Previsão de Variáveis Macroeconômicas usando Índice de Difusão, Dados Textuais e Aprendizado de Máquina.

Alexandro de Gois Oliveira* Lucas Lúcio Godeiro‡

2 de Julho de 2020

Resumo

O desemprego e a produção industrial são variáveis importantes nos relatórios econômicos. A cobertura jornalística dessas variáveis podem apontar medidas de previsão baseadas em um grande conjunto de dados e fatores, e as previsões geradas por esse grande conjunto de dados é conhecida com Índice de Difusão (DI). Nesse paper mostramos que, quando aplicamos o índice de difusão a dados de texto, os modelos DI apresentam um bom desempenho. O artigo mostra que o simples uso de dados de texto nos modelos de DI não melhoram as previsões da taxa de desemprego e produção industrial em relação ao modelo AR(1), contudo ganhos são obtidos quando selecionamos as palavras mais preditivas antes de calcular os fatores, permitindo que o dicionário seja atualizado ao longo do tempo.

Palavras-Chave: Dados de Texto; Aprendizado de Máquina, Índice de Difusão, Desemprego, Produção Industrial, Previsão.

Abstract

Unemployment and industrial production are important variables in economic reports. The news coverage of these variables can lead to forecasting measures based on a large set of data and factors, and the predictions of these large data sets are known as the Diffusion Index (DI) . In this paper We show that, when applied to text data, DI models do not perform well. We conclude that the simple use of text data in the DI models does not improve the predictions of the unemployment rate and industrial production in relation to the historical average model, however gains are obtained when we select the most predictive words before calculating the factors, allowing the dictionary to be updated over time.

Keywords: Text Data ; Machine Learning; ; Diffusion Index, Unemployment; Industrial Production; Forecasting.

Classificação Código JEL: C53, C55, E24, E37, E47

Área 4: Macroeconomia, Economia Monetária e Finanças

*Mestrando em Economia Aplicada pela Universidade do Estado do Rio Grande do Norte - UERN

†Doutor em Economia e Professor na Universidade Federal Rural do Semi-Árido - UFERSA.

‡Autor para Correspondência. E-mail: lucasgodeiro@ufersa.edu.br

1 Introdução

O desemprego e produção industrial são variáveis importantes para a economia e política, visto que, os formuladores de políticas públicas e o governo avaliam essas variáveis e relatam o estado da economia. Os agentes econômicos como investidores, consumidores e firmas avaliam suas expectativas quanto a economia, dado que os jornais noticiam as informações através de relatórios para informar o desempenho da economia. A cobertura de notícias pelos jornais tende a ter um viés negativo de relatórios, Soroka (2006) afirma que a mídia tende a relatar mais anúncios negativos quanto a taxa de desemprego do que relatórios positivos. Portanto, a literatura confirma que há viés de relatórios mais negativos do que positivos, dado que a informação ocorre de forma muito rápida, os agentes avaliam e formam suas expectativas a respeito do ambiente econômico.

A literatura empírica para previsão de preditores macroeconômicos já utilizam dados de texto de notícias econômicas como Feuerriegel and Gordon (2019), Garz (2018), Hendry and Madeley (2010), Hollanders and Vliegenthart (2011) entre outros. Os preditores usados neste artigo são calculados usando informações textuais que incluem dados não estruturados de notícias econômicas publicadas em jornais brasileiros como O Estadão, O Globo, Valor Econômico e Folha de São Paulo. Umas das metodologias utilizadas é o dicionário fixo, contudo quando o dicionário é usado para predição de preditores baseados em notícias, o mesmo não apresentou bom desempenho para previsão da taxa de desemprego e produção industrial. Loughran and McDonald (2011) já haviam mostrado que o uso das palavras não é idêntica na mídia e que as mesmas podem mudar.

A discussão sugerida é que o uso do dicionário utilizado para mineração de textos é atualizado no tempo para incorporar alterações que possam ocorrer no poder preditivo das palavras. A metodologia proposta em nosso artigo propõe e explica as duas possibilidades, visto que ao introduzir o dicionário fixo, a mineração de texto irá resultar em um grande conjunto de dados de séries temporais, com cada um representando a frequência em que uma determinada palavra aparece nas notícias e relatórios publicados ao longo do tempo. No entanto, nem todas as palavras ajudam a prever a taxa de desemprego e produção industrial, portanto, visamos atingir as palavras mais preditivas antes de calcular os fatores. Nessa parte do artigo é utilizado o aprendizado de máquina supervisionado introduzido por Bai and Ng (2008a). O uso do aprendizado de máquina quando é aplicado melhora a previsão fora da amostra da taxa de desemprego e produção industrial onde as palavras podem mudar ao longo do tempo, visto que o dicionário é variável no tempo. Além disso, a utilização desse dicionário implica em melhorar a interpretabilidade, onde podemos classificar as palavras mais preditivas em grupos de notícias positivas e negativas.[1] A classificação de palavras em positivas e negativas é importante para saber qual tipo de informação tem correlação e apresenta poder de previsão da taxa de desemprego e produção industrial. Investigamos a importância das notícias negativas, palavras carregadas de sentimento e o método utilizado para selecionar as palavras mais preditivas, o mesmo fizemos com as palavras positivas observando a relação dela com a taxa de desemprego e produção industrial.

Nossa metodologia de DI com preditores permite estimativas recursivas com bons desempenhos, tornando possível a estimação do modelo de previsão. Além disso, nossa abordagem permite uma análise de componentes principais baseadas em Hendry (2012) que utilizaram fatores com dados textuais. Portanto, a abordagem permite direcionar as palavras mais preditivas antes de calcular os fatores. Por fim, mostramos que o método proposto que seleciona as palavras teve poder preditivo e afeta o desempenho das previsões fora da amostra. Além disso, o uso do *LASSO* considerando um número de preditores muito grande deixa de fora uma quantidade considerável de informações. A utilização do *LASSO* para direcionar as palavras melhora o desempenho da previsão do modelo de DI em relação ao benchmark[2], contudo a melhora não é tão grande. No entanto, o oposto ocorre quando utilizamos Elastic Net para direcionar os dados de texto, o desempenho do método depende de nossa capacidade de selecionar corretamente as palavras mais preditivas e calcular os fatores. O nosso artigo está divido da seguinte forma; A Seção 2 e 3 é a revisão de literatura, a seção 4 apresenta o modelo e a seção 5 a nossa metodologia. A seção 6 são os nossos resultados e seção 7 as conclusões.

2 As notícias econômicas afetam a decisão dos agentes?

As expectativas dos agentes quanto ao estado da economia são de grande interesse econômico e político. Os agentes envolvidos podem ser famílias, firmas, investidores com o objetivo de investir, decisões quanto a economizar,

[1] Nossa metodologia classifica uma palavra (termo) como positiva (negativa) se estiver positivamente (negativamente) correlacionada com a taxa de desemprego e produção industrial.

[2] AR(1) é nosso modelo benchmark.

poupar e consumir. Portanto, as percepções dos agentes afetam as suas decisões. De acordo com Garz (2018), as famílias pessimistas podem ter gastos de consumos mais baixos do que famílias otimistas. A percepção das famílias mudam de acordo com o que ocorre com o estado da economia, e a cobertura da mídia nas notícias econômicas afetam essas percepções?

Na literatura empírica, Garz (2014), investiga o conteúdo da mídia em notícias econômicas, os dados indicaram um viés substancial em termos de quantias de relatórios positivos e negativos, em comparação com o desenvolvimento real do desemprego. De acordo com Garz (2014) os relatórios de desemprego noticiados pela mídia tendem a ter maior destaque negativo do que positivo. Blanchflower (1990) e Carroll and Dunn (1997), argumenta que as expectativas de desemprego podem ter efeitos negativos consideráveis no consumo agregado e no crescimento dos salários, pois, consumidores e funcionários se restringem quando são incertos ou pessimistas quanto ao futuro. A informação pode ser tida como um bem econômico, pois, diversos agentes podem se beneficiar através dela. Devido aos interesses das pessoas envolvidas na mídia, as notícias quanto à taxa de desemprego podem não ser transmitidas com precisão, a veracidade das informações pode não ser coerente com as informações verdadeiras, dada a alta relevância política que a taxa de desemprego proporciona, a cobertura da mídia pode ter efeitos cruciais sobre as eleições MacKuen et al. (1992), Hetherington (1996), Nadeau et al. (1999) e Easaw (2010).

De acordo com Soroka (2006), que testa a assimetria dos efeitos da cobertura noticiosa da mídia em questões econômicas, e descobre que a mídia tem maior probabilidade de relatar os aumentos do que as quedas na taxa de desemprego. Essa suposição de que a cobertura econômica da mídia afeta as percepções dos agentes quanto ao desemprego surgiu a partir da teoria da dependência da mídia (Ball Rokeach & DeFleur, 1976; Palmgreen & Clarke, 1977; Zucker, 1978). A teoria afirma que as notícias transmitidas pela mídia podem afetar a cognição individual de forma mais ampla, quando há menos informações em fontes alternativas.

Como o desemprego é uma variável de grande interesse, pois, os números relacionados a desemprego são indicadores da situação econômica do país que levam a formação das expectativas e incerteza que pode ser gerada na economia. De acordo com Garz (2016), as notícias podem afetar a percepção das famílias em dois sentidos, boas notícias sobre desemprego podem fazer as pessoas perceberem a economia de uma maneira mais favorável, por outro lado, notícias ruins sobre desemprego podem ter o efeito oposto. O consumo de informações geralmente é não rival, imperfeições como essa implica que o governo precise coletar e fornecer informações em algumas situações, porque não haveria oferta privada suficiente. o governo coleta e fornece essas informações por meio de comunicados a imprensa. Segundo (Mutz 1992; Brosius e Weimann 1995; Carroll 2003; Curtin 2003; Birz e Lott 2011), investigar notícias sobre desemprego é importante porque a cobertura afeta variáveis econômicas importantes, como consumo, investimento ou retorno de ações, por sua influência nas percepções das pessoas em relação à segurança no emprego e às expectativas de desemprego.

Andina-Díaz (2007), Gasper (2009) mostram que pode ser rentável para a mídia influenciar o público, diferenciando a cobertura das notícias, visto que, dependendo do viés de comunicação pode aumentar a demanda por seus produtos, atendendo suas preferências nas notícias. Besley e Prat (2006) afirmam que o viés político pode surgir quando as empresas de mídia se beneficiam da colaboração com o governo na forma de melhor acesso a funcionários públicos, decisões administrativas, ou intervenções legislativas. Garz (2014) afirma que as possíveis consequências da cobertura de notícias tendenciosas, como percepções/expectativas econômicas distorcidas constituem uma perda de eficiência macroeconômica. As empresas de mídia tem fortes incentivos para influenciar as políticas públicas no sentido de garantir sua liberdade de reportar e os consumidores tem pequenos incentivos para pressionar por mais regulamentação na cobertura das notícias.

Boyd et al. (2005), investiga a resposta de curto prazo dos preços das ações a chegada de notícias macroeconômicas. A resposta das ações depende se a economia está expandindo ou contraindo. Em média, o mercado de ações responde positivamente as notícias do aumento de desemprego em expansões, e negativamente em contrações. Gasper (2009), propõe o modelo formal para notícias políticas, sob condições razoáveis de mercado a mídia desempenha um papel crucial na maneira como as pessoas adquirem as informações. O modelo racional de escolha da teoria econômica veem a demanda por notícias apenas como uma demanda por informações. Os consumidores assistem ou ouvem para se tornarem mais informados. Se a informação não for precisa, isso valerá menos para o consumidor.

Boyd et al. (2005), chega a conclusão que as notícias sobre o aumento de desemprego nas contrações levam a menores retornos, ou ganhos esperados e, que resultam em menores preços das ações. Por outro lado, nas expansões, as mesmas notícias sobre aumento de desemprego acarretam em menores taxas de juros dos títulos do governo, levando ao aumento do preço das ações.

3 A cobertura das notícias econômicas reflete mudanças reais e que efeito exerce nas condições econômicas?

Birz and Lott Jr (2011), propõe uma abordagem para encontrar o efeito de notícias macroeconômicas nos preços das ações. Ele utiliza histórias de jornais para medir notícias macroeconômicas. Além de as reportagens enfatizarem os dados estatísticos, apontam também como a economia se saiu em relação as expectativas. Portanto, conclui-se que as divulgações estatísticas que representam as diversas condições econômicas podem ser um indicador das notícias reais associadas a divulgação.

De acordo com Hollanders and Vliegenthart (2011), a relação empírica entre a economia real, a confiança do consumidor e a cobertura de notícias econômicas em jornais da Holanda entre 1990 - 2009, chegando a conclusão que a atenção da mídia para o desenvolvimento econômico está associada à confiança do consumidor, com mais notícias negativas diminuindo a confiança do consumidor. Blood and Phillips (1995), descobrem que as manchetes da recessão influenciaram o sentimento do consumidor dos EUA em 1989 - 1993. Starr (2012) argumenta que a relevância das notícias econômicas para o público pode variar ao longo do tempo, de modo mais geral, as pessoas prestam mais atenção ao estado da economia nos meses que antecedem as eleições presidenciais, de modo que uma determinada notícia possa se tornar mais conhecida do que em outros momentos. Hollanders and Vliegenthart (2011) que durante uma recessão, a confiança do consumidor diminui ainda mais o consumo, levando ao declínio da demanda agregada, contraindo produção e aumentando o desemprego. McQueen and Roley (1993) encontram um resultado semelhante para índices de produção industrial e no desemprego. Considerando que boas notícias sobre essas variáveis nas expansões econômicas aumentam o preço das ações, embora, o efeito não seja estatisticamente significante e não se mantém em outros estágios dos ciclos de negócios.

De acordo com Carroll et al. (1994) explica que o sentimento do consumidor pode muito bem prever gastos sem ser um fator determinante; quando as perspectivas para a economia real são positivas, os ciclistas da vida futura fornecem leitura otimistas sobre o sentimento do consumidor. Em média, o otimismo será confirmado e a renda aumentará. Alexopoulos et al. (2009) afirma que mesmo que os indivíduos não participem do mercado acionário, eles tomam como indicador para avaliar o estado da economia.

4 Modelo Teórico

Neste trabalho, queremos comparar a capacidade preditiva de um modelo de regressão linear benchmark[3] com um modelo linear alternativo que inclui alguns fatores latentes. Estamos interessados em prever a taxa de desemprego e a produção industrial (y)[4] um passo a frente. Para isso, dividimos a amostra total de $T = R+P$ em observações dentro e fora da amostra. As observações na amostra abrangem de 1 a R, considerando que as observações fora da amostra abrangem $R+1$ através de T para um total de P previsões a um passo a frente. Dado que $X_t = (X_{1t}, ..., X_{Nt})'$, onde X_{it},i = 1,..., N, mede o número de vezes (frequência) que uma determinada palavra "i" aparece nos documentos publicados pelos jornais no tempo t^2. A equação de previsão que utiliza a informação X_t é dada por:

$$y_{t+1} = \alpha + \Gamma' X_t + \varepsilon_{t+1} \tag{1}$$

onde ε_{t+1} é um erro com média zero e variância limitada, α é o intercepto e Γ é um vetor $N \times 1$ dos coeficientes de inclinação. Se $N < T$, então para cada origem de previsão $t = R, ..., T-1$, geramos previsões de y_{t+1} regredindo recursivamente y_{s+1} nos $N \times 1$ preditores observados (palavras) X_s para $s = 1, ..., t-1$. Portanto, uma previsão que utilize todos os preditores disponíveis é $\widehat{y}_{1,t+1} = \widehat{\alpha}_t + \widehat{\Gamma}_t X_t$, onde $\widehat{\alpha}_t$ e $\widehat{\Gamma}_t$ são estimativas OLS recursivas de α e Γ sobre $t = R, ..., T-1$. Embora as estimativas OLS de α e Γ são $\sqrt{T}-$consistente, é sabido que o erro de previsão ao quadrado médio (MSPE) está aumentando N. Para resolver o problema de alta dimensionalidade que afeta o desempenho do modelo de previsão, Stock and Watson (2002) propõe a abordagem do índice de difusão (DI). Portanto, neste trabalho mostramos que a abordagem de DI combinada com dados de texto pode ser usada para melhorar as previsões fora da amostra da taxa de desemprego e produção industrial, se conseguirmos encontrar as palavras mais preditivas antes de calcular os fatores comuns. A abordagem fatorial para uma previsão um passo a frente é obtida por meio da estimação da equação de previsão

[3] Nosso modelo benchmark será o AR(1).

[4] y é uma matriz com 2 colunas. Neste sentido, separado efetuamos a previsão da produção industrial e depois do desemprego.

$$y_{t+1} = \alpha + \beta' \widetilde{f}_t + \varepsilon_{t+1} \tag{2}$$

onde $\widetilde{f}_t \subset \widetilde{F}_t$, β são coeficientes relativos a $\widetilde{f}_t$, $\widetilde{F}_t$ que por sua vez são estimativas do componente principal do vetor $\ell \times 1$ F_t no modelo de fatores.

$$X_{it} = \lambda_i' F_t + e_{it} \tag{3}$$

λ_i é a exposição da palavra i ao fator F_t, e e_{it} é a frequência idiossioncrática da palavra i. Em uma aplicação empírica, uma palavra pode ter seu peso λ_i igual ou próximo de zero.[5]

As equações (2) e (3) constituem a estrutura de previsão do Índice de Difusão (DI) de Stock e Watson (2002) sem defasagens $\widetilde{f}_t$. A previsão de DI é $\widetilde{y}_{T+1} = \alpha + \beta' \widetilde{f}_T$. Como demonstrado por Bai and Ng (2008a) nós podemos tratar $\widetilde{f}_t$ na equação de previsão como um vetor de regressores observados. As previsões geradas por essa metodologia tendem a superar outros modelos. De fato, as avaliações baseadas nas principais variáveis macroeconômicas constatam que as previsões de DI tendem a funcionar tão bem quanto costumam ser melhores que métodos alternativos, como combinação de previsão.

Deixe $\Lambda = (\lambda_1, ..., \lambda_N)'$ ser $N \times \ell$ a matriz de cargas fatoriais, e $e_t = (e_{1t}, ..., e_{Nt})'$ ser um vetor $N \times 1$. Então, na forma vetorial, o modelo de fator (3) pode ser escrito como

$$X = F\Lambda' + e$$

onde $e = (e_1', e_2', ..., e_N')$ e $T \times N$.

Sob grandes N e grandes T podemos estimar Λ e F por análises de componentes principais (PCA). Como foi demonstrado por Bai and Ng (2008a), aplica-se uma maneira sobre o modelo de fatores que eles fornecem uma classificação natural para N combinações lineares mutuamente ortogonais de X_t. Em outras palavras, o PCA produz uma solução que $\widetilde{F} = X\widetilde{\Lambda}/N$ com $\widetilde{\Lambda}'\widetilde{\Lambda}/N = I_\ell$, o que significa que as estimativas de fatores dos componentes principais são apenas combinações lineares das palavras em X_t. Essa proposta foi utilizada por Hendry and Madeley (2010) e Hendry (2012), que estudaram os efeitos das comunicações do Banco Central sobre variáveis macroeconômicas, no entanto, eles não direcionaram as palavras mais preditivas antes de estimar os fatores através do PCA.

Como o componente principal estima $\widetilde{F}_t$ são apenas combinações lineares de X_t, Bai and Ng (2008a) considera que a equaçao DI (2) pode ser escrita como

$$y_{t+1} = \alpha + \overline{\Gamma}' X_t + \varepsilon_{t+1}$$

onde $\overline{\Gamma}$ é uma versão restrita Γ na equação (1). Portanto, podemos interpretar as previsões de DI como usando todos N preditores da previsão na medida que $\overline{\Gamma}$ não possui nenhum elemento igual a zero. Nesse caso, temos um modelo denso em que todos os termos (séries temporais) têm um papel a desempenhar na previsão de y_{t+1}. Bai and Ng (2008a), consideraram refinamentos na metodologia de DI usando o que eles chamam de 'previsões do índice de difusão direcionadas'. Mais precisamente, busca-se uma equação de previsão

$$y_{t+1} = \alpha + \gamma' X_t^* + \varepsilon_{t+1}$$

onde o vetor $K \times 1$ X_t^* é um subvetor de X_t com $K << N$. Nesse caso, temos um modelo escasso em que nem todos os termos (séries temporais) têm um papel a desempenhar na previsão de y_{t+1}. Reescrevendo, temos;

$$y_{t+1} = \alpha + \Gamma^{*\prime} X_t + \varepsilon_{t+1}$$

onde o vetor $N \times 1$ Γ^* efetivamente atribui um peso zero aos preditores (palavras) que não são úteis na previsão y_{t+1}. Neste trabalho, usaremos o aprendizado de máquina supervisionado (SML) para identificar X_t^* (Γ^*), e mostraremos que essa ideia pode ser adaptada com objetivo de prever taxa de desemprego e produção industrial com dados de texto.

Além disso, a consideração por dados de texto leva a questão de quais preditores devem ser usados para formar índices de difusão. Portanto, utilizaremos o aprendizado de máquina superviosionado (SML) para identificar três

[5] Em nossa seção empírica, se o valor absoluto de λ_i for menor que 0.01, então λ_i será definido com zero.

formas diferentes de Γ^*. A primeira forma coloca um peso zero em termos que não são úteis para prever a taxa de desemprego e produção industrial.[6] A segunda (terceira) forma Γ^* coloca um peso zero naqueles termos que não estão negativos (positivamente) correlacionados com previsões futuras (y_{t+1}). Ao fazer isso, poderemos estimar fatores carregados com notícias positivas e negativas, bem como fatores carregados apenas com um tipo de notícia. O importante da metodologia proposta é que a identificação de X_t^* será adaptado ao problema da previsão da taxa de desemprego e produção industrial.

5 Metodologia

Nós empregaremos o método desenvolvido por Bai and Ng (2008a) para maximizar o poder preditivo de um dicionário, tornando-o mais útil para o problema de previsão de taxa de desemprego e produção industrial. Nossa metodologia pode ser resumida em três etapas: na primeira, transformamos as palavras que aparecem nos jornais publicados ao longo do tempo em valores numéricos (séries temporais) sem usar um dicionário pré-especificado (fixo). Essa representação numérica tem alta dimensionalidade, portanto, a redução da dimensionalidade deve ser empregada na próxima etapa. Na segunda etapa, usamos o aprendizado de máquina supervisionado (SML) para selecionar as palavras mais preditivas (séries temporais) e as usamos para calcular fatores por meio da análise de componentes principais. Finalmente, na terceira etapa, usamos os fatores em nosso modelo de índice de difusão para fazer previsões fora da amostra da taxa de desemprego e produção industrial. Esse procedimento em três etapas é repetido recursivamente no final da amostra (previsão fora da amostra recursiva), implicando que o conteúdo do dicionário mais preditivo pode variar ao longo do tempo. A seguir, descrevemos como os dados de texto entram em nosso banco de dados.

5.1 Dados Textuais

Utilizamos dados textuais mensais coletados no "Banco de Dados FACTIVA". Essas notícias são textos publicados no Estadão, O Globo, Folha de São Paulo e Valor Econômico, de março de 2002 a dezembro de 2019. Essa suposição de que a cobertura através dos veículos de comunicação observam eventos do mundo real e depois escolhem o que enfatizar em seu relatório, com o objetivo de construir sua reputação. Essa ideia é sugerida por Tetlock (2007) e Manela and Moreira (2017).

Primeiro, os jornais divulgam textos que não estão relacionados a economia. Portanto, aplicamos os seguintes filtros para descartar notícias não financeiras (econômicas):

- **Filtro de Pesquisa:** "O Globo, Folha de São Paulo - Todas as fontes ou Estadão, Valor Econômico - Todas as fontes". Esse filtro especifica que queremos apenas as notícias"O Globo, Estadão "ou "Folha de São Paulo".
- **Filtro de Assunto:** "Notícias sobre Mercado de Trabalho / Commodities/ Mercado Financeiro ou Notícias Econômicas". Esse filtro especifica que queremos apenas as notícias sobre "Mercado de Trabalho"/ "Commodities/ Mercado Financeiro ou Notícias Econômicas" para evitar notícias não relacionadas ao mercado de trabalho e notícias econômicas.
- **Filtro de Idioma:** "Português". Esse filtro especifica que queremos apenas os textos em Português.

Fizemos o download de texto por mês e os salvamos no que a literatura define como "um corpus", que é uma coleção de textos escritos, ou seja, um conjunto de notícias sobre o mercado de trabalho e notícias financeiras do O Globo, Estadão, Folha de São Paulo e Valor Econômico. Antes de realizar qualquer contagem de palavras, pré-processamos o texto bruto em várias etapas. O objetivo é reduzir o vocabulário a um conjunto de palavras mais significativas. Seguindo Hansen et al. (2017), identificamos colocações ou sequência de palavras que têm um significado específico. Por exemplo, "mercado de trabalho", corresponde a um único conceito econômico, mas é composto por duas palavras separadas. Para calcular as colocações das palavras, usamos o algorítimo desenvolvido por Straka and Straková (2017).

Também removemos palavras de parada. A função "stopword" no pacote "tm" do R remove *stopwords*[7] como: o, aquilo, qual, o que, etc. Outras funções padrão do pacote "tm' do R foram usadas para limpar os dados textuais. Descrevemos logo abaixo:

[6] Neste trabalho, "palavra"e "termo"serão usados de forma intercambiável. Um "termo"indica uma única palavra ou uma colocação (combinação de palavras).

[7] *stopwords* são palavras que não interessam na análise, como artigos e preposições por exemplo.

tolower: Como o R faz distinção entre maiúsculas e minúsculas, usamos o comando " tolower" para classificar palavras com letras maiúsculas e minúsculas (como casa e CASA) como palavras iguais.

removePunctuation: este comando foi usado para remover a pontuação.

removenumbers: usamos este comando para remover números.

stripWhitespace: este comando é usado para recolher vários espaços, para que "economia " e "economia " sejam tratadas como tendo o mesmo significado.

stemming: Esta função é usada para garantir que nosso banco de dados inclua apenas o radical das palavras. Por exemplo, palavras como " Economia" e " Economista" são contadas como " Econom". Se ambos aparecerem em um documento, sua soma será mostrada. Caso contrário, uma única palavra é contada. Neste artigo, consideramos apenas palavras derivadas. Observe que stemming identifica a raiz linguística de um termo, o que significa que o resultado de stemming não é necessariamente uma palavra em português. [8]

Por fim, assim como Hansen et al. (2017), classificamos as palavras restantes usando o frequência do termo em relação ao inverso da frequência do documento, conhecido por (tf-idf), que é uma medida que pune palavras raras e muito frequentes e eliminará todos os termos classificados com valor de 100 ou abaixo.

Após o pré-processamento acima, contamos as palavras restantes (colocações) no mês t e criamos um vetor $N \times 1$, $X_t = (X_{1t}, ..., X_{Nt})'$ onde cada elemento mostra a frequência que uma determinada palavra $i = 1, ..., N$ aparece em textos publicados no momento t. Para tornar esses dados de séries temporais adequados à estimativa de componentes principais, transformamos cada série para que se tornassem estacionarias e padronizamos os dados (diminuímos a média e dividimos pelo desvio padrão) antes de estimar o modelo, para evitar o viés de antecipação[9]. Essas transformações são realizadas usando observações de séries temporais de X_t disponível em cada origem de previsão $t = R, ..., T-1$. [10]

Assim, o processo explicado acima transforma as palavras em uma matriz $N \times T$ com elementos X_{is}, $i = 1, ..., N$ e $s = 1, ...t$ onde a dimensão transversal N, é maior que o número de períodos de tempo, T. Como usar essas informações não é imediatamente óbvio, porque, a menos que tenhamos uma maneira de ordenar a importância da série N na formação de expectativas condicionais (como em uma regressão automática), há potencialmente 2^N combinações possíveis para considerar. Além disso, com X_t denotando o vetor $N \times 1$ de observações no tempo t, os estimadores da regressão (1) são inviáveis quando $N >> T$, ou seja, a estimativa de

$$y_{t+1} = \alpha + \Gamma' X_t + \varepsilon_{t+1}$$

rapidamente se depara com o problema dos graus de liberdade devido à alta dimensão de X_t. Assim, as técnicas de redução de dimensão devem ser empregadas na próxima etapa. A segunda etapa usa o aprendizado de máquina supervisionado para segmentar os termos em X_t que não são importantes para prever y_{t+1} na origem de previsão t.

5.2 Selecionando os Termos mais Preditivos

Nesta seção, explicamos como selecionamos as palavras mais preditivas da taxa de desemprego e produção industrial. Nossa abordagem baseia-se na estimativa recursiva e é implementada aplicando *Elastic Net* para estimar a seguinte equação de previsão linear

$$y_{s+1} = X_s'\beta + \epsilon_{s+1} \tag{4}$$

onde y_{s+1} é o valor da taxa de desemprego e da taxa de crescimento da produção industrial no momento $s+1$, X_s é o vetor $N \times 1$ definido anteriormente e o vetor de coeficiente $\beta = (\beta_1, ..., \beta_N)'$ é estimado minimizando a seguinte função objetivo:

$$\min_{\beta} \sum_{s} (y_{s+1} - X_s'\beta)^2 + \lambda_1 \parallel \beta \parallel_{\ell_1} + \lambda_2 \parallel \beta \parallel_{\ell_2} \tag{5}$$

[8] Para a palavra stemming, usamos o algoritmo de stemming de Porter, disponível no pacote "tm" do R .

[9] Conhecido por *look ahead bias* ou viés antecipado, ocorre quando a amostra inteira é usada para calcular um parâmetro de modelo que é subsequentemente usado para gerar amostras fora da amostra previsões sobre um futuro em que o parâmetro estima extraiu informações.

[10] Por exemplo, se a origem da previsão for de $t = R$, as transformações acima usarão observações de séries temporais para cada $X_{it}, i = 1, ..., N$ com $t = 1, ..., R$.Em uma origem de previsão diferente $S > R$, as mesmas transformações são refeitas usando observações de séries temporais $t = 1, ..., R, ..., S$.

com $\|\cdot\|_{\ell_1}$ e $\|\cdot\|_{\ell_2}$ denotando o ℓ_1 e ℓ_2 norma, respectivamente. Estimamos a equação (5) regredindo y_{s+1} nos preditores X_s por $s = 1, ..., t-1$ e $t \in R, ..., T-1$. [11] A notação $t \in (R, ..., T-1)$ indica que a equação (5) é otimizada usando observações até uma origem de previsão específica. Selecionamos o $i-ésimo$ termo com (séries temporais) se seu coeficiente estimado, $\hat{\beta}_i$, for diferente de zero.

Em outras palavras, uma solução da equação (5) retornará os termos mais preditivos selecionados fora do vetor original $N \times 1$ de palavras e colocações. Salvamos os termos selecionados em $X_t^* = (X_{1t}^*, ..., X_{Kt}^*)'$ onde X_{it}^* mede a frequência dos termos mais preditivos $i = 1, ..., K$, para $K << N$, em documentos publicados até o momento $t \in (R, ..., T-1)$. Observe que o número de termos selecionados (K) é muito menor que o número de termos originais. O vetor X_t^* pode ser interpretado como o dicionário mais preditivo disponível até o momento $t \in (R, ..., T-1)$. Na prática, o dicionário X_t^* pode ser atualizado (equação 5 pode ser otimizada) em qualquer origem de previsão $R, ...T-1$ mas, para facilitar a interpretação de nossos resultados empíricos, selecionaremos o dicionário na origem de previsão R isto é, $t \in (R)$. Isso implica que as palavras serão selecionadas apenas uma vez, ou seja, no instante de tempo R (nossa primeira origem de previsão) e dando origem ao dicionário X_R^*. Este dicionário atualizado permanecerá o mesmo até a última observação fora da amostra prevista, T. Destacamos que os dicionários mais preditivos X_t^*, $t \in (R)$, são selecionados usando apenas observações dentro da amostra. Isso é necessário para evitar o viés de antecipação, como sugerido por Kalamara et al. (2020).

A estimativa da equação (5) depende dos parâmetros de ajuste λ_1 and λ_2. Há quatro casos a serem considerados. Se $\lambda_1 = \lambda_2 = 0$, então a função objetivo (5) torna-se igual à soma usual dos resíduos ao quadrado, o que não é possível porque o número de parâmetros (N) é muito maior que o tamanho da amostra (T). Se $\lambda_2 = 0$, então (5) torna-se um estimador do LASSO. Apesar do *LASSO* ser bem-sucedido na seleção de variáveis, no caso específico em que o número de palavras é maior que o tamanho da amostra, ou seja, $N >> T$, *LASSO* seleciona no máximo T palavras antes de saturar, excluindo, portanto, grandes porções do conjunto de informações condicional e potencialmente reduzindo a precisão das previsões. O caso com $\lambda_1 = 0$ corresponde à regressão de ridge, que não faz seleções de modelo porque não define coeficientes para zero. Por estas razões, Zou and Hastie (2005) sugere usar uma combinação de restrição-ℓ_1 e restrição-ℓ_2 que corresponde ao chamado estimador *elastic net*. Nesse caso, os coeficientes em β são reduzidos a zero de duas maneiras diferentes, promovendo a escassez e a estabilidade. Isso evita o super ajuste dos dados, definindo coeficientes sem importância para zero e identifica apenas as palavras preditivas mais relevantes. Os valores ótimos de λ_1 e λ_2 são obtidos a partir do procedimento sugerido no pacote GLMNET R (seção de regressão linear) desenvolvido por Trevor Hastie e Junyang Qian. [12]

Este procedimento de validação cruzada realiza uma seleção ex ante dos parâmetros de ajuste λ_1 e λ_2, essencial para evitar o excesso de ajustes. Na literatura de previsão, uma pequena lista de artigos que empregaram com sucesso *elastic net* inclui Bai and Ng (2008a), Li et al. (2015) e Lima et al. (2019) utilizaram *elastic net* para selecionar as palavras mais preditivas das atas publicadas pelo (FOMC).

Assim, a seleção dos termos mais preditivos (séries temporais) do vetor original $N \times 1$ é realizado através de *elastic net*. Além disso, podemos separar as palavras mais preditivas em dois grupos. O primeiro inclui apenas termos correlacionados positivamente com taxas de desemprego e produção industrial futuras y_{t+1}, e o segundo grupo inclui apenas termos correlacionados negativamente com y_{t+1}. Nós rotulamos esses sub dicionários como $X_{t^*}^{pos}$ e $X_{t^*}^{neg}$, respectivamente. Em suma, a cada período de atualização $t \in (R)$, selecionamos o dicionário mais preditivo de acordo com a seguinte regra:

(a) excluímos termos neutros, ou seja, descartamos a i-ésima série temporal se o seu coeficiente estimado na equação (5) for igual a zero, ou seja, $\hat{\beta}_i = 0$. Este é um procedimento padrão no aprendizado de máquina. Os termos restantes serão salvos em $X_t^* = (X_{1t}^*, ..., X_{Kt}^*)$ com $K << N$. Os elementos K de X_t^* são os termos mais preditivos disponíveis até $t \in (R)$;

(b) classificamos o $i-ésimo$ termo (séries temporais) como positivamente correlacionado com as taxas futuras se o seu correspondente ao coeficiente estimado na equação (5) é positivo, ou seja, $\hat{\beta}_i > 0$. Os termos correlacionados positivamente com taxas futuras são rotulados como notícias positivas e salvos no vetor $K_{pos} \times 1$ $X_{t^*}^{pos}$;

(c) classificamos o $i-ésimo$ termo (séries temporais) como negativamente correlacionado com taxas futuras se o valor estimado correspondente ao coeficiente na equação (5) é negativo, ou seja, $\hat{\beta}_i < 0$. Os termos que estão correlacionados negativamente com retornos futuros são rotulados como notícias negativas e salvos no vetor $K_{neg} \times 1$ $X_{t^*}^{neg}$.

Por fim, observe que $K = K_{pos} + K_{neg}$ para que a soma dos termos positivo e negativo deva ser igual à quantidade

[11] Terminamos em $T-1$ porque precisamos usar a observação T para avaliar as previsões feitas em $T-1$

[12] Empregamos validações cruzadas para dados dependentes como Elliott and Timmermann (2013)

total de termos selecionados pelo *elastic net* na equação (5).

5.3 Índice de Difusão com os Termos Selecionados

O modelo de índice de difusão é uma aplicação de procedimentos estatísticos desenvolvidos para o caso onde o número de séries temporais econômicas usadas para construir fatores comuns, K, e o número de períodos, T, são grandes e convergem para o infinito. O modelo de DI pressupõe que a correlação entre séries temporais econômicas seja capturada por alguns fatores comuns não observados.

Stock and Watson (2002) mostram que estimativas consistentes do espaço ocupado pelos fatores comuns podem ser construídas pela análise de componentes principais. Bai and Ng (2008b) mostram que as estimativas de mínimos quadrados das regressões de previsão aumentadas por fatores são $\sqrt{T}$ consistentes e assintoticamente normal e a pré-estimativa dos fatores não afeta a consistência das estimativas dos parâmetros do segundo estágio ou de seus erros padrão, o que significa que podemos tratar os fatores estimados como se representassem os fatores observados. Na prática, isso significa que o procedimento de estimativa para obter fatores pode ser executado independentemente de qualquer estimativa subsequente do modelo.

Assim, dados os termos selecionados $X_t^* = (X_{1t}^*, ..., X_{Kt}^*)'$, estimamos os fatores latentes recursivamente em cada origem da previsão $t = R, ..., S, ..., T-1$ pelo método dos componentes principais. Mais especificamente, em cada origem de previsão t calculamos uma matriz $K \times \ell$ de cargas fatoriais definidas como $\widetilde{\Lambda}_t = (\widetilde{\lambda}_{1,t}, ..., \widetilde{\lambda}_{K,t})'$ e a matriz $t \times \ell$ correspondente de fatores $\widetilde{F}_t = (\widetilde{f}_{1,t}, ..., \widetilde{f}_{t,t})'$. Como a estimativa ocorre recursivamente sobre $t = R, ..., S, ..., T-1$, nós indexamos $\widetilde{\Lambda}_t$ e $\widetilde{F}_t$ por t. Portanto, $\widetilde{f}_{s,t}$ denotam as i-ésimas observações sobre o vetor $\ell \times 1$ de fatores estimados usando dados até a origem da previsão t.

Os estimadores de componentes principais são definidos como

$$\left(\widetilde{\Lambda}_t, \widetilde{F}_t\right) = \{\lambda_k, f_s\} \arg\min \frac{1}{Kt} \sum_{k=1}^{K} \sum_{s=1}^{t} \left(X_{k,s}^* - \lambda_k' f_{s,t}\right)^2 \tag{6}$$

sujeito a condição de ortogonalidade $\frac{F_t' F_t'}{t} = I_\ell$, onde I_ℓ é uma matriz de identidade de dimensão ℓ. Intuitivamente, os fatores estimados no tempo t $\widetilde{F}_t$ são combinações lineares de cada elemento do vetor $K \times 1$ $X_t^* = (X_{1t}^*, ...X_{Kt}^*)$ onde a combinação linear é escolhida idealmente para minimizar a soma dos resíduos quadrados, com os resíduos sendo definidos como $X_t^* - \widetilde{\Lambda}_t \widetilde{F}_t$. Esta definição deixa claro que o ℓ fatores em $\widetilde{F}_t$ são carregados apenas com informações do dicionário mais preditivo X_t^*. Esse é um aspecto importante da metodologia proposta neste trabalho e será essencial para melhorar a previsão fora da amostra da taxa de desemprego e da produção industrial.

5.4 Previsão Fora da Amostra: Dados, Procedimento e Avaliação

Utilizamos dados mensais de Março de 2002 a Dezembro de 2019 da taxa de desemprego e da taxa de crescimento do índice de produção industrial. Além de construir os fatores do dicionário mais preditivo X_t^*, nós também construímos fatores $f_{s,t}$ usando um banco de dados de preditores macroeconômicos M_t. Este conjunto de dados foi projetado para ser representativo dos preditores mais utilizados na literatura nos modelos de previsão de DI.

Especificamente, nós utilizamos dados do Ipeadata, SGC do Banco Central do Brasil na web página [13] e realizamos as transformações sugeridas dos dados (explicadas no apêndice 1) para induzir a normalidade e a estacionaridade. O conjunto de dados é dividido em 8 grupos: produto e renda; mercado de trabalho; habitação; consumo; dinheiro e crédito; juros e taxas de câmbio; preços; mercado de ações. Nosso período abrange de 2002:03 a 2019:12. Excluímos as séries temporais com valores ausentes (detalhes no Apêndice 1) que nos deixaram com 52 séries temporais.

Assim, existem dois grupos de variáveis: um representando dados textuais e outro representando dados estruturados (macroeconômicos). Os fatores serão calculados separadamente para cada grupo e o elastic net será usado para selecionar as variáveis mais preditivas antes do cálculo dos fatores. Portanto, estamos aplicando a metodologia desenvolvida por Bai and Ng (2008a) para cada grupo de dados separadamente. Embora possamos segmentar os preditores em cada horizonte de previsão t, consideraremos uma data de seleção: $t \in (2013.03)$. A data escolhida antecede a maior recessão ocorrida na economia brasileira, entre 2014 e 2016. Isso implica que os termos mais preditivos serão selecionados em 2013.03. Da mesma forma, as variáveis macroeconômicas serão selecionadas

[13] Ipeadata e SGC Banco Central

apenas 2013.03. Adotamos esse procedimento para garantir a estabilidade dos parâmetros de ajuste e facilitar a interpretação dos resultados empíricos, mas outros períodos de seleção também podem ser considerados.

Para cada origem de previsão, t, temos $\ell_{max} = 8$ fatores e, em seguida, aplicamos a abordagem desenvolvida por Ahn and Horenstein (2013) para determinar o valor de ℓ. Então, seguindo, Bai and Ng (2008a), estimamos a seguinte equação de previsão pelo MQO[14]:

$$y_{t+1} = \alpha + \phi_t y_t + \beta' \widetilde{f}_{t,t} + \eta_{t+1} \tag{7}$$

que retenha apenas os fatores para os quais o valor p na equação acima é menor que 0.05. A equação de previsão é então estimada novamente, incluindo apenas os fatores selecionados. Uma previsão fora da amostra de y_{t+1} terá o formato:

$$\widehat{y}_{t+1|I_t} = \widehat{\alpha}_t + \widehat{\phi}_t y_t + \widehat{\beta}_t{}' \widehat{f}_{t,t} \tag{8}$$

onde I_t é o banco de dados (conjunto de informações) usado para estimar os fatores $\widetilde{f}_{t,t}$, e $\widehat{f}_{t,t} \subset \widetilde{f}_{t,t}$ são os fatores estatisticamente significativos na equação da previsão (7).

Este procedimento é repetido recursivamente até o final da amostra e, por esse motivo, $\widehat{\alpha}_t$, $\widehat{\phi}_t$ e $\widehat{\beta}_t$ são estimativas recursivas obtidas por MQO de α, ϕ e β, o que deixa claro que esses coeficientes variam ao longo do período fora da amostra. Esse ambiente de previsão é o que é usado na prática pelos agentes econômicos: uma reavaliação recursiva com previsões diretas em uma etapa à frente da amostra do desemprego e produção industrial. Em nossa aplicação empírica, geramos previsões fora da amostra recursivas da taxa de desemprego e produção industrial, y_{t+1} usando a equação (8) com fatores extraídos dos seguintes bancos de dados I_t:

(i) Termos originais (séries temporais), X_t. Nesse caso, não selecionamos os termos antes de calcular os fatores. Pode-se interpretar esse caso como um modelo denso, em que todos os termos (séries temporais) têm um papel a desempenhar. Nós rotulamos essa previsão como $\widehat{y}_{t+1|X_t}$;

(ii) Termos selecionados (séries temporais selecionadas), X_t^*.Nesse caso, temos um modelo escasso em que nem todos os termos (séries temporais) têm um papel a desempenhar. Aplicamos *elastic net* para selecionar os termos mais preditivos antes de calcular os fatores. Nós rotulamos essa previsão como $\widehat{y}_{t+1|X_t^*}$;

(iii) Termos selecionados correlacionados positivamente com previsões futuras, $X_{t^*}^{pos}$. Esse banco de dados inclui apenas os termos em X_t^* que foram correlacionados positivamente com y_{t+1}. Nós rotulamos essa previsão como $\widehat{y}_{t+1|X_{t^*}^{pos}}$;

(iv) Termos selecionados que estão correlacionados negativamente com previsões futuras, $X_{t^*}^{neg}$. Esse banco de dados inclui apenas os termos em X_t^* que foram correlacionados negativamente com y_{t+1}. Nós rotulamos essa previsão como $\widehat{y}_{t+1|X_{t^*}^{neg}}$;

(v) Banco de dados do Ipeadata e SGC do Banco Central, M_t. Usamos esse banco de dados para calcular os fatores comuns das 52 séries temporais listadas no Apêndice 1. A previsão que depende desse banco de dados será rotulada como $\widehat{y}_{t+1|M_t}$;

(vi) Preditores estruturados macroeconômicos, M_t^*. Primeiro usamos *elastic net* (equação 5) para identificar as variáveis mais preditivas no banco de dados do Ipeadata e SGC do Banco Central. Salvamos esses preditores no banco de dados M_t^* e depois o usamos para calcular fatores. A previsão que se baseia nesse banco de dados será rotulada como $\widehat{y}_{t+1|M_t^*}$.

As previsões de DI descritas acima seguem o mesmo procedimento e diferem entre si apenas em termos do banco de dados usado para estimar os fatores comuns; Todo o resto é idêntico nos modelos: as mesmas datas de seleção para X_t^* e M_t^*; mesma estimativa recursiva de fatores; mesmo procedimento usado para selecionar os fatores que inserem a equação de previsão (8); previsão direta em uma etapa; A única fonte de diferenciação entre essas previsões se deve exclusivamente ao banco de dados usado para calcular os fatores. Assim, nosso interesse é verificar

[14]Nosso objetivo é descobrir se os dados textuais ou macroeconômicos acrescentam poder preditivo ao benchmark AR(1).

se as informações disponíveis nos dados de texto são relativamente úteis para a previsão de desemprego e produção industrial fora da amostra e, se houver, quais grupos de palavras (termos) ajudam a aumentar o poder preditivo ao máximo.

Se as informações disponíveis nos dados de texto selecionados forem úteis para prever taxa de desemprego e produção industrial, então esperamos que os modelos de previsão baseados em X_t^* tenham desempenho superior. Tomar a previsão do AR(1) como referência também facilita a comparação das previsões baseadas no AR(1) mais dados de texto propostas com outras previsões que dependem de conjuntos de informações diferentes, como o AR(1) e dados macro, por exemplo.

A primeira medida de avaliação fora da amostra é o $RMSE$ relativo, que compara a previsão $\widehat{y}_{t+1|I_t}$ ao valor das previsões do modelo de referência $\overline{y}_{t+1}$, onde $I_t = (X_t; X_t^*; X_{t^*}^{pos}; X_{t^*}^{neg}; M_t; M_t^*)$ são os conjuntos de informações listadas acima. Reportamos o valor de $RMSE$ em termos relativos ao benchmark. O $RMSE$ pode ser calculado como:

$$RMSE = \frac{1}{R}\sum_{t=R}^{T-1}\frac{\left(y_{t+1} - \widehat{y}_{t+1|I_t}\right)^2}{\left(y_{t+1} - \overline{y}_{t+1}\right)^2} \tag{9}$$

Para testar a hipótese nula $RMSE < 1$, aplicamos o teste de Clark and West (2007). Este teste é realizado usando a estatística de teste desenvolvida por Clark and West (2007) que é obtida pela primeira estimativa.

$$g_{t+1} = \left(y_{t+1} - \overline{y}_{t+1}\right)^2 - \left[\left(y_{t+1} - \widehat{y}_{t+1|I_t}\right)^2 - \left(\overline{y}_{t+1} - \widehat{y}_{t+1|I_t}\right)^2\right], \tag{10}$$

depois regredimos $\{g_{t+1}\}_{t=R}^{T-1}$ sobre o intercepto e calculamos sua estatística t. O p-valor para um teste unilateral (cauda superior) é obtido com a distribuição normal padrão, conforme demonstrado por Clark and West (2007). Descobertas recentes de Gonçalves et al. (2017) justificam o uso dos valores críticos usuais quando fatores estimados são incluídos na equação de previsão (7). Em outras palavras, a incerteza da estimativa fatorial desaparece assintoticamente ao realizar inferência em fatores aumentados modelos de regressão. Portanto, a diferença entre as estatísticas de teste CW fora da amostra, com base nos fatores estimados, e aquelas baseadas nos fatores latentes, é assintoticamente desprezível [15]

6 Resultados

Nossas estimativas iniciais começam em 2003.3 e terminam em 2013.3 (132 Observações). A previsão fora da amostra é de 2013.4 a 2019.12, totalizando $P = 81$ observações fora da amostra. As observações fora da amostra não são iguais as observações dentro da amostra, visto que os dados não foram suficientes. Utilizamos o esquema recursivo de previsão fora da amostra.

A seguir, nos voltaremos para a análise fora da amostra. A tabela (1) relata os RMSE (Root Mean Squared Error), onde nossos resultados confirmam que a previsão (8) da abordagem sugerida por Bai and Ng (2008a) é beneficiada pois, os preditores são selecionados primeiramente pelo aprendizado de máquina supervisionado antes de calcular os fatores comuns. A afirmação é verdadeira, independentemente de a previsão (8) ser baseada em dados estruturados (M_t) ou não estruturados (texto). As previsões baseadas no modelo denso (M_t e X_t) são facilmente superadas pelo modelo de benchmark AR(1). As previsões baseadas em modelos esparsos usando dados textuais (M_t^* e X_t^*) tiveram resultados melhores para a previsão da produção industrial, no entanto, M_t^* não superou a previsão da equação (8) gerada por M_t para a variável emprego. Como citado anteriormente, a previsão baseada X_t^* superou outras previsões baseadas em DI. O resultado sugere que as informações contidas nos dados de texto (selecionados) podem ser úteis para melhorar as previsões da taxa de desemprego e produção industrial.

A tabela (2), que relata os resultados do teste de Clark West para os modelos propostos, corrobora ainda mais a ideia que o modelo preditivo sugerido por Bai and Ng (2008a) é beneficiado, levando a melhores resultados nas previsões. Neste trabalho, não afirmamos ou reivindicamos que as palavras que carregam os fatores mais preditivos tenham um significado de sentimento, mas podemos mostrar o que acontece com as previsões fora da amostra da taxa de desemprego e produção industrial quando o fator preditivo é calculado usando palavras carregadas de

[15] As estatísticas de Diebold e Mariano (1995) e West (1996) são frequentemente usadas para testar a hipótese nula, $R_{OS}^2 \leq 0$, entre modelos não aninhados. Para modelos aninhados, como os deste trabalho, Clark and McCracken (2001) e Clark and West (2007) mostram que essas estatísticas têm distribuição fora do padrão. Assim, o teste de Diebold-Mariano (DM) pode ser severamente subdimensionado sob a hipótese nula e ter baixo poder sob a hipótese alternativa.

sentimento dos dicionários sugeridos por Loughran and McDonald (2011), Tetlock (2007) e Correa et al. (2017).[16] Aplicamos o mesmo procedimento realizado para calcular as previsões anteriores baseadas em DI, mas substituímos os conjuntos de informações X_t e X_t^* pelo seu análogos de sentimento X_t^S e $X_{t^*}^S$, onde X_t^S mede a frequência normalizada de palavras carregadas de sentimentos que aparecem em notícias publicadas no Estadão, Folha de São Paulo, O Globo e Valor Econômico, enquanto que $X_{t^*}^S$ inclui apenas as palavras carregadas de sentimento selecionadas (ou seja, $X_{t^*}^S \subset X_t^S$). Portanto, o objetivo aqui é ilustrar se esses três dicionários são úteis para fazer previsões fora da amostra da taxa de desemprego e produção industrial. Os resultados da Tabela (1) sugerem que as previsões baseadas nos conjuntos de informações X_t^S e $X_{t^*}^S$ são superadas pelo modelo AR(1). Os resultados apontados não são surpreendentes, visto que Loughran and McDonald (2011) já demonstravam que, o uso das palavras não é idêntico na cobertura das notícias. Portanto, tomamos esse resultado como uma motivação empírica para a metodologia proposta no nosso trabalho, na qual o dicionário pode variar ao longo do tempo, incluindo ou excluindo novas palavras e colocações que ajudam a prever o desemprego e a produção industrial.

Os resultados sugerem que o método desenvolvido por Bai and Ng (2008a)[17], corroboram a ideia que, quando combinamos os dados de texto, aprendizado de máquina e índice de difusão, é o único jeito que consistentemente melhora as previsões fora da amostra do desemprego e produção industrial, de acordo com os resultados encontrados. O próximo passo é tentar explicar o que está impulsionando as previsões baseadas em dados de texto. Todas as previsões (8) são idênticas, entretanto algumas dependem de bancos de dados diferentes para calcular fatores. Dessa forma, o método utilizado para direcionar as palavras mais preditivas tem um papel fundamental neste trabalho. As tabelas (1) e (2), relatam as previsões baseadas em dados de texto que usa *LASSO* para selecionar as palavras mais preditivas. Embora o *LASSO* seja bem sucedido para selecionar as palavras mais preditivas, no caso onde há um grande número de preditores N, mas grande parte das informações é deixado de fora. O *elastic net* não apenas seleciona as palavras mais preditivas, como também garante que os termos igualmente preditivos não serão descartados de forma aleatória por que são altamente correlacionados com outros termos preditivos. Como visto anteriormente nas tabelas (1) e (2) os resultados mostram que o uso do *LASSO* embora melhore as previsões em relação ao AR(1) para a produção industrial, o desempenho da previsão não é tão grande, de acordo com o RMSE relativo. Portanto, o método proposto neste trabalho para prever a taxa de desemprego e produção industrial depende antes de selecionarmos corretamente os termos mais preditivos antes de calcular os fatores. Dessa forma, o *Elastic Net* como mostrado nesse trabalho sugere bons desempenhos para previsão da taxa de desemprego e produção industrial.

Em seguida, nós avaliamos as notícias positivas e negativas da cobertura dos jornais para melhorar as previsões fora da amostra da taxa de desemprego e produção industrial. Nós dividimos o banco de dados de notícias (X_t^*) em positivas ($X_{t^*}^{pos}$) e negativas ($X_{t^*}^{neg}$), respectivamente. O banco de dados anterior (último) são apenas as palavras que estão positivamente (negativamente) correlacionadas com a taxa de desemprego e produção industrial (y_{t+1}). De acordo com as tabelas (1) e (2) o RMSE e o Teste de Clark West do modelo de notícias negativas foram menores que o de palavras positivas. Portanto, as previsões da taxa de desemprego e produção industrial que dependem de notícias tem valores previstos melhores para o período fora da amostra. Isso implica que as notícias negativas desempenham um papel importante para melhorar a previsão fora da amostra, Soroka (2006) afirma que os efeitos da cobertura noticiosa da mídia em questões econômicas, tendem a relatar mais aumentos do que quedas na taxa de desemprego. Garz (2016) afirma que as más notícias sobre desemprego levam a um aumento no índice de percepções, o que implica que as pessoas percebem a economia de maneira mais negativa, e o mesmo não encontrou significância estatística para notícias boas. Embora os resultados das notícias negativas pareçam ter um bom poder preditivo ao longo do período fora da amostra, contudo as previsões que dependem do banco de dados completo (X_t^*) tem performance melhor em relação a bancos de dados parciais ($X_{t^*}^{neg}$ e $X_{t^*}^{pos}$). Portanto os resultados sugerem que o conjunto completo de informações direcionados pelo aprendizado de máquina (Elastic Net) possui mais conteúdo preditivo do que o conjunto parcial de informações com apenas notícias positivas ou negativas.

Além disso, queremos identificar os fatores com maior conteúdo preditivo para o desemprego e produção industrial (Gonçalves et al., 2017). Considerando os 8 fatores de dados de texto $\widetilde{f}_{t,t} = (\widetilde{f}_{1t},...,\widetilde{f}_{8t})$ como preditores em potencial ($\ell_{max} = 8$). Em vez de estimarmos todas as especificações possíveis de $2^8 - 1 = 255$ do modelo de previsão (7) aplicamos o teste de Ahn and Horenstein (2013) para determinar o número de fatores ($\kappa_{optimal} \leq 8$) e em seguida, reteve apenas os fatores $\ell_{optimal}$ com um p-valor menor que 0.05 na equação de previsão (7). Esse procedimento levou fatores $\widehat{f}_{t,t} \subset \widetilde{f}_{t,t}$ usados para gerar a previsão (8). Esse processo é repetido recursivamente durante o período

[16] Dividimos esses dicionários em palavras positivas e negativas

[17] Índice de difusão associado ao aprendizado de máquina

fora da amostra, e em cada origem de previsão $t = R, ..., T-1$ observamos quais dos 8 fatores potenciais foram selecionados para gerar a previsão (8).

A tabela 3 informa a porcentagem de vezes que um determinado fator foi selecionado para gerar a previsão (8). Os números mostrados na tabela 3 devem ser interpretados como uma fração da quantidade de observações fora da amostra. Consideramos um intervalo fora da amostra e calculamos os fatores usando o banco de dados (dicionários) disponíveis em $t^* = 2013.03$ para o período fora da amostra (81 observações), nossos resultados sugerem que o segundo fator é selecionado para taxa de desemprego, em seguida o primeiro fator. Para produção industrial, os resultados mostram que o primeiro fator foi selecionado em 100% seguido pelo segundo fator e terceiro fator, respectivamente. Os demais fatores não tiveram relevância para gerar a previsão (8). Portanto, a maior parte das informações textuais úteis para fazer previsões fora da amostra taxa de desemprego e produção industrial foram representada por poucos fatores.

Além disso, podemos classificar palavras que carregam os fatores mais preditivos usando dicionários que foram empregados para prever a taxa de desemprego e a produção industrial. Esses índices de sentimentos são construídos usando dicionários de palavras de sentimentos publicados por Tetlock (2007), Loughran and McDonald (2011) e Correa et al. (2017). Entretanto, não reivindicamos que as palavras que carregam os fatores mais preditivos tenham um significado de sentimento.

7 Conclusão

O modelo de DI (Índice de Difusão) empregado nesse trabalho têm sido amplamente utilizado na literatura de previsão macroeconômica devido a capacidade de lidar com grandes bancos de dados macroeconômicos. A utilização de dados de texto usados para construir novos preditores macroeconômicos para prever a taxa de desemprego e produção industrial. Esse trabalho mostra que o uso de dados de textos em um modelo de DI não melhorou as previsões da taxa de desemprego e produção industrial em relação ao modelo de benchmark, contudo houve ganhos quando selecionamos as palavras mais preditivas antes de calcular os fatores, permitindo que o dicionário seja atualizado no tempo.

Além disso, esses preditores são calculados usando informações de texto que inclui dados não estruturados de fontes de notícias dos jornais e relatórios econômicos. A metodologia utilizada permite extrair informações de textos e utiliza contagens de palavras com base em dicionários fixos. Contudo, quando aplicamos a lista de palavras para capturar informações preditivas acessíveis, o preditor que resulta dessa aplicação não consegue prever a taxa de desemprego e produção industrial. Loughran and McDonald (2011) aponta uma limitação, visto que a natureza das informações obtidas via mídia podem ou não ser válidas para amostras acima de 10Ks.

Nosso método também permite a interpretação e avaliação das notícias positivas e negativas na previsão fora da amostra da taxa de desemprego e produção industrial. Os resultados encontrados mostram que o método proposto gera previsões que superam outras previsões da produção industrial e da taxa de desemprego, tal qual os modelos gerados por uma variedade de modelos, como DI com preditores macroeconômicos e modelo de combinação com todos os outros modelos . Portanto, esse resultado é válido quando usamos medidas estatísticas como o ($RMSE$).

Os resultados apresentados neste artigo sugerem que combinação de modelo DI, aprendizado de máquina e dados textuais pode ser uma boa ferramenta para prever taxa de desemprego e produção industrial e outras variáveis macroeconômicas. As aplicações e os resultados encontrados mostram que é possível implementar outros tipos de modelos para prever variáveis macroeconômicas. É importante destacar também porque os fatores mais preditivos no modelo de DI são carregados com poucas palavras, umas das possibilidades é que esses dicionários de palavras de sentimento existente inclua uma quantidade razoável de colocações de palavras para prever a taxa de desemprego e produção industrial. Outra extensão que pode ser aplicada é o efeito do horizonte de previsão no longo prazo das notícias econômicas sobre a efetividade do que acontecesse na economia para tomada de decisão de investidores, consumidores, famílias e firmas de modo geral.

Referências

Ahn, S. C. and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3):1203–1227.

Alexopoulos, M., Cohen, J., et al. (2009). Uncertain times, uncertain measures. *University of Toronto Department of Economics Working Paper*, 352.

Andina-Díaz, A. (2007). Reinforcement vs. change: The political influence of the media. *Public Choice*, 131(1-2):65–81.

Bai, J. and Ng, S. (2008a). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146(2):304–317.

Bai, J. and Ng, S. (2008b). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146(2):304–317.

Birz, G. and Lott Jr, J. R. (2011). The effect of macroeconomic news on stock returns: New evidence from newspaper coverage. *Journal of Banking & Finance*, 35(11):2791–2800.

Blanchflower, D. G. (1990). Fear, unemployment and pay flexibility. Technical report, National Bureau of Economic Research.

Blood, D. J. and Phillips, P. C. (1995). Recession headline news, consumer sentiment, the state of the economy and presidential popularity: A time series analysis 1989–1993. *International Journal of Public Opinion Research*, 7(1):2–22.

Boyd, J. H., Hu, J., and Jagannathan, R. (2005). The stock market's reaction to unemployment news: Why bad news is usually good for stocks. *The Journal of Finance*, 60(2):649–672.

Carroll, C. D. and Dunn, W. E. (1997). Unemployment expectations, jumping (s, s) triggers, and household balance sheets. *NBER macroeconomics annual*, 12:165–217.

Carroll, C. D., Fuhrer, J. C., and Wilcox, D. W. (1994). Does consumer sentiment forecast household spending? if so, why? *The American Economic Review*, 84(5):1397–1408.

Clark, T. E. and McCracken, M. W. (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of econometrics*, 105(1):85–110.

Clark, T. E. and West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of econometrics*, 138(1):291–311.

Correa, R., Garud, K., Londono, J. M., and Mislang, N. (2017). Sentiment in central banks' financial stability reports. *Available at SSRN 3091943*.

Easaw, J. (2010). It's all 'bad'news! voters' perception of macroeconomic policy competence. *Public Choice*, 145(1-2):253–264.

Elliott, G. and Timmermann, A. (2013). *Handbook of economic forecasting*. Elsevier.

Feuerriegel, S. and Gordon, J. (2019). News-based forecasts of macroeconomic indicators: A semantic path model for interpretable predictions. *European Journal of Operational Research*, 272(1):162–175.

Garz, M. (2014). Good news and bad news: evidence of media bias in unemployment reports. *Public Choice*, 161(3-4):499–515.

Garz, M. (2016). Effects of unemployment news on consumers. Technical report, Working Paper.

Garz, M. (2018). Effects of unemployment news on economic perceptions–evidence from german federal states. *Regional Science and Urban Economics*, 68:172–190.

Gasper, J. T. (2009). Reporting for sale: the market for news coverage. *Public Choice*, 141(3-4):493.

Gonçalves, S., McCracken, M. W., and Perron, B. (2017). Tests of equal accuracy for nested models with estimated factors. *Journal of econometrics*, 198(2):231–252.

Hansen, S., McMahon, M., and Prat, A. (2017). Transparency and deliberation within the fomc: a computational linguistics approach. *The Quarterly Journal of Economics*, 133(2):801–870.

Hendry, S. (2012). Central bank communication or the media's interpretation: What moves markets? Technical report, Bank of Canada Working Paper.

Hendry, S. and Madeley, A. (2010). Text mining and the information content of bank of canada communications. *Available at SSRN 1722829*.

Hetherington, M. J. (1996). The media's role in forming voters' national economic evaluations in 1992. *American Journal of Political Science*, pages 372–395.

Hodrick, R. J. and Prescott, E. C. (1997). Postwar us business cycles: an empirical investigation. *Journal of Money, credit, and Banking*, pages 1–16.

Hollanders, D. and Vliegenthart, R. (2011). The influence of negative newspaper coverage on consumer confidence: The dutch case. *Journal of Economic Psychology*, 32(3):367–373.

Kalamara, E., Turrell, A., Kapetanios, G., Kapadia, S., and Redl, C. (2020). Making text count for macroeconomics: What newspaper text can tell us about sentiment and uncertainty. *2019 European Economic Association Meeting*.

Li, J., Tsiakas, I., and Wang, W. (2015). Predicting exchange rates out of sample: Can economic fundamentals beat the random walk? *Journal of Financial Econometrics*, 13(2):293–341.

Lima, L. R., Godeiro, L., and Mohsin, M. (2019). Time-varying dictionary and the predictive power of fed minutes. *Available at SSRN 3312483*.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65.

MacKuen, M. B., Erikson, R. S., and Stimson, J. A. (1992). Peasants or bankers? the american electorate and the us economy. *American Political Science Review*, 86(3):597–611.

Manela, A. and Moreira, A. (2017). News implied volatility and disaster concerns. *Journal of Financial Economics*, 123(1):137–162.

McQueen, G. and Roley, V. V. (1993). Stock prices, news, and business conditions. *The Review of Financial Studies*, 6(3):683–707.

Nadeau, R., Niemi, R. G., Fan, D. P., and Amato, T. (1999). Elite economic forecasts, economic news, mass economic judgments, and presidential approval. *The Journal of Politics*, 61(1):109–135.

Soroka, S. N. (2006). Good news and bad news: Asymmetric responses to economic information. *The journal of Politics*, 68(2):372–385.

Starr, M. A. (2012). Consumption, sentiment, and economic news. *Economic Inquiry*, 50(4):1097–1111.

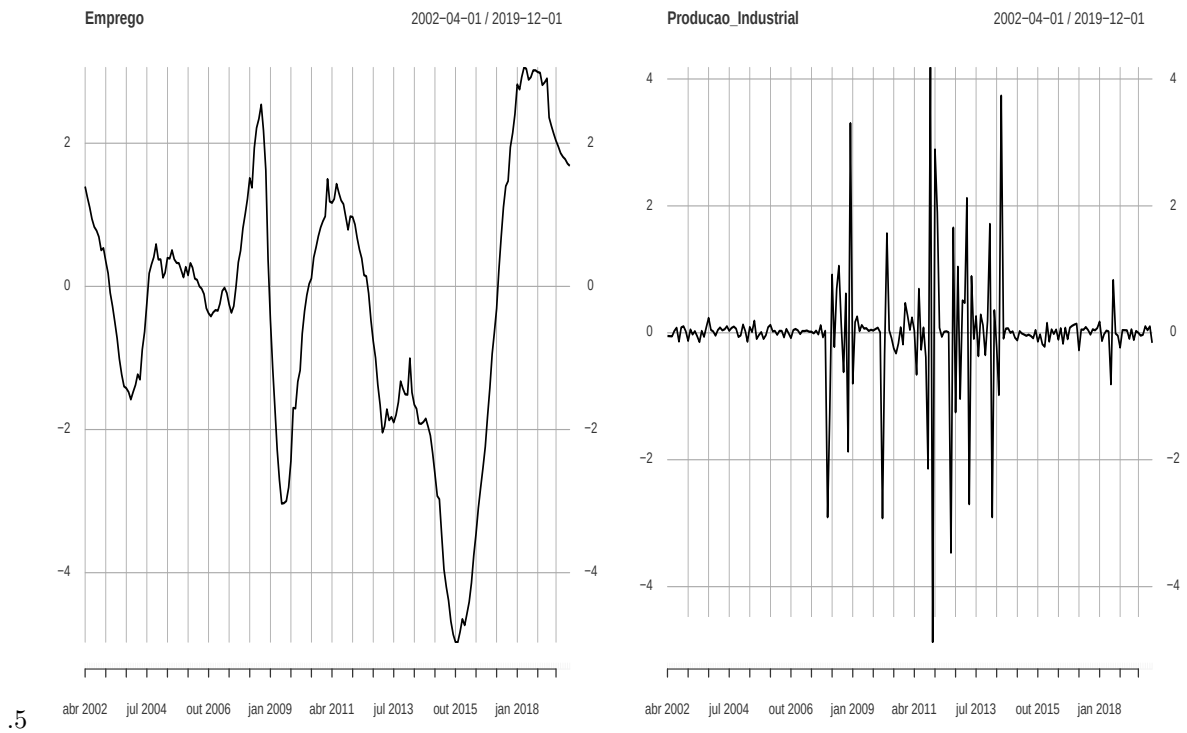
Stock, J. H. and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, 97(460):1167–1179.

Straka, M. and Straková, J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, 62(3):1139–1168.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Tabelas e Figuras



.5

Figura 1: Série Temporal da Taxa de Desemprego

Figura 2: Série Temporal da Taxa de Crescimento da Produção Industrial

O gráfico mostra a série temporal da taxa de desemprego, usamos o filtro HP de Hodrick and Prescott (1997) para remover o componente cíclico de uma séries temporal dos dados brutos.

O gráfico mostra a série temporal taxa de crescimento do Índice de Produção Industrial do Brasil para o período de 2002.3 a 2019.12.

Tabela 1: Previsão da Taxa de Desemprego e Produção Industrial fora da amostra

	Painel A: RMSE	
Modelo	Tx_desemp	Pro_idt
$\widehat{y}_{t+1\|X_{t^*}}$	0.9786	0.9741
$\widehat{y}_{t+1\|M_t}$	1.1325	1.0186
$\widehat{y}_{t+1\|X^S_{t^*}}$	1.0068	1.0089
$\widehat{y}_{t+1\|X^*_t}(LASSO)$	1.0111	0.9956
$\widehat{y}_{t+1\|M^S_{t^*}}$	1.1653	1.0125
$\widehat{y}_{t+1\|M^*_t}$	1.1557	1.0141
$\widehat{y}_{t+1\|X^{pos}_{t^*}}$	1.0244	1.0389
$\widehat{y}_{t+1\|X^{neg}_{t^*}}$	1.0011	1.0007
$\widehat{y}_{t+1\|X^t}$	1.0064	1.0078
$\widehat{y}_{t+1\|X_t}$	1.0080	1.0381
$\widehat{y}_{t+1\|X^{CP}_t}$	1.0298	0.9941

A tabela 1 mostra o RMSE relativo para os modelos de previsão em relação ao benchmark AR(1). Um RMSE menor que 1 indica que o modelo de previsão proposto supera o bechmark.

Tabela 2: Previsão da Taxa de Desemprego e Produção Industrial fora da amostra

Painel B: Teste de Clark West		
Modelo	Tx_desemp	Pro_idt
$\widehat{y}_{t+1\|X_{t^*}}$	0.00	0.17
$\widehat{y}_{t+1\|M_t}$	0.69	0.40
$\widehat{y}_{t+1\|X^S_{t^*}}$	0.96	0.71
$\widehat{y}_{t+1\|X^*_t}(LASSO)$	0.95	0.27
$\widehat{y}_{t+1\|M^S_{t^*}}$	0.65	0.67
$\widehat{y}_{t+1\|M^*_t}$	0.75	0.79
$\widehat{y}_{t+1\|X^{pos}_{t^*}}$	0.99	0.40
$\widehat{y}_{t+1\|X^{neg}_{t^*}}$	0.57	0.15
$\widehat{y}_{t+1\|X_t}$	0.98	0.66
$\widehat{y}_{t+1\|X_t}$	0.93	0.56
$\widehat{y}_{t+1\|X^{CP}_t}$	0.52	0.32

A tabela 2 mostra o teste Clark and West (2007) para os modelos de previsão em relação ao benchmark AR(1). A hipótese nula é que o modelo de previsão proposto é estatisticamente igual ao benchmark AR(1).

Apêndice 1: Dados do IPEADATA e SGC Banco Central

Obtemos o conjunto de dados do IPEADATA e SGC Banco Central na página da web[18]. Em seguida, realizamos as transformações sugeridas para uma série x_t: (1) nenhuma transformação; (2) Δx_t; (3) $\Delta^2 x_t$; (4) $\log x_t$; (5) $\Delta \log x_t$; (6) $\Delta^2 \log x_t$; (7) $\Delta(x_t/x_{t-1} - 1.0)$. O Conjunto de dados é composto por 8 grupos: Produto e Renda, Mercado de Trabalho, Habitação, Consumo, Dinheiro e Crédito, Taxas de Juros e de Câmbio, Preços e Mercado de Ações. O período abrange 2002:3 a 2019:12.

[18] Ipeadata e SGC Banco Central

Tabela 3: $\widehat{Y}_{t+1|X_{t^*}}$ Seleção de Fatores

	F1	F2	F3	F4	F5	F6	F7	F8
Taxa de Desemprego	32.09	80.24	0.00	0.00	0.00	0.00	0.00	0.00
Produção Industrial	100.00	18.51	11.11	0.00	0.00	0.00	0.00	0.00

A tabela 3 mostra a porcentagem de vezes que um fator foi incluído na equação de previsão (8). Os números mostrados devem ser interpretados como uma fração do número de observações fora da amostra. Os fatores foram extraídos com base em X_{t^*}. Nosso período fora da amostra vai de 2013.4 a 2019.12.

Tabela 4: Grupo 1: Produto e Renda

Código	SGC	Descrição
28504	IDPEM	Indicadores da Produção Extrativa Mineral
28505	IDPIT	Indicadores da Produção Indústria de Transformação
28506	IDPBC	Indicadores da Produção de Bens de Capital
28507	IDPBI	Indicadores da Produção Bens Intermediários
28508	IDPBCO	Indicadores da Produção de Bens de Consumo
28509	IDPBCD	Indicadores da Produção de Bens de Consumo Duráveis
28510	IDPSND	Indicadores da Produção de Semiduráveis e Não Duráveis
27574	IDC-BR	Índice de Commodities Brasil
27575	ICA	Índice de Commodities Agropecuária
27576	ICM	Índice de Commodities Metal
27577	ICE	Índice de Commodities Energia
22707	SBC	Saldo da Balança Comercial
28527	PTAV	Produção Total de Autoveículos
-	PIME	Produção Industrial Máquinas e Equipamentos
-	PIMV	Produção Industrial Móveis
-	UCII	Utilização da Capacidade Instalada
-	IDVNV	Índice de Vendas Nacionais do Varejo
-	IDVN-VM	Índice de Vendas Nacionais do Varejo - Motos, Veículos, Peças
-	PACL	Produção de Autoveículos e Comerciais Leves
-	PRC	Produção de Caminhões

Tabela 5: Grupo 2: Mercado de Trabalho

Código	IPEA/SGC	Descrição
-	SMR	Salário Mínimo Real
-	SM-PPC	Salário Mínimo - Paridade do Poder de Compra
-	SM	Salário Mínimo
-	ICEA	Índice de Condições Econômicas Atuais
-	IDEFIT	Índice de Emprego Formal - Indústria de Transformação
-	IDEFC	Índice de Emprego Formal - Comércio
-	IDEFCC	Índice de Emprego Formal - Construção Civil
-	IDEFEM	Índice de Emprego Formal -Extrativa Mineral

Tabela 6: Grupo 3: Habitação

Código	SGC	Descrição
1645	INPC-Habitação	Índice Nacional de Preços ao Consumidor - Habitação

Tabela 7: Group 4: Consumo, Pedidos e Estoques

Código	IPEA	Descrição
4393	ICC	Índice de Confiança do Consumidor
-	CBCA	Consumo de Bens de Capital
-	CBINT	Consumo de Bens Intermediários
-	CBCO	Consumo de Bens de Consumo
-	CBCD	Consumo de Bens de Consumo Duráveis
-	CBCSND	Consumo de Bens de Consumo Semi e Não-Duráveis
-	PPC-CF	Paridade do Poder de Compra - Consumo das Famílias

Tabela 8: Grupo 5: Moeda e Crédito

Código	IPEA/SGC	description
27841	M1	Meios de Pagamento M1
27842	M2	Meios de Pagamento M2
27813	M3	Meios de Pagamento M3
27815	M4	Meios de Pagamento M4

Tabela 9: Grupo 6: Juros e Taxas de Câmbio

Código	IPEA/SGC	Descrição
11	TXJ-Selic	Taxa de Juros - Selic
-	TXCB-BR/CN	Taxa de Câmbio Bilateral - BR/CN
-	TXCB-BR/AL	Taxa de Câmbio Bilateral - BR/AL
-	TXCB-BR/EUA	Taxa de Câmbio Bilateral - BR/EUA
-	CCC	Câmbio Contratado Comercial
20360	ITXCEN	Índice da Taxa de Câmbio Efetiva Nominal

Tabela 10: Grupo 7: Preços

Código	IPEA	Descrição
189	IGP-M	Índice Geral de Preços do Mercado
191	IPC-Br	Índice de Preços ao Consumidor Brasil
433	IPCA	Índice de Preços ao Consumidor Amplo
-	EXPI	Expectativas de Inflação

Tabela 11: Grupo 8: Mercado de Ações

Código	IPEA	Descrição
-	EMBI+BRASIL	EMBI+BRASIL - Risco País
-	RIBOV	Retorno da Ibovespa
-	ICA	Índice de Ações Ibovespa

Apêndice 2: Dicionário do Desemprego

Dicionário Positivo

adicion; alemã; alemanha; ameaça; atacado; barreira; bce; bug; chile; classificação; contração; csn; custa; derrubar; destinada; européia; europeu; extraordinária; humor; ien; informação; itália; legenda; leilõ; neto; onlin; recessão; reduziram; respectivament; saíram; sentimento; teme; colapso; seguradora; financi; preservar; ruin; microsoft; equador; quebra; rosa; tonelada; hipoteca; vaivém; álvaro; mineradora; madeira; ágora; boicot; irã; vivavoz; mercadoec; subprim; frisch; rangel; de o moeda; o bolsa de; em dólar; de o bolsa; em junho; perda de

Dicionário Negativo

acumulada; bahia; barri; cotada; eletrônica; estrangeiro; fusão; grécia; ibovespa; inglê; iniciou; novidad; observou; óleo; oscilação; reún; sessão; sexta; telecomunicaçõ; tend; tendem; terminou; terreno; troux; volátil; voltando; perfil; cruzeiro; estagflação; nasdaq; mostraram; avança; cauteloso; crescerá; sigla; rúa; pouso; verão; leia; nordest; vieira; cvm; lucent; cisco; turquia; rato; sênior; fabricio; cristina; tam; riscobrasil; pág; bernank; bmfbovespa; moeda único; de o estado; de o preço; política econômico; o fazenda;

Apêndice 3: Dicionário da Produção Industrial

Dicionário Positivo

britânica; exposição; grécia; inglê; irão; libra; madri; mostrado; precisará; quebrar; rússia; voltando; alimentar; aparelho; tragédia; bloomberg; sigla; vindo; sessõ; ipo; ribeiro; cristina; casai; dimitri; em o bolsa; não estar; october estar; words october estar; october estar de; o política; a o mercado; em o europa; grau de investimento"grau de; october folha; words october folha"reino unir; october folha de; paulo sérgio; paulo sérgio; dimitrir; paulo; bom para; ano; ano de; january globo globo; january globo; words january; globo;

Dicionário Negativo

aproveitar; comprado; emissõ; exportadora; generalizada; matériasprima; miséria; perderam; ponta; reún; unida; vendido; ajudando; fraqueza; fomc; eletrobrá; jpmorgan; cortou; falha; lehman; ivan; vasconcelo; onu; sadia; em alta de; taxa de desemprego; de pessoa; ficar em; september estar; palavras; september estar; september estar de; desde de; de o zona; dois dia; dizer; um dia; em todo; ano em; o região; ganho de;