CARTA DE
CONJUNTURA
NÚMERO 52 — NOTA DE CONJUNTURA 26 — 3 ° TRIMESTRE DE 2021

ATIVIDADE ECONÔMICA

Previsão de indicadores de atividade e preço com modelo algorítmico[1]

Ajax Moreira

Coordenador da Diretoria de Estudos e Políticas Macroeconômicas (Dimac) do Ipea

ajax.moreira@ipea.gov.br

Leonardo Carvalho

Técnico de planejamento e pesquisa na Dimac do Ipea

leonardo.carvalho@ipea.gov.br

Izabel Nolau

Bolsista na Dimac do Ipea

izabel.souza@ipea.gov.br

Divulgado em 17 de setembro de 2021.

1 Introdução

O modelo algorítmico prevê o valor de uma variável dado um conjunto de explicativas, identificando padrões na relação entre o alvo e as explicativas sem supor formas funcionais para esse padrão, além de sempre manter o valor previsto fora da amostra de estimação. Nesta abordagem, são estimados múltiplos algoritmos que determinam padrões segundo diferentes critérios, e o algoritmo que maximiza a exatidão da previsão – minimizando um critério de perda – é o escolhido. A qualidade do resultado depende da variedade dos algoritmos testados e do conjunto de explicativas.

Os modelos econométricos utilizam a teoria econômica para escolher o conjunto de explicativas. A abordagem algorítmica não dispõe de um critério para a especificação do conjunto de variáveis explicativas, mas tem recursos para selecionar as variáveis utilizadas na previsão. Por isso são, em geral, incluídas como explicativas a mais ampla cobertura de variáveis relacionadas ao objeto da previsão, o que pode implicar em conjuntos com mais do que quatrocentas variáveis. No entanto, conjuntos de explicativas com dimensão elevada tendem a dificultar o funcionamento dos algoritmos (Bolhuis e Rayner, 2020). Para lidar com essa questão, propomos a introdução de um procedimento de pré-seleção de explicativas como um elemento adicional na escolha do previsor.

Os modelos econométricos são especificados admitindo um processo gerador dos dados, e suas previsões têm distribuições de probabilidade conhecidas. O modelo algorítmico obtém previsão pontual, o que é uma limitação importante para a sua utilização na análise econômica. Admitindo que o desvio entre o valor previsto e o observado é uma realização de uma distribuição empírica, propomos uma metodologia para obter, de forma não paramétrica, realizações da previsão, e, assim, calcular o intervalo de confiança da previsão e das estatísticas derivadas.

A abordagem dos modelos algorítmicos com as duas extensões, pré-seleção e incerteza da previsão foi utilizada para prever: i) o Índice de Preços ao Consumidor Amplo (IPCA) e cinco dos seus componentes; ii) o produto interno bruto (PIB) medido com o monitor do PIB construído na Fundação Getulio Vargas (FGV) e dez de seus componentes; e iii) o produto dos setores da indústria, comércio e serviços e seus respectivos segmentos que tem repsectivamente (23, 9 6) elementos.

1. Os autores agradecem os comentários feitos por Marco Cavalcanti em uma versão desta *Nota*.

Os resultados são heterogêneos: o previsor dos componentes do PIB e do produto da indústria geral e do comércio e seus segmentos apresentam um desempenho satisfatório para previsões de até seis meses à frente. O previsor do IPCA e do produto e dos serviços tem desempenho insatisfatório para previsões para horizonte superior a dois meses. No caso do produto de comércio e serviços, a variável-alvo possui uma amostra muito curta, o que limita o funcionamento desta abordagem.

2 Metodologia

O algoritmo (k) prevê o valor da variável-alvo (a) no horizonte h $(y(a,h))$ dado um conjunto de explicativas (x), determinando padrões estimados em uma amostra de treinamento, para prever $yp(a,h,k) = A_k(y(a,h)|x)$, para $k \in K$. A seguir, vamos discutir como:

1) Construir o conjunto de explicativas $\{x(i), i \in s1\}$.
2) Pré-selecionar $s1(a,h) = \{x(i), i \in s2 \subset s1\}$ das explicativas por alvo e horizonte.
3) Determinar para cada (a,h) o previsor (k^*, s^*) que minimiza a perda.
4) Aleatorizar a previsão $\{yp^w(a,h,k^*)$, para $w = 1, \ldots, 500$.

2.1 Construção do conjunto de explicativas

As variáveis explicativas $x(i)$ e a variável-alvo $y(a)$ são divulgadas com diferentes atrasos em relação ao seu fato gerador. Para medir isso, seja $at(i)$ a atualidade de uma variável i, definida como a diferença entre o período (t), para o qual essa variável $x(i)$ é apurada, e o período (td), quando ela é divulgada, ou seja, $at(i) = t(i)-td(i)$. Seja o atraso relativo da variável i em relação à variável-alvo a como $d(i,a)=at(a)-at(i)$. Quando $x(i)$ é divulgada antes da variável-alvo a, $(d(i,a)<0)$, ela descreve fatos que irão condicionar a variável-alvo no futuro. A divulgação pode ser de forma sincrônica $(d(i,a)=0)$, ou atrasada $(d(i,a)>0)$. Esses grupos de variáveis trazem informações de naturezas distintas para a previsão e a forma com que o modelo é especificado estabelece a relação entre a previsão da alvo $(y(t+h))$ e a explicativa, mostrada na tabela 1. Por exemplo, se a explicativa é divulgada com três meses de atraso, estaremos relacionado, a observação $y(t)$ com $x(t-3)$, ou, se é divulgada dois meses adiantada, estaremos relacionando $y(t)$ com $x(t+2)$.

TABELA 1
Relação entre a variável-alvo (y) e a explicativa (x(i))

$d(i,a)$	3	2	1	0	-1	-2
$H=1$	$Y(t+1)\|x(i,t-3)$	$Y(t+1)\|x(i,t-2)$	$Y(t+1)\|x(i,t-1)$	$Y(t+1)\|x(i,t)$	$\boldsymbol{Y(t+1)\|x(i,t+1)}$	$Y(t+1)\|x(i,t+2)$
$H=2$	$Y(t+2)\|x(i,t-3)$	$Y(t+2)\|x(t,i-2)$	$Y(t+2)\|x(i,t-1)$	$Y(t+2)\|x(i,t)$	$Y(t+2)\|x(i,t+1)$	$\boldsymbol{Y(t+2)\|x(i,t+2)}$

Elaboração: Grupo de Conjuntura da Dimac/Ipea.

Consideramos que é irrelevante o conteúdo de informação de uma variável divulgada com atraso relativo superior a dois meses – primeira coluna – e por essa razão foram descartadas. Além disso, não ocorreram casos que antecipam a variável-alvo em mais do que dois meses. As variáveis que antecipam um mês, $d(i,a)=-1$, são

sincrônicas às previsões com h=1, e as que antecipam em dois meses antecipam a previsão h=1 e são sincrônicas às previsões com h=2. Essa dupla função, mostrada em negrito na tabela, motivou a inclusão das explicativas desse tipo duas vezes, ou seja, como $x(i,t+1)$ e como $x(i,t+2)$ – as demais, com $d(i,a)>-2$, são incluídas uma única vez no conjunto de explicativas.

2.2 Escolhendo o previso

Dado o conjunto de informação $(y(t), x_s(t)$, para $t = 1, \dots, N)$, onde $x_s(t) = (x(i,t), i \in S)$, o algoritmo (k) prevê $yp(t+h) = A_k(x(t)|Z_s(t))$ com a explicativa $x_s(t)$, dada a amostra de treino $Z_s(t) = (Y(t), X_s(t))$ onde $Y(t) = (y(t-n+1), \dots, y(t))$ e $X_s(t) = (x_s(t-h-n+1), \dots, x_s(t-h)]$, obtendo uma sequência de previsões $yp(a, t+h, k) = A_k(x(t)|Z_s(t))$, para $t = n-h, \dots, N$. Então, para cada alvo (a), horizonte (h), algoritmo $(k \in K)$[2] e conjunto de informação (s), obtemos uma sequência de previsões $yp(a, t+h, k, s)$.

A partir destas previsões, é selecionado o algoritmo (k) e o conjunto de explicativas (s) tais que o algoritmo $(k(s,a,h) = min_k\{rmse(a,h,k,s)\}$ e $(s(a,h) = min_s\{rmse(a,h,k(s,a,h),s)\}$ onde $rmse(a,h,k,s) = \sum_{t=1..n}\big(yp(a,t+h,k,s) - y(a,t+h)\big)^2/n$.

2.3 Pré-seleção

Os algoritmos utilizados dispõem de recursos para selecionar as variáveis relevantes, mas a eficiência destes recursos tem sido questionada (Bolhuis e Rayner, 2020) e verificada empiricamente. O dilema entre fazer uma seleção arbitrária, com base no conhecimento de um analista, ou utilizar todas as explicativas que incluem a seleção arbitrária mencionada anteriormente, ou utilizar critério para pré-selecionar o conjunto anterior só pode ser resolvido, no momento, testando empiricamente todas as alternativas.

Aqui vamos apresentar um procedimento, proposto por Bolhuis e Rayner (2020), para pré-selecionar explicativas, o que identifica, de um conjunto de variáveis explicativas $s1$, aquelas que têm maior potencial explicativo para a previsão, utilizando o teste F $(R(i,a,h))$ da exclusão de $x(i)$ em $y(a,t+h) = A(L)y(a,t+h-1) + B(L)x(i,t)$ como medida do conteúdo de informação de $x(i)$ para prever $y(a,h)$. Admite-se que, quanto maior o valor desse teste, maior o valor da perda potencial de informação. A pré-seleção determina o conjunto $s1^*(a,h) = \{i \in s1 \text{ tal que } R(i,a,h) > R0\}$, onde $R0$ é um valor crítico escolhido arbitrariamente. Para evitar que esse procedimento utilize informação fora da

2. Fatores autoregressivos, vetor autorregressivo (VAR) bayesiano, *randon forest*, *complete selectiom regression*, *suport vector machine* (SVM) e *gradient boosted trees* (GBT); exponencial, auto-ARIMA e os cinco algoritmos da família lasso (*lasso, ridge, adaptative lasso, elastic-lasso* e *ad-elatic lasso*). onde os parâmetros foram determinados segundo dois critérios, o BIC e o *cross validation*. Além destes, consideramos três regras para lidar com a degeneração: *random walk*, *randon walk* de doze meses e constante.

amostra de treinamento, o modelo foi estimado utilizando os dados da primeira janela deslizante. Serão considerados três conjuntos de explicativas: i) a seleção do especialista; ii) o conjunto de todas as explicativas, incluindo a seleção do especialista; e iii) o conjunto anterior reduzido utilizando o algoritmo de pré-seleção descrito anteriormente. Na implementação, serão discutidos os procedimentos adotados.

2.4 Aleatorizando a previsão

Para cada alvo e horizonte (a, h) temos um previsor ótimo $(k(a, h), s(a, h))$ com o qual se obtém uma sequência de previsões $yp(a, t + h)$, para $t = t_0, \dots, N$, e os seus respectivos erros $e(a, t + h) = y(a, t + h) - yp(a, t + h)$, para $t = t_0, \dots, N - h + 1$. Admitindo que $e(a, t + h)$ são realizações de um processo aleatório descrito com o conjunto $\Omega(a, h) = \{e(a, t + h), t = t_0, \dots, N - h + 1\}$, podemos obter as realizações da previsão $yp^w(t + h) = yp(t + h) + e^w(a, t + h)$, onde $e^w(a, t + h)$ é sorteado aleatoriamente de $\Omega^*(a, h)$, onde $\Omega^*(a, h)$ é a massa central de $\Omega(a, h)$ descartando os erros menores (maiores) do que o percentil 5% (95%) de $\Omega(a, h)$.

Esse procedimento é uma versão do algoritmo de *bootstrap*[3] que aleatoriza de forma não paramétrica a série temporal dos erros $(e(t + h))$ e, portanto, pode não ser simétrica em relação ao zero, o que reflete características da previsão. Obtido $\{yp^w(t + h), w = 1, \dots, 500\}$, são calculadas as correspondentes realizações das transformações $fyp^w(a) = F\big(yp^w(a, T + h)\big)$ obtendo a distribuição empírica das transformações associadas ao alvo (a) $\Omega f(a) = \{fyp^w(a), w = 1, \dots, 500\}$ com a qual se obtém o intervalo de variação apresentado nas tabelas. Vale dizer que esse procedimento requer processamento intensivo.[4]

3 Resultados

As variáveis previstas – PIB, IPCA e o produto da indústria, serviço e comércio e seus componentes – têm início em períodos diferentes e características diferentes. A tabela 2 apresenta, para cada caso, o número de componentes (ou desagregações), de observações da série temporal, a data da primeira observação, o tamanho da janela (), o número de janelas deslizantes () e o conjunto de explicativas. Cada grupo de variáveis começou a ser observado em datas diferentes, o que condiciona o tamanho da amostra e limita a escolha do tamanho da janela deslizante. Por exemplo, o produto do serviço começou a ser apurado depois de janeiro de 2010,

3. Essa construção é uma adaptação do algoritmo de *bootstrap* proposto por Bergmeir, Hyndman e Benitez (2016), onde o valor previsto é o resultado do algoritmo ótimo de previsão.

4. Tempo de processamento em horas

	Fgv(p0=0)	FGV(p0=0.8)	Com	Ser	ind
Rw	36	11	4	2	84

gerando cerca de 130 observações, o que coloca um dilema entre ter janelas pequenas demais ou poucas janelas para medir a exatidão do previsor. Os números apresentados na tabela são a escolha arbitrária feita.

A tabela 2 mostra um resumo das características de cada conjunto e os principais resultados. Apresenta, para cada caso, o número de variáveis-alvo, de observações, a data inicial dos dados, o tamanho da janela deslizante oferecida aos algoritmos, e o número de janelas utilizadas para medir o desempenho do algoritmo e o número de casos previstos, ou seja o número de variáveis alvo multiplicado com o número de horizontes (6).

TABELA 2
Características

	FGV	IPCA	IND	COM	SER
1.Componentes	11	6	23	9	6
2.Observações	220	220	220	136	126
3.Início dos dados	Jan-03	Jan-03	Jan-03	Mar-10	Jan-11
4.Tamanho da janela	112	112	112	80	80
5.Janelas estimadas	110	110	110	56	46
6.Total de casos	66	36	138	54	36

Elaboração: Grupo de Conjuntura da Dimac/Ipea.

Além das explicativas descritas, foram incluídas onze variáveis indicadoras de sazonalidade mensal e uma medida de dias úteis em cada período. Foram considerado 21 algoritmos, dos quais: i) cinco são da família *lasso* – *lasso, ridge, adaptative-lasso, elastic-lasso* e *adaptative elastic lasso* – com seus parâmetros escolhidos com BIC; ii) os mesmos cinco da família lasso, mas com os parâmetros escolhidos por *cross-validation* (Friedman, Hastie e Tibshirani, 2008); iii) modelos univariados: auto-ARIMA e *exponential smoothing;* iv) modelos lineares multivariados: *large bayesian*, VAR e fatores; e v) modelos não lineares: CSR, *randon forest*, SVM e GBT.

No caso dos grupos que utilizam o conjunto de explicativas (s1), adotamos o procedimento de pré-seleção escolhendo as 20% explicativas com maior potencial preditivo e foi testada a hipóteses da igualdade entre as previsões $yp(a,h|s1)$ e $yp(a,h|s2)$. As tabelas 4 e 5 apresentam resultados para o PIB e IPCA e mostram por componente e horizonte: i) o melhor conjunto de explicativas,; ii) o melhor algoritmo; iii) o Theil-U; iv) uma medida da acurácia da previsão; e v) os mesmos itens (iii e iv) medidos para a previsão acumulada até o mês.

O índice Theil-U é a razão entre o *root mean squared error* (RMSE) do modelo selecionado e o RMSE do passeio aleatório, e funciona como uma medida de exatidão relativa: quanto menor este valor, mais exato é o algoritmo. Finalmente, a medida acurácia é a razão entre o domínio de variação da observação e o domínio[5] de variação do erro. Quanto maior esta medida, mais acurácia tem a previsão, ou seja, quando a faixa de variação do erro é pequena diante da variação da observação.

A escolha das variáveis explicativas é uma questão relevante para a qual não dispomos de um procedimento conclusivo. Os modelos algorítmicos que selecionam internamente as explicativas mais relevantes têm mostrado que essa seleção não é eficiente, especialmente diante de conjuntos de explicativas de dimensão elevada.

5. Medido como a distância entre os percentis 95% e 5% de cada caso.

Entretanto, algumas das variáveis-alvo dispõem de indicadores de antecipação específicos, caso dos segmentos dos setores para os quais são construídas medidas de antecipação de atividade, como as sondagens realizadas pela FGV. Diante dessa questão adotamos a seguinte estratégia: i) construir um conjunto com cerca de quatrocentas explicativas para prever de forma indistinta os indicadores de preço e quantidade (s1) descrito no apêndice; ii) utilizar o procedimento estatístico descrito nesta *Nota* para selecionar de um conjunto (s1) as explicativas as mais relevantes (s1*); e iii) sempre que possível, utilizar a seleção de um especialista para cada alvo s3(a). Quando se tem o conjunto s3(a), o conjunto amplo é redefinido para ($s1(a)=s1 \cup s3(a)$), e o conjunto restrito para s1(a)*. Nos resultados apresentados dispomos de explicativas do especialista apenas para os resultados setoriais.

O grupo do IPCA e do PIB referem-se a quantidades macroeconômicas para as quais não se dispõe ainda de uma seleção do especialista, por isto adotamos o conjunto *s*1. Na pré-seleção (s1*), o produto setorial e seus segmentos têm associado a cada variável-alvo um conjunto definido por um especialista que inclui medidas da sondagem da FGV, as quais antecipam em até dois meses a divulgação desse produto.

Para entender a importância da escolha do conjunto de explicativas, a escolha do previsor se dá em duas etapas. Inicialmente, obtemos o melhor previsor, dado um conjunto de explicativas (s), k(s,a,h), e posteriormente o melhor previsor k(s(a,h),a,h). Para cada alvo e horizonte (a,h), pode ocorrer que as previsões obtidas segundo o melhor previsor y(a,t+h|k(a,h)) seja similar à previsão obtida com o conjunto de explicativas (s), ou seja, não é rejeitada a hipótese da igualdade[6] entre y(a,t+h|k(a,h)) e y(a,t+h|k(s,a,h)). Neste caso, seria indiferente utilizar o conjunto de explicativas (s) ou o melhor conjunto s(a,h).

A tabela 3 apresenta, para cada conjunto de variáveis-alvo, o número de casos onde : i) o conjunto de explicativas (s1, s1*, s3) foi selecionado como melhor; ii) é indiferente utilizar um dos três conjuntos de explicativas; iii) o conjunto indicado é o melhor e estatisticamente diferente das demais previsões; iv) mais do que uma previsão é similar; e v) resultados relativos à robustez dos resultados, casos em que a seleção (algoritmo) ótima não se manteve a mesma nos últimos três meses, e o casos em que o modelo algoritmo não funcionou, pois a constante é a melhor previsão.

TABELA 3
Resultados

	FGV	IPCA	IND.	COM	SER
Total de casos	66	36	138	54	36
tipos de seleção	S1,s2	S1,s2	S1,s2,S3	S1,s2,s3	S1,s2,S3
s1	30	21	48	31	20
Ss1*	36	15	76	13	11
s3			14	10	5
Indiferente	55	34	52	35	13
s1-excl	5	1	9	5	6
s1*-excl	6	1	9	6	4
s3-excl			3	0	1
s1 ou s1* ou s3			65	8	12
Não robusta					
Seleção	1	1	4	1	11
Algoritmo	1	2	4	1	11
Prev.constante	0	8	0	1	2

Elaboração: Grupo de Conjuntura da Dimac/Ipea.

6. O teste de igualdade entre previsões foi realizado utilizando o teste de Diebold-Mariano com significância de 7% (Diebold, 2012).

Os resultados mostram que no caso do PIB e seus componentes (FGV) em 55 dos 66 ca é irrelevante o procedimento de pré-seleção, mas para cinco deles, o melhor previsor utili conjunto (S1), e para seis, o conjunto (s1*), ou seja, os algoritmos não foram capazes identificar adequadamente as explicativas relevantes em 6 dos 66 casos estudados, mostra a importância da pré-seleção. No caso do IPCA, a pré-seleção parece irrelevante, e també dificuldade de realizar a previsão, uma vez que para 8 dos 36 a melhor previsão é u constante. No caso da indústria (comércio), dos 138 (54) casos em 52 (35) é irrelevan escolha do conjunto de explicativas, mas para 9 (5) o conjunto s1(a)=s1∪s3(a estatisticamente melhor do que os demais, assim como para 9 (6) o s1(a)*. Chama a aten que para apenas 3 (0) dos casos analisados o conjunto do especialista é melhor e distinto demais. Para estes grupos de variáveis-alvo, a maioria dos previsores não são estáveis. No c do setor serviços, que dispõe da menor amostra de dados, os previsores tendem a ser instáv para 11 do 66 casos, sugerindo fragilidade dos resultados. Os demais resultados es apresentados na tabela 3.

Esses resultados mostram que: i) a pré-seleção estatística é relevante e não deveria ser dispensada; ii) a seleção do especialista ou não foi relevante ou os algoritmos foram capazes de escolhê-la dentro do conjunto s1(a); e iii) o conjunto s1(a), que é o mais oneroso em termos computacionais, não pode ser dispensado.

Seguem os resultados detalhados por grupo de explicativas. Por alvo e horizonte são apresentados o melhor algoritmo e conjunto de explicativas, o Theil-U, a Acurácia. Para sintetizar a informação indicamos com valores negativos os algoritmos e escolha de explicativas que não são estáveis nos últimos três meses, e no caso da escolha de explicativas também foi utilizada uma notação para indicar se aquela seleção é estatisticamente diferente das demais seleções. Para isso, indicamos com unidades quando a hipótese de igualdade com as outras seleções não é rejeitada. Indicamos com dezenas quando a hipótese de igualdade entre a melhor seleção e uma das outras seleções é rejeitada, e no caso em que a seleção do especialista é considerada, indicamos com centenas quando a hipótese de igualdade entre a melhor seleção e as duas outras seleções é rejeitada.

TABELA 4
Previsor do PIB e seus componentes

	Seleção de explicativas						Algoritmo						Theil-U					Acurácia				Acumulado			
	h1	h2	h3	h4	h5	h6	a1	a2	a3	a4	a5	a6	tu1	tu2	tu3	tu4	tu6	ac1	ac3	ac4	ac6	tu3	tu6	ac3	ac6
va	2	2	1	1	20	2	20	21	18	20	7	7	.35	.47	.60	.66	.58	3.4	1.7	1.8	1.5	.45	.52	3.3	4.9
imp	20	2	2	2	2	1	14	14	14	7	7	5	.25	.43	.51	.57	.58	4.3	2.0	1.7	1.5	.35	.45	3.9	2.7
pib	2	1	2	2	2	2	20	13	6	6	7	7	.32	.49	.55	.67	.63	3.7	2.4	1.7	1.5	.43	.52	3.0	4.8
cf	1	1	2	2	2	20	13	10	20	13	17	20	.27	.38	.51	.54	.50	4.4	2.0	2.1	1.4	.36	.44	3.7	3.8
cg	1	2	2	2	2	1	7	17	14	19	17	13	.68	.51	.44	.43	.60	2.5	1.9	2.0	1.4	.49	.47	2.0	4.2
fbkf	2	10	1	1	1	20	13	13	13	13	20	12	.55	.66	.76	.72	.72	3.9	2.1	1.8	1.6	.58	.65	3.2	2.3
Exp	1	1	2	2	-20	2	10	3	17	17	-6	20	.46	.60	.57	.53	.47	2.3	1.8	1.3	1.8	.55	.44	1.5	1.6
Imp	2	10	2	10	1	1	20	13	21	20	13	17	.40	.64	.67	.51	.71	2.8	1.9	1.8	1.2	.46	.44	3.0	2.4
agro	1	1	2	20	2	1	1	7	7	7	7	7	.27	.26	.23	.22	.21	4.1	3.3	3.3	3.3	.26	.17	3.1	2.9
ind	2	10	1	10	1	2	20	18	5	7	7	19	.34	.49	.50	.46	.45	4.2	2.2	2.3	1.9	.42	.43	3.6	3.6
serv	1	1	1	1	1	2	20	13	18	7	5	7	.36	.51	.63	.66	.62	3.7	1.9	1.3	1.4	.48	.58	3.0	4.1

Elaboração: Grupo de Conjuntura da Dimac/Ipea.

Os resultados mostram que:

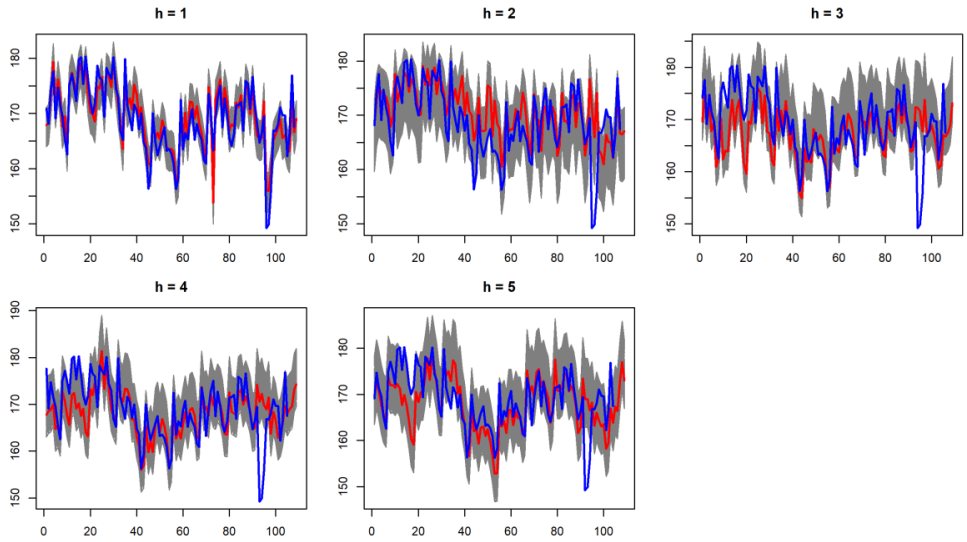
- A exatidão diminui (Theil-u aumenta) com o aumento do horizonte de previsão.

- O mesmo vale para a acurácia que aumenta com o horizonte.

- Os gráficos para o PIB mostram que os algoritmos conseguem prever os movimentos futuros, mesmo que para seis meses à frente.

As figuras subsequentes apresentam para o PIB e para cada horizonte o gráfico do valor observado $y(t+h)$ em linha azul, o valor previsto $yp(t+h|t)$ em vermelho e as realizações desta previsão – em cinza- obtida com a metodologia indicada, e mostra em cada caso se a previsão antecipou o valor futuro da variável-alvo, e qual a incerteza desta previsão.

GRÁFICO 1
Previsão do PIB



Elaboração: Grupo de Conjuntura da Dimac/Ipea.

TABELA 5
Previsor do IPCA e seus componentes

	Seleção de explicativas						Algoritmo						Theil-U					Acurácia				Acumulado			
	h1	h2	h3	h4	h5	h6	a1	a2	a3	a4	a5	a6	tu1	tu2	tu3	tu4	tu6	ac1	ac3	ac4	ac6	tu3	tu6	ac3	ac6
ipca	2	2	1	1	10	1	19	9	8	12	12	12	.53	.79	.80	.70	.69	2.1	1.2	1.1	1.1	.64	.59	1.4	1.1
mon	2	1	1	1	1	1	18	16	16	16	16	16	.81	.79	.79	.75	.75	1.1	1.0	1.0	1.0	.70	.58	1.1	1.0
bin	-2	2	2	20	1	2	-12	3	12	12	12	12	.80	.82	.79	.74	.72	1.4	1.2	1.2	1.0	.73	.63	1.2	1.2
alim	1	2	1	1	1	1	5	12	16	16	16	15	.72	.76	.75	.73	.70	1.4	1.0	1.0	1.0	.62	.55	1.1	1.1
educ	1	2	1	1	1	2	1	14	13	13	13	13	.27	.27	.27	.26	.27	4.0	4.4	4.5	5.0	.18	.13	2.5	3.2
ser	2	2	2	2	1	1	3	3	12	3	3	3	.69	.68	.78	.66	.75	1.2	1.2	1.2	1.3	.51	.41	1.5	1.8

Elaboração: Grupo de Conjuntura da Dimac/Ipea.

A tabela 5 e o gráfico 2 apresentam os resultados correspondentes para o IPCA que mostram:

- Para alguns horizontes de monitorados e alimentos, os algoritmos não funcionaram.

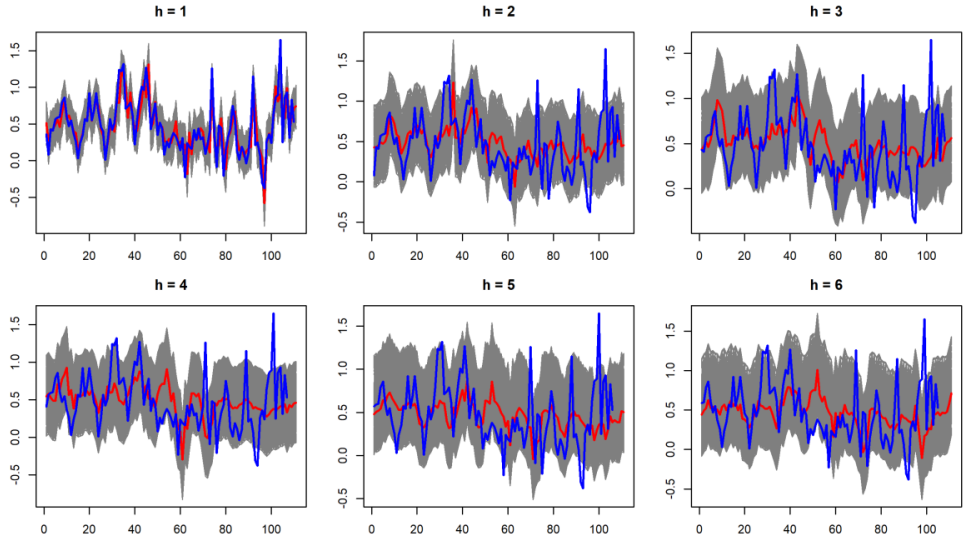
- Redução importante da exatidão (Theil-U).

- Redução da acurácia que em alguns casos alcança o valor 1, ou seja, o domínio do erro tem a mesma amplitude do domínio da observação sugerindo a baixa capacidade preditiva.

- A exatidão da previsão acumulada até o mês é maior do que a previsão para o mês, Theil-U menor.

O gráfico 2 mostra que, a partir do horizonte de três meses, a maioria das flutuação do observado fica dentro intervalo de variação da previsão anterior, sugerindo a baixa capacidade preditiva para horizontes maiores do que 2.

GRÁFICO 2
Previsão do IPCA



Elaboração: Grupo de Conjuntura da Dimac/Ipea.

A seguir, são apresentados os resultados para o produto setorial e seus segmentos. Eles seguem o mesmo formato dos casos anteriores, exceto que a escolha do conjunto de explicativas não é feita nesses casos.

TABELA 6
Previsor do produto do setor de serviço e seus componentes

	Seleção de explicativas						Algoritmo						Theil-U						Acurácia			
	h1	h2	h3	h4	h5	h6	a1	a2	a3	a4	a5	a6	tu1	tu2	tu3	tu4	tu5	tu6	ac1	ac3	ac5	ac6
Serviço	-1	-1	3	-2	-1	1	-18	-21	5	-7	-16	16	.39	.53	.62	.62	.65	.63	2.9	1.6	1.0	1.0
Família	1	10	10	10	2	1	20	10	18	13	7	7	.53	.72	.84	.80	.77	.78	4.2	1.4	1.2	1.1
Informação	2	-3	-30	300	30	-20	20	-19	-8	8	14	-21	.46	.44	.48	.44	.47	.55	2.1	1.9	1.6	1.6
Profissionais	100	-10	10	100	100	-1	13	-20	7	7	5	-5	.30	.40	.55	.56	.52	.59	3.7	1.6	1.5	1.2
Transporte	100	-20	200	-200	200	200	18	-8	14	-21	19	19	.49	.72	.68	.58	.56	.52	2.3	1.4	1.3	1.2
Outros	10	100	200	20	100	1	13	13	18	8	3	13	.57	.66	.72	.71	.74	.83	2.0	1.4	1.3	1.5

Elaboração: Grupo de Conjuntura da Dimac/Ipea.

TABELA 7
Previsor do produto do setor de comércio e seus componentes

	Seleção de explicativas						algorítmo						Theil-U						Acurácia			
	h1	h2	h3	h4	h5	h6	a1	a2	a3	a4	a5	a6	tu1	tu2	tu3	tu4	tu5	tu6	ac1	ac3	ac4	ac5	ac6
Var.ampl.	1	1	3	30	3	3	18	18	12	12	17	17	.32	.45	.53	.47	.51	.43	3.3	1.3	1.6	1.2	1.7
Var.restr.	2	3	30	30	2	2	5	21	14	6	19	19	.29	.31	.34	.33	.34	.31	3.5	2.6	2.3	2.6	2.8
Veículos	100	10	3	3	1	1	18	5	8	14	16	12	.49	.64	.76	.70	.69	.66	2.7	1.1	1.0	1.1	1.1
M.constru.	10	1	1	2	-1	1	10	10	5	19	-20	3	.55	.74	.78	.61	.73	.73	2.7	1.5	1.5	1.1	1.3
Hiper&super.	10	1	1	1	1	1	7	7	20	5	7	7	.32	.33	.33	.36	.36	.31	2.9	2.4	2.7	2.5	2.4
Tecido	200	30	2	200	2	2	20	13	18	18	18	18	.20	.23	.34	.33	.36	.38	5.6	2.5	2.9	2.8	2.7
Móveis	1	1	20	2	1	20	10	14	12	14	10	5	.59	.57	.56	.53	.54	.56	1.9	1.3	1.3	1.2	1.2
Combust.	100	100	1	1	1	1	10	5	3	3	20	5	.64	.75	.83	.77	.74	.69	2.2	1.0	1.1	1.2	1.2
Art.farmaceu.	100	10	100	200	10	1	18	5	7	18	5	18	.41	.46	.53	.43	.55	.59	3.9	3.4	3.3	2.9	2.3

Elaboração: Grupo de Conjuntura da Dimac/Ipea.

TABELA 8

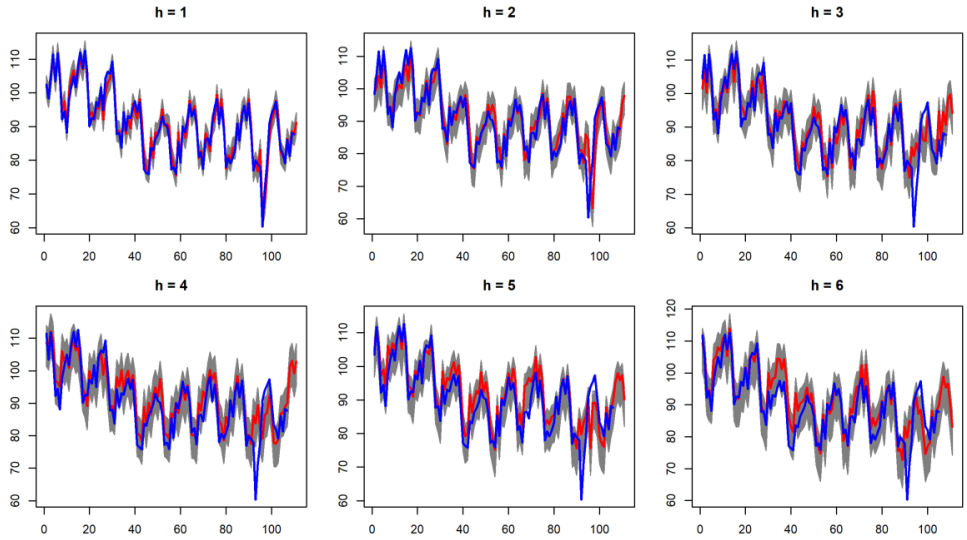
Previsor do produto do setor da indústria e seus componentes

	Seleção de explicativas						Algoritmo						Theil-U						Acurácia				
	h1	h2	h3	h4	h5	h6	a1	a2	a3	a4	a5	a6	tu1	tu2	tu3	tu4	tu5	tu6	ac1	ac3	ac4	ac5	ac6
1	2	1	20	2	2	20	17	5	21	8	14	8	.30	.48	.37	.39	.36	.37	4.7	3.1	2.9	2.5	2.5
2	200	20	20	1	3	3	19	7	7	7	7	7	.63	.73	.64	.60	.63	.57	2.3	1.9	1.6	1.4	1.3
3	20	2	200	20	1	20	20	18	14	8	5	19	.28	.49	.39	.40	.40	.38	5.8	2.6	3.0	2.8	2.0
4	-200	20	20	100	100	1	-20	17	14	20	5	13	.37	.62	.62	.54	.51	.63	5.4	2.6	2.2	2.5	2.6
5	2	2	2	30	300	300	7	7	7	7	7	7	.51	.33	.24	.20	.18	.17	2.7	2.6	2.6	2.6	2.8
6	10	10	1	100	100	100	5	13	13	13	5	5	.46	.56	.67	.65	.62	.61	3.3	2.7	1.7	1.7	1.6
7	10	20	20	3	-20	20	20	13	13	18	-13	6	.49	.48	.43	.49	.43	.42	2.8	2.2	2.2	2.1	2.1
8	10	2	200	-2	2	1	13	5	18	-7	7	20	.52	.48	.51	.58	.55	.53	2.6	2.1	1.7	1.2	1.3
9	2	10	10	20	20	2	10	3	10	20	20	6	.62	.69	.61	.61	.57	.55	1.7	1.6	1.5	1.5	1.5
10	2	10	10	10	100	-100	20	7	7	18	5	-3	.74	.64	.53	.44	.40	.39	2.2	1.8	1.7	2.0	1.8
11	10	1	20	10	10	20	18	13	7	7	7	17	.42	.39	.32	.28	.25	.23	3.6	2.6	2.3	2.5	2.4
12	2	2	20	2	20	2	5	3	3	3	7	7	.66	.72	.66	.66	.64	.60	1.2	1.0	1.0	1.0	1.0
13	2	3	1	2	2	10	7	6	7	17	19	3	.67	.55	.50	.49	.49	.50	1.6	1.7	1.7	1.6	1.7
14	200	10	2	2	2	10	5	18	13	21	12	12	.38	.61	.58	.65	.71	.70	3.5	2.2	1.7	1.5	1.5
15	1	30	2	20	2	20	18	14	21	21	12	12	.49	.70	.69	.63	.65	.63	3.3	2.4	2.4	1.9	1.7
16	20	1	2	200	2	2	17	20	18	8	6	18	.55	.63	.57	.60	.72	.78	3.0	2.5	2.7	2.3	2.2
17	10	10	20	20	20	200	13	13	14	14	12	12	.60	.72	.75	.72	.70	.68	3.1	2.2	1.9	1.7	2.2
18	3	300	1	200	20	1	20	5	3	6	6	13	.63	.64	.74	.65	.73	.92	2.6	1.7	1.6	1.6	1.5
19	10	100	20	1	10	20	5	18	8	5	7	6	.60	.67	.69	.68	.73	.69	2.5	2.2	1.8	1.5	1.7
20	10	100	3	3	3	20	5	20	8	12	12	12	.50	.60	.68	.70	.72	.68	3.4	2.1	1.7	1.5	1.6
21	10	10	10	20	20	10	5	20	7	17	13	5	.22	.68	.72	.71	.68	.69	7.3	2.3	1.9	1.9	2.0
22	20	1	200	10	2	20	13	5	13	18	20	5	.68	.69	.65	.68	.76	.77	2.5	2.4	2.6	2.8	2.5
23	20	3	20	20	20	20	8	18	8	14	21	14	.61	.55	.50	.47	.48	.46	1.9	2.0	2.1	2.1	1.9

Elaboração: Grupo de Conjuntura da Dimac/Ipea.

GRÁFICO 3

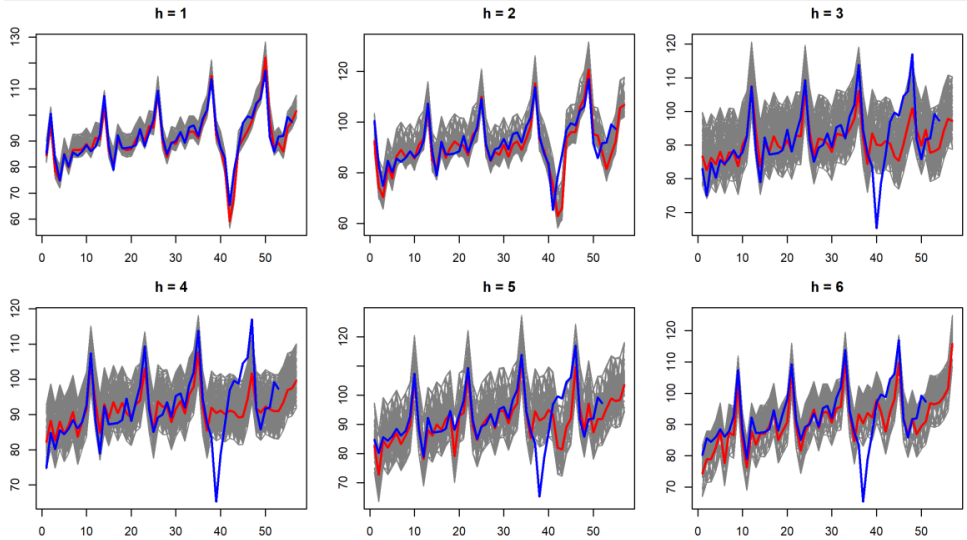
Previsão do produto da indústria geral



Elaboração: Grupo de Conjuntura da Dimac/Ipea.

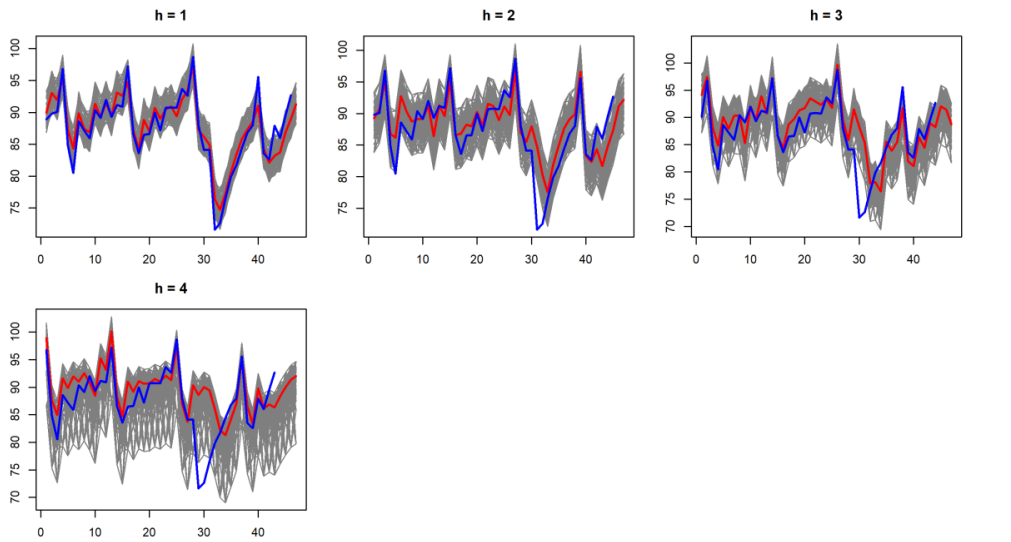
GRÁFICO 4

Previsão do produto do setor comércio



Elaboração: Grupo de Conjuntura da Dimac/Ipea.

GRÁFICO 5
Previsão do produto do setor serviço



Elaboração: Grupo de Conjuntura da Dimac/Ipea.

TABELA 9
Previsão do PIB

	Δyd(m)				ΔYd(T)			Δy(m)				ΔY(T)			ΔY(A)		
	21-06	21-07	21-08	21-09	T2	T3	T4	21-06	21-07	21-08	21-09	T2	T3	T4	T2	T3	T4
PIB	-.07	.17	1.64	-1.34	-.04	1.37	-1.13	1.10	5.18	6.87	3.68	12.08	5.24	1.77	1.68	4.06	4.83
Min.		-1.18	-2.00	-5.40		-.01	-4.29		3.77	4.81	1.88		3.81	-1.12		3.70	4.03
Max.		1.88	5.57	3.45		3.42	2.34		6.98	9.52	6.80		7.36	5.21		4.60	5.86
	Δyd(m)=yd(m)/yd(m - 1) – 1				ΔYd(T)=Yd(T(/Yd(T- 1) – 1			Δy(m)=y(m)/y(m - 12) – 1				ΔY(T)=Y(T(/Y(T- 4) – 1			ΔY(A)=Y(A(/Y(A- 1) – 1		

Elaboração: Grupo de Conjuntura da Dimac/Ipea.

TABELA 10
Previsão do IPCA

	Δy(m)				ΔY(T)			ΔY(A)		
	21-06	21-07	21-08	21-09	T2	T3	T4	T2	T3	T4
IPCA	.53	.74	.45	.56	1.68	1.77	1.69	8.67	9.23	7.39
Min.		.43	-.04	.08		.58	.41		7.95	5.35
Max.		1.03	.98	1.14		3.19	3.56		1.75	9.55
	Δy(m)=y(m)/y(m - 12) – 1				ΔY(T)=Y(T(/Y(T- 4) – 1			ΔY(A)=Y(A(/Y(A- 1) – 1		

Elaboração: Grupo de Conjuntura da Dimac/Ipea.

3.1 Previsões

A previsão necessita de transformações para ser útil no monitoramento da conjuntura econômica, e para isso apresentamos as transformações de variáveis: i) em nível que tem padrão sazonal, como os índices de atividade; e ii) em taxas, como os índices de preço. Em todos os casos, os dados são definidos pela série (y), que concatena os dados observados e previstos: $y(t)=yo(t)$,se $t \leq N$ e $yp(N+h)$,para $h=1,\ldots,6$. O valor observado (yo) da série em nível é dessazonalizado extraindo fatores sazonais que serão aplicados sobre os elementos de (y) e suas realizações, quando consideramos os resultados estocásticos. Seja (yd) a série (y) dessazonalizada, e as correspondentes agregações por trimestre calendário $Yd(T)$ e $Y(T)$, e a agregação para os últimos doze meses terminando no final do trimestre calendário indicado $Y(A)$. Na tabela está indicado o período para o qual foi calculada a taxa de variação.

4 Conclusão

Esta *Nota* apresenta a metodologia adotada e as previsões realizadas para o monitor do PIB e seus componentes; para o IPCA e seus componentes; e para o produto dos setores da indústria, comércio e serviços e seus segmentos, em um total de 42 variáveis-alvo, todas previstas para até seis meses à frente, utilizando modelos algorítmicos. Neste texto, lidamos com duas questões metodológicas: a escolha do conjunto de variáveis explicativas e uma estimativa do intervalo de confiança das previsões. A escolha das explicativas é, até onde sabemos, uma questão em aberto na abordagem de modelos algorítmicos. Estes algoritmos têm recursos internos que selecionam do conjunto de explicativas oferecido as mais adequadas para a previsão. Admitindo a possibilidade de os algoritmos não funcionarem bem quando o conjunto de explicativas é suficientemente grande, utilizamos um procedimento de pré-seleção proposto por Bouilhis e Rayner (2020), e ainda consideramos uma seleção de explicativas realizada por especialistas. O resultado da escolha do conjunto de explicativas é bastante informativo e mostra que nem sempre o maior conjunto de explicativas é o que prevê melhor, e que a seleção dos especialistas apresenta desempenho inferior ao conjunto amplo de variáveis.

Os modelos algorítmicos obtêm previsões pontuais, que podem ser insuficientes para o acompanhamento da conjuntura econômica. Admitindo que o erro da previsão é uma realização de um processo estocástico representado pelo conjunto de erro observados, este erro aleatorizado permite calcular o intervalo de confiança das previsões e suas transformações. Os resultados obtidos para os indicadores de atividade parecem satisfatórios, mas as previsões dos índices de preço apresentam um resultado insuficiente, o que entendemos ser uma questão que deveria ser revisitada no futuro,

No Brasil existem outros fornecedores de previsão econômica, e uma pergunta interessante é avaliar o desempenho desta ferramenta em comparação com esses outros fornecedores. Para isso, bastaria coletar a sequência de previsões realizadas para os diferentes horizontes e alvos, e computar por exemplo a exatidão e a acurácia destas outras previsões para compará-las com os resultados aqui obtidos.[7]

7. Planilhas anexas terminadas em G contêm os gráficos, e as terminadas em T, as previsões para as variáveis-alvo aqui consideradas.

Referências

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

BERGMEIR, C.; HYNDMAN, R. J.; BENITEZ, J. M. Bagging exponential smoothing methods using stl decomposition and box-cox transformation. **International Journal of Forecasting**, n. 32, p. 303-312, 2016.

BOLHUIS, M.; RAYNER, B. **Deus ex Machina?** A Framework for macro forecasting with machine learning. [s.l.]: IMF, 2020. (Working Paper, n. 20/45)

DIEBOLD, F. **Comparing predictive Accuracy**. Cambridge: NBER, 2012. (Working Papers, n. 18391).

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Regularization paths for generalized linear models via coordinate descent. **Journal of Statistical Software**, v. 33, n. 1, p. 1-22, 2008.

GARCIA, M.; MEDEIROS, M.; VASCONCELOS, G. Real-time inflation forecasting with high-dimensional models: the case of Brazil. **International Journal of Forecasting**, n. 33, p. 679–693, 2017.

HANSEN, P.; LUNDE, A. The model confidence set. **Econometrica**, v. 79, n. 2, p. 453-497, 2011.

Apêndice

Temas considerados:

SETOR REAL
- Indicadores de volume de atividade econômica setorial
- Receita e gastos do governo
- Volume de importações e exportações de bens e serviços
- Indicadores de utilização de capacidade produtiva

VARIÁVEIS QUALITATIVAS
- Indicadores de confiança empresarial
- Indicadores de confiança dos consumidores
- Indicadores de incerteza econômica
- Expectativas sobre inflação, renda e atividade econômica

SETOR FINANCEIRO / MONETÁRIO
- Oferta de moeda
- Taxas de juros
- Taxas de câmbio
- Demanda e oferta de crédito

PREÇOS
- Índices de preços ao produtor
- Índices de preços ao consumidor
- Volume de importações e exportações de bens e serviços
- Indicadores de utilização de capacidade produtiva

MERCADO DE TRABALHO
- Taxa de ocupação
- Taxa de desocupação
- Rendimento médio do trabalho

VARIÁVEIS EXTERNAS
- Indicadores de atividade econômica
- Índices de preços de commodities
- Volume de importações mundiais

Principais fontes:

	Nº
ASSOCIAÇÃO Brasileira de Concessionárias de Rodovias (ABCR)	6
Associação Nacional dos Fabricantes de Veículos Automotores (ANFAVEA)	26
Confederação Nacional da Indústria (CNI)	38
Fundação Getúlio Vargas (FGV)	555
Empresa de Pesquisa Energética (EPE)	12
Federal Reserve Economic Data - St. Louis Fed (FRED)	100
Fundação Centro de Estudos do Comércio Exterior (FUNCEX)	20
Secretaria de Comércio Exterior (SECEX)	25
Instituto Brasileiro de Geografia e Estatistica (IBGE)	3
Ipeadata	22
National Oceanic and Atmospheric Administration U.S. Department of Commerce	2
Banco Central do Brasil (SGS/BCB)	130
World Bank Commodity Price Data (The Pink Sheet) - FMI	15
Agência Nacional de Aviação Civil (ANAC)	3
Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP)	8