UNIVERSIDADE DO ESTADO DE SANTA CATARINA – UDESC
CENTRO DE CIÊNCIAS DA ADMINISTRAÇÃO E SOCIOECONÔMICAS – ESAG
DEPARTAMENTO DE CIÊNCIAS ECONÔMICAS - DCE

CAIRÊ BRITTO BARLETTA

TEXT MINING NAS ATAS DO COPOM PARA CRIAÇÃO DE UM ÍNDICE DE SENTIMENTO E AVERIGUAÇÃO DE CAUSALIDADE COM VARIÁVEIS MACROECONÔMICAS

FLORIANÓPOLIS
2022

CAIRÊ BRITTO BARLETTA

TEXT MINING NAS ATAS DO COPOM PARA CRIAÇÃO DE UM ÍNDICE DE SENTIMENTO E AVERIGUAÇÃO DE CAUSALIDADE COM VARIÁVEIS MACROECONÔMICAS

Trabalho de Conclusão de Curso apresentado ao curso de Ciências Econômicas, do Centro de Ciências da Administração e Socioeconômicas, da Universidade do Estado de Santa Catarina, como requisito parcial para a obtenção do grau de Bacharel em Ciências Econômicas.

Orientador: Prof.ª Dr.ª Marianne Zwilling Stampe

FLORIANÓPOLIS
2022

CAIRÊ BRITTO BARLETTA

TEXT MINING NAS ATAS DO COPOM PARA CRIAÇÃO DE UM ÍNDICE DE SENTIMENTO E AVERIGUAÇÃO DE CAUSALIDADE COM VARIÁVEIS MACROECONÔMICAS

Trabalho de Conclusão de Curso apresentado ao curso de Ciências Econômicas, do Centro de Ciências da Administração e Socioeconômicas, da Universidade do Estado de Santa Catarina, como requisito parcial para a obtenção do grau de Bacharel em Ciências Econômicas.

Orientador: Prof.ª Dr.ª Marianne Zwilling Stampe

BANCA EXAMINADORA:

Nome do Orientador e Titulação
Nome da Instituição

Membros:

Nome do Orientador e Titulação
Nome da Instituição

Nome do Orientador e Titulação
Nome da Instituição

Nome do Orientador e Titulação
Nome da Instituição

Florianópolis, 05 de julho de 2022

Dedico este trabalho aos meus pais,
sem vocês, eu nada seria.

AGRADECIMENTOS

Agradeço à minha família por ser o elemento de maior importância na minha vida. Em especial, à minha mãe Fabiane, ao meu pai Rodrigo, ao meu irmão Cainã, às minha avós Deuza e Rose, e aos meus avôs Antônio e Walter.

Agradeço à minha orientadora, Professora Doutora Marianne Zwilling Stampe, por todos os direcionamentos dados na elaboração do presente trabalho. Não só isso, agradeço pelos incentivos, comprometimento e parceria durante toda a jornada.

Agradeço também aos professores, mestres e doutores, que foram incondicionalmente importantes na minha trajetória, me marcando significativamente. Em especial, minha gratidão para Thais Waideman Niquito, Fernando Pozzobon, Marcello Beckert Zapelini, Marcos Vinicio Wink Junior, Analucia Vieira Fantin, Patrícia Bonini, Cassiano Ricardo Dalberto, e minha orientadora, Marianne Zwilling Stampe. Tenho imenso respeito e admiração pelo trabalho de todos vocês.

Agradeço também ao Fernando Zatt Schardosin, que me forneceu minha primeira e única bolsa de iniciação científica, logo no início da minha jornada acadêmica.

Agradeço aos amigos que já possuía antes de entrar na faculdade, por fazerem meus dias mais leves e agradáveis. Em especial, Peter Krause, Thales Zirbel, Yan Romano, Gabriel Lima, Rafael Aché, Pedro Aché e Arthur Masseli. Vocês são muito importantes para mim.

Agradeço às amizades que construí durante a graduação, por também fazerem meus dias mais leves e agradáveis, além de enfrentarem o percurso ao meu lado — na medida do possível. Em especial, Leonardo Oviedo, Bernardo Couto, José Marcelo Felleti, Isadora Zamprogna, Arthur Vier, Sabrina Bacaicoa, Nelson Ambros, Gabriel Akira, Henrique Matte e Arthur Rockenbach. A jornada não teria sido a mesma sem vocês.

Agradeço também pela oportunidade e experiência que o Clube de Finanças, projeto de extensão de liga acadêmica de mercado financeiro, me proporcionou.

Por fim, porém não menos importante, agradeço à Universidade do Estado de Santa Catarina e todos os colaboradores que nela trabalham.

Muito obrigado.

RESUMO

Este estudo apresenta a importância da comunicação e transparência de bancos centrais para a condução da política monetária. Dada esta relevância, foram empregadas técnicas de mineração textual e análise de sentimentos nas atas do Comitê de Política Monetária para criação de um índice, contemplando o período de 2006 até 2022. O índice foi então comparado com variáveis macroeconômicas, e após isso foram efetuados testes estatísticos, para contribuir com a robustez do trabalho. Como resultados, aplicando Vetores Autorregressivos (VAR), através de Funções de Impulso-Resposta (IRF), pode-se dizer que dado um choque positivo no índice, a taxa básica de juros real é influenciada negativamente, enquanto que para a atividade econômica e para a produção industrial, percebe-se que a resposta é positiva, onde todas se estabilizam em torno de zero no horizonte analisado, uma vez que os choques não possuem efeitos permanentes em séries estacionárias.

Palavras-chave: Mineração Textual. Análise de Sentimentos. Atas do Copom. Vetores Autorregressivos (VAR). Função Impulso-Resposta (IRF).

ABSTRACT

This research presents the importance around central banking communication and transparency to the conduction of monetary policy. Given this relevance, text mining and sentiment analysis techniques were applied to the Copom minutes in search of an index creation, covering the period from 2006 to 2022. The index was then compared with macroeconomic variables, and after that, statistical tests were performed to contribute to the robustness of the work. As a result, applying Autoregressive Vectors (VAR), through Impulse-Response Functions (IRF), it can be said that given a positive shock on the index, the real basic interest rate is negatively influenced, while for economic activity and for industrial production, it can be seen that the answer is positive, where they all stabilize around zero in the analyzed horizon, since the shocks do not have permanent effects in stationary series.

Keywords: Text Mining. Sentiment Analysis. Copom Minutes. Vector Autoregressive (VAR). Impulse Response Function (IRF).

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BCB	Banco Central do Brasil
Copom	Comitê de Política Monetária
VAR	Vetores Autorregressivos
IRF	Função de Impulso-Resposta
EUA	Estados Unidos da América
Fed	Federal Reserve
FOMC	Federal Open Market Committee
BCNZ	Banco Central da Nova Zelândia
BCI	Banco Central da Inglaterra
BCN	Banco Central da Noruega
BCS	Banco Central da Suécia
BCE	Banco Central Europeu
RMI	Regime de Metas de Inflação
NLP	Processamento de Linguagem Natural
ETTJ	Estrutura a Termo da Taxa de Juros
OF	Fator de Otimismo
IIE-Br	Indicador de Incerteza da Economia-Brasil
FGV	Fundação Getúlio Vargas
BCG	Banco Central da Gana
ADF	Dickey-Fuller Aumentado
AIC	Critério de Informação de Akaike
BIC	Critério de Informação Bayesiano
PP	Phillips–Perron
KPSS	Kwiatkowski–Phillips–Schmidt–Shin
HQ	Hannan-Quinn
VMA	Vetor de Médias Móveis
MQO	Mínimos Quadrados Ordinários
DW	Durbin-Watson
BG	Breusch–Godfrey
IBGE	Instituto Brasileiro de Geografia e Estatística

IPCA	Índice Nacional de Preços ao Consumidor Amplo
IBC-Br	Índice de Atividade Econômica do Banco Central-Brasil
PIM-PF	Pesquisa Industrial Mensal-Produção Física
SGS	Sistema Gerenciador de Séries
SIDRA	Sistema IBGE de Recuperação Automática

SUMÁRIO

1 INTRODUÇÃO

Com cada vez mais disponibilidade, variedade, volume e velocidade da produção de dados e avanço nos recursos tecnológicos, fez-se necessário a criação de técnicas que armazenem, tratem e analisem esses dados de maneira eficaz. No âmbito da ciência da economia, apesar de tais técnicas serem ainda poucos exploradas, pode-se esperar uma grande evolução na implicação de como mensurar os efeitos econômicos, levando em consideração que cada vez mais grandes quantidades de dados estão sendo dispostas, tanto por instituições públicas quanto privadas (COSTA, 2016).

Mais especificamente falando, pode-se utilizar de dados obtidos através dos comunicados oficiais da autoridade monetária brasileira, que visa principalmente conduzir a política monetária do país, controlando a inflação em determinado nível estável, além de afetar diretamente a situação da atividade econômica. Dessa forma, como problema de pesquisa, o trabalho por se tratar de método quantitativo-estatístico, se propõe a quantificar os textos emitidos, a fim de mensurar os impactos da comunicação nas expectativas dos agentes econômicos em variáveis macroeconômicas, permitindo o estabelecimento de relações e causalidades existentes.

Este estudo tem como objetivo principal a criação de um índice de sentimentos, com base nos comunicados oficiais do Banco Central do Brasil (BCB), para medir quantitativamente e inferir se de fato, e como, a comunicação da autoridade monetária surte efeitos em variáveis macroeconômicas brasileiras, sendo elas a taxa básica de juros descontada da inflação, isto é, a taxa básica de juros real; a atividade econômica e a produção industrial. O período da análise foi ínicio 2006 até final de 2022, incluindo da ata 116 até a ata 246.

Como objetivos secundários, engloba-se o procedimento de raspagem de dados *web* para coleta dos dados, bem como técnicas de mineração textual (quantificação de textos), aplicados aos comunicados de condução da política monetária brasileira.

Fazendo uso destas ferramentas, se acessou e extraiu-se os arquivos referentes às atas do Comitê de Política Monetária (Copom), disponibilizados no site do BCB, de forma automatizada. Em seguida, os dados foram tratados e com base em dicionários léxicos, proposto primeiramente por Stone, Dunphy e Smith (1966), foram atribuídos *scores* de teor positivo ou negativo para as atas, sendo possível a criação de um índice de sentimento ao longo do período analisado.

Por fim, foram aplicadas técnicas econométricas tradicionais em séries temporais, abordadas em Gujarati e Porter (2011), Bueno (2011) e outros autores, visando contribuir para a robustez dos resultados empíricos, utilizando modelos de Vetores Autorregressivos (VAR) e testes estatísticos necessários no índice de sentimentos criado, relacionado-o à variáveis macroeconômicas de interesse, obtendo as respectivas Funções de Impulso-Resposta (IRF) e inferindo ou não a causalidade de Granger, a partir dos choques efetuados.

Com isso, considerando que o objetivo geral é analisar quantitativamente os comunicados do BCB e mensurar os impactos nas variáveis macroeconômicas sugeridas, de forma enumerada, pode-se organizar os objetivos específicos da seguinte maneira:

a) criação de um algoritmo de raspagem de dados *web* que baixará de forma automatizada todas as atas do Copom, que serão utilizadas como uma das base de dados, organizando-as em um *Corpus* (conjunto de documentos);
b) re-agrupar as palavras em padrões (*token*), atribuindo um sentido positivo ou negativo;
c) criar o índice de sentimentos, com base no conjunto de palavras, quantificando em valores;
d) relacionar o índice defasado com as variáveis macroecônomias, utilizando modelos VAR, para quantificar os impactos dos comunicados determinado tempo à frente.

Objetivos estes que basearam-se na importância da comunicação do Banco Central, que conforme elucidam Blinder et al. (2008), pode ser definida como qualquer informação disponibilizada para o público em relação à condução da política monetária.

Nos EUA, anterior ao período de 1987, era considerado que ao pegar o mercado de surpresa, a política monetária teria efeitos mais eficazes. Woodford (2005), demonstrou em seu estudo, que uma comunicação bem feita por parte dos bancos centrais é pré-requisito básico na condução das políticas, uma vez que os principais tomadores de decisão em uma economia olham para o futuro, dando bons motivos para o comprometimento com a explicação ao público das decisões tomadas.

Além disso, a inferência que os agentes do mercado possuem sobre o Banco Central se concentra no fato da confiança estabelecida na capacidade de condução da instituição manter os preços e a economia estáveis, onde essa credibilidade passa por processo de construção, preservada e afirmada por meio de ações e explicações fornecidas ao passar do tempo. Quanto melhor esse papel for cumprido, de defensor do equilíbrio da moeda, maior a reputação da entidade (ISSING; WOOD, 2001).

Por conseguinte, no entendimento de Winkler (2000), o primeiro ato realizado por um Banco Central que visa o aumento de transparência deveria ser em tornar suas ações e visão de mundo entendida por todos, além de fornecer a informação de modo que fosse compreendida pelos diferentes agentes.

Por isso, assume-se de grande importância a pesquisa acerca do impacto dos comunicados emitidos sobre as variáveis macroeconômicas, a fim de ter uma quantificação da influência, e outras características dos informes (relações, correlações, causalidades etc).

2 REFERENCIAL TEÓRICO

Neste capítulo serão apresentados em detalhe as definições, referências e teorias necessárias para o desenvolvimento e compreensão dos estudos e análises desenvolvidas no presente trabalho, bem como a apresentação de pesquisas em que técnicas de mineração textual foram aplicadas em problemas econômicos e na análise de sentimentos de bancos centrais. Na Seção 2.1 e em suas Subseções é discursado sobre a importância da comunicação e transparência do banco central, assim como elucidação sobre regimes de regras de inflação e atas do Copom. Já na Seção 2.2 é apresentado o processo e relação da mineração textual com aplicações no âmbito econômico.

2.1 IMPORTÂNCIA DA COMUNICAÇÃO DE UM BANCO CENTRAL

Pode-se definir a comunicação do banco central como qualquer quantidade e qualidade de informação disponibilizada ao público que está relaciona à condução da política monetária, em relação à atividade econômica e em relação à sinais da trajetória de políticas futuras, por parte do banco central (BLINDER et al., 2008).

A partir de 1987 até meados de 2004, foi observado na economia dos Estados Unidos da América (EUA), um período no *Federal Reserve (Fed)*[1] em que foi caracterizado por grande adesão por parte das declarações do *Federal Open Market Committee (FOMC)*[2] à comunicados sobre o momento vigente e perspectivas futuras. Anteriormente à esse período, era considerado que ao pegar o mercado de surpresa, teria-se um efeito mais eficaz da política monetária (WOODFORD, 2005).

Como enunciam Blinder et al. (2008, p. 7), é cabível se chamar de uma "revolução no pensamento", o fato de partir-se de um ponto (início da década de 80) onde o banco central praticamente não se comunicava com os agentes, para outro ponto em que proclamou-se que um aumento na comunicação aumentaria a efetividade da política monetária (meados da década de 90) e depois para outro momento em que foi apontado que a essência da política monetária seria a "arte de gerenciar expectativas"(ou pelo menos em partes). Essas mudanças no jeito de pensar sobre a comunicação do banco central afetou a prática dele em si.

Nos EUA, o início da jornada acerca de maior transparência iniciou-se em 1994, quando o FOMC anunciou pela primeira vez suas decisões à respeito dos alvos-meta da taxa de juros. Em 1999, começou-se a publicar a tendenciosidade do Fed acerca das mudanças futuras e também a elaboração de comunicados mais completos. Em 2003, o Fed iniciou a anunciar os votos atrelados aos votantes após cada reunião, como também a explicitamente gerenciar expectativas ao declarar abertamente que manteria a taxa de juros em níveis baixos por um período considerável.

Entretanto, o Fed não lidera o caminho nessa questão, haja vista que outros bancos centrais ao redor do mundo ao longo do tempo vieram dando grande valor para sua comunicação,

[1] Equivalente ao Banco Central no Brasil.
[2] Equivalente ao Copom no Brasil.

podendo ser citados o Banco Central da Nova Zelândia (BCNZ), o Banco Central da Inglaterra (BCI), o Banco Central da Noruega (BCN), o Banco Central da Suécia (BCS) e o Banco Central Europeu (BCE). Uma comunicação mais transparente e efetiva por parte dos bancos centrais realmente é uma convergência internacional (BLINDER et al., 2008).

Uma importante razão do crescimento de transparência é a noção estabelecida de que bancos centrais deveriam deter maior responsabilidade, tendo o dever de explicar não só seus atos, como também os motivos por detrás deles. Além disso, conforme ficou mais evidente que a condução das expectativas é um aparato útil da política monetária, a comunicação passou de um fardo a ser lidado para um instrumento chave dentre as ferramentas do Banco Central (BLINDER et al., 2008).

Além disso, Woodford (2005), concluiu que uma comunicação eficaz é pré-requisito básico na condução de política monetária por parte de qualquer banco central que visa ter sucesso em seus objetivos:

> Como os principais tomadores de decisão em uma economia olham para o futuro, os bancos centrais afetam a economia tanto por sua influência nas expectativas quanto por quaisquer efeitos mecânicos diretos da negociação do banco central no mercado por dinheiro de um dia para o outro. Como consequência, há boas razões para um banco central se comprometer com uma abordagem sistemática da política, que não apenas fornece uma estrutura explícita à tomada de decisões dentro do banco, mas que também é usada para explicar as decisões do banco ao público. (WOODFORD, 2005).

De acordo com a literatura desenvolvida com o passar do tempo, é de conhecimento geral o fato de que a comunicação do Banco Central é uma ferramenta muito poderosa em diversos sentidos, podendo ser citados alguns deles, como a ancoragem de expectativas do mercado, a expectativa no preço de ativos financeiros, a melhora na transparência e a previsibilidade da política monetária e um redutor de caminho para se obter uma maior estabilidade econômica (BLINDER, 1999; MISHKIN, 2000; BERNANKE; REINHART; SACK, 2004).

Não somente, por meio do trabalho desenvolvido por Coenen et al. (2017), observa-se que o crescimento e a popularização dos comunicados emitidos por bancos centrais, devem-se principalmente pela necessariedade de uma maior transparência na condução da política monetária, levando em consideração a busca em atingir níveis de inflação que estão dentro das metas, assim como o crescente nível de autonomia dos bancos centrais ao redor do mundo.

2.1.1 A teoria por trás da comunicação do banco central

Atualmente, é amplamente aceito na literatura que o poder de um banco central afetar a economia depende fortemente de sua capacidade de influênciar as expectativas do mercado acerca do trajeto futuro das taxas de juros, e não apenas de seus níveis no período em questão. De acordo com as teorias que tangem a estrutura a termo, as taxas de juros de longo prazo deveriam refletir a sequência de expectativas sobre o futuro das taxas (BLINDER et al., 2008).

Sendo assim, como aborda Blinder et al. (2008), tem-se que a taxa de juros (R_t) do dia n deve ser aproximadamente igual a:

$$R_t = \alpha_n + \left(\frac{1}{n}\right)\left[r_t + \sum_{i=1}^{N} E(r_{t+i})\right] + \varepsilon_t \tag{1}$$

em que α_n é um termo de prêmio; r_t é a taxa atual de juros; $\sum_{i=1}^{N} E(r_{t+i})$ é o somatório das expectativas atuais acerca das taxas nos períodos $t+i$; e ε_t é um termo de erro aleatório. A Equação (1) demonstra que a taxa de longo prazo depende das expectativas do público sobre a política futura do banco central, além de apontar a baixa relevância da taxa atual. Caso tiver-se que a taxa de juros atual seja próximo de zero, a comunicação sobre a taxa futura esperada se transforma na essência da política monetária (BERNANKE; REINHART; SACK, 2004; WOODFORD, 2005; BLINDER et al., 2008).

Adicionando uma estrutura macroeconômica criada para ilustrar o papel da comunicação do banco central; e denotando r_t como a taxa de juros de curto prazo e R_t como a taxa de longo prazo, ambas da Equação (1), tem-se que a demanda agregada (D) depende desses dois fatores, expectativa da inflação (π_t^e) e outros fatores não citados, conforme:

$$Y_t = D(r_t - \pi_t^e, R_t - \pi_t^e, \ldots) + \varepsilon_t \tag{2}$$

Sendo que, a oferta agregada pode (não necessariamente) ser definida como a Curva de Phillips:

$$\pi_t = \beta E(\pi_{t+1}) + \gamma(Y_t - Y_t^*) + \varepsilon_t \tag{3}$$

onde π_t representa a inflação no período t; $E(\pi_{t+1})$ é a inflação esperada em $t+1$; Y_t é o produto interno bruto real atual; e Y_t^* é o produto interno bruto real potencial. Completa-se então o modelo adicionando uma função de reação do banco central, como por exemplo a aplicação da Regra de Taylor (1993):

$$(r_t - r_t^*) = \pi_t + \beta_\pi(\pi_t - \pi_t^*) + \beta_y(y_t - y_t^*) + \varepsilon_t \tag{4}$$

tal qual r_t é a taxa nominal de curto prazo; r_t^* indica a taxa real de equilíbrio; π_t^* é a taxa meta de inflação do banco central; y_t é o $\ln(Y_t)$; e y_t^* é o $\ln(Y_t^*)$, isto é, o desvio do produto em relação ao potencial. Tome como exemplo o caso em que o ambiente econômico faz com que as Equações (1), (2) e (3) sejam constantes ao longo do tempo (estacionárias); que o banco central seja confiavelmente comprometido com sua função de reação demonstrada pela Equação (4); e que as expectativas sejam racionais.

Nesta situação hipótetica irrealista, como levantado em Woodford (2005), a comunicação do Banco Central não teria nenhum papel a desempenhar e dessa forma qualquer padrão observado de como a política monetária fosse conduzida seria corretamente deduzido pelos agentes de mercado, fazendo com que toda comunicação explícita do BC fosse redundante.

Sendo assim, considerando transparência do Banco Central abordada em Faust e Svensson (2001) como a facilidade com que o público consegue deduzir as intenções e metas a partir dos dados observáveis, teria-se um BC totalmente transparente sem divulgação de comunicado algum (BLINDER et al., 2008).

Esta conjuntura extrema aponta para algumas características que possuem o potencial de fazer a comunicação do banco central ser importante, sendo elas:

a) a não estacionariedade (seja da economia ou da regra de condução da política monetária);
b) o constante aprendizado no ambiente (do e sobre o Banco Central);
c) expectativas não-racionais e/ou assimetria de informação entre os agentes econômicos e o BC

Uma vez que uma ou mais dessas condições são alcançadas, a comunicação do BC pode importar, e levando em consideração que essas situações são comuns - e não exceções - destaca-se a importância da comunicação.

Além do mais, é intrinsecamente inevitável o fato de que o BC sabe mais sobre o próprio jeito de pensar do que os agentes econômicos. Não somente isso, como aponta Svensson (2003), as decisões tomadas acerca da política monetária dependem de muitas outras coisas além da inflação corrente e hiatos do produto como apontado na Equação (4). É também extremamente improvável o cenário no qual o Banco Central se agarraria a uma política sem quaisquer mudanças por muito tempo (BLINDER et al., 2008).

Conforme levantado por Bernanke (2004b) e abordado em Blinder et al. (2008), tem-se também o motivo da comunicação afetar a efetividade da política monetária: quando o público não conhece a função de reação do banco central, e por consequência precisa estimá-la, não há garantia de que a economia convergirá para o equilíbrio de expectativas racionais, uma vez que o processo de aprendizado dos agentes externos afeta o comportamento da economia. Não obstante, os autores apontam que não é prático especificar e explicitar uma regra de política monetária ao qual o BC nunca desviaria independente da circunstância, e neste caso, o problema está no fato de que o número de eventualidades às quais a política do BC deve responder são infinitas e na grande maioria imprevisíveis.

Costa Filho e Rocha (2009) apontam que, de acordo com Haan, Eijffinger e Rybiński (2007), há três razões para a comunicação de um Banco Central ser relevante. Primeiramente, as expectativas não são racionais. Depois, tem-se informação assimétrica (o BC possui mais informação sobre a economia do que os agentes de mercado), o que justifica a importância que os agentes dão aos comunicados como referência de se ajustar as expectativas. Por fim, na ausência de regras de política e credibilidade da autoridade, a comunicação é o canal de fornecimento de informações aos agentes sobre a condução da política monetária.

Entretanto, há na literatura argumentos favoráveis e desfavoráveis acerca da adoção de uma maior transparência. Entre ideias a favor pode-se citar uma política monetária mais previsível, aumentando assim sua eficácia, assim como aponta Bernanke (2004a) e um aumento

na credibilidade do banco central no médio e longo prazo, de acordo com Issing (2005).

Enquanto para ideias contrárias, destaca-se o fato de que transparência é desejável somente se possuir uma relação positiva com uma política monetária mais eficaz, como elucidado por Issing (2005). Além disso, Woodford (2005) e Issing (2005), levantam que uma maior previsibilidade advinda de mais transparência, não necessariamente torna a política monetária mais eficaz, uma vez que a ação da política monetária é sempre contingente às condições econômicas (COSTA FILHO; ROCHA, 2009).

2.1.2 Atas do Copom como mecanismo de comunicação e transparência: uma revisão da literatura

O Banco Central possui uma função principal explícita, que é a de controlar a inflação, sendo o controle da atividade econômica, um objetivo implícito. Para poder controlar a taxa inflacionária, o BCB se utiliza da taxa básica de juros, a taxa Selic. Quanto mais elevada se está a taxa de inflação, mais se faz necessário elevar a taxa básica de juros de uma economia, estimulando assim uma maior poupança, resultando em menor consumo ao passo que o custo do crédito aumenta, interferindo diretamente na atividade econômica.

A disseminação de uma abordagem mais transparente na condução da política monetária possui, com grande relevância, a adoção de práticas do Regime de Metas de Inflação (RMI) como arcabouço principal, mesmo não sendo restrito a esse regime. Apesar disso, há exemplos de bancos centrais transparentes que não adotam o regime de metas de inflação, como o BCE e o Fed (ISSING, 2005).

Em regimes de política monetária orientados por regras, se os agentes do mercado possuem um entendimento suficiente da regra vigente, o nível requisitado de transparência é menor. Neste caso, os simples atos da condução da política monetária seriam explicações o bastante, reduzindo assim a necessidade da autoridade monetária se pronunciar. Dessa forma, considerando que o RMI implica arbitrariedade aos que conduzem a política monetária, a comunicação do Banco Central deve possuir papel fundamental em coordenar e administrar as expectativas dos agentes de mercado (COSTA FILHO; ROCHA, 2009).

Conforme Issing (2005) aponta, vale salientar que por mais que seja explicitamente informado um alvo numérico para a taxa de inflação, caso incorra-se em possíveis desvios da inflação em relação a meta, a forma e trajeto de ajuste são selecionados discricionariamente. Na verdade, quando trata-se da relação entre a discrição e regra, raramente tem-se um Banco Central seguindo rigorosamente as regras, por causa do chamado viés inflacionário, isto é, o impulso de se diminuir o desemprego ou aumentar o produto. Dessa forma, quando o RMI não possui credibilidade, a tendência é resultar na discrição, aumentando a oferta de moeda e por consequência gerando uma taxa de inflação maior.

No caso brasileiro, as atas do Copom servem como a principal ferramenta de comunicação do BCB, não só no âmbito nacional, como também para o exterior. É por meio destas que a autoridade monetária discorre acerca do processo de tomada de decisão da política monetária,

mantendo maior previsibilidade sobre as expectativas dos agentes econômicos, apresentando dados macroeconômicos pertinentes, discorrendo sobre a inflação, decidindo sobre o nível da taxa de juros básica e salientando sobre as perspectivas futuras (COSTA FILHO; ROCHA, 2010).

2.2 MINERAÇÃO TEXTUAL

Mineração textual[3] é a dinâmica que utiliza da tecnologia de Processamento de Linguagem Natural (NLP) para organizar e converter dados textuais disponíveis em documentos e grandes bases de dados, por meio de ferramentas estatísticas e computacionais, possibilitando a quantificação do texto. O principal objetivo desta técnica é extrair significados de uma forma que um ser humano sozinho não teria capacidade, mas uma máquina sim, obtendo padrões nos textos que não seriam encontrados *a priori* (BHOLAT et al., 2015).

Com a grande quantidade de informações disponíveis na internet, a ciência econômica - que ainda não possui a devida inserção e aderência de técnicas de mineração textual - pode se beneficiar, e muito, cada vez mais dessa disponibilidade toda. Um dos possíveis usos, é por parte dos bancos centrais, extraindo informações por métodos não usuais, de fontes diversas, para inferir de forma mais completa conjunturalmente sobre questões monetárias e financeiras. Como enunciado em Bholat et al. (2015):

> A mineração textual pode valer o investimento dos bancos centrais porque essas técnicas tornam tratável uma série de fontes de dados que são importantes para avaliar a estabilidade monetária e financeira e não podem ser analisadas quantitativamente por outros meios. Os principais dados de texto para bancos centrais incluem artigos de notícias, contratos financeiros, mídia social, supervisão e inteligência de mercado e relatórios escritos de vários tipos. (BHOLAT et al., 2015).

Alguns estudos já aplicaram técnicas de mineração textual - e suas derivações - para inferir sobre documentos textuais no âmbito econômico. Chague et al. (2015) elaboraram um estudo para a economia brasileira, onde analisaram como a comunicação do BCB afeta a Estrutura a Termo da Taxa de Juros (ETTJ) estimada. Por meio da criação de um Fator de Otimismo (OF), quando as atas do Copom indicavam otimismo, a OF aumentava, e as taxas de juros de mais longo prazo diminuiam e vice-versa. Seus resultados sugeriram que a comunicação da autoridade monetária possuem impactos efetivos nas expectativas de mercado.

Como aborda Costa (2016), em busca de uma nova medida de inflação para países cujos índices das autoridades oficiais perderam muita credibilidade, Cavallo (2013) coletou com mineração textual dados de 2007 a 2011 das páginas dos principais supermercados do Brasil, Chile, Colômbia, Venezuela e Argentina; e combinando com pesos oficiais das categorias dos produtos, criou um índice de inflação alternativo ao divulgado oficialmente. Para os quatro primeiros países incluídos na lista, o índice se aproximou tanto em nível como em dinâmica

[3] Também conhecido como: (i) *text mining*, (ii) mineração de textos

temporal da inflação oficial. Para a Argentina, encontrou-se grande discrepância entre o índice de preços online e a divulção oficial, que mostrou-se persistente em todo o período analisado.

Baseado na literatura internacional sobre mensuração de incerteza e seus efeitos na economia, Ferreira et al. (2017) desenvolveram o Indicador de Incerteza Econômica - Brasil (IIE-Br), cuja Fundação Getúlio Vargas (FGV) divulga mensalmente. O indicador criado apontou forte relação com grandes momentos de incerteza vividos pelo Brasil, e após abordagem econométrica, observou-se que choques de incerteza produzem efeitos negativos sobre a atividade econômica e produção industrial.

Já no artigo elaborado por Omotosho (2019), são avaliados os comunicados do comitê de política monetária do Banco Central da Gana (BCG) no período entre 2018 e 2019. Aplicando técnicas de mineração textual, análise de sentimentos e modelagem de tópicos (*Alocação Latente de Dirichlet*) proposto por Blei, Ng e Jordan (2003), obteve resultados de que, a partir da amostra utilizada, foi possível perceber evidências de uma melhora de transparência da política monetária por parte do BCG.

Destaca-se também que a composição das frases e utilização das palavras ficaram menos complexas, mostrando que os comunicados ficaram mais fáceis de serem lidos com o passar do tempo. Além disso, termos como *'inflação*, *'pib'*, *'monetário'* ficaram ressaltados, indicando consistência entre a estratégia de comunicação do BCG e seus objetivos. Por fim, por meio da análise de sentimentos e modelagem de tópicos, observou-se um *score* líquido médio no período analisado de $3,90\%$, representando que os comunicados do BCG indicam uma perspectiva positiva para a economia (OMOTOSHO, 2019).

Enquanto na maioria da literatura que faz do uso dessas técnicas, assume-se que banqueiros centrais fazem seus comunicados e depois observa-se as reações do mercado em relação à comunicação, a análise de Zahner (2021), por exemplo, foca nas mudanças na comunicação em resposta à variação da atividade econômica.

Levando em consideração duas das funções mais importantes de um banco central, que é controlar a inflação (sendo essa a principal) e controlar o nível da atividade econômica, são difíceis de quantificar por padrão, Zahner (2021), a partir do uso de mineração textual, extraiu dados de comunicados públicos de 2002 até 2020 do BCE, criou um índice de sentimentos e aplicou metodologias econométricas sugeridas por Shapiro e Wilson (2019) para estimar os objetivos relacionados às principais funções do BCE.

Seus estudos resultaram na descoberta de que os comunicados estão mais em linha com a função de bancos centrais preferirem favorecer melhores condições econômicas, independente do nível da economia e de que a inflação alvo ao longo do período rodeia levemente acima de $2,00\%$. Não só isso, seu trabalho revelou que a comunicação do BCE responde igualmente em variações da atividade econômica, assim como em variações da taxa de inflação, sendo que os comunicados ficam cada vez mais pessimistas com o passar do tempo (ZAHNER, 2021).

Dessa maneira, tem-se que técnicas de mineração textual cada vez mais vem sendo usadas na esfera macroeconômica, seja por meio de aplicações estatísticas ou seja por meio

de métodos de *Machine Learning*. Sendo assim, com o crescente nível de transparência dos comunicados emitidos pelos bancos centrais, estes verificam-se como uma propícia fonte de análise de sentimentos textuais, não só por sua influência nas expectativas dos agentes, mas também pela disponibilidade e avanço tecnológico, possibilitando a criação de algoritmos que facilitem e avancem na pesquisa focada no uso de mineração textual (e suas ramificações) para a extração de padrões, significados e relações úteis a partir de documentos que discorrem sobre a condução da política monetária (SHAPIRO; WILSON, 2019).

3 METODOLOGIA

Nesse Capítulo serão abordadadas as classificações da pesquisa, bem como aplicações das técnicas utilizadas no desenvolvimento do presente trabalho para obtenção e tratamento dos dados, assim como o ferramental metodológico econométrico empregado para conseguir-se os resultados.

3.1 CARACTERIZAÇÃO DA PESQUISA

Considerando o escopo geral do trabalho, define-o como classificado na categoria de pesquisa descritiva, onde uma das definições que pode-se dar é aquela que busca descobrir a existência de relações entre variáveis, ou até mesmo indo adiante, pretendendo determinar a natureza dessa relação, e neste caso, aproxima-se de uma pesquisa explicativa. Pesquisas descritivas em conjunto com as exploratórias são habitualmente realizadas por pesquisadores interessados com a atuação prática (GIL, 2008).

Além disso, levando em consideração a natureza do estudo, pode-se enquadrar o método científico como uma pesquisa quantitativa, dado o uso de técnicas que utilizam da quantificação desde a coleta, tratamento e apresentação das informações, baseando-se em ferramental estatístico. Ao longo de sua elaboração, são tomados os devidos mecanismos para dirimir ao máximo distorções na interpretação da análise, se apoiando em maior segurança à respeito das inferências então feitas. Ainda, conforme enunciado por Gil (2008):

> Mediante a utilização de testes estatísticos, torna-se possível determinar, em termos numéricos, a probabilidade de acerto de determinada conclusão, bem como a margem de erro de um valor obtido. Portanto, o método estatístico passa a caracterizar-se por razoável grau de precisão, o que o torna bastante aceito por parte dos pesquisadores com preocupações de ordem quantitativa. Os procedimentos estatísticos fornecem considerável reforço às conclusões obtidas, sobretudo mediante a experimentação e a observação. Tanto é que os conhecimentos obtidos em alguns setores da psicologia e da economia devem-se fundamentalmente à utilização do método estatístico. (GIL, 2008).

3.2 MÉTODO

Nessa Seção serão abordadas as técnicas para obtenção e tratamento dos dados. Após isso será apresentada a demonstração da abordagem estatística (econométrica) ao conjunto de dados de interesse, na missão de se chegar aos objetivos propostos, considerando que as explicações obtidas por tal método não são absolutamente verdadeiras, e sim muito próximas de serem verdadeiras (GIL, 2008).

3.2.1 Coleta e tratamento dos dados

Nessa Subseção serão abordados os processos para obtenção dos dados do presente estudo, bem como elucidar sobre o tema de análise de sentimentos, que possibilita a transformação de

informações qualitativas em quantitativas.

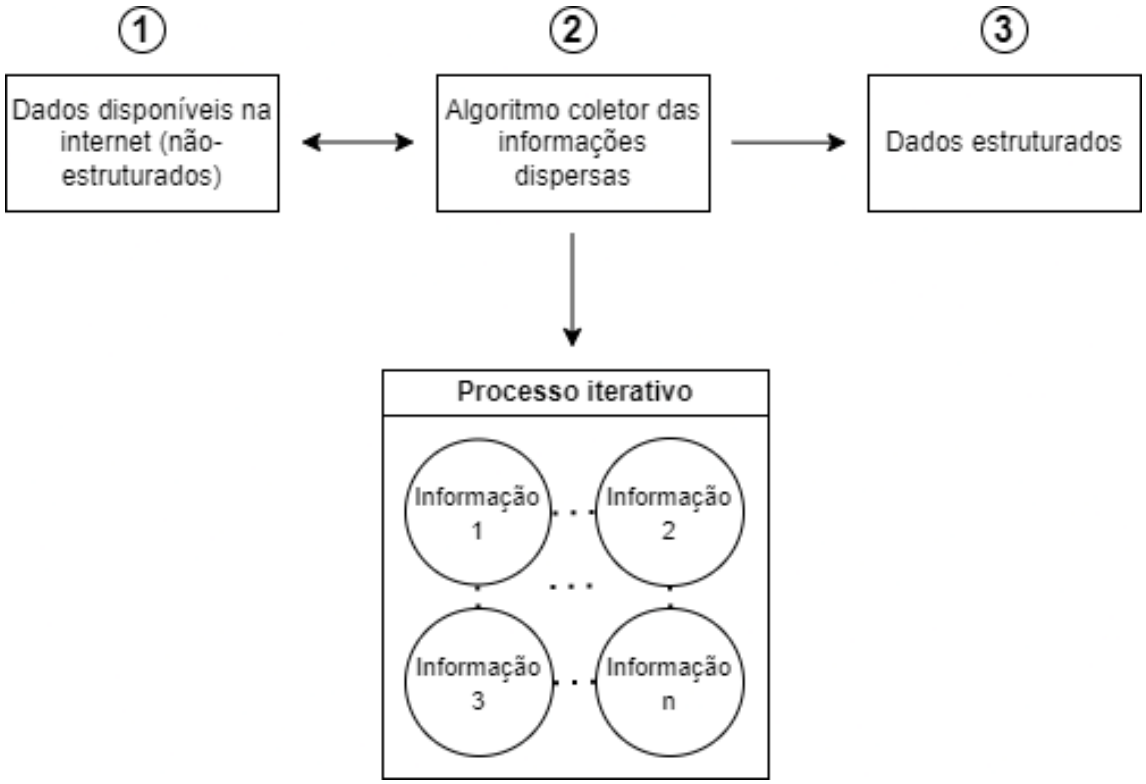
3.2.1.1 Raspagem de dados web

A partir da quantidade de informação que é gerada cada vez mais por meio da internet, observa-se a vasta disponibilidade de dados oriundos de diversas fontes diferentes. Mesmo que as informações possam ser muito úteis, na grande maioria das vezes elas não estão prontas para serem usadas, isto é, são dados não-estruturados.

Pode-se definir a raspagem de dados *web* [2] como um processo que facilita a execução de obtenção dos dados que estão dispersos pela internet, transformando-os em dados estruturados (com uma estrutura definida, organizados), possibilitando o uso dos mesmos em análises (MANNING; RAGHAVAN; SCHÜTZE, 2010).

A contraposição à essa técnica seria um indivíduo ir manualmente - página por página na internet - copiando e colando ou baixando as informações até ter coletado tudo aquilo que será necessário. Entretanto, essa maneira não é escalonável nem iterativa como o processo de raspagem de dados *web*, que conta com a criação de um algoritmo em determinada linguagem de programação para a simulação dessa captura dos dados que porventura seria manual, podendo extrair grande quantidades de informações de maneira automatizada.

Figura 1 – Fluxo de raspagem de dados *web*



Fonte: Elaboração do autor. Adaptado de Costa (2016, p. 96)

[2] Também conhecido como: (i) coleta de dados *web*, (ii) raspagem *web*.

A maior parte dos dados disponíveis na internet possuem um arcabouço que não muda ao longo do tempo, apesar das informações contidas dentro do mesmo mudarem. O algoritmo de raspagem de dados *web*, interagindo com os sites, localiza as informações, extrai os dados relevantes, estrutura e armazena os mesmos em formato painel, possibilitando futuras consultas.

Isto posto, a Figura 1 nos permite observar o fluxo iterativo do coletor de dados, onde primeiro o algoritmo interpreta a esquematização da página *web*. Por seguinte, iterativamente acessa o site, coletando as informações pré-definidas, extraindo os dados com base na disposição de como estão disponíveis. Essa dinâmica se repercute até que todos os dados foram coletados (COSTA, 2016).

O uso desta técnica nos permite obter um grande volume de dados de forma estruturada, não só poupando tempo, como também permitindo a escalabilidade. Além disso, guarda as informações em um formato organizado, promovendo a reprodutibilidade, um princípio importante no método científico.

3.2.1.2 Análise de sentimentos

Diversas aplicações de análises de sentimento vem sendo desenvolvidas ao longo dos anos, levando em consideração que técnicas de mineração textual permitem a conversão de informações qualitativas em informações quantitativas estruturadas, facilitando a relação com pesquisas de cunho econômico, permitindo incorporá-las na modelagem econométrica. Como tratam Liu et al. (2010), a análise de sentimentos possui como principal objetivo definir técnicas automáticas para se extrair informações subjetivas de textos em linguagem natural para categorizar e identificar sentimentos em textos de forma estruturada para que possa ser utilizado por um sistema de apoio ou tomador de decisão.

A determinação do sentimento está intimamente relacionada ao formato como o texto é escrito, em que a partir de um conjunto de palavras, é possível estabelecer sua tonalidade. A maioria das abordagens de análises de sentimentos baseiam-se em dicionários léxicos, que são listas extensas de palavras que são categorizadas como positivas ou negativas de acordo com sua orientação semântica (LIU et al., 2010).

A criação e validação dessas listas de palavras pode-se definir como um dos métodos mais robustos e confiáveis para a geração de dicionários léxicos, entretanto também é uma das maneiras que mais requisitam tempo. Dessa forma, diversas pesquisas aplicadas no tocante de análise de sentimentos baseia-se em dicionários léxicos previamente criados (HUTTO; GILBERT, 2014).

Um dos dicionários léxicos mais antigos criado foi introduzido por Stone, Dunphy e Smith (1966), denominado *General Inquirer*, onde foi escrito manualmente. Foi desenvolvido como ferramenta para *análise de conteúdo*, técnica usada por cientistas sociais, cientistas políticos e psicólogos para identificação objetiva de características específicas de mensagens. Dentre o total de palavras, 1915 categorizam-se como positivas e 2291 categorizam-se como negativas (HUTTO; GILBERT, 2014).

Ainda, tem-se que a semântica de um documento de texto é de extrema importância para

sua análise, onde existem diversas técnicas para se lidar com a questão da semântica, podendo ser citadas a remoção de *stopwords* e remoção de palavras e caracteres sem relevância. Muitos tipos de palavras não agregam de forma pertinente para o sentido de um documento textual, como preposições e conjunções, aparecendo repetidamente diversas vezes, podendo prejudicar o algoritmos de processamento de textos. Essas palavras são chamadas de *stopwords* e, na etapa de limpeza dos dados, elas são removidas.

3.2.1.2.1 A estatística tf-idf (term frequency-inverse document frequency)

Uma das maneiras costumeiras de representar-se um texto é dividi-lo em *tokens*, isto é, em palavras e outros elementos que deseja-se mapear no processamento de um documento. Conforme enunciam Salton, Wong e Yang (1975), entende-se que um documento é um conjunto de termos indexados que podem possuir pesos a partir da importância desse termo no documento, sendo assim pode-se representar Um documento como um vetor de n dimensões como:

$$D_i = (d_{i1}, d_{i2}, ..., d_{in}) \tag{5}$$

em que D_{in} representa o peso do n-ésimo termo. Sendo assim, tem-se que um *Corpus* de documentos é um conjunto de documentos dispostos em um espaço vetorial, isto é, como se fosse um documento de documentos.

O método da estatística de *term frequency-inverse document frequency* (tf-idf) consiste em considerar que a frequência dos *tokens* de determinado documento importa para a elaboração da análise.

A frequência de determinado termo em um documento é obtido através da quantidade de vezes que o termo surge, onde transforma-se cada documento em uma matriz de tuplas, sendo o primeiro elemento um determinado termo e o segundo elemento a frequência desse mesmo termo no documento. Para se ajustar a frequência dos termos no *Corpus*, conforme apontam Rajaraman e Ullman (2011), é interessante considerar a proporcionalidade das frequências em relação ao tamanho do documento, tem-se então:

$$tf_{ij} = \frac{f_{ij}}{max_k f_{kj}} \tag{6}$$

onde tf é a frequência do termo; i é um termo em determinado documento j; e k é o termo com maior frequência no documento. Ainda conforme Rajaraman e Ullman (2011), tem-se:

$$idf_i = \ln\left(\frac{N}{n_i}\right) \tag{7}$$

em que idf_i é a frequência inversa de um termo i que aparece n vezes em N documentos. Multiplicando a Equação (6) pela Equação (7), chega-se então na estatística tf-idf de um termo i em um documento j.

Em busca de considerar a influência de termos que são menos frequentes em um documento, entretando podem ser relevantes para a análise, calcula-se a frequência inversa do

termo em determinado documento, minimizando a importância de palavras muito comuns e aumentando a importância de palavras mais raras.

3.2.2 Modelagem econométrica

Nesta Subseção serão apresentados os processos da modelagem econométrica. Primeiramente é discursado sobre processos estocásticos e estacionariedade, apresentando os devidos testes, seguido pela apresentação do modelo VAR, que possibilita tanto verificar a causalidade de Granger, como elaborar a IRF. Após isso, são feitos testes de diagnóstico acerca de correlação serial e heterocedasticidade. Por fim, é apresentada a base de dados das atas do Copom e das variáveis macroeconômicas utilizadas na elaboração desta pesquisa.

3.2.2.1 Processos aleatórios e séries estacionárias

Conforme enunciado em Gujarati e Porter (2011), define-se um processo aleatório como uma coleção de variáveis estocásticas ordenadas no tempo. Sendo assim, sugere-se que dados macroeconômicos, genericamente, são categorizados como processos aleatórios (denominados também como séries temporais). Da mesma maneira que se utiliza amostras de dados para efetuar inferências sobre uma determinada população, no campo de séries temporais, utiliza-se a realização do processo estocástico em questão para extrações inferenciais (GUJARATI; PORTER, 2011).

De modo geral, caso a série temporal se enquadrar como estacionária, sua média e sua variância deverão ser constantes ao longo do tempo. Além disso, o valor da covariância entre dois períodos de tempo deverá depender somente da defasagem entre eles; de outro modo, a série não será estacionária. Tal processo estocástico é conhecido na literatura como fracamente estacionário (GUJARATI; PORTER, 2011; MORETTIN; BUSSAB, 2017).

A partir de um modelo generalizado, pode-se representar diversas séries temporais e suas características. Sendo Y_t um processo estocástico fracamente estacionário, como segue:

$$Y_t = \beta_0 + \beta_1 t + \beta_2 Y_{t-1} + \mu_t \tag{8}$$

tem-se que β_0 é o coeficiente constante (de intercepto); β_1 é o coeficiente relacionado à medição do tempo t (de tendência); β_2 é o coeficiente do termo autorregressivo (AR), que é a série no período anterior $(t-1)$; e μ_t é um termo de erro do tipo ruído branco, a parte que o modelo não consegue captar (GUJARATI; PORTER, 2011).

A depender dos valores obtidos para os estimadores de β_k, pode ser que a série temporal possua características de estacionariedade ou não. Supondo que μ_t seja um termo de erro do tipo ruído branco, com média igual à zero e variância σ^2 (constante); tem-se então que a série Y_t é um passeio aleatório puro (sem deslocamento), conforme:

$$Y_t = Y_{t-1} + \mu_t \tag{9}$$

em que $\beta_0 = \beta_1 = 0$ e $\beta_2 = 1$. Constata-se que a variância da Equação (9) crescerá indefinitivamente conforme o aumento do tempo t, uma vez que $var(Y_t) = \sigma^2 t$, violando assim uma das condições de estacionariedade fraca (GUJARATI; PORTER, 2011).

Subtraindo Y_{t-1} de ambos os lados da Equação (9), chega-se em:

$$\begin{aligned}(Y_t - Y_{t-1}) &= (Y_{t-1} - Y_{t-1}) + \mu_t \\ (Y_t - Y_{t-1}) &= \mu_t \\ \Delta Y_t &= \mu_t \end{aligned} \tag{10}$$

onde Δ é chamado de operador de diferença, isto é, o valor da série no tempo t subtraído do valor da série no tempo $t-1$. Assim como demonstrado por Gujarati e Porter (2011), as primeiras diferenças de séries temporais de um passeio aleatório são estacionárias, e neste caso, são integradas de ordem 1, podendo ser representadas por $I(1)$, ou genéricamente como integradas de ordem $d \rightarrow I(d)$.

3.2.2.1.1 Teste de raiz unitária de Dickey-Fuller Aumentado (ADF)

Em modelos de séries temporais, a unidade de raiz é uma importante característica dos processos de realização da série que, caso não for adequadamente tratada, pode causar problemas de inferência. Reescrevendo a Equação (9) como:

$$Y_t = \rho Y_{t-1} + \mu_t \qquad (-1 \leq \rho \leq 1) \tag{11}$$

tem-se μ_t como termo de erro ruído branco; e ρ como coeficiente de autocorrelação entre Y_t e Y_{t-1}. Caso $\rho = 1$, a Equação (11) será um processo de passeio aleatório e se terá problema de raiz unitária, ou seja, problema de estacionariedade na série. Subtraindo Y_{t-1} de ambos os lados da Equação (11) e aplicando o operador de diferença, é possível chegar em:

$$\begin{aligned} Y_t - Y_{t-1} &= \rho Y_{t-1} - Y_{t-1} + \mu_t \\ \Delta Y_t &= (\rho - 1) Y_{t-1} + \mu_t \\ \Delta Y_t &= \delta Y_{t-1} + \mu_t \end{aligned} \tag{12}$$

em que $\delta = (\rho - 1)$.

Para se efetuar o teste de raiz unitária, deve-se estimar os parâmetros da Equação (12), testando a hipótese nula (H_0)) contra a hipótese alternativa (H_1), como segue:

$$H_0 : \delta = 0 \tag{13}$$

$$H_1 : \delta < 0 \tag{14}$$

Caso o resultado do teste for $\delta = 0$, terá-se então um $\rho = 1$, isto é, raiz unitária. Dessa forma, o processo estócastico testado trata-se de uma série não estacionária. Como demonstrado por Dickey e Fuller (1979) e abordado em Gujarati e Porter (2011), o valor estimado de t para o

coeficiente Y_{t-1} da Equação (12) segue a estatística τ (tau). Os valores fundamentais da estatística τ foram computadas a partir de simulações de Monte Carlo, onde o teste ficou conhecido como teste de Dickey-Fuller. Com o objetivo de englobar diversas possibilidades, a estimação do teste é feito de três formas:

$$\Delta Y_t = \delta Y_{t-1} + \mu_t \quad \textbf{(passeio aleatório)} \tag{15}$$

$$\Delta Y_t = \beta_0 + \delta Y_{t-1} + \mu_t \quad \textbf{(passeio aleatório com deslocamento)} \tag{16}$$

$$\Delta Y_t = \beta_0 + \beta_1 t + \delta Y_{t-1} + \mu_t \quad \textbf{(passeio aleatório com deslocamento e tendência)} \tag{17}$$

A estatística tau calculada ($\tau_{calculada}$) pode ser obtida ao dividir o delta estimado ($\hat{\delta}$) pelo seu desvio-padrão ($\sigma_{\hat{\delta}}$), e caso seu módulo for maior do que o módulo da estatística tau crítica ($|\tau_{critica}|$), se rejeita a hipótese nula (H_0) de que não há estacionariedade, caso contrário, trata-se de uma série estacionária.

Para as situações em que os resíduos da série temporal (μ_t) são correlacionados, ou seja, problema de autocorrelação, utiliza-se do teste de Dickey-Fuller Aumentado (ADF), em que trata de adicionar valores defasados da variável Y_t na estimação, chegando-se na seguinte equação:

$$\Delta Y_t = \beta_0 + \beta_1 t + \delta Y_{t-1} + \sum_{p=1}^{n} \alpha_p \Delta Y_{t-p} + \varepsilon_t \tag{18}$$

em que ε_t é um termo de ruído branco e $p_{1,\ldots,n}$ refere-se ao número de defasagens.

O objetivo é adicionar termos suficientes na Equação (18) de maneira que o erro ε_t seja serialmente não correlacionado, para que seja possível obter-se uma estimativa não viesada de δ, que é o coeficiente defasado de Y_{t-1}. Embora haver mais parâmetros no teste de Dickey-Fuller Aumentado, ainda é testado se $\delta = 0$, seguindo a mesma distribuição assintótica da estatística de Dickey-Fuller, de forma que os mesmos valores de τ podem ser utilizados (GUJARATI; PORTER, 2011).

Dada determinada ordem de defasagem máxima (ρ_{maxima}), a partir do critério de informação de Akaike (AIC) ou critério Bayesiano de Schwarz (BIC), pode-se selecionar a melhor extensão de defasagens para o modelo. Conforme apresentado em Bueno (2011), a escolha do número máximo máximo de extensões dá-se pelo número inteiro obtido atráves de:

$$\rho_{maxima} = 12 \left(\frac{N}{100} \right)^{\frac{1}{4}} \tag{19}$$

em que N é o tamanho da amostra. Dado um nível de significância (α) de $5,00\%$, exibe-se na Tabela 1 os valores críticos para a estatística τ.

Como ponto de observação, na primeira coluna, PA refere-se à passeio aleatório (sem constante), PAD refere-se à passeio aleatório com deslocamento (com constante) e PADT refere-se à passeio aleatório com deslocamento e tendência (com constante e tendência).

Tabela 1 – Valores críticos para a estatística τ dado um $\alpha = 5,00\%$

Modelo	N = 100	N = 250	N = 500	N = ∞
PA	-1,95	-1,95	-1,95	-1,95
PAD	-2,89	-2,88	-2,87	-2,86
PADT	-3,45	-3,43	-3,42	-3,41

Fonte: Elaboração do autor. Adaptado de Fuller (1976, p. 373).

3.2.2.1.2 Teste de raiz unitária de Phillips-Perron (PP)

Outro teste conhecido é o teste desenvolvido por Phillips e Perron (1988), em que propõem um método alternativo e não-paramétrico de controlar a autocorrelação ao testar-se a raiz unitária, permitindo que seja consistente mesmo havendo variáveis dependentes defasadas e correlação serial nos resíduos. Sendo assim, o teste de Phillips-Perron (PP) faz com que seja desnecessária a especificação de um modelo com ordem suficientemente autorregressivo para a eliminação do problema de autocorrelação (BUENO, 2011).

Considerando a seguinte alternativa de regressão ΔY_t e a sua respectiva estatística Z_t associada, tem-se:

$$\Delta Y_t = \omega + \alpha Y_{t-1} + \mu_t \qquad \rightarrow \qquad Z_{t,\omega} \tag{20}$$

em que Δ é o operador da diferença e μ_t é um processo estacionário. Estima-se então o parâmetro $\hat{\alpha}$ de interesse:

$$\hat{\alpha} = \frac{\sum_{t=1}^{N}(Y_{t-1} - \overline{Y}_{-1})(Y_t - \overline{Y})}{\sum_{t=1}^{N}(Y_{t-1} - \overline{Y}_{-1})^2} - 1 \tag{21}$$

onde $\overline{Y}$ representa a média, sendo possível estimar a variância populacional $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \frac{\sum_{t=1}^{N}(\Delta Y_t - \hat{\omega} - \hat{\alpha} Y_{t-1})^2}{N} \tag{22}$$

e a partir da Equação (22) pode-se calcular a variância de longo prazo $\hat{\upsilon}^2$:

$$\hat{\upsilon}^2 = \hat{\sigma}^2 + \frac{2}{N}\sum_{j=1}^{M}\phi\left(\frac{j}{M+1}\right)\sum_{t=j+1}^{N}\hat{\mu}_t\hat{\mu}_{t-j} \tag{23}$$

podendo-se chegar afinal na estatística de PP, dada por:

$$\hat{Z}_{t,\omega} = \hat{\tau}_\omega\left(\frac{\hat{\sigma}}{\hat{\upsilon}}\right) - \frac{1}{2}\left(\frac{\hat{\upsilon}^2 - \hat{\sigma}^2}{\hat{\upsilon}\sqrt{N^{-2}\sum_{t=1}^{N}Y_{t-1}^2}}\right) \tag{24}$$

em que na variância de longo prazo estão inclusas todas as autocovariâncias do processo μ_t.

A interpretação do teste de PP é a mesma do que a do teste de ADF, uma vez que possuem a mesma distribuição assintótica, tendo-se que a hipótese nula (H_0) é de que a série apresenta raiz unitária contra a hipótese alternativa (H_1) de que a série é estacionária (BUENO, 2011; GUJARATI; PORTER, 2011).

3.2.2.1.3 Teste de raiz unitária de Kwiatkowski–Phillips–Schmidt–Shin (KPSS)

Devido à facilidade de implementação e bom funcionamento dos testes de raiz unitária ADF e PP, eles são popularmente utilizados. Entretanto, tanto o teste ADF quanto o teste de PP possuem como H_0 a presença de raiz unitária, isto é, a não estacionariedade. De qualquer maneira, a literatura acerca do tema frequentemente aceita a H_1, caso inexistam evidências contrárias a H_0, ainda mais quando trata-se de séries macroeconômicas. Devido a esse fator, faz-se de bom uso a aplicação de um teste que tenha como H_0 a estacionariedade e outro que a tenha como H_1 (KENNEDY, 2008).

O teste desenvolvido por Kwiatkowski et al. (1992), conhecido como teste KPSS (mnemônico dos autores), possui a estacionariedade da série como hipótese nula (H_0), frente a não estacionariedade como hipótese alternativa (H_1), como segue:

$$H_0 : Y_t \sim I(0) \tag{25}$$

$$H_1 : Y_t \sim I(1) \tag{26}$$

O objetivo é usar o teste como complementar aos testes tradicionais, distinguindo assim a raiz unitária de séries cujos dados não são suficientemente conclusivos (como séries macroeconômicas), caracterizando-se assim mais assertivamente a série (BUENO, 2011).

O teste KPSS decompõe a série em questão em três partes, conforme:

$$Y_t = \eta t + \phi_t + \varepsilon_t \tag{27}$$

onde ηt é o elemento de tendência determinística, ϕ_t é o elemento de passeio aleatório e ε_t é o elemento de erro estacionário. É efetuado então um processo de soma parcial dos resíduos, como:

$$S_t = \sum_{t=1}^{N} \hat{e}_t \tag{28}$$

no qual $\hat{e}_t$ são os erros da Equação (27). Tendo a variância de longo prazo ($\hat{\upsilon}^2$) como definida na Equação (22), chega-se à estatística do teste KPSS, dada por:

$$KPSS = \sum_{t=1}^{N} \frac{S_t^2}{N^2 \hat{\upsilon}^2} \tag{29}$$

Caso Y_t for um processo estacionário, então $S_t \sim I(1)$ e o numerador da estatística KPSS será um estimador da variância de S_t, tendo assim um limite assintótico. Já o denominador observado na Equação (29) assegura que a distribuição seja livre de ruídos. Na outra mão, caso $Y_t \sim I(1)$, o numerador crescerá ilimitadamente, fazendo a estatística KPSS ficar muito elevada. (BUENO, 2011).

Compara-se então a estatística calculada com os valores críticos a determinado nível de significância, sendo que se o valor calculado for maior que o valor crítico, rejeita-se a H_0 de

estacionariedade. Os valores críticos para a realização do teste de KPSS, que foram calculados e apresentados em Kwiatkowski et al. (1992), são utilizados como parâmetros de comparação.

Já em relação a ordem de defasagem máxima (ρ_{maxima}) na regressão, como apontado em Trapletti e Hornik (2022), será utilizado o número inteiro obtido através de:

$$\rho_{maxima} = 4\left(\frac{N}{100}\right)^{\frac{1}{4}} \tag{30}$$

alternativamente à Equação (19).

3.2.2.2 Vetores autorregressivos (VAR)

A modelagem econômica em geral é caracterizada por haver diversas variáveis, sendo assim, modelos univariados são limitados para expressar modelos econômicos. O VAR permite que sejam expressos modelos econômicos e se estimem seus parâmetros. Genericamente falando, pode-se definir um modelo autorregressivo de ordem p, isto é, um VAR(p) como:

$$AY_t = B_0 + \sum_{i=1}^{p} B_i Y_{t-i} + \varepsilon_t \tag{31}$$

onde A representa uma matriz $n \times n$ que define as restrições contemporâneas; Y_t é um vetor $n \times 1$ das variáveis endógenas no período t; B_0 é um vetor de constantes $n \times 1$; B_i é uma matriz $n \times n$ dos parâmetros referentes ao vetor Y_{t-i}, que são as variáveis endógenas defasadas i vezes, possuindo dimensão $n \times 1$; e por fim ε_t é um vetor $n \times 1$ de erros do tipo ruído branco (BUENO, 2011).

A Equação (31) demonstra as relações entre as variáveis endógenas em sua forma estrutural, onde os choques ε_t são chamados de choques estruturais, uma vez que afetam de forma individual cada uma das variáveis endógenas. Por causa disso, o modelo costumeiramente é estimado na sua forma reduzida, multiplicando a equação por A^{-1}, conforme:

$$Y_t = A^{-1}B_0 + \sum_{i=1}^{p} A^{-1}B_i Y_{t-i} + A^{-1}\varepsilon_t \tag{32}$$

onde $C \equiv A^{-1}B_0$; $\phi_i \equiv A^{-1}B_i$ e $e_t \equiv A^{-1}\varepsilon_t$, chegando-se então em:

$$Y_t = C + \sum_{i=1}^{p} \phi_i Y_{t-i} + e_t \tag{33}$$

Convenientemente, conforme Enders (2009), será considerado o modelo disposto na Equação (33) como um modelo bivariado com defasagem de ordem igual a 1, chegando no sistema:

$$y_t = b_{10} - a_{12}z_t + b_{11}y_{t-1} + b_{12}z_{t-1} + \sigma_y\varepsilon_{yt} \tag{34}$$

$$z_t = b_{20} - a_{21}y_t + b_{21}y_{t-1} + b_{22}z_{t-1} + \sigma_z\varepsilon_{zt} \tag{35}$$

Nota-se que as variáveis são mutuamente influenciadas umas pelas outras, tanto com seus valores atuais quanto com seus valores defasados. O VAR não pode ser diretamente estimado, considerando que as variáveis atuais z_t e y_t são isoladamentes correlacionadas aos erros ε_{yt} e ε_{zt}, respectivamente. Isso acontece, pois cada uma dessas variáveis depende do comportamento da outra, tendo como objetivo do VAR encontrar a trajetória da variável de interesse a partir da efetuação de um choque nesses erros (BUENO, 2011).

3.2.2.2.1 Causalidade de Granger

A causalidade no sentido de Granger acontece quando uma variável é capaz de prever outra variável e em sob quais condições, isto é, quando é possível utilizar uma variável, seja defasada ou não, para prever outra variável em determinado período de tempo. O objetivo é determinar se y ajuda a prever z e caso isso não aconteça, tem-se que y não Granger-causa z. O teste feito então é:

$$z_t = \phi_{20} + \sum_{i=1}^{p} \phi_{i,21} y_{t-i} + \sum_{i=1}^{p} \phi_{i,22} z_{t-i} + e_{2t} \tag{36}$$

onde a seguinte hipótese nula (H_0) é testada frente a hipótese alternativa (H_1):

$$H_0 : \phi_{i,21} = 0 \qquad \text{(ausência de Granger-causalidade)} \tag{37}$$

$$H_1 : \phi_{i,21} \neq 0 \qquad \text{(presença de Granger-causalidade)} \tag{38}$$

em que a estatística do teste (S_1) é dada por:

$$S_1 = \frac{\frac{(e_\gamma^2 - e_\mu^2)}{p}}{\frac{e_\mu^2}{N-2p-1}} \qquad \rightarrow \qquad F(p; N-2p-1) \tag{39}$$

tal qual γ representa restrito e μ representa irrestrito.

Caso $S_1 > F$ dado um nível de significância (α), rejeita-se a hipótese nula de que y não Granger-causa z, ou seja, se conclui que existe uma relação de causalidade no sentido de Granger (BUENO, 2011; GUJARATI; PORTER, 2011).

Para se definir a quantidade ideal de defasagens a serem inseridas dentro do sistema de equações, pode-se utilizar os critérios de informação, em que atribuem penalidade conforme aumenta-se o número de regressores, por impactar na soma dos resíduos do modelo. Estes critérios então minimizam uma função que baseia-se nos resíduos da regressão estimada, penalizada pela quantidade de regressores que foram utilizados (BUENO, 2011).

O modelo mais ideal será o com maior parcimônia, isto é, com menor número de parâmetros, satisfazendo com que os resíduos sejam os menores possíveis. Geralmente, o critério de especificação possui o seguinte formato:

$$CE = \ln \hat{\sigma}^2(N) + c_N \varphi(N) \tag{40}$$

onde $\hat{\sigma}^2(N) = \frac{\sum_{t=1}^{N} \hat{\varepsilon}_t^2}{N}$ é a variância estimada dos resíduos; c_N é o número de parâmetros estimados; e $\varphi(N)$ é a ordem do processo, que penaliza a falta de parcimoniosidade. Dessa forma, existem três principais critérios de informação, onde a estatística de Critério de Informação Bayesiano (BIC) se dá por:

$$BIC(p,q) = \ln \hat{\sigma}^2 + n\frac{lnN}{N} \tag{41}$$

no qual $n = p+q$ caso o modelo não possuir constante ou $n = p+q+1$ caso houver constante.

Outro critério é a estatística de Critério de Informação de Akaike (AIC), em que:

$$AIC(p,q) = \ln \hat{\sigma}^2 + n\frac{2}{N} \tag{42}$$

Por fim, a estatística Hannan-Quinn (HQ) é conforme segue:

$$HQ(p,q) = \ln \hat{\sigma}^2 + n\frac{2}{N}\ln(\ln N) \tag{43}$$

A primeira parte da Equação (40) mede a adequação do processo, em que quanto menor a variância dos resíduos, melhor. Entretanto, uma redução de variância que foi obtida ao introduzir mais parâmetros é penalizada pela segunda parte da equação. Dessa maneira, deseja-se o menor AIC, BIC ou HQ possível (BUENO, 2011).

3.2.2.2.2 Função de impulso-resposta (IRF)

Assim como existe a possibilidade de se retratar um modelo autorregressivo como seu componente de média móvel, pode-se representar um modelo VAR(p) como um Vetor de Média Móvel (VMA), isto é, um VMA(∞). Dessa forma, como abordado em Sims (1980), elabora-se projeções distintas a partir da efetuação de choques nas variáveis introduzidas no modelo VAR(p). Considerando uma abordagem matricial para um modelo de VAR(1), tem-se:

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} + \begin{bmatrix} \phi_{11} y_{1,t-1} \\ \phi_{12} y_{1,t-1} \end{bmatrix} + \begin{bmatrix} \phi_{21} y_{2,t-1} \\ \phi_{22} y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \mu_{1t} \\ \mu_{2t} \end{bmatrix} \tag{44}$$

no qual, reescrevendo a Equação (44), chega-se em:

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \end{bmatrix} + \sum_{i=0}^{\infty} \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix}^i \begin{bmatrix} \mu_{1,t-i} \\ \mu_{2,t-i} \end{bmatrix} \tag{45}$$

em que a Equação (45) representa $\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix}$ em termos de $\begin{bmatrix} \mu_{1t} \\ \mu_{2t} \end{bmatrix}$. Entretranto, reescrevendo em relação a $\begin{bmatrix} \mu_{y_1 t} \\ \mu_{y_2 t} \end{bmatrix}$, conforme aborda Enders (2009), obtém-se:

$$\begin{bmatrix} \mu_{1t} \\ \mu_{2t} \end{bmatrix} = \frac{1}{1 - b_{12} b_{21}} \begin{bmatrix} 1 & -b_{12} \\ -b_{21} & 1 \end{bmatrix} \begin{bmatrix} \mu_{y_1 t} \\ \mu_{y_2 t} \end{bmatrix} \tag{46}$$

onde ao combinar a Equação (45) e a Equação (46), terá-se:

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \end{bmatrix} + \frac{1}{1-b_{12}b_{21}} \sum_{i=0}^{\infty} \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix}^i \begin{bmatrix} 1 & -b_{12} \\ -b_{21} & 1 \end{bmatrix} \begin{bmatrix} \mu_{y_1 t} \\ \mu_{y_2 t} \end{bmatrix} \tag{47}$$

em que, para simplificar, define-se a matriz Θ_i como:

$$\Theta_i = \frac{A_1^i}{1-b_{12}b_{21}} \begin{bmatrix} 1 & -b_{12} \\ -b_{21} & 1 \end{bmatrix} \tag{48}$$

e com isso, a representação de média móvel introduzida na Equação (45) pode ser reescrita conforme:

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \end{bmatrix} + \sum_{i=0}^{\infty} \begin{bmatrix} \Theta_{11}(i) & \Theta_{12}(i) \\ \Theta_{21}(i) & \Theta_{22}(i) \end{bmatrix}^i \begin{bmatrix} \mu_{y_1 t} \\ \mu_{y_2 t} \end{bmatrix} \tag{49}$$

sendo assim, de maneira mais compacta:

$$Y_t = \overline{Y}_t + \sum_{i=0}^{\infty} \Theta_i \mu_{t-i} \tag{50}$$

em que os elementos da matriz Θ_i são denominados multiplicadores de impacto de um choque sobre as variáveis endógenas. Dessa forma, o impacto total de um choque de $\mu_{y_1 t}$ sobre Yt_{t+n} é dado por $\sum_{i=0}^{n} \Theta_{11}(i)$, sendo que caso y_1 e y_2 forem estacionárias, os valores de $\Theta_{xz}(i)$ convirjirão para zero conforme aumentar-se i, implicando que os choques não possuem efeitos permanentes em séries estacionárias (BUENO, 2011).

3.2.2.3 Autocorrelação serial

Define-se autocorrelação serial como a correlação entre o termo de erro μ em determinado período t e o termo de erro μ em seu período anterior $t-1$, isto é, a relação de μ_t e μ_{t-1}. Levando em consideração uma série temporal na forma:

$$Y_t = \beta_0 + \beta_1 X_t + \mu_t \tag{51}$$

supõe-se que o termo de erro μ_t é gerado pelo mecanismo:

$$\mu_t = \lambda \mu_{t-1} + \varepsilon_t \qquad (-1 < \lambda < 1) \tag{52}$$

em que λ é o coeficiente de autocovariância e ε_t é o termo de erro estocástico, ao qual atende à hipótese padrão dos Mínimos Quadrados Ordinários (MQO):

$$\begin{aligned} &E(\varepsilon_t) = 0 \\ &var(\varepsilon_t) = \sigma^2 \\ &cov(\varepsilon_t, \varepsilon_{t+s}) = 0 \qquad (s \neq 0) \end{aligned} \tag{53}$$

sendo que por possuir tais propriedades, o termo de erro ε_t é denominado de ruído branco.

A Equação (52) é conhecida como processo autorregressivo de primeira ordem de Markov e aponta que o valor do erro no período atual t depende de uma proporção λ do seu valor no período anterior, somado de um termo de erro totalmente aleatório ε_t (GUJARATI; PORTER, 2011).

3.2.2.3.1 Teste de Durbin-Watson (DW)

Para verificação da existência - e se cabível o grau - de autocorrelação, um popular teste que é empregado é a estatística d introduzida por Durbin e Watson (1950), definida como:

$$d = \frac{\sum \hat{\mu}_t^2 + \sum \hat{\mu}_{t-1}^2 - 2\sum \hat{\mu}_t \hat{\mu}_{t-1}}{\sum \hat{\mu}_t^2} \tag{54}$$

considerando que $\sum \hat{\mu}_t^2 \approx \sum \hat{\mu}_{t-1}^2$, tem-se então:

$$d \approx 2\left(1 - \frac{\sum \hat{\mu}_t \hat{\mu}_{t-1}}{\sum \hat{\mu}_t^2}\right) \tag{55}$$

em que $\approx$ significa aproximadamente. Definindo:

$$\hat{\lambda} = \frac{\sum \hat{\mu}_t \hat{\mu}_{t-1}}{\sum \hat{\mu}_t^2} \tag{56}$$

pode-se expressar a Equação (55) como:

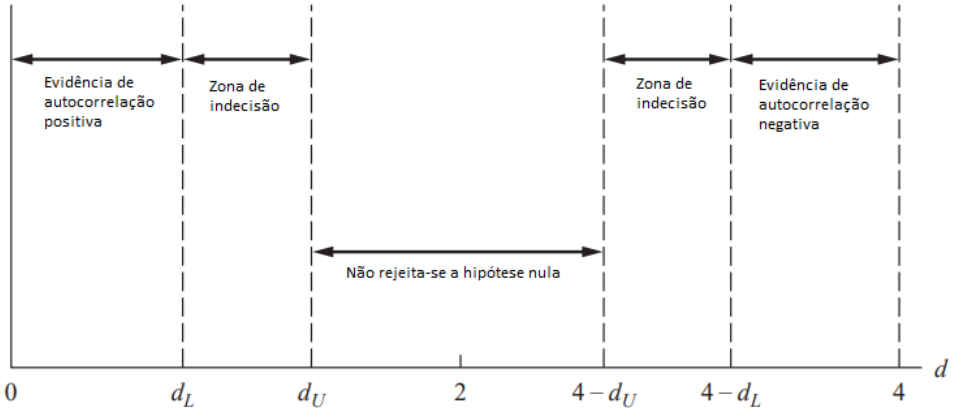
$$d \approx 2(1 - \hat{\lambda}) \tag{57}$$

onde $\hat{\lambda}$ é o coeficiente de covariância estimado da Equação (52).

Considerando que $-1 < \lambda < 1$, a Equação (57) denota que $0 \leq d \leq 4$. Valores próximos de zero para a estatística d indicam correlação positiva, ficando próximo de 2 de acordo com menor evidência de autocorrelação.

A Figura 2 apresenta as regras de decisão determinadas por Durbin e Watson (1950), em que d_L denota o limite inferior da estatística d e d_U denota o limite superior da estatística d.

Figura 2 – Zonas da estatística d de Durbin-Watson



Fonte: Elaboração do autor. Adaptado de Gujarati e Porter (2011).

É testado então a hipótese nula (H_0) contra a hipótese alternativa (H_1), tal qual:

$$H_0 : \text{nenhuma autocorrelação, positiva ou negativa} \tag{58}$$

$$H_1 : \text{presença de autocorrelação} \tag{59}$$

Salienta-se que d_L e d_U dependem do tamanho da amostra e da quantidade de variáveis explicativas do modelo; e caso $d_U < d < (4 - d_U)$, não rejeita-se a hipótese nula (H_0) de ausência de autocorrelação serial (GUJARATI; PORTER, 2011).

3.2.2.3.2 Teste de Breusch-Godfrey (BG)

Conforme apontado em Gujarati e Porter (2011), para se evitar algumas armadilhas do teste d de Durbin-Watson, Breusch (1978) e Godfrey (1978) desenvolveram um teste de autocorrelação que é genérico, não permitindo regressores não estocásticos (como valores defasados do regressando); esquemas autorregressivos de ordem superior; e médias móveis simples ou de ordem mais elevada de resíduos do tipo ruído branco, como ε_t na Equação (52).

O teste BG, também conhecido como teste do Multiplicador de Lagrange (LM), pressupõe que o termo de erro μ_t da Equação (51) siga um esquema autorregressivo de ordem p, como evidenciado em:

$$\mu_t = \lambda_1 \mu_{t-1} + \lambda_2 \mu_{t-2} + [...] + \lambda_p \mu_{t-p} + \varepsilon_t \tag{60}$$

em que ε_t é um termo de erro do tipo ruído branco, tal qual apresentado em (53). O teste LM baseia-se em efetuar a regressão original, obtendo-se assim os resíduos estimados $\hat{\mu}_t$ e após isso introduzir os valores defasados dos resíduos como regressores adicionais no modelo:

$$\hat{\mu}_t = \alpha_0 + \alpha_1 X_t + \hat{\lambda}_1 \hat{\mu}_{t-1} + \hat{\lambda}_2 \hat{\mu}_{t-2} + [...] + \hat{\lambda}_p \hat{\mu}_{t-p} + \varepsilon_t \tag{61}$$

Tendo N como o tamanho da amostra e R^2 como o coeficiente de determinação da Equação (61), Breusch (1978) e Godfrey (1978) demonstraram que para amostras grandes ($N \to \infty$), a relação $(N-p)R^2$ seguirá uma distribuição qui-quadrado com p graus de liberdade, sendo assim:

$$(N-p)R^2 \sim \chi^2_p \tag{62}$$

Realiza-se então a testagem de hipóteses a determinado nível de significância (α), tendo a hipótese nula (H_0) a ser testada como segue:

$$H_0 : \lambda_1 = \lambda_2 = [...] = \lambda_p = 0 \quad \rightarrow \quad \text{ausência de autocorrelação nos resíduos} \tag{63}$$

Desse modo, caso $(N-p)R^2$ exceder o valor crítico ao nível determinado de α, rejeita-se a H_0, em que pelo menos um dos λ_i na Equação (61) é estatísticamente significantemente diferente de zero (GUJARATI; PORTER, 2011; BUENO, 2011).

O fato de uma série possuir autocorrelação faz com que a eficiência dos esimadores de MQO seja afetada, subestimando assim a variância - e por consequência, os erros padrão - do modelo. Contudo, estimadores MQO permanecem não tedenciosos, convergindo assim para seu valor populacional verdadeiro (GUJARATI; PORTER, 2011).

3.2.2.4 Heterocedasticidade da variância dos resíduos

Não só autocorrelação serial é um problema, como os resíduos de determinada regressão podem ser heterocedásticos, isto é, sua variância não será constante ao longo do tempo (indiferentemente dos valores assumidos pela(s) variável(is) independente(s)). Os fatos de linearidade

e parâmetros do modelo não serem tendenciosos não são alterados pela aparição de heterocedasticidade nas estimativas, apesar disso, caso tiver-se resíduos que forem heterocedásticos, terá-se perda de eficiência, uma vez que distorcem as variâncias do modelo e dos estimadores (GUJARATI; PORTER, 2011).

Assim como é levantado em Gujarati e Porter (2011), diversos testes de heterocedasticidade foram propostos com o decorrer dos anos, podendo ser citados o teste de Park (1966), o teste de Glejser (1969), o teste de Goldfeld e Quandt (1972), o teste de Breusch e Pagan (1979) e o teste geral de White (1980).

3.2.2.4.1 Teste ARCH-LM

Com o VAR(p) estimado, assim como aponta Pfaff (2008), é de extrema importância verificar se os resíduos atendem às hipóteses norteadoras do modelo, ou seja, além de checar a autocorrelação serial, é necessário atentar-se para a heterocedasticidade da variância dos resíduos.

O teste ARCH-LM, conforme Bueno (2011), identifica sinais de heterocedasticidade condicional. Considerando a regressão como segue:

$$\hat{\varepsilon}_t^2 = \beta_1 \hat{\varepsilon}_{t-1}^2 + \beta_2 \hat{\varepsilon}_{t-2}^2 + [...] + \beta_i \hat{\varepsilon}_{t-i}^2 + \mu_t \tag{64}$$

em que testa-se a seguinte hipótese nula (H_0) contra a hipótese alternativa (H_1):

$$H_0 : \beta_i = 0 \qquad \text{(ausência de autocorrelação)} \tag{65}$$

$$H_1 : \text{pelo menos um } \beta_i \neq 0 \quad \text{(presença de autocorrelação)} \tag{66}$$

onde a estatística segue uma distribuição qui-quadrado χ_i^2:

$$\text{ARCH-LM}_i = NR^2 \quad \rightarrow \chi_i^2 \tag{67}$$

em que N é o tamanho da amostra e R^2 é o coeficiente de determinação e neste caso, rejeita-se H_0 caso o valor da estatística caculada for maior que o valor tabelado.

3.3 BASE DOS DADOS

O período considerado na análise foi do ínicio de 2006 até meados de 2022, contemplando da ata número 116 até a ata número 246 (resultando em um total de 131 documentos), uma vez que anteriormente a esse período, as atas eram emitidas mensalmente, além de que em 2002 foram publicadas treze atas, ao invés de doze.

3.3.1 Atas do Copom

Conforme abordado na parte metodológica, as informações referentes às atas do Copom foram adquiridas por meio da utilização de técnicas de raspagem de dados *web*, que abrange a elaboração de linhas de código em determinada linguagem de programação para detalhar ao

computador o que deve ser feito (imitando o manuseio de baixar as atas no site do BCB, porém automaticamente), a partir do site do BCB[1], sendo elas na versão em inglês para facilitar o tratamento das palavras presentes nas atas com as dispostas no dicionário léxico utilizado no presente estudo. Todos os processos e tratamentos dos dados e documentos foram feitos por meio do *software* de código aberto R (R Core Team, 2020).

Primeiramente, por meio do pacote *jsonlite* desenvolvido por Ooms (2014), descobre-se e salva-se informações referentes a: data, número, URL e texto das reuniões. Com essa etapa realizada, utilizando-se do pacote *pdftools* de Ooms (2022) é iniciado o processo iterativo de coleta dos textos de cada ata. Dessa forma, tem-se os dados brutos das atas do Copom.

A partir dos dados brutos reunidos, um processo de tratamento e limpeza das informações é iniciado, onde são removidos quebras de linha, espaçamentos indesejados, caracteres especiais ASCII, pontuações, números e; todas as palavras presentes nos textos são reduzidas a sua forma minúscula, para melhor unicidade. Além disso, é feito um processo de *tokenização*, isto é, divide-se o texto completo por unidades menores de texto (geralmente palavras) para posterior análise (SILGE; ROBINSON, 2017).

Tabela 2 – Palavras *stopwords* no compilado *tidytext*

a	afterwards	it	there	its
a's	again	its	when	itself
able	against	itself	where	just
about	ain't	they	why	keep
above	all	them	how	keeps
according	allow	their	all	kind
accordingly	allows	theirs	any	knew
across	almost	themselves	both	know
actually	alone	what	each	known
after	along	which	few	knows

Fonte: Elaboração do autor.

Para cada documento de corpo textual das atas do Copom, foi aplicado o processo de *stopwords* (remoção de palavras comuns), provenientes de uma compilação presente no pacote *tidytext* reunido por Silge e Robinson (2016), que contém 1149 palavras de três diferentes dicionários léxicos: *SMART*, *snowball* e *onix*. Para fins de exemplificação, algumas dessas *stopwords* podem ser observadas na Tabela 2.

Não obstante, palavras adicionais que para a análise desta pesquisa julgou-se irrelevante foram desconsideradas, tais quais meses do ano (como *'january'*, *'month'*, *'monthly'*), nomes próprios (como *'aloisio'*, *'guardado'*, *'rumenos'*), numerais (como *'one'*, *'two'*, *'three'*). Também desconsiderou-se palavras com erros de interpretação devido a quebras de linha (como *'nmonetary'*, *'nexecutive'*, *'npandemic'*).

[1] https://www.bcb.gov.br/en/publications/copomminutes

3.3.2 Variáveis macroeconômicas

Os dados macroeconômicos utilizados para comparação do índice de sentimentos foram reunidos do BCB e do Instituto Brasileiro de Geografia e Estatística (IBGE).

Conforme inicialmente propõe Bloom (2009) e posteriormente Ferreira et al. (2017) para o Brasil, baseando-se nesses estudos e considerando algumas alterações, o trabalho englobou as seguintes variáveis:

a) para representar a taxa básica de juros da economia brasileira, usou-se taxa Selic Meta, definida nas reuniões do Copom a cada quarenta e cinco dias;
b) para a medida de inflação, o Índice Nacional de Preços ao Consumidor Amplo (IPCA), que contempla uma cesta de bens e serviços referentes ao consumo pessoal das famílias de um até quarenta salários mínimos das principais regiões metropolitanas do país;
c) para a atividade econômica, usou-se do Índice de Atividade Econômica do Banco Central (IBC-Br), que serve como indicador agregado de atividade econômica com frequência mensal;
d) para representar a produção industrial, utilizou-se a Pesquisa Industrial Mensal de Produção Física (PIM-PF)

A relação resumida das variáveis, sua descrição, seu código nas bases de dados oficiais e sua fonte podem ser observadas na Tabela 3:

Tabela 3 – Resumo das variáveis macroeconômicas

Variável	Descrição	Código	Fonte
Selic	Taxa básica de juros	4189	BCB
IBC-Br	Atividade econômica	24364	BCB
IPCA	Inflação	1737	IBGE
PIM-PF	Produção Industrial	8159	IBGE

Fonte: Elaboração do autor.

Os dados de taxa de juros e atividade econômica foram obtidos pelo do Sistema Gerenciador de Séries Temporais (SGS) do BCB, sendo a de código 4189 a taxa de juros Selic acumulada no mês, anualizada base 252, como %a.a., com periodicidade mensal; e a de código 24364 o "Índice de Atividade Econômica do Banco Central (IBC-Br), aplicado ajuste sazonal, com periodicidade mensal.

Os dados de inflação e produção industrial foram adquiridos do Sistema IBGE de Recuperação Automática (SIDRA), com código 1737 e 8159, respectivamente.

Como a taxa básica de juros estava anualizada com periodicidade mensal, transformou-se ela em mensal e acumulou-se em número-índice. Da inflação e da taxa básica de juros então calculou-se o valor acumulado em doze meses, de acordo:

$$Índice_{12m,t} = \frac{Índice_t}{Índice_{t-12}} - 1 \tag{68}$$

em que o $\acute{I}ndice_t$ é o valor do Índice no tempo t; o $\acute{I}ndice_{t-12}$ é o valor do Índice doze meses após t; e $\acute{I}ndice_{12m,t}$ é o Índice acumulado no tempo t. Com isso, foi possível obter a série da taxa de juros real, descontando a inflação da taxa básica de juros.

Para as séries da inflação e taxa básica de juros haviam-se dados até maio de 2022; para a série da produção industrial até abril de 2022 e para a série de atividade econômica até fevereiro de 2022. Dessa maneira, optou-se por trabalhar com dados até fevereiro de 2022.

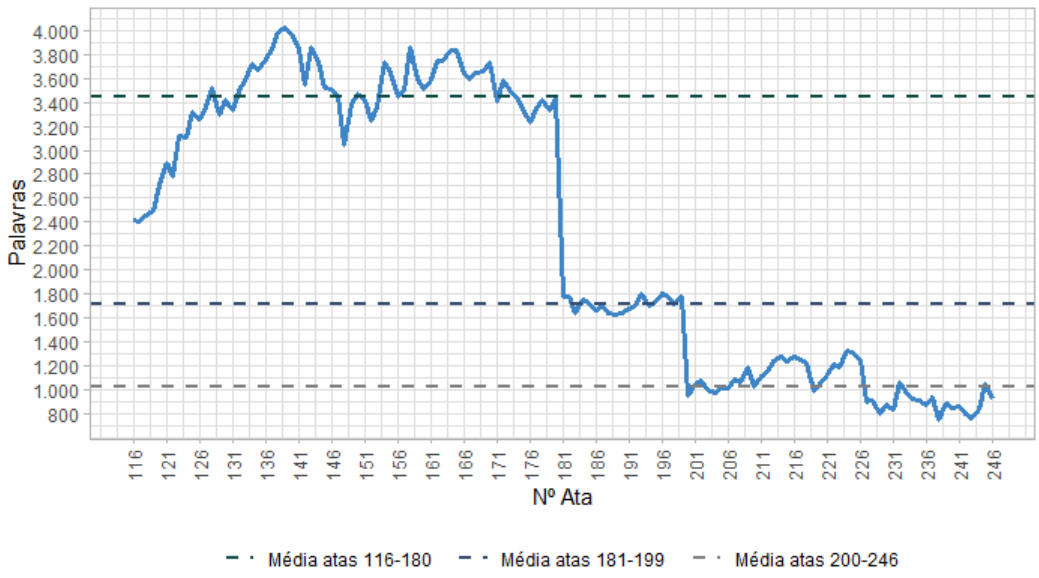
4 RESULTADOS

Nesta parte serão apresentados os resultados obtidos através da abordagem metodológica abordada anteriormente. Na Seção 4.1 são abordadas estatísticas descritivas acerca das atas do Copom. Já na Seção 4.2 é abordado o índice de sentimentos criado e comparação dele com as variáveis macroeconômicas presentes neste trabalho. Na Seção 4.3 são expostos os resultados dos testes de estacionariedade para as variáveis macroeconômicas e para o índice. Por fim, a Seção 4.4 trata da modelagem econométrica, onde por meio da aplicação de VAR, obtém-se as IRF.

4.1 ESTATÍSTICAS DESCRITIVAS

Com os dados limpos e tratados, algumas observações iniciais podem ser apontadas. Conforme demonstra-se na Figura 3, no período de Henrique Meirelles na presidência do BCB, tivemos o ápice (4027 palavras) em uma ata do Copom, filtrada por *stopwords*. Outro ponto de destaque é a mudança de média de palavras por ata em três períodos específicos, sendo que o primeiro contempla as atas de número 116 até a de número 180 (média de 3444 palavras por ata), o segundo contempla as atas 181 até 199 (média de 1715 palavras por ata) e o terceiro período as atas de 200 até 246 (média de 1027 palavras por ata).

Figura 3 – Número total de palavras por ata, filtrado por *stopwords*



Fonte: Elaboração do autor.

A alteração do tamanho a partir da ata 181 veio por retirada relevante de seções nos comunicados, onde o cargo de presidência do BCB estava ocupado por Alexandre Tombini. Por fim, a última quebra é referente a ata de número 200 em diante, onde a presidência do BCB se encontrava sob o comando de Ilan Goldfajn, em que houveram mudanças significativas nas estruturas e seções dos comunicados. Vale ressaltar que a quantidade de palavras nas atas do

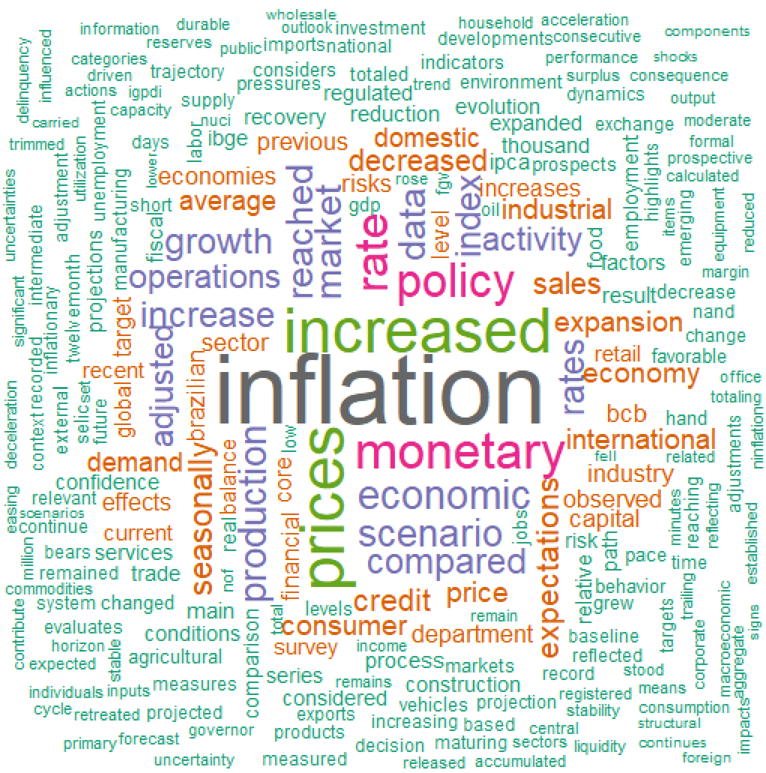
Copom não necessariamente está relacionado a maior qualidade das informações. Neste caso, a redução dos comunicados temporalmente é, em definitivo, uma questão interessante de ser abordada, ainda mais quando o tocante é legibilidade, transparência e credibilidade (maiores detalhes abordados em Omotosho (2019)).

Outra questão interessante que merece atenção é em relação as palavras mais frequentes (absolutamente falando) nas atas do Copom em todo o período analisado, conforme a Figura 4 apresenta, sendo que quanto maior a disposição da palavra, mais frequentemente ela é observada.

De antemão, destaca-se termos como *'inflation'*, *'prices'*, *'rate'*, *'monetary + policy'*, *'market + expectations'*, entre outras palavras relevantes. Vale notar-se também a alta correlação de $0,81$ entre os termos *'inflation'* e *'increased'*, denotando um alto aparecimento conjunto dessas palavras, isto é, aumento da inflação é um tópico muito recorrente nas atas do Copom, conforme o esperado.

Tais palavras estão em consonância com os objetivos tanto principais como secundários de um Banco Central, isto é, foco na estabilidade de preços, nível de atividade econômica e sinais acerca de políticas posteriores, fornecendo expectativas ao mercado (BLINDER, 1999; MISHKIN, 2000; BERNANKE; REINHART; SACK, 2004).

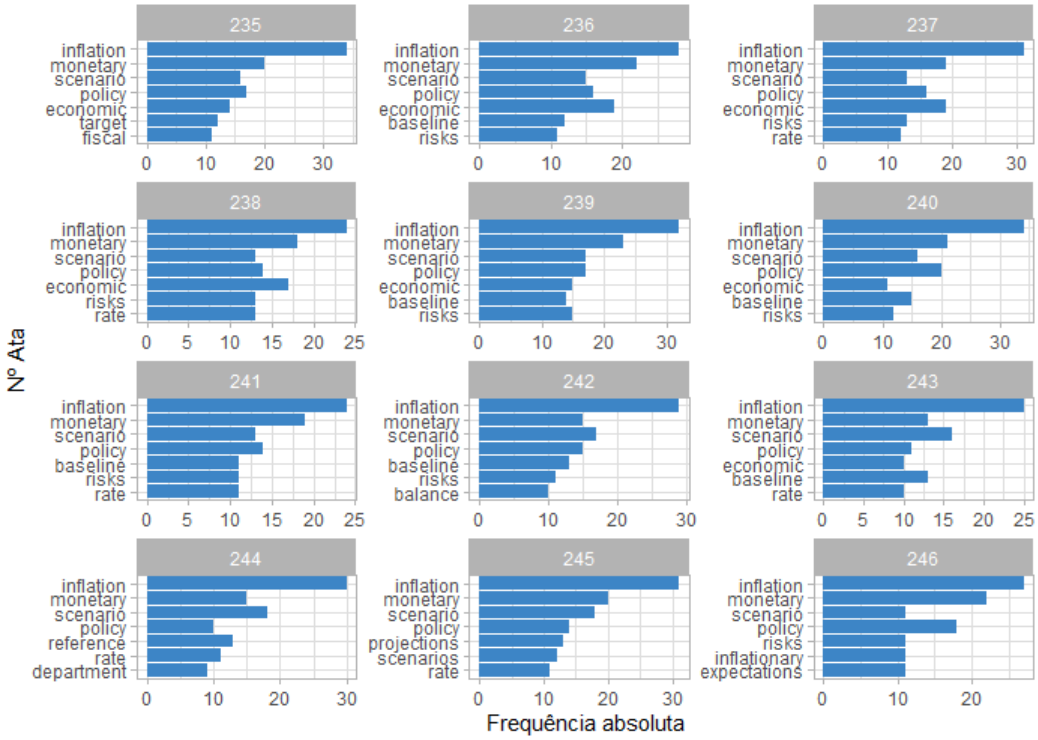
Figura 4 – Núvem de palavras mais frequentes nas atas



Fonte: Elaboração do autor.

Após ter-se visualizado de forma ampla os termos mais recorrentes englobando todas as atas desde janeiro de 2006 até maio de 2022, separou-se as palavras que mais apareceram

Figura 5 – Palavras mais frequentes nas atas do Copom (atas nº 235 até nº 246)



Fonte: Elaboração do autor.

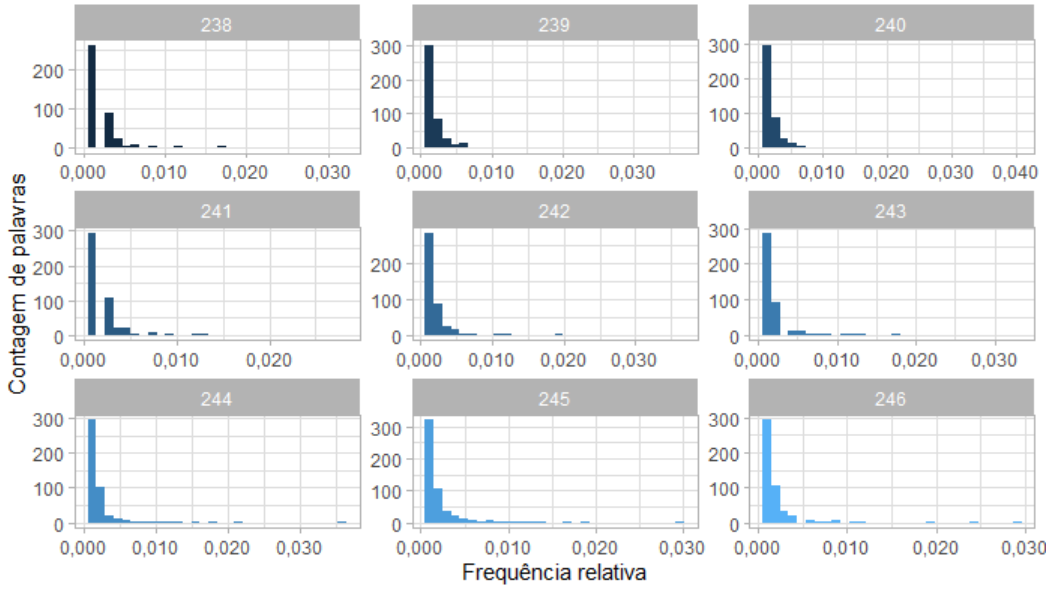
em cada ata. Como por exemplo, conforme observado na Figura 5, tem-se as oito palavras mais frequentes de uma sub-amostra, da ata 235 até a ata 246. Tal qual o esperado e apontado anteriormente pela Figura 4, *'inflation'* é o termo que aparece como mais corriqueiro para todas as atas, confirmado pelas atas da sub-amostra. Termos como *'monetary'*, *'scenario'*, *'policy'*, *'economic'* e *'rate'* também são homogeneamente comuns. Detalhe para o surgimento no pódio entre as oito mais citadas da palavra *'risks'* na ata 236 até a ata 242, saindo e reaparecendo na ata 246.

De qualquer maneira, apesar de ser interessante a reafirmação dos termos discutidos nas atas em relação aos objetivos de um comunicado de fonte oficial desta grandeza, pode-se constatar que em linhas gerais as palavras se repetem muito. Uma maneira de contornar essa questão e tentar extrair as palavras com maior importância de cada ata, é tornar a atenção para as frequências dos termos (*tf*) em relação à frequência inversa do documento (*idf*), denominada estatística tf-idf, abordada anteriormente na seção metodológica.

Como enunciado em Silge e Robinson (2017), a parte *idf* desta relação de frequências diminui o peso para palavras que aparecem muitas vezes nos documentos e aumenta o peso para palavras que não aparecem muitas vezes, sendo que uma vez multiplicado pela parte *tf* nos resulta na frequência do termo ajustado por quão raramente ele aparece nos comunicados, isto é, extrai-se as palavras mais relevantes para cada ata dentro do período de análise proposto.

Mediante demonstrado pelos histogramas na Figura 6, por meio da frequência relativa (número de vezes que determinada palavra aparece em determinada ata dividido pelo número de

Figura 6 – Frequências relativas, visualização das caudas longas (atas nº 238 até nº 246)



Fonte: Elaboração do autor.

palavras naquela ata) pode-se perceber que existem muitos termos nas caudas da distribuição (que aparecem poucas vezes), sendo assim, poucas palavras que aparecem muitas vezes e muitas palavras que aparecem poucas vezes.

Salienta-se que distribuições de cauda longa como as representadas na Figura 6, são comuns no estudo estatístico-linguístico dado um *Corpus* (neste caso os documentos das atas do Copom), onde há uma relação inversamente proporcional entre a frequência em que determinada palavra aparece e o rank (ordenação da maior para menor frequência) dessa mesma palavra. A versão clássica dessa relação é chamada Lei de Zipf (PIANTADOSI, 2014).

Tabela 4 – Estatísticas *tf-idf* para palavras da ata nº 246

Nº Ata	Palavra	n	tf	idf	tf-idf
246	inflation	27	0,0298	0,0000	0,0000
246	monetary	22	0,0243	0,0000	0,0000
246	policy	18	0,0199	0,0000	0,0000
246	expectations	11	0,0122	0,0000	0,0000
246	scenario	11	0,0122	0,0000	0,0000
246	department	10	0,0110	0,0000	0,0000
246	economic	10	0,0110	0,0000	0,0000
246	prices	8	0,0088	0,0000	0,0000
246	rate	8	0,0088	0,0000	0,0000
246	risk	7	0,0077	0,0000	0,0000

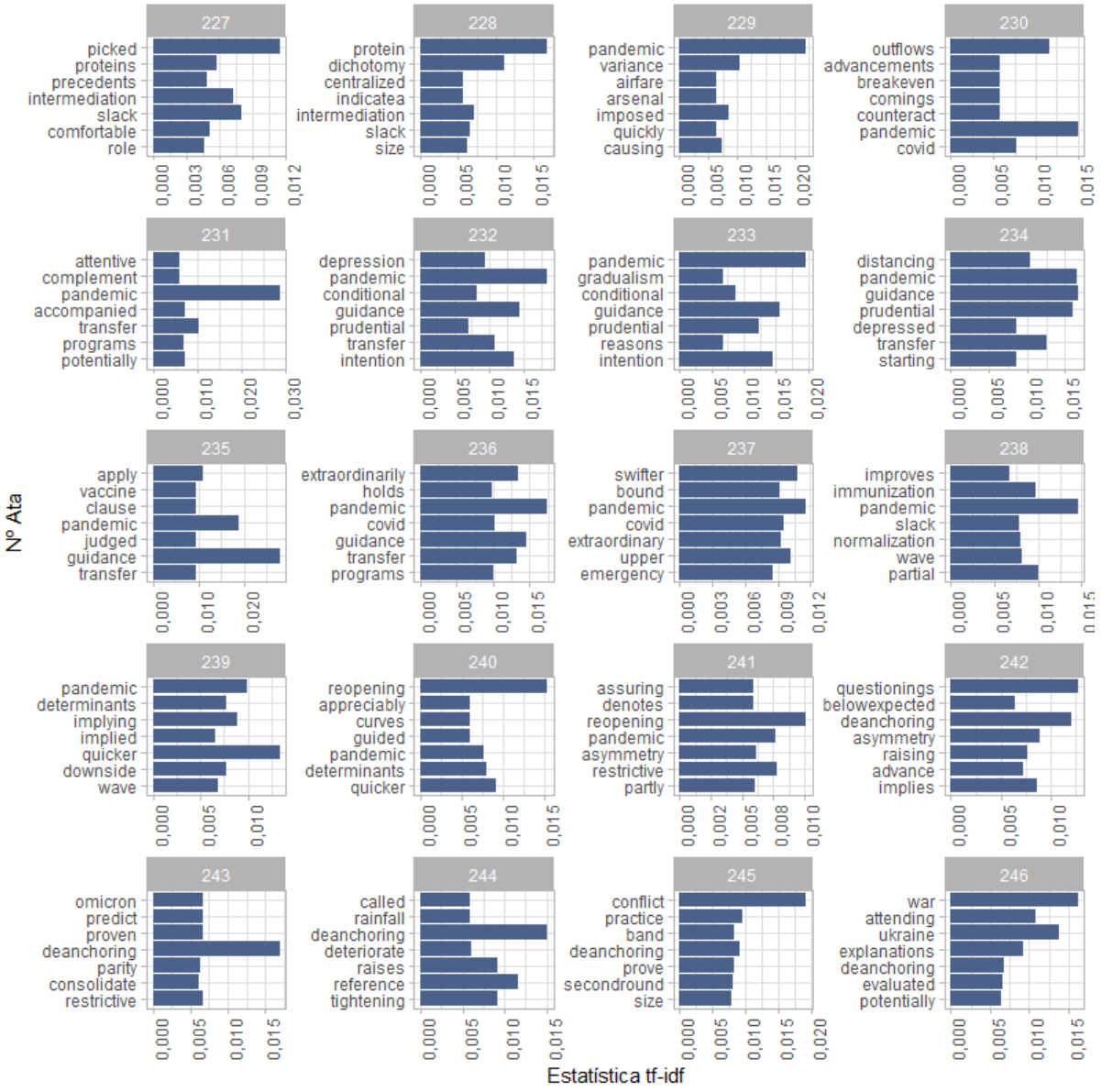
Fonte: Elaboração do autor.

Tomando a ata nº 246 como exemplo, conforme apresentado na Tabela 4, tem-se a palavra, a quantidade de vezes que ela aparece na determinada ata, a frequência do termo (*tf*), frequência inversa do documento (*idf*) e a estatística *tf-idf*. Como é possível contemplar, para palavras que

são muito comuns entre todas as atas (*'inflation'*, *'expectations'*, *'prices'* etc.), a *idf* será zero (e por consequência a estatística *tf-idf* também), uma vez que $\ln(1) = 0$.

Já, quando olha-se para as maiores estatísticas *tf-idf*, por exemplo, da ata nº 227 até a ata nº 246, consoante a Figura 7, tem-se resultados muito promissores e interessantes. Primeiramente, percebe-se de forma notória que as palavras se diferem bastante entre si e em relação àquelas presentes na Figura 5.

Figura 7 – Estatísticas *tf-idf* (atas nº 227 até nº 246)



Fonte: Elaboração do autor.

Além disso, se nota não só o surgimento da palavra *'pandemic'* a partir da ata nº 229, como sua persistência até a ata 240, com alta relevância. Na ata nº 229, pode-se perceber o aparecimento da palavra *'airfare'* (referente à variação nos preços de passagens aéreas devido à retração do preço internacional do barril de petróleo), assim como a palavra *'imposed'*, em

relação às restrições impostas; ambas referências como consequência do estourar da pandemia do Coronavírus em 2020.

Na ata nº 230, tem-se o surgimento da palavra *'outflows'*, referente a fortes saídas de capital no Brasil (um país emergente) em momentos incertos e de alto risco financeiro. Também rassalta-se o surgimento de *'transfer + programs'* na ata nº 231, acerca da implementação de programas de crédito e transferência de renda a famílias vulneráveis. No intermédio, destacam-se termos como *'prudential'*, *'gradualism'*, *'distancing'*, *'vaccine'*, *'emergency'*, *'immunization'*, *'partial + normalization'*, *'restrictive'*, *'omicron'*, todos relacionados à crise sanitária.

Outrossim, observa-se na ata 240, a inauguração do termo *'reopening'* em alusão ao início da reabertura da economia como um todo. Atenta-se também para o surgimento de *'deanchoring'* na ata nº 242 (perdurando até a 246), tratando-se da desancoragem das expectativas de inflação. Por fim, tem-se termos como *'conflict'* e *'ukraine'* devido à invasão da Ucrânia por parte da Rússia, iniciada em fevereiro de 2022.

Feitas análises de estatísticas descritivas referente ao período proposto, categorizou-se as palavras presentes nos documentos em positivas ou negativas e criou-se um índice de sentimentos das atas do Copom.

4.2 ÍNDICE DE SENTIMENTOS DAS ATAS DO COPOM

No presente trabalho, foi utilizado o dicionário léxico proposto por Hu e Liu (2004), que conta com 6.786 palavras classificadas com teor positivo ou negativo, onde alguns desses termos podem ser observados na Tabela 5. É importante destacar que esse método desconsidera bigramas (*'no good'*, *'not true'*, *'no risks'*, *'no volatility'*), ou seja, ele é baseado somente em unigramas, que são palavras isoladas (SILGE; ROBINSON, 2017).

Tabela 5 – Palavras presentes no dicionário de sentimentos proposto por Hu e Liu (2004)

Palavra	**Sentimento**
recovery	Positivo
consistent	Positivo
stability	Positivo
advanced	Positivo
robust	Positivo
risks	Negativo
inflationary	Negativo
worsening	Negativo
volatility	Negativo
inconsistency	Negativo

Fonte: Elaboração do autor.

Cruzando as informações das atas do Copom que foram adquiridas por meio da raspagem de dados *web* e posteriormente tratadas, com as informações presentes no dicionário léxico,

encontrou-se o número de palavras categorizadas como positivas e negativas para cada documento de texto.

Tabela 6 – Sentimentos encontrados nas atas do Copom

Data	Nº Ata	Positivo	Negativo	Sentimento	Total	Índice de Sentimento
11/12/2019	227	26	22	4	48	0,5417
05/02/2020	228	23	30	-7	53	0,4340
18/03/2020	229	21	36	-15	57	0,3684
06/05/2020	230	28	45	-17	73	0,3836
17/06/2020	231	29	39	-10	68	0,4265
05/08/2020	232	41	56	-15	97	0,4227
16/09/2020	233	47	48	-1	95	0,4947
28/10/2020	234	38	44	-6	82	0,4634
09/12/2020	235	41	33	8	74	0,5541
20/01/2021	236	33	36	-3	69	0,4783
17/03/2021	237	37	44	-7	81	0,4568
05/05/2021	238	20	32	-12	52	0,3846
16/06/2021	239	27	38	-11	65	0,4154
04/08/2021	240	27	39	-12	66	0,4091
22/09/2021	241	26	35	-9	61	0,4262
27/10/2021	242	17	35	-18	52	0,3269
08/12/2021	243	18	36	-18	54	0,3333
02/02/2022	244	20	37	-17	57	0,3509
16/03/2022	245	25	57	-32	82	0,3049
04/05/2022	246	19	46	-27	65	0,2923

Fonte: Elaboração do autor.

Pós cruzamento dos dados e categorização das palavras, se baseando em Bholat et al. (2015), criou-se o índice de sentimentos a seguir:

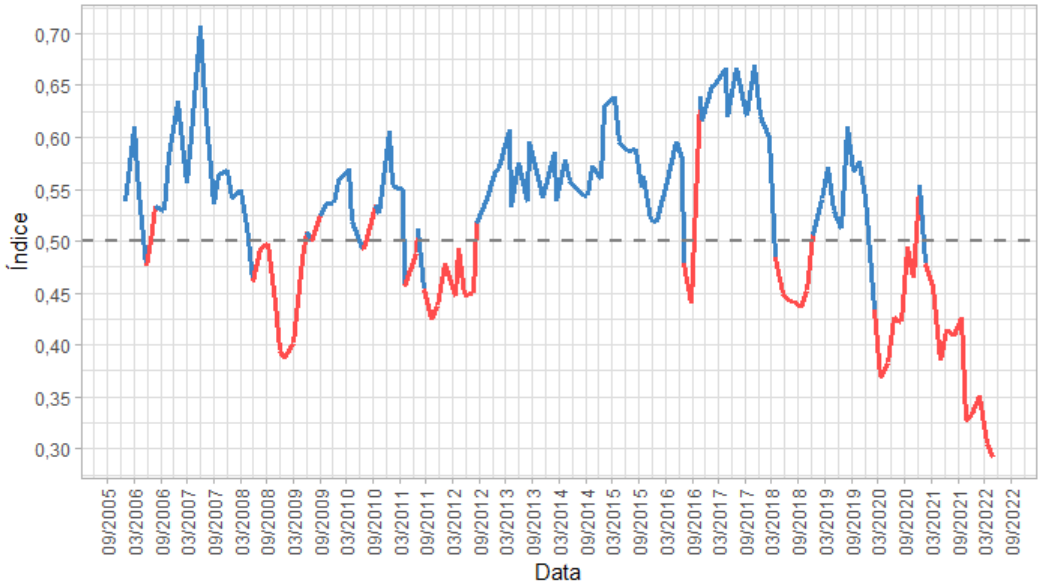
$$IS_t = \frac{NP_t}{NP_t - NN_t} \tag{69}$$

onde, na equação (69), tem-se que IS_t é o índice de sentimento para a ata no tempo t; NP_t é a quantidade de palavras categorizadas como positivas para a ata no tempo t; e NN_t é a quantidade de palavras categorizadas como negativas para a ata no tempo t.

Sendo assim, como $0 \leq IS_t \leq 1$, quando o valor for acima de $0,5$ tem-se uma perspectiva otimista para a economia, caso contrário, temos a definição do sentimento de pessimismo. Destaca-se também que quanto mais próximo de 1 o valor, mais otimista a ata, assim como quanto mais próximo de 0, mais pessimista.

Para exemplificar, na Tabela 6, estão dispostas informações do índice de sentimentos criado, da ata nº 227 até a de nº 246. A coluna *Positivo* é o NP_t, a coluna *Negativo* o NN_t, a coluna *Sentimento* é o saldo (positivas subtraídas das negativas), a coluna *Total* é a soma das palavras categorizadas e a última coluna é referente ao índice. Como é possível observar, das atas presentes na tabela, somente a 227 e 235 tiveram um saldo tido como otimista (e com baixo grau, de $0,5417$ e $0,5541$ respectivamente), o restante teve um teor negativo, com destaque para a ata nº 246 com um $IS_{246} = 0,2923$.

Figura 8 – Índice de sentimento das atas do Copom



Fonte: Elaboração do autor.

Por fim, foi elaborado um gráfico, conforme a Figura 8, com a série temporal do IS_t criado, em que $t = 116, ..., 246$. A coloração da linha em azul representa períodos com um $IS_t > 0,5$ e a coloração em vermelho, um $IS_t \leq 0,5$. Na maior parte do tempo (82 das atas), observa-se o índice indicando teor positivo, enquanto no restante, inclusive para o período mais recente, nota-se o teor negativo (49 das atas).

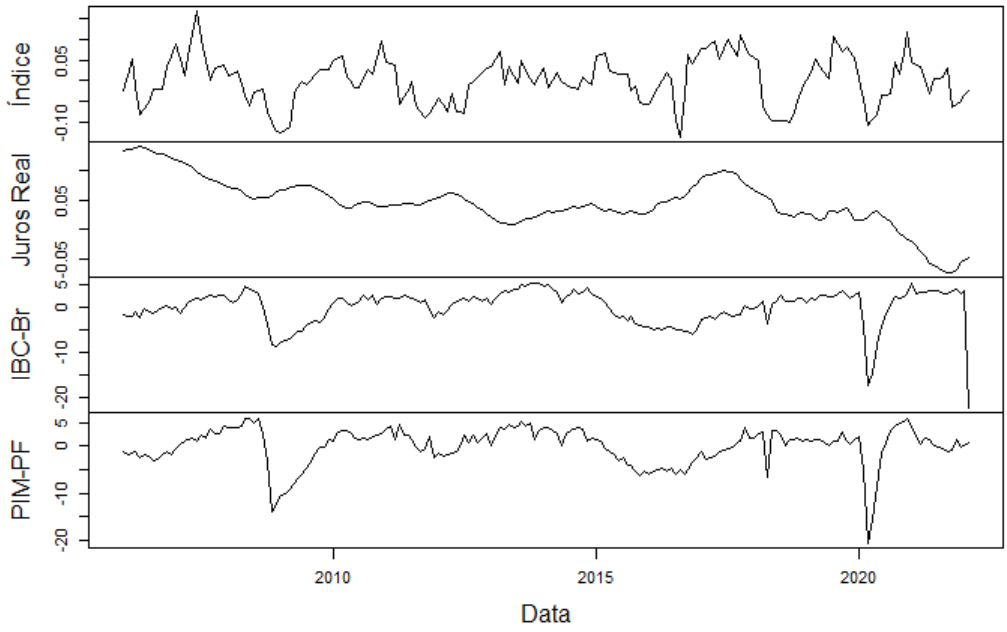
Chega-se na conclusão de que, de forma geral, o Copom aborda seus comunicados acerca da condução de política monetária e assuntos relacionados à atividade e conjuntura econômica no Brasil de maneira otimista. Com o índice elaborado, comparou-se ele com variáveis macroeconômicas de interesse.

4.2.1 Comparando o índice com variáveis macroeconômicas

Assim como nos exercícios realizados por Bloom (2009) e Ferreira et al. (2017), para as séries do Índice de Sentimentos, IBC-Br e PIM-PF foram extraídos os componentes cíclicos por meio do filtro de Hodrick e Prescott (1997).

A Figura 9 demonstra as séries temporais das quatro variáveis que posteriormente serão inclusas no modelo VAR. Atenta-se para o fato de que o índice de sentimentos criado aparenta acompanhar as demais variáveis selecionadas em momentos críticos na economia, como, por exemplo, em 2008 com a Crise Financeira do Subprime e em 2020 com a Crise Sanitária do Covid-19. Além disso, é possível observar uma tendência de queda da taxa de juros real em nível, enquanto as demais séries aparentam ser estáveis.

Figura 9 – Séries macroecônomicas pós-tratamento



Fonte: Elaboração do autor.

4.3 RESULTADOS DOS TESTES DE ESTACIONARIEDADE

Conforme mencionado anteriormente, as variáveis inclusas em modelos VAR(p) precisam possuir a característica de estacionariedade. Dessa maneira, foram realizados testes de raiz unitária de Dickey-Fuller (DF) e Dickey-Fuller Aumentado (ADF) - com uma defasagem - e de Kwiatkowski–Phillips–Schmidt–Shin (KPSS).

Tabela 7 – Resultados dos testes ADF sem constante e sem tendência

Variável	Estatística teste	1,0%	5,0%	10,0%	H-1,0%	H-5,0%	H-10,0%
Índice	-4,99	-2,58	-1,95	-1,62	Estacionária	Estacionária	Estacionária
Selic Real	-2,16	-2,58	-1,95	-1,62	Não-estacionária	Estacionária	Estacionária
IBC-Br	-3,52	-2,58	-1,95	-1,62	Estacionária	Estacionária	Estacionária
PIM-PF	-4,56	-2,58	-1,95	-1,62	Estacionária	Estacionária	Estacionária

Fonte: Elaboração do autor.

Primeiramente, conforme observa-se na Tabela 7, elaborou-se o teste de ADF sem constante e sem tendência. A primeira e segunda coluna representam a variável e sua respectiva estatística do teste; já as colunas "1,0%", "5,0%"e "10,0%"representam os níveis de significância (α); e por fim as colunas "H-1,0%", "H-5,0%"e "H-10,0%"indicam se a série é estacionária ou não através da aceitação ou não da hipótese nula.

Por meio deste teste constata-se que praticamente todas as séries rejeitam a hipótese nula (H_0) de não-estacionariedade, portanto são estacionárias para todos os níveis de significância

propostos, exceto para a série dos juros reais a um $\alpha = 1,0\%$.

Tabela 8 – Resultados dos testes ADF com constante

Variável	Estatística teste	1,0%	5,0%	10,0%	H-1,0%	H-5,0%	H-10,0%
Índice	-4,97	-3,46	-2,88	-2,57	Estacionária	Estacionária	Estacionária
Selic Real	-1,77	-3,46	-2,88	-2,57	Não-estacionária	Não-estacionária	Não-estacionária
IBC-Br	-3,49	-3,46	-2,88	-2,57	Estacionária	Estacionária	Estacionária
PIM-PF	-4,55	-3,46	-2,88	-2,57	Estacionária	Estacionária	Estacionária

Fonte: Elaboração do autor.

Quando feito o teste de ADF com constante, conforme a Tabela 8 demonstra, a série dos juros reais não só é tida como não-estacionária a um $\alpha = 10,0\%$, mas como para $\alpha = 5,0\%$ e $\alpha = 1,0\%$ também.

Elaborando o teste de ADF com constante e tendência, como contempla-se na Tabela 9, além das taxas de juros reais não serem estacionárias, tem-se que para níveis de significância de $5,0\%$ e $1,0\%$, o IBC-Br também apresenta não-estacionariedade.

Tabela 9 – Resultados dos testes ADF com constante e com tendência

Variável	Estatística teste	1,0%	5,0%	10,0%	H-1,0%	H-5,0%	H-10,0%
Índice	-4,96	-3,99	-3,43	-3,13	Estacionária	Estacionária	Estacionária
Selic Real	-2,32	-3,99	-3,43	-3,13	Não-estacionária	Não-estacionária	Não-estacionária
IBC-Br	-3,43	-3,99	-3,43	-3,13	Não-estacionária	Não-estacionária	Estacionária
PIM-PF	-4,54	-3,99	-3,43	-3,13	Estacionária	Estacionária	Estacionária

Fonte: Elaboração do autor.

Por fim, por meio do teste de KPSS, em que a hipótese nula (H_0) é a de estacionariedade, tem-se que somente a série das taxas de juros reais rejeita H_0, conforme está disposto na Tabela 10.

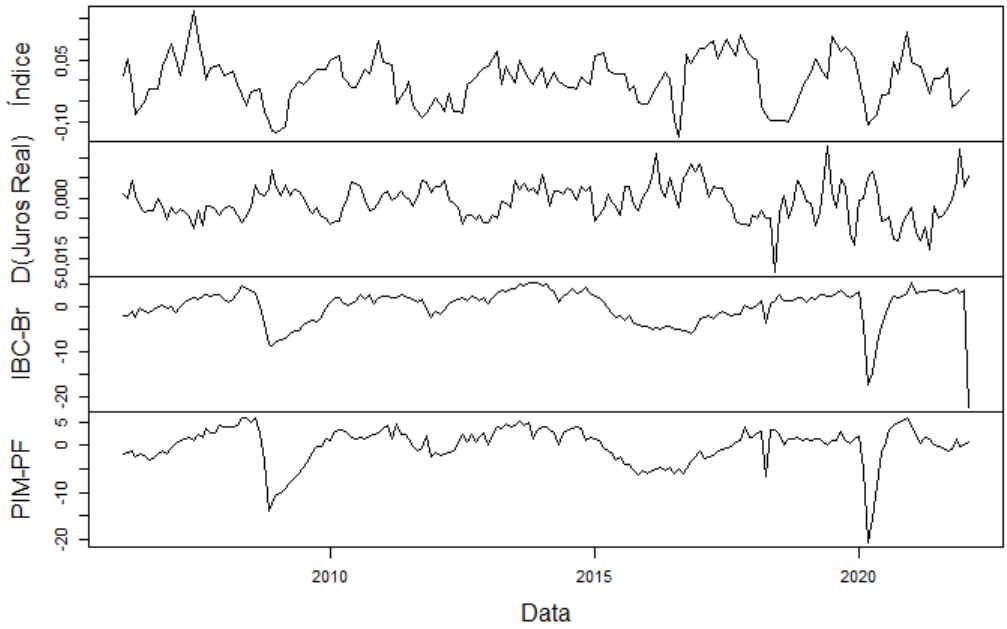
Tabela 10 – Resultados dos testes KPSS

Variável	Estatística teste	1,0%	5,0%	10,0%	H-1,0%	H-5,0%	H-10,0%
Índice	0,04	0,22	0,15	0,12	Estacionária	Estacionária	Estacionária
Selic Real	0,33	0,22	0,15	0,12	Não-estacionária	Não-estacionária	Não-estacionária
IBC-Br	0,11	0,22	0,15	0,12	Estacionária	Estacionária	Estacionária
PIM-PF	0,08	0,22	0,15	0,12	Estacionária	Estacionária	Estacionária

Fonte: Elaboração do autor.

Desta maneira, optou-se por manter todas as séries em seus níveis, exceto para a série dos juros reais, em que foi feita uma diferenciação, como pode ser observado na Figura 10.

Figura 10 – Séries macroeconômicas pós-tratamento, com juros real diferenciado



Fonte: Elaboração do autor.

Após isso, todos os testes foram refeitos e para um $\alpha = 5,0\%$, que é o nível de significância desejado, constatou-se que todas as séries são estacionárias, prontas para serem inclusas no modelo VAR.

4.4 RESULTADOS DO MODELO DE VETORES AUTORREGRESSIVOS (VAR)

Detendo o conjunto de variáveis estacionárias, a estimação do modelo VAR(p) pode ser efetuada, onde os critérios de decisão abordados nas Equações (41), (42) e (43) são utilizados para se definir as ordens ótimas de defasagens. Após isso, visto que a estabilidade do modelo e de seus parâmetros foram garantidas, foi selecionada a ordem que apresentou os melhores testes de diagnóstico.

A partir dos critérios $BIC(p,q)$ e $HQ(p,q)$ o número de defasagens ótimo apontado foi de 1, enquanto para o critério $AIC(p,q)$ o número aumenta para 2. Dessa forma, foram estimados dois modelos, contemplando as duas possibilidades de defasagens propostas pelos critérios de informação. Tanto para uma quanto para duas defasagens, as raízes do polinômio característico ficaram abaixo de um, ou seja, fora do círculo unitário, além dos testes de flutuação empírica, abordado em Ploberger e Krämer (1992), evidenciarem parâmetros estáveis, isto é, sem quebras estruturais (os resultados podem ser constatados no Apêndice A).

Para o modelo VAR(1), o teste conjunto de autocorrelação serial de Breusch-Godfrey falhou, dado que o p-valor = $0,007$, abaixo de p-valor = $0,05$. Considerando os testes individuais, somente a série do D(Juros Real) e a série do IBC-Br entregaram um p-valor > $0,05$. Enquanto, para o teste conjunto ARCH de heterocedasticidade, o modelo apresentou estar estável, já que retornou um p-valor = $0,10$.

Dessa maneira, optou-se trabalhar com o VAR(2), onde o teste conjunto de autocorrelação serial de Breusch-Godfrey apresentou como p-valor = $0,10$, acima de p-valor = $0,05$, denotando assim que não há autocorrelação nos resíduos das séries do modelo estimado. Considerando os testes individuais, todas as séries apresentaram um p-valor > $0,05$, isto é, individualmente também não apresentam correlação serial. Enquanto que, para o teste conjunto ARCH de heterocedasticidade, o modelo apresentou estar estável, retornando um p-valor = $0,10$.

No VAR(2), considerando todas as variáveis como endógenas, obteve-se então o seguinte sistema de equações:

$$IS_t = IS_{t-1} + \Delta JR_{t-1} + AE_{t-1} + PI_{t-1} + IS_{t-2} + \Delta JR_{t-2} + AE_{t-2} + PI_{t-2} + \varepsilon_{IS_t} \quad (70)$$

$$\Delta JR_t = IS_{t-1} + \Delta JR_{t-1} + AE_{t-1} + PI_{t-1} + IS_{t-2} + \Delta JR_{t-2} + AE_{t-2} + PI_{t-2} + \varepsilon_{\Delta JR_t} \quad (71)$$

$$AE_t = IS_{t-1} + \Delta JR_{t-1} + AE_{t-1} + PI_{t-1} + IS_{t-2} + \Delta JR_{t-2} + AE_{t-2} + PI_{t-2} + \varepsilon_{AE_t} \quad (72)$$

$$PI_t = IS_{t-1} + \Delta JR_{t-1} + AE_{t-1} + PI_{t-1} + IS_{t-2} + \Delta JR_{t-2} + AE_{t-2} + PI_{t-2} + \varepsilon_{PI_t} \quad (73)$$

onde a equação do índice de sentimentos (IS) chegou a um coeficiente de determinação ajustado no valor de $R^2_{adj} = 0,667$; ao passo que para a equação da taxa de juros real diferenciada (ΔJR), o $R^2_{adj} = 0,429$. Para a equação da atividade econômica (AE), resultou-se num $R^2_{adj} = 0,626$; enquanto a equação da produção industrial (PI) decorreu em um $R^2_{adj} = 0,722$, sendo o maior entre todos.

4.4.1 Causalidade de Granger

No modelo VAR(2), para os testes individuais de causalidade de Granger, nenhuma variável apresentou Granger-causar o índice de sentimentos criado para um nível de $\alpha = 5,0\%$. Considerando o índice Granger-causar as variáveis, conforme observa-se na Figura 11, tem-se que apenas a PIM-PF é afetada, para um $\alpha = 5,0\%$, sendo seu p-valor = $0,02$.

Figura 11 – Testes individuais de causalidade de Engle-Granger para o VAR(2)

Variável	Estatística F	P-valor
D(Juros) -> Índice	0,334	0,717
IBC-Br -> Índice	1,184	0,309
PIM-PF -> Índice	2,189	0,115
Índice -> D(Juros)	1,112	0,331
Índice -> IBC-Br	1,632	0,198
Índice -> PIM-PF	4,042	0,019

Fonte: Elaboração do autor.

Para o teste em conjunto de causalidade de Granger, tem-se um p-valor > 0,05 para as variáveis de taxa de juros real e produção industrial, assim essas variáveis Granger-causam o índice, apesar de que para a atividade econômica o p-valor = 0,04, fazendo com que o IBC-Br não Granger-cause o índice criado, como apresentado na Figura 12.

Figura 12 – Teste conjunto de causalidade de Engle-Granger para o VAR(2)

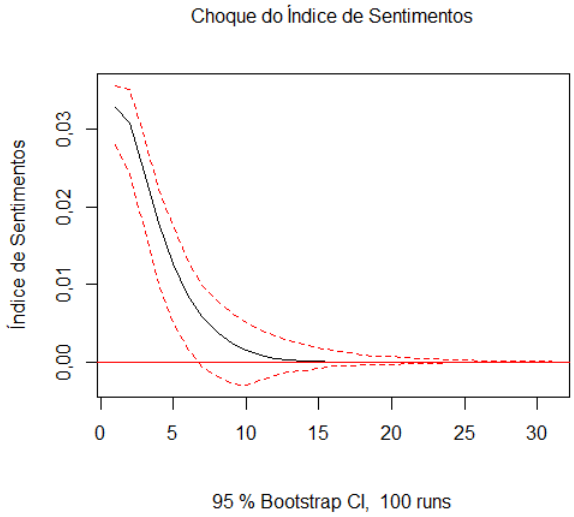
Variável	Estatística F	P-valor
D(Juros) -> Índice	1,650	0,131
IBC-Br -> Índice	2,248	0,037
PIM-PF -> Índice	0,969	0,445

Fonte: Elaboração do autor.

4.4.2 Funções impulso-resposta (IRF)

A análise das Funções de Impulso-Resposta demonstrou, conforme a Figura 13 aponta que, dado um choque positivo no índice, a resposta do próprio índice sofre positivamente segundo o tamanho do choque dado, de forma significativa. Salienta-se que para o VAR(2), a partir do 6º período, o choque deixa-se de ser significante para um intervalo de confiança de 95,0%.

Figura 13 – IRF do VAR(2): Índice → Índice

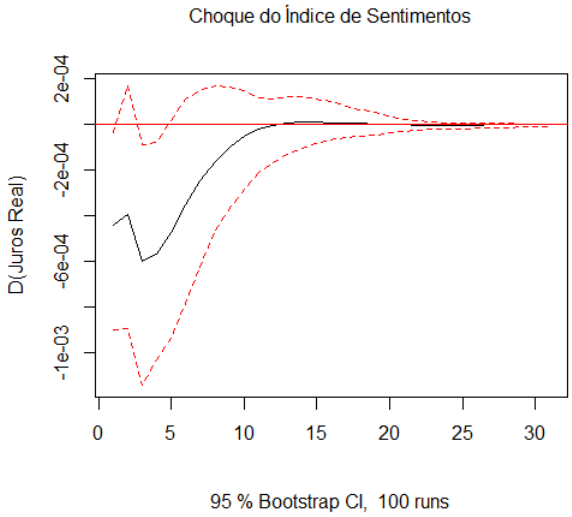


Fonte: Elaboração do autor.

Já, a partir de um choque efetuado no índice em relação à diferença dos juros reais, como observa-se na Figura 14, tem-se uma resposta negativa, que aumenta com o passar dos períodos. Entretanto, para o VAR(2), o choque é somente significativo entre o 4º e 5º período, para 95,0% de confiança.

Outrossim, quando se atenta à resposta da atividade econômica dado um choque no índice de sentimentos, percebe-se conforme a Figura 15 aponta, que os impactos são sentidos a partir

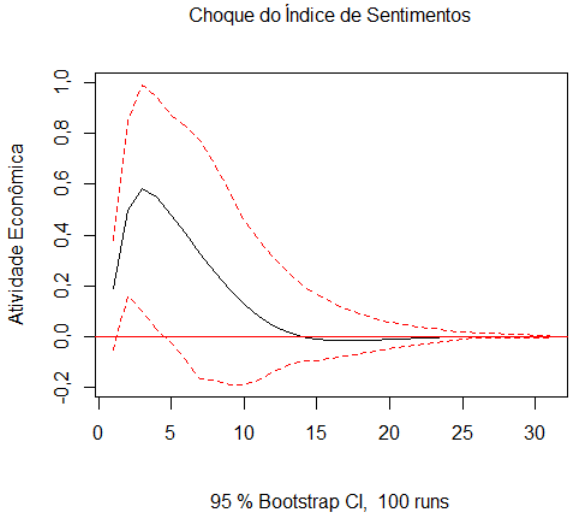
Figura 14 – IRF do VAR(2): Índice → D(Juros)



Fonte: Elaboração do autor.

do 2º mês, perdurando significativamente (intervalo de confiança de 95,0%) até o 5º período para o VAR(2). Do choque até o quinto período, tem-se uma reação positiva, sendo que a partir do 5º, ela decai.

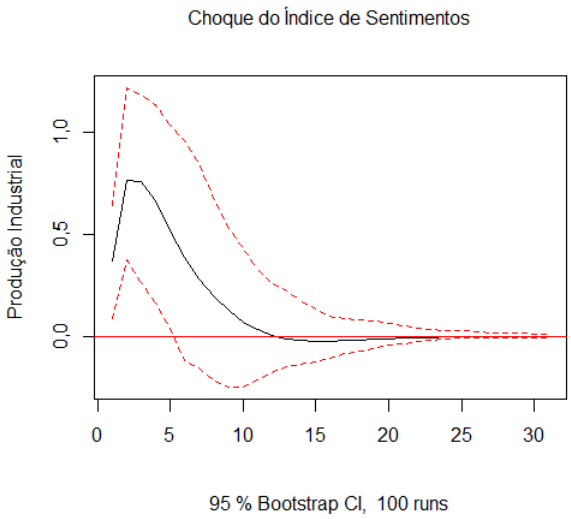
Figura 15 – IRF do VAR(2): Índice → IBC-Br



Fonte: Elaboração do autor.

Por fim, dado um impulso no índice, a resposta da produção industrial se assemelha com a da atividade econômica, onde como se observa na Figura 16, há um pico no 5º período, onde a partir dele, os choques deixam de ser significativos.

Figura 16 – IRF do VAR(2): Índice → PIM-PF



Fonte: Elaboração do autor.

Em Ferreira et al. (2017), são disponibilizados apenas os resultados das IRF de atividade econômica e produção industrial, dado um choque no IIE-Br, onde ambas respondem negativamente, com pequenos intervalos significantes, semelhantes aos resultados do trabalho de Bloom (2009). As possíveis divergências com os resultados obtidos neste trabalhao, provavelmente se dão, pois, os autores criaram índices de incerteza da economia e não um índice de sentimentos das atas do Copom. Não só isso, os autores contemplaram um número maior de variáveis, como câmbio, desemprego e preços de commodities, além de impor uma restrição de ordenamento das variáveis, o que ia além da proposta deste trabalho. Interessante apontar que, por ainda serem incipientes os estudos que relacionam mineração textual, análise de sentimentos e economia, os resultados podem servir como norteadores de alguma forma.

Dessa maneira, conclui-se que por meio de choques aos sentimentos dos comunicados das atas do Copom, a taxa de juros básica acumulada em doze meses, descontada da taxa de inflação (também acumulada em doze meses), sofre negativamente, ao qual se estabiliza conforme se avança nos períodos. Entretanto, quando se olha para a atividade econômica e para a produção industrial, percebe-se que a resposta é positiva, também se estabilizando em torno de zero conforme o avanço do tempo, visto que os choques não possuem efeitos permanentes em séries temporais que possuem a característica de estacionariedade.

5 CONSIDERAÇÕES FINAIS

Este estudo buscou elucidar sobre a importância da comunicação e transparência de um Banco Central, isto é, toda informação disponibilizada para os agentes externos, que envolve a condução da política monetária, em relação à atividade econômica e em relação à sinais de trajetórias futuras. Foi apontado, que os níveis atuais de comunicação e transparência foi um trajeto percorrido ao longo dos anos, assim como um movimento internacional.

Dessa forma, como ferramenta de comunicação, o BCB se utiliza das atas do Copom, além da abordagem de transparência adotada por meio do RMI. Com isso, quando se fala da relação entre a discrição e regra, de forma rara um Banco Central seguirá rigorosamente as regras, uma vez que tem-se o chamado viés inflacionário, isto é, o impulso de se diminuir o desemprego ou aumentar o produto. Nesse sentido, quando o RMI não possui credibilidade, a tendência é resultar na discrição, aumentando a oferta de moeda e por consequência gerando uma taxa de inflação maior.

Com isso em mente, apresentaram-se estudos que aplicaram na economia, técnicas de mineração textual, que é um processo para organizar e transformar dados textuais disponíveis em documentos e grandes bases de dados, para extrair significados e padrões. Como as atas do Copom são disponibilizadas no site do BCB, o objetivo deste trabalho foi aplicar essas técnicas para extrair as todas as atas lá presentes, filtrar para o período de 2006 até 2022 e, por meio de análise de sentimentos, criar um índice para os comunicados.

A partir da elaboração do índice, comparou-se ele com variáveis macroeconômicas de interesse, e com os devidos testes econométricos feitos para asseguração da consistência dos resultados, aplicou-se modelagem de vetores autorregressivos, podendo assim inferir ou não causalidade de Granger das variáveis, como também a elaboração de funções de impulso-resposta.

Os critérios de informação para escolha do número de defasagens do modelo apontou para uma e duas, entretanto, como o VAR(1) falhou no teste de autocorrelação serial dos resíduos, trabalhou-se somente com o VAR(2). Dessa maneira, foi possível sair de uma simples correlação para uma análise de causalidade, onde obteve-se que o índice Granger-causa a produção industrial e as variáveis de taxa de juros real e produção industrial Granger-causam o índice. Analisando as IRF, para os juros reais, tem-se uma resposta negativa, sendo os choques significativos para o 4º e 5º período; para a atividade econômica, os impactos são sentidos no 2º mês, perdurando significativamente até o 5º período; e por fim, a resposta da produção industrial é afetada positivamente, com um pico de impacto no 5º período.

Enfim, reitera-se que este trabalho procurou contribuir com a aplicação da intersecção de técnicas de mineração textual, análise de sentimentos e macroeconomia, além de considerar técnicas estatísticas-econométricas para contribuir com a robustez das resultantes, com o VAR cumprindo o proposto no trabalho. Como trabalhos futuros, pode-se sugerir modelagens preditivas, levando em consideração que o VAR é um ótimo modelo para tal, assim como a construção de novos índices de sentimentos, com diferentes fontes de mineração textual e/ou utilização de

diferentes dicionários léxicos. Não só isso, sugere-se também que pode ser estudada a relação de cointegração entre as variáveis e aplicar um modelo de Vetor de Correção de Erros (VECM) para entender o tempo que o sistema retorna para seu nível de equilíbrio, ou até mesmo, aplicação de técnicas de *Machine Learning*, como, por exemplo, a modelagem de tópicos.

REFERÊNCIAS

BERNANKE, B.; REINHART, V.; SACK, B. Monetary policy alternatives at the zero bound: An empirical assessment. **Brookings papers on economic activity**, Brookings Institution Press, v. 2004, n. 2, p. 1–100, 2004. Citado 3 vezes nas páginas 16, 17 e 43.

BERNANKE, Ben S. Remarks by governor ben s. bernanke: Central bank talk and monetary policy. **Japan Society Corporate Luncheon**, 2004. Citado na página 18.

BERNANKE, Ben S. Remarks by governor ben s. bernanke: Fedspeak. **Meetings of the American Economic Association**, 2004. Citado na página 18.

BHOLAT, D. et al. Text mining for central banks. **SSRN 2624811**, 2015. Disponível em: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2624811>. Citado 2 vezes nas páginas 20 e 48.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **J. Mach. Learn. Res.**, JMLR.org, v. 3, p. 993–1022, 2003. Disponível em: <http://portal.acm.org/citation.cfm?id=944937>. Citado na página 21.

BLINDER, A. S. Central-bank credibility: Why do we care? how do we build it? **American economic review**, v. 90, n. 5, p. 1421–1431, 1999. Citado 2 vezes nas páginas 16 e 43.

BLINDER, A. S. et al. Central bank communication and monetary policy: A survey of theory and evidence. **Journal of economic literature**, v. 46, n. 4, p. 910–45, 2008. Citado 5 vezes nas páginas 14, 15, 16, 17 e 18.

BLOOM, Nicholas. The impact of uncertainty shocks. **econometrica**, Wiley Online Library, v. 77, n. 3, p. 623–685, 2009. Citado 3 vezes nas páginas 40, 49 e 56.

BREUSCH, Trevor S. Testing for autocorrelation in dynamic linear models. **Australian economic papers**, Wiley Online Library, v. 17, n. 31, p. 334–355, 1978. Citado na página 37.

BREUSCH, Trevor S; PAGAN, Adrian R. A simple test for heteroscedasticity and random coefficient variation. **Econometrica: Journal of the econometric society**, JSTOR, p. 1287–1294, 1979. Citado na página 38.

BUENO, R. De-Losso. **Econometria de séries temporais**. [S.l.]: Cengage Learning, 2011. Citado 10 vezes nas páginas 13, 29, 30, 31, 32, 33, 34, 35, 37 e 38.

CAVALLO, Alberto. Online and official price indexes: Measuring argentina's inflation. **Journal of Monetary Economics**, Elsevier, v. 60, n. 2, p. 152–165, 2013. Citado na página 20.

CHAGUE, F. et al. Central bank communication affects the term-structure of interest rates. **Revista Brasileira de Economia**, SciELO Brasil, v. 69, p. 147–162, 2015. Citado na página 20.

COENEN, G. et al. Communication of monetary policy in unconventional times. ECB working paper, 2017. Citado na página 16.

COSTA FILHO, Adonias Evaristo; ROCHA, Fabiana. Comunicação e política monetária no brasil. **Revista Brasileira de Economia**, SciELO Brasil, v. 63, p. 405–422, 2009. Citado 2 vezes nas páginas 18 e 19.

COSTA FILHO, A. E.; ROCHA, F. Como o mercado de juros futuros reage à comunicação do banco central? **Economia aplicada**, SciELO Brasil, v. 14, p. 265–292, 2010. Citado na página 20.

COSTA, H. C. **ENSAIOS EM MACROECONOMIA APLICADA**. Tese (Doutorado) — Universidade Federal do Rio Grande do Sul, Porto Alegre, 2016. Disponível em: <https://lume.ufrgs.br/handle/10183/158135>. Citado 4 vezes nas páginas 13, 20, 24 e 25.

DICKEY, David A; FULLER, Wayne A. Distribution of the estimators for autoregressive time series with a unit root. **Journal of the American statistical association**, Taylor & Francis, v. 74, n. 366a, p. 427–431, 1979. Citado na página 28.

DURBIN, James; WATSON, Geoffrey S. Testing for serial correlation in least squares regression: I. **Biometrika**, JSTOR, v. 37, n. 3/4, p. 409–428, 1950. Citado na página 36.

ENDERS, Walter. **Applied econometric time series**. [S.l.]: John Wiley & Sons, 2009. Citado 2 vezes nas páginas 32 e 34.

FAUST, Jon; SVENSSON, Lars EO. Transparency and credibility: Monetary policy with unobservable goals. **International Economic Review**, Wiley Online Library, v. 42, n. 2, p. 369–397, 2001. Citado na página 18.

FERREIRA, Pedro Costa et al. Medindo a incerteza econômica no brasil. 2017. Citado 4 vezes nas páginas 21, 40, 49 e 56.

FULLER, Wayne A. **Introduction to statistical time series**. [S.l.]: John Wiley & Sons, 1976, p. 373. Citado na página 30.

GIL, Antonio Carlos. **Métodos e técnicas de pesquisa social**. [S.l.]: 6. ed. Ediitora Atlas SA, 2008. Citado na página 23.

GLEJSER, Herbert. A new test for heteroskedasticity. **Journal of the American Statistical Association**, Taylor & Francis, v. 64, n. 325, p. 316–323, 1969. Citado na página 38.

GODFREY, Leslie G. Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. **Econometrica: Journal of the Econometric Society**, JSTOR, p. 1293–1301, 1978. Citado na página 37.

GOLDFELD, Stephen M; QUANDT, Richard E. Nonlinear methods in econometrics. North-Holland Pub. Co., 1972. Citado na página 38.

GUJARATI, D. N.; PORTER, D. C. **Econometria básica**. [S.l.]: Amgh Editora, 2011. Citado 10 vezes nas páginas 13, 27, 28, 29, 30, 33, 35, 36, 37 e 38.

HAAN, Jakob De; EIJFFINGER, Sylvester CW; RYBIŃSKI, Krzysztof. **Central bank transparency and central bank communication: Editorial introduction**. [S.l.]: Elsevier, 2007. 1–8 p. Citado na página 18.

HODRICK, Robert J; PRESCOTT, Edward C. Postwar us business cycles: an empirical investigation. **Journal of Money, credit, and Banking**, JSTOR, p. 1–16, 1997. Citado na página 49.

HU, Minqing; LIU, Bing. Mining and summarizing customer reviews. In: **Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining**. [S.l.: s.n.], 2004. p. 168–177. Citado 2 vezes nas páginas 8 e 47.

HUTTO, Clayton; GILBERT, Eric. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: **Proceedings of the international AAAI conference on web and social media**. [S.l.: s.n.], 2014. v. 8, n. 1, p. 216–225. Citado na página 25.

ISSING, Otmar. Communication, transparency, accountability: monetary policy in the twenty-first century. **Review**, v. 87, n. Mar, p. 65–83, 2005. Disponível em: <https://EconPapers.repec.org/RePEc:fip:fedlrv:y:2005:i:mar:p:65-83:n:v.87no.2,pt.1>. Citado na página 19.

ISSING, O.; WOOD, G. E. **Should we have faith in central banks?** [S.l.]: Institute of Economic Affairs London, 2001. Citado na página 14.

KENNEDY, Peter. **A guide to econometrics**. [S.l.]: John Wiley & Sons, 2008. Citado na página 31.

KWIATKOWSKI, Denis et al. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? **Journal of econometrics**, Elsevier, v. 54, n. 1-3, p. 159–178, 1992. Citado 2 vezes nas páginas 31 e 32.

LIU, Bing et al. Sentiment analysis and subjectivity. **Handbook of natural language processing**, Oxfordshire, v. 2, n. 2010, p. 627–666, 2010. Citado na página 25.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. Introduction to information retrieval. **Natural Language Engineering**, Cambridge University Press, v. 16, n. 1, p. 100–103, 2010. Citado na página 24.

MISHKIN, F. S. **Moeda, bancos e mercados financeiros**. [S.l.]: Livros Técnicos e Científicos Editora, 2000. Citado 2 vezes nas páginas 16 e 43.

MORETTIN, Pedro Alberto; BUSSAB, Wilton Oliveira. **Estatística básica**. [S.l.]: Saraiva Educação SA, 2017. Citado na página 27.

OMOTOSHO, B. S. Central bank communication in ghana: Insights from a text mining analysis. **SSRN Electronic Journal**, Elsevier BV, 2019. Disponível em: <https://doi.org/10.2139/ssrn.3526451>. Citado 2 vezes nas páginas 21 e 43.

OOMS, J. The jsonlite package: A practical and consistent mapping between json data and r objects. arXiv, 2014. Disponível em: <https://arxiv.org/abs/1403.2805>. Citado na página 39.

OOMS, J. Text extraction, rendering and converting of pdf documents. 2022. Disponível em: <https://cran.r-project.org/web/packages/pdftools/pdftools.pdf>. Citado na página 39.

PARK, Rolla E. Estimation with heteroscedastic error terms. **Econometrica (pre-1986)**, Blackwell Publishing Ltd., v. 34, n. 4, p. 888, 1966. Citado na página 38.

PFAFF, Bernhard. **Analysis of integrated and cointegrated time series with R**. [S.l.]: Springer Science & Business Media, 2008. Citado na página 38.

PHILLIPS, Peter CB; PERRON, Pierre. Testing for a unit root in time series regression. **Biometrika**, Oxford University Press, v. 75, n. 2, p. 335–346, 1988. Citado na página 30.

PIANTADOSI, Steven T. Zipf's word frequency law in natural language: A critical review and future directions. **Psychonomic bulletin & review**, Springer, v. 21, n. 5, p. 1112–1130, 2014. Citado na página 45.

PLOBERGER, Werner; KRÄMER, Walter. The cusum test with ols residuals. **Econometrica: Journal of the Econometric Society**, JSTOR, p. 271–285, 1992. Citado na página 52.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2020. Disponível em: <https://www.R-project.org/>. Citado na página 39.

RAJARAMAN, Anand; ULLMAN, Jeffrey David. **Mining of massive datasets**. [S.l.]: Cambridge University Press, 2011. Citado na página 26.

SALTON, Gerard; WONG, Anita; YANG, Chung-Shu. A vector space model for automatic indexing. **Communications of the ACM**, ACM New York, NY, USA, v. 18, n. 11, p. 613–620, 1975. Citado na página 26.

SHAPIRO, A.H.; WILSON, D. J. Taking the fed at its word: A new approach to estimating central bank objectives using text analysis. **Federal Reserve Bank of San Francisco, Working Paper Series**, p. 01–74, 6 2019. Disponível em: <http://www.frbsf.org/economic-research/publications/working-papers/2019/02>. Citado 2 vezes nas páginas 21 e 22.

SILGE, Julia; ROBINSON, David. tidytext: Text mining and analysis using tidy data principles in r. **Journal of Open Source Software**, v. 1, n. 3, p. 37, 2016. Citado na página 39.

SILGE, Julia; ROBINSON, David. **Text mining with R: A tidy approach**. [S.l.]: "O'Reilly Media, Inc.", 2017. Citado 3 vezes nas páginas 39, 44 e 47.

SIMS, Christopher A. Macroeconomics and reality. **Econometrica: journal of the Econometric Society**, JSTOR, p. 1–48, 1980. Citado na página 34.

STONE, P. J.; DUNPHY, D. C.; SMITH, M. S. The general inquirer: A computer approach to content analysis. MIT press, 1966. Citado 2 vezes nas páginas 13 e 25.

SVENSSON, Lars EO. What is wrong with taylor rules? using judgment in monetary policy through targeting rules. **Journal of Economic Literature**, v. 41, n. 2, p. 426–477, 2003. Citado na página 18.

TAYLOR, John B. Discretion versus policy rules in practice. In: ELSEVIER. **Carnegie-Rochester conference series on public policy**. [S.l.], 1993. v. 39, p. 195–214. Citado na página 17.

TRAPLETTI, Adrian; HORNIK, Kurt. **tseries: Time Series Analysis and Computational Finance**. [S.l.], 2022. R package version 0.10-51. Disponível em: <https://CRAN.R-project.org/package=tseries>. Citado na página 32.

WHITE, Halbert. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. **Econometrica: journal of the Econometric Society**, JSTOR, p. 817–838, 1980. Citado na página 38.

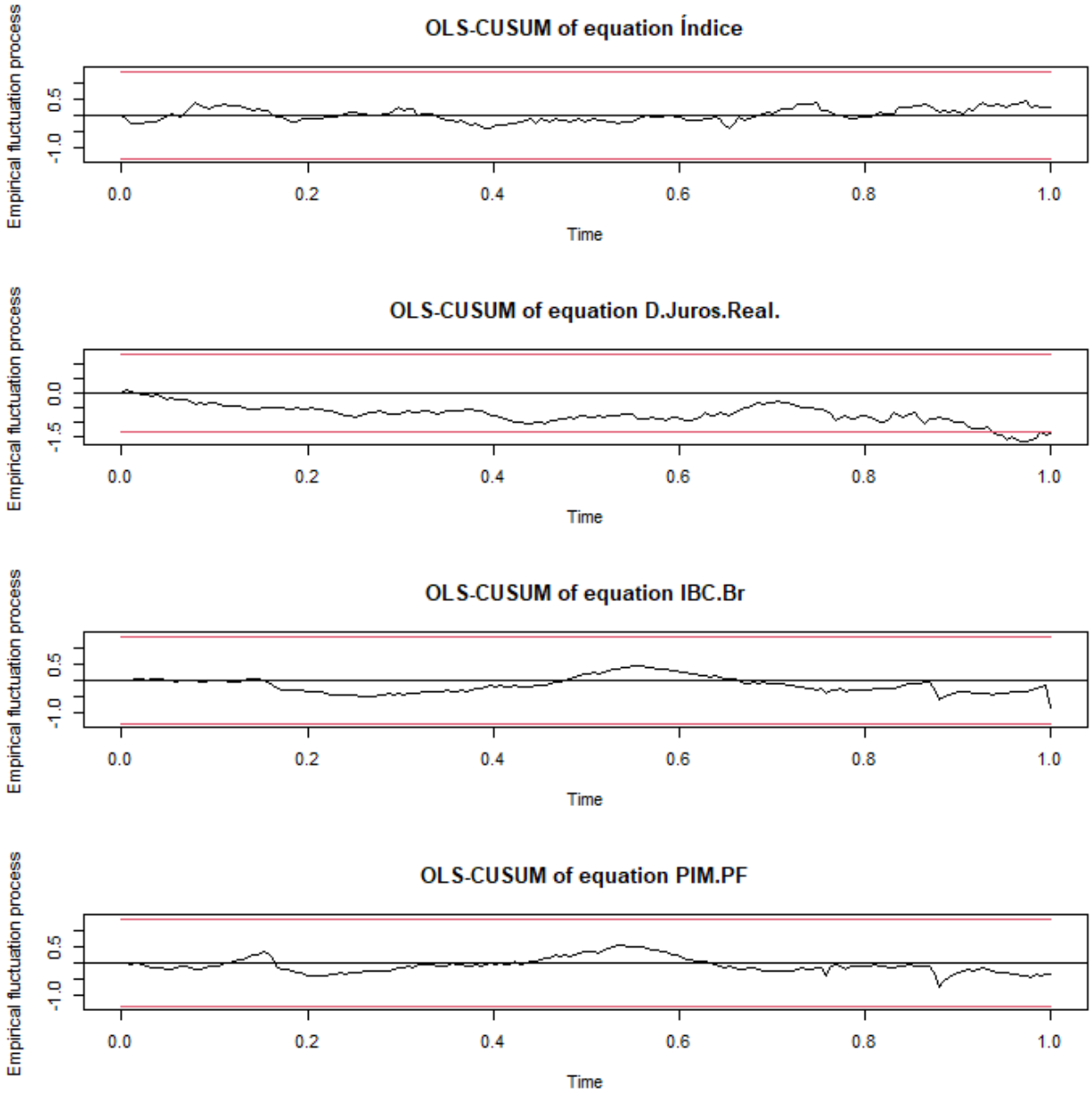
WINKLER, Bernhard. Which kind of transparency? on the need for clarity in monetary policy-making. **On the Need for Clarity in Monetary Policy-Making (August 2000)**, 2000. Citado na página 14.

WOODFORD, M. Central bank communication and policy effectiveness. National Bureau of Economic Research Cambridge, Mass., USA, 2005. Citado 5 vezes nas páginas 14, 15, 16, 17 e 19.

ZAHNER, J. **Essays on Natural Language Processing and Central Banking**. Tese (Doutorado) — Universidade de Marburgo, Marburgo, 2021. Disponível em: <https://archiv.ub.uni-marburg.de/diss/z2021/0497/>. Citado na página 21.

APÊNDICE A – SAÍDAS AUXILIARES

Figura 17 – Verificação de quebras estruturais: OLS-CUSUM VAR(2)



Fonte: Elaboração do autor.

Tabela 11 – Raízes do polinômio característico do VAR(2)

Variável	Raiz
ISt-1	0,764
ISt-2	0,764
D(JR)t-1	0,695
D(JR)t-2	0,695
AEt-1	0,339
AEt-2	0,335
PIt-1	0,335
PIt-2	0,311

Fonte: Elaboração do autor.

Tabela 12 – Coeficientes da equação do índice de sentimentos do VAR(2)

Variável	Coeficiente	Erro-padrão	Significância
ISt-1	0,92	0,07	0,00
ISt-2	-0,15	0,07	0,04
D(JR)t-1	0,65	0,07	0,35
D(JR)t-2	-0,08	0,07	0,91
AEt-1	0,00	0,00	0,74
AEt-2	0,00	0,00	0,96
PIt-1	0,00	0,00	0,13
PIt-2	0,00	0,00	0,33

Fonte: Elaboração do autor.

Tabela 13 – Coeficientes da equação da diferença da taxa de juros real do VAR(2)

Variável	Coeficiente	Erro-padrão	Significância
ISt-1	0,00	0,00	0,97
ISt-2	0,00	0,00	0,44
D(JR)t-1	0,69	0,07	0,00
D(JR)t-2	0,00	0,07	0,98
AEt-1	0,00	0,00	0,83
AEt-2	0,00	0,00	0,62
PIt-1	0,00	0,00	0,11
PIt-2	0,00	0,00	0,21

Fonte: Elaboração do autor.

Tabela 14 – Coeficientes da equação da atividade econômica do VAR(2)

Variável	Coeficiente	Erro-padrão	Significância
ISt-1	9,07	5,36	0,09
ISt-2	-6,90	5,38	0,20
D(JR)t-1	2,62	49,90	0,96
D(JR)t-2	-89,80	50,38	0,08
AEt-1	0,98	0,20	0,00
AEt-2	-0,35	0,19	0,07
PIt-1	0,05	0,14	0,72
PIt-2	0,09	0,14	0,52

Fonte: Elaboração do autor.

Tabela 15 – Coeficientes da equação da produção industrial do VAR(2)

Variável	Coeficiente	Erro-padrão	Significância
ISt-1	12,93	4,85	0,00
ISt-2	-9,99	4,86	0,04
D(JR)t-1	-62,40	45,11	0,17
D(JR)t-2	19,24	45,50	0,67
AEt-1	0,60	0,18	0,00
AEt-2	-0,54	0,17	0,00
PIt-1	0,54	0,13	0,00
PIt-2	0,18	0,13	0,16

Fonte: Elaboração do autor.

APÊNDICE B – CÓDIGOS

Todos os códigos utilizados no presente trabalho estão disponíveis no repositório online dessa monografia, sendo possível ser acessado por meio deste link.