*Artigo de Pesquisa* *DOI: http://dx.doi.org/10.1590/0101-41615013rrt*

# Previsão de Variáveis Macroeconômicas Brasileiras usando Modelos de Séries Temporais de Alta Dimensão

**Rafael B. Barbosa**[1]
**Roberto Tatiwa Ferreira**[2]
**Thibério Mota da Silva**[3]

## Resumo

Este artigo analisa o desempenho de modelos fatoriais de alta dimensão para prever quatro variáveis macroeconômicas brasileiras: duas variáveis reais, taxa de desemprego e o índice de produção industrial, e duas variáveis nominais, IPCA e IPC. Os fatores são estimados a partir de um conjunto composto por 117 variáveis macroeconômicas. Visando aumentar a performance dos modelos fatoriais são empregadas diferentes formas de extração e de utilização dos fatores. Três tipos de técnicas de aprendizado estatístico foram aplicados: métodos de *shrinkage*, combinações de previsões e seleção de previsores. Os fatores são extraídos de forma supervisionada e não supervisionada. Os resultados indicam que métodos de aprendizado estatístico melhoram o desempenho preditivo das variáveis econômicas brasileiras. Além disso, a combinação de técnicas de aprendizagem estatística e supervisão fatorial produzem melhores previsões que modelos que não utilizam fatores, modelos fatoriais com ou sem supervisão e modelos que utilizam apenas o aprendizado estatístico sem supervisão dos fatores. Única exceção a estas conclusões foram a variável índice de produção industrial que foi melhor prevista pelo modelo não supervisionado de fatores.

## Palavras-Chave

Previsão. Modelos de Fatores. Métodos de Shrinkage. Combinação de Previsão. Variáveis Macroeconômicas Brasileiras.

## Abstract

This paper analyzes the performance of high-dimensional factor models to forecast four Brazilian macroeconomic variables: two real variables, unemployment rate and industrial production index, and two nominal variables, IPCA and IPC. The factors are estimated from a data set containing 117 macroeconomic variables. We applied techniques to improve factor models

[1] Professor – Universidade Federal do Ceará
Endereço: Av. da Universidade, 2486 – Benfica – Fortaleza/CE – Brasil – CEP: 60020-180.
E-mail: rafael.barbosa@ufc.br – ORCiD: https://orcid.org/0000-0002-7365-9223.

[2] Professor – Universidade Federal do Ceará
Endereço: Av. da Universidade, 2762 – Benfica – Fortaleza/CE – Brasil – CEP: 60020-181.
E-mail: rtf2@uol.com.br – ORCiD: https://orcid.org/0000-0002-2529-686X.

[3] Professor – Universidade Federal do Piauí
Endereço: Rua João Crisóstomo e Silva, 159-309 - Ininga – Teresina/PI – Brasil – CEP: 64048-435.
E-mail: thiberiomota@gmail.com – ORCiD: https://orcid.org/0000-0002-5639-8658.
Recebido: 09/01/2018. Aceite: 27/08/2019.
Editor Responsável: Bruno de Paula Rocha

forecasts. Methods of statistical learning are applied aims to increase the performance of factors models. Three types of statistical learning techniques are used: shrinkage methods, forecast combinations, and selection of preditors. The factors are extracted using supervised and unsupervised version. The results indicate that statistical learning improves forecasts performance. The combination of statistical learning and supervised factor models is more accurate than all other models, with exception to the industrial production index which is best forecasted by unsupervised factor model without statistical learning.

**Keywords**
Forecast. Diffusion Index, Shrinkage Methods, Forecast Combination, Brazilian Macroeconomics

**Classificação JEL**
C1. C22. C52. C58.

## 1. Introdução

A recente disponibilidade de grandes conjuntos de dados econômicos requer a aplicação de técnicas capazes de utilizar essas informações de forma eficiente. Na realização de previsões de variáveis econômicas, diversos trabalhos têm apontado o poder de previsão de modelos fatoriais. Tais modelos extraem um pequeno número de variáveis, chamados de fatores, representativos da variabilidade dessas grandes bases de dados e que podem ser utilizados como previsores.

Os trabalhos de Stock e Watson (1999, 2002) mostram que modelos de fatores podem gerar previsões mais eficientes para a inflação e outras importantes variáveis macroeconômicas dos E.U.A. quando comparadas com as de modelos tradicionais como: modelos autorregressivos (AR), modelos vetoriais autorregressivos (VAR) e a curva de Phillips. Posteriormente a esses resultados, vários outros estudos têm indicado conclusões semelhantes, como, por exemplo, em Marcellino *et al.* (2008), Artis *et al.* (2005) e Dias, Pinheiro e Rua (2010).

Para o Brasil, Ferreira, Bierens e Castelar (2005) usam modelos fatoriais lineares e não lineares para prever a taxa de crescimento do PIB brasileiro e reportam que esses modelos geram previsões com menores erros quadráticos médios de previsão (EQMP) do que as dos modelos VAR e AR. Figueiredo (2010) aponta que modelos fatoriais produzem melhores previsões para a taxa de inflação brasileira, principalmente em horizontes de previsão mais longos.

Diferentes formas de aperfeiçoar o desempenho dos modelos de fatores têm sido testadas, como em Kim e Swanson (2014), Cheng e Hansen (2015), Stock e Watson (2012). Tais técnicas promovem um uso mais eficiente das variáveis através de uma pré-seleção ou por meio da combinação dos fatores. Estes métodos podem ser divididos em três grandes categorias: métodos de *shrinkage*, que reduzem o valor dos parâmetros estimados de acordo com uma penalidade ótima; métodos de combinação de variáveis, que combinam diversas variáveis visando aumentar o poder preditivo; ou métodos de agregação, que agregam as previsões de diversos modelos.

Poucos trabalhos têm investigado o desempenho preditivo desses modelos no Brasil. Medeiros *et al.* (2016), por exemplo, aplicam métodos *shrinkage* baseado no Lasso, para prever a taxa de inflação brasileira, mensurada pelo IPCA e IGP. Os autores mostram que este método gera melhores previsões que modelos autorregressivos. Garcia *et al.* (2016) estudam o poder de previsão de alguns métodos de *shrinkage*, como o LASSO, o ADA-LASSO, POST-LASSO em relação à previsão de especialistas para a inflação. Os resultados apontam que, a curto prazo, os métodos de *shrinkage* não são melhores que os especialistas. Entretanto, a longo prazo sempre existe algum modelo que prevê melhor a inflação do que os especialistas.

Apesar desses poucos trabalhos sobre o tema, nenhum deles investiga o desempenho preditivo de modelos fatoriais combinados com técnicas de aprendizado estatístico para as variáveis macroeconômicas brasileiras. O presente artigo, portanto, busca preencher essa lacuna ao investigar se modelos de fatores podem ter seu desempenho preditivo aprimorado por meio da introdução de técnicas de aprendizado estatístico aplicados ao contexto brasileiro. Os modelos serão utilizados para prever quatro variáveis econômicas, sendo duas nominais e duas reais: a taxa de inflação medida a partir do índice de preço ao consumidor (IPC) e do índice de preço ao consumidor amplo (IPCA), taxa de desemprego (Desemp) e índice de produção industrial (IPI).

Como *benchmank* será utilizado o modelo autorregressivo de quarta ordem (AR(4)). Dentre os modelos fatoriais, dois modelos não utilizam técnicas de aprendizado estatístico: modelo fatorial tradicional (FAT) e modelo fatorial autorregressivo aumentado (FAAC). A diferença decorre da inclusão de defasagens das variáveis-alvo no último. Os demais métodos serão obtidos a partir da aplicação de diferentes técnicas de aprendizagem estatística sobre os fatores estimados, divididos em: i. métodos de ponderação

como o bagging (BA), modelos bayesianos ponderados (BMA) e modelo de ponderação simples (SMA); ii. métodos de seleção de previsores como o critério de informação de Mallows (CMA), de jackknife (CJA) e de validação cruzada *leave-h-out* (LHO); e iii. métodos de *shrinkage*, como *Least angle regressions* (LARS), *elastic-net* (EN) e o *non-negative garotte* (NNG).

Outra forma de aperfeiçoar o poder de previsão dos modelos fatoriais é por meio da supervisão. Modelos fatoriais em sua abordagem tradicional extraem os fatores de forma independente da variável que será alvo da previsão. Assim, os mesmos fatores poderiam ser utilizados para prever variáveis com dinâmicas temporais distintas. Na abordagem supervisionada, os fatores são estimados de forma dependente da variável- alvo da previsão. Diferentes trabalhos analisam os ganhos da supervisão de modelos fatoriais como Bair *et al*. (2006), Boivin e Ng (2006), Bai e Ng (2008), Tu e Lee (2014), Boldrini e Hillenbrand (2015) e Giovannelli e Proietti (2014).

Antes da extração dos fatores, portanto, são aplicadas duas técnicas de supervisão: a seleção de variáveis por meio do algoritmo *Least Angle Regression* (LARS) e a supervisão com o método de componentes principais usando combinação de previsões (CFPC), desenvolvido por Tu e Lee (2014). Após o pré-tratamento das variáveis, os fatores são estimados e as demais técnicas aplicadas, como no caso sem supervisão.

Os fatores são estimados pelo método dos componentes principais (PC) a partir de uma ampla base de dados contendo 117 variáveis macroeconômicas brasileiras, com frequência mensal no período de 1996.5 a 2015.12. Todas as variáveis são transformadas em estacionárias.[1] As previsões são comparadas em termos de raiz quadrada dos erros quadráticos médios de previsão (REQMP) relativo ao modelo autorregressivo de ordem 4.

Os resultados indicam que a incorporação de aprendizado estatístico aos modelos fatoriais aumenta o poder de previsão de quase todas as variáveis, com exceção do IPI, que não apresentou ganhos no caso não supervisionado. A introdução de supervisão também aumentou significativamente a performance dos modelos fatoriais, com destaque para os métodos de seleção de variável. O IPI apenas aumentou seu poder preditivo no caso da supervisão por LARS. De forma geral, tanto a supervisão de modelos

[1] No apêndice encontra-se a descrição das variáveis utilizadas e dos testes realizados para estacionarizar as séries temporais.

fatoriais quanto a incorporação de aprendizado estatístico geraram previsões melhores.

Este trabalho contribui para a investigação de métodos de previsão para as variáveis brasileiras. A dinâmica das variáveis macroeconômicas em países emergentes, como o Brasil, apresenta características próprias que podem prejudicar o desempenho de modelos de previsão. Assim, espera-se que este trabalho contribua para gerar modelos de previsão mais adaptados ao comportamento das variáveis brasileiras, como em: Medeiros *et al.* (2016), Garcia *et al.* (2016), Ferreira, Bierens e Castelar (2005), entre outros. Além disso, este artigo contribui para compreensão de como modelos fatoriais podem ter seu poder preditivo melhorado ao se introduzirem técnicas de aprendizado estatístico como: Kim e Swanson (2014), Cheng e Hansen (2015), Stock e Watson (2012).

O restante do artigo está organizado da seguinte forma: a seção 2 apresenta os modelos fatoriais e a forma pela qual métodos de aprendizagem estatística podem ser utilizados na previsão. A seção 3 descreve os dados utilizados e a seção 4 discute os principais resultados. Por fim, a seção 5 apresenta as conclusões.

## 2. Modelos Fatoriais e Aprendizado Estatístico

### *2.1. Modelos Fatoriais*

Considere um painel de variáveis macroeconômicas representada por $X_{it}$ em que $i = 1,\ldots,N$ corresponde as $N$ variáveis dispostas no tempo $t = 1,\ldots,T$. Assuma que um pequeno número de fatores $r \ll N$ possa representar a maior parte da variabilidade de $X_{it}$, isto é:

$$X_{it} = \lambda_i' F_t + e_{it} \tag{1}$$

Onde $F_t$ é uma matriz de fatores de dimensão $r \times T$ , $\lambda_i$ é uma matriz de fatores de carga de dimensão $r \times N$ e $e_{it}$ é uma matriz de componentes idiossincráticos que representam o erro de aproximação de por $X_{it}$ por $\lambda_i' F_t$.

Os fatores serão estimados por meio do método de componentes principais (PC).[2] Bai (2003) apresenta uma propriedade interessante deste método. Mesmo diante de limitada dependência transversal entre os dados e o erro, e na presença de heteroscedasticidade,[3] a estimação por PC é consistente se $1/\sqrt{(NT)} \rightarrow 0$. Ou seja, ambas as dimensões de $X_{it}$ contribuem para a consistência das estimativas dos fatores.

Seja $y_t$ uma variável alvo para a previsão. Uma vez que os fatores tenham sido estimados $(\widehat{F_t})$, a equação de previsão para $h > 0$ períodos à frente é obtida por:

$$y_{t+h} = \beta(L)\widehat{Ft} + \gamma(L)Wt + e_{t+h|t} \quad (2)$$

Em que: $\beta(L)\ e\ \gamma(L)$ são defasagens polinomiais com ordens $p\ e\ q$, respectivamente, $W_t$ é um conjunto de variáveis exógenas definidas pelas defasagens de $y_t$. O modelo fatorial tradicional (FAT), também chamado de modelo de índice de difusão (STOCK e WATSON (2002)), considera que $W_t = 0$. Por sua vez, o modelo fatorial aumentado (FAAP) adota a formulação em (2).

Para a estimação dos modelos fatoriais é necessário assumir um número de fatores dado *a priori*. Esse número normalmente é selecionado através de critérios de informação, como o proposto por Bai e Ng (2002). Aqui será utilizado este critério para estabelecer o número máximo de fatores a ser extraído de $X_{it}$.

O objetivo deste artigo é verificar se métodos de aprendizado estatístico sobre a equação (2) geram ganhos em termos de poder de previsão. Para facilitar a exposição dos métodos de aprendizado estatístico, considere novamente a Equação (2). Considere que exista um número máximo de defasagens, tal que: $< p < p_{max}\ e\ 0 < q < q_{max}$[4]. Suponha que todos os regressores de (2) possam ser incluídos numa matriz de previsores $z_t$, tal que $z_t' = [F_t, W_t]$. Então, (2) pode ser reescrita como função de $z_t$,

$$y_{t+h} = z_t'b + \epsilon_{t+h|t} \quad (3)$$

[2] Ver Stock e Watson (2012) e Bai e Ng (2009) para mais detalhes sobre o método de componentes principais e formas alternativas de estimação dos modelos fatoriais.
[3] Ambas as características estão presentes em dados macroeconômicos.
[4] Serão permitidas no máximo quatro defasagens para p e q .

Em que: $b' = [\beta(L)\ \gamma(L)]$. Considere, adicionalmente, que existam $M$ possíveis modelos derivados da combinação de todas variáveis contidas em $z_t$. Isto é, cada modelo $m = 1, \ldots, M$ contém um subconjunto de $z_t$. Técnicas de aprendizado estatístico selecionam as variáveis e escolhem o modelo $m^*$que produz a melhor performance de previsão. A equação de previsão é dada por:

$$y_{t+h}(m^*) = z_t'(m^*)b(m^*) + \epsilon_{t+h|t}(m^*) \quad (4)$$

A previsão de $y_{t+h}$ considerando um conjunto de informações em $_t$ é obtida ao substituir $z_t$ por sua versão estimada $\widehat{z_t} = [\widehat{F_t}\ W_t]$ e pela estimação do número ótimo de defasagens $\hat{b}$.

$$\hat{y}_{t+h}(m^*) = \hat{z}_t'(m^*)\hat{b}(m^*) \quad (5)$$

O erro de previsão é dado por: $\hat{u}_{t+h|t}(m^*) = y_{t+h} - \hat{y}_{t+h|t}(m^*)$.

Os métodos de ponderação de previsões, por sua vez, obtêm pesos ótimos que ponderam as previsões $\hat{y}_{t+h|t}(m)$. Neste caso, a previsão de $y_t$ para o horizonte $h$ dado o conjunto de informação em $t$ é definida por

$$\hat{y}_{t+h|t}(w) = \sum_{m=1}^{M} w(m)\hat{y}_{t+h|t}(m) \quad (6)$$

Onde: $w(m) = \big(w(1), \ldots, w(M)\big)$ denota o vetor de pesos. Assume-se que e que $0 \leq w(m) \leq 1$ e que $\sum_{m=1}^{M} w(m) = 1$. Assim, métodos de combinação de previsões utilizam critérios específicos para escolher os pesos ótimos.

## *2.2. Técnicas de Aprendizado Estatístico*

### *2.2.1. Métodos de Shrinkage*

Os métodos *shrinkage* realizam uma seleção das variáveis com o objetivo de aumentar o desempenho preditivo da variável alvo. Esses procedimentos usualmente consideram um processo de estimação com um critério de penalização dos coeficientes. Nesse processo, os coeficientes estimados das variáveis consideradas irrelevantes para a previsão são reduzidos (encolhem-se) em direção a zero.

Ao se reduzir os valores absolutos dos parâmetros os métodos de *shrinkage* podem realizar a seleção das variáveis que produzem a melhor previsão para a variável alvo, procedimento chamado de regularização.

Em essência, a maior parte dos métodos de *shrinkage* realizam o seguinte processo de minimização penalizada:

$$min_b \left|\left|y_t - z_t' b\right|\right|^2 \tag{7}$$

$$sujeito\ a\ g(b) \leq s$$

Em que: $\left|\left|y_t - z_t' b\right|\right|^2$ corresponde a soma do quadrado dos resíduos, $s$ é um parâmetro de ajuste da penalização e $g(b)$ é uma função que penaliza os valores dos coeficientes $b$ da regressão em questão.

Esse trabalho utiliza três métodos de *shrinkage*: método do *Elastic Net* (EN), método *Non-Negative Garotte* (NNG) e o método do *Least Angle Regression* (LARS). Basicamente, a diferença entre os três métodos decorre da forma funcional da função de penalização e $g(b)$.

O método do *Elastic Net* (EN) foi inicialmente proposto por Zou e Hastie (2005) e adota uma forma ponderada de penalização dos parâmetros.

$$min_b \left|\left|y_t - z_t' b\right|\right|^2 \tag{8}$$

$$sujeito\ a\ \alpha \sum_{j=1}^{N} (b_j)^2 + (1-\alpha) \sum_{j=1}^{N} |b_j| \quad \leq 0$$

Em que $\alpha$ é o parâmetro de ajuste da penalização. O método do EN tem a vantagem de incorporar dois tipos de penalizações específicas: a penalização normada de ordem 2, dada por $\sum_{j=1}^{N} (b_j)^2$, e a penalização normada de ordem 1, dada por $\sum_{j=1}^{N} |b_j|$. No primeiro caso, se $\alpha = 1$, tem-se o método de shrinkage de Ridge e no segundo caso, se $\alpha = 0$, tem-se o método do LASSO, proposto por Tibshrani (1996).

Os valores de $\alpha$ são obtidos por validação cruzada. Entretanto, para séries temporais este procedimento não é apropriado devido à existência de autocorrelação entre as variáveis. Bai and Ng (2008) propõe utilizar critérios de validação como o BIC ou AIC para determinar os valores dos parâmetros de ajuste e esta será a abordagem adotada neste artigo.

O segundo método de *shrinkage* é o método do *Non-Negative Garotte* (NNG) proposto por Breiman (1995). Este método realiza uma regressão com penalização sobre o estimador de mínimos quadrados.

$$min_b \left|\left|y_t - z_t' b^{MQO}\right|\right|^2 + t\xi \sum_{j=1}^{N} v_j \qquad (9)$$
$$sujeito\ a\ v_i > 0; j = 1, \ldots, N$$

Em que: $b^{MQO}$ é a estimativa de mínimos quadrados ordinários (MQO), $\xi$ é o parâmetro de ajuste e $v(\xi) = \left(v_1(\xi), \ldots, v_N(\xi)\right)'$ é o fator de regularização que permite ponderar as estimativas produzidas pelo estimador de MQO. Isto é, o estimador de NNG é definido por: $b_j^{NNG}(\xi) = v_j(\xi) b_j^{MQO}, j = 1, \ldots, p.$

Tal estimador possui uma limitação importante referente à dependência explícita em relação ao estimador de MQO. Portanto, situações não adequadas ao estimador de MQO podem afetar negativamente as propriedades estatísticas do estimador de NNG. Por exemplo, em pequenas amostras o estimador de NNG pode ser prejudicado devido à inconsistência do MQO.

Yuan e Lin (2006) mostram que, em geral, o NNG produz estimativas consistentes dos parâmetros e produz adequada seleção de variáveis. Zhang *et al*. (2014), usando técnicas de simulação, encontraram que o estimador de NNG funciona apropriadamente para previsão, entretanto, a sua habilidade de seleção de variáveis é prejudicada em modelos com alta dimensão. Zou (2006) mostra que o NNG é consistente, se T → ∞ com N fixo. Kim e Swanson (2014) aplicam o NNG similarmente ao que será aplicado no presente artigo.

Por fim, o último método de *shrinkage* considerado neste artigo é o método do *Least Angle Regression* (LARS), desenvolvido por Effron *et al* (2004). O método do LARS consiste em um método de seleção de variáveis similar ao *Foward Stepwise Regression*.[5]

Esta técnica, a princípio, ordena as variáveis candidatas a participar do modelo de acordo com seu poder preditivo. Posteriormente, seleciona um modelo parcimonioso entre as variáveis melhor ranqueadas e usa este para

[5] *Forward Stepwise Regression* seleciona modelos por meio da adição de novas variáveis a cada processo de estimação. Após a inclusão de uma nova variável no conjunto de variáveis (conjunto ativo) são utilizados critérios para decidir qual conjunto de variáveis possui maior poder preditivo.

realizar a previsão. Este procedimento de seleção ocorre por meio da verificação da variável que é mais correlacionada com os resíduos de modelos anteriormente selecionados.

Em comparação com outros métodos de seleção, o LARS é considerado mais eficiente computacionalmente do que o LASSO. O algoritmo do LARS é rápido e possui previsão acurada em diferentes estruturas de dados.[6]

Gelper e Croux (2008) estendem o algoritmo do LARS para o contexto de séries temporais. A ideia principal da abordagem é selecionar os previsores em blocos compostos por séries temporais defasadas e não defasadas. Portanto, ao invés de considerar apenas um conjunto de variáveis, a seleção é feita sobre tais blocos. Isto implica que, caso um bloco de variáveis seja selecionado, então todas as variáveis contidas nele serão utilizadas para a previsão.

Kim e Swanson (2014, 2016) usam a metodologia de Gelper and Croux (2008) para regressores gerados, como os fatores. Para aplicar o LARS, primeiro se ordena os preditores incluindo um por vez. A cada inclusão é computado o erro de estimação por MQO.

Cada variável é incluída no conjunto ativo de acordo com sua correlação com a variável a ser predita. Então, o modelo estimado é obtido por critérios de informação. Seguindo Gelper e Croux (2008), o número de defasagens é escolhido ao se minimizar o critério de Schwartz (BIC). Aqui será adotado metodologia semelhante a Gelper e Croux (2008).

### *2.2.2. Métodos de Ponderação de Previsões*

Os métodos de ponderação de previsões aplicam um peso para cada previsão gerada a partir dos subconjuntos de variáveis como em (6) e com isso obtém-se uma nova previsão ponderada. A literatura empírica indica que existem ganhos em termos de poder de previsão devido à adoção de procedimentos de ponderação de previsões; ver Elliot e Timmermann (2016) para uma revisão.

[6] Ver Efron *et al.* (2004), Boveslstad *et al.* (2007), Saigo *et al.* (2007), Gelper e Croux (2008).

Aqui serão utilizados três procedimentos de ponderação de previsões: Ponderação via *bagging*, ponderação bayesiana e ponderação simples. No caso da ponderação simples será atribuído um mesmo peso para cada previsão gerada por algum subconjunto de variáveis. [7]

*Bagging* consiste num método de agregação via re-amostragem (*bootstrap aggregation-bagging*). Uma equação é reestimada em um conjunto de treino e sua previsão é realizada em uma amostra de teste. A amostra de dentro é obtida por meio de um procedimento de *bootstrap*. Seja $B^*$ o número de amostra de *bootstrap*, então a previsão de *bagging* é obtida ao se combinar com pesos iguais a tais previsões, isto é: $y_{t+h}^{bag} = \frac{1}{B}\sum_{i=1}^{B} \hat{y}_i^*$. Em que: $y_{t+h}^{bag}$ é a estimação de *bagging*, $\hat{y}_i^*$ é a estimação de MQO em cada amostra de *bootstrap*.

$$y_{t+h}^{bag} = W_t\hat{\beta}^{MQO} + \sum_{j=1}^{r} \psi(t_j)\,\hat{\beta}_{F_j}\hat{F}_j$$

Em que: $\hat{\beta}^{MQO}$ é a estimação de MQO de $y_{t+h}$ contra $W_t$; $\hat{\beta}_{F_j}$; é a estimação de MQO dos resíduos $e_{t+h} = y_{t+h} - W_t\hat{\beta}^{MQO}$ sobre $F_{t+h,j}, j = 1,\ldots,r$ representa o número de fatores estimados e $t_j$ é uma estatística $t - student$ associada a $\hat{\beta}_{F_j}$ e definida por $\sqrt{t}\hat{\beta}_{F_j}/s_e$, com $s_e$, com o estimador de Newey-West do desvio padrão de $\hat{\beta}_{F_j}$. No caso específico do *bagging* o parâmetro $\psi(t_j)$ é definido por:

$$\psi(t_j) = 1 - \Phi(t + c) + t - 1[\phi(t - c) - \phi(t + c)]$$

Onde: c é um valor crítico previamente definido, aqui será adotado c= ± 1.96, φ é densidade normal padrão e Φ é função de distribuição acumulada da distribuição normal.

A ideia é que as previsões usando as variáveis fatoriais sejam incorporadas considerando a sua relevância estatística, mensurada pela significância de $\hat{\beta}_{F_j}$.

O segundo tipo de modelo de ponderação adotado é modelo bayesiano ponderado (BMA). Estes modelos buscam combinar diversas previsões utilizando o paradigma bayesiano no qual os pesos de cada previsão são dados pela distribuição *a posteriori* de cada submodelo. Diversos autores

[7] Visando reduzir o custo computacional, será utilizado o procedimento de *Forward Stepwise Regression*, isto é, serão incluídas uma a uma as variáveis e geradas as previsões. Posteriormente, cada previsão será ponderada igualmente.

têm apontado o relativo sucesso do BMA em realizar previsões para variáveis macroeconômicas. Raftery *et al.* (1997), Fernandéz *et al.* (2001) e Ley e Steel (2009) são alguns exemplos.

Seja $M$ um conjunto total de modelos e considere $P(m)$ a distribuição *a priori* de cada modelo. O conjunto de dados fornecerá a distribuição *a posteriori* $P(m|\hat{z}_t)$. Segue do teorema de Bayes que: $P(m|\hat{z}_t) \propto I_{\hat{z}_t} \times P(m)$ , em que $I_{\hat{z}_t}$ é a função de verossimilhança do modelo $m$ , definida por:

$$I_{\hat{z}_t} = \int p(\hat{z}|b(m), m) \times p(b(m)|m)\, db(m)$$

Com $p(\hat{z}|b(m), m)$ sendo a distribuição marginal de $\hat{z}_t$ dado $b(m)$ e $m$; $p(b(m)|m)$ a distribuição marginal de $b(m)$ dado $m$. Como $M$ pode ser grande, então, utiliza-se o método *Markov Chain Monte Carlo (MCMC)* para simular e escolher os pesos.

A distribuição *a posteriori* do modelo m dado o conjunto de observações $\hat{z}_t$ é, portanto:

$$p(m|\hat{z}_t) = \frac{p(\hat{z}|m)p(m)}{\sum_{m=1}^{M} p(\hat{z}|m)p(m)}$$

Dessa forma, a média ponderada pela distribuição *a posteriori* para cada previsão $\hat{y}_{t+h}$ é dada por:

$$E(\hat{y}_{t+h}|\hat{z}_t) = \sum_{m=1}^{M} \hat{y}_{t+h} p(m|\hat{z}_t) \quad (10)$$

Os métodos bayesianos de ponderação de previsões possuem a vantagem de atribuir pouco peso a modelos improváveis. Serão adotadas duas formas funcionais possíveis para a distribuição *a priori*: $P(m) = (1/T)$ e $P(m) = (1/N^2)$. Para o primeiro tipo de distribuição *a priori* tem-se o modelo BMA 1. Enquanto o segundo tipo de distribuição *a priori*, o modelo BMA 2. Tais especificações foram adotadas em outros exercícios similares para dados econômicos, como Kim e Swanson (2014, 2016) e Koop e Portter (2004).

Fernandez, Ley e Steel (2001) examinam as propriedades estatísticas de diversas formas funcionais para as probabilidades *a priori*. No caso do

BMA 1, os autores mostram que tal probabilidade *a priori* gera uma estimação consistente do melhor modelo de previsão e que assintoticamente possui um comportamento semelhante aos critérios BIC e Hannan-Quinn.

Por sua vez, a segunda forma funcional para a probabilidade *a priori*, BMA 2, implica um comportamento assintoticamente semelhante ao critério de risco de inflação (RIC) de Foster e George (1994). Além desta relação com critérios de informação, tais formas funcionais são baseadas nos dados, seguindo de perto os princípios dos métodos bayesianos empíricos.

### *2.2.3.* Métodos de *Seleção de Previsores*

Os métodos de seleção de previsores utilizados neste artigo aplicam critérios de informação especialmente desenvolvidos para variáveis geradas, como é o caso dos fatores. A introdução de variáveis não diretamente observadas, mas sim estimadas, na equação de previsão inviabiliza a aplicação direta de critérios de informação tradicionais, como o BIC, o AIC e o Hannan-Quinn. Isso se deve a introdução de erros obtidos durante o processo de estimação.

Além disso, especificamente para modelos fatoriais, Bai e Ng (2008) argumentam que a aplicação de critérios de informação tradicionais sobre os fatores não é apropriada, pois não existe uma ordenação natural entre os fatores, ao contrário do que ocorre em relação a variáveis observáveis e suas defasagens. Assim, os critérios de informação utilizados neste artigo foram desenvolvidos visando a melhor adequação a modelos fatoriais.

Cheng e Hansen (2015) estudam o comportamento do critério de informação de Mallows (CMA)[8] utilizando regressores gerados, como os fatores comuns considerados neste artigo. Os autores mostraram que a utilização de fatores previamente estimados eleva o erro de previsão devido a possíveis erros de estimação. Todavia, como o critério de Mallows depende do erro de previsão, este também se eleva, mantendo a capacidade comparativa sobre o melhor subconjunto de previsores.

[8] Hansen (2007) propôs a seleção de modelos baseado no critério de Mallows (1973). Segundo Hansen (2008) as previsões realizadas pela seleção de modelos feita pelo critério de Mallows (CMA) possui relativo sucesso mesmo em previsões em que h>1 e quando há heteroscedasticidade condicional.

O critério de Mallows é definido como a média quadrática do erro de previsão mais um termo que depende do tamanho da amostra, da variância do erro de previsão e do número de regressores contidos no *m-ésimo* modelo.

$$C_t(m) = \frac{1}{T}\sum_{t=1}^{T} \hat{\varepsilon}_t(m)^2 + \frac{2\hat{\sigma}_T^2}{T} k(m) \tag{11}$$

Em que: $k(m)$ é o número de regressores em cada modelo $m$; $\hat{\sigma}_T^2$ é a variância estimada do erro de previsão dada por $\hat{\sigma}_T^2 = E(\hat{\varepsilon}(m)^2$. Seguindo Cheng e Hansen (2015), o estimador da variância será definido por: $\hat{\sigma}_T^2 = \big(T - k(m)\big)^{-1} \sum_{t=1}^{T} \hat{\varepsilon}_t\,(M^*)^2$ em que $M^*$ é o modelo com maior número de parâmetros.

O modelo selecionado será aquele que minimiza o critério de Mallows, isto é:

$$m^* = \arg\min C_t(m) \tag{12}$$

Critérios de informação de Jacknife (CJA) foram propostos por Hansen e Racine (2012) que utilizam o critério de validação cruzada *leave-one-out*. Isto é, *m* modelos são estimados e a última observação é deixada fora da amostra em cada procedimento. Adicionalmente os autores mostram que CJA e CMA são aproximadamente equivalentes em presença de homocedasticidade. Entretanto, CJA tem melhor performance do que o CMA diante de heteroscedasticidade.

O comportamento desses critérios de informação não é ideal em horizontes de longo prazo h > 1. Assumindo que os erros seguem um processo MA(h−1) , Hansen (2010) mostra que o erro quadrático médio de previsão (EQMP) depende do horizonte de previsão (h), dos parâmetros de MA (h−1) e do tamanho da amostra.

Para superar esse problema, Hansen (2010) propôs o método de validação cruzada *leave-h-out*, em que são deixados de fora da amostra de treino h períodos. No mesmo artigo, Hansen demostrou que tal critério tem performance superior que o CJA, CMA e outros critérios clássicos diante de heteroscedasticidade e múltiplos horizontes de previsão.

Os três critérios de informação foram testados por Cheng e Hansen (2015), Rahal (2015), Liu e Kuo (2016) e Yin, Liu e Lin (2016) em modelos fatoriais apresentando resultados interessantes em termos de poder de previsão.

De forma sintética, o Quadro 1 apresenta todos os métodos utilizados, suas respectivas siglas e as técnicas de aprendizado estatístico que representam.

**Quadro 1 - Modelos de Previsão**

| Modelos | Descrição | Técnica de Aprendizado |
|---|---|---|
| AR(4) | Modelo autorregressivo de ordem 4 | *Benchmark* |
| FAAR | Modelo fatorial aumentado | Fatorial |
| FAT | Modelo fatorial | Fatorial |
| LARS | Modelo least angle regression | *Shrinkage* |
| EN | Modelo elastic net | *Shrinkage* |
| NNG | Modelo non-negative garotte | *Shrinkage* |
| Bagging | Modelo de ponderação bagging | Ponderação |
| BMA 1 | Modelo de ponderação bayesiana P( m ) =( 1/T ) | Ponderação |
| BMA 2 | Modelo de ponderação bayesiana P( m ) =( 1/ N 2 ) | Ponderação |
| SMA | Modelo de ponderação simples | Ponderação |
| CMA | Critério de informação de Mallows | Seleção |
| CJA | Critério de informação de Jackniffe | Seleção |
| LHO | Critério de informação leave-h-out | Seleção |

Nota: A Tabela 1 apresenta todos os métodos utilizados para previsão neste artigo. Os métodos foram escolhidos de acordo com a sua adequação a modelos fatoriais. Cada um destes métodos será utilizado para prever as quatro variáveis alvo e serão aplicados em duas formas de extração dos fatores: estimação supervisionada e não supervisionada.

## *2.3. Métodos de Supervisão de Fatores*

Bair *et al*. (2006) e Boivin e Ng (2006) mostram que a seleção das variáveis por meio da análise da correlação entre as variáveis-alvo e as variáveis contidas na base de dados pode melhorar a eficiência preditiva de modelos fatoriais.

Em Bai e Ng (2008) a seleção de variáveis, por sua vez, é feita através de métodos *hard* e *soft thresholds*. No primeiro caso, as variáveis são selecionadas por critérios de informação como o de Akaike (AIC) e o Bayesiano (BIC). No segundo caso, são aplicados métodos de regularização como o LASSO (*Least Absolute Selector and Shrinkage* Operator) e o LARS (*Least*

*Angle Regression*) para selecionar previamente sobre de quais variáveis serão extraídos os fatores.

Outra denominação utilizada para esse procedimento é a de modelos de fatores supervisionados (Hastie *et al.* (2013)). Na técnica de supervisão proposta por Tu e Lee (2019), por exemplo, os fatores são estimados de uma base de dados transformada, obtida pela estimação da influência de cada variável contida em X it sobre a variável alvo da previsão. Este método é chamado de Componentes Principais usando Combinação de Previsões (CFPC). Hillebrand *et al.* (2018) utilizam forma semelhante de supervisão para prever inflação e produto dos EUA.

O método do CFPC realiza uma regressão para cada uma das variáveis contidas em $X_{it}$ sobre a variável-alvo $y_t$. A partir destas regressões são feitas previsões para os valores futuros de $y_t$ a partir de cada variável contida em $X_{it}$. Os fatores são extraídos desse conjunto de previsões. O CFPC, portanto, seleciona os fatores a partir da contribuição individual que cada variável $x_{it} \subset X_{it}$ proporciona sobre a previsão da variável-alvo.

A principal vantagem da adoção de procedimentos supervisionados deve-se a estimação dos fatores considerando qual variável será alvo da previsão. Nos métodos não supervisionados de estimação dos fatores, estes podem ser utilizados para prever qualquer tipo de variável-alvo.

Neste trabalho são usadas para previsão versões supervisionadas dos modelos fatoriais com e sem aprendizado estatístico pelos métodos LARS e CFPC. O objetivo é verificar se o poder de previsão pode ser melhorado com o emprego de supervisão dos fatores, obtida ou pela prévia redução do número de variáveis contidas em $X_{it}$ ou pela extração dos fatores de uma base de dados transformada.

### *2.4. Avaliação da Previsão*

Seja a função perda quadrática associada a cada modelo k=1,...,K :

$$L_{kt} = \left(y_{t+h} - \hat{y}_{k,t+h|t}\right)^2 \tag{13}$$

A raiz do erro quadrático médio de previsão (REQMP) é definida por:

$$REQMP_k = \sqrt{\frac{1}{P}\sum_{j=t+h}^{T}\left(y_j + \hat{y}_{k,j|t}\right)^2}$$

Em que: P é o tamanho da amostra de fora.

Os melhores modelos de previsão são aqueles que apresentam menor REQMP. Para comparar o poder de previsão frente ao modelo utilizado como *benchmark*, será utilizada a raiz do erro quadrático médio de previsão relativo ao modelo AR(4), [9] isto é:

$$REQMP_{AR(4),K} = \sqrt{\frac{\frac{1}{P}\sum_{j=t+h}^{T}\left(y_j+\hat{y}_{k,j|t}\right)^2}{\frac{1}{P}\sum_{j=t+h}^{T}\left(y_j+\hat{y}_{AR(4),j|t}\right)^2}} \quad (14)$$

Onde: $\hat{y}_{AR(4),j|t}$ é a previsão gerada pelo modelo autorregressivo de ordem 4. Desta forma, todas as previsões serão comparadas ao modelo AR(4). Como o *benchmark* não se modifica em diferentes estimações dos fatores supervisionados e não supervisionados, então o $REQMP_{AR(4),K}$ permitirá comparar os modelos em ambas as formas de estimação.

## 3. Base de Dados

A base de dados contém 117 variáveis macroeconômicas com frequência mensal incluindo: índices de preços setoriais, dívidas externas e governamentais, alguns componentes da balança de pagamentos, importação e exportação de bens setoriais, salário, taxa de juros de longo prazo, indicadores financeiros, índices da atividade econômica, taxa de desemprego, algumas variáveis de economia internacional como importação e exportação dos EUA, taxa de juros de longo prazo dos EUA, consumo de energia elétrica e combustíveis, índices de produção setorial e outros. [10]

[9] A escolha do AR(4) foi realizada por meio do critério de informação BIC. Este modelo foi o que melhor se ajustou as variáveis alvo da previsão.

[10] No apêndice, A.1, encontra-se a tabela com a descrição de cada variável, da fonte e do tipo de transformação realizada. A base de dados utilizada e os códigos utilizados na pesquisa podem ser solicitados por email para replicação dos resultados.

Espera-se que tais variáveis possam representar adequadamente o comportamento macroeconômico da economia brasileira. O procedimento de escolha das variáveis considerou a disponibilidade das séries temporais e seguiu de perto estratégias adotadas por outros estudos, como Stock e Watson (2002) e Ng e McCraken (2015).

Será adotado o esquema recursivo de previsão no qual as séries temporais serão particionadas em duas subamostras: amostra de dentro e amostra de fora. [11] Nos trabalhos de previsão, usualmente a amostra é separada em duas partes. A amostra de dentro (*in sample*) é usada para estimativas iniciais dos parâmetros, seleção de modelos, dentre outros processos. A amostra de fora (*out-of-sample*) é usada para avaliação das previsões dos modelos.

Os modelos fatoriais serão inicialmente estimados na amostra de dentro e a previsão será realizada para as observações na amostra de fora. A cada previsão realizada na amostra de fora, esta será incorporada a amostra de dentro para gerar a nova previsão. Este procedimento irá continuar até que todas as observações da amostra de fora tenham sido previstas.

A base de dados inicia-se em 1996.5 e termina em 2015.12, perfazendo 251 observações temporais. A amostra de dentro inicia-se em 1996.5 e termina em 2006.12, contendo 142 meses. Já a amostra de fora, inicia-se me 2007.1 e termina em 2015.12, contendo 109 meses.

As variáveis-alvo incluem dois índices de preços, o índice de produção industrial e a taxa de desemprego. Os índices de preços são: índice geral de preços ao consumidor (IPC) e índice de preço ao consumidor amplo (IPCA).

Foram realizadas transformações nas variáveis para obter a estacionariedade das séries temporais antes da estimação dos fatores. O procedimento para alcançar a estacionariedade foi baseado em Stock e Watson (2002, 2012). Primeiro, as variáveis são testadas por meio do teste Dickey-Fuller Aumentado (ADF). Se a hipótese nula não é rejeitada, então é aplicada uma transformação sobre a séries e novamente é verificada a estacionariedade por meio do teste ADF. Esse procedimento repete-se até a variável rejeitar a hipótese nula de não estacionariedade. Dentre as transformações aplicadas estão a primeira ou a segunda diferença, logaritmo e logaritmo

[11] Nos trabalhos de previsão, usualmente a amostra é separada em duas partes. A amostra de dentro (*in sample*) é usada para estimativas iniciais dos parâmetros, seleção de modelos, dentre outros processos. A amostra de fora (*out-of-sample*) é usada para avaliação das previsões dos modelos.

da diferença. Ao final, a robustez da estacionariedade da série temporal é verificada por meio do teste de estacionariedade KPSS, desenvolvido por Kwiatkowski *et al* (1992).

## 4. Resultados

### *4.1. Previsões não Supervisionadas*

Esta seção apresenta os resultados para a previsão das variáveis macroeconômicas brasileiras utilizando o esquema recursivo de previsão. Neste esquema, uma a uma as observações são incorporadas ao modelo após cada procedimento de previsão. Os horizontes de previsão considerados são $h = 1; 3; 6$ e $12$ meses. A Tabela 2 mostra a REQMP relativo ao AR(4) para cada variável e para cada modelo. Em negrito estão marcados os modelos que apresentaram melhor previsão frente ao AR(4).

É possível extrair algumas afirmações gerais da Tabela 1. Primeiro, o índice de produção industrial (IPI) foi melhor previsto pelo modelo que utiliza apenas os fatores estimados (FAT) para todos os horizontes considerados. Este resultado está de acordo com as evidências encontradas na literatura como Stock e Watson (2002, 2004, 2006 e 2012), Marcellino (2008) e Eickmeier e Ziegler (2008). Ademais, a introdução de métodos de aprendizado estatístico não implicou melhores previsões para esta variável, nem mesmo em horizontes de previsão de curto prazo.

A variável taxa de desemprego (*Desemp*) apresentou resultados diferentes. Os horizontes de curto prazo foram melhor previstos por modelos fatoriais sem a introdução de aprendizado estatístico. Porém, para horizontes mais longos, a incorporação de aprendizado estatístico gerou as melhores previsões. A literatura empírica aponta que modelos fatoriais são bons previsores para variáveis reais, em curto e longo prazos; ver Eickmeier e Ziegler (2008).

Com relação às variáveis de inflação, a introdução do aprendizado estatístico gerou as melhores previsões se comparados aos modelos fatoriais tradicionais e aumentado. Com exceção do horizonte h=12 para a variável IPC, todos os demais horizontes e em ambas variáveis, as melhores previsões foram obtidas por modelos que incorporam aprendizado estatístico aos

modelos fatoriais. Em especial, destacam-se os modelos que realizam seleção de variáveis. Estes modelos tiveram desempenho superior aos demais.

Como resultado geral, apesar de os ganhos não terem sido generalizados a todas as variáveis, a introdução de aprendizado estatístico aos modelos fatoriais contribuiu para gerar melhores previsões nas variáveis nominais e na taxa de desemprego, considerando diferentes horizontes de previsão.

**Tabela 1 - Raiz do erro quadrático médio relativo ao AR(4) (Esquema Recursivo)**

| Horizonte de Previsão h=1 | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | FAAR | FAT | Bagging | BMA 1 | BMA 2 | LARS | EN | NNG | CMA | CJA | LHO | SMA |
| Desemp | **1.013** | 1.063 | **1.013** | 1.050 | 1.038 | **1.013** | **1.013** | **1.013** | 1.038 | 1.050 | 1.050 | 1.063 |
| IPC | 1.002 | 1.032 | 0.996 | 1.005 | 1.008 | 1.017 | 1.017 | 1.047 | **0.973** | 0.977 | 0.977 | 1.013 |
| IPCA | 1.027 | 0.977 | 1.007 | 1.004 | 1.006 | 1.010 | 1.009 | 1.013 | 0.976 | **0.972** | **0.972** | 1.019 |
| IPI | 0.891 | **0.820** | 0.957 | 0.949 | 0.937 | 0.990 | 0.990 | 0.977 | 0.953 | 0.946 | 0.946 | 1.048 |
| Horizonte de Previsão h=3 | | | | | | | | | | | | |
| | FAAR | FAT | Bagging | BMA 1 | BMA 2 | LARS | EN | NNG | CMA | CJA | LHO | SMA |
| Desemp | **0.003** | **0.003** | 0.004 | **0.003** | **0.003** | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.398 |
| IPC | 1.002 | 1.032 | 0.996 | 1.005 | 1.008 | 1.017 | 1.017 | 1.047 | **0.973** | 0.977 | 0.977 | 1.013 |
| IPCA | 1.039 | 1.095 | 0.988 | 1.052 | 1.074 | 1.029 | 1.028 | 1.058 | 0.967 | **0.966** | 0.987 | 1.036 |
| IPI | 0.858 | **0.832** | 0.869 | 0.845 | 0.847 | 0.921 | 0.921 | 0.917 | 0.870 | 0.880 | 0.872 | 0.902 |
| Horizonte de Previsão h=6 | | | | | | | | | | | | |
| | FAAR | FAT | Bagging | BMA 1 | BMA 2 | LARS | EN | NNG | CMA | CJA | LHO | SMA |
| Desemp | 0.949 | 0.918 | 0.969 | **0.847** | 0.939 | 0.959 | 0.959 | 0.959 | 0.918 | 0.918 | 0.939 | 1.010 |
| IPC | 1.107 | 1.132 | 1.010 | 1.020 | 1.032 | 0.995 | 0.995 | **0.980** | 0.993 | 0.990 | 0.991 | 1.011 |
| IPCA | 0.955 | 0.917 | 0.980 | 0.983 | 0.979 | 0.988 | 0.988 | 0.976 | 0.901 | 0.896 | **0.880** | 0.960 |
| IPI | 0.905 | **0.858** | 1.003 | 0.990 | 0.978 | 1.001 | 1.001 | 0.991 | 0.971 | 0.974 | 0.948 | 1.004 |
| Horizonte de Previsão h=12 | | | | | | | | | | | | |
| | FAAR | FAT | Bagging | BMA 1 | BMA 2 | LARS | EN | NNG | CMA | CJA | LHO | SMA |
| Desemp | 1.090 | 1.112 | 0.978 | 1.000 | 1.011 | 1.000 | 1.000 | 1.011 | 0.888 | **0.876** | 0.888 | 1.000 |
| IPC | 0.880 | **0.871** | 1.054 | 0.940 | 0.930 | 0.949 | 0.949 | 0.897 | 0.959 | 0.967 | 0.985 | 0.931 |
| IPCA | 0.959 | 0.946 | 1.033 | 0.987 | 0.972 | 0.990 | 0.990 | **0.945** | 0.998 | 0.979 | 0.976 | 0.972 |
| IPI | 0.927 | **0.922** | 0.996 | 0.991 | 0.982 | 0.983 | 0.983 | 0.949 | 0.985 | 0.991 | 0.973 | 1.018 |

Nota: A Tabela 2 apresenta a raiz da razão do erro quadrático médio de previsão (REQMP) dos modelos contidos na Tabela 1 em relação ao modelo de *benchmark*. Em negrito estão destacados os modelos que geraram melhor previsão, menor REQMP, para cada variável alvo considerada.

### *4.2. Previsões Supervisionadas*

As Tabelas 2 e 3 apresentam os resultados de previsão para todos os modelos e variáveis, assim como a Tabela 1, porém, tais estimações foram realizadas por meio da supervisão dos modelos fatoriais. No primeiro caso, Tabela 2, foi aplicado um método de supervisão que estima os fatores de uma base de dados composta por previsões lineares de todas as variáveis, chamado de componentes principais usando combinação de previsões (CFPC). No segundo caso, Tabela 3, foi aplicado o algoritmo do LARS para selecionar algumas variáveis que tinham maior poder de previsão da variável-alvo. Posteriormente a esta seleção, foram extraídos os fatores e reaplicados todos os métodos descritos nas subseções 2.1 e 2.2.

Em negrito estão destacados os modelos que geraram previsões melhores que os modelos não supervisionados, reportados na Tabela 1. Por exemplo, para a variável IPC no horizonte h=1 para o modelo FAAR gerou melhor previsão no caso supervisionado por CFPC (Tabela 2) do que o caso não supervisionado (Tabela 1). Entretanto, a mesma variável, no mesmo horizonte e para o mesmo modelo teve performance pior no caso supervisionado por LARS (Tabela 3) do que no caso supervisionado (Tabela 1). [12]

[12] Observe que tal comparação é possível devido a adoção de um mesmo *benchmark* que é independente dos procedimentos de estimação fatorial e aprendizado estatístico adotados em ambos os casos, com e sem supervisão.

**Tabela 2 - Raiz do erro quadrático médio relativo ao AR(4) - Supervisionado por CFPC**

| Horizonte de Previsão h=1 | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | FAAR | FAT | Bagging | BMA 1 | BMA 2 | LARS | EN | NNG | CMA | CJA | LHO | SMA |
| Desemp | **0.941** | **0.976** | **0.925** | **0.971** | **0.971** | 1.003 | 1.003 | **1.002** | **0.775** | **0.774** | **1.023** | **1.032** |
| IPC | **0.967** | **0.901** | **0.772** | 1.011 | **0.986** | **1.001** | **1.001** | **0.994** | **0.851** | **0.858** | **0.599** | **1.001** |
| IPCA | 1.028 | **0.921** | **0.755** | **0.997** | **0.960** | **0.999** | **0.999** | **0.995** | **0.904** | **0.905** | **0.481** | 1.026 |
| IPI | 0.978 | 0.980 | 0.976 | 1.001 | 1.001 | 1.003 | 1.003 | 1.004 | **0.855** | **0.860** | 0.963 | 1.078 |
| Horizonte de Previsão h=3 | | | | | | | | | | | | |
| | FAAR | FAT | Bagging | BMA 1 | BMA 2 | LARS | EN | NNG | CMA | CJA | LHO | SMA |
| Desemp | 0.760 | 0.636 | 0.851 | 0.720 | 0.730 | 0.899 | 0.900 | 0.992 | 0.631 | 0.634 | 0.862 | 0.904 |
| IPC | **0.941** | **0.940** | **0.556** | **0.481** | **0.478** | **1.004** | **1.004** | **1.005** | **0.885** | **0.891** | **0.606** | 1.020 |
| IPCA | **0.895** | **0.903** | **0.487** | **0.456** | **0.449** | **1.014** | **1.014** | **1.009** | **0.780** | **0.794** | **0.472** | 1.098 |
| IPI | 0.956 | 0.940 | 1.002 | 0.994 | 0.988 | 1.001 | 1.001 | 0.997 | 0.978 | 0.988 | 0.981 | 1.004 |
| Horizonte de Previsão h=6 | | | | | | | | | | | | |
| | FAAR | FAT | Bagging | BMA 1 | BMA 2 | LARS | EN | NNG | CMA | CJA | LHO | SMA |
| Desemp | **0.749** | **0.754** | **0.892** | **0.764** | **0.767** | 1.030 | 1.030 | 1.013 | **0.724** | **0.725** | 1.015 | 0.950 |
| IPC | **0.891** | **0.879** | **0.750** | **0.777** | **0.769** | **1.007** | **1.007** | **1.009** | **0.813** | **0.827** | **0.619** | **0.975** |
| IPCA | **0.795** | **0.770** | **0.895** | **0.541** | **0.566** | **0.987** | **0.987** | **0.995** | **0.783** | **0.786** | **0.464** | **0.945** |
| IPI | 0.891 | **0.852** | **0.983** | **0.930** | **0.923** | **0.969** | **0.969** | **0.971** | **0.954** | **0.956** | 1.016 | **0.995** |
| Horizonte de Previsão h=12 | | | | | | | | | | | | |
| | FAAR | FAT | Bagging | BMA 1 | BMA 2 | LARS | EN | NNG | CMA | CJA | LHO | SMA |
| Desemp | **1.035** | **0.992** | 0.995 | 1.004 | **1.007** | 1.007 | 1.007 | 1.029 | **0.862** | **0.866** | 1.018 | 1.028 |
| IPC | 0.948 | 0.938 | **0.867** | 0.999 | 0.998 | 0.999 | 0.996 | 0.981 | 0.991 | 0.987 | **0.599** | 0.953 |
| IPCA | **0.906** | **0.890** | **0.965** | 0.991 | 0.982 | **0.988** | **0.988** | 0.954 | 0.971 | **0.963** | **0.439** | **0.952** |
| IPI | **0.919** | **0.910** | **0.991** | **0.951** | **0.943** | 0.996 | 0.996 | 0.991 | 0.985 | **0.990** | 1.029 | 1.031 |

Nota: A Tabela 2 apresenta a raiz da razão do erro quadrático médio de previsão (REQMP) dos modelos contidos na Tabela 1 em relação ao modelo de *benchmark*, considerando que os modelos fatoriais tenham sido supervisionados pelo método do CFPC. Em negrito estão destacados os modelos que geraram melhor previsão, menor REQMP, quando comparado ao caso supervisionado, apresentado na Tabela 1.

A introdução da supervisão dos modelos fatoriais aumentou a performance dos modelos de previsão em 60,9% para a supervisão com CFPC e 56,7% para a supervisão com o LARS. O horizonte h=6 foi o que apresentou maior aumento absoluto de desempenho (85,4% em ambos os tipos de supervisão) quando comparado ao caso não supervisionado. Já o horizonte h=12 foi o que apresentou menor aumento de performance (43,7% - CFPC, 45,8% - LARS).

**Tabela 3 - Raiz do erro quadrático médio relativo ao AR(4) - Supervisionado por LARS**

| Horizonte de Previsão h=1 | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | FAAR | FAT | Bagging | BMA 1 | BMA 2 | LARS | EN | NNG | CMA | CJA | LHO | SMA |
| Desemp | **0.907** | **0.928** | **0.977** | **0.965** | **0.954** | **0.993** | **0.993** | **0.988** | **0.787** | **0.787** | **1.023** | **1.015** |
| IPC | 1.071 | **1.012** | 1.033 | 1.037 | 1.046 | 1.025 | 1.025 | 1.033 | **0.910** | **0.908** | **0.599** | 1.038 |
| IPCA | **1.000** | **0.867** | 1.017 | 1.014 | 1.012 | **1.008** | 1.008 | 1.013 | **0.918** | **0.905** | **0.481** | 1.019 |
| IPI | 0.973 | 0.968 | 0.973 | 1.004 | 1.003 | 1.007 | 1.007 | 1.008 | **0.841** | **0.853** | 0.963 | 1.068 |
| Horizonte de Previsão h=3 | | | | | | | | | | | | |
| | FAAR | FAT | Bagging | BMA 1 | BMA 2 | LARS | EN | NNG | CMA | CJA | LHO | SMA |
| Desemp | 0.757 | 0.680 | **0.909** | **0.794** | **0.800** | 0.986 | 0.986 | 0.994 | **0.628** | **0.627** | **0.862** | 0.921 |
| IPC | **0.988** | **1.006** | **0.865** | 1.071 | 1.059 | 1.012 | 1.012 | 1.005 | **0.878** | **0.883** | **0.472** | 1.068 |
| IPCA | **0.895** | **0.903** | **0.487** | **0.456** | **0.449** | 1.014 | 1.014 | 1.009 | **0.780** | **0.794** | **0.472** | 1.098 |
| IPI | 0.938 | 0.906 | 1.014 | 1.003 | 0.998 | 1.011 | 1.011 | 1.011 | 0.948 | 0.966 | 0.981 | 1.001 |
| Horizonte de Previsão h=6 | | | | | | | | | | | | |
| | FAAR | FAT | Bagging | BMA 1 | BMA 2 | LARS | EN | NNG | CMA | CJA | LHO | SMA |
| Desemp | **0.655** | **0.642** | **0.828** | **0.672** | **0.685** | 1.010 | 1.010 | 1.009 | **0.636** | **0.639** | 1.015 | 0.890 |
| IPC | **0.910** | **0.904** | 1.019 | **0.947** | **0.950** | **0.984** | **0.984** | 0.996 | **0.877** | **0.888** | **0.619** | 0.964 |
| IPCA | **0.861** | **0.836** | **0.920** | **0.896** | **0.897** | **0.956** | **0.956** | 0.996 | **0.758** | **0.775** | **0.464** | **0.943** |
| IPI | 0.966 | 0.978 | **1.001** | 0.997 | 0.991 | **1.000** | **1.000** | 0.996 | **0.945** | **0.950** | 1.016 | 1.029 |
| Horizonte de Previsão h=12 | | | | | | | | | | | | |
| | FAAR | FAT | Bagging | BMA 1 | BMA 2 | LARS | EN | NNG | CMA | CJA | LHO | SMA |
| Desemp | **0.980** | **0.929** | 1.020 | **0.997** | **0.994** | **0.998** | **0.998** | **0.996** | **0.858** | **0.861** | 1.018 | 1.008 |
| IPC | 0.993 | 0.974 | 0.994 | 1.006 | 1.010 | 1.004 | 1.004 | 1.014 | 0.991 | 0.989 | **0.599** | 0.974 |
| IPCA | **0.918** | **0.897** | **0.994** | 0.992 | 0.987 | 0.986 | **0.986** | **0.952** | **0.956** | **0.951** | **0.439** | **0.947** |
| IPI | 0.985 | 1.033 | 1.006 | **0.989** | **0.986** | **0.990** | **0.990** | **0.987** | 0.970 | 0.979 | **1.029** | **1.045** |

Nota: A Tabela 3 apresenta a raiz da razão do erro quadrático médio de previsão (REQMP) dos modelos contidos na Tabela 1 em relação ao modelo de *benchmark*, considerando que os modelos fatoriais tenham sido supervisionados pelo método do LARS. Em negrito estão destacados os modelos que geraram melhor previsão, menor REQMP, quando comparado ao caso supervisionado, apresentado na Tabela 1.

A Figura 1 apresenta a proporção do número de modelos supervisionados que tiveram performance superior aos modelos não supervisionados em comparação com o número total de modelos. Por exemplo, para o horizonte h=1, para a variável desemprego, a supervisão por CFCP teve desempenho superior em 80% dos modelos comparado ao número total de previsões geradas. Isto é, apenas em 20% das previsões com diferentes

modelos para este horizonte e esta variável, a não supervisão gerou melhores previsões.

Destaca-se que as variáveis que são mais sensíveis à melhoria de desempenho quando supervisionadas são as variáveis nominais: IPC e IPCA. A supervisão por CFPC tende a gerar melhores resultados do que a supervisão por LARS. Em relação ao IPC para o horizonte h=12 , a supervisão não foi tão relevante, tendo um aumento de desempenho apenas em 18,3% dos casos.

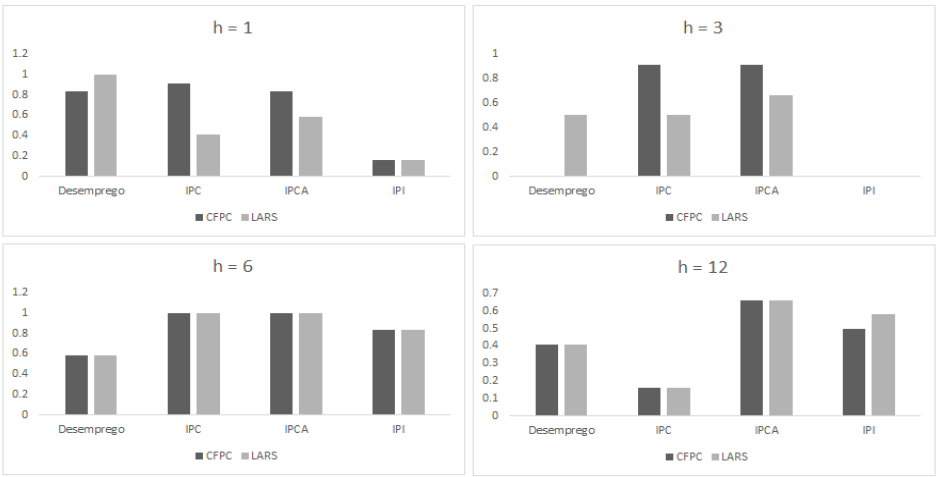


**Figura 1 - Modelos supervisionados vs. modelos não supervisionados**

Nota: A Figura 1 apresenta a proporção de modelos obtidos por supervisão por LARS e CFPC que tiveram desempenho superior, menor REQMP, do que as previsões não supervisionadas apresentadas na Tabela 1, para cada variável alvo e para cada horizonte de previsão considerado.

Variáveis reais são menos sensíveis à supervisão. Entretanto, para a variável IPI a supervisão apresentou resultados superiores à medida que o horizonte de tempo se tornava mais longo. Ou seja, previsões de maior longo prazo apresentaram melhor desempenho para os modelos supervisionados do que para os modelos sem supervisão.

Por sua vez, a variável desemprego teve comportamento semelhante as variáveis IPC e IPCA, entretanto, a proporção de ganho com tal variável foi menor do que com as variáveis nominais.

Estes resultados indicam que existem elevados ganhos de previsão para variáveis como IPCA, por exemplo, ao considerar não apenas modelos fatoriais com a incorporação de técnicas de aprendizado estatístico, mas também a supervisão dos modelos fatoriais em si. Assim, uma possível área de pesquisa sobre este tema é investigar a natureza e o potencial ganho de métodos alternativos de supervisão de modelos fatoriais.

## 5. Conclusões

Este artigo comparou treze modelos para a previsão de quatro variáveis macroeconômicas brasileiras: taxa de desemprego, índice de produção industrial, índices de preços mensurados pelo IPCA e IPC.

Os métodos utilizados foram Modelo de Fator Autorregressivo Aumentado (FAAR), Modelo de Fator Tradicional (FAT), *Bagging*, dois tipos de Modelos Bayesianos Ponderado (BMA), *Least Angle Regression* (LARS), *Elastic Net* (EN) , *Non-Negative Garotte* (NNG), Modelo Mallows Ponderado (CMA), Modelo Jacknife Ponderado (CJA), Modelo *Leave-h-out* (LHO) e Modelo de ponderado por média simples (SMA). Apenas o modelo AR(4) não inclui fatores estimados por componentes principais e serviu como *benchmark*. Tais métodos combinavam fatores estimados seja por meio da seleção de variáveis, por meio da ponderação das previsões ou ao se adotar técnicas de *shrinkage*.

Duas formas de extração dos fatores foram consideradas: estimação não supervisionada e a estimação supervisionada. A última realizada por meio de dois algoritmos diferentes: CFPC e LARS. No caso dos modelos não supervisionados, os fatores são extraídos sem considerar a variável-alvo que será prevista. Já no caso dos modelos supervisionados, o comportamento da variável-alvo pode afetar a estimação dos fatores.

Os resultados indicam que métodos de aprendizado estatístico melhoram o desempenho preditivo das variáveis econômicas brasileiras. A única exceção foi o índice de produção industrial (IPI) que foi melhor previsto por métodos fatoriais sem adoção de aprendizado estatístico.

Além disso, a combinação de técnicas de aprendizagem estatística e supervisão fatorial produzem previsões com melhor desempenho do que modelos sem fatores, do que modelos fatoriais com ou sem supervisão e do que modelos que utilizam apenas o aprendizado estatístico sem supervisão fatorial. Este resultado foi válido para ambos os índices de preços (IPCA e IPC) e para a taxa de desemprego (*Desemp*). O IPI somente teve seu desempenho melhorado ao se considerar a supervisão em maiores horizontes de tempo.

Dentre todos os modelos com supervisão e aprendizado estatístico, os métodos de seleção de variáveis (CMA, LHO e CJA) foram os que tiveram melhor desempenho para índices de preços e taxa de desemprego. Este resultado mostra que tal método é uma opção robusta para prever variáveis macroeconômicas brasileiras.

## Referências

Artis, Michael, Marcellino e Proietti. "Business cycles in the new EU member countries and their conformity with the euro area". *Journal of business cycle measurement and analysis* 1: 7-41, 2005

Bai, J. "Inferential theory for factor models of large dimensions". *Econometrica* 71:135–171, 2003.

Bai, J. and Ng, S. "Determining the number of factors in approximate factor models". *Econometrica* 70: 191–221, 2002.

Bai, J. and Ng, S. "Confidence intervals for diffusion index forecasts and inference for factoraugmented regressions". *Econometrica* 74 1133–1150, 2006.

Bai, J. and Ng, S. "Forecasting economic time series using targeted predictors" *Journal of Econometrics* 146: 304–317, 2009.

Bai, J. and Ng, S. "Boosting diffusion indices". *Journal of Applied Econometrics* 4: 607–629, 2008.

Bair, Eric, Hastie, T., Paul, D. e Tibsharani, R. "Prediction by supervised principal components". *Journal of the American Statistical Association* 101, no. 473, 2006.

Boivin, J., and Ng, S. "Are more data always better for factor analysis?" *Journal of Econometrics* 132: 169–194, 2006.

Breiman, L. "Better subset regression using the nonnegative garrote". *Technometrics* 37, no. 4: p. 373- 384, 1995.

Cheng, X. and Hansen, B. "Forecasting with factor-augmented regression: A frequentist model averaging approach". *Journal of Econometrics* 186: 280-293, 2015.

Dias, F., Pinheiro, M. e Rua, A. "Forecasting using targeted diffusion indexes", *Journal of Forecasting* 29, no.3: 341-352, 2010.

Efron, B., Hastie, T., Johnstone, L., and Tibshirani, R. "Least angle regression". *Annals of Statistics* 32: 407-499, 2004.

Eickmeier, S. and Ziegler, C. "How successful are dynamic factor models at forecasting output and inflation? a meta-analytic approach". *Journal of Forecasting* 27, no.3: 237-265, 2008.

Elliot, G. e Timmermann, A. "Economic forecasting". *Princeton University Press*, New Jersey, 2016.

Fernandez, C., lLey, E. e Steel, M. "Benchmark priors for Bayesian model averaging". *Journal of Econometrics* 100: 381-427, 2001.

Ferreira, R., Bierens, H. e Castelar, I. "Forecasting quarterly Brazilian GDP growth rate with linear and nonlinear diffusion index models". *EconomiA* 6, no.3: 261-292, 2005.

Figueiredo, F. M. R. "Forecasting Brazilian inflation using a large data set" *Brazilian Central Bank*, working paper series 228, 2010.

Foster, D. e George, E. "The risk inflation criterion for multiple regression" *The Annals of Statistics* 22: 1947-1975, 1994.

Garcia, M. Medeiros, M. e Vasconcelos, G. 2016 "Real-time inflation forecasting with high dimensional models: The case of Brazil". *XVI Encontro de Finanças*, Rio de Janeiro.

Gelper, S. and Croux, C.2008 "Least angle regression for time series forecasting with many predictors", *Working paper. technical report, Katholieke Universiteit Leuven*.

Hansen, B. "Least squares model averaging". *Econometrica* 75: 1175–1189, 2007.

Hansen, B. "Least squares forecasting averaging". *Journal of Econometrics* 146: 342–350, 2008.

Hansen, B. e Racine, J.S. "Jackknife model averaging". *Journal of Econometrics* 167: 38–46, 2012.

Hansen, P. Lunde, A. e Nason, J. M. "The model confidence set". *Econometrica* 79: 453-497, 2011.

Hastie, T.; Tibshirani, R. e Friedman, J. "The elements of statistical learning: data mining, inference, and prediction". *Springer-Verlag New York*. 2ª ed, 2009.

Hillebrand, e.; Huang, Y.; Lee, t.; Li, C. "Using the entire yield curve in forecasting output and inflation", *Econometrics* 6, no. 40, 2018.

Inoue, A., e Kilian, L. "How useful is bagging in forecasting economics time series? a case study of us cpi inflation". *J. Amer. Statist. Assoc*. 103: 511–522, 2008.

Kim, H., e Swanson, N. "Forecasting financial and macroeconomic variables using data reduction methods: new empirical evidence". *Journal of Econometrics* 178: 352–367, 2014.

Kim, H., e Swanson, N. "Mining big data using parsimonious factor machine learning, variable selection, and shrinkage methods", *Rutgers University*, working paper, 2016.

Koop, G. and Potter, S. "Forecasting in dynamic factor models using Bayesian model averaging". *Econometrics Journal* 7: 550-565, 2004.

Liu, C. E Kuo, B. "Model Averaging In Predictive Regressions". *Econometrics Journal* 19, no.2: 203-231, 2016.

Kwiatkowski, D; Phillips, P.; Schmidt, P. e Shin, Y. "Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root". *Journal of Econometrics* 54, no.9: 159-178, 1992.

Mallows, C.L. "Some Comments on Cp" *Technometrics* 15: 661–675, 1973.

Marcellino, M. A. "Comparison of Time Series Model Forecasting GDP Growth and Inflation". *Journal of Forecasting* 27: 305-340, 2008.

Medeiros, M. C.; Vasconcelos, G. and Freitas, E. "Forecasting Brazilian Inflation with High-dimensional Models". *Brazilian Econometric Review* 36, no 2. 2016.

Pesaran, H. and Timmermann, A. "Selection of Estimation Window in The Presence of Breaks", *Journal of Econometrics* 137, no.1: 134-161, 2007.

Pesaran, H.; Petenuzzo, D. e Timmermann, A. "Forecasting Time Series Subject To Multiple Structural Breaks" *Review of Economic Studies* 73: 1057-1084, 2006.

Rahal, C. "Housing Market Forecasting With Factor Combinations", Discussion Papers 15-05r, 2015, *Department of Economics, University Of Birmingham.*

Rossi, B. Advances in Forecasting Under Instability. In Elliott, G. and Timmermann, A., Editors, *Handbook of Economic Forecasting*, Volume 2b, Chapter 21: 1203-1324, 2012.

Rossi, B. e Inoue, A. "Out-of-sample Forecast Tests Robust to The Window Size Choice". *Journal of Business and Economics Statistics* 30, no.3: 432-453, 2012.

Saigo, H.; Uno, T. and Tsuda, K. "Mining Complex Genotypic Features for Predicting Hiv-1 Drug Resistance", *Bioinformatics* 23: 2455–2462, 2007.

Stock, J. H. and Watson, M. W. "A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series". In: Engle, R., White, H. (Eds.), *Cointegration, Causality and Forecasting: A Festschrift for Clive W.J. Granger.* Oxford University Press, 1999.

Stock, J. H. and Watson, M. W. "Forecasting Using Principal Components from a Large Number of Predictors". *Journal of the American Statistical Association* 97: 1167-1179, 2002.

Stock, J. H. and Watson, M. W. "Combination Forecasts of Output Growth in a Seven Country Data Set". *Journal of Forecasting* 23: 405-430, 2004.

Stock, J. H. and Watson, M. W. "Implications of Dynamic Factor Models for VAR Analysis". *Nber Working Papers 11467,* 2005.

Stock, J. H. and Watson, M. W. "Forecasting With Many Predictors" In Elliott, G., Granger, C., and Timmermann, A., Editors, *Handbook of Economic Forecasting* 1, Chapter 10: 515-554, 2006.

Stock, J. H. and Watson, M. W. "Generalized Shrinkage Methods for Forecasting Using Many Predictors" *Journal of Business and Economic Statistics* 30, no.4: 481-493, 2012.

Tibshirani, R. "Regression Shrinkage and Selection via the Lasso". *Journal of the Royal Statistical Society, Series B*, 58 p. 267-288, 1996.

Tu, Y., and Lee, T.H. "Forecasting Using Supervised Factor Models" *Journal of Management Science and Engineering* 4: 12-27, 2019.

Watson, M. And Amengual, D. "Consistent Estimation Of The Number Of Dynamic Factors In A Large N And T Panel" *Journal of Business and Economic Statistics* 25, no. 1: 91-96, 2007.

Yin, Shou-yung, Liu, Chu-an e Lin, C. "Focused Information Criterion And Model Averaging For Large Panels With A Multifactor Error Structure", Ideas Working Paper: 16-a016, *Institute Of Economics, Academia Sinica,* 2016.

Yuan, M. and Lin, Y. 2007 "On the Non-negative Garrotte Estimator." *Journal of the Royal Statistical Society* 69, no.2: 143-161.

Zou, H. "The Adaptive Lasso and Its Oracle Properties" *Journal of the American Statistical Association* 101: 1418-1429, 2006.

Zou, H. e Hastie, T. "Regularization and Variable Selection Via The Elastic Net", *Journal of the Royal Statistical Society, Series B* Vol. 67, Part 2: 301–320, 2005.

Zhang, Ke; Yin, Fan E Xiong, S. "Comparisons of Penalized Least Squares Methods by Simulations". *Working Paper, Chinese Academy Of Sciences,* 2014.

# Apêndice

## A.1 Descrição das variáveis utilizadas

As Tabelas 5, 6 e 7 apresentam as 117 variáveis macroeconômicas usadas para estimar os fatores utilizados para realização das previsões. Em cada tabela está indicada a referência da variável na base de dados, o nome da série, as fontes de onde as variáveis foram extraídas [13] e a transformação que cada variável foi submetida para obter a estacionariedade. Tais transformações são apresentadas em forma de código, de modo que: 1 - variável em nível, 2 - Primeira diferença, 3 - Logaritmo da primeira diferença, 4 - Segunda diferença, 5 - Logaritmo da segunda diferença. A base de dados completa pode ser requerida por email aos autores.

| Ref. | Nomes das séries | Simbolo | Fonte | Transformação |
|---|---|---|---|---|
| 1 | Salário real - médio- SP | Sal.R.IND | IPEA | 2 |
| 2 | Salário real - indústria - SP | Sal.R.Med | IPEA | 2 |
| 3 | Rendimento médio real dos assalariados RMSP | Rend.Med.Ass | IPEA | 2 |
| 4 | Salário mínimo (PPC) | Sal.Min | IPEA | 2 |
| 5 | Folha de pagamento - indústria geral | F.pag.Ind | IBGE/PIMES | 2 |
| 6 | IGP-10 índice | IGP_10 | IPEA | 2 |
| 7 | IGP-OG - geral - índice | IGP_OG | IPEA | 2 |
| 8 | INCC-Geral | INCC_geral | IPEA | 2 |
| 9 | INPC alimentos e bebidas | INPC_ali. | IPEA | 1 |
| 10 | INPC artigos de residência | INPC_res | IPEA | 1 |
| 11 | INPC-Geral | INPC_geral | IPEA | 3 |
| 12 | INPC Comunicação | INPC_com | IPEA | 1 |
| 13 | INPC - educação. leitura e papelaria | INPC_educ | IPEA | 4 |
| 14 | INPC - despesas pessoais | INPC_desp | IPEA | 1 |
| 15 | INPC - habitação | INPC_hab | IPEA | 2 |
| 16 | INPC - saúde e cuidados pessoais | INPC_saude | IPEA | 2 |
| 17 | INPC - transportes | INPC_transp | IPEA | 1 |
| 18 | INPC - vestuário | INPC_vest | IPEA | 1 |
| 19 | IPA origem - produtos agropecuários | IPA_orig_agro | IPEA | 2 |
| 20 | IPA origem - produtos industriais | IPA_orig_ind | IPEA | 2 |
| 21 | IPA-10 - índice (ago. 1994 = 100) | IPA_10 | IPEA | 2 |
| 22 | IPCA - alimentos e bebidas | IPCA_ali | IPEA | 1 |

[13] A maioria das variáveis foram extraídas do site do Ipeadata (http://www.ipeadata.gov.br/) e está indicado nas tabelas por IPEA. Outras fontes foram o Banco Central (Bacen) (https://www3.bcb.gov.br/sgspub/localizarseries/localizarSeries.do?method=prepararTelaLocalizarSeries) e o IBGE (https://www.ibge.gov.br/).

| Ref. | Nomes das séries | Simbolo | Fonte | Transformação |
|---|---|---|---|---|
| 23 | IPCA - comunicação - var. | IPCA_com | IPEA | 1 |
| 24 | IPCA - artigos de residência - var. | IPCA_res | IPEA | 1 |
| 25 | IPCA - despesas pessoais - var. | IPCA_desp | IPEA | 2 |
| 26 | IPCA - educação, leitura e papelaria - var. | IPCA_educ | IPEA | 2 |
| 27 | IPCA - transportes - var. | IPCA_transp | IPEA | 1 |
| 28 | IPCA - vestuário - var. | IPCA_vest | IPEA | 4 |
| 29 | IPCA - habitação - | IPCA_hab | IPEA | 2 |
| 30 | IPCA - preços livres - bens duráveis - var | IPCA_nao_duraveis | IPEA | 1 |
| 31 | IPCA - preços livres - bens semi duráveis | IPCA_semi_duraveis | IPEA | 2 |
| 32 | IPCA - preços livres - bens duráveis | IPCA_duraveis | IPEA | 1 |
| 33 | IPCA - preços livres - serviços - var | IPCA_serviços | IPEA | 2 |
| 34 | IPCA - preços livres - comercializáveis | IPCA_comerc | IPEA | 1 |
| 35 | IPCA - preços livres -não comercializáveis | IPCA_n_comerc | IPEA | 1 |
| 36 | IPCA - preços livres | IPCA_livre | IPEA | 2 |
| 37 | IPCA- preços monitorados | IPCA_monit | IPEA | 2 |
| 38 | IPCA_Geral | IPCA_geral | Bacen | 1 |
| 39 | IPCA - saúde e cuidados pessoais - var. | IPCA_saude | IPEA | 2 |
| 40 | IPPA | IPPA | Bacen | 1 |
| 41 | IPC - geral | IPC_geral | IPEA | 3 |
| 42 | IGP-M - geral | IGP_M | IPEA | 2 |
| 43 | Reservas bancárias (final de período) | Res_banc | Bacen | 2 |
| 44 | Base monetária restrita (final de período) | BM | Bacen | 2 |
| 45 | M0 - base monetária - média | M0 | IPEA | 2 |
| 46 | M1 - depósitos à vista - média | M1 | IPEA | 2 |
| 47 | Ibov - Fechamento Ajustado | IBOV | IPEA | 2 |
| 49 | SP 500 | SP500 | IPEA | 2 |
| 50 | Ouro - rendimento nominal | Ouro | IPEA | 2 |
| 51 | Taxa de juros - TJLP | TJLP | IPEA | 1 |
| 52 | Taxa de juros - Over / Selic | Tx_Over_selic | IPEA | 2 |
| 53 | Volatidade tx_juros_over_selic | vol_Tx_Over_selic | IPEA | 2 |
| 54 | Dívida estados município por pib | div_est_mun | IPEA | 2 |
| 55 | Dívida total setor público | div_total | IPEA | 4 |
| 56 | ICMS | ICMS | IPEA | 2 |
| 57 | Imposto de Importação | II | IPEA | 2 |
| 58 | Imposto Territorial Rural | ITR | IPEA | 2 |
| 59 | Imposto de Renda | IR | IPEA | 3 |
| 60 | IOF | IOF | IPEA | 2 |

| Ref. | Nomes das séries | Simbolo | Fonte | Transformação |
|---|---|---|---|---|
| 61 | IPI | IPI | IPEA | 2 |
| 62 | NFSP Estados e municípios | NFSP_est_mun | IPEA | 2 |
| 63 | Horas Trabalhadas na Indústria | H_ind | IPEA | 1 |
| 64 | Nível de emprego na Indústria-SP | N_emp_ind_sp | IPEA | 2 |
| 65 | Taxa de desemprego aberto na RMSP | Tx_dese_aberto_RMSP | IPEA | 2 |
| 66 | Taxa de desemprego RMSP | Tx_desemp_RMSP | IPEA | 2 |
| 67 | NASDAQ | NASDAQ | IPEA | 2 |
| 68 | EUA_IPA | EUA_IPA | IPEA | 2 |
| 69 | EUA_IPC | EUA_IPC | IPEA | 2 |
| 70 | EUA_T-note-2 | EUA_T_NOTE_2 | IPEA | 2 |
| 71 | EUA_T-Note 10 | EUA_T_Note_10 | IPEA | 2 |
| 72 | Exportação FOB | Exp_FOB | IPEA | 2 |
| 73 | Exp_bens de capital | Exp_Cap | IPEA | 2 |
| 74 | Exp_Bens_duraveis | Exp_Dur | IPEA | 2 |
| 75 | Exp_Não_Duraveis | Exp_n_dur | IPEA | 2 |
| 76 | Exp_bens Interm. | Exp_intermed | IPEA | 2 |
| 77 | Importação_FOB | Imp_FOB | IPEA | 2 |
| 78 | Import_bens capital | Imp_Cap | IPEA | 2 |
| 79 | Import_cons_duravel | Imp_Dur | IPEA | 2 |
| 80 | Import_cons_nao duravel | Imp_n_dur | IPEA | 4 |
| 81 | Import_bens interm | Imp_intermed | IPEA | 2 |
| 82 | Tx_cambio efetiva (INPC) | tx_cambio_efetiva | IPEA | 2 |
| 83 | Tx_cambio comercial | tcamb_dolar | IPEA | 2 |
| 84 | Cons_Energ_Ind | Cons_Energ_Ind | IPEA | 4 |
| 85 | Cons_Ener_comer | Cons_Ener_comer | IPEA | 2 |
| 86 | Consu_Petroleo | Consu_Petroleo | IPEA | 2 |
| 87 | ICEA | ICEA | IPEA | 4 |
| 88 | Vendas Industriais Sp | V_ind_SP | IPEA | 2 |
| 89 | Dív_externa_est_mun | Div_ext_est_mun | IPEA | 2 |
| 90 | Divida externa governo federal | Div_ext_gov_fed | IPEA | 4 |
| 91 | Produção industrial - bens de capital | Prod_ind_b_cap | IPEA | 2 |
| 92 | Produção industrial - bens de consumo | Prod_ind_cons | IPEA | 2 |
| 93 | Produção industrial - bens de consumo duráveis | Prod_ind_cons_dur | IPEA | 2 |
| 94 | Produção industrial - bens de intermed. | Prod_ind_intermed | IPEA | 2 |
| 96 | Emprego formal - índice geral | Emp_ext_min | IPEA | 2 |

| Ref. | Nomes das séries | Simbolo | Fonte | Transformação |
|---|---|---|---|---|
| 97 | Emprego formal - Extrativa mineral | Emp_ext_min | IPEA | 4 |
| 98 | Emprego formal - Indústria de transformação (total) | Emp_ind_to | IPEA | 4 |
| 99 | Emprego formal - Minerais não-metálicos - Índice | Emp_min_metál | IPEA | 4 |
| 100 | Emprego formal - Metalurgia - Índice | Emp_met_Metalu | IPEA | 4 |
| 101 | Emprego formal - Mecânica - Índice | Emp_mec_Mecâ | IPEA | 4 |
| 102 | Emprego formal - Material de transporte - Índice | Emp_mat_transp | IPEA | 5 |
| 103 | Emprego formal - Mobiliário - Índice | Emp_mob_Mobili | IPEA | 4 |
| 104 | Emprego formal - Construção civil - Índice | Emp_Cont_civil | IPEA | 2 |
| 105 | Emprego formal - Comércio - Índice | Emp_Comr_Comércio | IPEA | 4 |
| 106 | Emprego formal - Serviços - Índice | Emp_Seri_Serviços | IPEA | 4 |
| 107 | Emprego formal - Agropecuária | Emp_Agrp_pesca | IPEA | 4 |
| 108 | IPI-EUA | IPI_EUA | IPEA | 2 |
| 109 | Reservas internacionais - Conceito liquidez | Res_inter | IPEA | 2 |
| 110 | Transações correntes | Tra_corre | IPEA | 3 |
| 111 | Conta capital - mensal | Con_capit | IPEA | 2 |
| 112 | Conta financeira - mensal | Con_finan | IPEA | 1 |
| 113 | Ind_prod_ind | Ind_prod_ind | IPEA | 2 |
| 114 | Utilização da capacidade instalada - indústria - (%) | Cap_pro_ind | IPEA | 2 |
| 115 | Utilização da capacidade instalada - indústria - (%) sp | Cap_pro_ind_sp | IPEA | 2 |
| 116 | Utilização da capacidade - extrativa mineral - média - (%) | U_Cap_prod_ext | IPEA | 2 |
| 117 | cap_prod_extrat | Cap_prod_ext | IPEA | 1 |