequently, improving the predictive accuracy of the DFM. The results suggest that DFMs estimated using variables selected by RF MBB and RF CBB perform well in terms of *nowcasting* Brazilian GDP growth rates. In contrast, DFMs estimated using all variables (ALL) and those selected by LASSO and ENET exhibit lower predictive accuracy than the AR(1) model in almost all periods. Finally, it is noteworthy that the AR(1) performed considerably well for a simple technique requiring limited data and low computational capacity; in one of the periods, the AR(1) even outperformed the DFM with RF MBB, which was the best-performing specification overall.

Keywords: Nowcasting; Gross Domestic Product; Dynamic Factor Model; LASSO; Random Forest.

RESUMO

O presente trabalho possui como objetivo realizar exercícios de *nowcasting* das taxas de crescimento do PIB brasileiro. Para isso, utilizam-se técnicas de redução de dimensionalidade e/ou de encolhimento, de forma a selecionar as variáveis mais relevantes; feita esta seleção, são realizadas previsões do PIB por meio de diversas especificações de um Modelo de Fatores Dinâmicos (DFM), comparando-o com um Modelo Autoregressivo (AR). A contribuição deste trabalho está na utilização da importância por permutação em blocos (IPB) da *Random Forest* (RF) tanto com um *Moving Block Bootstrap* (MBB) quanto com um *Circular Block Bootstrap* (CBB) para selecionar as variáveis independentes mais relevantes. A hipótese do presente estudo é de que a aplicação dessas modernas técnicas de redução de dimensionalidade pode contribuir para a redução do Erro Quadrático Médio de Previsão (MSFE) e, assim sendo, aprimorar a acurácia preditiva do DFM. Os resultados encontrados sugerem que os DFMs ajustados com as variáveis selecionadas pela RF MBB e pela RF CBB desempenham bem em termos de *nowcasting* das taxas de crescimento do PIB brasileiro. Por outro lado, os DFMs ajustados com todas as variáveis (ALL) e com as variáveis selecionadas pelo LASSO e pelo ENET apresentam uma acurácia preditiva inferior àquela do AR(1) em quase todos os períodos. Por fim, destaca-se que o AR(1) desempenhou consideravelmente bem para uma técnica simples, que demanda poucos dados e pouca capacidade computacional; em um dos períodos, o AR(1) teve um desempenho superior até mesmo que o DFM com a RF MBB, que foi a melhor especificação.

Palavras-chave: *Nowcasting*; Produto Interno Bruto; *Dynamic Factor Model*; LASSO; *Random Forest*.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A compreensão do estado real e atual de uma economia está longe de ser tarefa simples. Quando considera-se a produção da economia, em termos monetários, a métrica mais conhecida e utilizada é o Produto Interno Bruto (PIB), que mensura o valor adicionado presente nos bens e serviços finais. Não obstante, o PIB é uma variável[1] que possui uma frequência de divulgação trimestral, isto é, possui uma baixa frequência, e é disponibilizado, no Brasil, com um atraso considerável, de cerca de 65 dias após o fechamento do trimestre (IBGE, 2025a; IBGE, 2025b). Em outras palavras, existe uma dificuldade de quantificar o estado corrente da economia brasileira antes da divulgação dessa variável trimestral, o que implica na necessidade de formulação de maneiras alternativas de estimar esse PIB antes que seja, de fato, disponibilizado. Nesse contexto, surgiu a literatura do *nowcasting*[2], na qual os pesquisadores almejam utilizar variáveis disponibilizadas em frequências maiores (mensais, de modo geral) e com atrasos menores para prever o PIB em tempo real. Essas estimações realizadas pelos pesquisadores que trabalham com o *nowcasting* são frequentemente efetuadas por meio de Modelos de Fatores Dinâmicos (DFMs, da sigla em inglês), originalmente formulados por Evans (2005) e Giannone, Reichlin and Small (2008).

Com o decorrer dos anos, diversos estudos buscaram estimar, em tempo real, a taxa de crescimento do PIB brasileiro por meio de DFMs com diferentes especificações. Cepni, Güney e Swanson (2019a, 2019b) utilizaram séries econômicas nacionais, bem como índices globais, para averiguar se há ganhos na previsão do PIB para cinco países emergentes, incluindo o Brasil; seus resultados sugerem que, de fato, esses índices incrementam a acurácia preditiva dos modelos. Bantis, Clements and Urquhart (2023) utilizaram tanto séries econômicas tradicionais quanto dados de pesquisa do *Google Trends* para avaliar se estes acarretam em ganhos na previsão do PIB para os Estados Unidos e o Brasil; seus resultados sugerem que sim, mas que depende do horizonte de previsão. Todos estes três estudos utilizam técnicas de encolhimento e/ou de seleção de variáveis, com o objetivo de diminuir o excesso de informação desnecessária, isto é, de informação que não contribui para uma melhoria no desempenho preditivo do modelo. Afinal, um conjunto maior de

[1] Os termos "variável" e "série" são utilizados, neste trabalho, de forma intercambiável e se referem a todo tipo de atributo, quantidade e/ou característica que assume, ou pode assumir, diferentes valores para diferentes observações. Por sua vez, uma série temporal, ou série de tempo, é definida como um conjunto de observações ordenadas no tempo. O salário mínimo real, a produção industrial e o índice de confiança do consumidor são exemplos de variáveis, mais especificamente de séries temporais.

[2] Neste trabalho, considera-se o termo "previsão em tempo real" uma razoável tradução de "*nowcasting*", o que está em acordo com a definição dada por Bańbura, Giannone and Reichlin (2010). Por sua vez, a definição de *nowcasting* dada por Giannone, Reichlin and Small (2008) está restrita somente à previsão do trimestre corrente, sendo, portanto, uma definição menos abrangente. Não obstante, opta-se por utilizar somente o termo "*nowcasting*", que é bem consolidado na literatura.

dados nem sempre implica em um aprimoramento na acurácia preditiva do modelo (Boivin; Ng, 2006). A literatura mostra que ocorrem ganhos em termos de acurácia preditiva quando ocorre uma seleção prévia das variáveis independentes do estudo (Bai; Ng, 2008; Kim; Swanson, 2018). Os demais trabalhos que realizam exercícios de *nowcasting* da taxa de crescimento do PIB brasileiro não utilizam nenhuma dessas técnicas de seleção de variáveis (Bragoli; Metelli; Modugno, 2015; Issler; Notini, 2016; Dahlhaus; Guénette; Vasishtha, 2017; Issler; Pimentel, 2019).

As florestas aleatórias são técnicas preditivas de *Machine Learning* que podem ser aplicadas tanto a problemas de classificação quanto aos de regressão. Por meio dessas florestas, é formado um *ensemble*[3] de árvores de decisão, com um componente aleatório presente, por meio do mecanismo de *bootstrap*[4], e funciona de forma que cada previsão seja o resultado, no caso da regressão, da média das previsões de cada árvore individual. Essas florestas apresentam, para além disso, métricas de importância de variáveis, o que permite avaliar quais variáveis independentes são as mais importantes para a previsão da variável dependente (Breiman, 2001; Strobl et al., 2008; James et al., 2013; Goehry et al., 2023). Nas últimas décadas, essas medidas das florestas aleatórias foram sugeridas para a seleção das variáveis independentes mais relevantes em várias áreas científicas, com destaque para as análises de dados de microarranjo, de sequenciamento de DNA e de previsão de séries temporais (Lunetta et al., 2004; Bureau et al., 2005; Huang et al., 2005; Qi; Bar-Joseph; Klein-Seetharaman, 2006; Strobl et al., 2008; Goehry et al., 2023). Não obstante, até onde se sabe, nenhum dos trabalhos de *nowcasting* da taxa de crescimento do PIB brasileiro utilizou métricas de importância de variáveis das florestas aleatórias para selecionar as variáveis independentes mais relevantes. Neste sentido, o presente estudo contribui com a utilização desses métodos para a seleção de variáveis, utilização esta que foi bem-sucedida até mesmo em problemas de previsão de séries temporais (Huang; Lu; Xu, 2016; Goehry et al., 2023; Fang et al., 2024), mas ainda não foi, até onde se sabe, aplicada a problemas de *nowcasting* do PIB brasileiro.

[3] Um *ensemble* é um conjunto de modelos que possui o objetivo de produzir resultados mais robustos. Em vez de utilizar somente um modelo preditivo, são criados vários modelos e depois combinadas suas previsões para obter um resultado final mais robusto. Essa combinação pode ocorrer, por exemplo, por meio de uma média das previsões de cada modelo singular. A ideia central é que, quando combinados muitos modelos fracos, pode-se obter um modelo mais forte e robusto (James et al., 2013; Goehry et al., 2023).

[4] O *bootstrap* é uma ferramenta estatística de reamostragem com reposição, com o objetivo de aproximar, a partir dos próprios dados utilizados, a distribuição amostral de um estimador ou estatística. Por meio da reposição, uma observação pode aparecer mais de uma vez e algumas podem sequer aparecer. Por meio do algoritmo de *bootstrap* surgem as observações *out-of-bag* (OOB), já que, em cada reamostragem de tamanho n com reposição, a probabilidade de uma observação não ser escolhida é de cerca de um terço. No caso das *Random Forests*, o *bootstrap* funciona de forma a produzir diferentes conjuntos de dados para treinar as árvores de decisão, enquanto o cálculo da média ou o voto majoritário pertencem à etapa de *bagging*, que é justamente a agregação do *bootstrap* (James et al., 2013).

Dadas essas questões, o presente trabalho tem como objetivo principal prever a taxa de crescimento trimestral do PIB brasileiro. Para isso, utilizam-se técnicas de redução de dimensionalidade e/ou de encolhimento, de forma a selecionar as variáveis mais relevantes; após essa fase de seleção, são realizadas previsões do PIB por meio de diversas especificações, utilizando tanto um Modelo de Fatores Dinâmicos (DFM) quanto um Modelo Autorregressivo (AR, da sigla em inglês) simples. A contribuição deste trabalho está na utilização de métricas de importância de variáveis derivadas de florestas aleatórias, a saber, da importância por permutação em blocos (IPB) da *Random Forest* (RF) tanto com um *Moving Block Bootstrap* (MBB) quanto com um *Circular Block Bootstrap* (CBB), para selecionar as variáveis independentes mais relevantes (Breiman, 2001; Strobl et al., 2008; Friedberg et al., 2020; Goehry et al., 2023). Esses novos métodos de *bootstrap* em blocos não foram, até o conhecimento deste autor, utilizados no *nowcasting* da taxa de crescimento do PIB, o que caracteriza esta pesquisa como potencialmente inovadora. Para a seleção de variáveis, a variável dependente[5] utilizada foi o Índice de Atividade Econômica do Banco Central do Brasil (IBC-Br), que é um índice coincidente[6] do PIB, dada a necessidade de que esta variável seja da mesma frequência das variáveis independentes[7].

A hipótese do presente estudo é de que a aplicação dessas modernas técnicas de redução de dimensionalidade pode contribuir para a redução do Erro Quadrático Médio de Previsão (MSFE, da sigla em inglês) e, assim sendo, aprimorar a acurácia preditiva do DFM. A questão de pesquisa que se coloca é se métodos baseados em florestas aleatórias voltados para a mensuração da importância das variáveis fornecem resultados melhores do que técnicas tradicionais de encolhimento como o *Least Absolute Shrinkage and Selection Operator* (LASSO) e o *Elastic Net* (ENET), já que aquelas métricas de florestas aleatórias apresentaram bons resultados em outros contextos científicos (Lunetta et al., 2004; Bureau et al., 2005; Huang et al., 2005; Qi; Bar-Joseph; Klein-Seetharaman, 2006; Strobl et al.,

[5] Variável dependente é a variável que almeja-se explicar, prever e/ou estimar por meio de um modelo. Esse termo possui inúmeros sinônimos, dentre os quais se destacam: variável alvo, variável resposta, variável objetivo, variável predita, variável explicada, variável de saída, regressando e *label*. No decorrer deste estudo, utiliza-se somente o termo variável dependente, tendo em vista que este explicita a relação de dependência entre a variável dependente e as variáveis independentes.

[6] Um índice coincidente é um indicador concebido para acompanhar o nível corrente de atividade econômica, isto é, para refletir as condições econômicas no período mais recente disponível. No caso da economia brasileira, o Índice de Atividade Econômica do Banco Central do Brasil (IBC-Br) foi desenvolvido com o objetivo de sintetizar, em frequência mensal, as informações provenientes de diferentes segmentos da economia, exercendo o papel de *proxy* do nível de atividade econômica e como instrumento de acompanhamento do ciclo econômico (Banco Central do Brasil, 2025d).

[7] Variáveis independentes são as variáveis usadas para explicar, prever e/ou estimar a variável dependente. Dentre os inúmeros sinônimos para esse termo, destacam-se: variável de entrada, variável explicativa, variável preditora, covariável, regressor, *feature* e preditor. No decorrer deste estudo, utiliza-se somente o termo variável independente, tendo em vista que este explicita a relação de dependência entre a variável dependente e as variáveis independentes.

2008; Huang; Lu; Xu, 2016; Fang et al., 2024). Destaca-se que as técnicas de encolhimento, como o LASSO e o ENET, já se consolidaram como ferramentas de seleção prévia de variáveis independentes na literatura de *nowcasting* das taxas de crescimento do PIB (Cepni; Güney; Swanson, 2019a, 2019b; Bantis; Clements; Urquhart, 2023). Em outras palavras, busca-se investigar se a seleção de variáveis a partir de métricas de importância de variáveis derivadas de florestas aleatórias pode efetivamente aprimorar a capacidade preditiva do DFM, refletindo-se em menores valores do MSFE e, consequentemente, em previsões mais acuradas para as dinâmicas do PIB brasileiro.

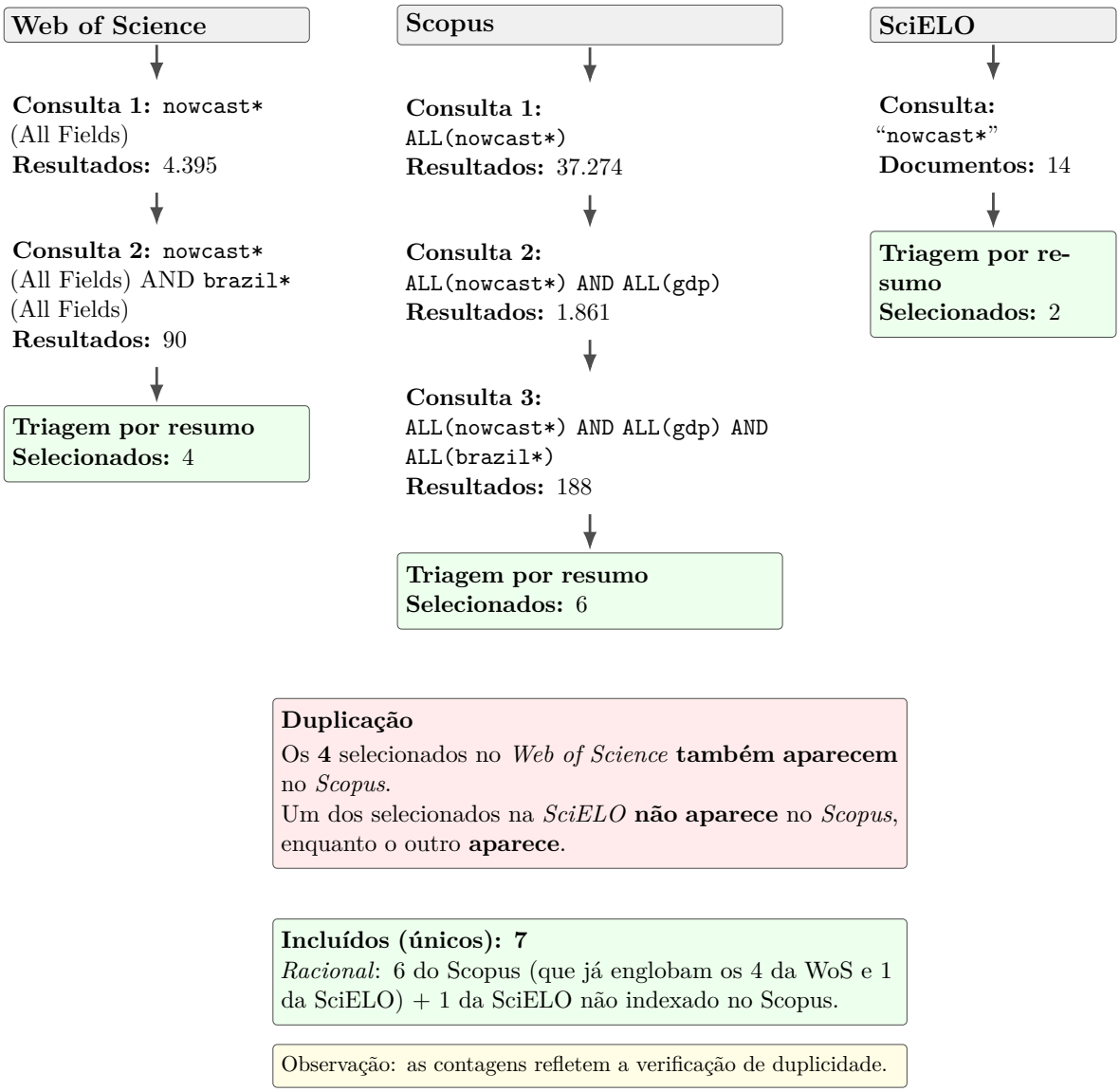
Os resultados encontrados sugerem que os DFMs ajustados com as variáveis selecionadas pela RF MBB e pela RF CBB desempenham bem em termos de *nowcasting* das taxas de crescimento do PIB brasileiro. Por outro lado, os DFMs ajustados com todas as variáveis (ALL) e com as variáveis selecionadas pelo LASSO e pelo ENET apresentam uma acurácia preditiva inferior àquela do AR(1) em quase todos os períodos. Destaca-se que o AR(1) desempenhou consideravelmente bem para uma técnica simples, que demanda poucos dados e pouca capacidade computacional; em um dos períodos, o AR(1) teve um desempenho superior até mesmo que o DFM com a RF MBB, que foi a melhor especificação. Em outras palavras, os resultados sugerem existir ganhos significativos com as novas metodologias propostas neste trabalho, a saber, a RF MBB e a RF CBB, com o AR(1) sendo uma alternativa viável e pouco custosa. Por fim, é prudente ressaltar que o bom desempenho do AR(1) está de acordo com o esperado, dado que a série da taxa de crescimento do PIB brasileiro é estacionária e, portanto, espera-se que suas dinâmicas sejam bem apreendidas por um modelo autorregressivo.

O presente trabalho é composto por este capítulo de Introdução, seguido pelo capítulo de Revisão de Literatura, pelo capítulo de Metodologia, pelo capítulo dos Dados, pelo capítulo de Resultados e, por fim, pelo capítulo de Considerações Finais. Após esses capítulos, estarão presentes as Referências e os Apêndices.

2 REVISÃO DE LITERATURA

O objetivo do presente capítulo é realizar uma revisão de literatura das principais contribuições para uma análise de *nowcasting* do PIB brasileiro, tanto em termos teóricos quanto aplicados. Para atingir esse objetivo, foram utilizadas algumas bases de pesquisas científicas, a saber, o *Web of Science* (WoS), o *SciELO* e o *Scopus*, com o objetivo de selecionar os artigos mais relevantes para esta revisão. A Figura 1 apresenta os critérios utilizados para as buscas nessas três bases, bem como seus resultados. Como o foco se concentrou em encontrar artigos que tratassem especificamente do *nowcasting* do PIB brasileiro, um número pequeno de documentos foi selecionado; todos os artigos que atenderam aos critérios foram selecionados e revisados no presente capítulo. Não obstante, artigos seminais, como os de Evans (2005) e Giannone, Reichlin and Small (2008), foram revisados, apesar de não terem sido selecionados pelas bases de pesquisas científicas.

Figura 1 – Fluxo de busca e seleção de estudos (WoS, Scopus e SciELO)



Fonte: Elaboração própria.

Em seu artigo seminal acerca das estimativas em tempo real do estado corrente da economia estadunidense, Evans (2005) argumenta que as informações em tempo real sobre o estado atual da atividade econômica estão dispersas entre inúmeros consumidores, empresas e formuladores de políticas públicas. Em decorrência dessa questão, os formuladores de políticas públicas não possuem informações precisas sobre a atividade econômica privada corrente, já que os dados rotineiramente disponibilizados pelos vários órgãos do setor público com frequência representam uma agregação de informações passadas, não sendo, portanto, adequados para uma análise da atividade econômica em tempo real. O autor afirma que essa escassez de informações é amplamente reconhecida pelos formuladores de políticas públicas, principalmente quando se trata do PIB, que é a mais ampla medida da atividade econômica real. Esse problema dificulta, inclusive, o processo de tomada de

decisão do *Federal Reserve*[1] acerca de sua política monetária.

O objetivo de Evans (2005) é descrever um método para a estimação do estado corrente da economia de forma contínua por meio da utilização das informações contidas em uma ampla gama de séries macroeconômicas. Essas estimativas em tempo real são computadas por meio de um modelo econométrico que permite a presença de defasagens nas publicações das variáveis, agregação temporal e outras dificuldades características do fluxo diário de informações macroeconômicas. Destaca-se que o modelo elaborado pode ser utilizado não somente para exercícios de *nowcasting* do PIB, como também da inflação, do desemprego ou de qualquer outra variável macroeconômica de interesse. Entretanto, o foco do autor se encontra justamente no *nowcasting* do PIB.

O trabalho de Evans (2005) se diferencia de parte da literatura (Chow; Lin, 1971; Liu; Hall, 2001; Mariano; Murasawa, 2003) por modelar a taxa de crescimento do PIB como o agregado trimestral de um processo diário não observado para a atividade econômica real. O modelo especifica, igualmente, a relação entre o PIB, as divulgações de dados do crescimento do PIB e as divulgações de dados para um conjunto de outras variáveis macroeconômicas de uma forma em que a complexa temporalidade dessas divulgações seja adequadamente incorporada. Além disso, a estrutura do modelo desenvolvido por ele permite computar estimativas do PIB em tempo real como a solução para um problema de inferência. Em termos práticos, ele obtém as estimativas em tempo real como um subproduto da estimação do modelo. Os parâmetros são estimados por uma quase máxima verossimilhança usando o algoritmo do Filtro de Kalman[2], de forma que as estimativas em tempo real sejam, então, obtidas por meio da aplicação do algoritmo ao modelo avaliado nas estimativas de máxima verossimilhança (MLEs, da sigla em inglês).

Em sua seminal contribuição para a análise preditiva de curto prazo, Giannone, Reichlin and Small (2008) ressaltam, em primeiro momento, que as decisões de política monetária são realizadas a partir de avaliações dos estados corrente e futuro da economia, avaliações essas que são feitas utilizando dados incompletos. Como a maioria dos dados

[1] O *Federal Reserve System* é o Banco Central dos Estados Unidos, sendo responsável por diversas funções, das quais se destacam a condução da política monetária, a promoção da estabilidade do sistema financeiro e a promoção da segurança das instituições financeiras (Board of Governors of the Federal Reserve System, 2025).

[2] O Filtro de Kalman é um algoritmo recursivo para modelos lineares gaussianos em espaço de estados, que auxilia sobremaneira em problemas que lidam com bordas irregulares. Em cada período t, o filtro combina a previsão do estado latente realizada pelo modelo com base no período $t-1$ com a informação contida nas observações disponíveis em t, produzindo $\hat{\alpha}_{t|t-1} = \mathbb{E}(\alpha_t \mid y_{1:t-1})$, onde $\mathbb{E}(\alpha_t \mid \cdot)$ é a esperança condicional do estado α_t dado um conjunto de informações e $y_{1:t-1}$ é o conjunto de todas as observações disponíveis até o período $t-1$. Quando existem observações ausentes, como em uma borda irregular, a etapa de atualização é realizada somente com o subconjunto observado, ignorando as entradas ausentes. Por sua vez, o Suavizador de Kalman utiliza toda a amostra $y_{1:T}$ para refinar os estados, produzindo $\hat{\alpha}_{t|T} = \mathbb{E}(\alpha_t \mid y_{1:T})$, o que permite obter estimativas para as observações ausentes e preencher a borda irregular na base de dados (Giannone; Reichlin; Small, 2008).

é divulgada com atrasos, sendo, posteriormente, revisada, as tarefas de *forecasting*[3] e de *nowcasting* possuem uma importância considerável para os bancos centrais. É justamente em decorrência dessas dificuldades apresentadas pelos dados que os bancos centrais e os mercados estão sempre atentos a determinadas variáveis de interesse, seja porque são disponibilizadas em tempo hábil ou porque estão profundamente relacionadas à variável que se quer prever. Um exemplo fornecido pelos autores é da coleta de dados relativos à taxa de desemprego e/ou à produção industrial para auxiliar na previsão do PIB, tendo em vista que essas três variáveis possuem uma profunda conexão. Os autores argumentam, de forma semelhante, que, a princípio, toda e qualquer divulgação de dados pode impactar nas estimativas sobre o trimestre corrente, bem como suas respectivas precisões. Para o pesquisador que trabalha com o *nowcasting*, não há motivos que justifiquem descartar informações, apesar de que seja importante compreender em que medida cada uma dessas informações é confiável como um sinal das atuais condições econômicas.

Dadas essas considerações, Giannone, Reichlin and Small (2008) desenvolvem um modelo formal de previsão que lida com diversos tópicos-chave que aparecem quando trabalha-se com um grande número de séries que são divulgadas em diferentes períodos e com diferentes atrasos. Além disso, os autores combinam a ideia de conectar informações mensais com o *nowcasting* do PIB trimestral e a ideia de utilizar um grande número de séries temporais em uma única estrutura estatística. A partir disso, essa estrutura incorpora, formalmente, a atualização do *nowcasting* do PIB à medida que dados mensais são disponibilizados com o decorrer do trimestre. De forma semelhante, essa abordagem pode ser usada, igualmente, para avaliar o impacto marginal de cada nova divulgação de dados sobre a previsão e sua acurácia. Em outras palavras, essa estrutura estatística pode ser compreendida como um grande modelo *bridge*[4] que combina três diferentes aspectos do *nowcasting*, a saber, (1) ela utiliza um elevado número de séries temporais, (2) ela atualiza as previsões e mensura sua acurácia de acordo com o calendário em tempo real das disponibilizações de dados e (3) ela conecta divulgações de dados mensais com a previsão do PIB trimestral em tempo real.

[3] No presente trabalho, utiliza-se o termo "*forecasting*", em detrimento de sua tradução para o português, quando o objetivo é contrastar o desempenho dos exercícios de *forecasting* e de *nowcasting*. O *forecasting* se refere ao exercício de prever o valor de uma variável em algum período seguinte, o que se contrapõe ao exercício de *nowcasting*, no qual o objetivo é prever o período em que se está inserido ou, pelo menos, prever o período utilizando somente informações que estavam disponíveis no período em consideração (Cepni; Güney; Swanson, 2019a, 2019b; Bantis; Clements; Urquhart, 2023).

[4] Um modelo *bridge* é um modelo econométrico que realiza uma ponte (*bridge*) entre variáveis independentes de maiores frequências (mensais, por exemplo) e uma variável dependente de menor frequência (o PIB trimestral, por exemplo). Esse modelo permite utilizar informações que chegam ao longo do trimestre, com seus vários atrasos de publicação, para atualizar continuamente o *nowcasting* do PIB. Em outras palavras, pode-se interpretar o modelo *bridge* como uma forma de construir uma ponte entre variáveis de diferentes frequências.

Outro ponto importante levantado por Giannone, Reichlin and Small (2008) é a forma parcimoniosa pela qual o modelo precisa ser especificado, em decorrência da utilização das informações contidas em um grande número de dados, de forma que mantenha seu poder preditivo. Esse objetivo é atingido por meio do resumo das informações de inúmeras variáveis em poucos fatores comuns. Dessa forma, o exercício de *nowcasting* é definido por eles como a projeção do PIB trimestral com base em fatores comuns estimados a partir de dados mensais. Em um exercício de *nowcasting*, algumas séries possuem observações para o período corrente, enquanto as observações mais recentes de outras séries se encontram disponíveis somente para um mês ou trimestre passados. Portanto, os conjuntos de dados não estão balanceados, de forma que lidar, de maneira apropriada, com a borda irregular[5] característica dessas séries de dados seja de extrema relevância para realizar exercícios de *nowcasting* que utilizem as informações contidas nas mais recentes divulgações de dados.

A partir deste momento prossegue-se para uma análise das contribuições aplicadas de *nowcasting* do PIB brasileiro. Em sua análise preditiva das taxas de crescimento do PIB nos Estados Unidos e no Brasil, Bantis, Clements and Urquhart (2023) usaram dados de pesquisa do *Google Trends* tanto para o *nowcasting* quanto para o *forecasting* dessas duas importantes economias. Com frequência, três diferentes tipos de dados são usados para a análise de *nowcasting*, a saber, (1) indicadores brutos, dos quais se destacam a produção industrial e as vendas no varejo; (2) pesquisas de opinião e de intenção; e (3) dados do mercado financeiro que possuam altas frequências. Não obstante, nos últimos anos fontes alternativas de dados surgiram como possíveis opções para o *nowcasting*, em decorrência dos avanços na computação e nos serviços de coleta de informações on-line. Essas fontes de dados são com frequência chamadas de *big data*[6], sendo os dados do *Google Trends* uma de suas principais bases para o *nowcasting*. O foco do trabalho realizado por eles se encontra na contribuição marginal da *big data* na forma de dados do *Google Trends* para além das variáveis independentes tradicionalmente usadas pelos especialistas. Em outras palavras, os autores almejavam determinar se conjuntos de dados de alta dimensionalidade, como as séries do *Google Search*, apresentam uma capacidade preditiva adicional em relação às tradicionais fontes de dados de elevada frequência.

Bantis, Clements and Urquhart (2023) utilizaram, igualmente, alguns métodos de seleção de variáveis, a saber, o *Least Absolute Shrinkage and Selection Operator* (LASSO), o

[5] No presente trabalho utiliza-se o termo "borda irregular" como tradução dos termos "*ragged edge*" e "*jagged edge*", que são frequentes na literatura de *nowcasting*. O problema da borda irregular ocorre quando uma ou mais variáveis possuem observações ausentes (NAs) no final da amostra, em decorrência dos diferentes atrasos em suas respectivas publicações (Bantis; Clements; Urquhart, 2023).

[6] A empresa *International Business Machines* (IBM) classifica os *big data* como bases de dados que possuem elevados (1) volume, (2) variedade, (3) velocidade e (4) veracidade. Essas quatro categorias são chamadas, por motivos de abreviação, de "os quatro Vs" (Bantis; Clements; Urquhart, 2023).

LASSO adaptativo e o *Elastic Net* (ENET). O principal conjunto de dados usado consistiu em 96 e 115 indicadores econômicos para o Brasil e os Estados Unidos, respectivamente. Esses dados foram agrupados em dez grupos: (1) atividade econômica, (2) setor externo, (3) setor governamental, (4) mercado imobiliário, (5) mercado de trabalho, (6) principais indicadores, (7) setor monetário, (8) preços, (9) setor de varejo e (10) indicadores de pesquisa. Todos esses indicadores econômicos foram coletados dos *Key Economic Indicators* da *Bloomberg* para o período de janeiro de 2005 até setembro de 2019. Apesar das variáveis financeiras fornecerem informações em tempo hábil, os autores optaram por não as utilizar em decorrência de suas elevadas volatilidades, que poderiam acrescentar um ruído considerável ao modelo. Destaca-se, de forma semelhante, que todas as variáveis foram padronizadas por meio da subtração de suas respectivas médias e pela divisão por seus respectivos desvios padrão, de forma a evitar sobrecarregar as variáveis independentes com elevadas variâncias quando os fatores fossem derivados.

Os resultados de Bantis, Clements and Urquhart (2023) corroboraram grande parte da literatura ao evidenciar que o DFM incorpora, com sucesso, novas informações à medida que elas são divulgadas, com os erros de previsão diminuindo à medida em que ocorre um deslocamento do *forecasting* para o *nowcasting* e depois para o *backcasting* (Dahlhaus; Guénette; Vashistha, 2017; Cepni; Güney; Swanson, 2019a, 2019b). Além disso, esses resultados mostram que as técnicas de seleção de variáveis contribuíram, na maioria dos casos, para a redução da Raiz do Erro Quadrático Médio de Previsão (RMSFE, da sigla em inglês). Tanto no caso brasileiro quanto no estadunidense, as previsões realizadas com a seleção prévia de variáveis usando o LASSO apresentaram, em muitos horizontes de previsão, o menor RMSFE. Não obstante, para o Brasil os DFMs superam o *benchmark*, o AR simples, principalmente nos horizontes temporais de *nowcasting* e de *backcasting*, enquanto que, para horizontes mais longos, os ganhos de previsão tendem a desaparecer. Por fim, destaca-se que os ganhos da seleção prévia de variáveis ocorrem principalmente em previsões para um trimestre à frente e que o desempenho dessas técnicas de seleção de variáveis diminui à medida que novas informações são incorporadas.

Em seu estudo de *nowcasting* para Brasil, Rússia, Índia, China e México, chamados pelos autores de BRIC+M, Dahlhaus, Guénette and Vasishtha (2017) utilizaram o DFM e um número de indicadores mensais que variou entre 13 e 41 para os países do BRIC+M, com 36 séries de dados para o Brasil. Destaca-se que em muitas das séries usadas para o *nowcasting* do PIB brasileiro foram realizadas transformações para atingir a estacionariedade, seja por meio da expressão dos dados em logaritmos e/ou por meio da aplicação da primeira diferença. No caso brasileiro, o período de estimação foi entre o primeiro trimestre de 1996 e o primeiro trimestre de 2014. Esses indicadores podem ser agrupados em nove categorias, a saber, (1) os índices dos gerentes de compras, (2) os indicadores *soft*, (3) a produção industrial, (4) a produção, as vendas e o uso de veículos,

(5) a balança de pagamentos, (6) os indicadores financeiros, (7) os indicadores do mercado de trabalho, (8) os preços e (9) as variáveis exógenas. Além disso, os autores converteram a frequência das variáveis diárias, como os preços do petróleo e diversos indicadores financeiros, para mensais, por meio da realização de suas respectivas médias mensais.

Os resultados encontrados por Dahlhaus, Guénette and Vasishtha (2017) sugerem que os DFMs apresentam, de modo geral, uma boa acurácia direcional e fornecem previsões em tempo real confiáveis do crescimento do PIB real para as economias do BRIC+M, de forma a capturar razoavelmente bem as dinâmicas do PIB nesses países. Esses resultados são frequentemente robustos[7] a outras especificações do modelo. Os autores descobriram, igualmente, que os DFMs possuem, por via de regra, um desempenho superior aos modelos univariados de *benchmark*, os ARs. Para além disso, as divulgações de variáveis financeiras e de indicadores *soft* desempenham um papel muito relevante na explicação das revisões de *nowcasting* que ocorrem no começo do período de previsão. Por sua vez, as novidades que decorrem de indicadores domésticos são o principal motor das mudanças nas revisões do *nowcasting* durante todo o período em que foi realizada a previsão, enquanto as variáveis exógenas (globais e dos EUA) aparentaram incrementar a acurácia preditiva somente no início do período.

Por sua vez, Cepni, Güney and Swanson (2019a) utilizaram DFMs para os exercícios de *nowcasting* e focaram em cinco economias emergentes: Brasil, Indonésia, México, África do Sul e Turquia. Os autores realizaram, igualmente, uma comparação das previsões incorridas pelas diferentes especificações dos modelos, comparação essa que é chamada de *forecasting horse race*, e que também é realizada neste trabalho. O conjunto de modelos examinado pelos autores inclui tanto os ARs quanto os DFMs, que são estimados usando diversos métodos de seleção de variáveis baseados no LASSO. Os autores utilizaram um grande conjunto de indicadores econômicos coletados da *Bloomberg*, conjunto este que consiste em 97, 87, 116, 109 e 102 variáveis para Brasil, Indonésia, México, África do Sul e Turquia, respectivamente, para o período entre janeiro de 2003 e junho de 2018. Esse conjunto de dados é composto pelos indicadores *hard*, que incluem tanto variáveis do lado da oferta quanto do lado da demanda, e pelos dados de pesquisas, que incluem o *Purchasing Managers' Index* (PMI), que é um dos indicadores mais observados dos ciclos econômicos. O conjunto de dados utilizado pelos autores pode ser dividido nas seguintes categorias: (1) variáveis de habitação e encomenda, (2) variáveis do mercado de trabalho, (3) preços, (4) variáveis financeiras, (5) agregados monetários, de crédito e de quantidade e (6) variáveis da atividade econômica real.

Os resultados encontrados por Cepni, Güney and Swanson (2019a) mostraram

[7] A robustez é a capacidade de um modelo de manter seu desempenho, bem como suas conclusões, estável quando os dados sofrem uma pequena alteração, as hipóteses do modelo não se verificam exatamente e/ou ocorrem pequenos deslocamentos de distribuição. Em outras palavras, um modelo é robusto quando resiste a perturbações plausíveis do problema.

que (1) ocorreu uma queda considerável do MSFE à medida em que novos dados se tornaram disponíveis e foram incorporados aos modelos, de forma que a acurácia dos DFMs geralmente aumenta com a incorporação de novas informações; (2) os métodos de seleção de variáveis empregados acarretaram grandes ganhos em termos de acurácia preditiva; e (3) o aprimoramento dos DFMs utilizados com defasagens temporais para a variável dependente, isto é, com a incorporação de um componente AR ao DFM, gerou ganhos de previsão, em relação aos modelos sem o termo AR, somente para a Indonésia e a África do Sul. Por fim, deve-se ressaltar que os autores consideraram este último resultado deveras interessante, já que existe uma preponderância de evidências na literatura de previsão de séries temporais no que se refere à importância de incluir componentes AR quando forem realizadas previsões de variáveis econômicas e, além disso, enfatiza a relevância das variáveis de incerteza e de surpresa em exercícios de previsão do PIB de economias em desenvolvimento.

Cepni, Güney and Swanson (2019b) repetiram o exercício para Brasil, Indonésia, México, África do Sul e Turquia utilizando métodos de *big data* ao analisarem diversos métodos de seleção de variáveis. O objetivo do estudo foi investigar a possibilidade de produzir sinais preliminares do estado atual da economia, antes mesmo que os dados oficiais fossem divulgados. Em relação aos dados, os autores utilizaram um grande conjunto de indicadores econômicos, com 103, 103, 117, 110 e 88 séries para Turquia, Brasil, México, África do Sul e Indonésia, respectivamente. Essas séries foram escolhidas para representarem amplas categorias de indicadores econômicos e se referem ao período entre janeiro de 2005 e setembro de 2017. Os métodos de seleção de variáveis usados foram (1) o LASSO, (2) o ENET, (3) o *Least Angle Regression* (LARS) e (4) o *Sparse Principal Component Analysis* (SPCA). Diversos estudos já mostraram que a implementação de métodos de seleção de variáveis antes do exercício de previsão pode ser muito positiva para a capacidade preditiva dos modelos (Kim; Swanson, 2018; Bai; Ng, 2008). Os resultados mostraram, dentre outras coisas, que, quando os métodos de seleção de variáveis são empregados, os DFMs produzem previsões superiores tanto aos ARs quanto aos DFMs sem a seleção prévia de variáveis.

A partir do presente momento o foco será direcionado para uma análise das contribuições de *nowcasting* para a atividade econômica somente do Brasil, isto é, o foco estará restrito à economia brasileira. Bragoli, Metelli and Modugno (2015) avaliam o desempenho do *nowcasting* dos especialistas, conforme apresentados no Sistema de Expectativas de Mercado do Banco Central do Brasil (BCB), e compararam-no com seus próprios exercícios de *nowcasting* realizados com o DFM. Para isso, foram utilizadas 12 variáveis, para além da taxa de crescimento do PIB. Essas variáveis foram transformadas de forma a atingir a estacionariedade de cada uma e podem ser agrupadas em cinco categorias: (1) indicadores de pesquisas, (2) trabalho, (3) produção, (4) demanda e (5) comércio.

Os resultados indicaram que tanto os exercícios de *nowcasting* do modelo quanto os dos especialistas são bem comportados, o que significa que, à medida que novas informações se tornam disponíveis, aumentam sua precisão e correlação com os valores observados. Entretanto, os resultados do modelo apresentam um desempenho ligeiramente superior às previsões institucionais, o que corrobora as descobertas de Giannone, Reichlin and Small (2008) e Liebermann (2014), porém contrasta com os resultados de Ang, Bekaert and Wei (2007), Clements (2010) e Jansen, Jin and Winter (2016).

Em seu trabalho de estimação do PIB mensal do Brasil, Issler and Notini (2016) afirmam contribuir de três maneiras para essa literatura. A primeira delas é a utilização de uma técnica mais avançada de interpolação do PIB brasileiro para a frequência mensal, de forma a auxiliar no processo de *nowcasting* fora da amostra analisada. A segunda contribuição se encontra na proposição, bem como no teste, de inúmeros modelos de interpolação e de séries auxiliares para habilitar a interpolação do PIB. A terceira contribuição se refere à comparação entre a série de PIB mensal construída pelos autores e o IBC-Br, que é rotineiramente utilizado pelo mercado financeiro em suas tomadas de decisão. Por meio dessa comparação, foi possível constatar que o seu indicador do PIB mensal estava mais próximo das variações do PIB dentro da amostra analisada e, para além disso, o previu de forma mais satisfatória fora dessa amostra.

Os resultados encontrados por Issler and Notini (2016) sugerem que a construção do IBC-Br seja pouco eficiente, já que necessita do armazenamento e do acompanhamento de centenas de séries com frequência mensal, enquanto que o modelo desenvolvido por eles requer o armazenamento e o acompanhamento de apenas duas séries auxiliares. Em sua descrição dos modelos de espaço de estados[8], os autores afirmam que o Filtro de Kalman, que foi usado por eles, forneceu um eficiente meio computacional e recursivo de estimar o estado de um processo estocástico, com frequência posto na forma de espaço de estados. Por sua vez, essa abordagem de espaço de estados incorpora as relações entre as séries observáveis e as não observáveis, sendo um tópico central nessa abordagem a estimação das variáveis não observáveis por meio das informações passadas e/ou correntes disponíveis (Hamilton, 1994; Harvey, 1989).

Em um trabalho mais recente, Issler and Pimentel (2019) afirmam que, apesar da evidente relevância de indicadores como o Indicador Antecedente Composto da Economia (IACE), o Indicador Coincidente Composto da Economia (ICCE) e o Índice de Atividade Econômica do Banco Central do Brasil (IBC-Br), eles possuem uma metodologia excessi-

[8] A abordagem em espaço de estados (*state-space approach*) é uma estrutura matemática usada para modelar sistemas dinâmicos em que as variáveis observadas são vinculadas a variáveis latentes (não observadas) que evoluem ao longo do tempo. Essa abordagem é composta por duas equações principais, a saber, (1) a equação de medição, que relaciona as variáveis observadas aos estados latentes; e (2) a equação de transição, que descreve como os estados latentes se comportam com o decorrer do tempo.

vamente simples e pouco fundamentada em termos econômicos. Em outras palavras, esses indicadores não estabelecem critérios adequados para a escolha das variáveis e de seus respectivos pesos. Em decorrência dessas questões, os autores se propuseram a desenvolver um PIB mensal do Brasil que fosse capaz de prever a sua variação em trimestres passados ainda não divulgados e, para além disso, prever a sua variação no trimestre corrente.

Para atingir esse objetivo, Issler and Pimentel (2019) construíram seu modelo com base na Equação de Apreçamento dos Ativos, conforme desenvolvida por Cochrane (2002), o que difere dos indicadores anteriormente apresentados. Em relação ao modelo econométrico utilizado, ele foi posto, pelos autores, na forma de espaço de estados, de acordo com o proposto por Bernanke et al. (1997) e Mönch and Uhlig (2005), sendo estimado por meio do Filtro de Kalman. Além disso, os autores utilizaram variáveis auxiliares que foram escolhidas conforme critérios estatísticos e com base em um modelo estrutural que relaciona os *spreads* dos retornos dos ativos com vencimentos diferentes, isto é, eles utilizaram o chamado *term spread*[9].

Por meio desse novo indicador, Issler and Pimentel (2019) esperavam ser capazes de acompanhar a atividade econômica de uma maneira mais eficaz, de forma a fornecer informações com maior frequência e com uma defasagem temporal consideravelmente menor. Os resultados mostraram que o *spread* corrente entre os retornos dos ativos com diferentes maturidades foi de grande importância para a previsão da taxa de crescimento futura do PIB e, para além disso, que seus exercícios de *nowcasting* foram superiores àqueles do IACE e do IBC-Br. Não obstante a capacidade preditiva da abordagem proposta por esses autores, o objetivo do presente trabalho não está na utilização, em um problema de *nowcasting*, de variáveis com elevadas frequências; foram utilizadas variáveis financeiras com frequência mensal, incluindo o saldo da carteira de crédito e a inadimplência da carteira de crédito.

Dada a apresentação da literatura pertinente ao *nowcasting* da taxa de crescimento do PIB brasileiro, reforça-se a hipótese deste estudo, a saber, que as métricas de importância de variáveis oriundas de florestas aleatórias podem contribuir na etapa de seleção das variáveis independentes mais relevantes ao problema em consideração (Lunetta et al., 2004; Bureau et al., 2005; Huang et al., 2005; Qi; Bar-Joseph; Klein-Seetharaman, 2006; Strobl et al., 2008; Huang; Lu; Xu, 2016; Fang et al., 2024).

---

[9] De forma mais precisa, o *spread* é a diferença entre duas taxas ou retornos comparáveis. Formalmente, o *spread* no período t é dado por $s_t = x_{1,t} - x_{2,t}$, onde $x_{1,t}$ é a taxa de retorno 1 e $x_{2,t}$ é a taxa de retorno 2. O *term spread* é um caso específico no qual a diferença é entre a taxa de um título de longo prazo e a de curto prazo, isto é, Term Spread$_t = y_t^{\text{longo}} - y_t^{\text{curto}}$. Esse indicador resume a inclinação da curva de juros, em que os valores positivos indicam uma curva normal, enquanto que valores próximos de zero ou negativos sugerem um aperto monetário e uma possível desaceleração.

3 METODOLOGIA

No presente capítulo, a metodologia utilizada neste trabalho será detalhada. Na primeira seção será apresentado o Modelo de Fatores Dinâmicos (DFM) e, na segunda seção, os métodos de seleção de variáveis serão descritos.

3.1 Modelo de Fatores Dinâmicos (DFM)

Neste momento, prossegue-se para uma descrição formal do problema a ser enfrentado por um modelo que tenha como objetivo realizar uma previsão do PIB em tempo real. O *vintage*[1] de dados disponíveis no período v será denotado por Ω_v, em que v se refere à data de divulgação de um dado específico, e a taxa de crescimento trimestral Q do PIB no tempo t será denotada por y_t^Q. O problema do *nowcasting* de y_t^Q é então definido como a projeção ortogonal[2] de y_t^Q por meio do conjunto informacional disponível Ω_v:

$$P\left[y_t^Q \,\middle|\, \Omega_v\right] \;=\; E\left[y_t^Q \,\middle|\, \Omega_v\right], \tag{3.1}$$

em que $E\left[y_t^Q \mid \Omega_v\right]$ é a esperança de y_t^Q condicionada a Ω_v. Em outras palavras, deseja-se prever o valor esperado do PIB a partir das informações disponíveis no conjunto de dados, sendo que estes foram coletados até o período v (Giannone; Reichlin; Small, 2008; Bańbura; Giannone; Reichlin, 2010; Bańbura; Modugno, 2014).

Um dos elementos que distinguem o *nowcasting* de outros métodos de previsão é a estrutura do conjunto informacional Ω_v. Uma característica particular desse conjunto é a sua borda irregular, que se refere ao fato de que, como as séries de dados são disponibilizadas de maneira assíncrona e com diferentes atrasos, o período $T_{i,v}$ da última observação disponível pode ser diferente para cada série i. Outro aspecto relevante dessa abordagem é que ela incorpora séries com frequências mistas, que, neste caso, são mensais e trimestrais. Dessa forma, tem-se

$$\Omega_v \;=\; \left\{ x_{i,t_i},\; t_i = 1, 2, \ldots, T_{i,v},\; i = 1, \ldots, n;\;\; y_{3k}^Q,\; 3k = 3, 6, \ldots, T_{Q,v} \right\}, \tag{3.2}$$

[1] Um *vintage* é um conjunto de dados disponível em um determinado momento no tempo, isto é, são os dados disponíveis em uma determinada data, o que pode incluir revisões de dados divulgadas até o momento em questão. Em outras palavras, o *vintage* de dados disponíveis no período v, Ω_v, representa todos os dados disponíveis até uma data v qualquer.

[2] Uma projeção ortogonal é uma projeção de um dado vetor em outro espaço, de forma que o erro seja ortogonal ao espaço de projeção. A projeção ortogonal de uma variável Y sobre o espaço de variáveis X é a melhor estimativa de Y baseada em X, no sentido de minimizar o erro quadrático. Em linguagem matemática, $P[Y \mid X] = E[Y \mid X]$, o que significa que $E[(Y - E[Y \mid X])X] = 0$. No caso deste trabalho, a projeção ortogonal é utilizada para obter a melhor estimativa do PIB trimestral a partir deste conjunto de informações.

onde x_{i,t_i} é a observação da série i no período t_i, $T_{i,v}$ corresponde ao último período de observação da série i disponível no *vintage* v, y_{3k}^{Q} são as observações trimestrais da taxa de crescimento do PIB e $3k$ representa os finais de trimestre[3], que se estendem até o período $T_{Q,v}$ (Giannone; Reichlin; Small, 2008; Bańbura; Giannone; Reichlin, 2010; Bańbura; Modugno, 2014).

Em decorrência da assincronia das disponibilizações de dados, $T_{i,v}$ não é o mesmo para todas as variáveis e, portanto, o conjunto de dados exibe uma borda irregular. Portanto, o problema de *nowcasting* precisa ser analisado em um contexto que impõe uma estrutura de probabilidade plausível em Ω_v e que explora, de forma parcimoniosa e informacionalmente ótima (no sentido de minimizar o MSFE, via Filtro de Kalman sob hipóteses lineares–gaussianas), esse conjunto de informações, onde o potencial número de variáveis independentes mensais $x_{i,t}$ é elevado. Outra característica relevante do processo de *nowcasting* é que raramente os especialistas realizam uma única projeção para o trimestre de interesse, realizando, ao contrário, uma sequência de exercícios, que são atualizados à medida que novos dados são divulgados. As primeiras projeções são, com frequência, aplicadas com pouca, ou talvez nenhuma, informação sobre o trimestre de referência. Com a subsequente disponibilização de novos dados, os exercícios de *nowcasting* são então revisados, acarretando em projeções mais precisas à medida que novas informações referentes ao trimestre são divulgadas. Em termos matemáticos, será realizada uma sequência de projeções

$$E\left[y_t^Q \,\middle|\, \Omega_v\right],\ E\left[y_t^Q \,\middle|\, \Omega_{v+1}\right],\ \ldots, \tag{3.3}$$

onde $v, v+1, \ldots$ se referem às datas das consecutivas divulgações de dados. Tipicamente, os intervalos entre duas consecutivas divulgações de dados são pequenos e mudam com o decorrer do tempo. Portanto, v possui uma alta frequência e é espaçado de forma irregular (Giannone; Reichlin; Small, 2008; Bańbura; Giannone; Reichlin, 2010; Bańbura; Modugno, 2014).

Para computar os exercícios de *nowcasting*, tudo que se necessita é, a princípio, realizar projeções lineares. Em termos práticos, há a necessidade de lidar com diversos problemas, como a frequência mista, a borda irregular, possivelmente outros casos de dados ausentes e a questão da dimensionalidade que decorre da riqueza das informações disponíveis que, caso sejam incluídas, podem acarretar em estimativas imprecisas e voláteis (Giannone; Reichlin; Small, 2008; Bańbura; Giannone; Reichlin, 2010; Bańbura; Modugno, 2014).

A abordagem aqui apresentada segue as elaborações de Giannone, Reichlin and Small (2008) e, portanto, oferece uma solução a esses problemas por meio da modelagem de

[3] Os finais de trimestre são os meses que fecham os trimestres, isto é, os meses $3, 6, 9$ e 12 dentro de um ano qualquer. Esses meses marcam o fim de cada trimestre até o período $T_{Q,v}$.

dados mensais como um Modelo de Fatores Dinâmicos (DFM) parametrizado apresentado em uma representação de espaço de estados. Quando obtida essa representação, as técnicas do Filtro de Kalman podem ser usadas para realizar as projeções dos fatores, já que elas automaticamente se adaptam à disponibilidade de dados em constante mudança. Destaca-se que a representação do modelo de fatores permite a inclusão de muitas variáveis, o que é desejável para este contexto, no qual inúmeras séries podem conter informações relevantes acerca da variável dependente. No que se refere à estimação, adota-se a abordagem de estimação do modelo por máxima verossimilhança de Bańbura and Modugno (2014), que é uma abordagem factível e robusta no contexto de modelos de fatores de larga escala (Doz; Giannone; Reichlin, 2006; Giannone; Reichlin; Small, 2008; Bańbura; Giannone; Reichlin, 2010; Bańbura; Modugno, 2014).

Seja $x_t = (x_{1,t}, x_{2,t}, \ldots, x_{n,t})'$ a representação das séries mensais, que foram transformadas de forma a satisfazer a premissa de estacionariedade. Mais precisamente, x_t são as taxas de crescimento mês a mês das variáveis originais. Assume-se, de forma semelhante, que x_t obedece à seguinte representação do modelo de fatores:

$$x_t \;=\; \mu \;+\; \Lambda f_t \;+\; \varepsilon_t, \tag{3.4}$$

onde μ é um vetor $n \times 1$ de médias incondicionais, Λ é uma matriz $n \times r$ que denota as cargas fatoriais para as variáveis mensais, f_t é um vetor $r \times 1$ de fatores latentes (não observáveis) comuns no tempo t e, por fim, ε_t é um vetor $n \times 1$ de componentes idiossincráticos (Giannone; Reichlin; Small, 2008; Bańbura; Giannone; Reichlin, 2010; Bańbura; Modugno, 2014).[4]

A Equação (3.4) pode ser chamada de equação de medição do modelo fatorial, já que descreve como as séries observadas x_t estão relacionadas com os fatores latentes f_t. Nesse contexto, pressupõe-se que os fatores comuns e os componentes idiossincráticos possuem média zero e, portanto, as constantes $\mu = (\mu_1, \mu_2, \ldots, \mu_n)'$ são as médias incondicionais. Portanto, os fatores comuns são modelados como um processo VAR de ordem p:

$$f_t \;=\; A_1 f_{t-1} \;+\; \cdots \;+\; A_p f_{t-p} \;+\; u_t, \qquad u_t \overset{i.i.d.}{\sim} \mathcal{N}(0, Q), \tag{3.5}$$

onde $A_1, \ldots, A_p$ são as matrizes $r \times r$ de coeficientes autorregressivos, que medem o quanto os valores passados dos fatores comuns latentes influenciam seus valores presentes, e u_t são os choques nos fatores comuns, que possuem uma distribuição normal de média zero e matriz de covariância Q (Giannone; Reichlin; Small, 2008; Bańbura; Giannone; Reichlin, 2010; Bańbura; Modugno, 2014).

[4] Considera-se importante enfatizar que n denota o número de variáveis independentes e r denota o número de fatores latentes. Dessa forma, a matriz Λ possui o número de linhas igual ao número de variáveis independentes e o número de colunas igual ao número de fatores latentes.

Finalmente, assume-se que o componente idiossincrático das variáveis mensais segue um processo autorregressivo de primeira ordem, um AR(1):

$$\varepsilon_{i,t} \;=\; \alpha_i\, \varepsilon_{i,t-1} \;+\; e_{i,t}, \qquad e_{i,t} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_i^2), \tag{3.6}$$

em que α_i é o coeficiente de autocorrelação da variável i e $e_{i,t}$ é um ruído branco normal com média zero e variância σ_i^2. Além disso, $E[e_{i,t}e_{j,s}] = 0$ para todo $i \neq j$, de forma que os choques idiossincráticos não estejam correlacionados para diferentes séries (Giannone; Reichlin; Small, 2008; Bańbura; Giannone; Reichlin, 2010; Bańbura; Modugno, 2014).

A relevância da Equação (3.6) encontra-se em sua capacidade de permitir que as séries possuam, em certa medida, uma dinâmica própria, de forma que não sejam explicadas somente por fatores comuns. Destaca-se que é importante considerar, de forma explícita, as dinâmicas dos fatores nas aplicações de *nowcasting*. O motivo para isso está no fato de que, em decorrência dos atrasos nas publicações, a informação acerca dos períodos mais recentes pode ser escassa e de que explorar as dinâmicas, em adição às contemporâneas relações, pode incrementar a precisão das estimativas dos fatores comuns (Giannone; Reichlin; Small, 2008; Bańbura; Giannone; Reichlin, 2010; Bańbura; Modugno, 2014).

Em relação à modelagem das variáveis trimestrais, segue-se a abordagem de Mariano and Murasawa (2003), de forma a incorporar as variáveis trimestrais no modelo por meio da construção, para cada uma delas, de um correspondente mensal parcialmente observado. Em outras palavras, trata-se de uma identidade de agregação temporal embutida em um modelo em espaço de estados, não sendo uma *proxy* nem uma técnica mecânica de interpolação. Considerando que estes exercícios são para o *nowcasting* da taxa de crescimento trimestral do PIB, adota-se a convenção segundo a qual o valor da variável trimestral é atribuído ao terceiro mês do respectivo trimestre. De acordo com essa convenção, o nível trimestral do PIB, que é denotado por GDP_t^Q, $t = 3, 6, 9, \ldots$, pode ser expresso como a soma de suas contribuições mensais não observadas, GDP_t^M:

$$GDP_t^Q \;=\; GDP_t^M \;+\; GDP_{t-1}^M \;+\; GDP_{t-2}^M, \qquad t = 3, 6, 9, \ldots \tag{3.7}$$

em que $Y_t^Q = 100 \times \log\Big(GDP_t^Q\Big)$ e $Y_t^M = 100 \times \log\Big(GDP_t^M\Big)$, de forma que Y_t^Q seja o logaritmo do nível do PIB trimestral e Y_t^M seja o logaritmo do nível do PIB mensal, que é uma variável latente. Em outras palavras, o PIB em nível de um dado trimestre GDP_t^Q é igual à soma dos PIBs mensais em nível dos três meses que compõem esse trimestre $\Big(GDP_t^M + GDP_{t-1}^M + GDP_{t-2}^M\Big)$, com os PIBs mensais sendo tratados como latentes e com o PIB trimestral observado atuando como restrição de agregação (Giannone; Reichlin; Small, 2008; Bańbura; Giannone; Reichlin, 2010; Bańbura; Modugno, 2014).

De forma semelhante, assume-se que a taxa de crescimento mensal não observada do PIB, $y_t = \Delta Y_t^M$, admite a mesma representação de modelo de fatores que as variáveis mensais:

$$y_t \;=\; \mu_Q \;+\; \Lambda_Q f_t \;+\; \varepsilon_t^Q, \tag{3.8}$$

$$\varepsilon_t^Q \;=\; \alpha_Q \varepsilon_{t-1}^Q \;+\; e_t^Q, \qquad e_t^Q \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma_Q^2\right), \tag{3.9}$$

onde μ_Q é a média incondicional do crescimento mensal, Λ_Q é a matriz de cargas fatoriais e e_t^Q é o erro idiossincrático, específico ao PIB. Conforme pode-se observar na Equação (3.9), o erro ε_t^Q possui uma estrutura AR(1) (Giannone; Reichlin; Small, 2008; Bańbura; Giannone; Reichlin, 2010; Bańbura; Modugno, 2014).

Para conectar y_t com os dados observados do PIB, é construída uma série mensal parcialmente observada:

$$y_t^* \;=\; \begin{cases} Y_t^Q - Y_{t-3}^Q, & \text{se } t = 3, 6, 9, \ldots, \\ \text{não observado}, & \text{caso contrário.} \end{cases} \tag{3.10}$$

Em outras palavras, a Equação (3.10) transforma, nos meses que fecham um trimestre (março, junho, setembro e dezembro), o valor trimestral do PIB em uma escala mensal. Por outro lado, os demais meses (janeiro, fevereiro, abril, etc.) permanecem sem observações para essa variável. Como o Filtro de Kalman é capaz de lidar com observações ausentes, mesmo com a conversão das observações trimestrais em uma série mensal com valores ausentes, o modelo consegue trabalhar com o PIB trimestral dentro do aparato mensal (Giannone; Reichlin; Small, 2008; Bańbura; Giannone; Reichlin, 2010; Bańbura; Modugno, 2014).

Após esse processo, utiliza-se a aproximação de Mariano and Murasawa (2003):

$$\begin{aligned} y_t^Q &= Y_t^Q - Y_{t-3}^Q \\ &\approx \left(Y_t^M + Y_{t-1}^M + Y_{t-2}^M\right) \\ &\quad - \left(Y_{t-3}^M + Y_{t-4}^M + Y_{t-5}^M\right) \\ &= y_t + 2y_{t-1} + 3y_{t-2} + 2y_{t-3} + y_{t-4}, \qquad t = 3, 6, 9, \ldots \end{aligned} \tag{3.11}$$

Por meio da aproximação realizada pela Equação (3.11), o crescimento trimestral pode ser aproximado por uma média móvel[5] em pirâmide dos cinco crescimentos mensais

[5] Uma média móvel é simplesmente uma média calculada usando diversos valores passados de uma série temporal. Por exemplo, uma média móvel simples de 3 meses para uma série y_t é dada por $\bar{y}_t = \frac{y_t + y_{t-1} + y_{t-2}}{3}$. Em outras palavras, combinam-se valores recentes para suavizar e/ou representar uma tendência.

mais recentes, com os pesos apresentados no formato 1–2–3–2–1. Em outras palavras, o crescimento trimestral é baseado em níveis acumulados ao longo do tempo, de forma que, quando realizados os cálculos, o crescimento trimestral se torna uma combinação suavizada dos crescimentos mensais. Trata-se de uma estrutura mais espalhada no tempo, onde o modelo consegue utilizar o crescimento trimestral observado para informar o crescimento mensal latente. Essa expressão é fundamental para o DFM, já que permite utilizar séries com diferentes frequências sem a necessidade de uma interpolação do PIB, fornecendo coerência temporal ao modelo (Giannone; Reichlin; Small, 2008; Bańbura; Giannone; Reichlin, 2010; Bańbura; Modugno, 2014).

Em relação à estimação e à previsão, define-se $\bar{x}_t = (x_t', y_t^Q)'$ e $\bar{\mu} = (\mu', \mu_Q)'$, de forma que o modelo conjunto possa ser representado na forma de espaço de estados:

$$\bar{x}_t \;=\; \bar{\mu} \;+\; Z(\theta)\,\alpha_t, \tag{3.12}$$

$$\alpha_t \;=\; T(\theta)\,\alpha_{t-1} \;+\; \eta_t, \qquad \eta_t \overset{i.i.d.}{\sim} \mathcal{N}\Big(0, \Sigma_\eta(\theta)\Big), \tag{3.13}$$

onde $\bar{x}_t$ é um vetor empilhado das variáveis observadas, incluindo as variáveis independentes mensais e a variável dependente trimestral; $\bar{\mu}$ é um vetor contendo os níveis médios associados às variáveis observadas; $Z(\theta)$ é uma matriz que conecta o estado latente da economia aos dados observados; α_t é um vetor de estados inclui os fatores comuns f_t e os componentes idiossincráticos ε_t; $T(\theta)$ é a matriz de transição dos estados, que mostra a persistência temporal dos fatores latentes; α_{t-1} é o vetor de estados no período anterior; e η_t é o vetor de choques do estado, representando tudo que ocorre de inesperado na economia entre os períodos $t-1$ e t (Giannone; Reichlin; Small, 2008; Bańbura; Giannone; Reichlin, 2010; Bańbura; Modugno, 2014).

Em outras palavras, a Equação (3.12) corresponde à equação de observação (ou de medição) do modelo em espaço de estados. Essa equação formaliza que o vetor $\bar{x}_t$ de variáveis observadas no período t pode ser decomposto em dois componentes: (1) um nível médio $\bar{\mu}$, que captura o comportamento típico das séries; e (2) um componente sistemático $Z(\theta)\alpha_t$, em que $Z(\theta)$ é a matriz de cargas que traduz o estado latente α_t na dinâmica das variáveis observáveis. Assim, essa equação descreve como a informação econômica observada é gerada a partir de um conjunto reduzido de fatores latentes, que sintetizam a movimentação conjunta entre as séries.

Por sua vez, a Equação (3.13) é a equação de transição (ou de estado) do modelo em espaço de estados. Essa equação descreve a dinâmica temporal do vetor de estados α_t, estabelecendo que o estado corrente depende, em certa medida, do estado passado α_{t-1} por meio da matriz de transição $T(\theta)$, que governa o grau de persistência temporal dos componentes latentes. A evolução do estado é, igualmente, afetada por um termo de

choques η_t, assumido como ruído branco com média zero e matriz de covariâncias $\Sigma_\eta(\theta)$. Em conjunto, essas duas equações permitem representar o DFM de modo a combinar muitas séries observadas com a evolução de alguns poucos estados latentes, possibilitando a atualização e a previsão em tempo real a partir do Filtro, e do Suavizador, de Kalman. No caso em que $p \leq 5$, sendo p a ordem do processo autorregressivo dos fatores comuns, tem-se:

$$\alpha_t = \left(f_t', f_{t-1}', f_{t-2}', f_{t-3}', f_{t-4}', \varepsilon_{1,t}, \ldots, \varepsilon_{n,t}, \varepsilon_t^Q, \varepsilon_{t-1}^Q, \varepsilon_{t-2}^Q, \varepsilon_{t-3}^Q, \varepsilon_{t-4}^Q\right)'. \tag{3.14}$$

Em outras palavras, a Equação (3.14) define, de forma clara e explícita, a composição do vetor de estados α_t, mostrando que é formado (1) pelos fatores latentes f_t contemporâneos e defasados; (2) pelos componentes idiossincráticos $\varepsilon_{i,t}$ das variáveis independentes mensais; e (3) pelos componentes idissioncráticos ε_t^Q associados à variável dependente trimestral, tanto contemporâneos quanto defasados.

Todos os parâmetros do modelo são coletados no vetor θ de parâmetros do modelo. Neste ponto da discussão é prudente destacar que θ é estimado por meio da máxima verossimilhança implementada no algoritmo *Expectation-Maximisation* (EM). Essa abordagem foi proposta para grandes conjuntos de dados por Doz, Giannone and Reichlin (2006) e aprimorada por Bańbura and Modugno (2014) para lidar com observações ausentes e dinâmicas idiossincráticas. Por outro lado, Giannone, Reichlin and Small (2008) usaram um procedimento de duas etapas, a saber, estimar os parâmetros do modelo como estimativas de fatores por meio de componentes principais para, depois, serem estimados novamente usando o Filtro de Kalman (Giannone; Reichlin; Small, 2008; Bańbura; Giannone; Reichlin, 2010; Bańbura; Modugno, 2014).

A estimação de máxima verossimilhança usando o algoritmo EM consiste na iteração da abordagem de duas etapas, isto é, em estimar os parâmetros condicionados às estimativas dos fatores da iteração anterior e vice-versa. O método de máxima verossimilhança permite lidar com características centrais do modelo, a saber, uma fração considerável dos dados estando ausente, as restrições sobre os parâmetros e a correlação serial dos componentes idiossincráticos. Dada uma estimativa de θ, os exercícios de *nowcasting*, bem como as estimativas dos fatores e de quaisquer observações ausentes em $\bar{x}_t$, podem ser obtidos por meio do Filtro, ou do Suavizador, de Kalman. A abordagem desses autores pressupõe que os dados possuem uma dinâmica dirigida por alguns poucos fatores comuns não observáveis. Em outras palavras, o Suavizador de Kalman utiliza o filtro previamente construído e refina, em cada período, as estimativas dos fatores latentes usando toda a amostra, produzindo uma imputação coerente e estatisticamente eficiente para observações ausentes na base de dados, o que é fundamental ao trabalhar com bordas irregulares (Giannone; Reichlin; Small, 2008; Bańbura; Giannone; Reichlin, 2010; Bańbura; Modugno,

2014).

Para realizar os exercícios foi utilizado o pacote `dfms`, disponível no R, que apresenta uma abordagem amigável e computacionalmente eficiente de estimar os DFMs. O pacote realiza uma estimação eficiente do DFM utilizando o algoritmo EM ou a estimação *Two-Step* (2S), permitindo o uso de conjuntos de dados com observações ausentes. De forma semelhante, assume-se que os fatores seguem um VAR estacionário de ordem p. As opções de estimação seguem os avanços na literatura econométrica: (1) por meio do Filtro de Kalman, e de seu Suavizador, com valores iniciais estimados por PCA-2S, conforme Doz, Giannone and Reichlin (2011); (2) por meio da iteração do Filtro de Kalman, bem como de seu Suavizador, até a convergência EM, seguindo Doz, Giannone and Reichlin (2012); ou (3) por meio do algoritmo EM adaptado de Bańbura and Modugno (2014), permitindo padrões arbitrários de observações ausentes (Krantz, 2025a; Krantz, 2025b).

O pacote também fornece um critério de informação para escolher o número ótimo de fatores latentes comuns, seguindo Bai and Ng (2002), e permite a estimação com frequência mista, a saber, mensal e trimestral, de acordo com as formulações de Mariano and Murasawa (2003) e Bańbura and Modugno (2014). Como trabalha-se com o *nowcasting* do PIB, enfrentando o problema da borda irregular e lidando com um conjunto de dados contendo variáveis independentes mensais e uma variável dependente trimestral, utiliza-se a abordagem de Bańbura and Modugno (2014), pois permite trabalhar com muitos valores ausentes e com diferentes frequências (Krantz, 2025a; Krantz, 2025b).

De forma semelhante, é prudente ressaltar que a literatura mostra que existem ganhos em termos de acurácia preditiva quando ocorre uma seleção prévia das variáveis independentes (Bai; Ng, 2008; Kim; Swanson, 2018). Dessa forma, optou-se por realizar o *nowcasting* utilizando o conjunto completo de variáveis e, para além disso, utilizando somente as variáveis selecionadas por quatro técnicas distintas de seleção, a saber, o LASSO, o ENET e a IPB da RF tanto com MBB quanto com CBB. Também deve-se ressaltar que foram empregados dados em tempo pseudo-real[6], isto é, a base de dados utilizada considerou atrasos constantes na divulgação de cada observação de cada variável. Na próxima seção, será apresentada uma descrição mais detalhada de cada uma dessas técnicas de seleção de variáveis.

[6] No contexto de *nowcasting*, os dados em tempo real são aqueles que realmente estavam disponíveis na data da previsão; isto é, são considerados os atrasos de publicação de cada observação de cada variável. Por outro lado, os dados em tempo pseudo-real (*pseudo-real time*) são uma aproximação, no sentido de que se assumem atrasos de publicação constantes ao longo do tempo (Evans, 2005; Giannone; Reichlin; Small, 2008; Bantis; Clements; Urquhart, 2023). Além disso, assume-se, no presente trabalho, que os dados utilizados sejam referentes às suas primeiras divulgações, e não às revisões, em decorrência da profunda dificuldade na obtenção das observações divulgadas antes de suas revisões.

3.2 Métodos de seleção de variáveis

No presente estudo, são utilizados quatro métodos de redução de dimensionalidade e/ou de encolhimento, a saber, o *Least Absolute Shrinkage and Selection Operator* (LASSO), o *Elastic Net* (ENET) e a seleção das k variáveis mais relevantes a partir da importância por permutação em blocos (IPB) da *Random Forest* (RF), tanto com *Moving Block Bootstrap* (MBB) quanto com *Circular Block Bootstrap* (CBB). O número de variáveis mais importantes selecionadas pelas florestas em cada período foi escolhido para ser igual ao número de variáveis selecionadas pelo LASSO no mesmo período, que é a técnica de referência para a redução de dimensionalidade na literatura de *nowcasting* do PIB. Esta seção possui como objetivo descrever os objetivos, fundamentos e implicações dessas técnicas.

3.2.1 LASSO

Agora o método LASSO será abordado, método este que é frequentemente utilizado na literatura de *nowcasting*, sendo também utilizado no presente estudo. Esse método foi originalmente formulado no seminal artigo de Tibshirani (1996) e, como seu nome sugere, ele opera de modo a encolher os coeficientes e, ao mesmo tempo, selecionar um conjunto de variáveis de um modelo linear, deixando o modelo em questão mais enxuto e menos complexo, evitando os problemas de ajuste excessivo[7]. O LASSO minimiza a Soma dos Quadrados dos Resíduos (RSS, da sigla em inglês)[8] sujeita à soma do valor absoluto dos coeficientes sendo menor que uma constante[9]. Em decorrência dessa restrição, esse operador tende a transformar alguns coeficientes em exatamente zero, tornando o modelo mais interpretável. O autor argumenta, igualmente, que o LASSO apresenta propriedades desejáveis tanto no que se refere à Seleção de Subconjuntos quanto à Regressão Ridge.

Considere-se a situação comum de regressão, na qual tem-se os dados (x^i, y_i), $i = 1, 2, \ldots, N$, onde $x^i = (x_{i1}, \ldots, x_{ip})^T$, com p variáveis independentes para a observação i, e y_i é a variável dependente correspondente à i-ésima observação.[10] O método padrão de estimação é por meio dos Mínimos Quadrados Ordinários (MQO ou OLS), que minimiza

[7] O problema do ajuste excessivo, também chamado de "*overfitting*" ou "sobreajuste", ocorre quando um modelo se adapta de forma excessiva aos dados de treinamento, de forma a generalizar mal para novos dados e, portanto, dificultar o processo de previsão. No presente trabalho será utilizado somente o termo "ajuste excessivo", já que é um termo em português que captura a essência do problema.

[8] Matematicamente, a Soma dos Quadrados dos Resíduos pode ser definida como RSS $= \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$, onde y_i é o valor observado da variável dependente e $\hat{y}_i$ é o valor previsto da variável dependente.

[9] Em linguagem matemática, a soma do valor absoluto dos coeficientes ser menor do que o de uma constante pode ser escrita como $\sum_{j=1}^{p}|\beta_j| \leq t$.

[10] O sobrescrito T no vetor $(x_{i1}, \ldots, x_{ip})^T$ significa que se trata de um vetor transposto, isto é, o vetor que originalmente era linha tornou-se coluna após a transposição.

o erro quadrático residual. Não obstante, de acordo com Tibshirani (1996), existem dois motivos pelos quais esse método por vezes não é satisfatório: (1) com frequência as estimativas de MQO possuem baixo viés, mas elevada variância, de forma que a acurácia preditiva possa, em algumas ocasiões, ser incrementada ao tornar alguns coeficientes iguais a zero; e (2) com um grande número de variáveis independentes, com frequência é apropriado determinar um menor subconjunto de variáveis que exibam os efeitos mais relevantes, de forma a garantir uma maior interpretabilidade ao modelo em questão.

Para além disso, as duas técnicas tradicionais para o aprimoramento das estimativas de MQO, a Seleção de Subconjuntos e a Regressão Ridge, possuem suas desvantagens. A primeira dessas técnicas fornece modelos interpretáveis, mas pode ser extremamente variável, no sentido de que pequenas mudanças nos dados possam acarretar em modelos consideravelmente diferentes, em decorrência de ser um processo discreto. A segunda delas é um processo contínuo que de fato encolhe os coeficientes e, portanto, é mais estável; todavia, essa técnica não zera nenhum dos coeficientes, de forma que não produza um modelo facilmente interpretável. Em decorrência dessas questões, o autor propõe o LASSO, que encolhe alguns coeficientes e transforma outros em zero, de forma a manter as propriedades desejáveis tanto da Seleção de Subconjuntos quanto da Regressão Ridge (Tibshirani, 1996).

De forma semelhante, Tibshirani (1996) supõe, conforme o procedimento padrão em análises de regressão, que as observações são independentes ou que os y_i sejam condicionalmente independentes dados os x_{ij}. O autor também assume que as variáveis independentes x_{ij} sejam padronizadas, de modo que $\frac{1}{N}\sum_{i=1}^{N} x_{ij} = 0$ e $\frac{1}{N}\sum_{i=1}^{N} x_{ij}^2 = 1$, isto é, cada variável independente foi centralizada por meio da subtração de sua respectiva média amostral e normalizada com variância unitária.[11] Sendo $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$, a estimativa do LASSO $(\hat{\alpha}, \hat{\beta})$ é definida por

$$\begin{aligned}(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha,\beta} \left\{ \sum_{i=1}^{N} \Big(y_i - \alpha - \sum_{j=1}^{p} \beta_j x_{ij} \Big)^2 \right\} \\ \text{sujeito a } \sum_{j=1}^{p} |\beta_j| \leq t.\end{aligned} \tag{3.15}$$

Nesse ponto da elaboração, Tibshirani (1996) destaca que $t \geq 0$ é o parâmetro de ajuste, isto é, o parâmetro responsável pelo grau de encolhimento (ou restrição) aplicado às

[11] Sendo mais preciso, centralizar um vetor de observações significa subtrair cada observação particular da média amostral, criando um novo vetor centralizado com somatório igual a 0. Em termos matemáticos, dado um vetor de observações $x = (x_1, ..., x_n)$, o vetor centralizado é dado por $x_i^c = (x_1 - \bar{x}, ..., x_n - \bar{x})$, onde $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, o que implica que $\sum_{i=1}^{n} x_i^c = 0$. Por sua vez, normalizar um vetor de observações significa reescalá-lo de modo que sua magnitude passe a ser igual a 1, preservando-se a sua direção. Em termos matemáticos, dado um vetor não nulo $v \in \mathbb{R}^n$, sua versão normalizada é definida por $\tilde{v} = \frac{v}{\|v\|_2}$, onde $\|v\|_2$ é a norma euclidiana de v, de forma que $\|\tilde{v}\|_2 = 1$.

estimativas. Agora, para todo t, a solução para α é $\hat{\alpha} = \bar{y}$, em decorrência da centralização dos dados, que passam a ter média igual a zero. O autor afirma que pode-se assumir, sem perda de generalidade, que $\bar{y} = 0$ e, portanto, omitir α, já que a centralização de y_i simplifica os cálculos necessários, sem alterar a solução para β_j. Em outras palavras, a Equação (3.15) define o estimador LASSO, que minimiza o somatório do erro residual quadrático da regressão linear — o quadrado da diferença entre o valor observado y_i e o valor previsto $\alpha + \sum_j \beta_j x_{ij}$ — ao mesmo tempo em que impõe uma restrição sobre o somatório dos valores absolutos dos coeficientes β_j das variáveis independentes.

Caso o parâmetro de ajuste t seja igual ao parâmetro de referência $t_0 = \sum_j |\hat{\beta}_j^0|$, os coeficientes permanecerão os mesmos e o modelo se comportará como uma tradicional regressão por MQO; por outro lado, se $t < t_0$, o tamanho dos coeficientes será reduzido, podendo, caso esse t seja suficientemente pequeno, tornar alguns dos coeficientes exatamente iguais a zero, de forma que suas respectivas variáveis sejam removidas do modelo. Em outras palavras, o parâmetro de referência t_0 representa o somatório dos valores absolutos dos coeficientes estimados presentes antes da inclusão de uma penalização, de forma que, quanto mais próximo t estiver de t_0, menor será a penalização aplicada aos coeficientes das variáveis independentes. Por exemplo, se $t = \frac{t_0}{2}$, o efeito será aproximadamente similar a encontrar o menor subconjunto de variáveis independentes de tamanho $p/2$. Em decorrência dessas características, pode-se afirmar que o LASSO realiza, simultaneamente, o encolhimento dos coeficientes das variáveis independentes e a seleção das variáveis relevantes (Tibshirani, 1996; James et al., 2013).

Em relação à Regressão Ridge, James et al. (2013) ressaltam que essa técnica possui uma clara desvantagem, a saber, incluir todas as variáveis independentes no modelo final, ao contrário de métodos como a Seleção de Melhor Subconjunto (*Best Subset Selection*). A penalização $\lambda \sum \beta_j^2$ da Regressão Ridge encolhe todos os coeficientes em direção a zero, sem, contudo, tornar nenhum deles exatamente zero. Este aspecto pode não ser um problema para a acurácia preditiva, mas impõe um desafio à interpretação quando o número de variáveis independentes é grande. A Regressão Ridge sempre construirá um modelo com todas as variáveis independentes; aumentar λ apenas reduz a magnitude dos coeficientes sem eliminar variáveis. James et al. (2013) afirmam que o LASSO é uma alternativa recente que supera essa desvantagem de não eliminar variáveis. De forma semelhante, pode-se expressar a Equação (3.15) da seguinte forma:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \; + \; \lambda\sum_{j=1}^{p}|\beta_j| \; = \; \text{RSS} \; + \; \lambda\sum_{j=1}^{p}|\beta_j|. \tag{3.16}$$

A Regressão Ridge e o LASSO possuem formulações similares, com a diferença de que o termo de penalização β_j^2 do primeiro é substituído pelo termo $|\beta_j|$ no segundo. Em linguagem estatística, o LASSO utiliza a penalidade ℓ_1 em vez da penalidade ℓ_2. A norma

ℓ_1 de um vetor de coeficientes β é dada por $\|\beta\|_1 = \sum_j |\beta_j|$. Assim como a Regressão Ridge, o LASSO encolhe as estimativas dos coeficientes em direção a zero; todavia, a penalidade ℓ_1 pode forçar algumas estimativas a serem exatamente zero quando o parâmetro de ajuste λ é suficientemente elevado. Em outras palavras, tal como a Melhor Seleção de Subconjuntos, o LASSO desempenha um papel de seleção de variáveis. Por isso, James et al. (2013) consideram os modelos obtidos via LASSO mais facilmente interpretáveis do que aqueles gerados pela Regressão Ridge, pois produzem os chamados modelos esparsos, isto é, modelos que envolvem apenas um subconjunto das variáveis originais.

Da mesma forma que ocorre na Regressão Ridge, no LASSO é fundamental escolher um valor adequado para λ. Quando $\lambda = 0$, o LASSO simplesmente resultará no mesmo conjunto de variáveis independentes originalmente disponível, resolvendo o mesmo problema de regressão linear por MQO. Por outro lado, quando λ é suficientemente elevado, o LASSO retorna um modelo nulo, no qual todas as estimativas dos coeficientes são iguais a zero. Dessa forma, a depender do valor do parâmetro de ajuste λ, o LASSO pode produzir um modelo que incorpore qualquer número de variáveis, variando de zero variáveis ao total de variáveis originalmente disponível. Por outro lado, a Regressão Ridge sempre manterá todas as variáveis independentes no modelo, apesar de que a magnitude dos coeficientes dependerá do parâmetro de ajuste (Tibshirani, 1996; James et al., 2013).

Para além disso, o grau de regularização do modelo decorrente da aplicação do LASSO pode ser observada a partir da razão $\frac{\|\hat{\beta}_\lambda\|_1}{\|\hat{\beta}^0\|_1}$, onde o numerador representa a norma ℓ_1 dos coeficientes estimados pelo LASSO, dado um parâmetro de ajuste λ, e o denominador representa a norma ℓ_1 dos coeficientes estimados por MQO, dado um $\lambda = 0$. Destaca-se que, quando $\lambda = 0$, o estimador do LASSO coincide com o estimador de MQO, de modo que a razão assume valor igual a 1. À medida que λ aumenta, a penalização torna-se mais intensa, reduzindo a magnitude dos coeficientes e possivelmente reduzindo alguns deles a zero. No limite, quando a penalização é suficientemente elevada, todos os coeficientes podem ser encolhidos a zero, resultando em um modelo nulo. Para valores da razão entre 0 e 1, o modelo reduzirá alguns coeficientes e possivelmente zerará outros; quanto menor a quantidade de variáveis independentes, mais parcimonioso e interpretável será o modelo em questão (Tibshirani, 1996; James et al., 2013).

Ambas as técnicas analisadas, o LASSO e a Regressão Ridge, possuem uma íntima conexão com a Seleção de Melhor Subconjunto, que, diferentemente desses dois métodos, não impõe uma restrição ℓ_1 nem uma ℓ_2. Essa técnica impõe uma condição de $\sum_{j=1}^{p} \mathbb{I}(\beta_j \neq 0) \leq s$, onde $\mathbb{I}(\beta_j \neq 0)$ é uma variável indicadora que assume o valor 1 caso $\beta_j \neq 0$ e o valor 0 caso $\beta_j = 0$, de forma que a restrição imponha que no máximo s variáveis podem ter coeficientes diferentes de zero. Em outras palavras, a variável indicadora é igual a 1 caso a variável j entre no modelo, dado seu coeficiente diferente de zero, e igual a 0 caso a variável j não entre, dado seu coeficiente igual a zero. Portanto, a Seleção de

Melhor Subconjunto almeja encontrar um conjunto de estimativas dos coeficientes em que o erro de ajuste RSS seja o menor possível, sujeito à restrição de que não mais do que s coeficientes possam ser não nulos, limitando o modelo a um total de s variáveis independentes (Tibshirani, 1996; James et al., 2013).

Não obstante, de acordo com James et al. (2013), a resolução da técnica de Seleção de Melhor Subconjunto não é factível, em termos computacionais, quando o número de variáveis independentes é muito grande, já que requer considerar todos os modelos $\binom{p}{s}$, onde p é a quantidade de variáveis originalmente disponíveis e s é a quantidade de variáveis que permanecerá após a seleção. Em decorrência dessa questão, pode-se interpretar o LASSO e a Regressão Ridge como alternativas computacionalmente factíveis à Seleção de Melhor Subconjunto. Evidentemente, o LASSO está muito mais próximo da Seleção de Melhor Subconjunto, já que desempenha uma seleção de variáveis independentes quando o s é suficientemente pequeno em sua restrição. James et al. (2013) afirmam estar evidente que o LASSO possui uma vantagem considerável sobre a Regressão Ridge, já que ele produz modelos mais simples e interpretáveis que contêm somente um subconjunto das variáveis independentes originais.

O LASSO acarreta um comportamento qualitativamente semelhante àquele da Regressão Ridge, no sentido de que, à medida que o parâmetro de ajuste λ aumenta, a variância diminui e o viés aumenta. Todavia, o LASSO pressupõe, implicitamente, que existem coeficientes das variáveis independentes que são exatamente iguais a zero. Em decorrência disso, é razoável que, em um contexto no qual todas as variáveis independentes possuem uma relação com a variável dependente, a Regressão Ridge possua um desempenho superior ao do LASSO em termos de erro de previsão. De forma semelhante, é razoável que, em um contexto no qual apenas uma pequena parcela das variáveis independentes possua uma relação com a variável dependente, o LASSO apresente um desempenho melhor que a Regressão Ridge em termos de viés, variância e MSE[12]. Em outras palavras, nenhuma das duas técnicas — o LASSO e a Regressão Ridge — domina, de forma universal, a outra. Não obstante, a quantidade de variáveis independentes que está relacionada à variável dependente nunca é conhecida, *a priori*, para conjuntos de dados reais, de forma que não seja conhecido, pelo menos não a princípio, qual dos dois métodos é o mais apropriado (Tibshirani, 1996; James et al., 2013).

Uma técnica como a validação cruzada pode ser utilizada com o objetivo de determinar qual das abordagens é mais adequada. Em relação à atribuição de um valor ao parâmetro de ajuste λ, é necessário que métodos sejam empregados, na implementação do LASSO e da Regressão Ridge, para encontrar o valor adequado a esse parâmetro. A validação cruzada fornece uma maneira simples de lidar com essa questão por meio

[12] O *Mean Squared Error* (MSE), sinônimo de *Mean Squared Forecast Error* (MSFE), é o erro quadrático médio, que mensura a média do quadrado da diferença entre o valor observado y_i e o valor $\hat{y}_i$ estimado pelo modelo. Em termos matemáticos: $\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

da escolha de um conjunto de valores de λ, do cálculo do erro de validação cruzada para cada valor de λ, da seleção do valor no qual esse erro foi o menor e do ajuste do modelo a partir das observações disponíveis e do valor escolhido do parâmetro λ. Inclusive, existem situações em que, em uma Regressão Ridge, o encolhimento é consideravelmente pequeno, com o ajuste resultante sendo muito similar à solução por MQO e com inúmeros valores de λ resultando em erros muito próximos. Nesse contexto, pode-se simplesmente adotar a solução por MQO. Por outro lado, existem casos nos quais o conjunto de dados possui inúmeras variáveis de ruído, irrelevantes, de forma que, à medida que a razão $\frac{\|\hat{\beta}_\lambda\|_1}{\|\hat{\beta}^0\|_1}$ diminui, pouquíssimas variáveis independentes permanecem no modelo, com o menor erro de validação cruzada estando no ponto em que esse quociente possui um valor bem reduzido (Tibshirani, 1996; James et al., 2013).

Apesar de suas importantes vantagens, o LASSO possui algumas limitações, que devem ser consideradas nesta análise preditiva. De acordo com Erp, Oberski and Mulder (2019), existem cinco dificuldades na utilização dessa técnica: (1) o LASSO não consegue selecionar mais variáveis independentes do que observações, o que é limitante quando, em um problema, a quantidade dos primeiros é maior que a dos segundos; (2) quando um conjunto de variáveis independentes possui uma correlação considerável, o LASSO frequentemente seleciona apenas uma dessas variáveis, problema esse que é resolvido pelo ENET; (3) o erro de previsão é maior do que aquele relativo à Regressão Ridge quando o número de observações é maior que o de variáveis independentes e estas são altamente correlacionadas; (4) ele pode acarretar em um encolhimento excessivo dos coeficientes grandes, já que o estimador encolhe os coeficientes para próximo de zero, gerando um viés adicional; e (5) o LASSO nem sempre possui a propriedade de oráculo[13], o que implica que essa técnica nem sempre possui um desempenho na seleção de variáveis tão bom quanto o que apresentaria caso o verdadeiro modelo fosse dado.

A propriedade de oráculo é apresentada pelo LASSO somente sob específicas e rigorosas condições. Essas questões estimularam o surgimento de diversas ramificações do LASSO, bem como de generalizações. Para Zou (2006), parece válido concluir que o LASSO é um procedimento de oráculo por atingir, de forma simultânea, uma seleção de variáveis consistente e uma estimação (previsão) ótima. Entretanto, existem também argumentos muito sólidos que se contrapõem a essa proposição, dos quais se destacam aqueles desenvolvidos por Fan and Li (2001), que demonstram que as duas condições para a propriedade de oráculo — a consistência na seleção de variáveis independentes e a eficiência assintótica — não podem ser simultaneamente satisfeitas no LASSO com a penalidade L_1, não possuindo, conforme a conjectura dos autores, essa propriedade. Por

[13] A propriedade de oráculo é caracterizada pela capacidade de uma técnica de selecionar, de forma correta, as variáveis independentes relevantes para o problema em questão e, além disso, de estimar seus coeficientes com uma eficiência assintótica, isto é, onde a acurácia preditiva é incrementada à medida que o tamanho da amostra aumenta (Erp; Oberski; Mulder, 2019).

sua vez, a penalidade *smoothly clipped absolute deviation* (SCAD), elaborada por Fan and Li (2001), e o LASSO adaptativo, elaborado por Zou (2006), satisfazem as condições necessárias para a propriedade de oráculo. Além disso, destaca-se que o LASSO adaptativo se compara de maneira favorável em relação a outras técnicas de modelagem esparsa, conforme as simulações realizadas por Zou (2006).

Não obstante, o LASSO é uma técnica padrão na literatura de *nowcasting* das taxas de crescimento do PIB, já que, mesmo com todas as dificuldades anteriormente apresentadas, possui ganhos consideráveis em termos de acurácia preditiva (Cepni; Güney; Swanson, 2019a, 2019b; Bantis; Clements; Urquhart, 2023). Dessa forma, considera-se prudente trabalhar também com o LASSO, de forma a avaliar seu desempenho frente às outras técnicas. Para séries temporais, entretanto, é prudente utilizar uma *expanding window* ou uma *rolling window*, de forma a manter a estrutura de dependência temporal, estrutura essa que é violada por uma validação cruzada tradicional, com *k-fold* (Rossi; Inoue, 2012; Feng; Zhang; Wang, 2023).[14] Por esses motivos, primeiro foi criada uma grade de 100 valores possíveis para o parâmetro de ajuste λ, espaçados igualmente entre 0,0001 e 1 na escala logarítmica; a criação desses valores é justificada pela necessidade de selecionar o valor ótimo de λ, isto é, que apresenta o menor MSFE.

Posteriormente, foi criada uma função de *expanding window 1-step-ahead cross-validation*, por meio da qual, para cada janela temporal t, foram estimados simultaneamente 100 modelos LASSO, um para cada λ, utilizando o IBC-Br[15] como variável dependente e as outras 42 séries mensais como variáveis independentes; depois, foi calculado o MSFE de cada um desses modelos para, então, encontrar o valor ótimo do parâmetro de ajuste λ. Por fim, foi estimado um modelo final do LASSO, usando todo o conjunto de treinamento, com o λ ótimo; depois, foram selecionadas as variáveis que não tiveram seus coeficientes reduzidos a zero. Dessa forma, por meio da utilização da *expansive window*, foi possível manter a estrutura temporal dos dados e, ao mesmo tempo, simular previsões em tempo real, já que a base de dados aumenta à medida que avançam o tempo e as previsões. Essas regressões foram realizadas usando o pacote `glmnet`, disponível no R.

[14] Neste ponto da discussão, é prudente esclarecer que os termos *recursive forecasting* e *expanding window forecasting* são frequentemente empregados como sinônimos. Ambos descrevem um procedimento no qual o modelo é reestimado sequencialmente à medida que novas observações se tornam disponíveis, de forma que o conjunto de estimação cresce ao longo do tempo. Formalmente, para gerar a previsão de y_{t+1}, o modelo é estimado com a amostra $\{1, \ldots, t\}$; para prever y_{t+2}, utiliza-se a amostra $\{1, \ldots, t+1\}$; e assim sucessivamente. Dessa forma, a janela de estimação é expandida progressivamente, sem descartar as observações passadas (Rossi; Inoue, 2012; Feng; Zhang; Wang, 2023).

[15] A escolha do IBC-Br como variável dependente decorre de sua frequência mensal. Como o LASSO e as demais técnicas de seleção de variáveis não trabalham com variáveis de diferentes frequências, tornou-se necessário escolher uma variável de frequência mensal para ser a variável dependente. Considerando que o IBC-Br é o principal indicador coincidente do PIB, considera-se prudente utiliza-lo para essa função.

3.2.2 ENET

Uma das principais técnicas de regularização e seleção de variáveis é o *Elastic Net* (ENET), originalmente formulado por Zou and Hastie (2005). Em seu trabalho seminal, os autores mostram que, em exercícios com dados reais e de simulação, o ENET apresentou desempenho frequentemente superior ao do LASSO, mantendo uma esparsidade semelhante àquela. Uma característica muito desejável do ENET é a capacidade de lidar com variáveis independentes correlacionadas por meio de um efeito de agrupamento, no qual variáveis altamente correlacionadas tendem a permanecer ou sair do modelo juntas. Além disso, o ENET é muito útil quando o número de variáveis independentes é consideravelmente maior do que o de observações, o que contrasta com o LASSO, que não apresenta esse aspecto desejável. Embora o estimador do LASSO continue bem definido nesse cenário, ele é capaz de selecionar, no máximo, uma quantidade de variáveis distintas igual ao número de observações, antes de saturar o ajuste. De forma semelhante, o LASSO tende a escolher apenas uma variável dentre um conjunto fortemente correlacionado, descartando as demais, o que decorre da natureza da penalização L_1, que impõe soluções esparsas, apesar de não estabilizar adequadamente na presença de uma forte multicolinearidade.

Os critérios para avaliar a qualidade de um modelo podem ser resumidos em sua capacidade preditiva sobre dados futuros e, ao mesmo tempo, em sua interpretabilidade. A partir de uma maior parcimônia do modelo decorre uma capacidade de iluminar melhor a relação entre a variável dependente e as independentes, conferindo maior interpretabilidade ao modelo em questão. De forma semelhante ao que ocorre no LASSO, a motivação dos criadores do ENET se encontra na superação do MQO, da Regressão Ridge e da Melhor Seleção de Subconjuntos, tanto em termos de previsão quanto de interpretabilidade; a diferença neste caso está no objetivo de superar, também, o LASSO, pelo menos em certos contextos. Apesar de trabalhos como os de Tibshirani (1996) e Fu (1998) terem descoberto que nenhuma das três técnicas analisadas (LASSO, Regressão Ridge e Regressão Bridge) domina universalmente as outras, os autores argumentam que, com a crescente importância da seleção de variáveis na moderna análise de dados, o LASSO se torna muito atrativo por sua esparsidade (Zou; Hastie, 2005).

Em outras palavras, em muitos problemas aplicados, os pesquisadores se deparam com a presença de muitas variáveis independentes, forte multicolinearidade e elevado ruído, principalmente em cenários de *nowcasting* com dezenas ou centenas de variáveis mensais. Quando a quantidade de variáveis independentes cresce em excesso, podem surgir problemas de ajuste excessivo, instabilidade dos coeficientes e baixa capacidade de generalização para fora da amostra; a seleção de variáveis surge exatamente para amenizar esses problemas. O objetivo desses autores é que o ENET funcione tão bem quanto o LASSO nas situações em que este é a melhor técnica disponível e, ao mesmo tempo, apresente, no cenário de variáveis independentes altamente correlacionadas, desempenho preditivo superior ao do

LASSO. Da mesma forma que o LASSO, o ENET realiza, simultaneamente, seleção de variáveis e encolhimento contínuo, podendo selecionar grupos de variáveis independentes correlacionadas (Zou; Hastie, 2005).

O método de ENET ingênuo de Zou and Hastie (2005) supõe que o conjunto de dados tenha n observações e p variáveis independentes, sendo $y = (y_1, \ldots, y_n)^T$ a variável dependente e $X = (x_1, \ldots, x_p)$ a matriz de variáveis independentes do modelo, onde $x_j = (x_{1j}, \ldots, x_{nj})^T$, com $j = 1, \ldots, p$, são as p variáveis independentes. Após uma transformação de escala e localização, assume-se que a variável dependente está centralizada e que as variáveis independentes estão padronizadas. Em termos matemáticos:

$$\sum_{i=1}^{n} y_i = 0, \qquad \sum_{i=1}^{n} x_{ij} = 0 \quad \text{e} \quad \sum_{i=1}^{n} x_{ij}^2 = 1, \quad j = 1, 2, \ldots, p, \tag{3.17}$$

onde $\sum_{i=1}^{n} y_i = 0$ significa que a variável dependente foi centralizada; $\sum_{i=1}^{n} x_{ij} = 0$ significa que cada variável independente também foi centralizada; e $\sum_{i=1}^{n} x_{ij}^2 = 1$ significa que cada variável independente foi escalada para ter norma igual a 1, de forma que todas as variáveis estejam na mesma escala. Para quaisquer parâmetros de regularização $\lambda_1, \lambda_2 \geq 0$ fixos, o ENET ingênuo é definido pela função de perda

$$L(\lambda_1, \lambda_2, \beta) = \|y - X\beta\|^2 \; + \; \lambda_2 \|\beta\|^2 \; + \; \lambda_1 \|\beta\|_1, \tag{3.18}$$

onde y é o vetor da variável dependente; X é a matriz de variáveis independentes; β é o vetor de coeficientes; $\|y - X\beta\|^2$ é a soma dos quadrados dos resíduos; $\lambda_2\|\beta\|^2$ é a parte da Regressão Ridge, que penaliza coeficientes grandes; e $\lambda_1\|\beta\|_1$ é a parte do LASSO, que penaliza a magnitude absoluta dos coeficientes. De forma semelhante, tem-se $\|\beta\|^2 = \sum_{j=1}^{p} \beta_j^2$ e $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$. Em outras palavras, a Equação (3.18) define a função de perda do ENET ingênuo, sendo composta pelo erro do ajuste e pelas penalizações Ridge e LASSO.[16]

Assim sendo, o estimador $\hat{\beta}$ do ENET ingênuo é o minimizador da Equação (3.18), sendo dado por

$$\hat{\beta} \; = \; \arg\min_{\beta} \Big\{ L(\lambda_1, \lambda_2, \beta) \Big\}. \tag{3.19}$$

[16] Neste ponto da discussão, é prudente ressaltar que a norma euclidiana, também conhecida como norma L_2, é uma medida de comprimento ou magnitude de um vetor no espaço $\mathbb{R}^n$. Dado um vetor $v = (v_1, v_2, \ldots, v_n)'$, sua norma euclidiana é definida por $\|v\|_2 = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2} = \sqrt{v'v}$. Geometricamente, essa quantidade representa a distância do ponto v até a origem, generalizando o Teorema de Pitágoras para dimensões superiores. Por sua vez, o quadrado da norma euclidiana é simplesmente $\|v\|_2^2 = v_1^2 + v_2^2 + \cdots + v_n^2 = v'v$. Ou seja, trata-se da soma dos quadrados dos elementos do vetor; essa expressão frequentemente aparece na soma dos quadrados dos resíduos, como no termo $\|y - X\beta\|^2$ da Equação (3.18).

No decorrer de seu trabalho, Zou and Hastie (2005) desenvolvem uma formulação aprimorada do ENET, de forma a resolver eficientemente as limitações do ENET ingênuo. Apesar deste superar as limitações do LASSO nos cenários (1) de um número de variáveis independentes maior que o de observações e (2) de variáveis independentes altamente correlacionadas, ele pode incorrer em um encolhimento excessivo dos coeficientes, por se tratar de um método automático de seleção de variáveis. As evidências apresentadas pelos autores sugerem que o ENET ingênuo só apresenta um bom desempenho quando está muito próximo do LASSO ou da Regressão Ridge, que são justamente os casos limite. No contexto preditivo de regressão, um método acurado de penalização adquire bom desempenho por meio do *trade-off* entre viés e variância, tendo em vista que a redução do erro sistemático frequentemente eleva a instabilidade do modelo, ao passo que a redução da instabilidade frequentemente aumenta o erro sistemático.

O estimador de ENET ingênuo é uma abordagem de duas etapas, a saber: (1) para cada λ_2 fixado, encontram-se primeiro os coeficientes da Regressão Ridge e, então, (2) realiza-se um encolhimento com o LASSO ao longo de suas trajetórias de solução. Esse procedimento acarreta no dobro de encolhimento, o que pouco ajuda na redução da variância e introduz viés adicional e desnecessário quando comparado ao LASSO ou à Regressão Ridge (Zou; Hastie, 2005). Em decorrência disso, os autores aprimoraram a técnica por meio da correção desse encolhimento duplo. Considerando dados (y, X) e (λ_1, λ_2), as estimativas $\hat{\beta}$ do ENET são dadas por

$$\hat{\beta} \;=\; \arg\min_{\beta} \; \beta^T \left(\frac{X^T X + \lambda_2 I}{1 + \lambda_2} \right) \beta \;-\; 2\, y^T X \beta \;+\; \lambda_1 \left\| \beta \right\|_1. \tag{3.20}$$

Em decorrência disso, pode-se observar que

$$\hat{\beta}(\text{LASSO}) \;=\; \arg\min_{\beta} \; \beta^T (X^T X) \beta \;-\; 2\, y^T X \beta \;+\; \lambda_1 \left\| \beta \right\|_1. \tag{3.21}$$

Portanto, essa nova e aprimorada formulação do ENET interpreta-o como uma versão estabilizada do LASSO. Em outras palavras, a Equação (3.20) do ENET aprimorado combina as penalizações L_1, do LASSO, e L_2, da Regressão Ridge; dessa forma, esse ENET aprimorado corrige o excesso de viés decorrente do encolhimento duplo do ENET ingênuo, que primeiro executava a Regressão Ridge e depois o LASSO. Por outro lado, a Equação (3.21) do LASSO escolhe os coeficientes β que minimizam a soma dos quadrados dos resíduos, dada por $\|y - X\beta\|^2$, e uma penalização $\lambda_1 \left\| \beta \right\|_1$ proporcional ao tamanho absoluto dos coeficientes. Por fim, é prudente ressaltar que, para séries temporais altamente correlacionadas, o LASSO pode escolher apenas uma variável do grupo, enquanto o ENET tende a selecionar grupos de variáveis correlacionadas; trata-se do efeito de agrupamento, encontrado pelos autores em seu trabalho (Zou; Hastie, 2005).

No presente trabalho, foi utilizado o pacote `glmnet`, disponível no R, para realizar as regressões do LASSO e do ENET com *expanding window 1-step-ahead cross-validation* para a escolha do valor ótimo do termo de penalização λ. Os procedimentos foram os mesmos que aqueles realizados no LASSO, alterando somente o argumento que define o tipo de regressão entre LASSO, ENET e Regressão Ridge. O pacote `glmnet` executa procedimentos extremamente eficientes para ajustar todo o caminho de regularização do LASSO ou do ENET para modelos de regressão linear, logística, multinomial, dentre outros. O algoritmo utiliza a descida de coordenadas, é extremamente rápido e explora a esparsidade na matriz de variáveis independentes quando ela existe. Uma variedade de previsões pode ser realizada a partir dos modelos ajustados. O algoritmo ajusta um modelo linear generalizado via máxima verossimilhança penalizada usando o método *k-fold* de validação cruzada; como as variáveis deste trabalho são todas séries temporais, o *k-fold* foi substituído por uma função de *expanding window 1-step-ahead cross-validation*. Salienta-se que a função utilizada para implementar o ENET realiza um ENET ingênuo, misturando as penalizações L_1 e L_2; a função resolve diretamente o problema do ENET, minimizando a função-objetivo por descida de coordenadas, utilizando o valor previamente calculado do parâmetro de ajuste λ. Além disso, optou-se por estabelecer pesos iguais para ambas as normas e, dessa forma, utilizar um ENET balanceado (Friedman et al., 2025; Hastie; Qian; Tay, 2025). Neste estudo, foi realizada uma regressão de ENET utilizando o IBC-Br como variável dependente e as outras 42 séries mensais como variáveis independentes.

3.2.3 Random Forest

Uma das principais técnicas de *Machine Learning* é a *Random Forest* (RF), que pode ser aplicada tanto em problemas de classificação quanto de regressão. Esse método consiste na construção de um *ensemble* de múltiplas árvores de decisão, geradas a partir de diferentes subconjuntos de dados e variáveis independentes. A ideia central é que, embora cada árvore individual seja um preditor relativamente instável e sujeito à alta variância, a combinação das previsões de muitas árvores, por meio da votação, no caso do problema de classificação, ou da média, no caso do problema de regressão, resulta em um modelo final mais robusto e preciso. Essa estratégia reduz a variância e melhora o desempenho em relação ao uso de árvores isoladas. Além disso, a RF fornece uma forma consistente de medir a importância das variáveis independentes, permitindo identificar quais delas contribuem de maneira mais significativa para a previsão. Neste trabalho explora-se essa propriedade de mensuração da relevância dessas variáveis para selecionar as variáveis independentes mais relevantes, auxiliando na tarefa de *nowcasting* da taxa de crescimento do PIB brasileiro (Breiman, 2001; James et al., 2013).

Em seu trabalho seminal, Breiman (2001) descreve a significativa melhoria na acurácia de classificação por meio da utilização de um conjunto de árvores que votam na

classe mais popular, com frequência gerando vetores aleatórios que governam o crescimento de cada árvore no conjunto. A característica comum a todas essas técnicas é que, para a k-ésima árvore, um vetor aleatório Θ_k é criado, possuindo a mesma distribuição que os vetores anteriores $\Theta_1, \ldots, \Theta_{k-1}$ ao mesmo tempo em que é independente deles. Assim sendo, a árvore é construída por meio de um conjunto de treinamento e de um vetor aleatório Θ_k, de forma a resultar no classificador $h(x, \Theta_k)$. Depois que muitas árvores são criadas elas votam na classe mais popular.

Assim, Breiman (2001) define uma floresta aleatória como um classificador consistindo de uma coleção de classificadores com estrutura de árvore $\{h(x, \Theta_k),\ k = 1, \ldots\}$, onde os vetores aleatórios $\{\Theta_k\}$ são independentes e identicamente distribuídos e, além disso, cada árvore fornece um voto unitário para a classe mais popular na variável independente x. A ideia central é que a combinação de inúmeras árvores, que são classificadores fracos, acarreta em um classificador forte e robusto, isto é, as RFs se fundamentam no conceito de *ensemble*. Em outras palavras, $h(x, \Theta_k)$ é uma árvore de decisão, ou classificador baseado em árvore, em que x é um vetor de variáveis independentes com uma observação e p variáveis independentes e Θ_k é um vetor aleatório que introduz um componente aleatório na construção da árvore em questão, de forma que cada árvore da RF seja construída com alguma aleatoriedade. Além disso, é importante enfatizar que a floresta é formada por inúmeras árvores, quantidade essa que é delimitada pelo subscrito k (James et al., 2013).

As árvores aleatórias apresentam um progresso relevante sobre as árvores criadas com o *bagging*[17], progresso esse que decorre de uma modificação que remove a correlação entre as árvores. De forma semelhante ao *bagging*, inúmeras árvores de decisão são construídas nos dados de treinamento utilizando um *bootstrap*. Não obstante, em todos os momentos da construção das árvores nos quais uma divisão em uma árvore é considerada, uma amostra aleatória das variáveis independentes é escolhida como candidata à divisão, ao invés do conjunto completo de variáveis independentes. Dessa forma, o algoritmo de RF sequer pode considerar a maioria das variáveis independentes disponíveis (James et al., 2013).

A lógica que fundamenta essa ideia é que, quando existe uma variável independente muito forte em um conjunto de dados, a maioria das árvores em um *bagging* utilizará essa variável como a divisão principal, de forma que as árvores fiquem muito parecidas, com suas previsões sendo altamente correlacionadas. Por sua vez, a média de muitas árvores com correlação elevada não acarreta em uma grande redução da variância; isto

[17] O *bagging*, ou *bootstrap aggregation*, é um método de *ensemble* cujo objetivo é diminuir a variância de um método de aprendizado estatístico, sendo frequentemente utilizado no contexto de árvores de decisão. Em linhas gerais: (1) o *bootstrap* gera uma quantidade de amostras dos dados de treinamento, cada uma do mesmo tamanho da amostra original, com reposição; (2) treina-se o modelo em cada amostra *bootstrap*; (3) realiza-se a agregação, que, no caso de regressão, é feita pela média das previsões; e, por fim, (4) avalia-se o desempenho nas observações que ficaram fora de cada *bootstrap* (James et al., 2013).

é, o *bagging* não reduzirá, neste caso e em grau considerável, a variância em relação a uma única árvore. Esse problema é resolvido pelas RFs quando elas obrigam que cada divisão considere apenas um subconjunto das variáveis independentes. Esse processo pode ser interpretado, inclusive, como uma descorrelação das árvores, tornando a média delas menos volátil e mais confiável (James et al., 2013). Essas florestas foram amplamente utilizadas, e com bons resultados, em diversos campos científicos (Svetnik et al., 2003; Díaz-Uriarte; Andrés, 2006; Cutler et al., 2007).

Dado um *ensemble* de classificadores $h_1(X), h_2(X), \ldots, h_K(X)$ e um conjunto de treinamento aleatoriamente retirado do vetor aleatório (Y, X), Breiman (2001) define a função de margem como

$$mg(X,Y) \;=\; \frac{1}{K}\sum_{k=1}^{K} I\Big(h_k(X) = Y\Big) \;-\; \max_{j \neq Y} \; \frac{1}{K}\sum_{k=1}^{K} I\Big(h_k(X) = j\Big). \tag{3.22}$$

em que X é o vetor de variáveis independentes; Y é o valor da variável dependente, sendo, portanto, um escalar; $h_k(x)$ é o classificador produzido pela k-ésima árvore da floresta, dado o vetor x; $I(\cdot)$ é a função indicadora, que assume o valor 1 caso a condição seja verdadeira e 0 caso contrário; e a margem $mg(X, Y)$ mede em que medida o voto médio para a classe correta excede o voto médio para qualquer outra classe. Quanto maior a margem, maior a confiança na classificação (Breiman, 2001).

Por sua vez, o erro de generalização é dado por

$$PE^* \;=\; P_{X,Y}\Big(mg(X,Y) < 0\Big), \tag{3.23}$$

onde os subscritos X, Y indicam que a probabilidade é calculada sobre o espaço conjunto (X, Y) e $P_{X,Y}(\cdot)$ calcula a probabilidade sobre a distribuição conjunta dos dados. Nas RFs, $h_k(X) = h(X, \Theta_k)$, isto é, cada árvore da floresta é um classificador que depende de um vetor aleatório Θ_k. Para um grande número de árvores, afirma Breiman (2001), pela Lei Forte dos Grandes Números e pela estrutura das árvores, tem-se que, à medida que o número de árvores aumenta, quase certamente todas as sequências $\Theta_1, \ldots, \Theta_K$ fazem PE^* convergir para

$$P_{X,Y}\left(P_{\Theta}\Big(h(X,\Theta) = Y\Big) - \max_{j \neq Y} P_{\Theta}\Big(h(X,\Theta) = j\Big) < 0\right). \tag{3.24}$$

Os resultados acima explicam por que, em vez de sofrerem um ajuste excessivo quando mais árvores são adicionadas, as RFs produzem um valor limitante para o erro de generalização. Embora o teorema seja formulado para classificação, em regressão com perda quadrática ocorre fenômeno análogo, a saber, o MSFE da floresta tende a um limite que depende do ruído irredutível e da correlação entre os erros das árvores (Breiman, 2001).

Um aspecto das RFs que é crucial para o presente trabalho é a métrica de importância das variáveis obtida via erro *out-of-bag*[18]. Suponha-se que existam M variáveis independentes e que, após cada árvore ser gerada, os valores da m-ésima variável nas observações OOB sejam embaralhados de forma aleatória e, em seguida, essas observações sejam passadas pela mesma árvore. Para cada observação x_n, a respectiva classificação (ou previsão) OOB é salva e compara-se a taxa de erro com e sem a perturbação da variável m; a perda média resultante mede a importância dessa variável (Breiman, 2001).

Um procedimento semelhante é usado para medir a importância das variáveis em um problema de regressão, com a diferença de que tanto as variáveis independentes da árvore $h(x, \Theta)$ quanto a variável dependente assumem valores numéricos ao invés de rótulos de classe. Além disso, assume-se que o conjunto de treinamento é retirado, de forma independente, da distribuição de um vetor aleatório (Y, X) (Breiman, 2001). O erro quadrático médio para qualquer preditor numérico h(x) é dado por

$$E_{X,Y}(Y - h(X))^2\,, \tag{3.25}$$

em que o preditor da floresta aleatória é formado ao tomar-se a média de k árvores $\{\, h(x, \Theta_k)\,\}$. Em outras palavras, no contexto de regressão cada árvore fornece um valor numérico, com a média desses resultados sendo a previsão final e o MSE sendo o critério de erro da floresta (Breiman, 2001). Assumindo que, para todo Θ, $E_Y = E_X h(X, \Theta)$, tem-se

$$PE^*(\text{floresta}) \;\leq\; \bar{\rho}\, PE^*(\text{árvore}), \tag{3.26}$$

em que $\bar{\rho}$ é a correlação ponderada entre os resíduos $Y - h(X, \Theta)$ e $Y - h(X, \Theta')$, com Θ e Θ' independentes. Em outras palavras, o erro da floresta é menor ou igual ao erro médio de uma única árvore, multiplicado por uma medida de correlação entre os resíduos das diferentes árvores. Para que as florestas de regressão sejam precisas, é desejável que os resíduos tenham baixa correlação e que as árvores individuais apresentem baixos erros (Breiman, 2001).

A lógica da importância por permutação (IP) está fundamentada na ideia de que, ao aleatoriamente permutar a variável independente X_j, sua associação original com a variável dependente Y é quebrada. Quando a variável permutada X_j é usada, em conjunto com as variáveis independentes restantes que não foram permutadas, para prever a variável dependente para as observações OOB, a acurácia preditiva decresce consideravelmente

[18] O termo *out-of-bag* (OOB) se refere às observações que ficam de fora de uma amostra *bootstrap* utilizada para treinar uma árvore em um modelo de RF. Como a reamostragem é realizada com reposição, em média cerca de um terço das observações de treinamento não são selecionadas em cada *bootstrap* e, portanto, permanecem disponíveis para a validação daquela árvore. Essas observações OOB funcionam como uma avaliação interna, pois cada árvore pode ser testada justamente nas observações que ela não utilizou no treinamento (James et al., 2013).

caso a variável original X_j esteja associada à variável dependente. Portanto, Breiman (2001) sugere que a diferença na acurácia preditiva antes e depois de permutar X_j, após feita a média para todas as árvores, seja tratada como uma medida de importância de variáveis (Breiman, 2001; Strobl et al., 2008). Em termos formais, seja $\bar{B}^{(t)}$ a amostra OOB para uma árvore t, com $t \in \{1, \ldots, ntree\}$, então, a importância da variável X_j na árvore t é

$$\mathrm{VI}^{(t)}(X_j) = \frac{\sum\limits_{i \in \bar{\mathcal{B}}^{(t)}} \mathbb{I}\Big\{ y_i = \hat{y}_i^{(t)} \Big\}}{|\bar{\mathcal{B}}^{(t)}|} - \frac{\sum\limits_{i \in \bar{\mathcal{B}}^{(t)}} \mathbb{I}\Big\{ y_i = \hat{y}_{i,\pi_j}^{(t)} \Big\}}{|\bar{\mathcal{B}}^{(t)}|}, \tag{3.27}$$

onde $\hat{y}_i^{(t)} = f^{(t)}(\mathbf{x}_i)$ é a classe prevista para a observação i pela árvore t antes da permutação e $\hat{y}_{i,\pi_j}^{(t)} = f^{(t)}(\mathbf{x}_{i,\pi_j})$ é a classe prevista para a observação i após permutar seus valores da variável X_j. Aqui, $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,p})$ denota o vetor de variáveis independentes da observação i, com $\mathbf{x}_{i,\pi_j} = (x_{i,1}, \ldots, x_{i,j-1}, x_{\pi_j(i),j}, x_{i,j+1}, \ldots, x_{i,p})$, isto é, o mesmo vetor de variáveis independentes de i, exceto pelo componente j, que foi substituído pelo valor correspondente a uma permutação π_j das posições. A função $\mathbb{I}\{\cdot\}$ é o indicador, que vale 1 quando a proposição é verdadeira e 0 caso contrário. Destaca-se que, por definição, $\mathrm{VI}^{(t)}(X_j) = 0$ se a variável X_j não é utilizada na árvore t (Breiman, 2001; Strobl et al., 2008). A importância bruta da variável X_j na floresta é então calculada como a média das importâncias nas árvores:

$$\mathrm{VI}(X_j) = \frac{1}{\texttt{ntree}} \sum_{t=1}^{\texttt{ntree}} \mathrm{VI}^{(t)}(X_j)\,. \tag{3.28}$$

Na implementação padrão das *Random Forests*, é fornecida uma versão adicional e escalada da importância por permutação, frequentemente chamada de z-score, que é obtida dividindo-se a importância bruta pelo seu erro-padrão. No entanto, como resultados indicam que a importância bruta $\mathrm{VI}(X_j)$ apresenta melhores propriedades estatísticas, considere-se apenas a versão não escalada (Díaz-Uriarte; Andrés, 2006; Strobl; Zeileis, 2008; Strobl et al., 2008).

Não obstante, a utilização das RFs em problemas de séries temporais requer um cuidado adicional, já que perde-se a estrutura de dependência temporal, em decorrência da premissa de observações independentes. Para lidar com esse obstáculo, Goehry et al. (2023) apresentaram algumas variantes para a aplicação das RFs em problemas de séries temporais, onde o *bootstrap* padrão é substitúido por um *bootstrap* em blocos dependentes, de modo a subamostrar as séries temporais durante a construção das árvores. Em outras palavras, utilizando blocos inteiros preserva-se, dentro de cada bloco, as relações temporais das variáveis, que são fundamentais para problemas de previsão de séries temporais.

Uma primeira variante é apresentada por Carlstein (1986), a saber, o *Non-overlapping Block Bootstrap* (NBB). A ideia que fundamenta esta abordagem está em

construir um determinado número de blocos não sobrepostos e, então, amostrar, de forma uniforme e com reposição, entre os blocos construídos. Mais precisamente, seja l_n o tamanho de um bloco, isto é, quantas observações consecutivas estão dentro do bloco em questão; e seja $B \geq 1$ o maior inteiro tal que $l_n B \leq n$, isto é, o número total de observações ocupadas pelos B blocos, dado por $l_n B$, não pode ultrapassar o número total n de observações da amostra. Dessa forma, os blocos são construídos da seguinte maneira:

$$B_b = \left((X_{(b-1)l_n+1}, Y_{(b-1)l_n+1}), \ldots, (X_{bl_n}, Y_{bl_n}) \right), \quad b = 1, \ldots, B. \tag{3.29}$$

O conjunto *bootstrap* $\mathcal{D}_n^\star$, que nada mais é do que a amostra *bootstrap* extraída a partir da amostra original, é então obtido ao se extrair K blocos, $(B_1^\star, \ldots, B_K^\star)$, de forma uniforme com reposição na coleção de blocos não sobrepostos $(B_b)_{1 \leq b \leq B}$, para um K escolhido de maneira apropriada. Em outras palavras, para criar uma amostra *bootstrap* $\mathcal{D}_n^\star$, são sorteados K blocos da coleção de blocos não sobrepostos da série original, com todos os blocos tendo a mesma chance de serem escolhidos e permitindo repetições. Posteriormente, esses K blocos são agrupados na ordem em que foram sorteados para criar uma nova série, de tamanho aproximadamente igual ao da amostra original (Goehry et al., 2023).

Por sua vez, Künsch (1989) e Liu and Singh (1992) introduziram o *Moving Block Bootstrap* (MBB), cuja ideia central é, em vez de sortear aleatoriamente uma única observação dentre as n observações, como no *bootstrap* padrão, sorteia-se aleatoriamente um bloco de l_n observações consecutivas. Repetindo esse procedimento e concatenando os blocos selecionados, isto é, agrupando os blocos de forma consecutiva, obtém-se uma série temporal *bootstrap* que preserva a estrutura de dependência temporal dentro de cada bloco. Mais precisamente, define-se o bloco

$$B_{i,l_n} = ((X_i, Y_i), \ldots, (X_{i+l_n-1}, Y_{i+l_n-1})), \tag{3.30}$$

que se inicia na observação (X_i, Y_i) para $i \in \{1, \ldots, n - l_n + 1\}$, onde n é o número total de observações da série, l_n é o tamanho do bloco *bootstrap* e i é a posição da primeira observação que entra no bloco. Ou seja, as observações do bloco se iniciam em $i = 1$ e terminam em $i = n - l_n + 1$, já que, para que o bloco esteja contido dentro da amostra, é necessário que $i + l_n - 1 \leq n$.[19] O método consiste em sortear aleatoriamente K índices $(I_j)_{1 \leq j \leq K}$ de forma uniforme no conjunto $\{1, \ldots, n - l_n + 1\}$ e associar a cada índice o bloco correspondente $(B_{I_j})_{1 \leq j \leq K}$. O conjunto *bootstrap* é então definido como

[19] De forma mais detalhada, as observações contidas no bloco B_{i,l_n} variam de (X_i, Y_i) até $(X_{i+l_n-1}, Y_{i+l_n-1})$. Dessa forma, a condição para que o bloco esteja contido dentro da amostra é que $i + l_n - 1 \leq n$, o que implica que $i \leq n - l_n + 1$. Portanto, o índice i de início do bloco só pode assumir valores no conjunto $i \in \{1, \ldots, n - l_n + 1\}$, que corresponde a todos as observações da série a partir das quais um bloco de comprimento l_n pode estar contido no interior da amostra.

$\mathcal{D}_n^\star = (B_{I_1}, \ldots, B_{I_K})$, o que significa que é formado por uma sequência de blocos, que foram sorteados de forma aleatória; a concatenação desses blocos forma a nova série temporal a ser usada como amostra *bootstrap* (Goehry et al., 2023).

Entretanto, pode-se observar que as extremidades da série recebem menos peso, já que as observações do meio aparecem em muitos blocos diferentes, enquanto as observações das pontas aparecem em menos blocos, o que pode gerar um viés não desprezível na média. Para corrigir esse problema, Politis and Romano (1992) introduzem o *Circular Block Bootstrap* (CBB), cuja ideia central é enrolar a série temporal em um círculo, de modo que, depois da última observação, retorne-se para a primeira observação. Assim, não existe mais um começo ou fim da amostra, de forma que todas as observações fiquem no meio. Por exemplo, se a série em questão possui apenas seis observações, de X_1 a X_6, então a série é tratada de forma que, após X_6, venha X_1. Dessa forma, a ordenação será $X_1, X_2, X_3, X_4, X_5, X_6, X_1, X_2, ...$, continuando indefinidamente, como em um círculo (Goehry et al., 2023).

Em termos matemáticos, $X_i := X_{i_n}$, onde $i_n = i \bmod n$, e $X_0 := X_n$, e então aplica-se o mesmo procedimento do *bootstrap* em blocos móveis, mas agora com o índice I sorteado uniformemente no conjunto $\{1, \ldots, n\}$. Em outras palavras, X_i é definido como X_{i_n}, com i_n sendo o resto da divisão inteira de i por n, e X_0 sendo definido como X_n, que é a última observação da série. Retornando ao exemplo de uma série com somente seis observações, $X_i := X_{i_6}$, com $i_6 = i \bmod 6$, e $X_0 := X_6$. Assim sendo, $X_7 = X_{7 \bmod 6} = X_1$, já que, em um círculo composto por $n = 6$ observações, a primeira observação após a última, que é X_6, é justamente X_1. Portanto, a série continua indefinidamente, com as observações sempre se repetindo (Goehry et al., 2023).

A proposta de Goehry et al. (2023) para incorporar a estrutura de dependência temporal é substituir a primeira etapa de construção de cada árvore aleatória no procedimento da *Random Forest*, isto é, substituir o *bootstrap* tradicional por uma variante de *bootstrap* em blocos. O algoritmo proposto considera a dependência apenas durante a fase de *bootstrap*, aplicada diretamente às entradas $(X_i, Y_i)_{1 \leq i \leq n}$. Uma vez obtida a amostra *bootstrap*, a etapa de divisão dos nós é realizada como no caso independente. O algoritmo adaptado encontra-se apresentado no Algoritmo 1, destacando-se as modificações relativas ao procedimento original da *Random Forest*.

Algorithm 1: Random Forest para séries temporais

Entrada: $(X_1, Y_1), \ldots, (X_n, Y_n)$

Parâmetros: M, α_n, m_{try}, τ_n, l_n

para $j \leftarrow 1$ **to** M **faça**

- Sortear $\alpha_n \leq n$ observações usando um *block bootstrap* com parâmetro l_n.
- Repetir recursivamente, em cada nó resultante, até que o critério de parada seja satisfeito:
  - Em cada nó, selecionar aleatoriamente m_{try} variáveis.
  - Escolher a melhor divisão segundo o critério da variância entre as variáveis selecionadas.
  - Dividir o nó de acordo com o ponto de corte escolhido.

Saída: Para uma nova observação x, a previsão é dada pela média das M previsões fornecidas pelas árvores para x.

Fonte: Adaptado de Goehry et al. (2023).

A partir das elaborações acima, pode-se observar que considerou-se o *bootstrap* aplicado diretamente às entradas $(X_i, Y_i)_{1 \leq i \leq n}$, preservando assim o caráter de caixa-preta da RF. Mesmo que a natureza temporal dos dados seja ignorada após a etapa de *bootstrap*, incluir o tempo como variável independente pode fornecer uma forma indireta de capturar, ainda que parcialmente, a dependência temporal. Trabalhos sobre *bootstrap* em blocos na literatura de previsão de séries temporais frequentemente aplicam o *bootstrap* aos resíduos após remover a tendência e a sazonalidade. No entanto, os experimentos de Goehry et al. (2023) mostraram que aplicar o *bootstrap* sobre resíduos gerados por uma RF preliminar pode acarretar em um desempenho inferior, o que reforça a abordagem adotada.

As *Random Forests* podem ser usadas para ordenar as variáveis de acordo com sua importância. Uma das medidas mais utilizadas é o *Mean Decrease Accuracy*, que se baseia na ideia de que, se uma variável não é importante, então permutar seus valores não deve alterar a acurácia preditiva do modelo. Para cada árvore, possui-se acesso às observações OOB, denotadas OOB_m, que correspondem às observações não incluídas no *bootstrap* usado para construir a árvore m. A amostra OOB_m é usada para estimar o erro de previsão, denotado por errOOB_m. Para calcular a importância da variável $X^{(j)}$, permutam-se os valores dessa variável apenas no conjunto OOB e computa-se o erro de previsão após a permutação. A importância final é dada pela diferença média de erros antes e depois da permutação (Goehry et al., 2023).

Seja $\widetilde{\text{errOOB}}_m^{\,j}$ o erro da árvore m após a permutação da variável $X^{(j)}$, a importância da variável é definida por:

$$VI\left(X^{(j)}\right) = \frac{1}{M} \sum_{m=1}^{M} \left(\widetilde{\text{errOOB}}_m^{\,j} - \text{errOOB}_m \right). \tag{3.31}$$

Quanto maior o aumento no erro de previsão causado pela permutação, mais importante é a variável em questão. Caso a permutação não altere substancialmente o erro, então pode-se concluir que a variável possui uma baixa importância para a previsão da variável dependente (Goehry et al., 2023).

No caso de observações dependentes, permutar aleatoriamente os valores da variável no conjunto OOB destrói a estrutura temporal. Em outras palavras, em séries temporais os valores do período corrente de uma dada variável dependem de seus valores de períodos passados, criando padrões como tendências e sazonalidade; dessa forma, permutar aleatoriamente os valores da variável em questão destrói essa temporalidade, prejudicando a acurácia preditiva do modelo (Goehry et al., 2023).

Para contornar esse problema, Goehry et al. (2023) propõem a importância por permutação em blocos (IPB), de forma que, ao invés de embaralhar cada observação individual, divide-se o conjunto OOB em blocos consecutivos de tempo. Em termos matemáticos, suponha que os conjuntos OOB possam ser divididos em blocos de tamanho l_n, denotados por B_m^*. Para calcular a importância da variável $X^{(j)}$, permutam-se apenas os blocos dentro de B_m^*, preservando a dependência temporal interna de cada bloco. Seja $\widetilde{\text{errOOB}}_m^{j,B}$ o erro após a permutação em blocos, a importância em blocos é definida por:

$$VI\left(X^{(j)}\right) = \frac{1}{M}\sum_{m=1}^{M}\left(\widetilde{\text{errOOB}}_m^{j,B} - \text{errOOB}_m\right). \tag{3.32}$$

Em outras palavras, a importância da variável $X^{(j)}$ é dada pela média, ao longo dos $m = 1, ..., M$ blocos, da diferença dos erros de previsão antes e depois da permutação em blocos, dada por $\widetilde{\text{errOOB}}_m^{j,B} - \text{errOOB}_m$. Se o valor resultante for elevado e positivo, as evidências sugerem que, ao embaralhar a variável $X^{(j)}$, a acurácia preditiva do modelo piora consideravelmente, sugerindo que a variável é relevante para prever a variável dependente; e, se for próximo de zero, as evidências sugerem que embaralhar a variável $X^{(j)}$ altera muito pouco o erro, sugerindo que a variável é pouco relevante para a previsão da variável dependente.

Os blocos provenientes de um *bootstrap* em blocos com parâmetro l_n nem sempre possuem exatamente esse tamanho, podendo ser um pouco maiores ou menores. Como a IPB compara os erros antes e após a permutação dos blocos, é fundamental que esses blocos tenham um tamanho comparável, de forma que a importância não seja injusta, já que um bloco maior tende a embaralhar mais e aumentar mais o erro. Assim sendo, os autores propõem um procedimento de ajuste. Se um bloco apresentar um tamanho de exatamente l_n, é mantido; se apresentar tamanho maior que l_n, mas menor que $2l_n$, escolhe-se aleatoriamente um subconjunto consecutivo de tamanho l_n; e se apresentar um tamanho inferior a l_n, é descartado. O conjunto final de blocos ajustados, isto é, padronizados para o tamanho l_n, é então usado para calcular a IPB (Goehry et al., 2023).

Por fim, destaca-se que os resultados da pesquisa de Goehry et al. (2023) mostraram que é possível aprimorar, de forma significativa, o desempenho de exercícios de previsão quando o tamanho do bloco é escolhido adequadamente. Não obstante, a variante NBB mostrou-se problemática para capturar corretamente a importância das variáveis, ignorando variáveis fundamentais para o problema de previsão. Por outro lado, as variantes MBB e CBB apresentaram um desempenho muito superior à *Random Forest* padrão quando o comprimento do bloco é bem escolhido. Os autores também mostraram que uma boa heurística para a escolha desse comprimento é tomá-lo como um múltiplo da menor sazonalidade presente na série.

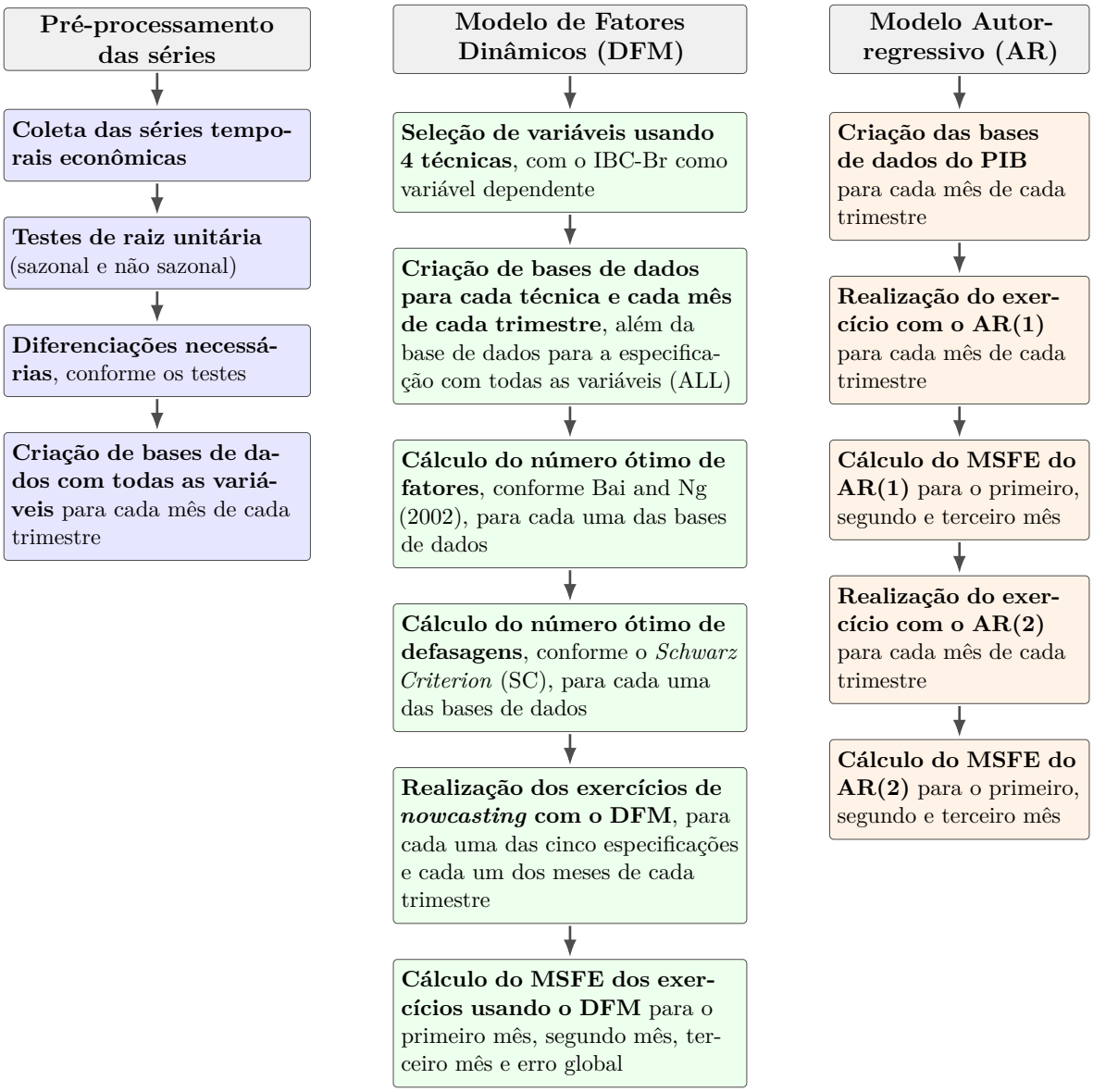
No presente trabalho, utiliza-se o pacote `rangerts`, disponível no R, para ajustar uma *Random Forest* (RF) com *bootstrap* em blocos e calcular a importância por permutação em blocos (IPB) das variáveis independentes (Wright, 2019). De forma mais precisa, foram utilizadas duas especificações de *bootstrap* em blocos, a saber, o *Moving Block Bootstrap* (MBB) e o *Circular Block Bootstrap* (CBB), que estão entre as especificações que apresentaram melhores resultados em outros exercícios de previsão de séries temporais (Goehry et al., 2023). Após a aplicação dessas duas especificações, foram calculadas as IPBs para todos os períodos nos quais foram realizados exercícios de *nowcasting*.

A ideia principal do pacote pacote `rangerts` é testar o algoritmo de *Random Forest* utilizando o *bootstrap* em blocos, de forma a contribuir na captura da dependência temporal dos dados, durante o período de construção das árvores, em vez do modo padrão de reamostragem independente. Esse pacote contém todas as funções do pacote `ranger` original, além de funcionalidades adicionais. Por meio desse pacote, pode-se escolher as seguintes abordagens de *bootstrap*: blocos não sobrepostos, móveis, estacionários, circulares e sazonais. Não obstante, algumas pesquisas mostraram que os modelos móvel e estacionário podem ser mais eficientes. Em outras palavras, o `rangerts` utiliza uma lógica similar ao `ranger`, mudando a maneira na qual os dados são reamostrados durante a construção das árvores, usando o *bootstrap* em blocos (Wright, 2019).

3.3 Procedimentos realizados

A presente seção possui como objetivo apresentar, de forma visual, todos os passos e procedimentos que foram realizados neste trabalho. Uma descrição ainda mais detalhada se encontra no Apêndice B. Segue o fluxograma com os procedimentos:

Figura 2 – Fluxo metodológico dos procedimentos realizados no trabalho.



Observação: Os exercícios foram realizados, em todos os casos, para os 8 trimestres entre o segundo trimestre de 2023 e o primeiro trimestre de 2025. Todas as bases de dados pressuporam dados em tempo pseudo-real, isto é, foram considerados atrasos constantes nas divulgações das variáveis. De forma semelhante, todas as bases para os exercícios com o DFM apresentaram bordas irregulares, que são aquelas observações ausentes ao final da amostra.

Fonte: Elaboração própria.

4 DADOS

No presente capítulo serão apresentadas as variáveis utilizadas, suas periodicidades, suas fontes e as transformações que foram realizadas para torná-las estacionárias.

4.1 Séries temporais

O conjunto de dados coletado para este trabalho inclui 44 variáveis econômicas[1], que podem ser divididas em oito grupos: (1) mercado de trabalho, (2) variáveis financeiras, (3) setor monetário, (4) índices de preços, (5) atividade econômica, (6) setor externo, (7) setor governamental e (8) setor de varejo. A maioria dessas séries foi coletada no Sistema Gerenciador de Séries Temporais (SGS) do Banco Central do Brasil (BCB), no Instituto Brasileiro de Geografia e Estatística (IBGE) e/ou no Ipeadata, apesar de suas fontes serem diversas. Todas as variáveis independentes utilizadas possuem frequência mensal, embora algumas variáveis coletadas sejam diárias. Assim, em alguns casos, tornou-se necessário realizar uma média mensal dos dados diários. Destaca-se que o presente estudo utiliza dados em tempo pseudo-real, seguindo as formulações de Bantis, Clements and Urquhart (2023).

Destaca-se que, para todas as variáveis independentes, o período temporal considerado foi de março de 2012 a março de 2025, já que a taxa de desocupação (desemprego) é uma série que se inicia somente em março de 2012. Não obstante, o período temporal depende da disponibilidade da variável na data de previsão, isto é, para muitas variáveis não havia observações de março de 2025 disponíveis em 31/01/2025, 28/02/2025 e/ou 31/03/2025. Todos os conjuntos de dados foram carregados em pastas de trabalho do *Microsoft Excel*, enquanto as transformações e as modelagens foram realizadas por meio do R e do RStudio. A Tabela 1 abaixo apresenta os nomes das séries usadas neste trabalho, bem como suas periodicidades e fontes. Por sua vez, os gráficos das séries temporais e das funções de autocorrelação (ACFs, da sigla em inglês)[2] estão contidos no Apêndice A.

[1] Neste ponto da discussão, torna-se prudente destacar alguns pontos: (1) o "novo" nas séries de Meios de Pagamento se refere à revisão metodológica realizada pelo Departamento de Estatísticas do BCB, com o objetivo de adequar as séries aos padrões internacionais mais recentes de estatísticas monetárias (Banco Central do Brasil, 2025b); (2) a variável coletada para o Índice Nacional de Preços ao Consumidor-Amplo (IPCA) se refere à variação percentual mensal desse índice (Banco Central do Brasil, 2025c); (3) a variável coletada para a Taxa de Juros Selic é o seu percentual ao dia (Banco Central do Brasil, 2025a), no qual foi aplicado uma média mensal; (4) a variável coletada para o PIB se refere ao PIB a preços de mercado - Taxa trimestre contra trimestre imediatamente anterior, disponível na Tabela 5932 - Taxa de variação do índice de volume trimestral das Contas Nacionais Trimestrais do SIDRA IBGE (IBGE, 2025c); e (5) em nenhuma das séries foram aplicados logaritmos, já que o objetivo é a previsão do PIB.

[2] A função de autocorrelação é uma ferramenta estatística que mede a correlação de uma série temporal com suas próprias defasagens, isto é, com seus valores em períodos anteriores.

Tabela 1 – Séries temporais: Nome, Frequência e Fonte

Variável utilizada	**Frequência**	**Fonte**
Índice nacional de custo da construção do mercado (INCC-M)	Mensal	FGV/Conj. Econ. - IGP
Índice de ações: Ibovespa - fechamento	Mensal	Anbima
Meios de pagamento - M1 (saldo em final de período) - Novo - sazonalmente ajustado	Mensal	BCB-DSTAT
Meios de pagamento amplos - M2 (saldo em final de período) - Novo - sazonalmente ajustado	Mensal	BCB-DSTAT
Meios de pagamento amplos - M3 (saldo em final de período) - Novo	Mensal	BCB-DSTAT
Meios de pagamento amplos - M4 (saldo em final de período) - Novo	Mensal	BCB-DSTAT
Salário mínimo (deflacionado com o IPCA)	Mensal	Ministério do Trabalho
Saldo da carteira de crédito - Total	Mensal	BCB-DSTAT
Taxa de câmbio - Livre - Dólar americano (venda)	Mensal	Sisbacen PTAX800
Taxa de câmbio - Livre - Euro (venda)	Mensal	PTAX
Taxa de câmbio - Livre - Iene (venda)	Mensal	PTAX
Taxa de desocupação das pessoas de 14 anos ou mais de idade, na semana de referência	Mensal	IBGE/PNADC
Percentual ao dia da Taxa de Juros Selic (média mensal)	Mensal	BCB-Demab
Variação percentual mensal do IPCA	Mensal	IBGE
Índice de confiança do empresário industrial geral (ICEI)	Mensal	CNI
Empregados no setor público e privado com carteira	Mensal	IBGE/PNADC
Índice de confiança do consumidor (ICC)	Mensal	Fecomercio SP
Operações de crédito - inadimplência da carteira de crédito	Mensal	BCB-DSTAT
Exportação de bens - Balanço de Pagamentos	Mensal	BCB-DSTAT
Importação de bens - Balanço de Pagamentos	Mensal	BCB-DSTAT
Balanço de pagamentos: transações correntes - saldo	Mensal	BCB-DSTAT
Investimentos diretos no país (IDP) líquido	Mensal	BCB-DSTAT
Dívida Líquida do Setor Público - Total	Mensal	BCB-DSTAT
Arrecadação das receitas federais - receita bruta	Mensal	Min. Economia/SRF
Resultado Primário do Governo Central	Mensal	MF-STN
Produção industrial - indústria geral: índice de quantum dessazonalizado	Mensal	IBGE/PIM-PF
Vendas reais no varejo de veículos, motos, partes e peças	Mensal	IBGE/PMC
Utilização da capacidade instalada - indústria - índice dessazonalizado	Mensal	CNI
Faturamento real - indústria - índice dessazonalizado	Mensal	CNI
Pessoal empregado - indústria - índice dessazonalizado	Mensal	CNI
Horas trabalhadas - indústria - índice dessazonalizado	Mensal	CNI
Produção industrial - veículos automotores, reboques e carrocerias - quantum - índice	Mensal	IBGE/PIM-PF
Emplacamento de autoveículos	Mensal	Fenabrave
Exportações - veículos automotores, reboques, carrocerias - quantum - índice	Mensal	Funcex
Vendas reais no varejo ampliado - índice dessazonalizado	Mensal	IBGE/PMC
Vendas reais - varejo - móveis e eletrodomésticos - índice dessazonalizado	Mensal	IBGE/PMC
IBC-Br - índice real dessazonalizado	Mensal	BCB/SGS
Índice de volume de serviços - total	Mensal	IBGE/PMS
Vendas reais no varejo de materiais de construção: índice dessazonalizado	Mensal	IBGE/PMC
Exportações - agricultura e pecuária - quantum: índice	Mensal	Funcex
Exportações - extração de petróleo e gás natural - quantum: índice	Mensal	Funcex
Massa de rendimento real de todos os trabalhos	Mensal	IBGE/PNADC
Energia elétrica referente ao consumo - quantidade	Mensal	Eletrobras
PIB a preços de mercado - Taxa trimestre contra trimestre imediatamente anterior	Trimestral	IBGE/SIDRA

Fonte: Elaboração própria.

4.2 Testes de raiz unitária e de sazonalidade

Nesta seção, serão apresentados os resultados da aplicação dos testes de raiz unitária e de sazonalidade para as variáveis utilizadas neste trabalho. Essas técnicas foram utilizadas para decidir quais séries são não estacionárias, bem como quais diferenciações são necessárias. Para atingir esse objetivo, foram realizados testes de raiz unitária para cada série temporal coletada, a saber, os testes de Dickey-Fuller Aumentado (ADF, da sigla em inglês), de Dickey-Fuller Generalized Least Squares (DF-GLS), de Kwiatkowski, Phillips, Schmidt e Shin (KPSS), de Phillips e Perron (PP), o *bootstrap* do ADF e o *bootstrap* do teste de união, bem como os testes de Kruskall-Wallis (KW) e de Hylleberg, Engle, Granger e Yoo (HEGY). Esses testes são relevantes para a identificação de quais variáveis apresentam um comportamento de processo não estacionário, de forma que seja possível realizar as transformações necessárias para torná-las estacionárias (Hamilton, 1994; Bueno, 2011; Barros et al., 2017).[3]

A Tabela 2 a seguir apresenta a lista das séries temporais coletadas e utilizadas neste estudo, juntamente com as diferenciações realizadas para torna-las estacionárias. Pode-se observar que a maior parte das séries exigiu algum tipo de diferenciação, seja não sazonal, por meio da primeira diferença com defasagem de 1 período, ou sazonal, com defasagem de 12 períodos, o que revela a presença de não estacionariedade em muitas variáveis. Por exemplo, a variável do índice nacional de custo de construção do mercado (INCC-M) necessitou de uma diferenciação não sazonal de primeira ordem, isto é, uma única diferenciação com uma defasagem de apenas um período.

Essa função auxilia na identificação de certos padrões, a saber, (1) a raiz unitária, que é uma espécie de memória de longo prazo; (2) a sazonalidade, que são padrões se repetem periodicamente, com frequência anualmente; e (3) uma estrutura de dependência temporal, onde o valor presente da variável depende de seus valores passados.

[3] Destaca-se que a interpretação gráfica da ACF colabora significativamente para o entendimento de se a série temporal em questão é ou não estacionária. Uma série não estacionária reduz lentamente sua autocorrelação à medida que o tempo passa, enquanto, no caso de uma série estacionária, sua autocorrelação decresce rapidamente para zero (Enders, 2008).

Tabela 2 – Séries temporais: Diferenciações realizadas

Variável utilizada	**Diferenciações**
Índice nacional de custo da construção do mercado (INCC-M)	Δ
Índice de ações: Ibovespa - fechamento	Nenhuma
Meios de pagamento - M1 (saldo em final de período) - Novo - sazonalmente ajustado	$\Delta\Delta_{12}$
Meios de pagamento amplos - M2 (saldo em final de período) - Novo - sazonalmente ajustado	$\Delta\Delta_{12}$
Meios de pagamento amplos - M3 (saldo em final de período) - Novo	Δ
Meios de pagamento amplos - M4 (saldo em final de período) - Novo	$\Delta\Delta_{12}$
Salário mínimo (deflacionado com o IPCA)	Nenhuma
Saldo da carteira de crédito - Total	$\Delta\Delta_{12}$
Taxa de câmbio - Livre - Dólar americano (venda)	Δ
Taxa de câmbio - Livre - Euro (venda)	Δ
Taxa de câmbio - Livre - Iene (venda)	Δ
Taxa de desocupação das pessoas de 14 anos ou mais de idade, na semana de referência	Δ
Percentual ao dia da Taxa de Juros Selic (média mensal)	Δ
Variação percentual mensal do IPCA	Δ
Índice de confiança do empresário industrial geral (ICEI)	Δ
Empregados no setor público e privado com carteira	Δ
Índice de confiança do consumidor (ICC)	Δ
Operações de crédito - inadimplência da carteira de crédito	Δ
Exportação de bens - Balanço de Pagamentos	$\Delta\Delta_{12}$
Importação de bens - Balanço de Pagamentos	$\Delta\Delta_{12}$
Balanço de pagamentos: transações correntes - saldo	$\Delta\Delta_{12}$
Investimentos diretos no país (IDP) líquido	$\Delta\Delta_{12}$
Dívida Líquida do Setor Público - Total	Δ
Arrecadação das receitas federais - receita bruta	$\Delta\Delta_{12}$
Resultado Primário do Governo Central	$\Delta\Delta_{12}$
Produção industrial - indústria geral: índice de quantum dessazonalizado	Δ
Vendas reais no varejo de veículos, motos, partes e peças	Δ
Utilização da capacidade instalada - indústria - índice dessazonalizado	Δ
Faturamento real - indústria - índice dessazonalizado	Δ
Pessoal empregado - indústria - índice dessazonalizado	Δ
Horas trabalhadas - indústria - índice dessazonalizado	$\Delta\Delta_{12}$
Produção industrial - veículos automotores, reboques e carrocerias - quantum - índice	$\Delta\Delta_{12}$
Emplacamento de autoveículos	Δ
Exportações - veículos automotores, reboques, carrocerias - quantum - índice	Δ
Vendas reais no varejo ampliado - índice dessazonalizado	Nenhuma
Vendas reais - varejo - móveis e eletrodomésticos - índice dessazonalizado	Nenhuma
IBC-Br - índice real dessazonalizado	Δ
Índice de volume de serviços - total	Δ
Vendas reais no varejo de materiais de construção: índice dessazonalizado	Δ
Exportações - agricultura e pecuária - quantum: índice	Δ
Exportações - extração de petróleo e gás natural - quantum: índice	$\Delta\Delta_{12}$
Massa de rendimento real de todos os trabalhos	$\Delta\Delta_{12}$
Energia elétrica referente ao consumo - quantidade	$\Delta\Delta_{12}$
PIB a preços de mercado - Taxa trimestre contra trimestre imediatamente anterior	Nenhuma

Fonte: Elaboração própria.

5 RESULTADOS

O presente capítulo apresenta os resultados da seleção de variáveis e dos exercícios de *nowcasting* da taxa de crescimento do PIB brasileiro.

5.1 Seleção de variáveis

No presente trabalho, foram realizadas seleções de variáveis com as informações disponíveis no último dia de cada mês de cada trimestre, o que significa que cada uma das 4 técnicas foi aplicada um total de 24 vezes. A título de exemplo, dentre as variáveis independentes coletadas, o LASSO selecionou 34 delas em 31/01/2025, que estão apresentadas na Tabela 3. Neste ponto da discussão, é prudente ressaltar que a seleção em 31/01/2025 é realizada somente com as observações que estavam disponíveis naquela data particular, dada a suposição de atrasos constantes nas divulgações das variáveis. Pode-se observar uma presença notável de variáveis (1) de varejo, dadas as séries de vendas reais no varejo de veículos, de vendas reais no varejo ampliado, de vendas reais no varejo de móveis e eletrodomésticos e de vendas reais no varejo de materiais de construção; (2) de comércio internacional, dadas as variáveis de taxa de câmbio, de exportação e importação de bens, de exportação de produtos agropecuários, de investimentos diretos no país e de transações correntes; (3) de emprego e renda, dadas as variáveis de empregados no setor público e no privado com carteira, de salário mínimo, de massa de rendimento real e de índice de volume de serviços; e (4) de variáveis monetárias e financeiras, dadas as variáveis de M1, M3, M4, variação percentual mensal do IPCA e de inadimplência da carteira de crédito.

A seleção dessas variáveis está, em grande medida, de acordo com o esperado nesta pesquisa. De fato, os setores de varejo, serviços, industrial e agropecuário são fundamentais para a economia brasileira, o que justifica a escolha das variáveis que mensuram a atividade nesses setores. De forma mais precisa, em 2023 o setor de serviços era responsável por impressionantes 67,8% do Valor Adicionado Bruto (VAB) a preços básicos do Brasil; mais especificamente, as outras atividades de serviços lideravam o setor de serviços, com uma participação de aproximadamente 16,7% do VAB total, seguidas pelas atividades de administração, defesa, saúde e educação públicas e seguridade social, com participação de cerca de 15,9%. A participação do agronegócio se limitava a cerca de 6,9% do VAB total. Por sua vez, a indústria apresentava, em 2023, uma participação de aproximadamente 25,4%, menos que a metade da participação do setor de serviços; mais especificamente, a indústria de transformação liderava o setor industrial, com uma participação de aproximadamente 15,2% do VAB total, seguida pela indústria extrativa, com uma participação de cerca de 4,2%. Em outras palavras, a seleção de variáveis efetuada pelo LASSO parece ser justificada, dadas as características da economia brasileira (IBGE, 2023).

Tabela 3 – Variáveis selecionadas pelo LASSO em 31/01/2025

Variáveis selecionadas
Índice nacional de custo da construção do mercado (INCC-M)
Meios de pagamento - M1 (saldo em final de período) - Novo - sazonalmente ajustado
Meios de pagamento amplos - M3 (saldo em final de período) - Novo
Meios de pagamento amplos - M4 (saldo em final de período) - Novo
Salário mínimo (deflacionado com o IPCA)
Saldo da carteira de crédito - Total
Taxa de câmbio - Livre - Dólar americano (venda)
Taxa de câmbio - Livre - Euro (venda)
Percentual ao dia da Taxa de Juros Selic (média mensal)
Variação percentual mensal do IPCA
Índice de confiança do empresário industrial geral (ICEI)
Empregados no setor público e privado com carteira
Índice de confiança do consumidor (ICC)
Operações de crédito - inadimplência da carteira de crédito
Exportação de bens - Balanço de Pagamentos
Importação de bens - Balanço de Pagamentos
Balanço de pagamentos: transações correntes - saldo
Investimentos diretos no país (IDP) líquido
Dívida Líquida do Setor Público - Total
Arrecadação das receitas federais - receita bruta
Resultado Primário do Governo Central
Produção industrial - indústria geral: índice de quantum dessazonalizado
Vendas reais no varejo de veículos, motos, partes e peças
Utilização da capacidade instalada - indústria - índice dessazonalizado
Produção industrial - veículos automotores, reboques e carrocerias - quantum - índice
Emplacamento de autoveículos
Exportações - veículos automotores, reboques, carrocerias - quantum - índice
Vendas reais no varejo ampliado - índice dessazonalizado
Vendas reais - varejo - móveis e eletrodomésticos - índice dessazonalizado
Índice de volume de serviços - total
Vendas reais no varejo de materiais de construção: índice dessazonalizado
Exportações - agricultura e pecuária - quantum: índice
Massa de rendimento real de todos os trabalhos
Energia elétrica referente ao consumo - quantidade

Fonte: Elaboração própria.

Novamente à título de exemplo, será apresentada a seleção de variáveis realizada pelo ENET para o período de 31/01/2025. Dentre as variáveis independentes coletadas, o ENET selecionou 37 delas nesse período, apenas três a mais que o LASSO, a saber, (1) o Ibovespa, (2) a taxa de desocupação e (3) o pessoal empregado na indústria. Uma lista completa das variáveis selecionadas é apresentada na Tabela 4. Pode-se observar uma presença notável de variáveis (1) de varejo, dadas as séries de vendas reais no varejo de veículos, de vendas reais no varejo ampliado, de vendas reais no varejo de móveis e eletrodomésticos e de vendas reais no varejo de materiais de construção; (2) de comércio internacional, dadas as variáveis de taxa de câmbio, de exportação e importação de bens, de exportação de produtos agropecuários, de investimentos diretos no país e de transações correntes; (3) de emprego e renda, dadas as variáveis de empregados no setor público e no privado com carteira, de empregados na indústria, de salário mínimo, de massa de rendimento real e de índice de volume de serviços; e (4) de variáveis monetárias e financeiras, dadas as variáveis de M1, M3, M4, variação percentual mensal do IPCA, de inadimplência da carteira de crédito e do Ibovespa.

Novamente, a seleção dessas variáveis está, em grande medida, de acordo com o esperado na pesquisa. De fato, os setores de varejo, serviços, indústria e agropecuário são fundamentais para a economia brasileira. Por fim, destaca-se que a presença marcante de variáveis industriais está longe de ser contraintuitiva, já que a participação da indústria na economia brasileira ainda é relevante, apesar da histórica perda de importância para o setor de serviços (IBGE, 2023). De forma semelhante, é prudente destacar que as regressões do LASSO e do ENET foram ambas realizadas usando o IBC-Br como variável dependente, escolhendo o valor ótimo do parâmetro de ajuste λ usando uma *expanding window 1-step-ahead cross-validation* e utilizando o pacote `glmnet`, disponível no *software* R. Por fim, a Figura 3 apresenta um Diagrama de Venn, que mostra as quantidades de variáveis selecionadas somente pelo LASSO, apenas pelo ENET e por ambas as técnicas. Por meio deste diagrama, pode-se observar (1) que todas as variáveis selecionadas pelo LASSO também o foram pelo ENET, o que já era esperado, tendo em vista que o ENET é uma técnica com uma penalização intermediária entre o LASSO e a Regressão Ridge; e (2) que o ENET selecionou apenas três variáveis a mais que o LASSO.

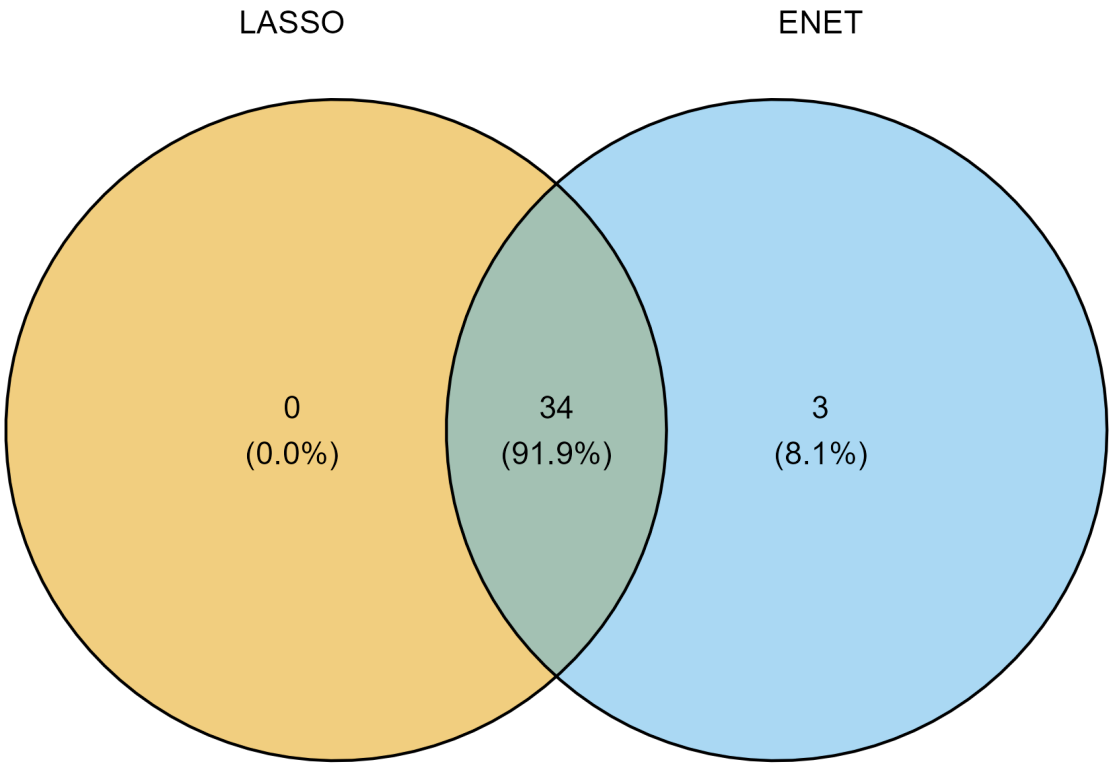
De forma semelhante, foram realizadas regressões da *Random Forest* (RF) com duas especificações, o *Moving Block Bootstrap* (MBB) e o *Circular Block Bootstrap* (CBB); realizados os ajustes desses modelos, foi calculada a importância por permutação em blocos (IPB) das florestas com ambas as especificações. Destaca-se que todas florestas foram regredidas com o IBC-Br como variável dependente, dada a necessidade de que seja de mesma frequência das variáveis independentes, e com as demais séries mensais como variáveis independentes. Após serem ordenadas, de forma decrescente, as variáveis pela IPB, selecionou-se a mesma quantidade de variáveis que o LASSO, que, para o exemplo

Tabela 4 – Variáveis selecionadas pelo ENET em 31/01/2025

Variáveis selecionadas
Índice nacional de custo da construção do mercado (INCC-M)
Índice de ações: Ibovespa - fechamento
Meios de pagamento - M1 (saldo em final de período) - Novo - sazonalmente ajustado
Meios de pagamento amplos - M3 (saldo em final de período) - Novo
Meios de pagamento amplos - M4 (saldo em final de período) - Novo
Salário mínimo (deflacionado com o IPCA)
Saldo da carteira de crédito - Total
Taxa de câmbio - Livre - Dólar americano (venda)
Taxa de câmbio - Livre - Euro (venda)
Taxa de desocupação das pessoas de 14 anos ou mais de idade, na semana de referência
Percentual ao dia da Taxa de Juros Selic (média mensal)
Variação percentual mensal do IPCA
Índice de confiança do empresário industrial geral (ICEI)
Empregados no setor público e privado com carteira
Índice de confiança do consumidor (ICC)
Operações de crédito - inadimplência da carteira de crédito
Exportação de bens - Balanço de Pagamentos
Importação de bens - Balanço de Pagamentos
Balanço de pagamentos: transações correntes - saldo
Investimentos diretos no país (IDP) líquido
Dívida Líquida do Setor Público - Total
Arrecadação das receitas federais - receita bruta
Resultado Primário do Governo Central
Produção industrial - indústria geral: índice de quantum dessazonalizado
Vendas reais no varejo de veículos, motos, partes e peças
Utilização da capacidade instalada - indústria - índice dessazonalizado
Pessoal empregado - indústria - índice dessazonalizado
Produção industrial - veículos automotores, reboques e carrocerias - quantum - índice
Emplacamento de autoveículos
Exportações - veículos automotores, reboques, carrocerias - quantum - índice
Vendas reais no varejo ampliado - índice dessazonalizado
Vendas reais - varejo - móveis e eletrodomésticos - índice dessazonalizado
Índice de volume de serviços - total
Vendas reais no varejo de materiais de construção: índice dessazonalizado
Exportações - agricultura e pecuária - quantum: índice
Massa de rendimento real de todos os trabalhos
Energia elétrica referente ao consumo - quantidade

Fonte: Elaboração própria.

Figura 3 – Variáveis selecionadas pelo LASSO e pelo ENET em 31/01/2025



Fonte: Elaboração própria.

de 31/01/2025, é um total de 34 variáveis. A decisão de selecionar o mesmo número de variáveis independentes que o LASSO justifica-se pela centralidade desta técnica para a seleção de variáveis dos conjuntos de dados em problemas de *nowcasting* das taxas de crescimento do PIB (Cepni; Güney; Swanson, 2019a, 2019b; Bantis; Clements; Urquhart, 2023).

Em relação às variáveis selecionadas pela RF MBB e pela RF CBB, a Tabela 5 apresenta as primeiras, enquanto a Tabela 6 mostra as segundas. Pode-se observar uma nítida presença de variáveis (1) de varejo, (2) de comércio internacional, (3) de emprego e renda e (4) industriais, o que suscita as mesmas considerações anteriormente apresentadas. Não obstante, nenhuma das duas especificações de florestas selecionou o índice de serviços, o que é questionável, já que o setor de serviços é o principal responsável pelo Valor Adicionado Bruto (VAB) a preços básicos do Brasil (IBGE, 2023). De modo geral, pode-se afirmar que essas duas técnicas selecionaram um conjunto de variáveis similar

ao LASSO e ao ENET, o que pode ser observado também na Figura 4, que apresenta diversas informações relevantes, como as interseções das variáveis selecionadas por cada técnica e a quantidade de variáveis selecionadas por cada método. Pode-se observar que, dentre as quatro técnicas, (1) 25 variáveis foram selecionadas por todas as quatro técnicas; (2) 6 apenas pelo LASSO e pelo ENET; (3) 4 somente pela RF MBB e pela RF CBB; (4) 3 variáveis apenas pela RF MBB, pela RF CBB e pelo ENET; (5) 2 apenas pelo LASSO, pelo ENET e pela RF CBB; (6) 1 apenas pela RF MBB; e (7) 1 apenas pela RF MBB, pelo LASSO e pelo ENET.

Tabela 5 – Variáveis selecionadas pela RF MBB em 31/01/2025

Variáveis selecionadas
Vendas reais no varejo de veículos, motos, partes e peças
Produção industrial - indústria geral: índice de quantum dessazonalizado
Vendas reais no varejo de materiais de construção: índice dessazonalizado
Faturamento real - indústria - índice dessazonalizado
Produção industrial - veículos automotores, reboques e carrocerias - quantum - índice
Vendas reais - varejo - móveis e eletrodomésticos - índice dessazonalizado
Utilização da capacidade instalada - indústria - índice dessazonalizado
Energia elétrica referente ao consumo - quantidade
Taxa de câmbio - Livre - Dólar americano (venda)
Operações de crédito - inadimplência da carteira de crédito
Exportações - agricultura e pecuária - quantum: índice
Emplacamento de autoveículos
Taxa de câmbio - Livre - Iene (venda)
Pessoal empregado - indústria - índice dessazonalizado
Vendas reais no varejo ampliado - índice dessazonalizado
Exportações - veículos automotores, reboques, carrocerias - quantum - índice
Exportações - extração de petróleo e gás natural - quantum: índice
Saldo da carteira de crédito - Total
Importação de bens - Balanço de Pagamentos
Índice de confiança do consumidor (ICC)
Variação percentual mensal do IPCA
Índice de ações: Ibovespa - fechamento
Índice nacional de custo da construção do mercado (INCC-M)
Horas trabalhadas - indústria - índice dessazonalizado
Índice de confiança do empresário industrial geral (ICEI)
Empregados no setor público e privado com carteira
Taxa de desocupação das pessoas de 14 anos ou mais de idade, na semana de referência
Percentual ao dia da Taxa de Juros Selic (média mensal)
Resultado Primário do Governo Central
Taxa de câmbio - Livre - Euro (venda)
Dívida Líquida do Setor Público - Total
Investimentos diretos no país (IDP) líquido
Meios de pagamento amplos - M2 (saldo em final de período) - Novo - sazonalmente ajustado
Balanço de pagamentos: transações correntes - saldo

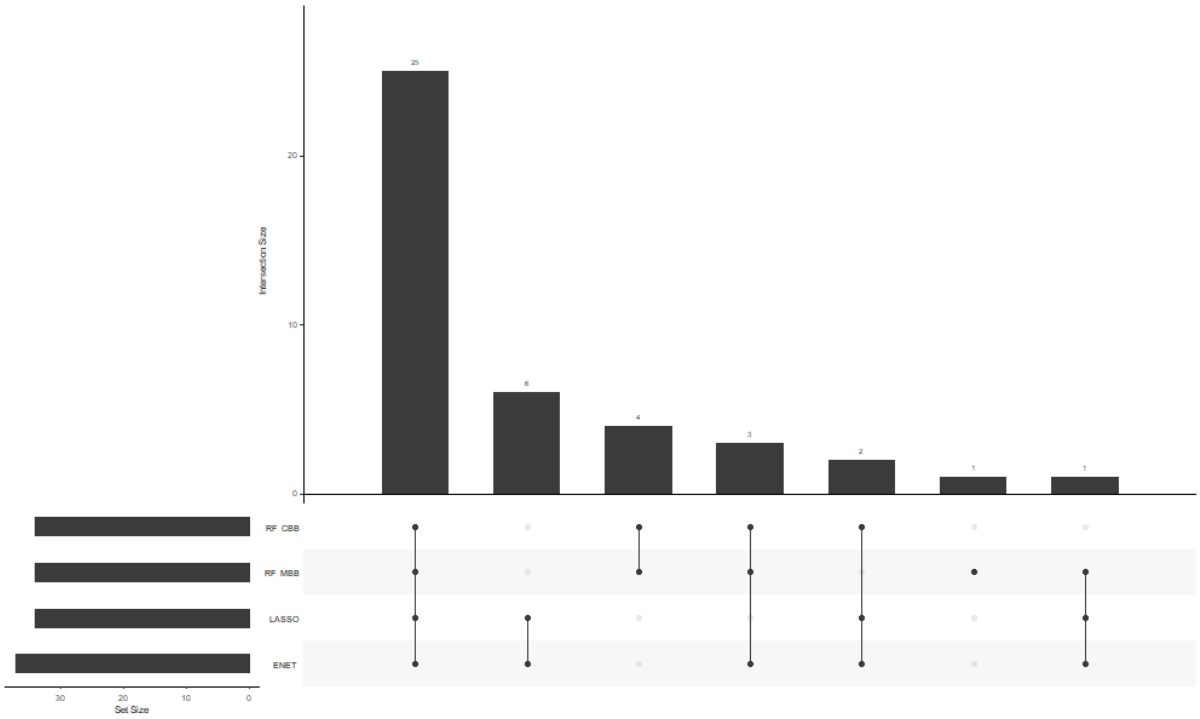
Fonte: Elaboração própria.

Tabela 6 – Variáveis selecionadas pela RF CBB em 31/01/2025

Variáveis selecionadas
Vendas reais no varejo de veículos, motos, partes e peças
Produção industrial - indústria geral: índice de quantum dessazonalizado
Vendas reais no varejo de materiais de construção: índice dessazonalizado
Faturamento real - indústria - índice dessazonalizado
Vendas reais - varejo - móveis e eletrodomésticos - índice dessazonalizado
Produção industrial - veículos automotores, reboques e carrocerias - quantum - índice
Utilização da capacidade instalada - indústria - índice dessazonalizado
Energia elétrica referente ao consumo - quantidade
Exportações - veículos automotores, reboques, carrocerias - quantum - índice
Taxa de câmbio - Livre - Dólar americano (venda)
Taxa de câmbio - Livre - Iene (venda)
Exportações - extração de petróleo e gás natural - quantum: índice
Exportações - agricultura e pecuária - quantum: índice
Emplacamento de autoveículos
Vendas reais no varejo ampliado - índice dessazonalizado
Operações de crédito - inadimplência da carteira de crédito
Horas trabalhadas - indústria - índice dessazonalizado
Pessoal empregado - indústria - índice dessazonalizado
Saldo da carteira de crédito - Total
Índice nacional de custo da construção do mercado (INCC-M)
Índice de confiança do empresário industrial geral (ICEI)
Índice de ações: Ibovespa - fechamento
Taxa de desocupação das pessoas de 14 anos ou mais de idade, na semana de referência
Variação percentual mensal do IPCA
Percentual ao dia da Taxa de Juros Selic (média mensal)
Importação de bens - Balanço de Pagamentos
Arrecadação das receitas federais - receita bruta
Índice de confiança do consumidor (ICC)
Balanço de pagamentos: transações correntes - saldo
Empregados no setor público e privado com carteira
Investimentos diretos no país (IDP) líquido
Resultado Primário do Governo Central
Massa de rendimento real de todos os trabalhos
Taxa de câmbio - Livre - Euro (venda)

Fonte: Elaboração própria.

Figura 4 – Variáveis selecionadas pelas quatro técnicas em 31/01/2025



Fonte: Elaboração própria.

5.2 *Nowcasting* do PIB

Na presente seção, serão apresentados os resultados dos exercícios de *nowcasting* das taxas de crescimento do PIB brasileiro, usando quatro métodos de seleção de variáveis, além dos exercícios com todas as séries, ou seja, sem a seleção prévia de variáveis. Os exercícios foram realizados utilizando um DFM, um AR(1) e um AR(2).

A Tabela 7 apresenta o Erro Quadrático Médio de Previsão (MSFE, da sigla em inglês) para os exercícios de *nowcasting* das taxas de crescimento do PIB brasileiro, utilizando um DFM, para cada período dos oito trimestres previstos, a saber, do 2° trimestre de 2023 ao 1° trimestre de 2025. Os exercícios foram realizados para os conjuntos de dados que estavam disponíveis, dadas as suposições de dados em tempo pseudo-real, no último dia do primeiro, segundo e terceiro meses. Por exemplo, para o 4° trimestre de 2024, o 1° mês se refere à previsão com dados disponíveis em 31/10/2024, o 2° mês em 30/11/2024 e o 3° mês em 31/12/2024. Também foram calculados MSFEs globais, que incluem todos os exercícios, independentemente do mês a que se referem. Além disso, os exercícios foram realizados para todas as variáveis (ALL), bem como para as variáveis selecionadas pelo LASSO, pelo ENET e pela IPB da RF com MBB e com CBB.

A partir desses resultados, é possível observar que, para os quatro casos considerados, a saber, os três meses e o valor global, a RF MBB apresentou o menor MSFE em todos

eles, com a RF CBB também apresentando um bom desempenho. Não obstante, para quase todas as especificações, o MSFE aumenta à medida que se desloca do primeiro para o segundo mês e diminui à medida que se desloca do segundo para o terceiro mês. Esse resultado é deveras contraintuitivo, já que o esperado é que, à medida que novas informações sejam divulgadas e, então, incorporadas à base de dados, a acurácia preditiva do modelo seja aprimorada, refletindo-se em uma redução do MSFE. Os resultados sugerem a necessidade de investigações posteriores, com a utilização de outras bases de dados e a realização de testes de robustez. Por fim, é prudente ressaltar dois pontos. O primeiro é que os métodos de seleção de variáveis foram aplicados em cada mês do trimestre; por exemplo, no caso do 1° trimestre de 2025, o LASSO foi executado em três momentos, a saber, em janeiro, em fevereiro e em março de 2025. O segundo é que todas as técnicas de seleção de variáveis foram utilizadas de forma compatível com a estrutura interna das séries temporais; por exemplo, o valor ótimo do parâmetro de ajuste no LASSO e no ENET foi escolhido utilizando uma *expanding window cross-validation 1-step-ahead*, de forma que manteve-se a estrutura temporal interna de cada variável.

Tabela 7 – Desempenho por método – DFM

Menor MSFE por último dia do mês do trimestre						
	ALL	LASSO	ENET	RF MBB	RF CBB	Melhor especificação
1° mês	0,7186	0,7203	0,8121	**0,3480**	0,4087	RF MBB
2° mês	0,8461	0,6009	0,9281	**0,5867**	0,5958	RF MBB
3° mês	0,6267	0,5026	0,4100	**0,3776**	0,417	RF MBB
Global	0,7305	0,6079	0,7167	**0,4374**	0,4738	RF MBB

Fonte: Elaboração própria.

De forma semelhante, foram realizados exercícios de previsão das taxas de crescimento do PIB brasileiro usando um AR(1) e um AR(2). A Tabela 8 apresenta os resultados para esses modelos autorregressivos de referência. A partir desses resultados, pode-se observar que o AR(1) apresentou um MSFE menor que o AR(2) em todos os casos e que o MSFE do primeiro mês é igual ao do segundo mês tanto no AR(1) quanto no AR(2), o que já era esperado, tendo em vista que pressupõe-se atrasos constantes na divulgação das variáveis, de forma que, ao final do segundo mês, a taxa de crescimento do PIB do trimestre anterior ainda não havia sido divulgada. Neste ponto da discussão, é prudente destacar que as previsões do AR(1) e do AR(2) nos dois primeiros meses do trimestre foram realizadas para dois passos a frente, já que foram pressupostos atrasos constantes na divulgação das variáveis, de forma que a observação do PIB do trimestre anterior ainda não estivesse disponível. Por outro lado, as previsões realizadas no 3° mês foram de um passo à frente, já que foi pressuposto que o PIB do trimestre anterior já fora disponibilizado.

Por exemplo, foi pressuposto que, em 31/01/2024 e em 29/02/2024, o PIB do 4° trimestre de 2023 ainda não fora disponibilizado, enquanto que, em 31/03/2024, foi pressuposto que essa observação já fora divulgada. Assim sendo, foi necessário que as previsões nos dois primeiros meses fossem para dois passos à frente, o que impossibilita uma comparação adequada entre as previsões dos dois primeiros meses com aquela do terceiro mês.

Tabela 8 – Desempenho por método – AR

Menor MSFE por último dia do mês do trimestre			
	AR(1)	AR(2)	Melhor especificação
1° mês	**0,5728**	0,6874	AR(1)
2° mês	**0,5728**	0,6874	AR(1)
3° mês	**0,5809**	0,7082	AR(1)

Fonte: Elaboração própria.

A partir dos resultados apresentados, pode-se observar que os DFMs com as variáveis selecionadas pela RF MBB e pela RF CBB apresentam menores MSFEs que o AR(1) e o AR(2) em todos os períodos, com exceção daquele relativo ao segundo mês. Por sua vez, o LASSO e o ENET apresentam um menor MSFE que o AR(1) somente no terceiro mês, sendo maiores nos outros períodos. Por fim, a especificação com todas as variáveis (ALL) apresentou MSFEs maiores que o AR(1) em todos os períodos. Esses resultados sugerem que (1) a RF MBB e a RF CBB desempenham bem como técnicas de seleção prévia de variáveis independentes em problemas de *nowcasting* das taxas de crescimento do PIB brasileiro; (2) o LASSO e o ENET não apresentam bons desempenhos, o que contrasta com os resultados previamente encontrados pela literatura (Cepni; Güney; Swanson, 2019a, 2019b; Bantis; Clements; Urquhart, 2023); e (3) o AR(1) consegue desempenhar consideravelmente bem para uma técnica simples, que exige poucas informações e pouca capacidade computacional. Os resultados sugerem que, em contextos de escassez de tempo e/ou recursos, o AR(1) aparenta ser uma técnica útil e eficaz para a previsão de séries temporais estacionárias, como a taxa de crescimento do PIB brasileiro.

6 CONSIDERAÇÕES FINAIS

O presente trabalho teve como objetivo principal prever, em termos de *nowcasting*, as taxas de crescimento trimestral do PIB brasileiro. Para isso, foram utilizadas técnicas de redução de dimensionalidade e/ou de encolhimento, de forma a selecionar as variáveis mais relevantes; após essa fase de seleção, foram realizados exercícios de previsão do PIB por meio de diversas especificações, utilizando um Modelo de Fatores Dinâmicos (DFM) e comparando-o com um Modelo Autorregressivo (AR).

A contribuição deste trabalho esteve na utilização de métricas de importância de variáveis derivadas de florestas aleatórias, a saber, da importância por permutação em blocos (IPB) da *Random Forest* (RF) tanto com um *Moving Block Bootstrap* (MBB) quanto com um *Circular Block Bootstrap* (CBB), para selecionar as variáveis independentes mais relevantes. Esses novos métodos de *bootstrap* em blocos não foram, até o conhecimento deste autor, utilizados no *nowcasting* das taxas de crescimento do PIB. Para a seleção de variáveis, a variável dependente utilizada foi o Índice de Atividade Econômica do Banco Central do Brasil (IBC-Br), que é um índice coincidente do PIB, dada a necessidade de que essa variável fosse da mesma frequência das variáveis independentes.

A hipótese do presente estudo foi de que a aplicação dessas modernas técnicas de redução de dimensionalidade poderia contribuir para a redução do Erro Quadrático Médio de Previsão (MSFE, da sigla em inglês) e, assim sendo, aprimorar a acurácia preditiva do DFM. A questão de pesquisa que se colocou foi a de que métodos baseados em florestas aleatórias voltados para a mensuração da importância das variáveis fornecem resultados melhores do que técnicas tradicionais de encolhimento como o *Least Absolute Shrinkage and Selection Operator* (LASSO) e o *Elastic Net* (ENET), já que aquelas métricas de florestas aleatórias apresentaram bons resultados em outros contextos científicos. Em outras palavras, buscou-se investigar se a seleção de variáveis a partir de métricas de importância de variáveis derivadas de florestas aleatórias poderia efetivamente aprimorar a capacidade preditiva do DFM, refletindo-se em menores valores do MSFE e, consequentemente, em previsões mais acuradas para as dinâmicas do PIB brasileiro.

Os resultados encontrados mostraram que os DFMs ajustados com as variáveis selecionadas pela RF MBB e pela RF CBB desempenharam bem em termos de *nowcasting* das taxas de crescimento do PIB brasileiro. Por outro lado, os DFMs ajustados com todas as variáveis (ALL) e com as variáveis selecionadas pelo LASSO e pelo ENET apresentaram uma acurácia preditiva inferior àquela do AR(1) em quase todos os períodos. Destaca-se que o AR(1) desempenhou consideravelmente bem para uma técnica simples, que demanda poucos dados e pouca capacidade computacional; em um dos períodos, o AR(1) teve um desempenho superior até mesmo que o DFM com a RF MBB, que foi a melhor especificação. Em outras palavras, os resultados sugerem existir ganhos significativos

com as metodologias utilizadas neste trabalho, a saber, a RF MBB e a RF CBB, com o AR(1) sendo uma alternativa viável e pouco custosa. Por fim, é prudente ressaltar que o bom desempenho do AR(1) está de acordo com o esperado, dado que a série da taxa de crescimento do PIB brasileiro é estacionária e, portanto, espera-se que suas dinâmicas sejam bem apreendidas por um modelo autorregressivo.

Não obstante, a literatura de *nowcasting* ainda apresenta lacunas, o que demanda pesquisas posteriores. Uma delas é a respeito da quantidade de variáveis originalmente coletadas para o estudo, já que a literatura é pouco conclusiva a respeito disso, com estudos utilizando cerca de 10 variáveis, enquanto outros utilizam mais de 100 (Bragoli; Metelli; Modugno, 2015; Cepni; Güney; Swanson, 2019b). Outra questão pertinente é a respeito da quantidade de variáveis mais importantes que devem ser selecionadas, dada a importância por permutação em blocos (IPB) da *Random Forest* (RF); a literatura de RFs não aponta uma quantidade de variáveis relevantes a serem utilizadas no modelo. Neste trabalho, optou-se por selecionar a mesma quantidade de variáveis que o LASSO, que é uma das técnicas mais bem consolidadas para a seleção prévia de variáveis na literatura de *nowcasting*. Por fim, uma terceira questão que necessita de estudos posteriores se refere ao desempenho do DFM ajustado com outros tipos de blocos de *bootstrap*, além do *Moving Block Bootstrap* (MBB) e do *Circular Block Bootstrap* (CBB) que foram utilizados neste trabalho.

REFERÊNCIAS

ANG, A.; BEKAERT, G.; WEI, M. Do macro variables, asset markets, or surveys forecast inflation better? *Journal of Monetary Economics*, v. 54, n. 4, p. 1163–1212, 2007.

BAI, J.; NG, S. Determining the number of factors in approximate factor models. *Econometrica*, Wiley, v. 70, n. 1, p. 191–221, jan. 2002. Available on: <https://doi.org/10.1111/1468-0262.00273>.

BAI, J.; NG, S. Forecasting economic time series using targeted predictors. *Journal of Econometrics*, v. 146, n. 2, p. 304–317, 2008.

BAŃBURA, M.; GIANNONE, D.; REICHLIN, L. *Nowcasting*. Frankfurt am Main, 2010. ECB Working Paper Series, n. 1275.

BAŃBURA, M.; MODUGNO, M. Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. *Journal of Applied Econometrics*, v. 29, n. 1, p. 133–160, 2014.

Banco Central do Brasil. *Série 11 – Taxa de juros – Selic*. 2025. Sistema Gerenciador de Séries Temporais – SGS. Acesso em: 5 nov. 2025. Dados fornecidos por André de Oliveira Castanheira Rodrigues (DEMAB/DICEL). Available on: <https://www3.bcb.gov.br/sgspub/localizarseries/localizarSeries.do?method=prepararTelaLocalizarSeries>.

Banco Central do Brasil. *Série 27841 – Meios de pagamento – M1 (saldo em final de período) – Novo – sazonalmente ajustado*. 2025. Sistema Gerenciador de Séries Temporais – SGS. Acesso em: 5 nov. 2025. Dados fornecidos por Helcio Magalhães Novaes (DSTAT/DIMOB/SUMON). Available on: <https://www3.bcb.gov.br/sgspub/localizarseries/localizarSeries.do?method=prepararTelaLocalizarSeries>.

Banco Central do Brasil. *Série 433 – Índice Nacional de Preços ao Consumidor Amplo (IPCA)*. 2025. Sistema Gerenciador de Séries Temporais – SGS. Acesso em: 5 nov. 2025. Dados fornecidos por Fernando Ryu Ramos Kawaoka (DEPEC/GEPRE). Available on: <https://www3.bcb.gov.br/sgspub/localizarseries/localizarSeries.do?method=prepararTelaLocalizarSeries>.

Banco Central do Brasil. *Índice de Atividade Econômica do Banco Central – IBC-Br*. 2025. <https://dadosabertos.bcb.gov.br/dataset/24363-indice-de-atividade-economica-do-banco-central---ibc-br>. Fonte: Banco Central do Brasil – Departamento Econômico. Conceito: indicador mensal contemporâneo da atividade econômica nacional. Disponível em: <https://dadosabertos.bcb.gov.br/dataset/24363-indice-de-atividade-economica-do-banco-central—ibc-br>. Acesso em: 4 out. 2025.

BANTIS, E.; CLEMENTS, M. P.; URQUHART, A. Forecasting gdp growth rates in the united states and brazil using google trends. *International Journal of Forecasting*, v. 39, n. 4, p. 1909–1924, 2023.

BARROS, A. C.; MATTOS, D. M. d.; OLIVEIRA, I. C. L. d.; FERREIRA, P. G. C.; DUCA, V. E. L. d. A. *Análise de séries temporais em R: curso introdutório*. 1. ed. Rio de Janeiro: GEN Atlas, 2017. 264 p. Instituto Brasileiro de Economia (FGV IBRE). ISBN 978-85-352-9087-5.

BERNANKE, B. S. et al. Systematic monetary policy and the effects of oil price shocks. *Brookings Papers on Economic Activity*, v. 1997, n. 1, p. 91–157, 1997.

Board of Governors of the Federal Reserve System. *About the Fed*. Washington, D.C.: , 2025. Available on: <https://www.federalreserve.gov/aboutthefed.htm>.

BOIVIN, J.; NG, S. Are more data always better for factor analysis? *Journal of Econometrics*, v. 132, n. 1, p. 169–194, 2006.

BRAGOLI, D.; METELLI, L.; MODUGNO, M. The importance of updating: Evidence from a brazilian nowcasting model. *OECD Journal: Journal of Business Cycle Measurement and Analysis*, v. 2015, n. 1, p. 5–22, 2015.

BREIMAN, L. Random forests. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001.

BUENO, R. d. L. d. S. *Econometria de séries temporais*. 2. ed. São Paulo: Cengage Learning, 2011.

BUREAU, A. et al. Identifying snps predictive of phenotype using random forests. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, v. 28, n. 2, p. 171–182, 2005.

CARLSTEIN, E. The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 14, n. 3, p. 1171–1179, 1986. Available on: <https://www.jstor.org/stable/3035565>.

CEPNI, O.; GÜNEY, I. E.; SWANSON, N. R. Forecasting and nowcasting emerging market gdp growth rates: The role of latent global economic policy uncertainty and macroeconomic data surprise factors. *Journal of Forecasting*, v. 39, n. 1, p. 18–36, 2019a.

CEPNI, O.; GÜNEY, I. E.; SWANSON, N. R. Nowcasting and forecasting gdp in emerging markets using global financial and macroeconomic diffusion indexes. *International Journal of Forecasting*, v. 35, n. 2, p. 555–572, 2019b.

CHOW, G. C.; LIN, A.-l. Best linear unbiased interpolation, distribution, and extrapolation of time series by related series. *The Review of Economics and Statistics*, The MIT Press, v. 53, n. 4, p. 372–375, nov. 1971. Available on: <https://www.jstor.org/stable/1928739>.

CLEMENTS, M. P. *Why are survey forecasts superior to model forecasts?* 2010.

COCHRANE, J. H. *Asset Pricing*. Princeton: Princeton University Press, 2002.

CUTLER, D. R. et al. Random forests for classification in ecology. *Ecology*, v. 88, n. 11, p. 2783–2792, 2007.

DAHLHAUS, T.; GUÉNETTE, J.-D.; VASISHTHA, G. Nowcasting bric+ m in real time. *International Journal of Forecasting*, v. 33, n. 4, p. 915–935, 2017.

DÍAZ-URIARTE, R.; ANDRÉS, S. Alvarez de. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, v. 7, p. 1–13, 2006.

DOZ, C.; GIANNONE, D.; REICHLIN, L. A maximum likelihood approach for large approximate dynamic factor models. *Review of Economics and Statistics*, 2006.

DOZ, C.; GIANNONE, D.; REICHLIN, L. A two-step estimator for large approximate dynamic factor models based on kalman filtering. *Journal of Econometrics*, v. 164, n. 1, p. 188–205, set. 2011. ISSN 0304-4076. Available on: <https://doi.org/10.1016/j.jeconom.2011.02.012>.

DOZ, C.; GIANNONE, D.; REICHLIN, L. A quasi–maximum likelihood approach for large, approximate dynamic factor models. *The Review of Economics and Statistics*, MIT Press, v. 94, n. 4, p. 1014–1024, nov. 2012. Available on: <https://doi.org/10.1162/REST_a_00225>.

ENDERS, W. *Applied Econometric Time Series*. 4. ed. New Jersey: John Wiley & Sons, 2008.

ERP, S. van; OBERSKI, D. L.; MULDER, J. Shrinkage priors for bayesian penalized regression. *Journal of Mathematical Psychology*, v. 89, p. 31–50, 2019.

EVANS, M. D. D. Where are we now? real-time estimates of the macroeconomy. *International Journal of Central Banking*, v. 1, n. 2, p. 127–175, 2005.

FAN, J.; LI, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, v. 96, p. 1348–1360, 2001.

FANG, W. et al. An evaluation of random forest-based input variable selection methods for one month ahead streamflow forecasting. *Scientific Reports*, Nature Portfolio, London, v. 14, p. 29766, 2024. Disponível em: <https://doi.org/10.1038/s41598-024-81502-y>. Acesso em: 4 nov. 2025. Available on: <https://doi.org/10.1038/s41598-024-81502-y>.

FENG, Y.; ZHANG, Y.; WANG, Y. Out-of-sample volatility prediction: Rolling window, expanding window, or both? *Journal of Forecasting*, 2023.

FRIEDBERG, R. et al. Local linear forests. *Journal of Computational and Graphical Statistics*, v. 30, n. 2, p. 503–517, 2020.

FRIEDMAN, J. et al. *Package 'glmnet': Lasso and Elastic-Net Regularized Generalized Linear Models*. 2025. Version 4.1-10. Available on: <https://glmnet.stanford.edu>.

FU, W. J. Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, v. 7, n. 3, p. 397–416, 1998.

GIANNONE, D.; REICHLIN, L.; SMALL, D. Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, v. 55, n. 4, p. 665–676, 2008.

GOEHRY, B.; YAN, H.; GOUDE, Y.; MASSART, P.; POGGI, J.-M. Random forests for time series. *REVSTAT – Statistical Journal*, Statistics Portugal, Lisboa, v. 21, n. 2, p. 283–302, jun. 2023. Open Access sob licença Creative Commons Attribution 4.0 International License. Available on: <https://revstat.ine.pt/index.php/REVSTAT/article/view/400>.

HAMILTON, J. D. *Time series analysis*. Princeton, NJ: Princeton University Press, 1994. ISBN 0-691-04289-6.

HARVEY, A. C. *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press, 1989.

HASTIE, T.; QIAN, J.; TAY, K. *An Introduction to glmnet*. 2025. Available on: <https://glmnet.stanford.edu>.

HUANG, N.; LU, G.; XU, D. A permutation importance-based feature selection method for short-term electricity load forecasting using random forest. *Energies*, MDPI, Basel, Switzerland, v. 9, n. 10, p. 767, 2016. ISSN 1996-1073. Disponível em: <https://doi.org/10.3390/en9100767>. Acesso em: 4 nov. 2025. Available on: <https://doi.org/10.3390/en9100767>.

HUANG, X. et al. A comparative study of discriminating human heart failure etiology using gene expression profiles. *BMC Bioinformatics*, v. 6, n. 1, p. 205, 2005.

IBGE. *SCN – Sistema de Contas Nacionais*. 2023. Portal do IBGE. Acesso em: 09 jan. 2026. Available on: <https://www.ibge.gov.br/estatisticas/economicas/contas-nacionais/9052-sistema-de-contas-nacionais-brasil.html>.

IBGE. *Próximas divulgações*. 2025a. <https://www.ibge.gov.br/calendario-de-divulgacoes-novoportal.html>. Acesso em: 25 set. 2025.

IBGE. *SCNT – Sistema de Contas Nacionais Trimestrais*. 2025b. <https://www.ibge.gov.br/estatisticas/economicas/industria/9300-contas-nacionais-trimestrais.html>. Acesso em: 25 set. 2025.

IBGE. *SIDRA – Tabela 5932: Taxa de variação do índice de volume trimestral (Contas Nacionais Trimestrais)*. 2025c. Sistema IBGE de Recuperação Automática – SIDRA. Acesso em: 5 nov. 2025. Available on: <https://sidra.ibge.gov.br/tabela/5932>.

ISSLER, J. V.; NOTINI, H. H. Estimating brazilian monthly gdp: A state-space approach. *Revista Brasileira de Economia*, v. 70, n. 1, p. 41–59, 2016.

ISSLER, J. V.; PIMENTEL, L. M. d. M. Uma medida de pib mensal para o brasil usando o term spread. *Revista Brasileira de Economia*, v. 73, n. 1, p. 53–75, 2019.

JAMES, G. et al. *An Introduction to Statistical Learning: With Applications in R*. New York: Springer, 2013.

JANSEN, W. J.; JIN, X.; WINTER, J. M. de. Forecasting and nowcasting real gdp: Comparing statistical models and subjective forecasts. *International Journal of Forecasting*, v. 32, n. 2, p. 411–436, 2016.

KIM, H. H.; SWANSON, N. R. Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods. *International Journal of Forecasting*, v. 34, n. 2, p. 339–354, 2018.

KRANTZ, S. *Package 'dfms': Dynamic Factor Models*. Vienna, 2025a. Version 0.3.1. Available on: <https://cran.r-project.org/web/packages/dfms/>.

KRANTZ, S. *Introduction to dfms*. 2025b. Available on: <https://cran.r-project.org/web/packages/dfms/vignettes/introduction.html>.

KÜNSCH, H. R. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 17, n. 3, p. 1217–1241, 1989. Available on: <https://www.jstor.org/stable/2241719>.

LIEBERMANN, J. Real-time nowcasting of gdp: A factor model vs. professional forecasters. *Oxford Bulletin of Economics and Statistics*, v. 76, n. 6, p. 783–811, 2014.

LIU, H.; HALL, S. G. Creating high-frequency national accounts with state-space modelling: a monte carlo experiment. *Journal of Forecasting*, Wiley, v. 20, n. 6, p. 441–449, set. 2001. Available on: <https://doi.org/10.1002/for.810>.

LIU, R. Y.; SINGH, K. Moving blocks jackknife and bootstrap capture weak dependence. In: LEPAGE, R.; BILLARD, L. (Ed.). *Exploring the Limits of Bootstrap*. New York: John Wiley & Sons, 1992. p. —. Chapter in edited volume.

LUNETTA, K. L. et al. Screening large-scale association study data: exploiting interactions using random forests. *BMC Genetics*, v. 5, n. 1, p. 32, 2004.

MARIANO, R.; MURASAWA, Y. A new coincident index of business cycles based on monthly and quarterly series. *Journal of Applied Econometrics*, v. 18, p. 427–443, 2003.

MöNCH, E.; UHLIG, H. *Towards a monthly business cycle chronology for the euro area*. 2005.

POLITIS, D. N.; ROMANO, J. P. A circular block resampling procedure for stationary data. In: LEPAGE, R.; BILLARD, L. (Ed.). *Exploring the Limits of Bootstrap*. New York: John Wiley & Sons, 1992. p. 263–270.

QI, Y.; BAR-JOSEPH, Z.; KLEIN-SEETHARAMAN, J. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins: Structure, Function, and Bioinformatics*, v. 63, n. 3, p. 490–500, 2006.

ROSSI, B.; INOUE, A. Out-of-sample forecast tests robust to the choice of window size. *Journal of Business & Economic Statistics*, v. 30, n. 3, p. 432–453, 2012.

STROBL, C. et al. Conditional variable importance for random forests. *BMC Bioinformatics*, v. 9, p. 1–11, 2008.

STROBL, C.; ZEILEIS, A. *Danger: High Power! – Exploring the Statistical Properties of a Test for Random Forest Variable Importance*. München, 2008. Available on: <https://epub.ub.uni-muenchen.de/2111/>.

SVETNIK, V. et al. Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of Chemical Information and Computer Sciences*, v. 43, n. 6, p. 1947–1958, 2003.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, v. 58, n. 1, p. 267–288, 1996.

WRIGHT, M. N. *rangerts: A Modified Version of ranger for Time Series*. 2019. <https://github.com/hyanworkspace/rangerts>. Acessado em: 13 nov. 2025.

ZOU, H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, v. 101, n. 476, p. 1418–1429, 2006.

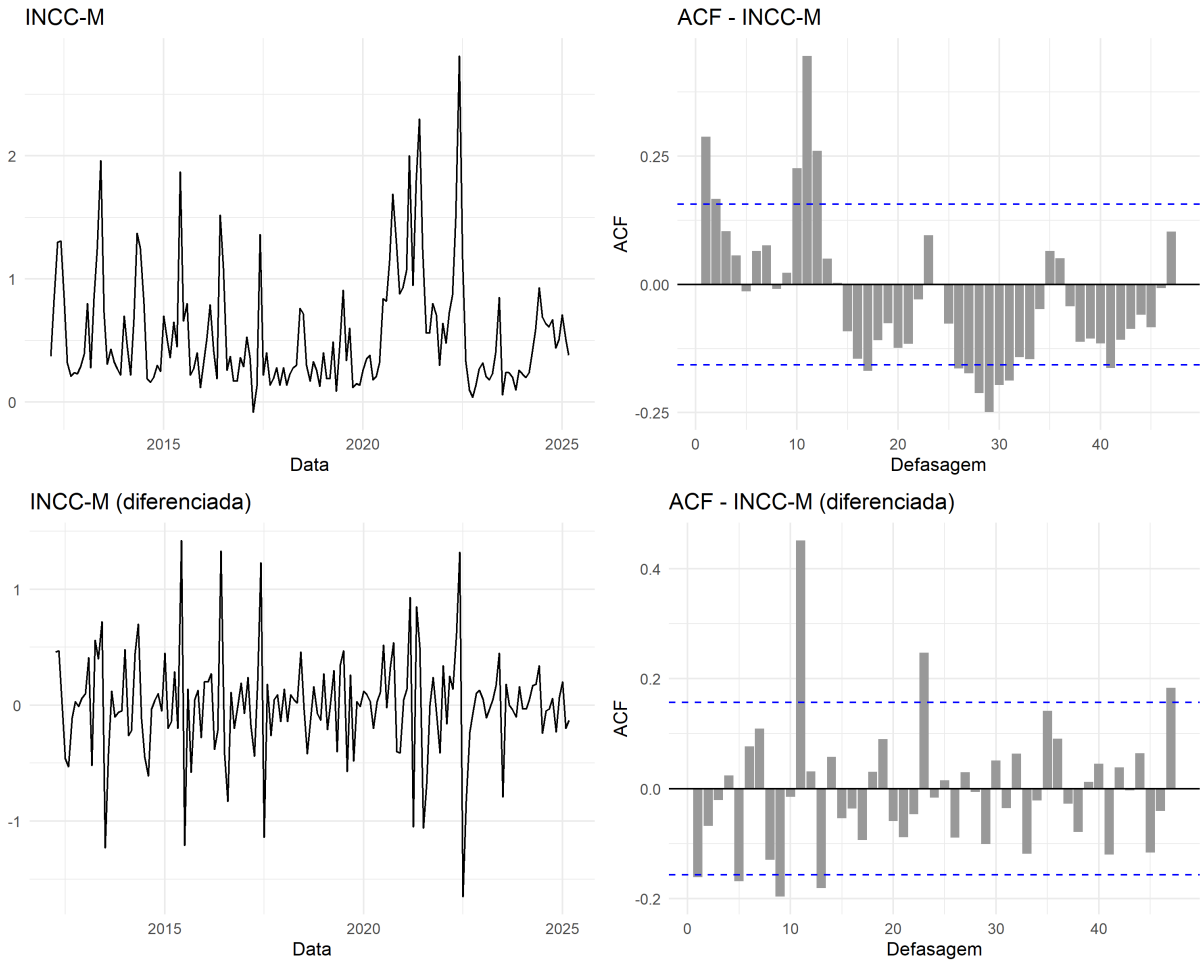
ZOU, H.; HASTIE, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, v. 67, n. 2, p. 301–320, 2005.

APÊNDICE A – Gráficos das séries e dos ACFs

No presente Apêndice serão apresentadas, visualmente, as 44 séries temporais originais e diferenciadas. Dentre essas variáveis, todas possuem uma periodicidade mensal, com exceção do PIB, cuja frequência é trimestral. É possível observar que muitas dessas séries apresentam tendências e/ou outros comportamentos não estacionários. Assim sendo, tornou-se necessário realizar diferenciações não sazonais e/ou sazonais.

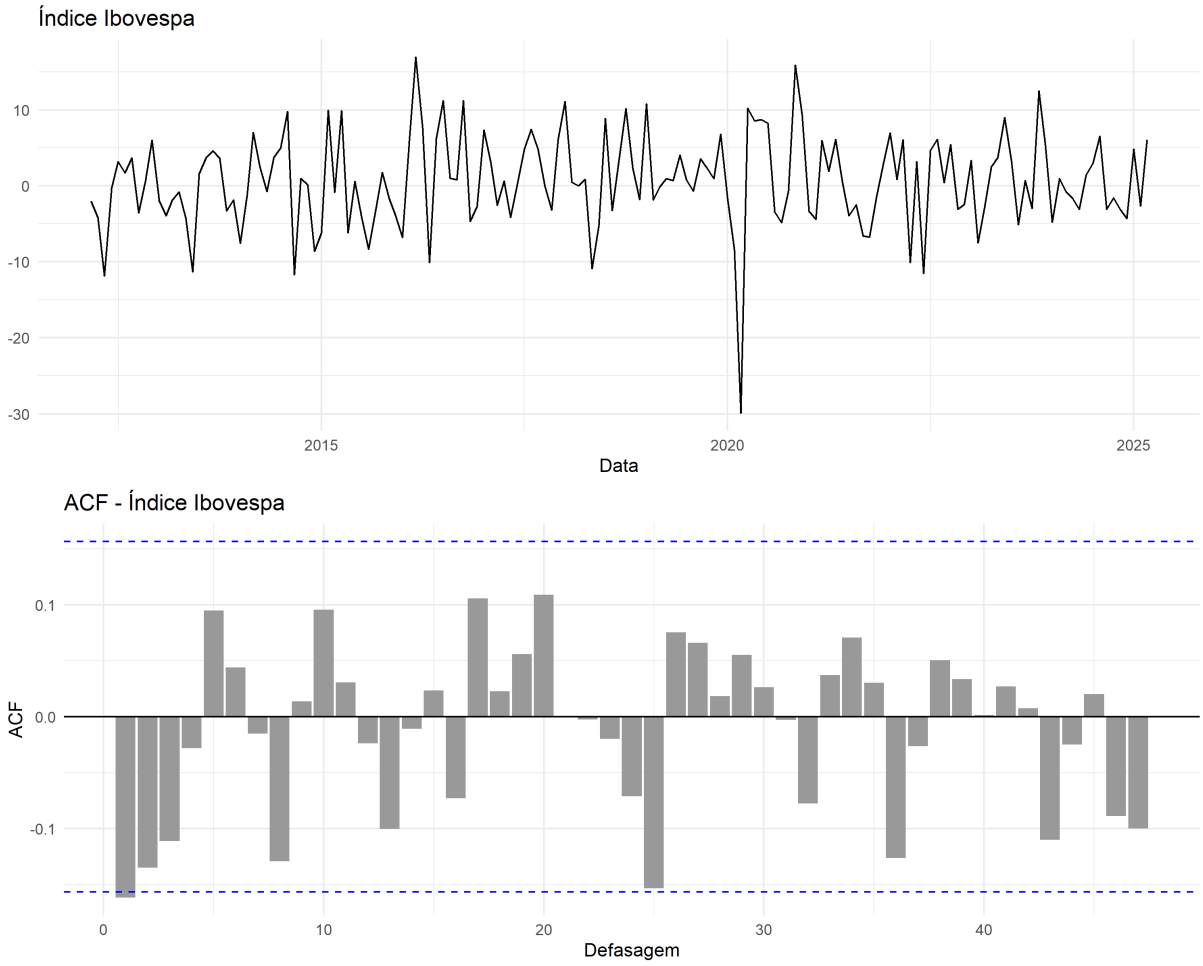
As Figuras de 5 a 48 apresentam os gráficos das séries temporais originais e diferenciadas, bem como seus respectivos ACFs.

Figura 5 – Gráficos - Índice Nacional de Custo da Construção (INCC-M)



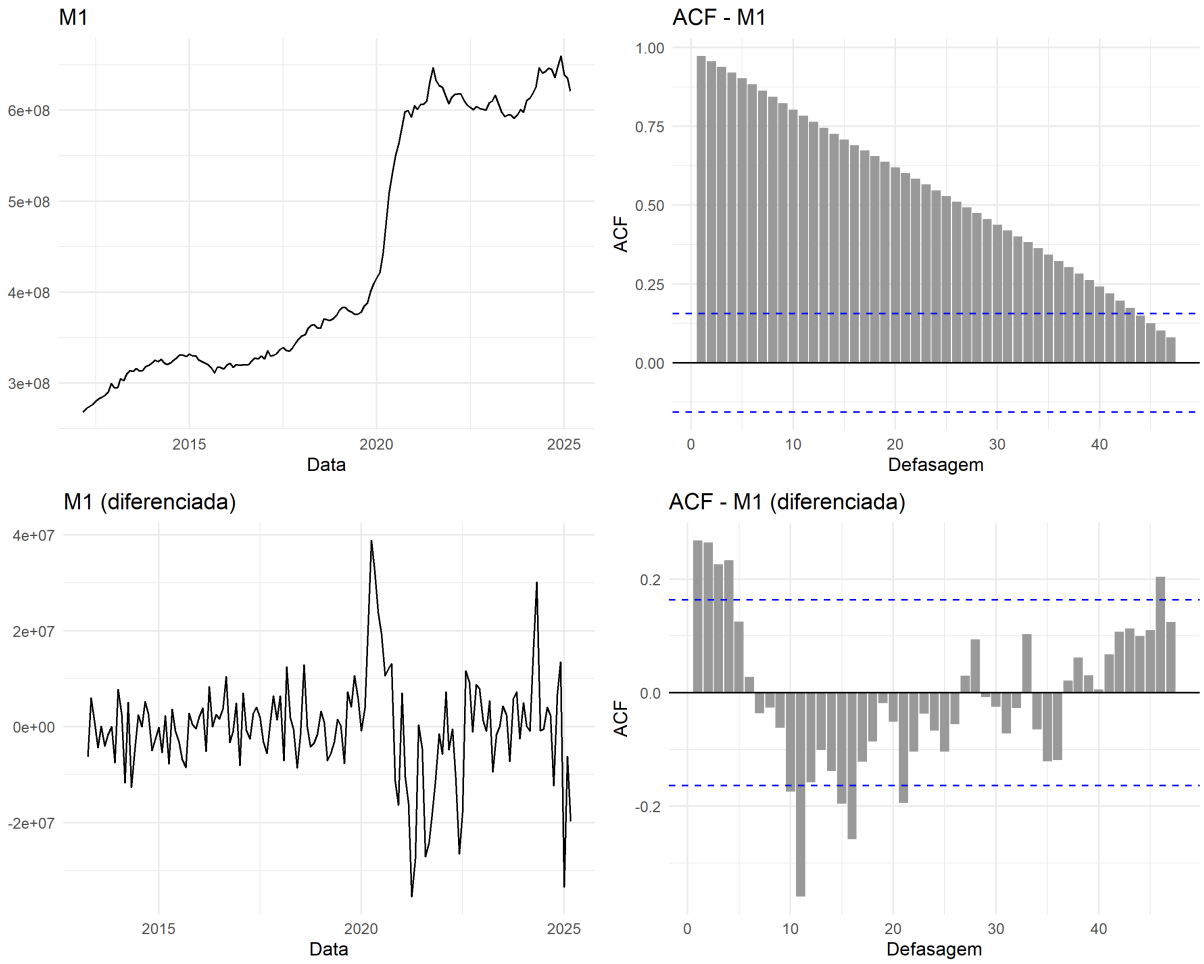
Fonte: Elaboração própria.

Figura 6 – Gráficos - Índice de ações: Ibovespa - fechamento



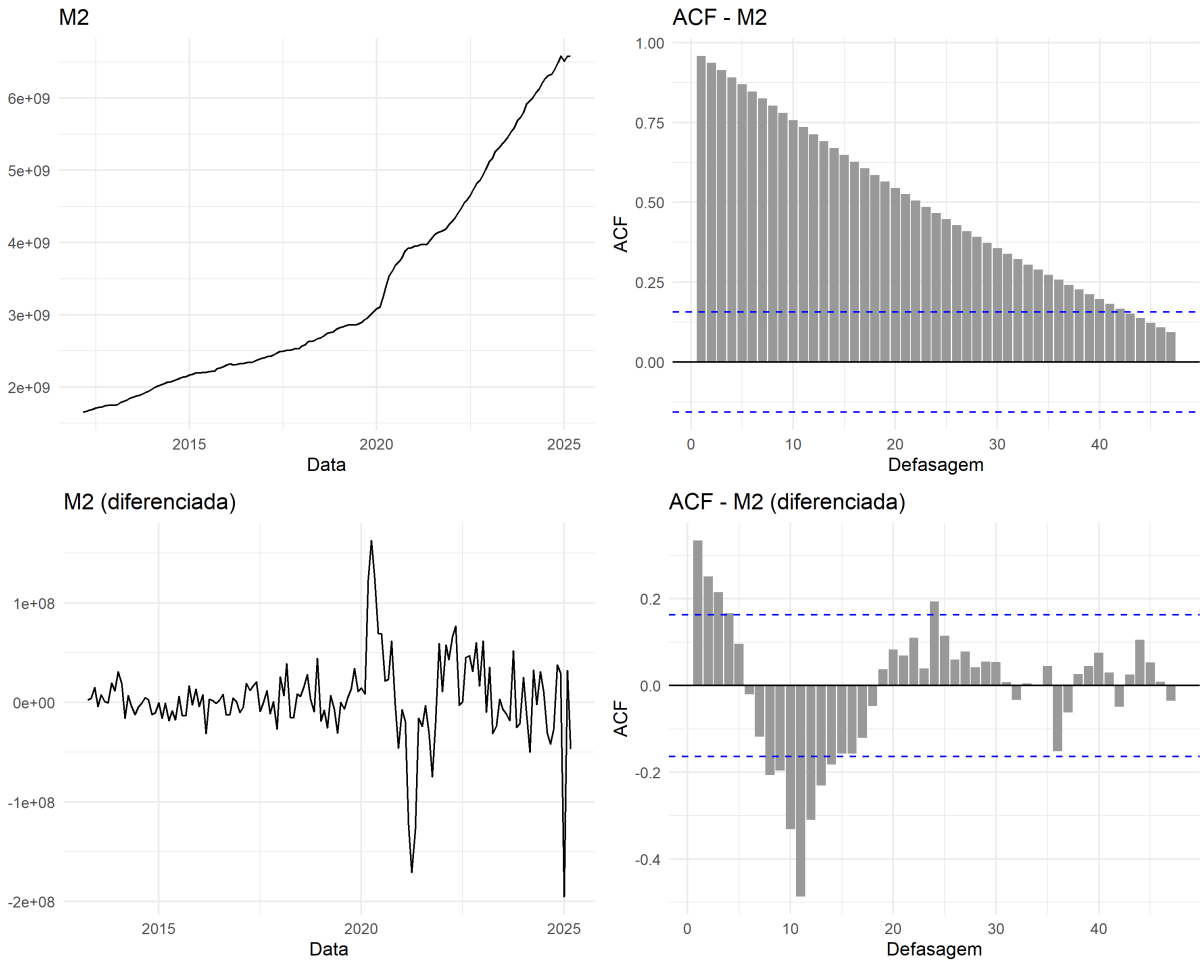
Fonte: Elaboração própria.

Figura 7 – Gráficos - Meios de pagamento - M1



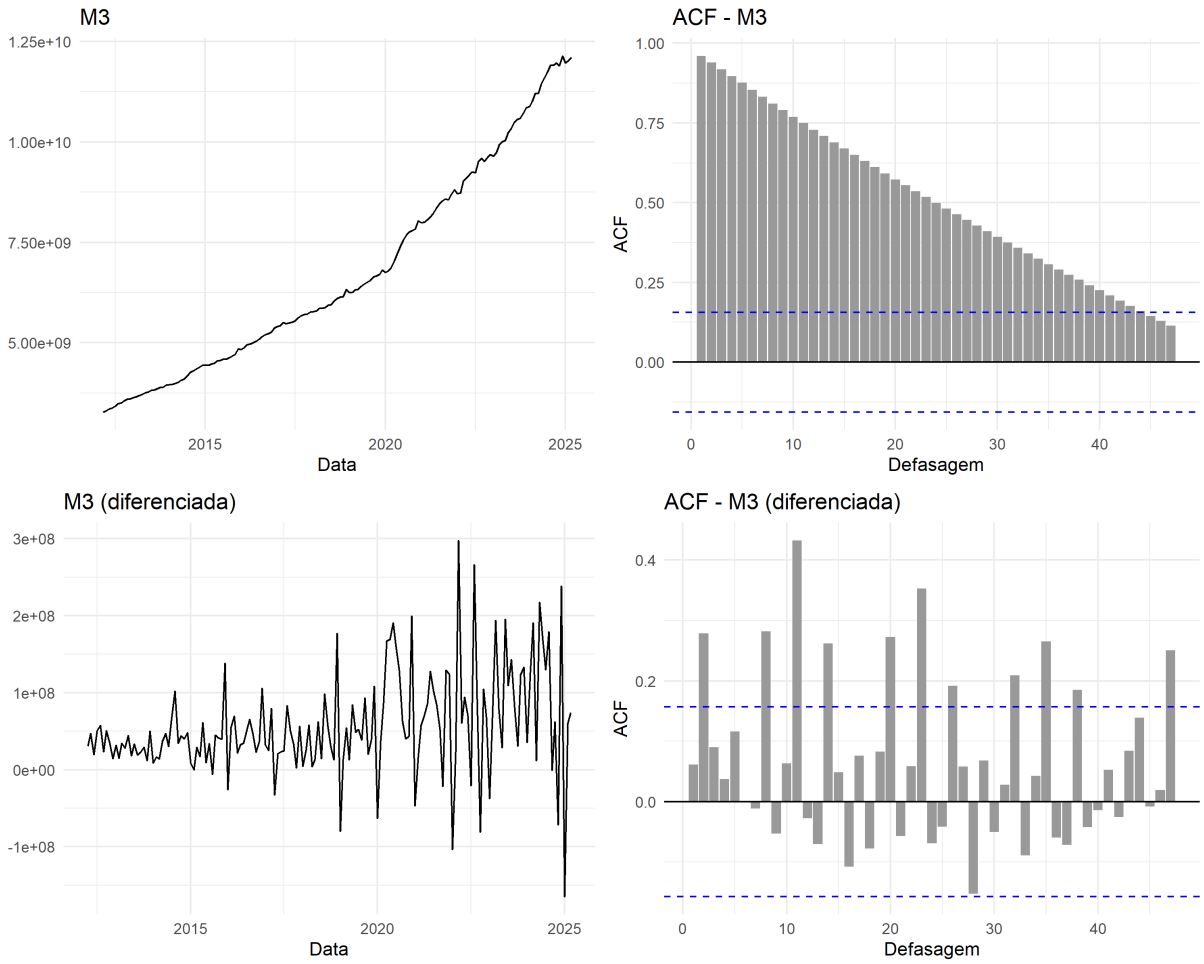
Fonte: Elaboração própria.

Figura 8 – Gráficos - Meios de pagamento amplos - M2



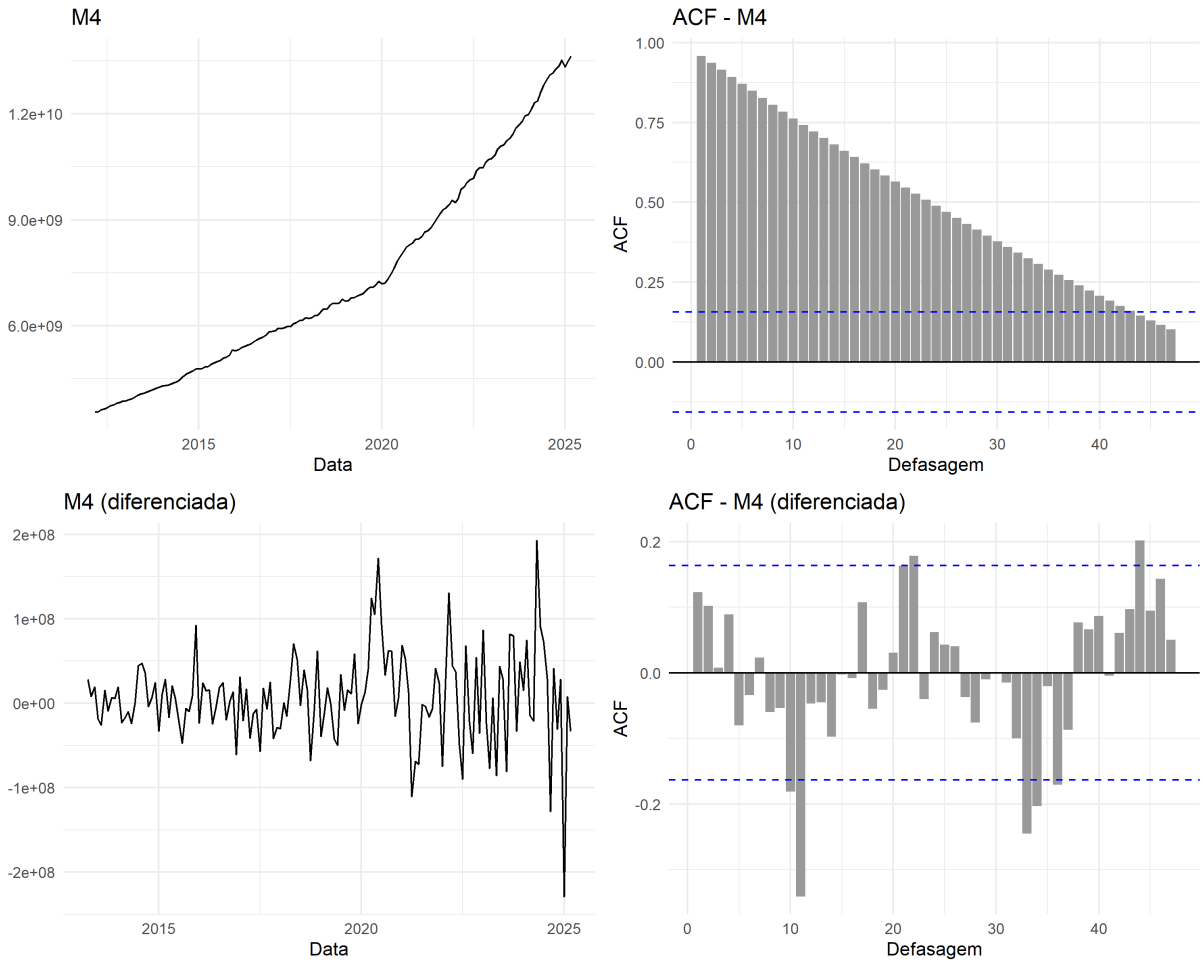
Fonte: Elaboração própria.

Figura 9 – Gráficos - Meios de pagamento amplos - M3



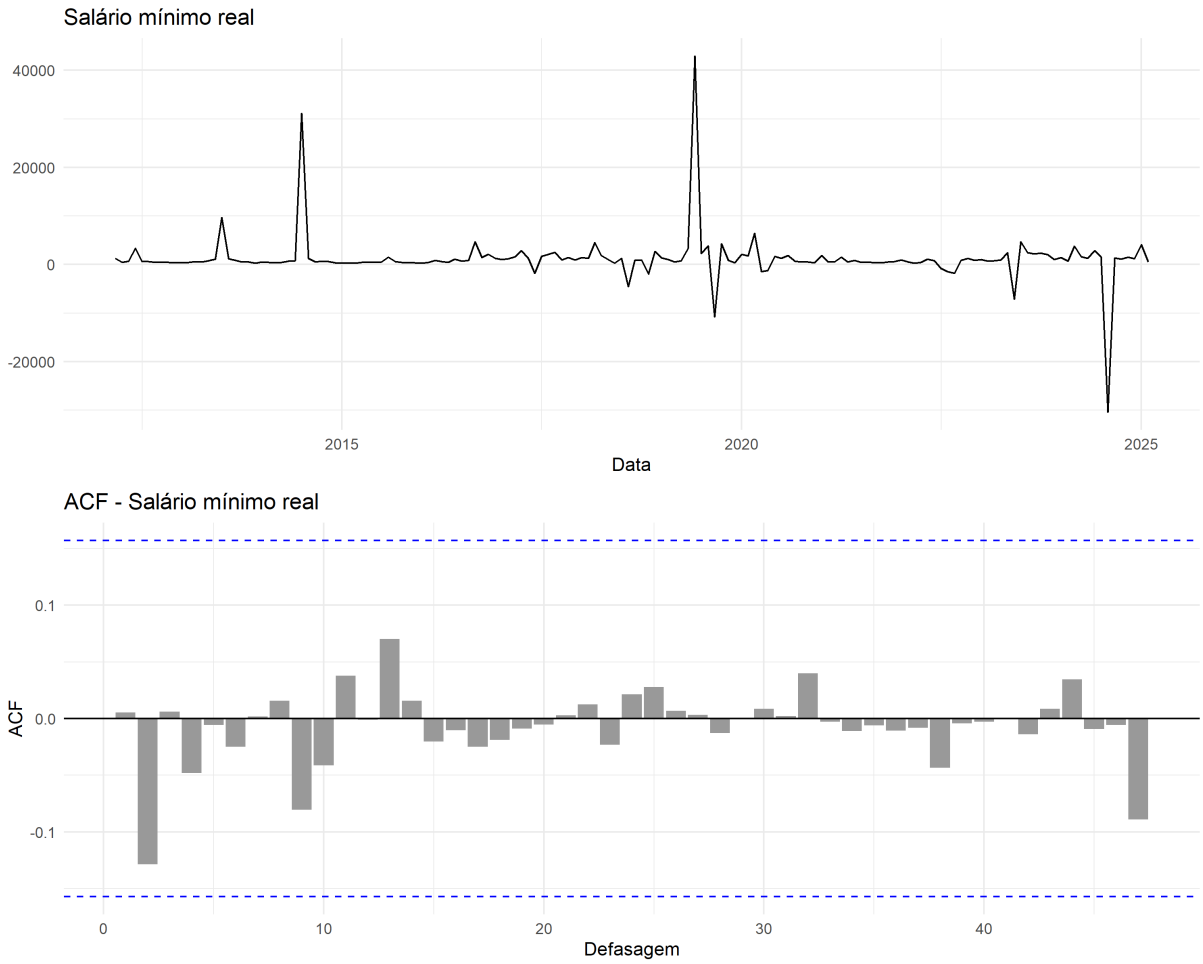
Fonte: Elaboração própria.

Figura 10 – Gráficos - Meios de pagamento amplos - M4



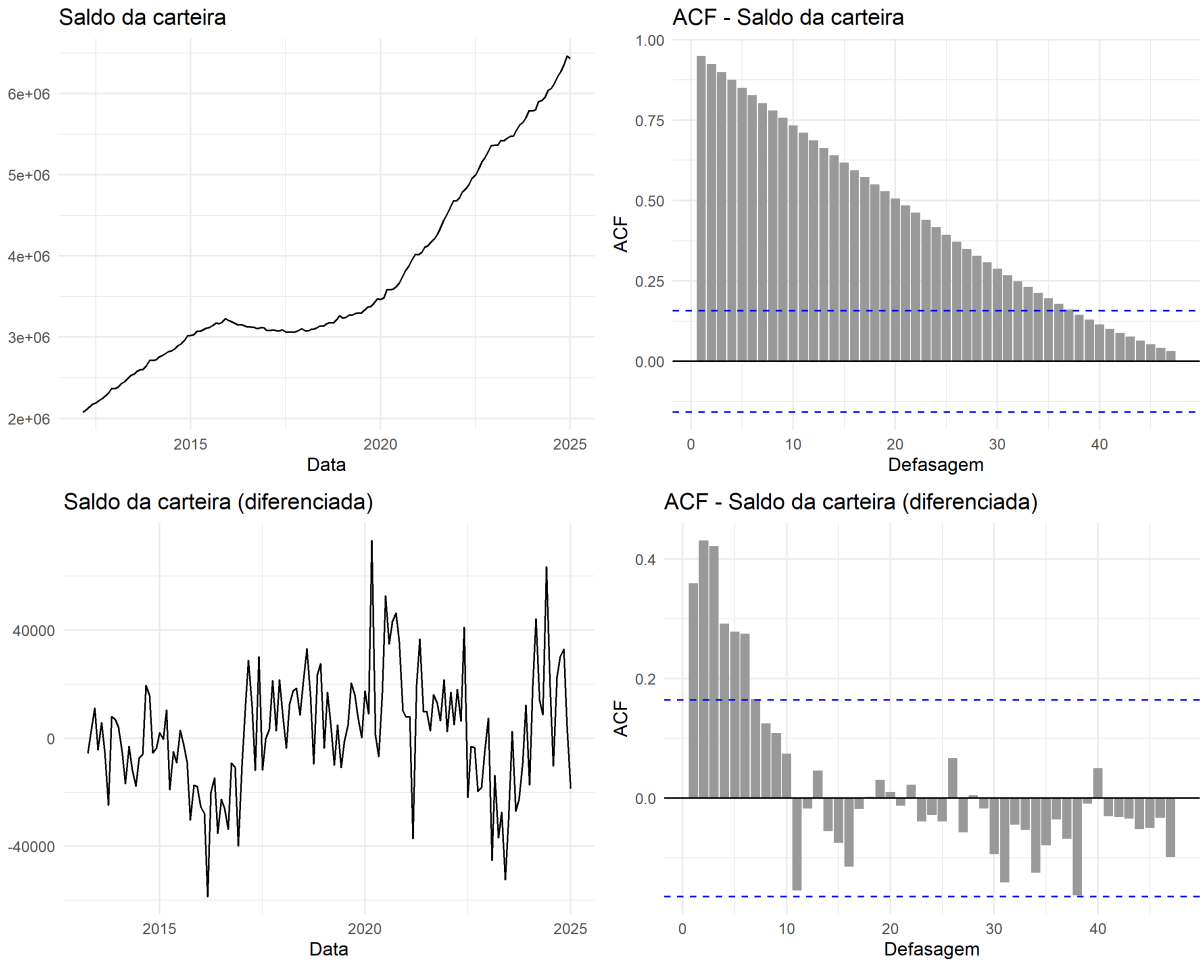
Fonte: Elaboração própria.

Figura 11 – Gráficos - Salário mínimo (deflacionado com o IPCA)



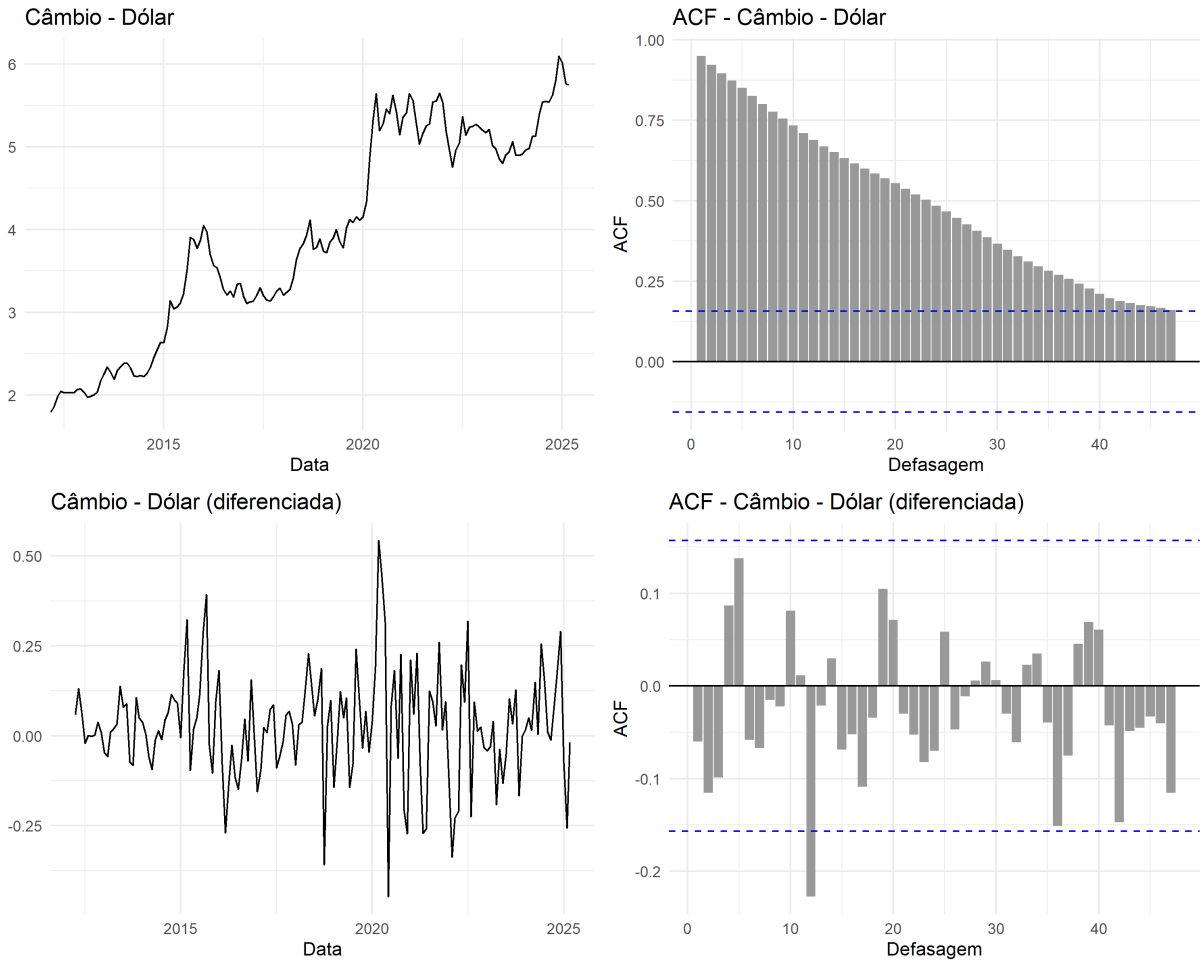
Fonte: Elaboração própria.

Figura 12 – Gráficos - Saldo da carteira de crédito - Total



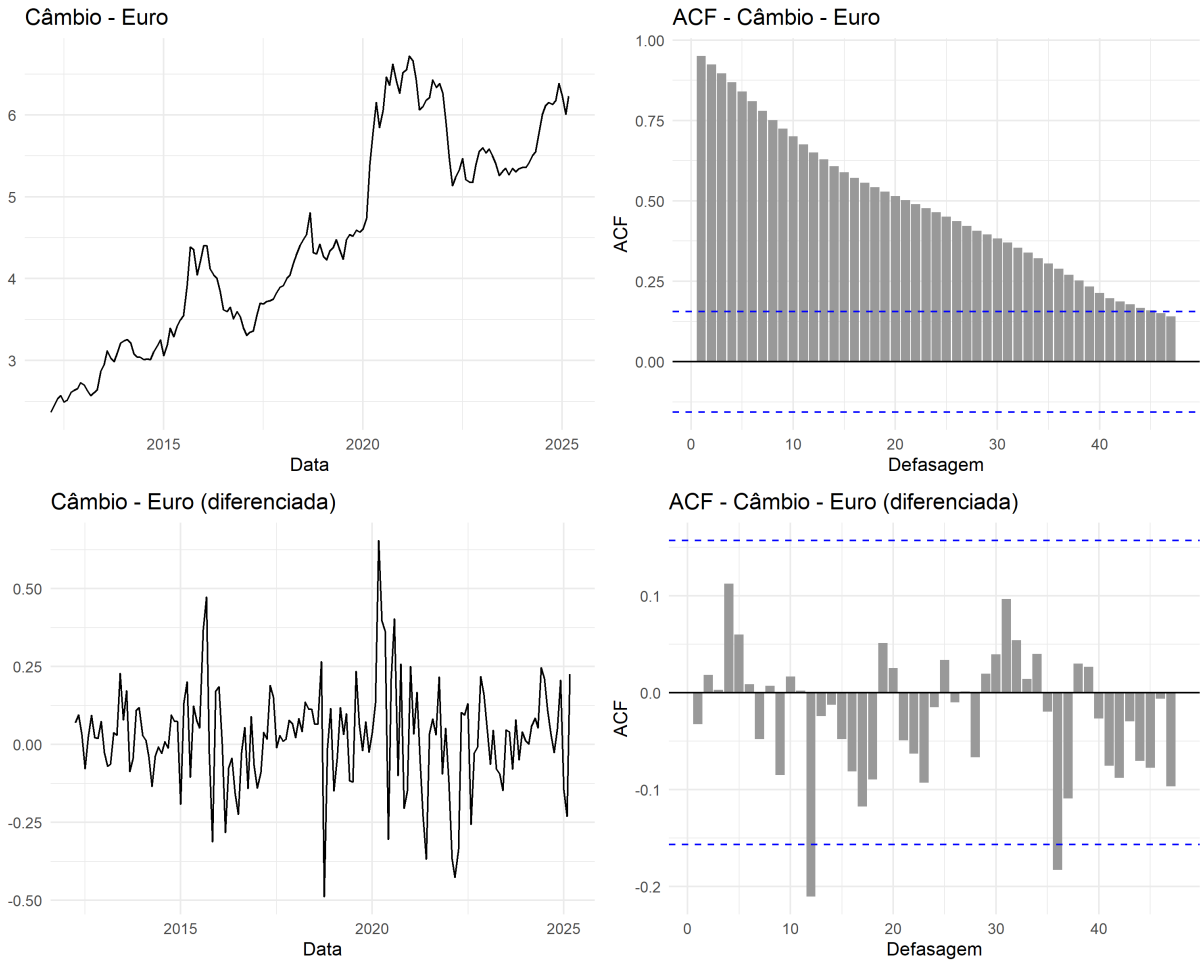
Fonte: Elaboração própria.

Figura 13 – Gráficos - Taxa de câmbio - Livre - Dólar americano (venda)



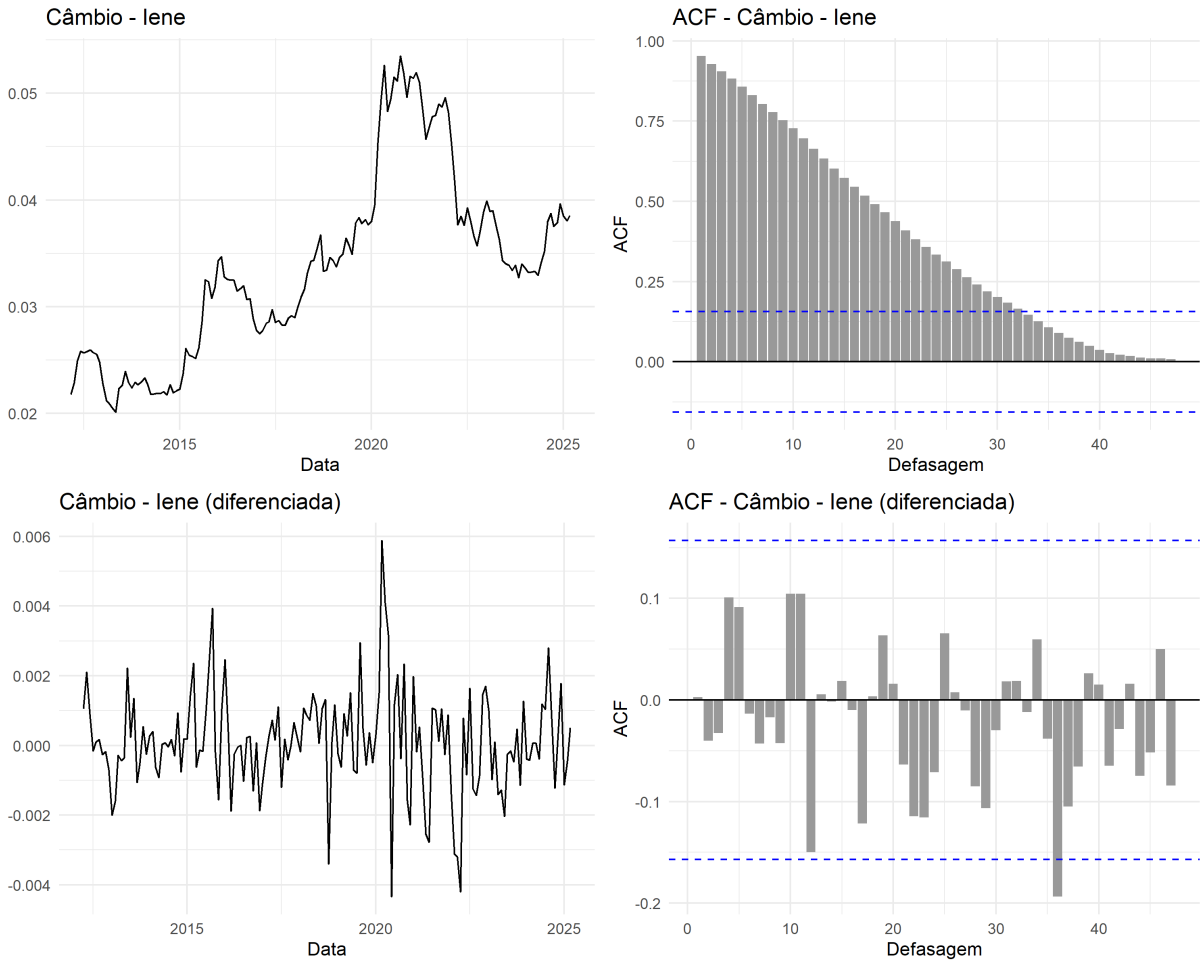
Fonte: Elaboração própria.

Figura 14 – Gráficos - Taxa de câmbio - Livre - Euro (venda)



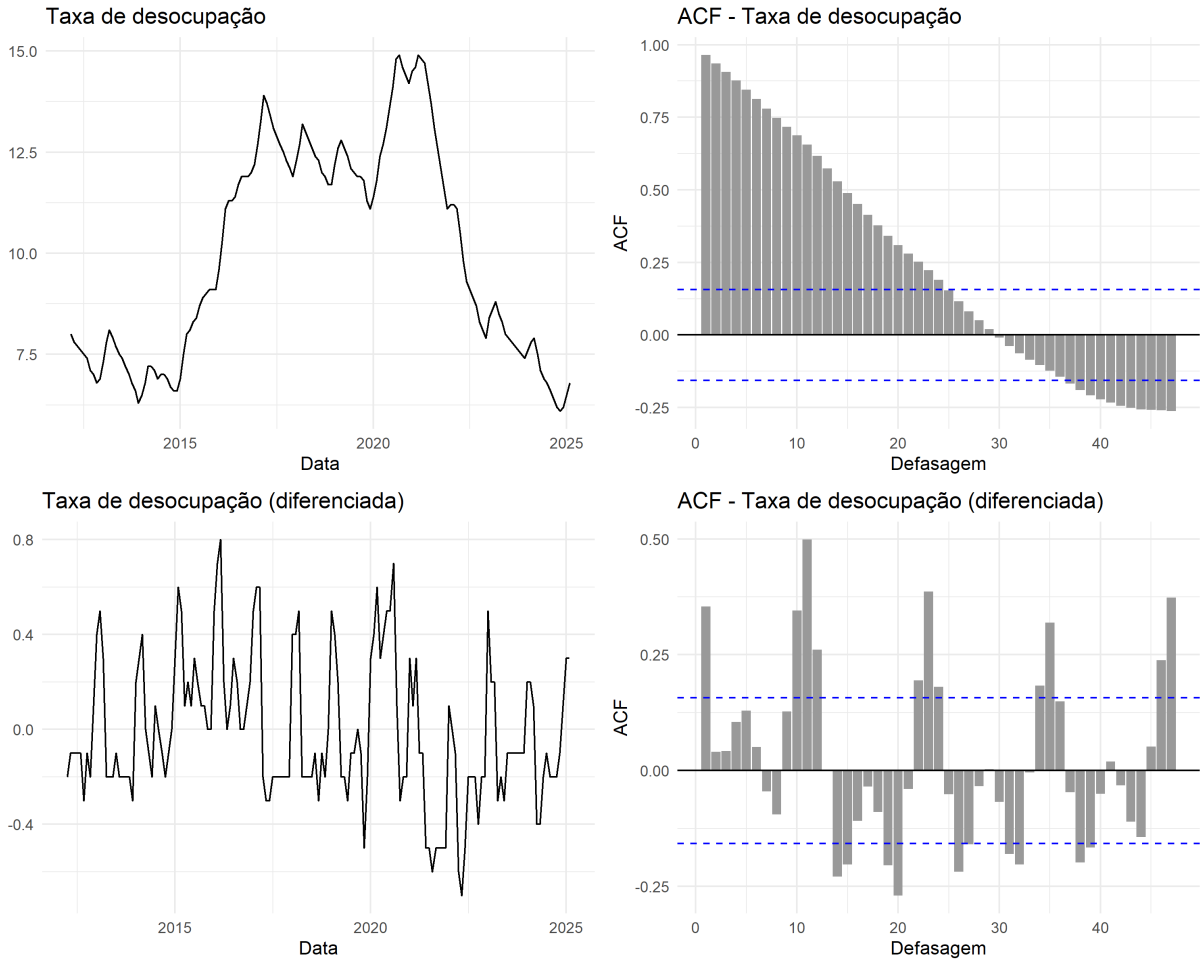
Fonte: Elaboração própria.

Figura 15 – Gráficos - Taxa de câmbio - Livre - Iene (venda)



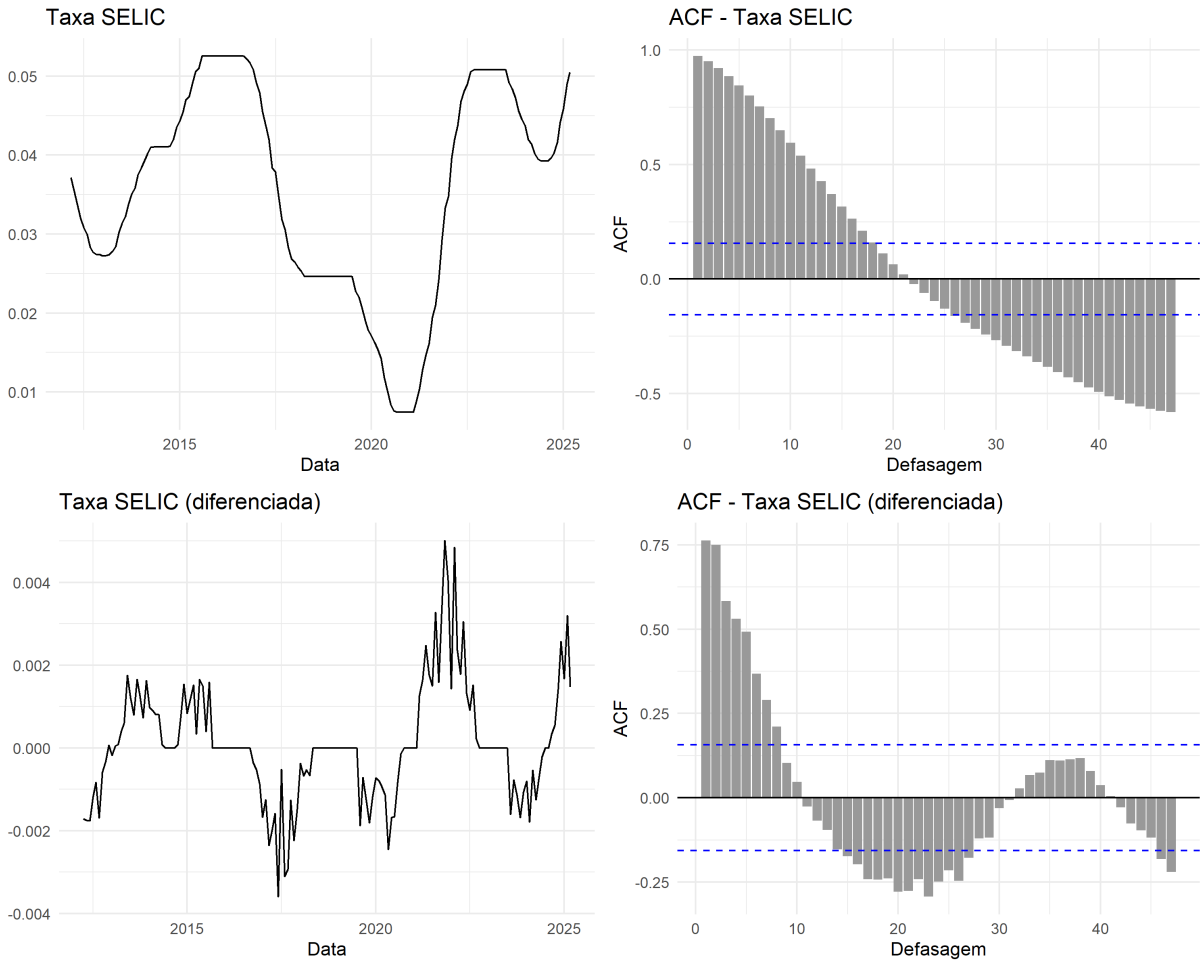
Fonte: Elaboração própria.

Figura 16 – Gráficos - Taxa de desocupação das pessoas de 14 anos ou mais de idade, na semana de referência



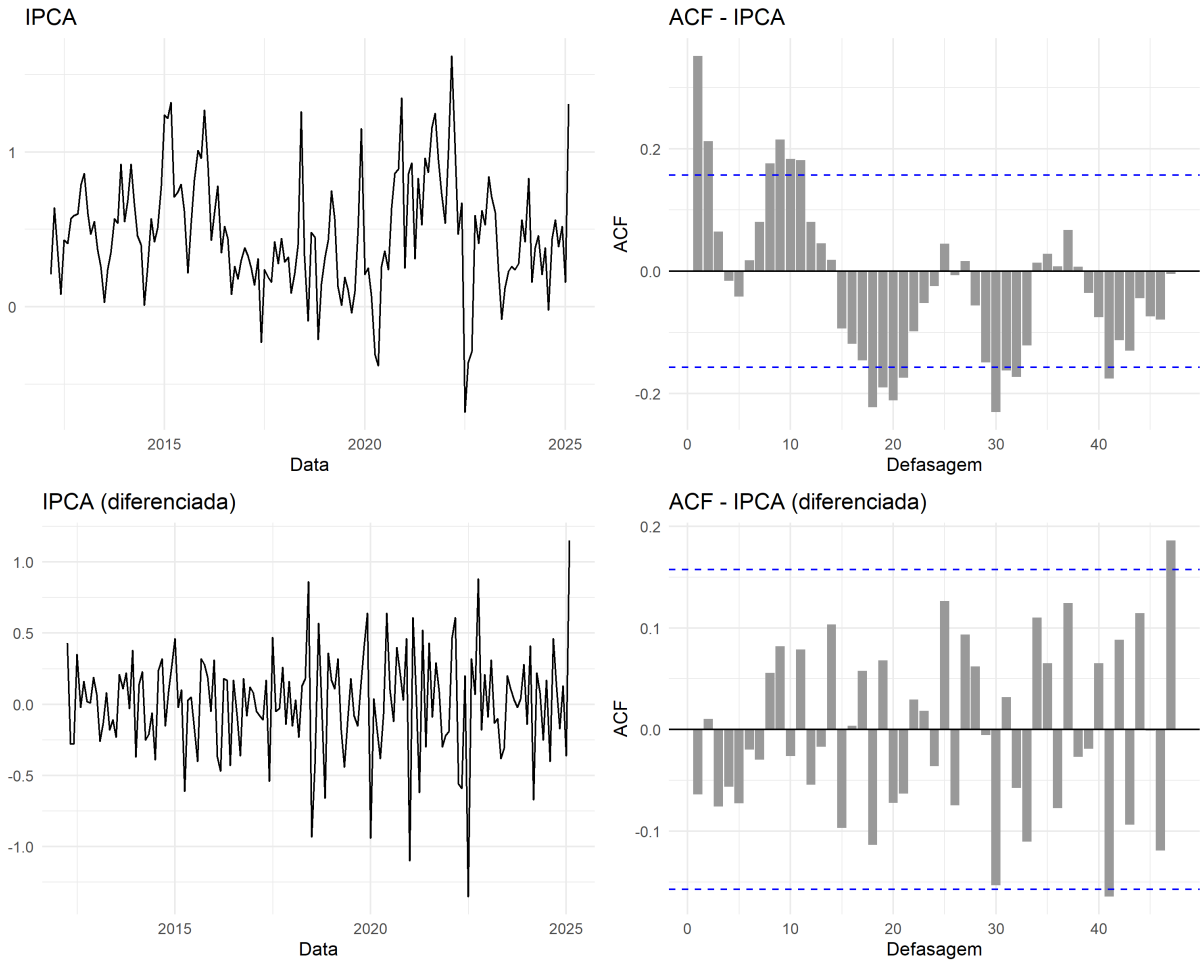
Fonte: Elaboração própria.

Figura 17 – Gráficos - Percentual ao dia da Taxa de Juros Selic (média mensal)



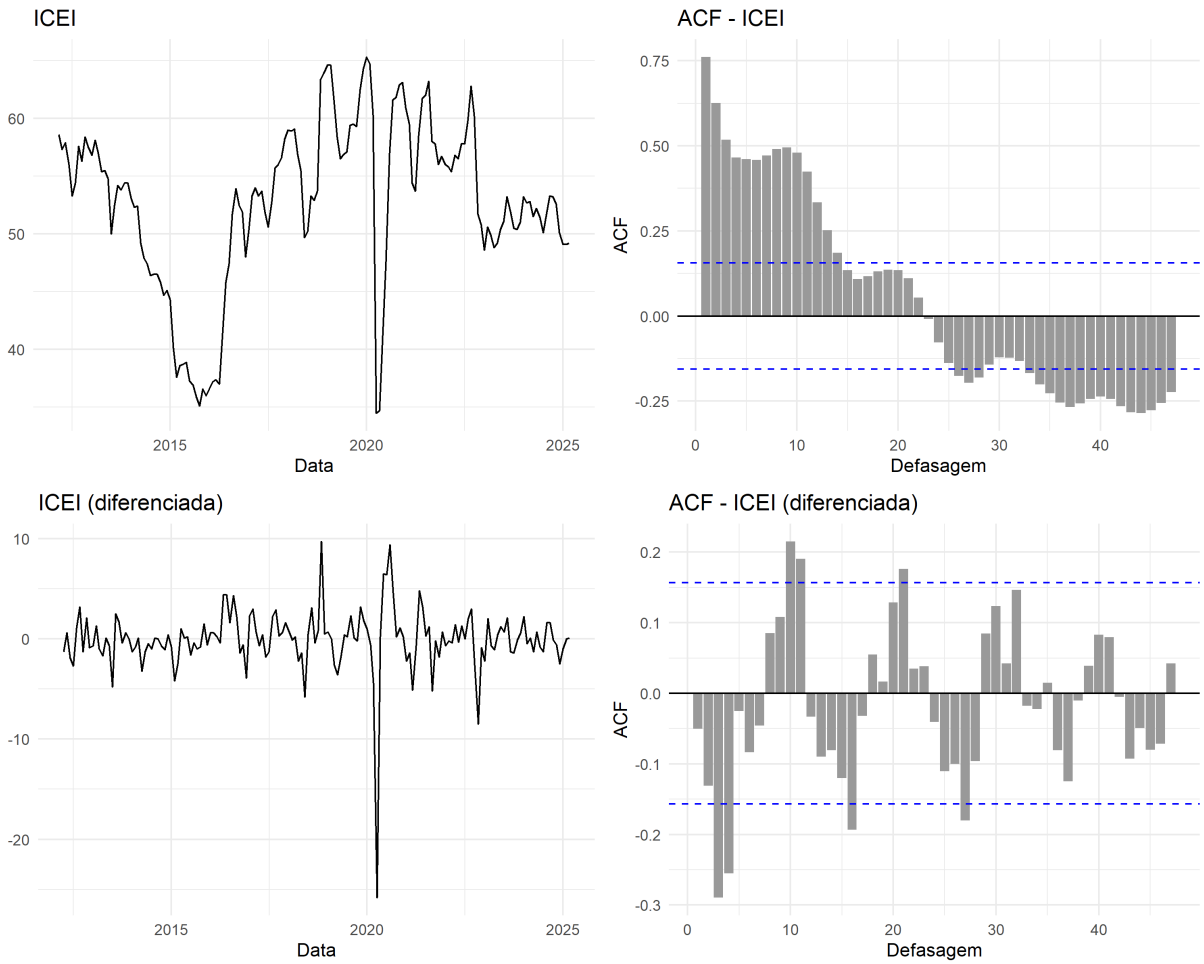
Fonte: Elaboração própria.

Figura 18 – Gráficos - Variação percentual mensal do IPCA



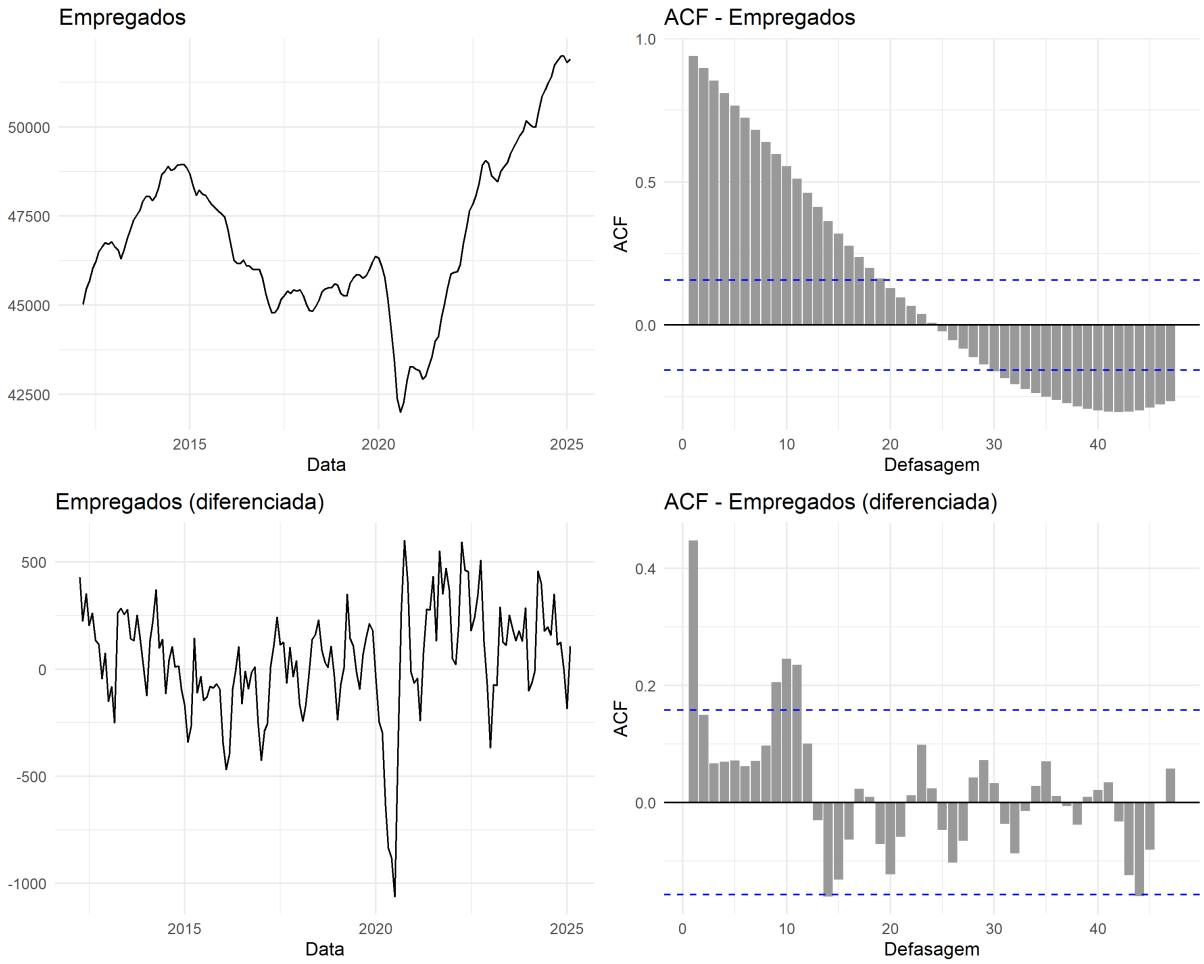
Fonte: Elaboração própria.

Figura 19 – Gráficos - Índice de confiança do empresário industrial (ICEI) geral



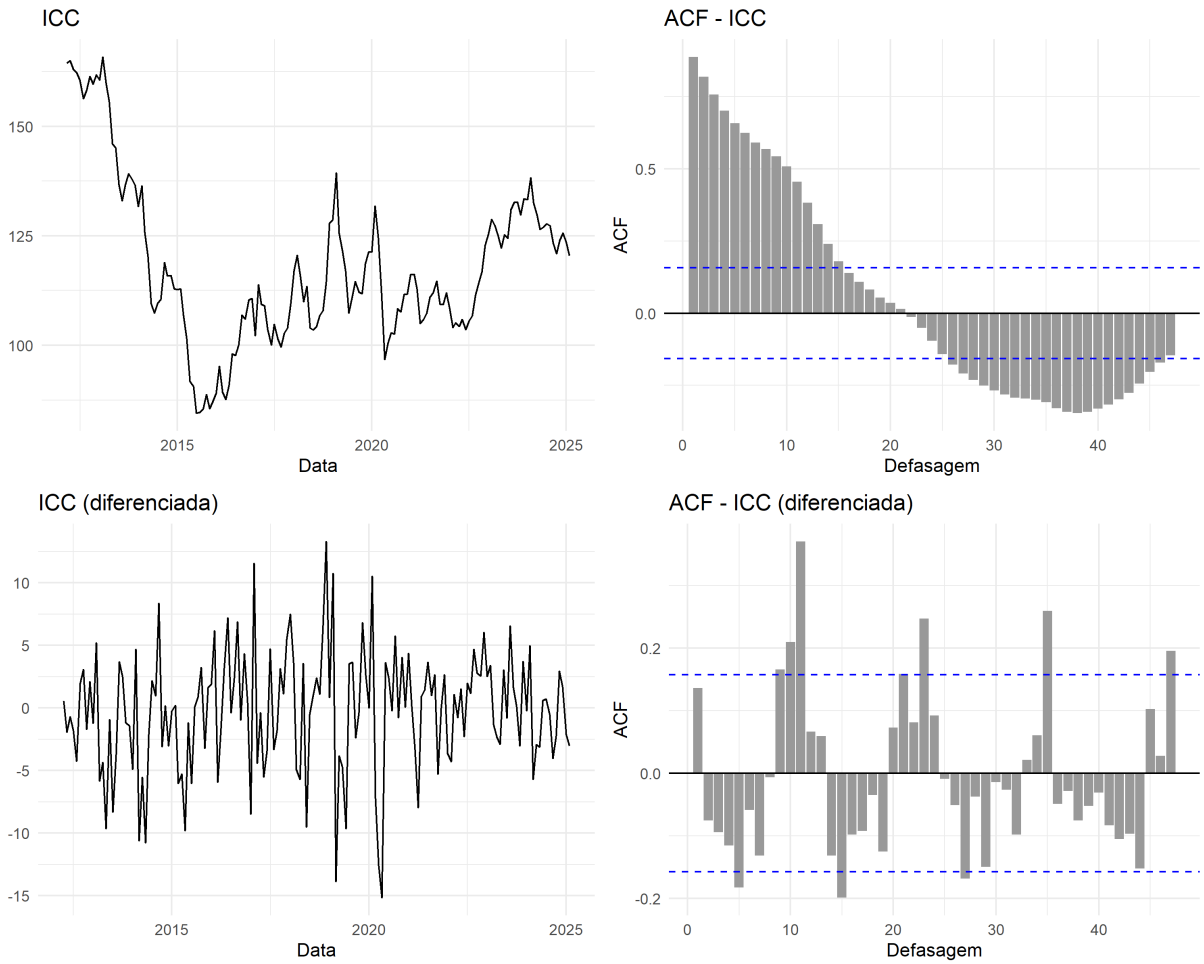
Fonte: Elaboração própria.

Figura 20 – Gráficos - Empregados no setor público e privado com carteira



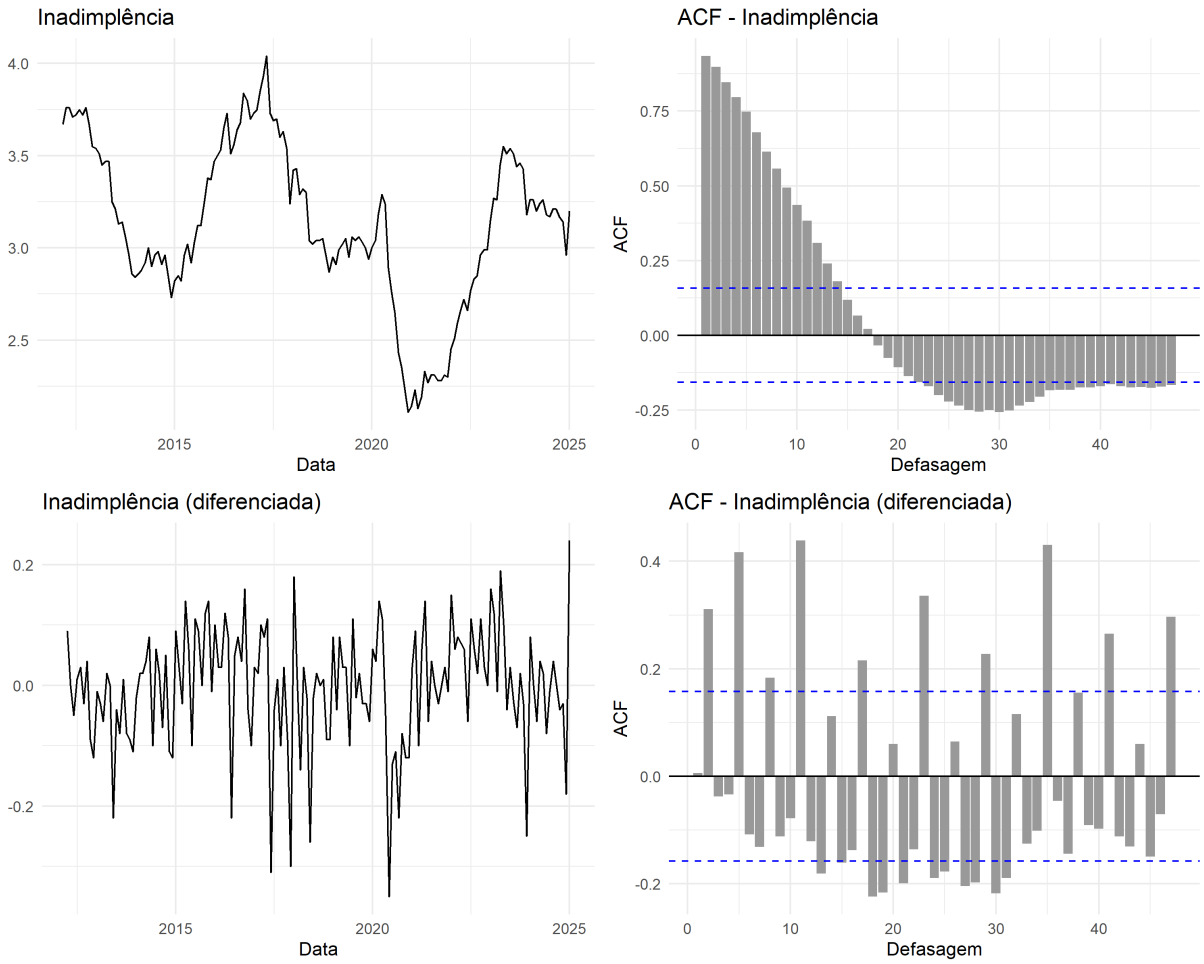
Fonte: Elaboração própria.

Figura 21 – Gráficos - Índice de confiança do consumidor (ICC)



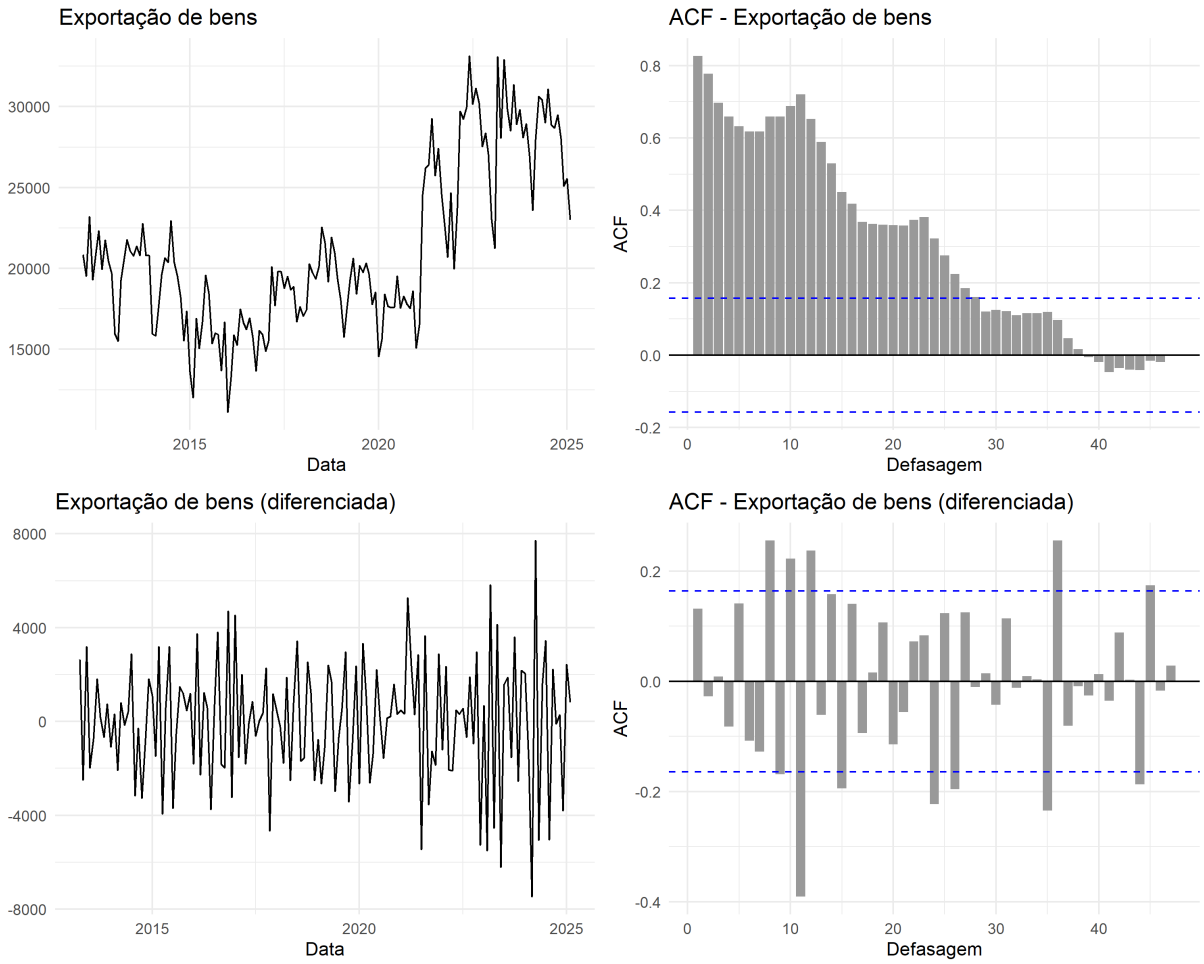
Fonte: Elaboração própria.

Figura 22 – Gráficos - Operações de crédito - inadimplência da carteira de crédito - total



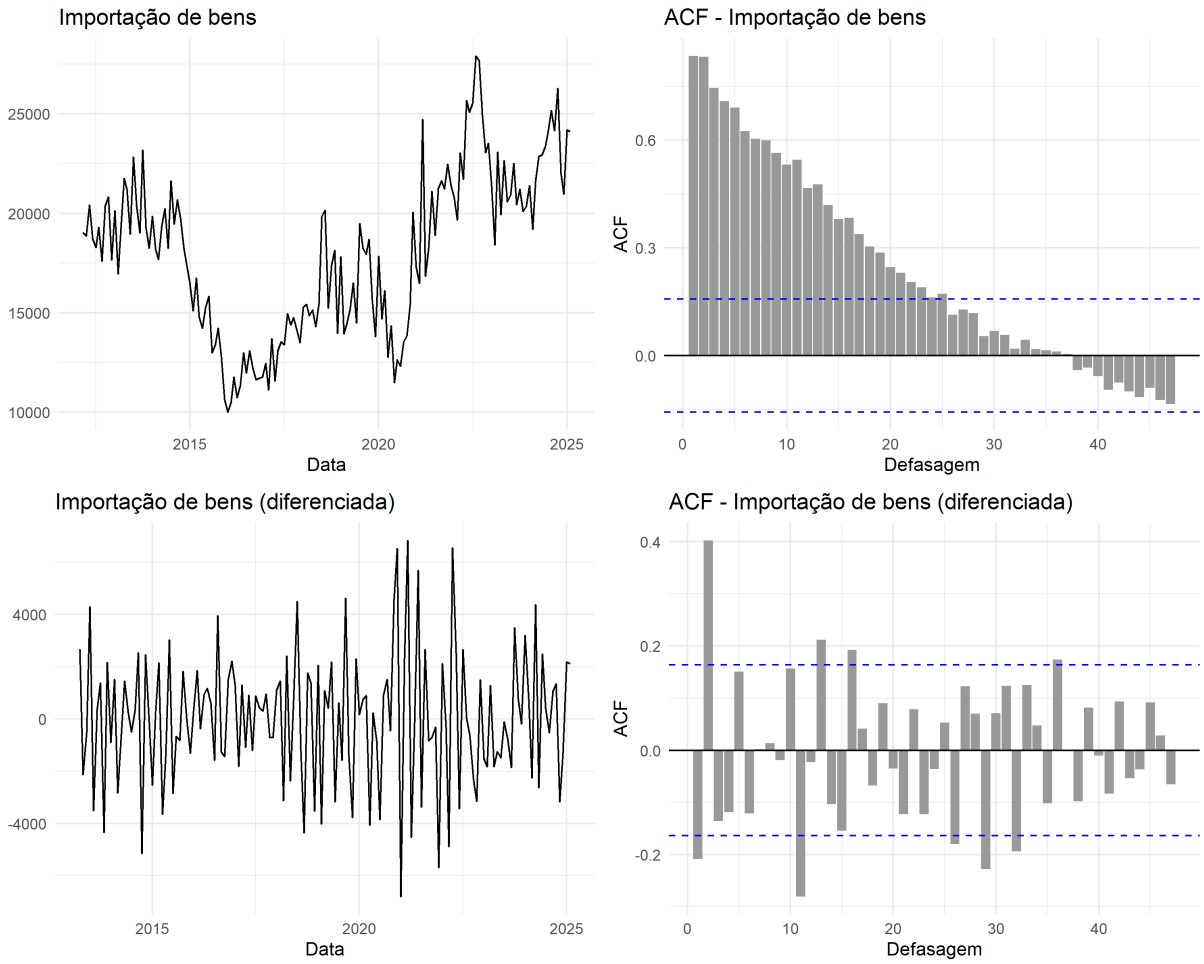
Fonte: Elaboração própria.

Figura 23 – Gráficos - Exportação de bens - Balanço de Pagamentos



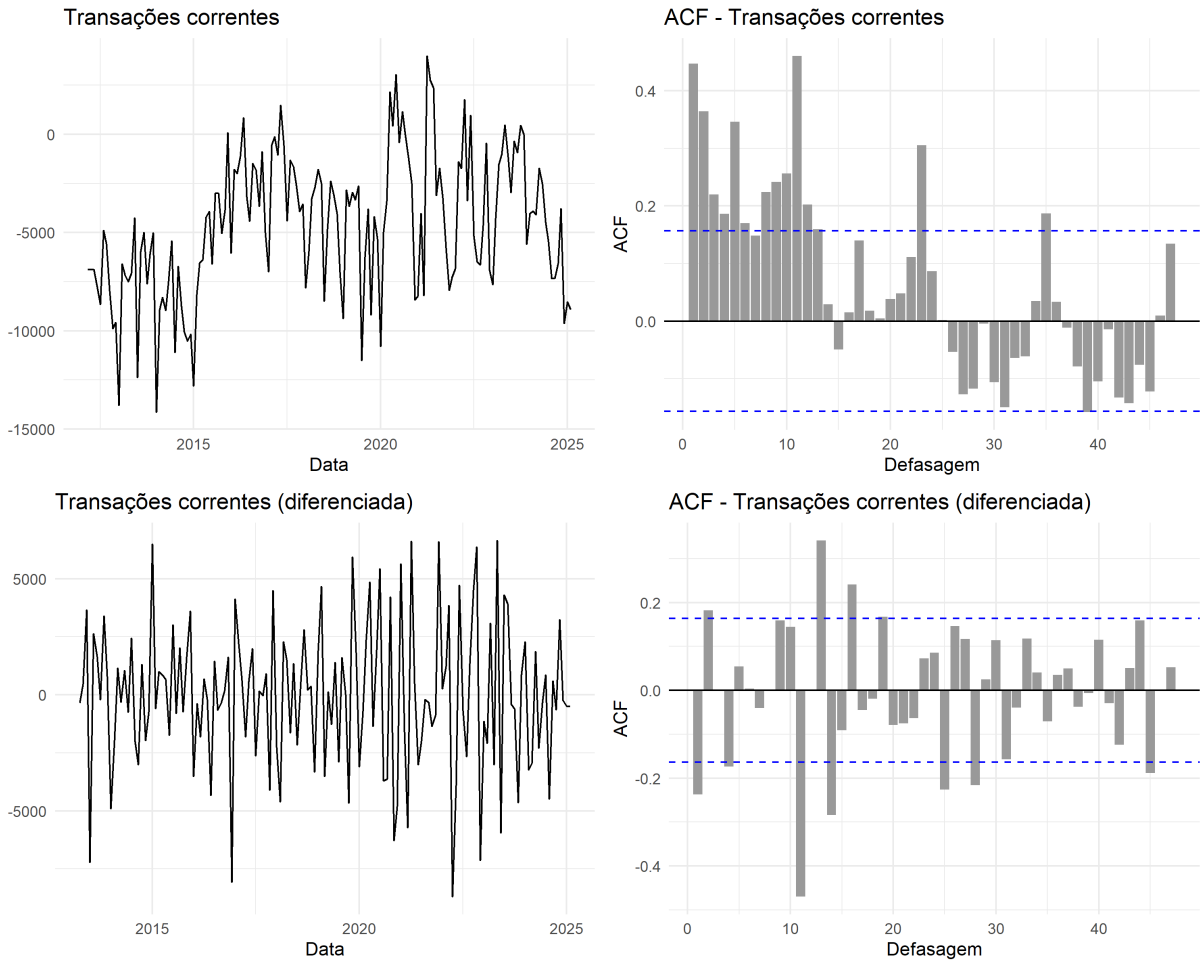
Fonte: Elaboração própria.

Figura 24 – Gráficos - Importação de bens - Balanço de Pagamentos



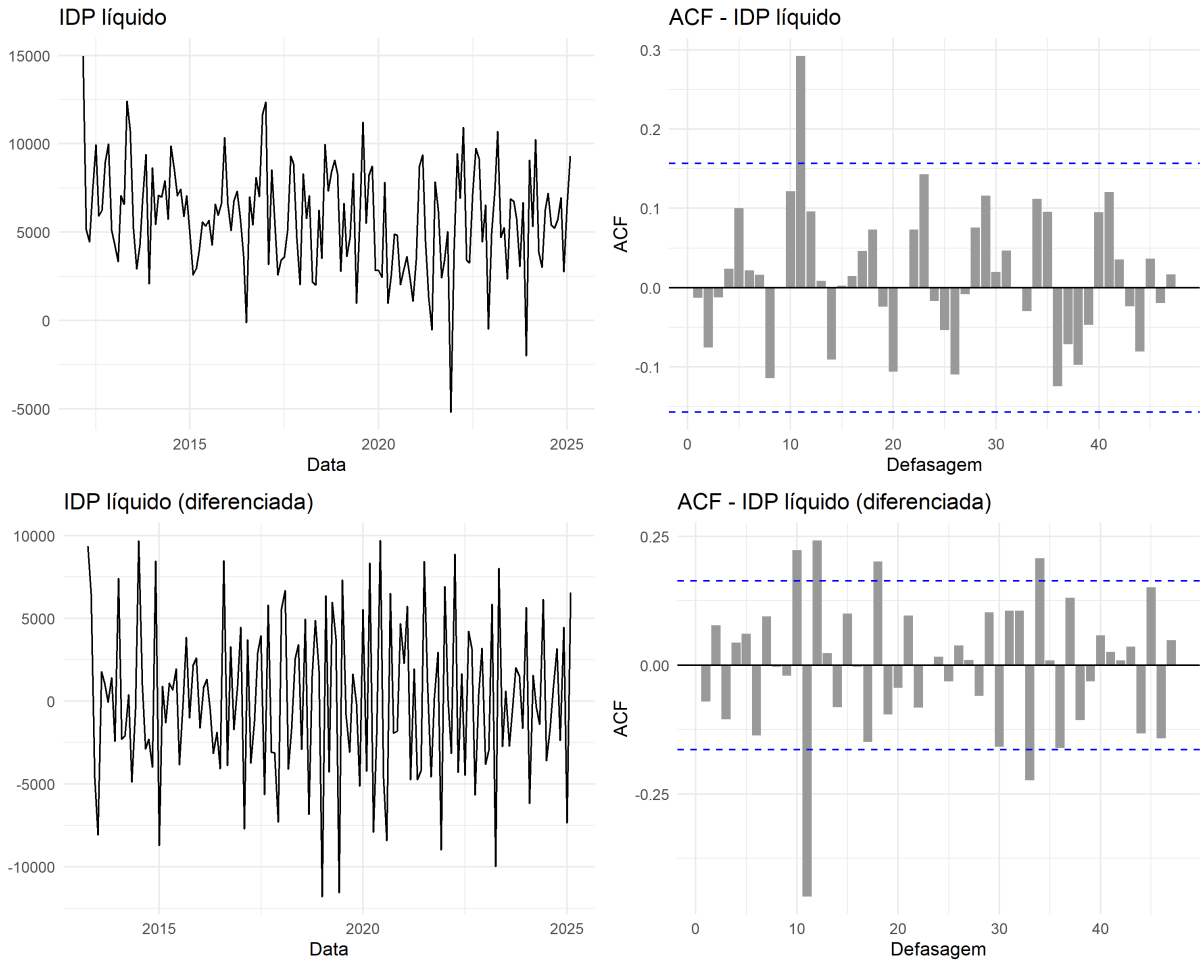
Fonte: Elaboração própria.

Figura 25 – Gráficos - Balanço de pagamentos: transações correntes - saldo



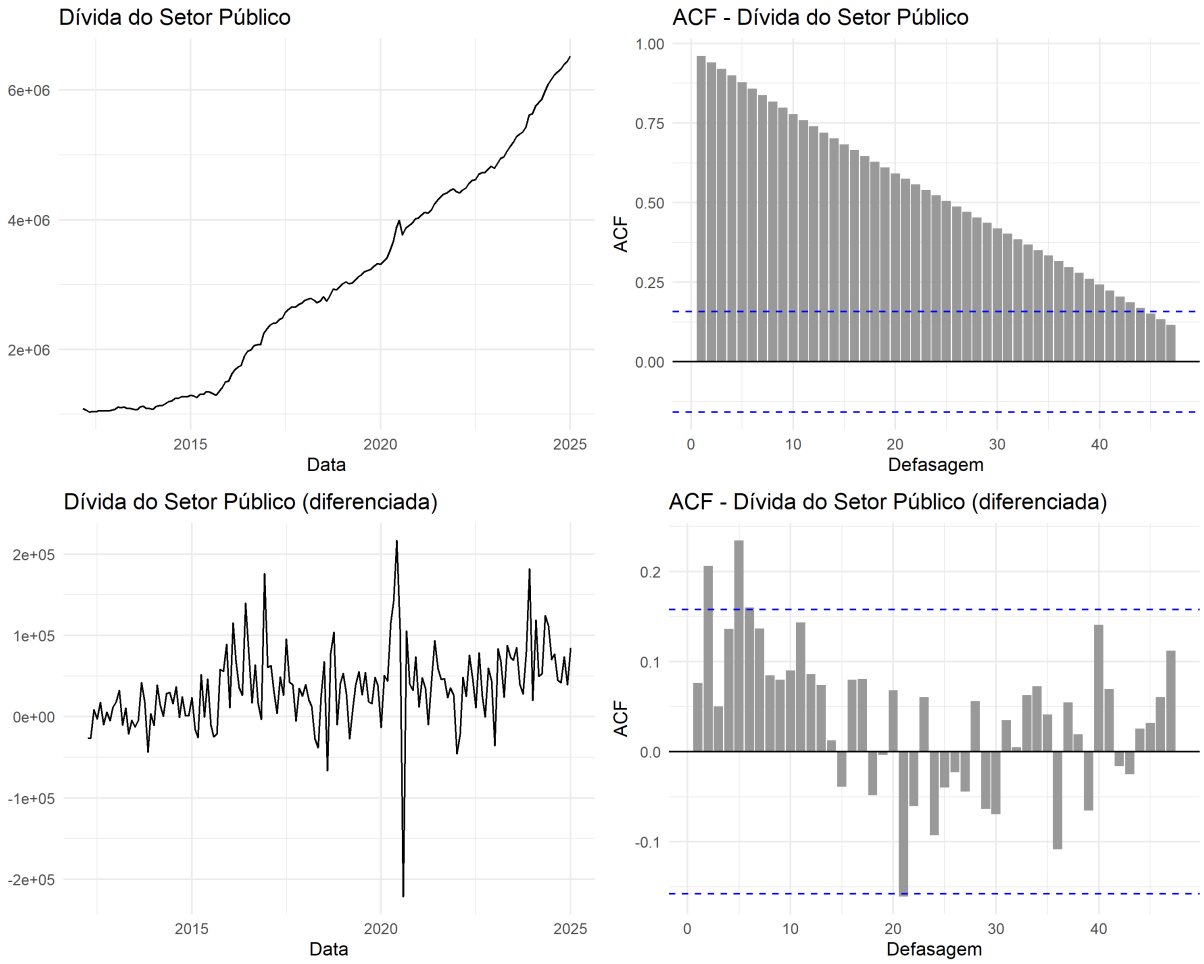
Fonte: Elaboração própria.

Figura 26 – Gráficos - Investimentos diretos no país (IDP) líquido



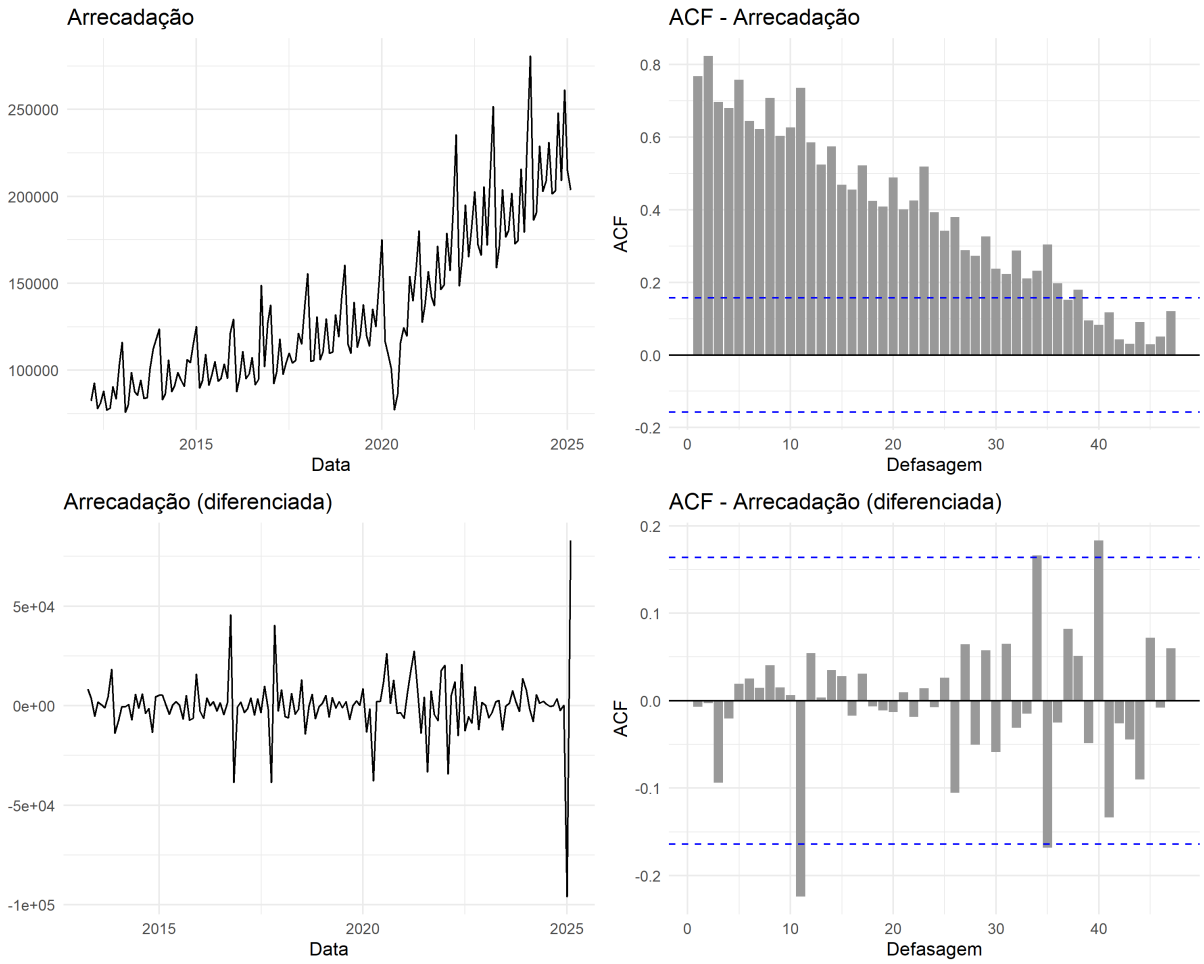
Fonte: Elaboração própria.

Figura 27 – Gráficos - Dívida Líquida do Setor Público - Saldos - Total - Governo Federal



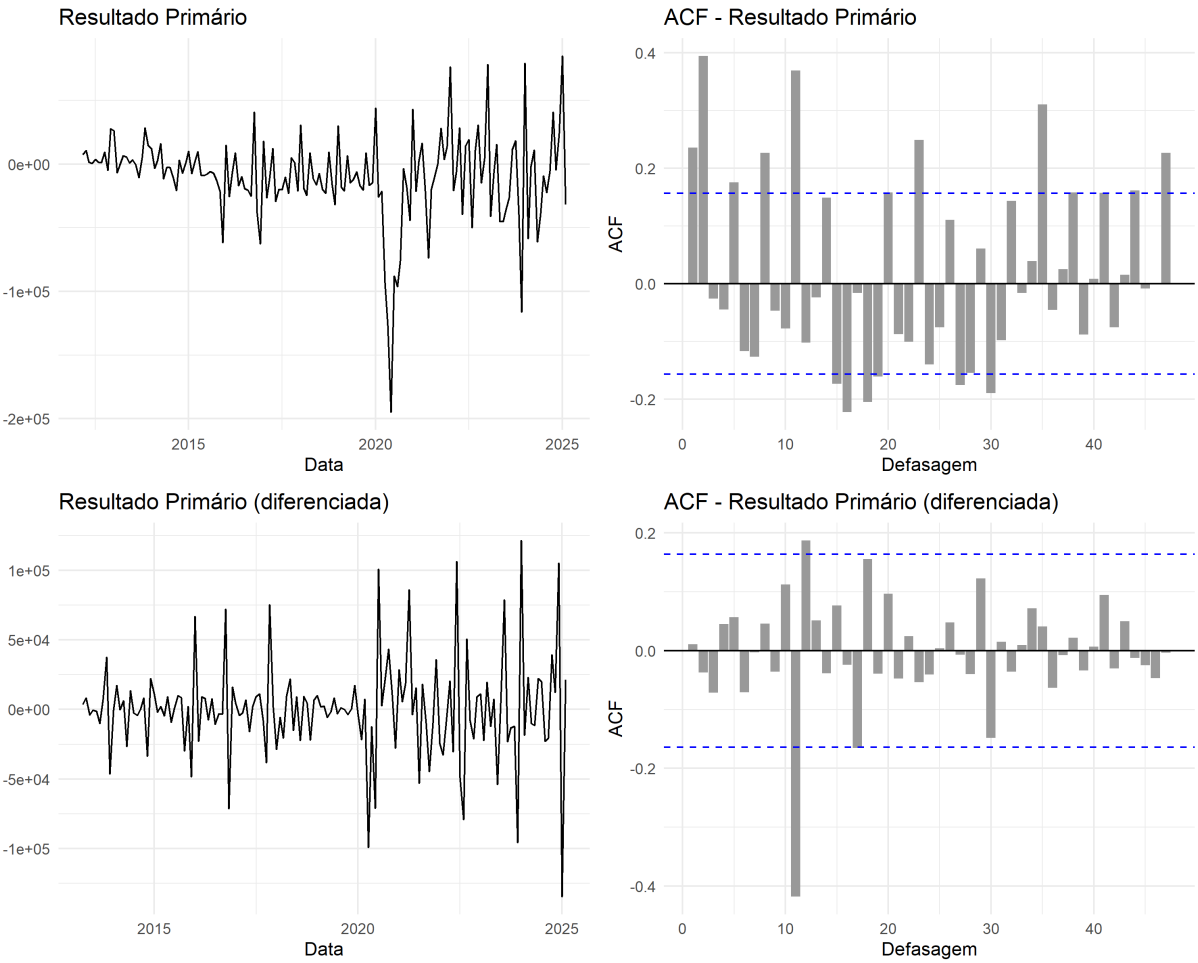
Fonte: Elaboração própria.

Figura 28 – Gráficos - Arrecadação das receitas federais - receita bruta



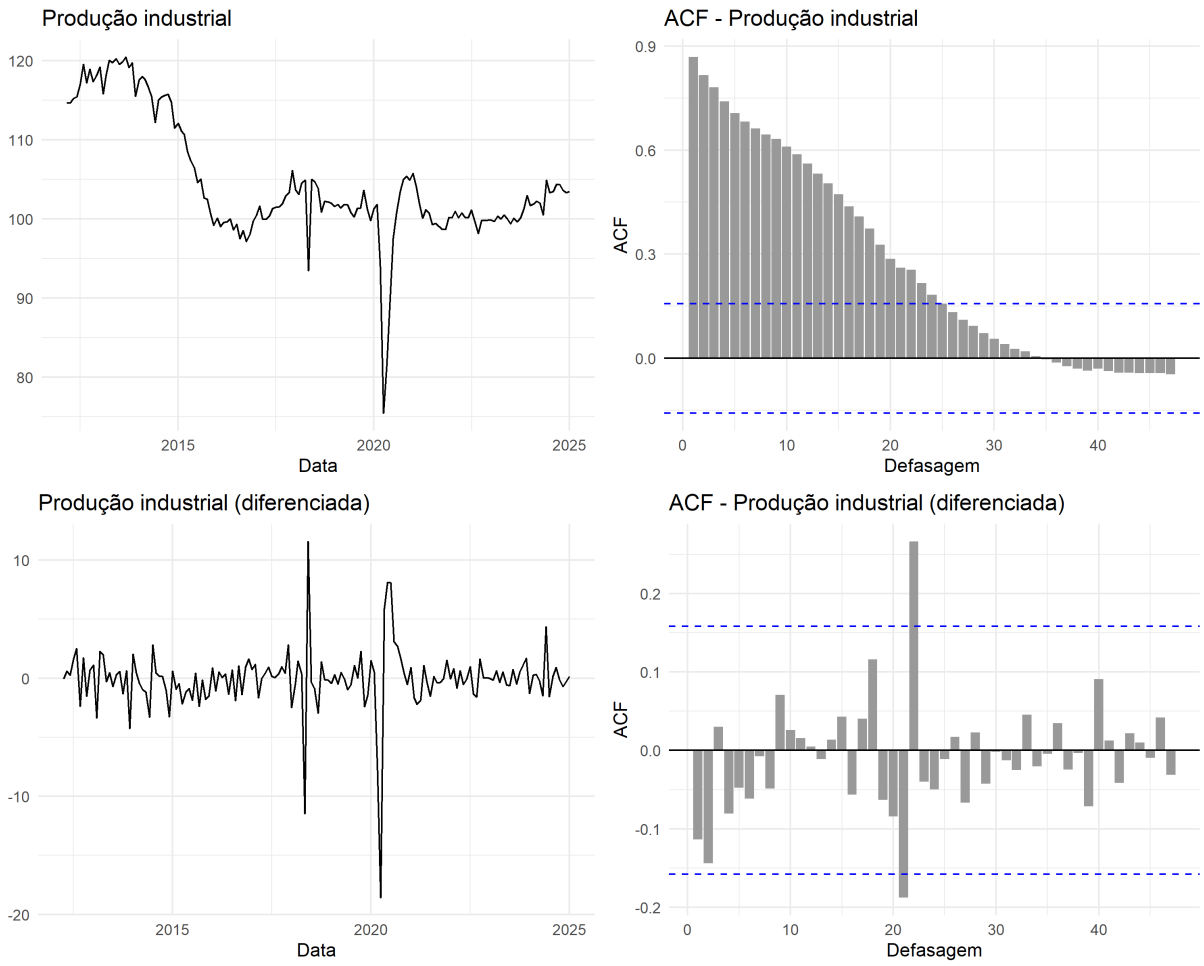
Fonte: Elaboração própria.

Figura 29 – Gráficos - Resultado Primário do Governo Central



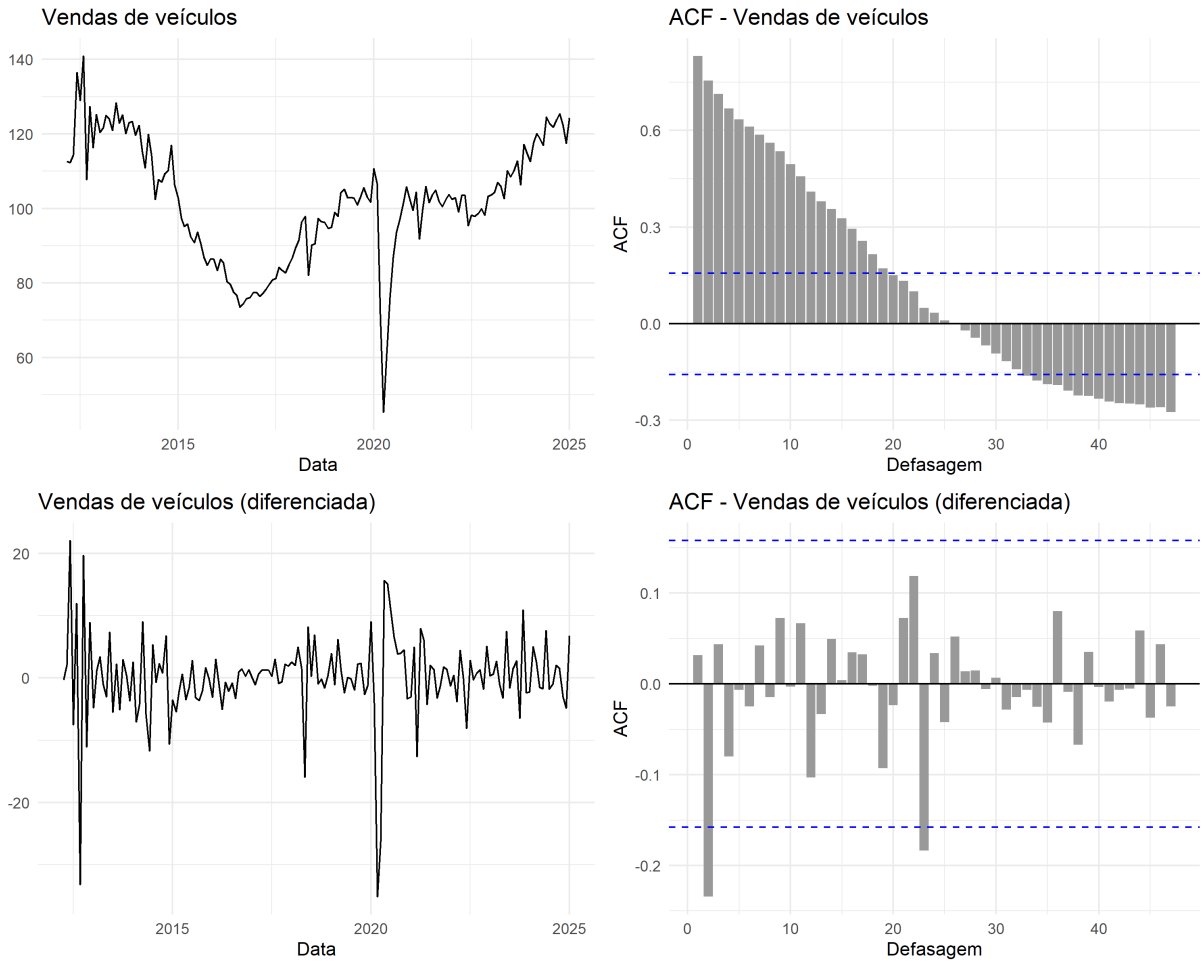
Fonte: Elaboração própria.

Figura 30 – Gráficos - Produção industrial - indústria geral: índice de quantum dessazonalizado



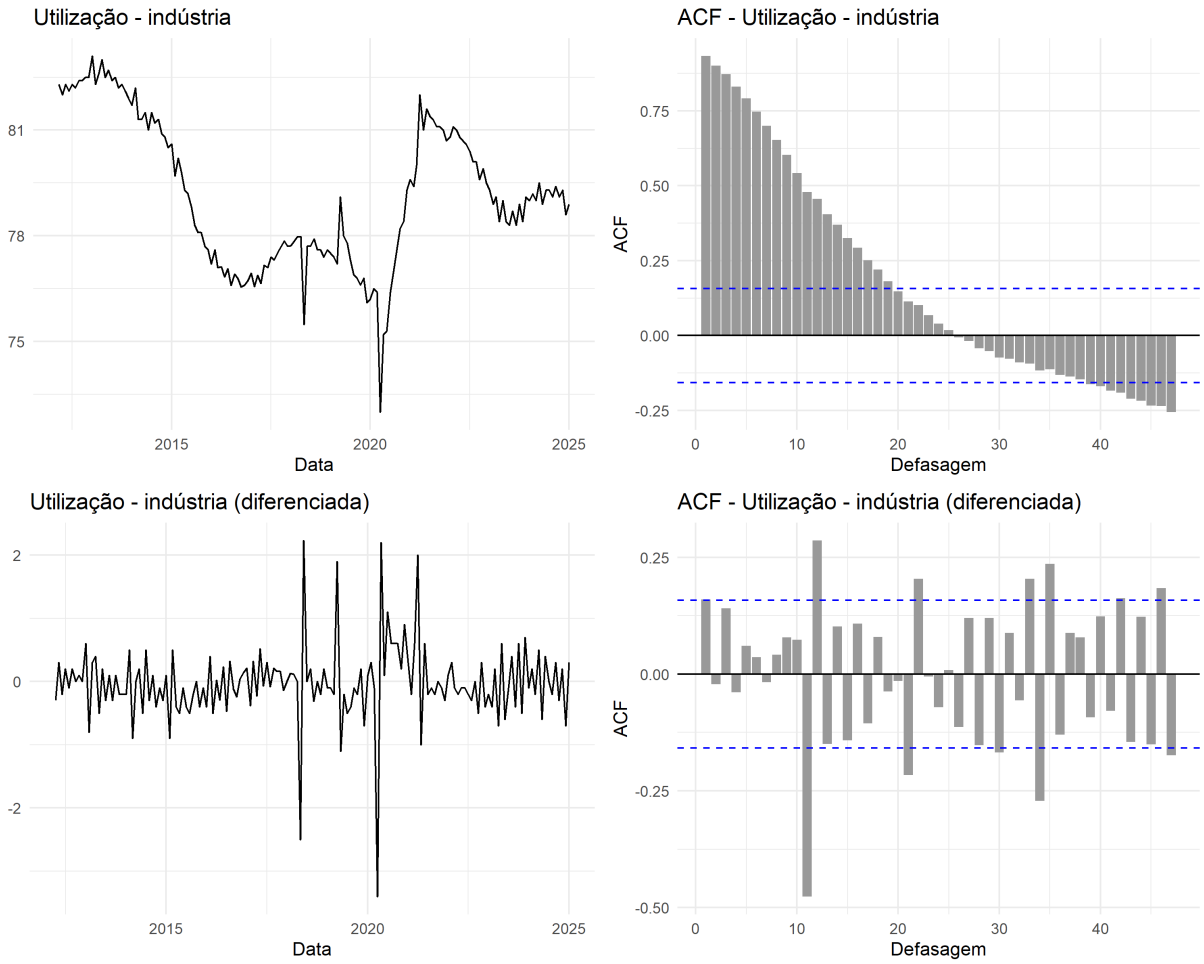
Fonte: Elaboração própria.

Figura 31 – Gráficos - Vendas reais no varejo de veículos, motos, partes e peças



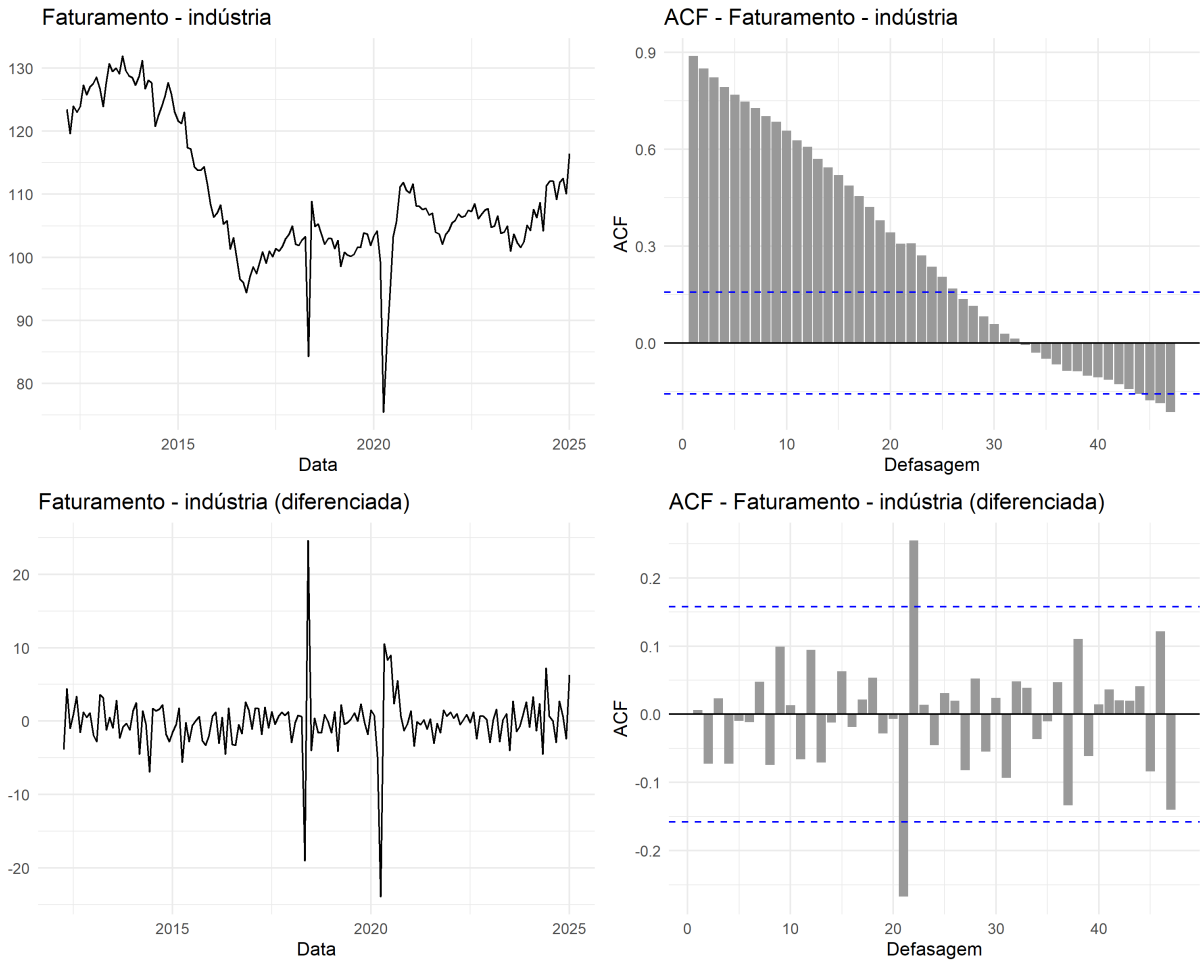
Fonte: Elaboração própria.

Figura 32 – Gráficos - Utilização da capacidade instalada - indústria - índice dessazonalizado



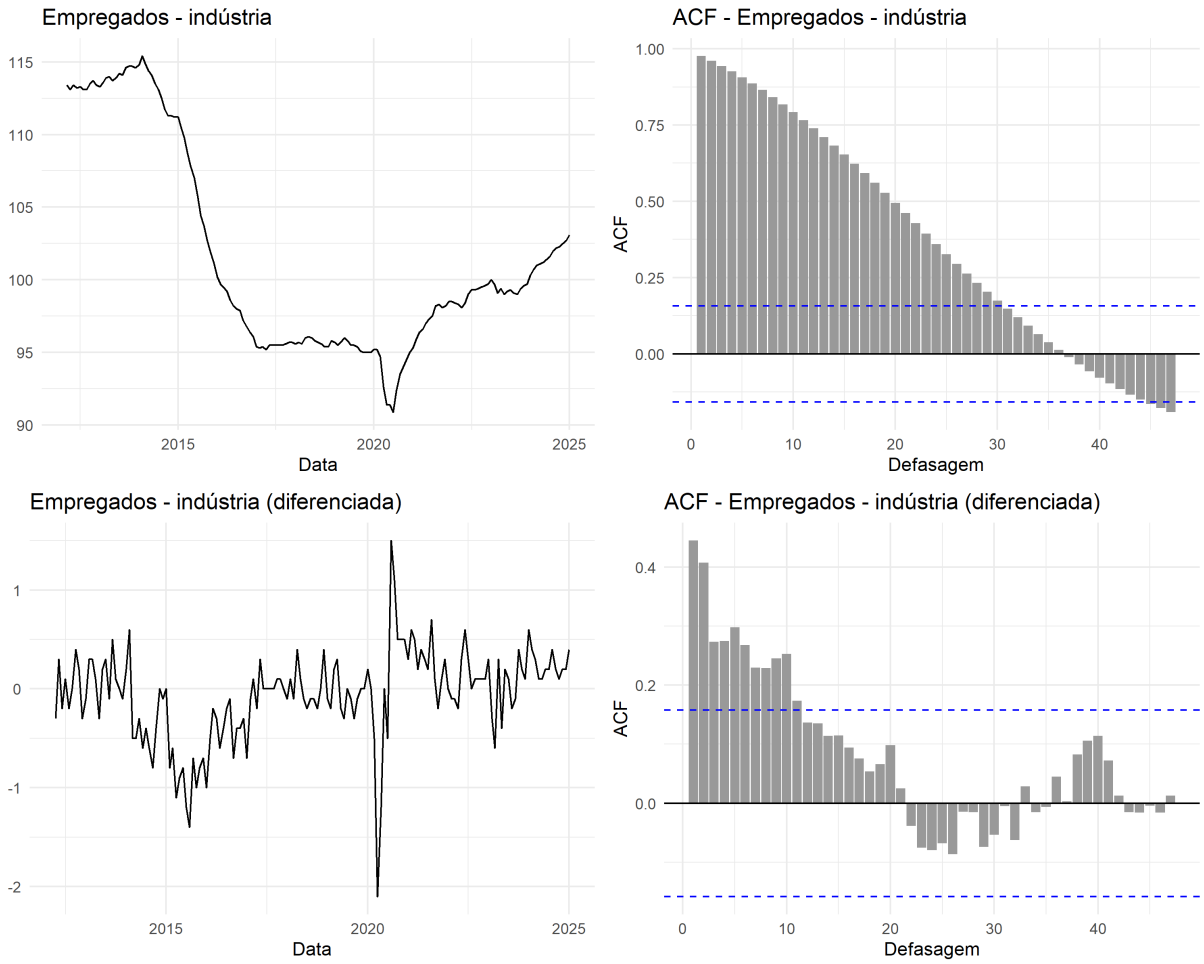
Fonte: Elaboração própria.

Figura 33 – Gráficos - Faturamento real - indústria - índice dessazonalizado



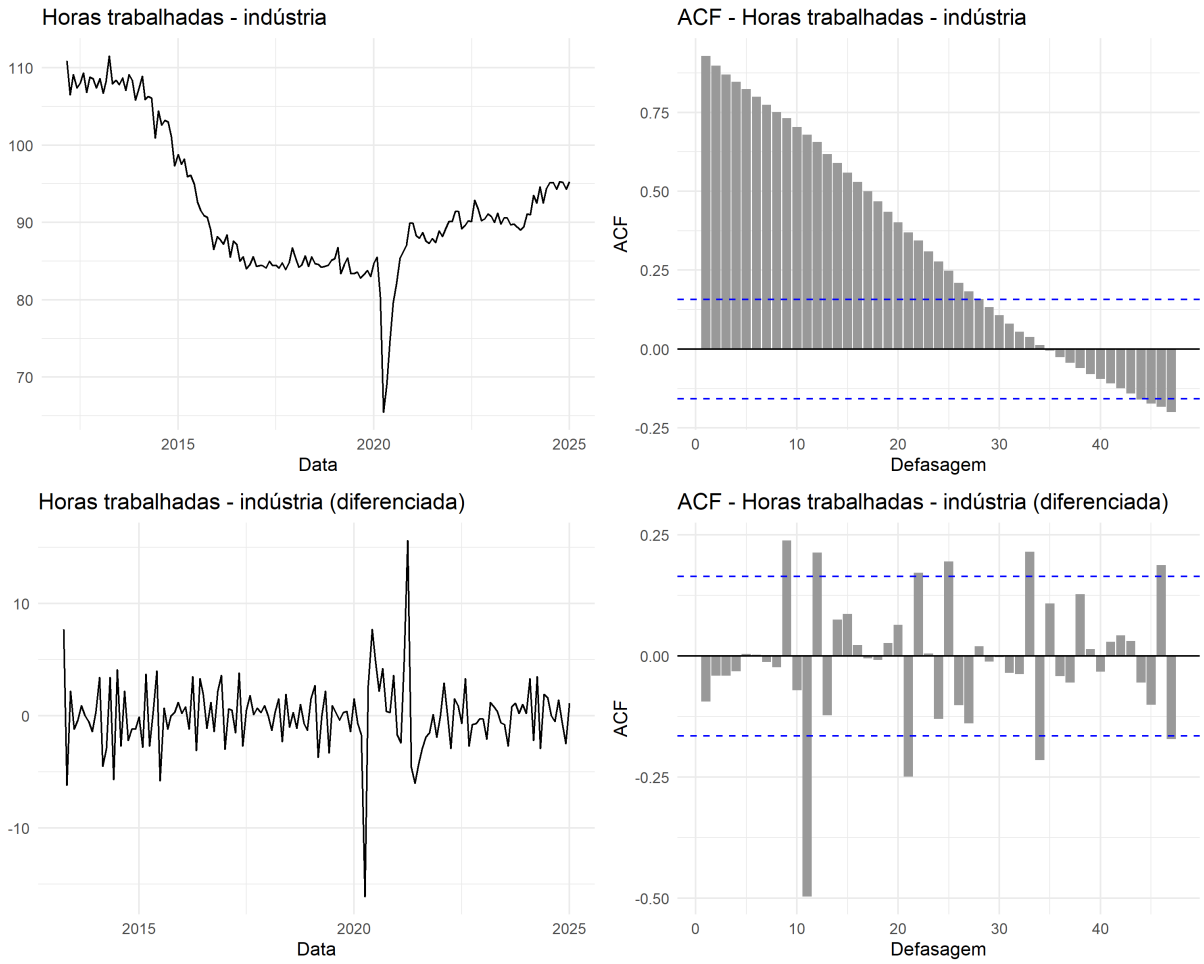
Fonte: Elaboração própria.

Figura 34 – Gráficos - Pessoal empregado - indústria - índice dessazonalizado



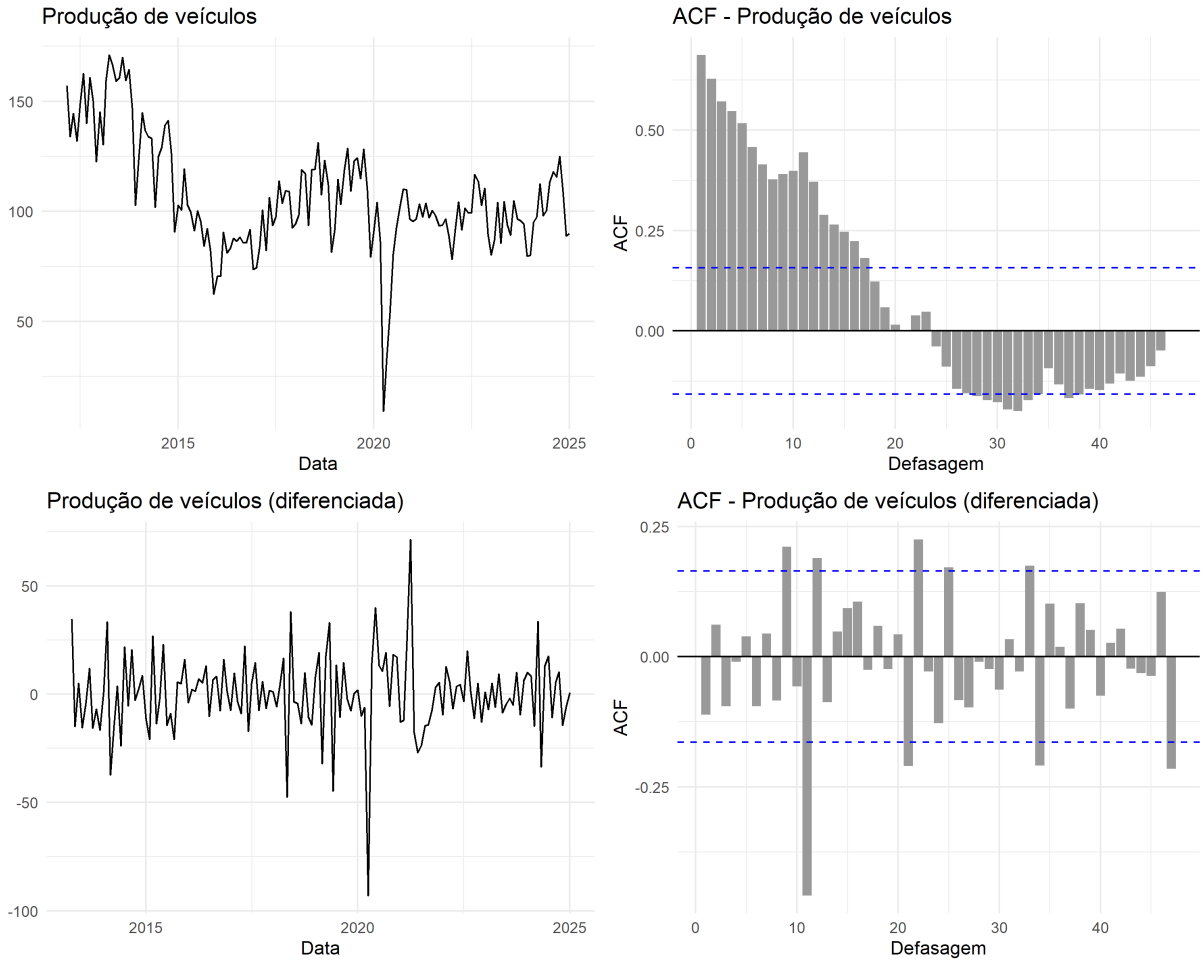
Fonte: Elaboração própria.

Figura 35 – Gráficos - Horas trabalhadas - indústria - índice dessazonalizado



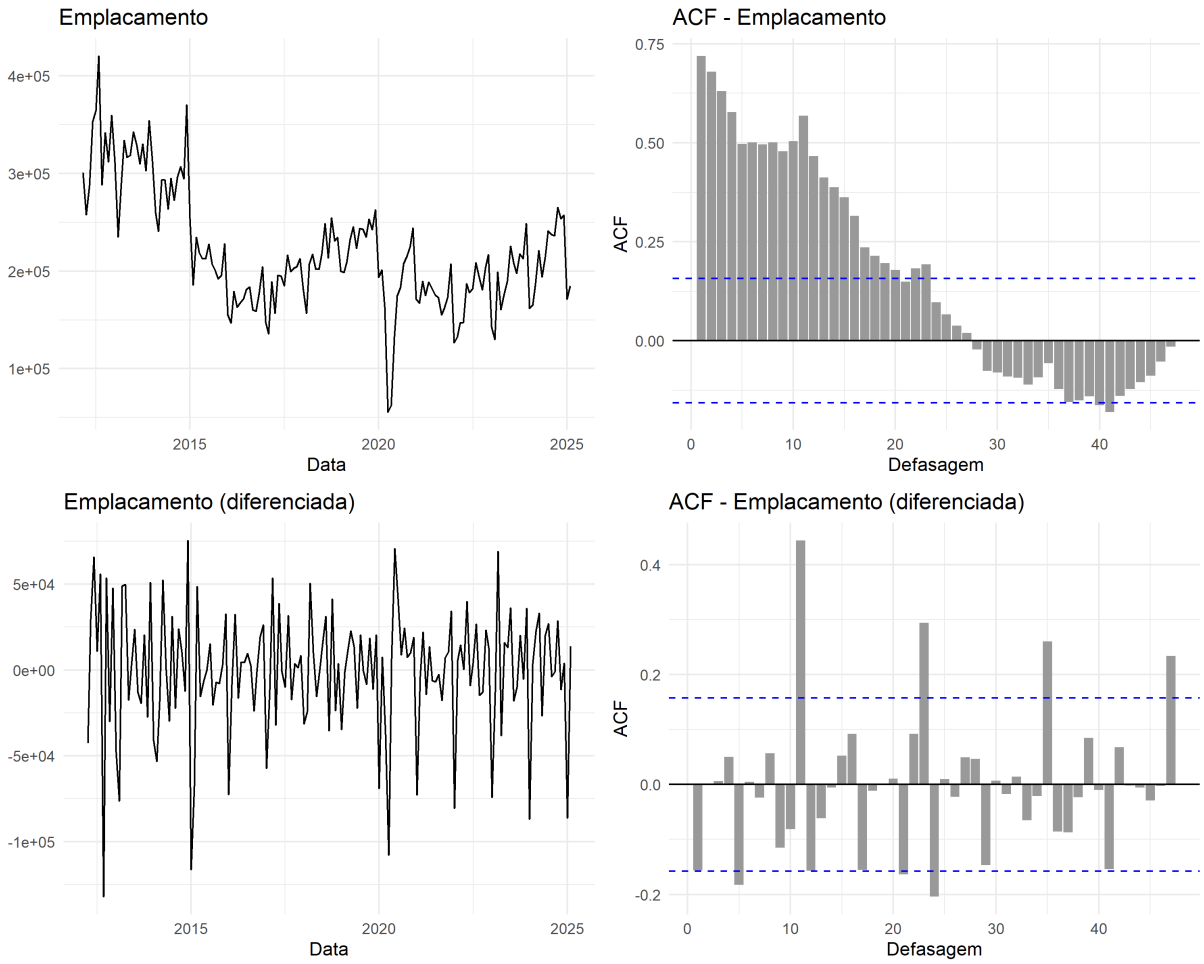
Fonte: Elaboração própria.

Figura 36 – Gráficos - Produção industrial - veículos automotores, reboques e carrocerias - quantum - índice



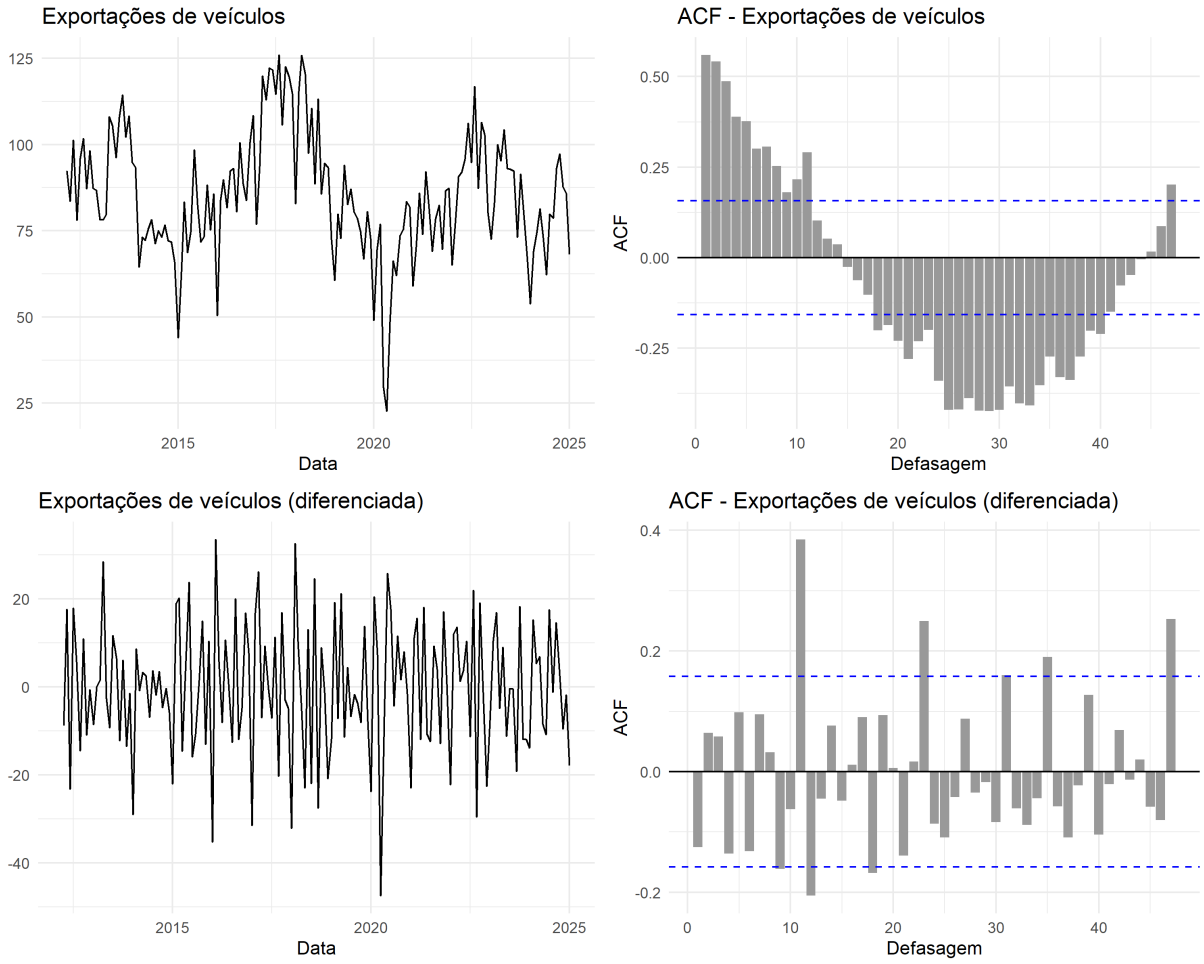
Fonte: Elaboração própria.

Figura 37 – Gráficos - Emplacamento de autoveículos



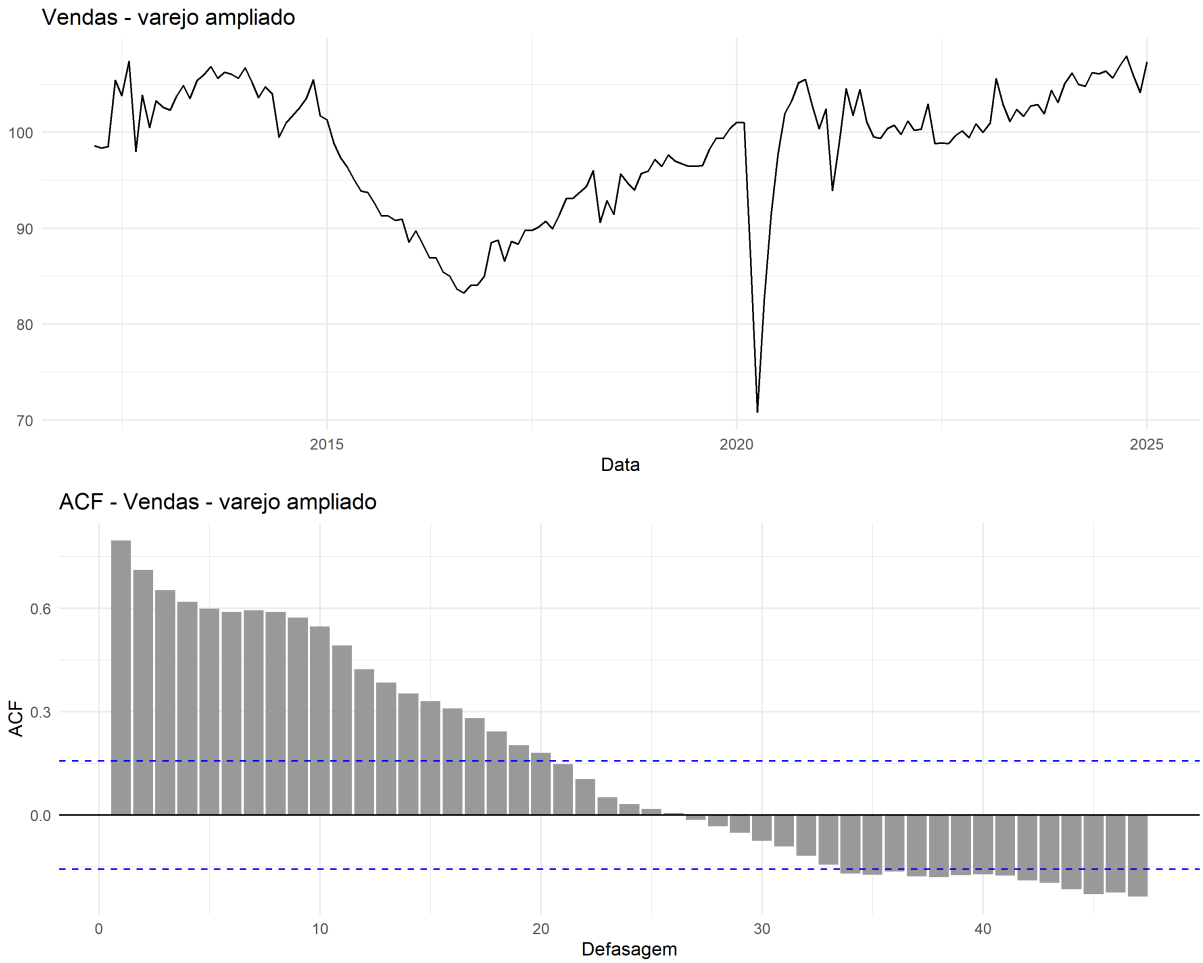
Fonte: Elaboração própria.

Figura 38 – Gráficos - Exportações - veículos automotores, reboques, carrocerias - quantum - índice



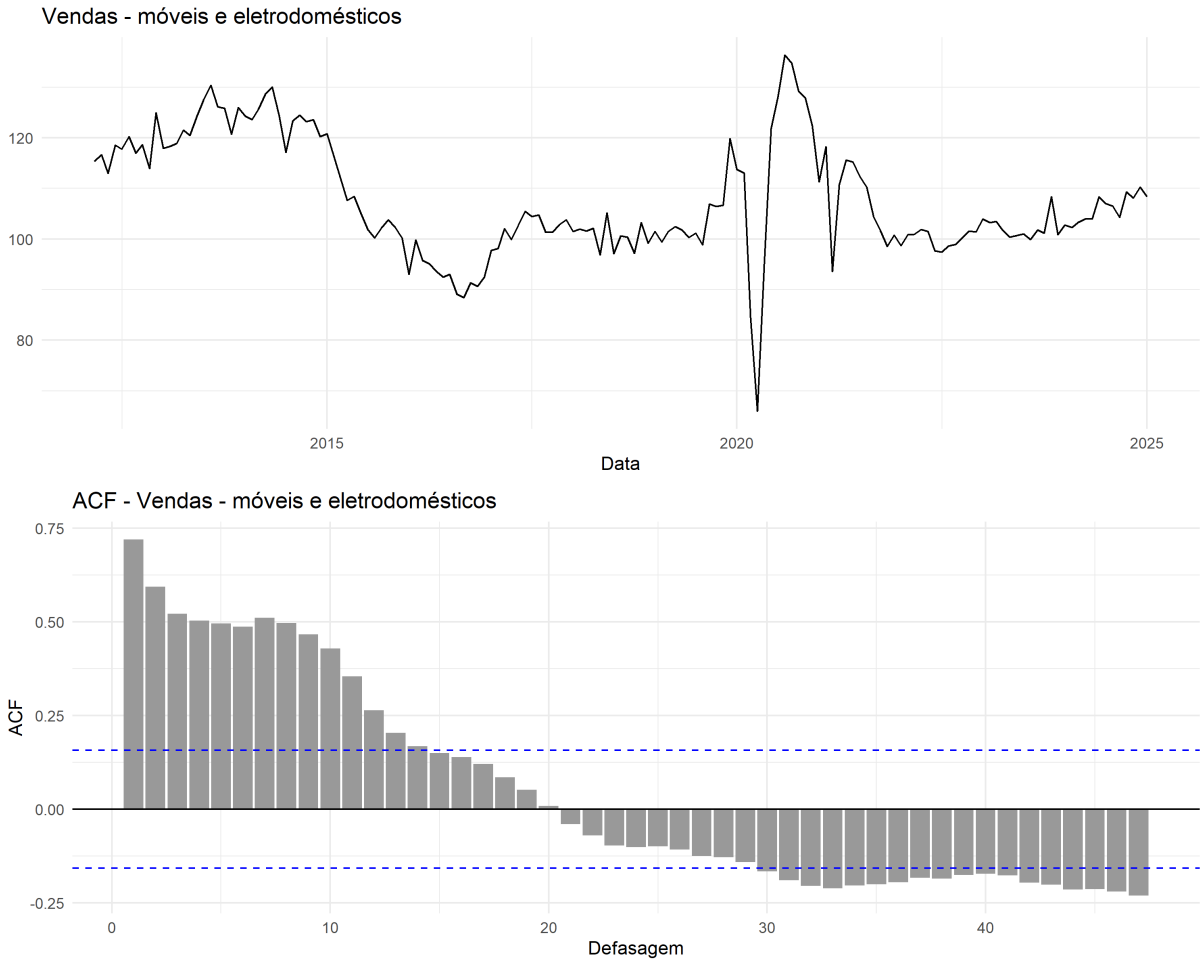
Fonte: Elaboração própria.

Figura 39 – Gráficos - Vendas reais no varejo ampliado - índice dessazonalizado



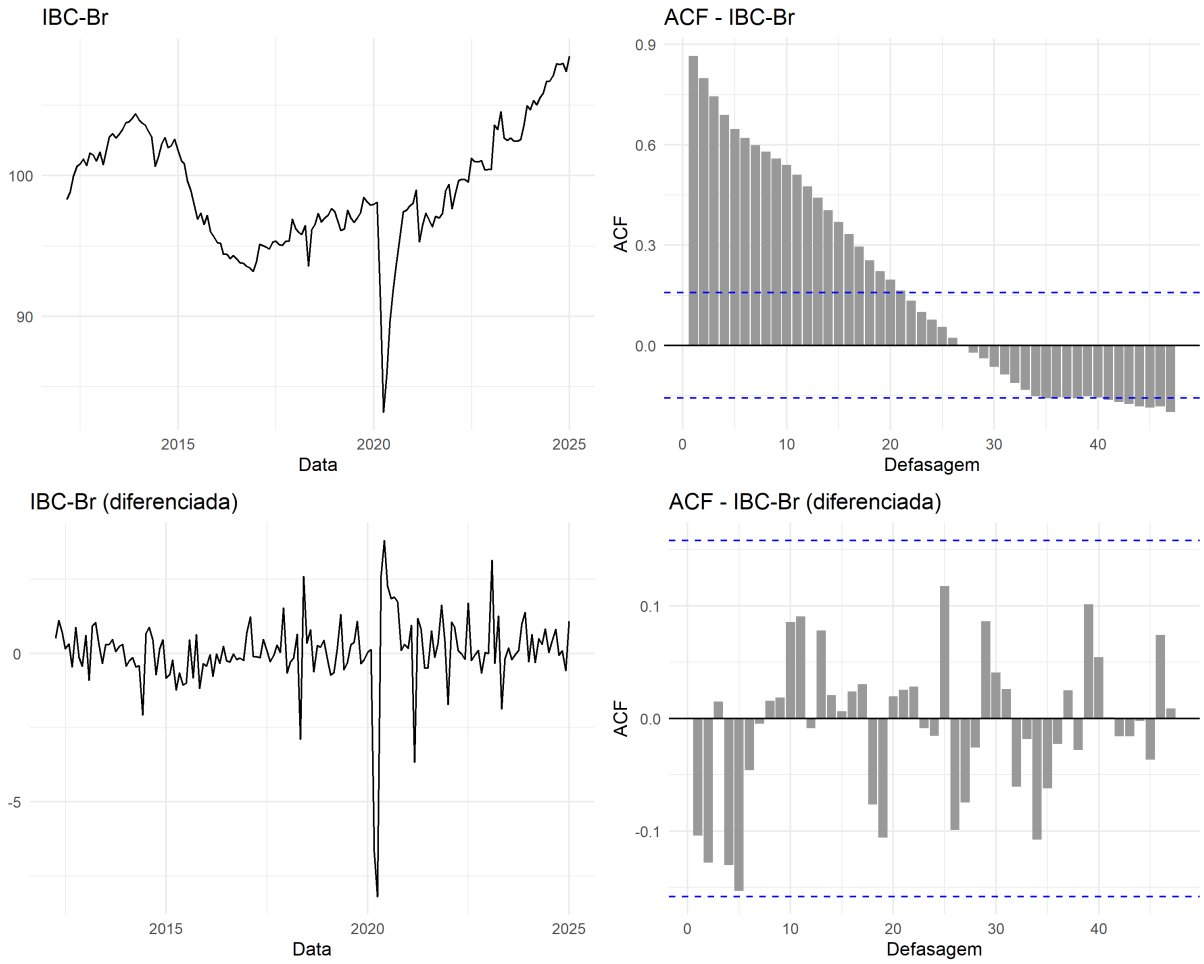
Fonte: Elaboração própria.

Figura 40 – Gráficos - Vendas reais - varejo - móveis e eletrodomésticos - índice dessazonalizado



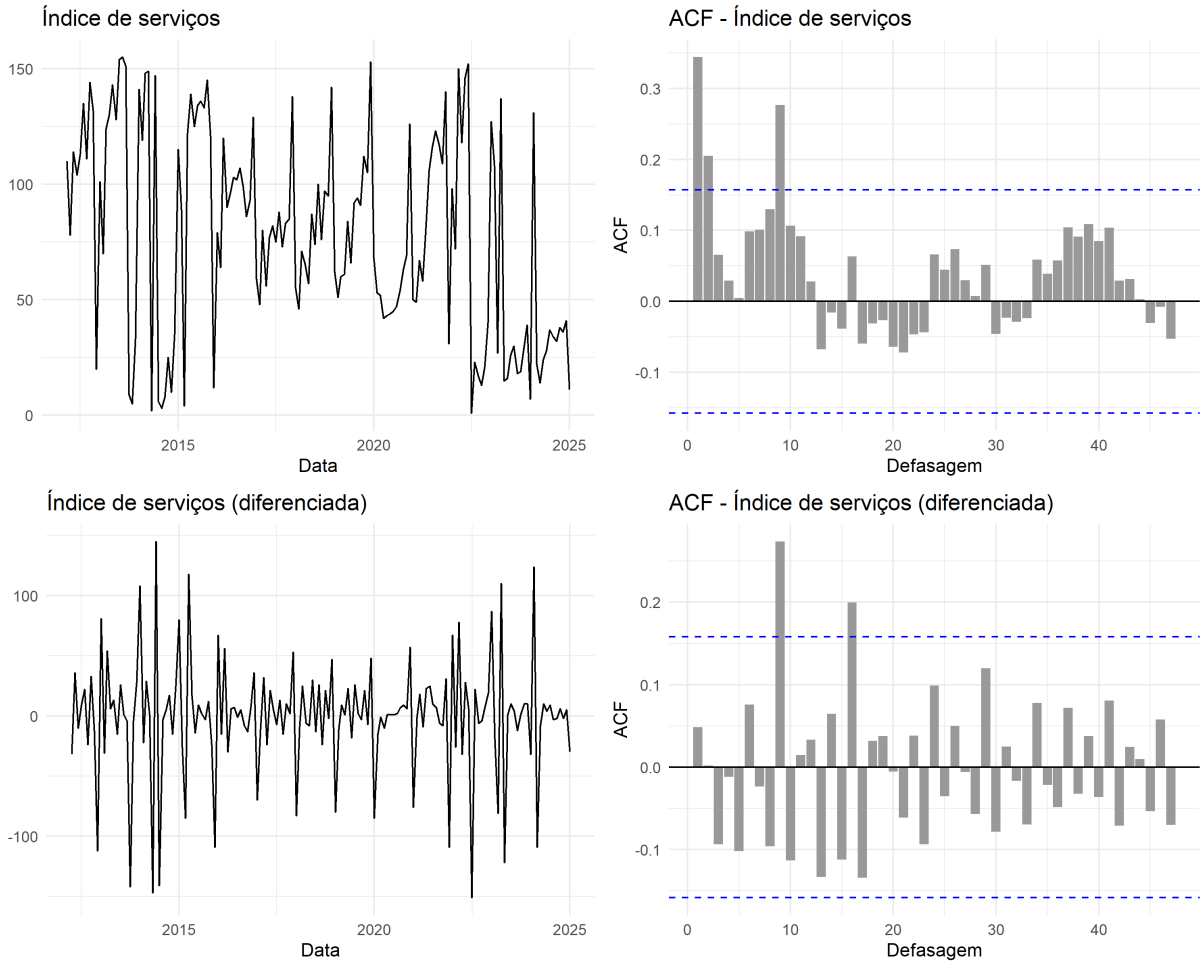
Fonte: Elaboração própria.

Figura 41 – Gráficos - IBC-Br - índice real dessazonalizado



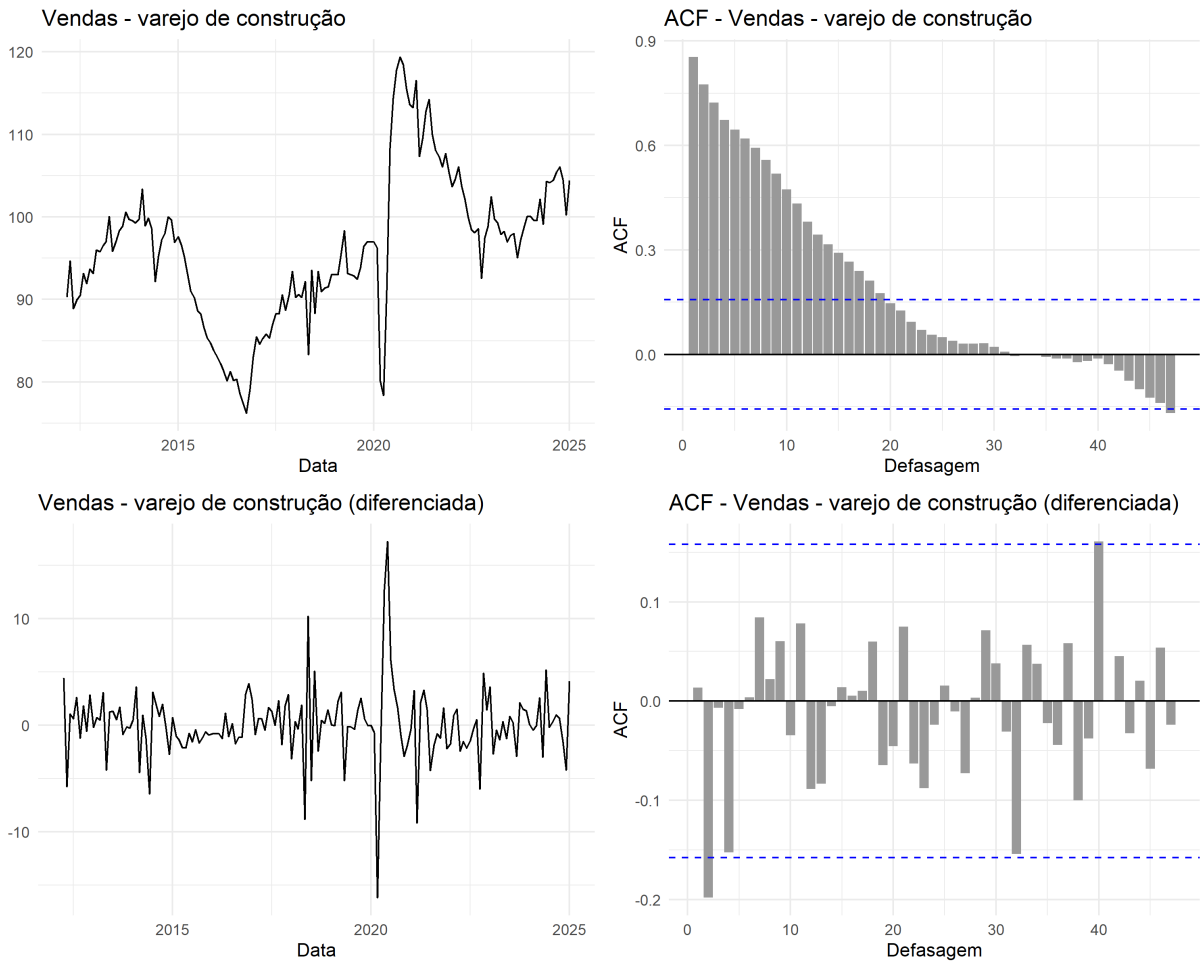
Fonte: Elaboração própria.

Figura 42 – Gráficos - Índice de volume de serviços - total



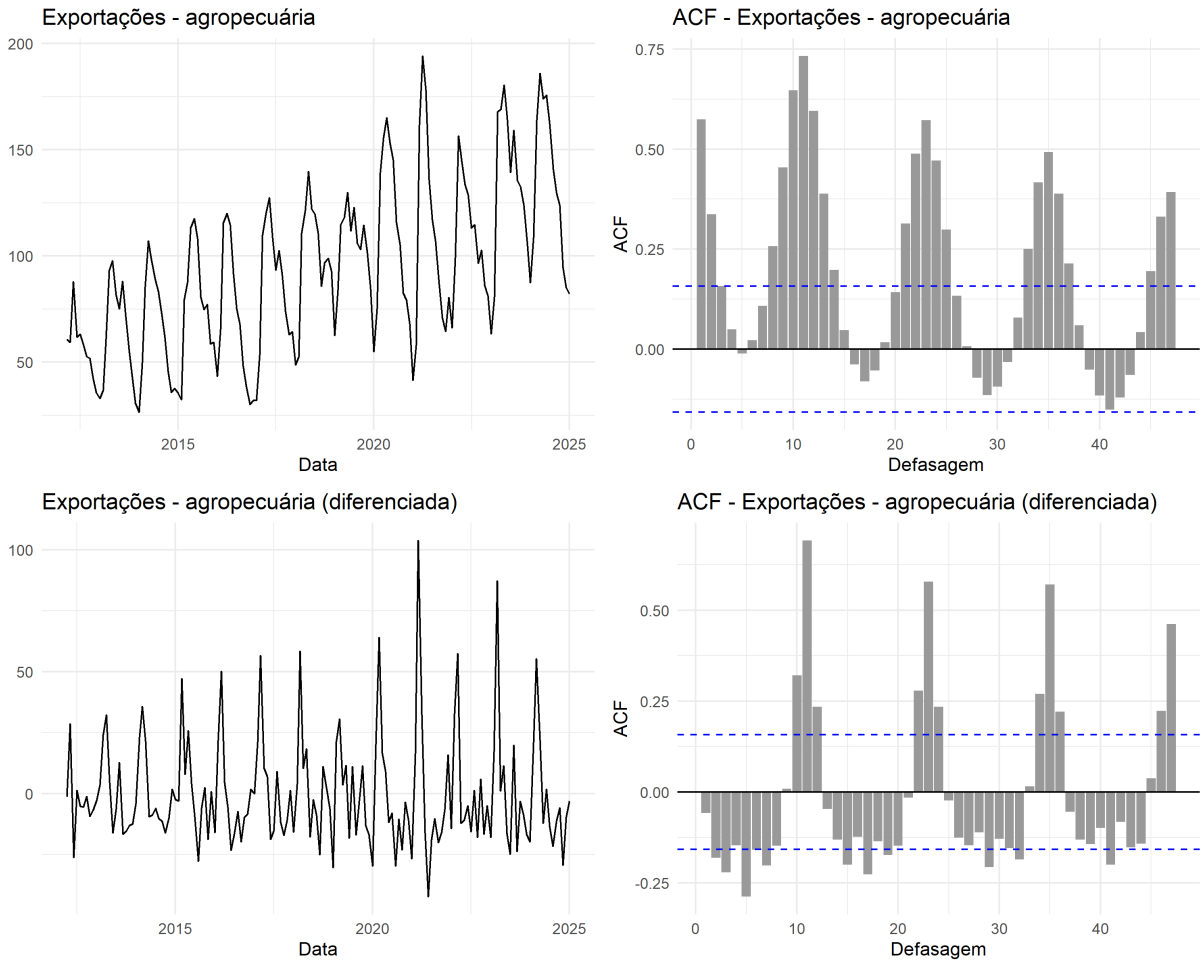
Fonte: Elaboração própria.

Figura 43 – Gráficos - Vendas reais no varejo de materiais de construção: índice dessazonalizado



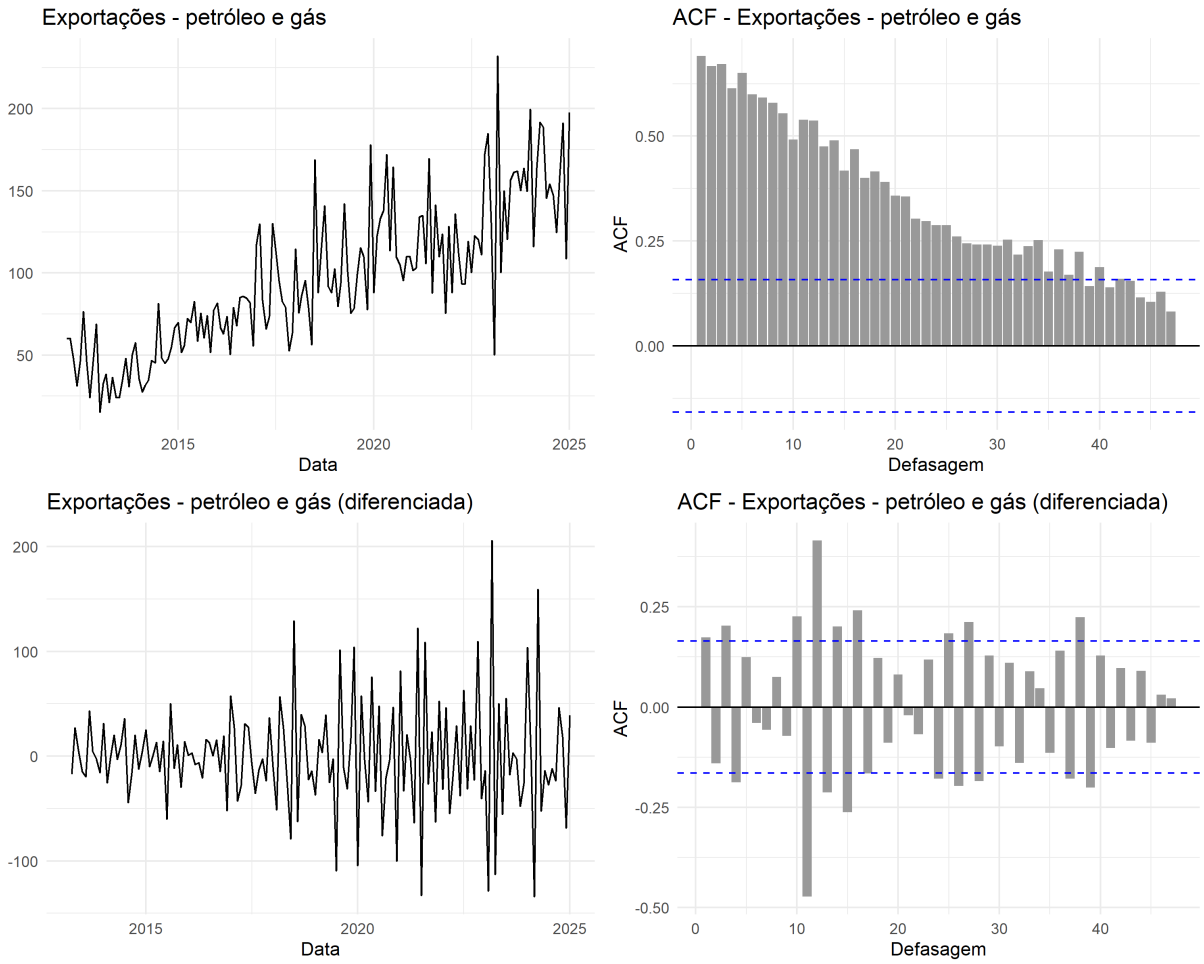
Fonte: Elaboração própria.

Figura 44 – Gráficos - Exportações - agricultura e pecuária - quantum: índice



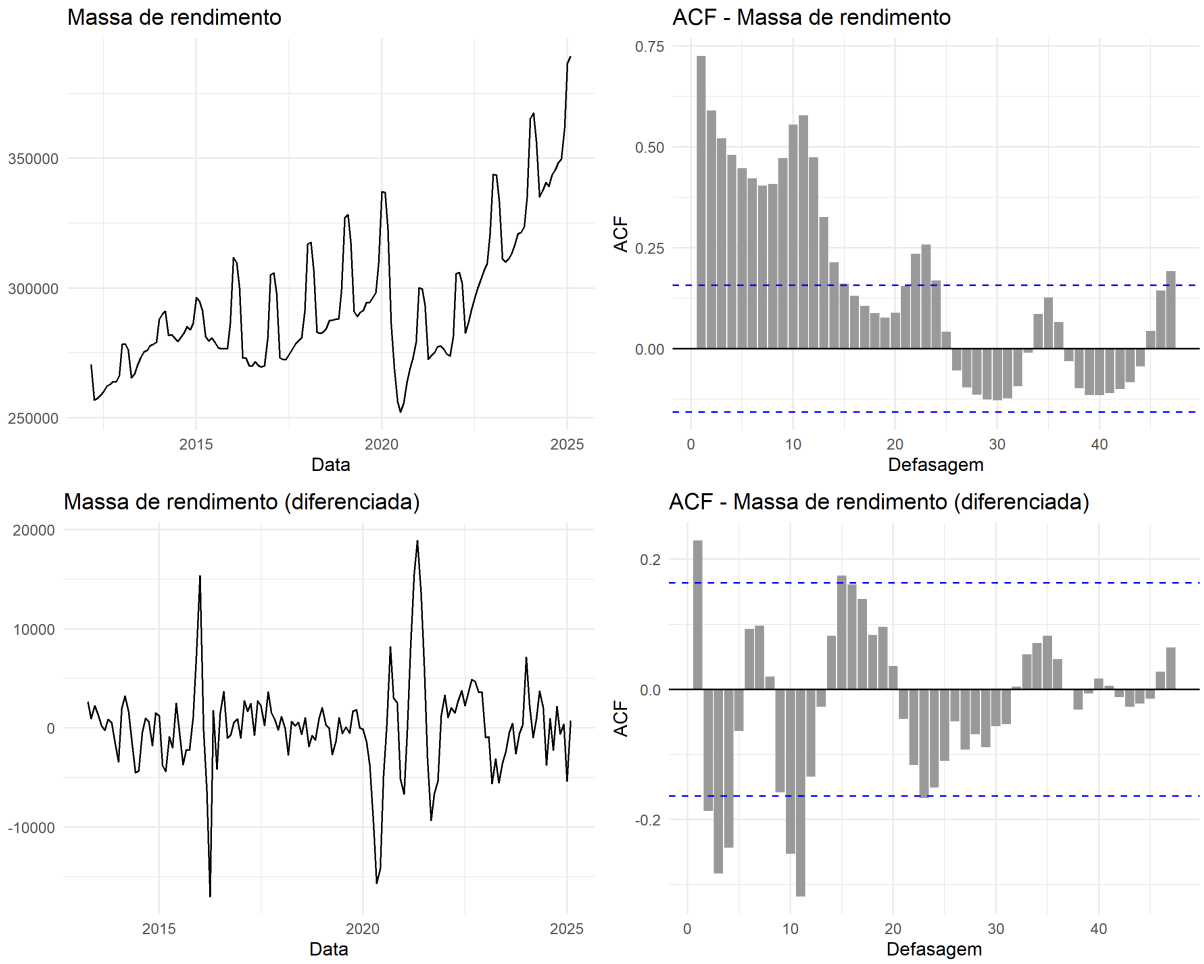
Fonte: Elaboração própria.

Figura 45 – Gráficos - Exportações - extração de petróleo e gás natural - quantum: índice



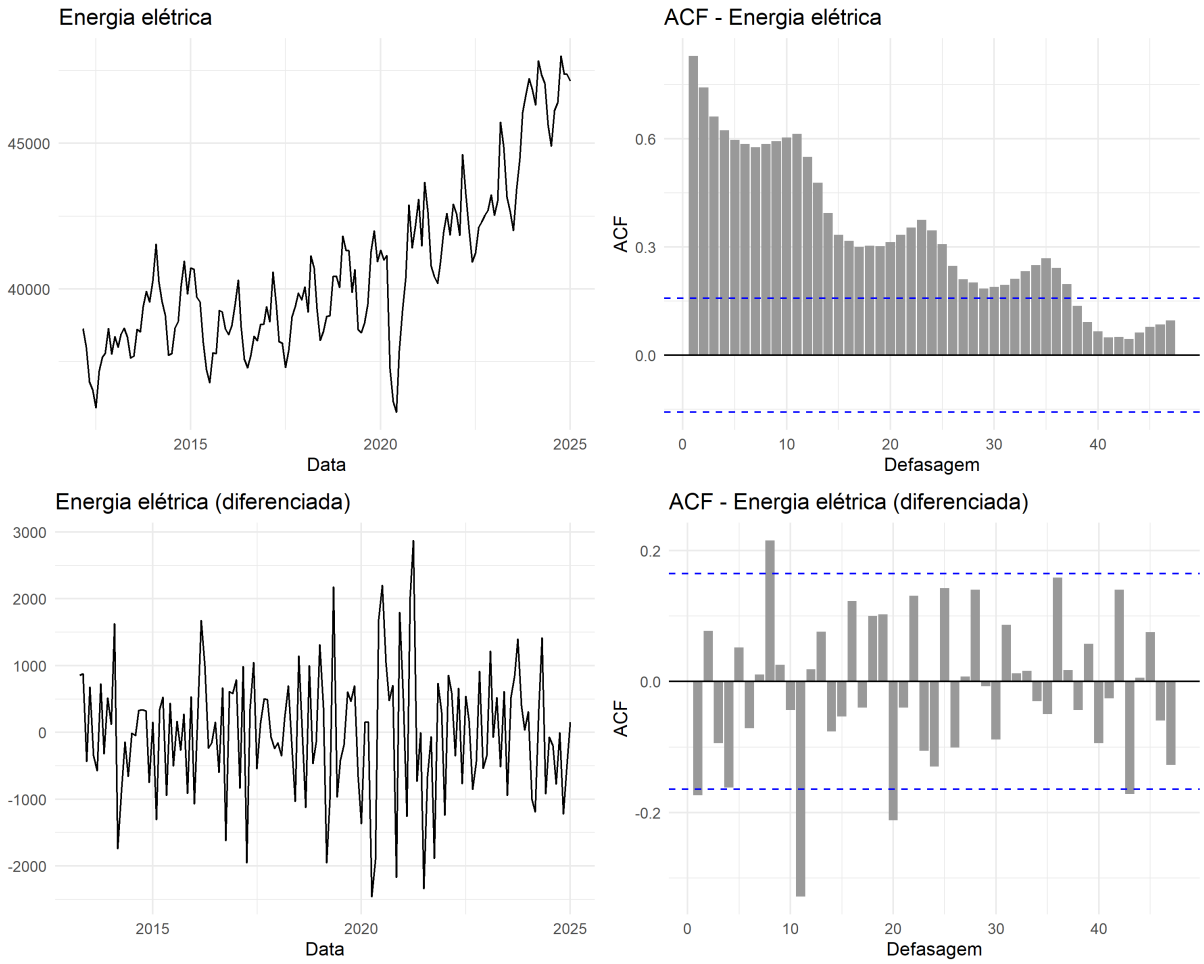
Fonte: Elaboração própria.

Figura 46 – Gráficos - Massa de rendimento real de todos os trabalhos



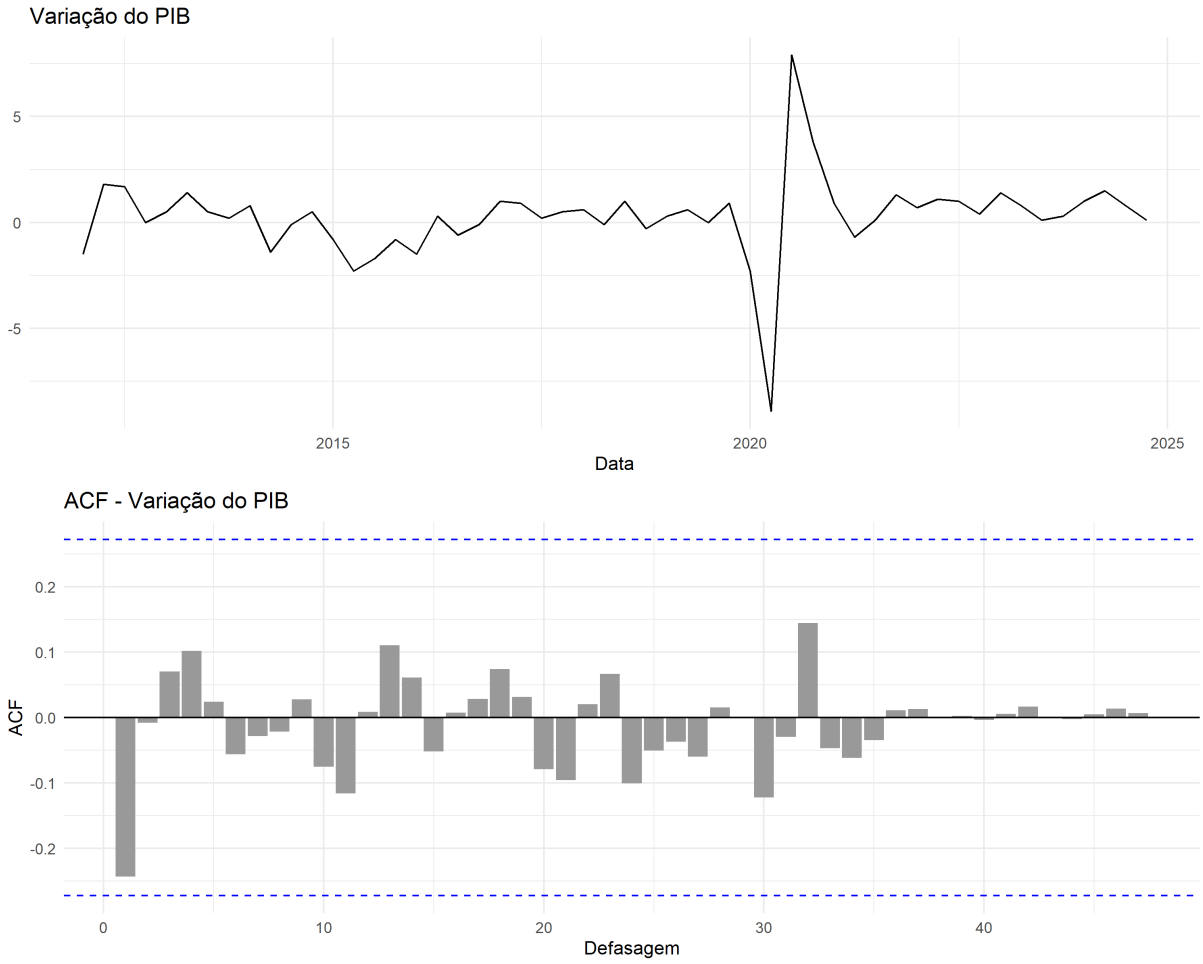
Fonte: Elaboração própria.

Figura 47 – Gráficos - Energia elétrica referente ao consumo - quantidade



Fonte: Elaboração própria.

Figura 48 – Gráficos - PIB a preços de mercado - Taxa trimestre contra trimestre imediatamente anterior



Fonte: Elaboração própria.

APÊNDICE B – Procedimentos realizados

O presente Apêndice possui como objetivo apresentar, de forma detalhada, todos os passos e procedimentos que foram realizados neste trabalho. Segue a lista dos procedimentos, divididos em tópicos:

a) Coleta das séries temporais econômicas;
   - Sistema Gerenciador de Séries Temporais (SGS) do Banco Central do Brasil (BCB);
   - Instituto Brasileiro de Geografia e Estatística (IBGE);
   - Ipeadata;
   - outros repositórios.

b) Realização dos testes de raiz unitária sazonal e não sazonal;
   - Dickey-Fuller Aumentado (ADF);
   - Dickey-Fuller *Generalized Least Squares* (DF-GLS);
   - Kwiatkowski, Phillips, Schmidt e Shin (KPSS);
   - Phillips e Perron (PP);
   - *bootstrap* do ADF;
   - *bootstrap* do teste de união;
   - Kruskall-Wallis (KW);
   - Hylleberg, Engle, Granger e Yoo (HEGY).

c) Realização das diferenciações necessárias;

d) Criação de várias bases de dados para cada um dos 3 meses que compõem cada um dos 8 trimestres;
   - 1° trimestre de 2025;
   - 4° trimestre de 2024;
   - 3° trimestre de 2024;
   - 2° trimestre de 2024;
   - 1° trimestre de 2024;
   - 4° trimestre de 2023;
   - 3° trimestre de 2023;
   - 2° trimestre de 2023;

e) Seleção de variáveis conforme as quatro técnicas, com o IBC-Br como variável dependente e as demais séries mensais como variáveis independentes;
   - *Least Absolute Shrinkage and Selection Operator* (LASSO);

  - *Elastic Net* (ENET);
  - importância por permutação em blocos (IPB) da *Random Forest* (RF) com *Moving Block Bootstrap* (MBB);
  - importância por permutação em blocos (IPB) da *Random Forest* (RF) com *Circular Block Bootstrap* (MBB).

f) Criação de várias bases de dados para cada uma das 4 técnicas de seleção de variáveis para cada um dos 3 meses que compõem cada um dos 8 trimestres;
  - 1° trimestre de 2025;
  - 4° trimestre de 2024;
  - 3° trimestre de 2024;
  - 2° trimestre de 2024;
  - 1° trimestre de 2024;
  - 4° trimestre de 2023;
  - 3° trimestre de 2023;
  - 2° trimestre de 2023;

g) Cálculo do número ótimo de fatores latentes, conforme Bai and Ng (2002), para cada uma das 4 técnicas de seleção de variáveis, além da especificação com todas as variáveis (ALL), para cada um dos 3 meses que compõem cada um dos 8 trimestres;
  - 1° trimestre de 2025;
  - 4° trimestre de 2024;
  - 3° trimestre de 2024;
  - 2° trimestre de 2024;
  - 1° trimestre de 2024;
  - 4° trimestre de 2023;
  - 3° trimestre de 2023;
  - 2° trimestre de 2023;

h) Cálculo do número ótimo de defasagens, conforme o *Schwarz Criterion* (SC), para cada uma das 4 técnicas de seleção de variáveis, além da especificação com todas as variáveis (ALL), para cada um dos 3 meses que compõem cada um dos 8 trimestres;
  - 1° trimestre de 2025;
  - 4° trimestre de 2024;
  - 3° trimestre de 2024;
  - 2° trimestre de 2024;

  - 1° trimestre de 2024;
  - 4° trimestre de 2023;
  - 3° trimestre de 2023;
  - 2° trimestre de 2023;

i) Realização dos exercícios de *nowcasting* com um DFM usando as variáveis selecionadas por cada uma dessas 4 técnicas, além da especificação com todas as variáveis (ALL), e para cada um dos 3 meses que compõem cada um dos 8 trimestres;
  - 1° trimestre de 2025;
  - 4° trimestre de 2024;
  - 3° trimestre de 2024;
  - 2° trimestre de 2024;
  - 1° trimestre de 2024;
  - 4° trimestre de 2023;
  - 3° trimestre de 2023;
  - 2° trimestre de 2023;

j) Cálculo do Erro Quadrático Médio de Previsão (MSFE) para o primeiro mês, segundo mês, terceiro mês e global dos DFMs com diferentes especificações;
  - último dia do 1° mês;
  - último dia do 2° mês;
  - último dia do 3° mês;
  - valor global.

k) Criação de várias bases de dados do PIB para cada um dos 3 meses que compõem cada um dos 8 trimestres;
  - 1° trimestre de 2025;
  - 4° trimestre de 2024;
  - 3° trimestre de 2024;
  - 2° trimestre de 2024;
  - 1° trimestre de 2024;
  - 4° trimestre de 2023;
  - 3° trimestre de 2023;
  - 2° trimestre de 2023;

l) Realização dos exercícios de previsão usando um AR(1);
  - 1° trimestre de 2025;

  - 4° trimestre de 2024;
  - 3° trimestre de 2024;
  - 2° trimestre de 2024;
  - 1° trimestre de 2024;
  - 4° trimestre de 2023;
  - 3° trimestre de 2023;
  - 2° trimestre de 2023;

m) Cálculo do Erro Quadrático Médio de Previsão (MSFE) para o primeiro, segundo e terceiro meses do AR(1);
  - último dia do 1° mês;
  - último dia do 2° mês;
  - último dia do 3° mês;

n) Realização dos exercícios de previsão usando um AR(2);
  - 1° trimestre de 2025;
  - 4° trimestre de 2024;
  - 3° trimestre de 2024;
  - 2° trimestre de 2024;
  - 1° trimestre de 2024;
  - 4° trimestre de 2023;
  - 3° trimestre de 2023;
  - 2° trimestre de 2023;

o) Cálculo do Erro Quadrático Médio (MSFE) para o primeiro, segundo e terceiro meses do AR(2);
  - último dia do 1° mês;
  - último dia do 2° mês;
  - último dia do 3° mês.