Previsão do IPCA Utilizando Árvores de Regressão com Variáveis Selecionadas por Dynamic Time Warping

Daiane Marcolino de Mattos[1], Karla Figueiredo[2]

[1] Instituto Brasileiro de Economia (FGV/IBRE), Rio de Janeiro RJ, Brazil
[2] Rio de Janeiro State University (UERJ), Rio de Janeiro RJ, Brazil
daimmattos@gmail.com

Abstract. This work has developed a methodology involving several Machine Learning and Data Mining techniques to forecast the Extended National Consumer Price Index (IPCA), which is used by the Central Bank of Brazil as the official measure of inflation in the country. The model used to select variables by Dynamic Time Warping (DTW) and performs the prediction of the single-step IPCA value with an uncertainty margin taken as the prediction of the class of that index. The results are promising and encourage the continuity of the study, especially with regard to the uncertainty margin for the index forecast.

Keywords: Inflation, Forecast, IPCA, Data Mining, Tree Regression

1 Introdução

A inflação é uma das variáveis de maior importância para a governabilidade de um país. Do ponto de vista econômico-financeiro, social e político, ela tem relação com fatores que se relacionam diretamente com a: natureza do fenômeno, à magnitude da taxa de elevação dos preços, à dimensão do fator tempo, ao caráter dinâmico do processo, a abrangência do fenômeno, aos fatores exógenos e aos mecanismos repressores [1]. O Índice de Preços ao Consumidor Amplo (IPCA), verificado mensalmente pelo Instituto Brasileiro de Geografia e Estatística (IBGE), foi criado com o objetivo de oferecer a variação dos preços no comércio para o público final. O IPCA é utilizado pelo Banco Central Brasil (BCB) como medida oficial da inflação do país. O governo também usa o IPCA como referência para verificar se a meta estabelecida para a inflação está sendo cumprida.

Os índices gerados com esse objetivo (não há apenas o IPCA) são calculados com base em cestas com centenas de produtos, que variam conforme o índice. No caso do IPCA são mais de 400 itens na cesta. Entre as diferenças dos índices estão: os dias em que os índices são apurados, os produtos incluídos nas cestas, o peso deles na composição geral e o local/faixa de população estudada.

Como tais índices afetam outros parâmetros e taxas, analisar o comportamento de tais séries temporais visando a previsão de seus comportamentos, é de grande interesse de pesquisadores e instituições governamentais e privadas [2,3,4,5]. Assim, nesse estudo, tem-se o objetivo de utilizar metodologias baseadas em Data Mining e Ma-

chine Learning para a previsão do IPCA. Destaca-se que os resultados obtidos foram muito encorajadores quando comparados com os valores publicados pelo BCB.

Nas duas próximas seções, respectivamente, apresenta-se resumidamente os fundamentos teóricos e a metodologia usada para previsão da série de interesse. A quarta seção destaca os resultados obtidos, e a seção 5 conclui o estudo.

2 Fundamentos de Data Mining e Machine Learning

2.1 Algoritmos de Agrupamento e Regressão

Nesse trabalho, visando modelagem para previsão do IPCA, utilizou-se técnicas de agrupamento e regressão. O uso de agrupamentos visa realizar a categorização dos valores do IPCA em intervalos para prever faixas ou classes de valores do IPCA. O objetivo é construir uma margem de erro ou intervalo de confiança para o índice a partir da previsão da faixa, além da previsão do valor do IPCA que será feita por um modelo baseado em Árvore de Regressão. Os intervalos que definem as faixas poderiam ser escolhidos de maneira ad-hoc, no entanto, optou-se por utilizar o algoritmo K-Means [6,7] para o agrupamento. Este tipo de método é satisfatório ao problema, pois os dados são adequados à formação elipsoide do método K-Means.

A Árvore de Regressão [8] é um método muito usado em mineração de dados e é semelhante à Árvore de Decisão, exceto porque a folha contém valores de previsões numéricas, ao invés de uma categoria ou classe. Uma Árvore de Regressão é um algoritmo baseado em árvore, o qual divide a amostra para encontrar subconjuntos semelhantes com relação ao atributo objetivo numérico. Essa decomposição de dados em subconjuntos objetiva encontrar folhas da árvore com subconjuntos de dados suficientemente pequenos ou uniformes. Os atributos que compõe a árvore são escolhidos visando reduzir a dispersão dos valores do atributo objetivo. Assim, cada nó terminal ou folha é um valor numérico (por exemplo, média) ou uma equação para o valor previsto de um determinado conjunto de dados. O algoritmo usado nesse trabalho é baseado no CART - “Classification and Regression Tree” [8]. O desenvolvimento da metodologia usou as ferramentas R e Matlab 2015.

Na metodologia adotada deseja se criar um modelo preditivo para o valor de uma variável resposta (IPCA nesse caso), com base em outras variáveis de entrada exógenas (distintas da variável de saída). A inserção de um grande número de variáveis exógenas demandou o uso de método de seleção de atributos, visando à redução do número de variáveis de entrada que deveria ser considerado no modelo. O método utilizado será brevemente discutido na próxima seção.

2.2 Seleção de Atributos

As variáveis entradas foram avaliadas a partir de um método de seleção de variáveis, visando selecionar as mais promissoras para o processo de previsão do índice IPCA. A medida de similaridade escolhida para selecionar variáveis explicativas foi o Dynamic Time Warping ou, simplesmente, DTW [9,10,11,12]. A medida é resultado

da aplicação de um algoritmo utilizado para encontrar o alinhamento não-linear ótimo entre duas sequências (séries temporais) de valores numéricos, onde o eixo do tempo é estendido ou comprimido para alcançar um ajuste razoável. Dessa maneira, é possível, sob certas restrições, encontrar padrões entre medições de eventos com diferentes frequências. Esta técnica, que originalmente foi usada em reconhecimento de fala, e que tem seu algoritmo baseado em programação dinâmica para alinhar séries temporais, considerando a minimização de uma medida de distância definida, não obedece a desigualdade triangular (não poderia ser considerado uma métrica), e por isso seu uso deve ser avaliado criteriosamente. No entanto, há evidências crescentes de que DTW é a melhor medida na maioria dos domínios [9].

3 Metodologia

A metodologia envolveu o desenvolvimento de um modelo de classificação, que possibilitou a construção de uma faixa de valores, interpretada como um intervalo de confiança. Assim, primeiro é feita a previsão da faixa de valores para o IPCA do mês imediatamente a frente, e em seguida é feita a previsão do valor para o IPCA do próximo mês.

3.1 Construção da Base de Informações

A base de informações foi definida a partir da variável de interesse, nesse caso IPCA, e de outras séries relacionadas ao poder de compra do consumidor (ver Tabela 1). Essas séries contêm 156 observações compreendidas de jan/2004 à dez/2016 e foram obtidas no site de coletadas da Sondagem do Consumidor (FGV/IBRE).

Tabela 1. Índices Economico-Financeiros Avaliados

Índice	Responsável	Descrição
IPCA	IBGE[1]	Índice de Preços ao Consumidor Amplo
IPCBr	FGV[2]	Índice de Preços ao Consumidor - Brasil
IPC-M 1	FGV	Índice de Preços ao Consumidor - Mercado
IPC-M 2	FGV	Índice de Preços ao Consumidor - Mercado
IPCA15	IBGE	Índice Nacional de Preços ao Consumidor Amplo-15 (com coleta do dia 15 do mês anterior a 15 do mês de referência)
IPC.Fipe	FIPE[3]	Índice de Preços ao Consumidor (coletados apenas no município de São Paulo)
Selic	BCB[4]	Taxa Selic
CoreIPCBr	FGV	Índice de Preços ao Consumidor - Brasil - CORE
expec_ipca_mediana	FGV	Indicador de Expectativa de Inflação dos Consumidores - Mediana.
expec_ipca_media	FGV	Indicador de Expectativa de Inflação dos Consumidores - Média
sit_econ_loc_atual	Sondagem do Consumidor - FGV	Situação econômica local atual

sit_econ_loc_fut	Sondagem do Consumidor - FGV	Situação econômica local futura
sit_financ_atual_fam	Sondagem do Consumidor - FGV	Situação financeira familiar atual
sit_financ_fut_fam	Sondagem do Consumidor - FGV	Situação financeira familiar futura
emp_loc_atual	Sondagem do Consumidor - FGV	Emprego local atual
emp_loc_fut	Sondagem do Consumidor - FGV	Emprego local futura
com_prev_bens_dur	Sondagem do Consumidor - FGV	Compras previstas de bens-duráveis
Poupança	Sondagem do Consumidor - FGV	Poupança
PIB	FGV	Produto Interno Bruto estimado mensalmente
juros	Sondagem do Consumidor - FGV	Expectativa de juros

[1]http://www.ibge.gov.br

[2]http://portalibre.fgv.br/

[3] http://www.fipe.org.br/

[4] http://dadosabertos.bcb.gov.br/

3.2 Processamento da Série IPCA

Na Figura 1 mostra-se a evolução temporal do IPCA (linha tracejada em azul), bem como a mesma corrigida após a remoção dos *outliers* (linha contínua em vermelho). Os *outliers* foram identificados utilizando *boxplot* e substituídos pela média dos valores imediatamente anterior e posterior.

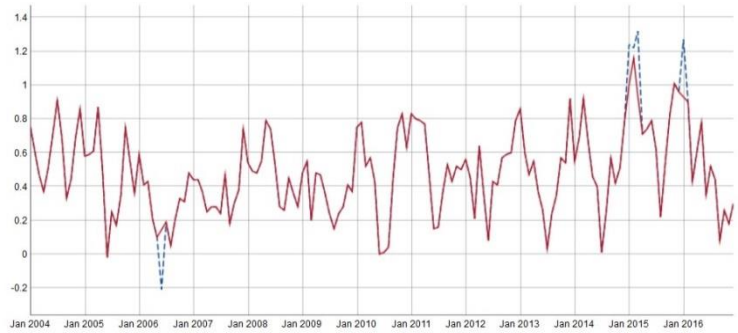

Fig. 1 - Índice Nacional de Preços ao Consumidor (IPCA) - variação percentual mensal com e sem *outliers* (fonte:IBGE)

3.3 Agrupamento dos Valores da Série IPCA

A partir do método K-Means pode-se identificar o número de grupamentos de valores da série do IPCA. O número de grupos foi escolhido de forma a minimizar o erro de classificação realizada para o IPCA, considerando diferentes números de grupos. Destaca-se que o modelo de classificação utilizado foi a Árvore de Decisão, e apenas as variáveis IPCA15, IPCM1 e IPCBr foram selecionadas através o método Correlation Based Feature Selection (CFS) [13] para serem avaliadas na composição da árvore.

Conforme observado na Tabela 2, o melhor número de grupos avaliado foi três . A Figura 2 e a Tabela 3 exibem a série do IPCA segundo as três faixas (grupos) de classificação e os valores dos intervalos para os grupos identificados como sendo o melhor agrupamento de dados, verificado por Árvore de Decisão.

Tabela 2. Grupos de Valores do IPCA identificados por clusterização (K-Means)

Nº de Clusters	Treino	Teste
2	86,80%	91,67%
3	80,55%	100%
4	75%	66,67%
5	65%	66,67%
6	55,55%	66,67%

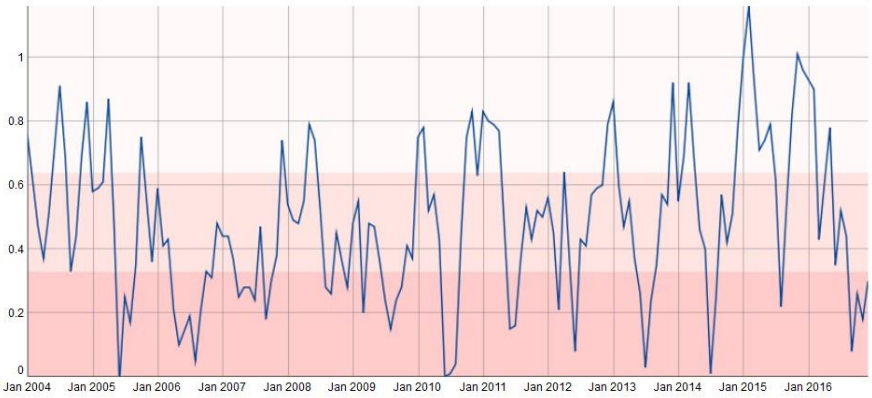

Fig 2- IPCA classificado em 3 faixas

Tabela 3. Intervalos das Faixas para o IPCA

Grupos	Mínimo	Máximo	Nº de observações
1	-0,02	0,34	39
2	0,35	0,66	70
3	0,67	1,32	35

3.4 Seleção de Séries para Previsão do IPCA

Com respeito às variáveis que poderiam ser utilizadas para a previsão do IPCA, restringiu-se o conjunto de opções àquelas que são divulgadas anteriormente ao IPCA. Por exemplo, suponha que se está interessado em prever o IPCA para o mês de junho

de 2017. Nesse caso é necessário que as variáveis explicativas no tempo t = junho/2017 estejam disponíveis antes do dia 07 de julho, data de divulgação do IPCA pelo IBGE. Utilizou-se essa restrição, pois caso contrário as variáveis explicativas também deveriam ser previstas antes de prever o IPCA. As variáveis são referentes a outros índices de inflação divulgados no Brasil, taxa de juros e a as variáveis que compõem a pesquisa de sondagem do consumidor (FGV/IBRE).

Dado o grande número de variáveis, utilizou-se uma medida de similaridade entre o IPCA e as 19 variáveis investigadas para um filtragem inicial.

A Figura 3 apresenta a distância DTW (brevemente apresentada na seção 2.2) entre o IPCA e outras 19 séries. As variáveis que mais se aproximam do desenvolvimento do IPCA ao longo do tempo, segundo essa medida de distância, são os índices de inflação IPCA-15 (IBGE), IPCBr, CoreIPCBr, IPC-M1, IPC-M2 (FGV/IBRE) e IPC-Fipe (Fipe). Em seguida aparecem os indicadores da sondagem do consumidor (FGV/IBRE) e por último a Selic. Considerando a distância DTW, avaliou-se o desempenho das nove primeiras variáveis na criação da Árvore de Regressão para prever o IPCA.

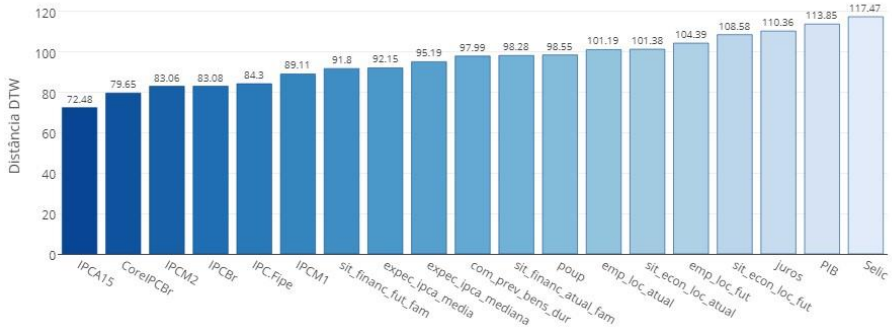

Fig. 3 - Distâncias DTW entre IPCA e outras séries temporais

3.5 Previsão dos Valores da Série IPCA

O objetivo do trabalho é inicialmente definir em que grupo: baixo, médio ou alto (já identificado no passo anterior na metodologia apresentada) o valor do IPCA do próximo mês se enquadra, dados os valores das variáveis de entrada avaliados no modelo de Árvore de Decisão. O segundo passo é a previsão do valor do IPCA oferecido pelo modelo de Árvore de Regressão, considerando variáveis selecionadas a partir do método DTW.

A base de dados com as variáveis utilizadas no estudo estão compreendidas no espaço de tempo de jan/2004 a dez/2016 (156 observações). A fim de ajustar os modelos, foram recortadas da base as últimas 12 observações (ano de 2016), restando 144 observações. Esses últimos dados foram utilizados para a validação do modelo e representam 7,7% do total de informação disponível. A previsão foi executada da seguinte forma: estima-se a Árvore de Regressão com as 144 observações disponíveis; faz-se a previsão para o tempo seguinte (observação 145) e esse valor é armazenado;

acrescenta-se a próxima informação observada (145), e faz-se a previsão da próxima observação. Dessa forma, tem-se a previsão 1 passo à frente (*single-step*) até a última observação (156). O metodologia foi construída dessa forma, à disponibilidade das informações das variáveis explicativas e visando a comparação com o Relatório Focus publicado pelo BCB.

4 Resultados

A Figura 4 apresenta a Árvore de Decisão modelada, a partir da base de treinamento, para a previsão das classes do IPCA definidas na Tabela 3. Inicialmente, no topo da árvore (nó raiz), a primeira decisão é tomada sobre a variável IPCA15. Se esta for inferior 0,29, a resposta é classe 1 (IPCA entre -0,02 e 0,33). Caso contrário, é testado se o valor do IPCA15 é menor do que 0,68. Se a resposta é positiva, a variável de saída é da classe 2, caso contrário ainda pode ser da classe 2, se o IPCBr for inferior a 0,62, ou da classe 3, se o IPCBr for igual ou superior a 0,62.

A base utilizada para o treinamento da árvore forneceu uma acurácia de 80,55%. O maior erro de predição ocorreu para a classe 3 (28,5%). Apesar disso, a acurácia de previsão (base de teste) foi de 100% (Tabela 3).

A Figura 5 apresenta a Árvore de Regressão identificada considerando o conjunto de treinamento. Dessa vez o índice IPC-Fipe foi incluído nas regras (pela avaliação do método Taxa de Ganho, intrínseco ao modelo Árvore de Regressão, na escolha dos atributos que compõem a árvore). Na raiz da árvore apresentada na Figura 5 é avaliado se o IPCA15 é inferior a 0,52. No lado esquerdo da árvore verifica-se que em 56% das observações os valores são inferiores, e no ramo direito 44% é superior a esse valor. Em seguida, a mesma variável é verificada. Caso a resposta seja positiva, questiona-se se o IPCBr é menor do que 0,11. Se a resposta a essa pergunta for verdadeira, tira-se a média das 16 observações do IPCA (11% da base de treino) e essa média será o valor definido para a variável resposta (0,12). Caso contrário, a resposta é de 0,26 (média de 24 observações). O número de nós externos (8) foi definido de forma a minimizar o erro de predição. O erro de previsão (base de teste) foi de 0,08 (*Root Mean Square Error (*eq. 1) e representa um valor pequeno.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \qquad \text{eq.1}$$

Onde: n é o número total de avaliações, $\hat{y}_i$ é o valor previsto e y_ié o valor observado.

A Figura 6 apresenta a Árvore de Regressão resultante, após se inserir a variável explicativa "classes IPCA" via Árvore de Decisão, que apresenta um erro *RMSE* igual a 0,07. O valor de entrada para a o atributo classes IPCA, no modelo regressivo, é a classe prevista pelo modelo de Árvore de Decisão para o mês objetivo. Assim, o melhor resultado de previsão para o IPCA, via Árvore de Regressão, incluiu as variáveis:

IPCA_CLASSES, IPCA15 e IPCBr. Ressalta-se que a variável IPCFipe foi excluída pelo método de seleção Taxa de Ganho intrínseco ao modelo de árvore.

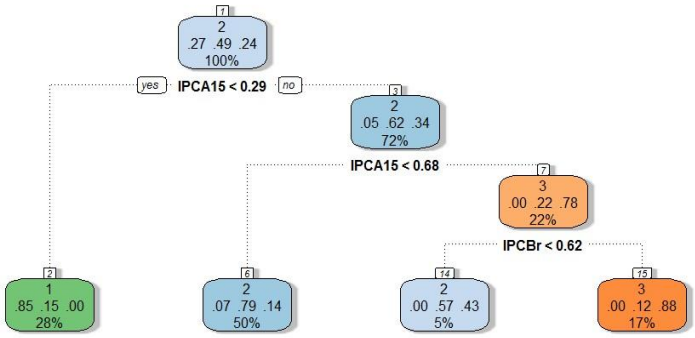

Fig. 4 - Árvore de Decisão para Classificação do IPCA por Faixas de Valores – Base Treinamento

Tabela 3 - Matriz de Confusão para a Classificação (Árvore de Decisão) - Base Teste

	Treino			Teste		
	1	**2**	**3**	**1**	**2**	**3**
1	34	6	0	4	0	0
2	5	57	10	0	5	0
3	0	7	25	0	0	3
Erro (%)	12,8	18,6	28,5	0	0	0

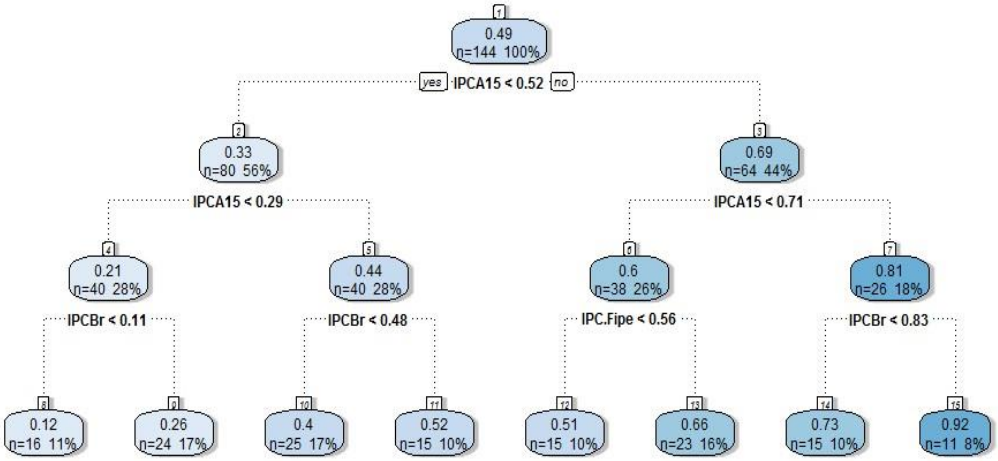

Fig. 5 - Árvore de Regressão – Base Treinamento

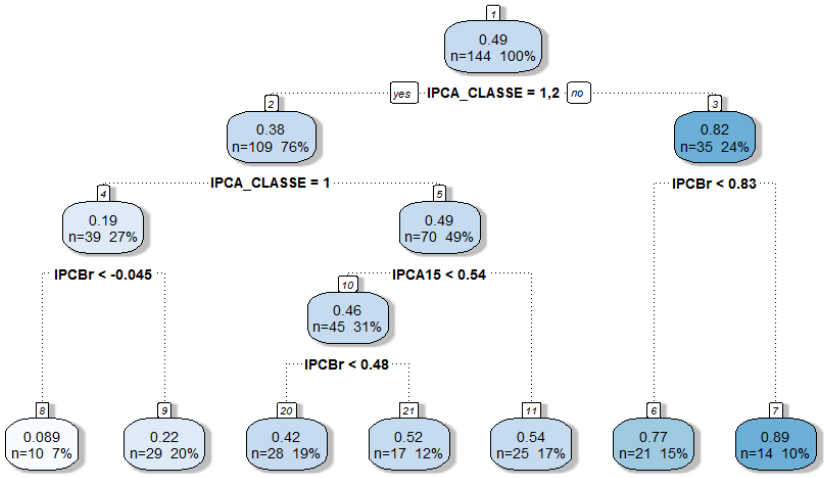

Fig. 6 - Árvore de Regressão – considerando o atributo "classe IPCA" como atributo fornecido à Árvore de Regressão - Base Treinamento

A visualização das classes e valores previstos (jan/2016 a dez/2016) pelas Árvores de Decisão e Regressão está disponível na Figura 7. A área hachurada em azul representa os valores previstos para a Árvore de Decisão, enquanto a linha sólida em vermelho representa a Árvore de Regressão. O valor observado do IPCA é indicado pela linha tracejada em azul.

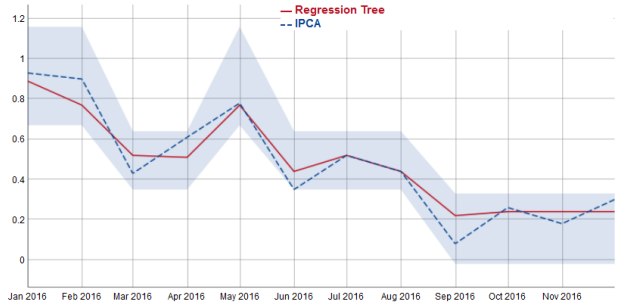

Fig. 7- Previsão para o ano de 2016 - Base de Teste

4.1 Análise dos Resultados

A partir do Relatório de Mercado Focus emitido pelo BCB pode-se comparar o desempenho do modelo proposto (ver Tabela 4). Destaca-se que o resultado obtido pelo modelo proposto é próximo à metade do erro obtido em diversas previsões publicadas pelo BCB.

Tabela 4. Resultados comparativos da previsão do IPCA

	IPCA Observado	Previsão Árvore de Regressão	Relatório Focus – Publicado pelo BCB		
			Mediana - agregado	Mediana top 5* - curto prazo	Média - top 5* - curto prazo
jan/16	0,93	0,89	0,85	0,85	0,85
fev/16	0,90	0,77	0,85	0,85	0,85
mar/16	0,43	0,52	0,55	0,46	0,46
abr/16	0,61	0,51	0,62	0,62	0,63
mai/16	0,78	0,77	0,50	0,59	0,57
jun/16	0,35	0,44	0,31	0,34	0,32
jul/16	0,52	0,52	0,40	0,45	0,44
ago/16	0,44	0,44	0,30	0,27	0,27
set/16	0,08	0,22	0,36	0,34	0,35
out/16	0,26	0,24	0,40	0,41	0,39
nov/16	0,18	0,24	0,39	0,35	0,35
dez/16	0,30	0,24	0,55	0,56	0,56
RMSE		0,0775	0,1519	0,1490	0,1688

*classificação mensal das 5 instituições com melhores previsões dentre aquelas participantes da pesquisa de mercado [13]

5 Conclusão

Nesse trabalho utilizou-se Ávores de Regressão para a previsão da série temporal do IPCA estimado mensalmente pelo IBGE. Foram utilizadas duas vertentes da técnica: Árvore de Decisão e Árvore de Regressão. A Árvore de Regressão apresentou melhores resultados para a previsão do IPCA, quando considerou como entrada a classe prevista pela Árvore de Decisão (classificação).

A separação dos valores do IPCA em classes foi identificada utilizando-se o método K-Means. O número de classes que maximizou a acurácia do modelo foi de três classes. A acurácia para a base de treinamento foi de 80,5%, enquanto para a base de teste, que considerava os últimos 12 meses do histórico de dados, foi de 100%.

A Árvore de Regressão, utilizada para prever o valor IPCA um passo à frente, foi criada utilizando 8 nós terminais, definidos de forma a minimizar o erro de predição. A acurácia foi medida pelo RMSE que retornou um valor baixo de 0,08.

Destaca-se que as variáveis que mais cooperaram para o aumento da acurácia foram do IPCA foram: IPCA-15, IPCBr e IPC-Fipe.

Os resultados apontados pelo Relatório Focus com previsão também um passo à frente indicam que a Árvore de Regressão pode ser uma boa alternativa para a previsão do IPCA, pois os resultados apontados superam a média e a mediana obtidas pelas cinco melhores previsões feitas por instituições financeiras dentre aquelas participantes da pesquisa de mercado. Ressalta-se que as variáveis explicativas utilizadas no estudo são divulgadas sempre antes do IPCA, o que torna factível a estimação das previsões. Isto é, com esse modelo só se pode obter previsões 1 passo à frente.

Como perspectiva futura de novos trabalhos deve-se aumentar o número de faixas de classificação do IPCA de forma a reduzir a margem de erro proposta para acompanhar a os valores de previsão do IPCA, reduzindo a margem de incerteza do valor previsto. Além disso, outros modelos de previsão, tais como os baseados em Reservoir Computers devem ser avaliados.

Referências

1. D.C. Ferreira, E.R.L Silva, J.P.R. Ferreira, M.S. Almeida, E. Ferreira: Causas e Efeitos da Inflação e sua significância no Cenário Econômico Brasileiro no Período de 1964 aos Dias Atuais, Disponível em <http://artigos.netsaber.com.br/artigos_letra_a/2>. Acesso em 01/06/2017
2. D.T. Mori: Construção de um Modelo de Regressão para Previsão da Inflação. Departamento de Engenharia de Produção Escola Politécnica USP. Trabalho Final de Graduação (2007).
3. N. Coelho Júnior: Utilização do Método Box-Jenkins para Previsão de Indicadores Econômicos (IPCA, SELIC, Câmbio e IBOVESPA). Programa Pós-Graduação em Macroeconomia e Finanças, Setor de Ciências Sociais Aplicadas, da Universidade Federal do Paraná (2014).
4. Ticona, W. M.; Vellasco M. B. R.; Figueiredo, K., Estudo de métodos de mineração de dados aplicados à gestão fazendária de municípios, Dissertação de Mestrado, Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro - PUC-Rio (2013).

5. G. Elliott, A. Timmermann,: Economic Forecasting, Princeton University Press (2016)
6. Steinhaus, H.: Sur la division des corps matériels en parties, Bull. Acad. Polon. Sci. (em francês). 4 (12): 801–804 (1957).
   MacQueen, J.: Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281–297 (1967).
7. Breiman, L., Friedman, J., Stone, C., Olshen, R, Classification and Regression Trees (Wadsworth Statistics/Probability) 1st ed. Chapman and Hall/CRC. (1984).
8. D.J.Bemdt, J.Clifford: Using Dynamic Time Warping to Find Patterns in Time Series, AAAI Technical Report - KDD Workshop, pp. 359–370 (www.aaai.org) (1994).
9. E. Keogh, C.A. Ratanamahatana, Exact indexing of dynamic time warping, V7, Issue 3, pp 358–386 (2005).
10. C.A.Ratanamahatana, E. Keogh: Three Myths about Dynamic Time Warping Data Mining, Society for Industrial and Applied Mathematics Philadelphia, Proceedings of the Fifth SIAM International Conference on Data Mining (2005).
11. T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria,, E. Keogh: Searching and mining trillions of time series subsequences under dynamic time warping, AAAI Press / Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence. pp.262-270 (2013).
12. M.A. Hall,: Correlation-based feature selection for machine learning. Diss. The University of Waikato (1999).
13. Acesso em 01/05/2017 - http://www4.bcb.gov.br/pec/gci/port/sobregerin.asp
14. Relatório de Mercado – Focus. Disponível em: <http://www.bcb.gov.br/pec/GCI/PORT/readout/readout.asp>. Acesso em 01/05/2017