Companhia de Planejamento do Distrito Federal

Texto para discussão

ALGORITMO PARA AUTOMATIZAR PREVISÕES

João Renato Leripio Gomes

nº 65/outubro de 2019
ISSN 2446-7502

codeplan
COMPANHIA DE PLANEJAMENTO DO DISTRITO FEDERAL

Texto para Discussão

nº 65/outubro de 2019

ALGORITMO PARA AUTOMATIZAR PREVISÕES

João Renato Leripio Gomes[1]

Brasília-DF, Outubro de 2019

[1] João Renato Leripio Gomes - Pesquisador na Diretoria de Estudos e Pesquisas Socioeconômicas da Companhia de Planejamento do Distrito Federal - DIEPS/Codeplan.

Texto para Discussão

Veículo de divulgação de conhecimento, análises e informações, sobre desenvolvimento econômico, social, político, gestão e política públicas, com foco no Distrito Federal, na Área Metropolitana de Brasília (AMB) e na Região Integrada de Desenvolvimento do Distrito Federal e Entorno (RIDE) e estudos comparados mais amplos, envolvendo os casos acima.

Os textos devem seguir as regras da **Resolução 143/2015,** que regem o Comitê Editorial da Codeplan, e não poderão evidenciar interesses econômicos, político-partidários, conteúdo publicitário ou de patrocinador. As opiniões contidas nos trabalhos publicados na série Texto para Discussão são de exclusiva responsabilidade do(s) autor(es), não exprimindo, de qualquer maneira, o ponto de vista da Companhia de Planejamento do Distrito Federal - Codeplan.

Companhia de Planejamento do Distrito Federal - Codeplan

Texto para Discussão

TD - n. 65 (2019) - . - Brasília: Companhia de Planejamento do Distrito Federal, 2019.

n. 65, outubro, 29,7 cm.

Periodicidade irregular.
ISSN 2446-7502

1. Desenvolvimento econômico-social. 2. Políticas Públicas 3. Área Metropolitana de Brasília (AMB). 4. Região Integrada de Desenvolvimento do Distrito Federal e Entorno (RIDE). I. Companhia de Planejamento do Distrito Federal. II. Codeplan.

CDU 338 (817.4)

RESUMO

A projeção de variáveis constitui importante subsídio para a formulação de políticas públicas como, por exemplo, as fiscais e demográficas, que são de grande relevância no desenho de sistemas previdenciários ou de benefícios focalizados. Adicionalmente, deve-se destacar seu papel para o planejamento, como é o caso de projeções para a atividade econômica e para a inflação na definição de parâmetros em projetos orçamentários. O objetivo deste trabalho consiste em propor um algoritmo[2] para a previsão de variáveis baseado em modelos estatísticos, porém sem arbitragem. Mais especificamente, o algoritmo implementa um conjunto amplo e diverso de métodos sobre a série temporal da variável de interesse e seleciona aquele com o melhor desempenho preditivo. A finalidade desta ferramenta é criar previsões rápidas e confiáveis de variáveis para as quais não existem referências metodológicas consolidadas ou, nos casos em que estas existam, gerar benchmarks de qualidade para comparação.

Palavras-chave: previsão, algoritmo, linguagem R.

[2] A versão de desenvolvimento está sendo construída com o nome *RAFA: R Automatic Forecasting Algorithm* e está disponível em: https://github.com/leripio/rafa. O autor agradece os comentários feitos por professores do Departamento de Estatística da Universidade de Brasília (UnB) em apresentação realizada no dia 9 de maio de 2019. Qualquer erro, contudo, é de responsabilidade do autor.

SUMÁRIO

1. INTRODUÇÃO

A projeção de variáveis constitui importante subsídio para a formulação de políticas públicas como, por exemplo, as fiscais e demográficas, que são de grande relevância no desenho de sistemas previdenciários ou de benefícios focalizados. Adicionalmente, deve-se destacar o seu papel para o planejamento, como é o caso de projeções para a atividade econômica e para a inflação na definição de parâmetros em projetos orçamentários.

Existem duas abordagens mais comuns no processo de gerar previsões. A primeira consiste na elaboração de um modelo estrutural, isto é, um sistema de equações capaz de emular o conjunto de relações que resultam no valor da variável de interesse. A segunda abordagem ignora o conjunto de relações estruturais e foca apenas nas propriedades estatísticas da série. O objetivo, neste caso, é identificar um padrão nas realizações da série que seja informativo de sua trajetória futura: tendência, sazonalidade e ciclo.

De modo geral, as referências sobre previsões econômicas envolvendo modelos estruturais baseiam-se nos agregados em nível nacional. O mais conhecido deles é o modelo SAMBA do Banco Central do Brasil (CASTRO *et al.*, 2011). Para o nível regional, entretanto, a necessidade de explorar aspectos específicos da dinâmica local requer um trabalho de pesquisa mais aprofundado – o que pode não se justificar, uma vez que o desempenho preditivo de modelos estatísticos pode ser superior ao de modelos estruturais, sobretudo no curto prazo (GURKAYNAK, KISACIKOGLU E ROSSI, 2013). Adicionalmente, nem todas as variáveis em nível nacional estão disponíveis ou têm contraparte em nível regional, o que dificulta a exata replicação dos modelos estruturais. Um exemplo importante é o Produto Interno Bruto (PIB), divulgado pelo Instituto Brasileiro de Geografia e Estatística (IBGE). Para o agregado nacional, existem dados para cada trimestre, divulgados com defasagem aproximada de dois meses. Para as unidades federativas, entretanto, as divulgações são anuais e a defasagem é de cerca de dois anos.

Neste sentido, a abordagem puramente estatística é mais popular, uma vez que não exige conhecimento prévio sobre as relações estruturais e pode oferecer boa capacidade preditiva. Todavia, esta abordagem não está isenta de desafios. Em particular, é preciso selecionar adequadamente o método que será ajustado aos dados. Esta etapa pode ser altamente custosa, dado que não existe um único método que seja superior em todos os casos (TALAGALA, THIYANGA E HYNDMAN, 2018). Portanto, é preciso testar cada um deles para a variável de interesse.

Com base no que foi exposto, o objetivo deste trabalho consiste em propor um algoritmo para automatizar a previsão de variáveis baseado em modelos estatísticos, porém sem arbitragem. Mais especificamente, o algoritmo implementa um conjunto amplo e diverso de métodos sobre a série temporal da variável de interesse e seleciona aquele com o melhor desempenho preditivo. A finalidade desta ferramenta é criar previsões rápidas e confiáveis de variáveis para as quais não existem referências metodológicas consolidadas ou, nos casos em que estas existam, gerar *benchmarks* para comparação. Em particular, o desenvolvimento da ferramenta teve como motivação as demandas por previsões para a economia do Distrito Federal, no âmbito da Codeplan, em que a escassez de relações estruturais documentadas e ausência de um amplo conjunto de variáveis dificultam a elaboração de modelos mais sofisticados. Neste sentido, trata-se de um primeiro esforço para a obtenção de previsões para a economia do Distrito Federal.

O trabalho está dividido em três partes, além desta Introdução. O Capítulo 2 descreve a metodologia proposta para o algoritmo de previsão. O Capítulo 3 traz uma aplicação da metodologia sobre a série da taxa de desocupação da Pesquisa do Emprego e Desemprego do Distrito Federal (PED-DF). O Capítulo 4 apresenta a conclusão e indicações para desenvolvimentos futuros.

2. METODOLOGIA

2.1. Pacote *forecast*

O pacote *forecast* (HYNDMAN *et al.*, 2019)[3] é uma das ferramentas mais populares para gerar previsões em linguagem R. Além de contar com os principais modelos embutidos, o pacote também traz diversos recursos complementares à tarefa de previsão. Por exemplo, funções para calcular erros de previsão por validação cruzada e gerar variações via *bootstrap* das séries temporais.

O algoritmo proposto é amplamente baseado em funções do pacote *forecast* e será descrito em detalhes na próxima seção. Entretanto, cabe destacar brevemente as principais utilidades do pacote-base que serão utilizadas, bem como o motivo pelo qual o algoritmo foi desenvolvido.

Em linhas gerais, a geração de previsões consiste em diversos passos. O primeiro deles é selecionar o modelo que será ajustado aos dados. Alguns modelos são mais adequados às características de uma determinada série do que outros. Portanto, a escolha do modelo ocorre pela sua capacidade de extrair informações dos dados que possam ser informativas da trajetória futura da série – o que costuma ser definido como a capacidade de generalização do modelo. O pacote *forecast* contém uma gama de modelos, mas não oferece nenhuma forma direta de selecionar o mais adequado para a série de interesse. Assim, o primeiro avanço do algoritmo *RAFA* consiste na seleção do melhor modelo a partir do seu desempenho fora da amostra (acurácia preditiva).[4] Adicionalmente, vale notar que as medidas de acurácia preditiva disponíveis no pacote-base só consideram a magnitude numérica dos erros. O algoritmo, além das medidas tradicionais de magnitude numérica dos erros, também contempla a possibilidade de escolher o melhor modelo através do erro direcional.

O segundo avanço do algoritmo em relação ao pacote-base diz respeito ao cálculo do intervalo de confiança. Na sua opção padrão, os modelos do pacote *forecast* retornam os intervalos de confiança gerados a partir da hipótese de que os erros de previsão seguem uma distribuição gaussiana (normal). Esta é uma limitação importante, pois nem sempre essa hipótese é válida. Além disso, esta forma de cálculo não considera a possibilidade de erro de estimação dos parâmetros do modelo selecionado. Por estas razões, o algoritmo calcula o intervalo de confiança a partir da distribuição das previsões geradas pelo modelo sobre um conjunto de variações da série de dados original obtidos por *bootstrap*. Este procedimento faz com que o intervalo obtido seja mais robusto àqueles problemas mencionados anteriormente. Ademais, a previsão pontual também é computada através da média desta distribuição -- procedimento conhecido na literatura como *bootstrap aggregating (bagging)*. Diversos estudos mostram que este método costuma performar melhor que a simples previsão sobre a série original (BREIMAN, 1996).

Em síntese, o algoritmo *RAFA* organiza as etapas de seleção, avaliação e refinamento do processo de previsão. Em relação ao pacote *forecast* que serve como base, o algoritmo constitui um avanço ao generalizar algumas tarefas e, sobretudo, implementar processos de refinamento dentro do próprio processo.

[3] O pacote *forecast* está disponível no CRAN, repositório oficial da linguagem R. Para maiores detalhes: https://cran.r-project.org/web/packages/forecast/index.html

[4] A forma como é feita a seleção será descrita na próxima subseção “O algoritmo RAFA”.

2.2. O algoritmo *RAFA*

O algoritmo foi desenhado para gerar as previsões a partir de um amplo conjunto de modelos de referência e sem qualquer tipo de arbitragem, isto é, a seleção do melhor modelo é feita com base no desempenho apurado fora da amostra. Isso é especialmente importante, pois reduz a chance de o modelo "aprender" os valores da própria amostra e foca em características com maior potencial de serem observadas no futuro. Adicionalmente, há a preocupação de gerar intervalos de confiança por meio de técnicas suficientemente robustas, visando à correta mensuração da incerteza associada às projeções.

A implementação do algoritmo é feita em linguagem R (R CORE TEAM, 2018) e tem como base as funções do pacote *forecast*. Entretanto, cabe notar que é flexível para receber futuramente novos modelos. O processo divide-se, basicamente, em quatro etapas. Estas etapas são descritas em maior detalhe abaixo e sumarizadas na Figura 4. São elas:

1. O usuário deve introduzir a série temporal (objeto "ts") para a qual deseja gerar previsões. Esta série servirá como insumo para um *loop* que ajustará todos os modelos incorporados ao algoritmo. A versão atual (1.0) conta com 12 modelos contidos no pacote *forecast*: **ARFIMA**, **ARIMA**, **ETS**, **HOLT**, **HOLT-WINTERS**, **MEANF**, **NNETAR**, **SES**, **SPLINEF**, **STRUCTS**, **TBATS** e **THETAF**. Em linhas gerais, os modelos consideram os termos autorregressivos da série ou alguma forma de suavização, o que faz com que as projeções sejam baseadas somente no padrão identificado a partir do histórico da série. Abaixo, uma breve descrição dos principais métodos empregados baseado em Spiliotis *et. al.* (2019). Cabe destacar que grande parte dos modelos é capaz de lidar com sazonalidade, o que dispensa tratamento sazonal prévio.

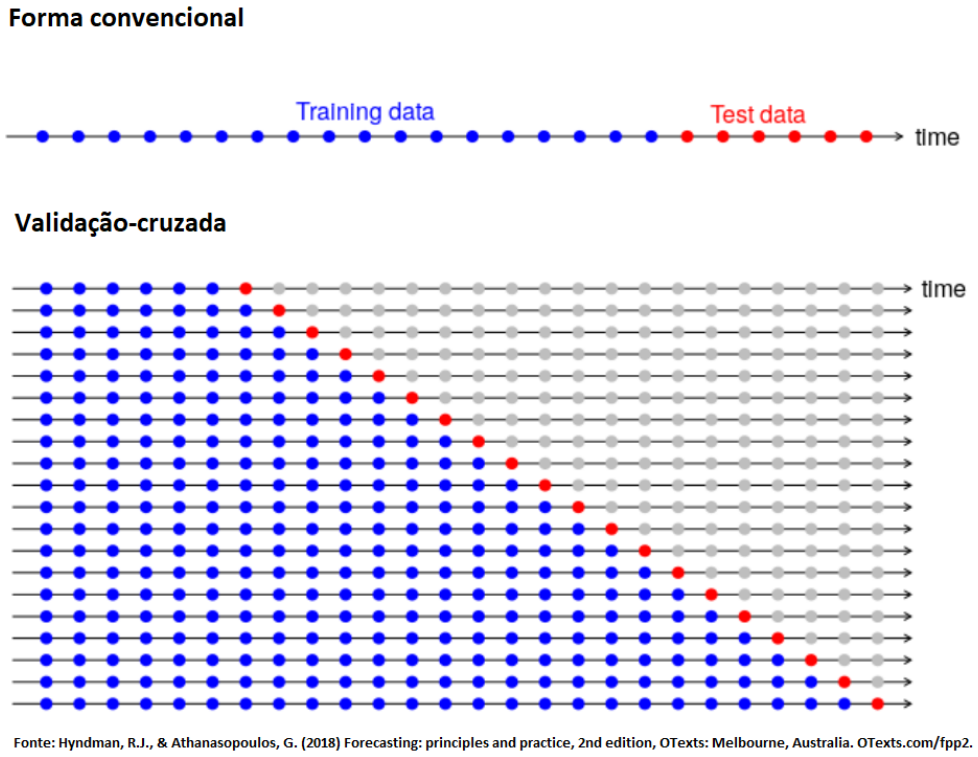
    - **ARIMA**: modelo autorregressivo integrado de média móvel. Trata-se de uma classe bem popular de modelos cuja principal característica é modelar séries que apresentam algum grau de persistência associado às realizações anteriores.
    - **ETS**: Suavização exponencial com representação em espaço-estado. É uma das ferramentas mais populares para previsão de séries genéricas dada sua capacidade em selecionar o melhor modelo de suavização exponencial.
    - **HOLT**: modelo de suavização exponencial especialmente voltado a séries com tendência linear.
    - **SES**: modelo de suavização exponencial simples, concebido para séries sem tendência linear.
    - **THETAF**: a extrapolação é feita através do método Theta, o qual considera que a série temporal é uma combinação linear de um componente de curto e outro de longo prazo.

Informações sobre os demais modelos podem ser obtidas na documentação do pacote *forecast* ou através de literatura específica.

2. Na sequência, são avaliadas a performance preditiva de cada um fora da amostra. Isto pode ser feito através de duas formas. A primeira é a forma convencional, na qual parte da amostra é dedicada para ajustar (treinar) o modelo e as observações restantes são utilizadas para teste. A segunda forma consiste no método de validação cruzada (HYNDMAN E ATHANASOPOULOS, 2018). A validação cruzada pode ser melhor entendida através da Figura 1. Ao

invés de separar uma parte da amostra para testar o poder preditivo do modelo, calcula-se o erro de previsão um (ou mais) passo(s) à frente, a partir de janelas da amostra. Essas janelas podem ser crescentes, com origem fixa como na ilustração, ou móvel com tamanho fixo. O padrão, para mimetizar uma situação mais comum, é utilizar a origem fixa e ir adicionando observações, conforme ilustrado. A opção por disponibilizar as duas formas é justificada pelo fato de que, embora apropriada para séries curtas, a validação cruzada é mais intensiva computacionalmente e pode tornar-se bastante lenta a depender do tamanho da janela e comprimento da série.

Figura 1 - Formas de testar a performance dos modelos



Fonte: R. Hyndman and G. Athanasopoulos, Forecasting: principles and practice. OTexts: Melbourne, Australia, second ed., 2018. Seção 3.4

3. Neste ponto, os modelos são ordenados a partir da estatística de erro definida pelo usuário. Duas estatísticas mais usuais estão disponíveis: Raiz do Erro Quadrático Médio (RMSE) e Erro Absoluto Médio (MAE). Adicionalmente, foi introduzida uma medida de acurácia direcional (DIR) que computa o número de acertos/erros na direção das previsões.

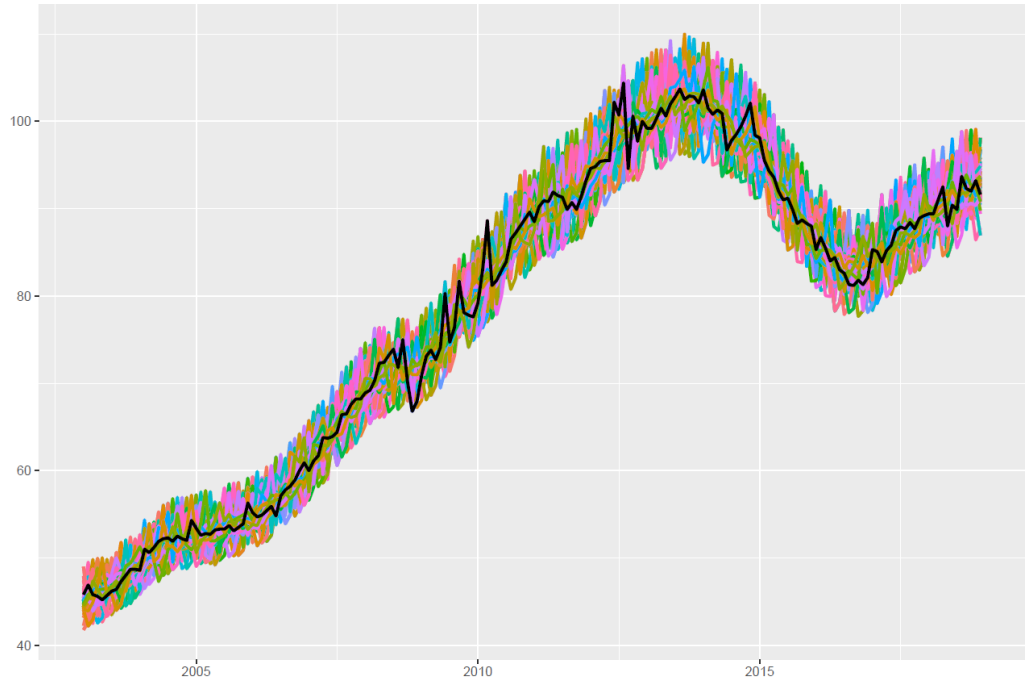
$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \qquad (1)$$

$$MAE = \frac{\left|\sum_{t=1}^{T}(y_t - \hat{y}_t)\right|}{n} \qquad (2)$$

$$DIR = \begin{cases} 1 & \text{se } \operatorname{sgn}(y_t - y_{t-1}) = \operatorname{sgn}(\hat{y}_t - y_{t-1}) \\ -1 & \text{se } \operatorname{sgn}(y_t - y_{t-1}) \neq \operatorname{sgn}(\hat{y}_t - y_{t-1}) \end{cases} \qquad (3)$$

4. Na última etapa de implementação, o modelo selecionado na etapa anterior sofre um processo de refinamento adicional conhecido como *bagging* - *bootstrap aggregating*. Trata-se, basicamente, de ajustar o modelo a um conjunto de variações da série original de dados, obtido por meio de *blocked bootstrap* e realizar as projeções para cada uma. Mais especificamente, a série temporal é decomposta em suas componentes: ciclo, tendência e sazonalidade. A parte restante (aleatória) passa por um processo de reamostragem e é reintroduzida àquelas componentes, formando variações da série original. A Gráfico 1 mostra um exemplo de séries obtidas desta maneira.[5] A série em preto é a original, ao passo que as coloridas são variações.

Gráfico 1 - Exemplo de séries geradas por *bootstrap*



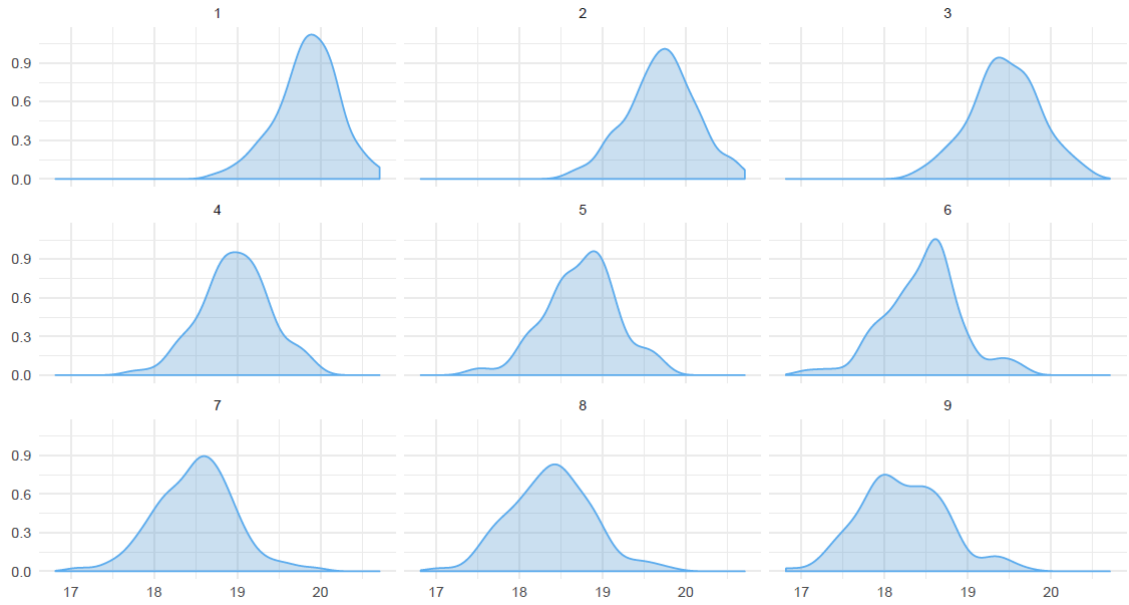
Fonte: Dados da PMC/IBGE
Elaboração: O Autor

A partir disto, obtém-se uma distribuição de projeções a partir da qual é possível extrair o valor central (média) e o intervalo de confiança (quantis) (Gráfico 2). Este procedimento resulta em estimativas mais precisas do valor central. Ademais, os intervalos de confiança computados desta maneira, além de não repousar na hipótese de normalidade dos resíduos, também capturam a incerteza associada aos parâmetros estimados pelo modelo utilizado.[6]

[5] A simulação teve como base a série do Índice dessazonalizado de volume do Comércio Varejista Ampliado para o Brasil (2014 = 100).

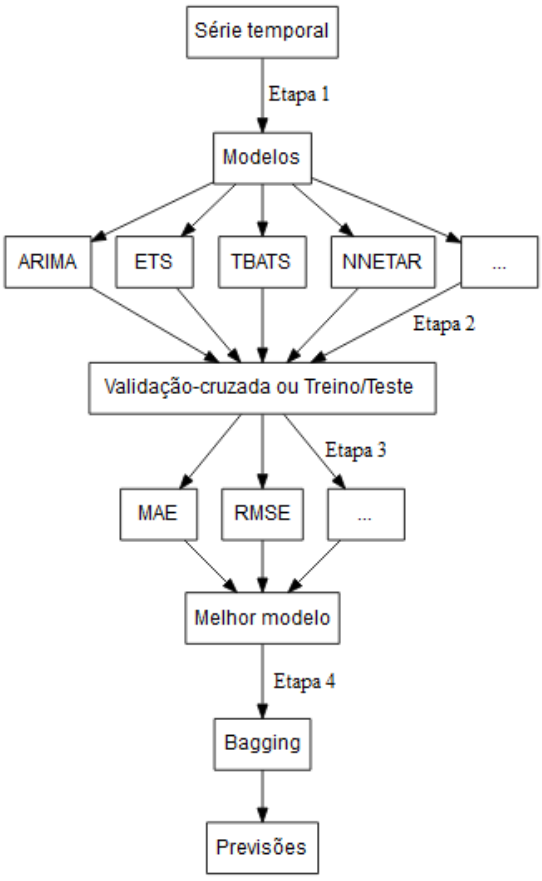
[6] Para maiores detalhes, ver Hyndman e Athanasopoulos (2018), seção 11.4.

Gráfico 2 - Exemplo de densidade das projeções para cada horizonte de tempo



Fonte: Dados do modelo
Elaboração: O Autor

Figura 2 - Etapas do algoritmo de previsão



Elaboração: O Autor

2.3. *auto_forecast*(): sintaxe da função e saídas

A função *auto_forecast*() faz parte do pacote RAFA (R Automatic Forecasting Algorithm)[7] e implementa o algoritmo descrito sobre a série temporal de interesse. Sua sintaxe na versão atual (1.0) é descrita abaixo e os valores padrão, quando definidos, estão entre parênteses:

```
auto_forecast(data, h, h_cv, window, n, level, acc, test, exclude)
```

onde:

- **data**: série temporal em formato "ts" (*time series*).
- **h**: número de períodos para a previsão (12).
- **h_cv**: número de períodos no cálculo da validação-cruzada (1).
- **window**: tamanho da janela para o cálculo da validação cruzada.
- **n**: número de séries geradas por *bootstrap* (100).
- **level**: nível de significância para o intervalo de confiança (0.05).
- **acc**: medida de acurácia para selecionar o melhor modelo (“MAE”). "RMSE" para a raiz do erro quadrático médio, "MAE" para erro absoluto médio ou "DIR" para a medida direcional.
- **test**: vetor no formato c (ano, mês) com a data ou número da primeira observação na amostra de teste. Se deixado em branco, os erros serão computados por validação cruzada.
- **exclude**: vetor com modelos que devem ser desconsiderados pelo algoritmo.

O objeto de saída da função é uma lista contendo quatro elementos:

1. ***Tibble***: contém os valores centrais, máximos e mínimos para o número de períodos (h) e de acordo com o nível de significância estipulado;
2. ***List***: contém os erros de previsão de cada modelo;
3. ***Tibble*** com as estatísticas MAE, RMSE e DIR para cada modelo; e
4. ***Ggplot*** com o gráfico da previsão com intervalos de confiança.

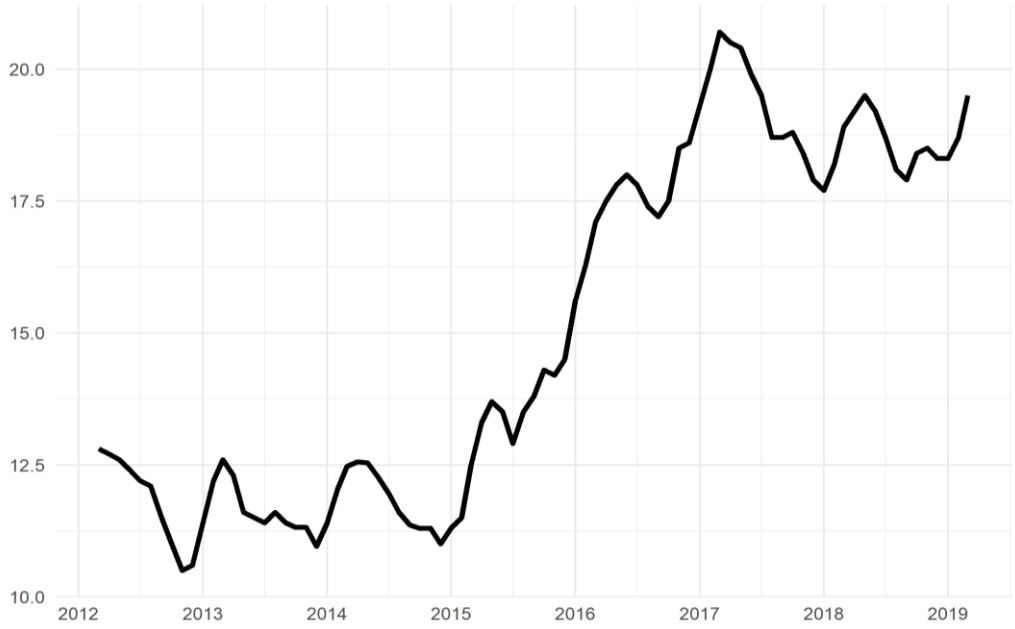
[7] A versão de testes está disponível em: https://github.com/leripio/rafa

3. APLICAÇÃO: TAXA DE DESOCUPAÇÃO DA PED/DF

Nesta Seção será apresentada uma aplicação do algoritmo utilizando a série da taxa de desocupação da Pesquisa do Emprego e Desemprego (PED) para o Distrito Federal.[8]

A amostra tem frequência mensal e compreende o período entre março de 2012 a março de 2019, totalizando 85 observações (Gráfico 3). Dada a interrupção da pesquisa entre outubro de 2013 e outubro de 2014, optou-se por preencher os valores ausentes por meio de interpolação linear levando em conta a sazonalidade da série.

Gráfico 3 - Taxa de desocupação (%) - Pesquisa do Emprego e Desemprego (PED)



Fonte: Dados da PED/DF
Elaboração: O Autor

A função foi invocada mantendo os argumentos em seus valores padrão, exceto pela janela para a validação cruzada que foi definida em 73 e a exclusão do modelo "auto.arima".[9] O objetivo, portanto, foi avaliar os modelos de acordo com os erros de previsão nos últimos 12 meses obtidos a partir de janelas móveis com 73 observações.

```
rafa::auto_forecast(ped_ts, window = 73, exclude = "auto.arima")
```

O primeiro objeto da saída a ser destacado é o quadro com as estatísticas de acurácia dos modelos (Figura 3).[10] Os modelos foram ordenados pelo valor do MAE, de

[8] A PED - Pesquisa do Emprego e Desemprego é uma pesquisa com frequência mensal realizada pelo convênio entre Seatrab-GDF, Codeplan, Seade-SP e Dieese.

[9] Em virtude de alguma característica dos dados, o modelo auto.arima reportou problemas na convergência do seu algoritmo de otimização. Por esta razão, optou-se por removê-lo do processo.

[10] Wrong dir = direção incorreta; Right dir = direção correta. Reportam o número de observações classificadas de acordo com a direção da variação.

acordo com a configuração padrão. É importante notar que, embora o modelo TBATS tenha apresentado melhor desempenho em critérios de magnitude (MAE e RMSE), o mesmo não ocorreu em termos direcionais: acertou nove vezes a direção da variação, ao passo que modelos como StrucTS e Splinef acertaram dez vezes. Portanto, caso o interesse fosse em acertar com maior frequência a direção das variações, o modelo StrucTS teria sido o melhor na combinação de direção e magnitude.

Figura 3 - Performance dos modelos ordenados pelo Erro Absoluto Médio

Model	MAE	RMSE	Wrong dir	Right dir
tbats	***0.21***	***0.26***	***3***	***9***
ets	0.26	0.32	3	9
arfima	0.28	0.31	3	9
hw	0.31	0.37	3	9
holt	0.32	0.38	3	9
StrucTS	0.34	0.39	2	10
nnetar	0.35	0.42	5	7
ses	0.35	0.41	5	7
thetaf	0.35	0.41	4	8
splinef	0.37	0.42	2	10
meanf	3.73	3.79	7	5

Fonte: Dados do modelo
Elaboração: O Autor

Na sequência, são apresentadas as projeções para os 12 meses à frente: de abril de 2019 a março de 2020 (Figura 4). As previsões indicam um leve aumento na taxa de desemprego nos dois meses seguintes, seguido de redução até encerrar o período ano (h = 9) em 18.2%. Vale notar que o modelo captura o padrão sazonal da série, o qual costuma apresentar elevações no início de cada ano.

Figura 4 - Projeções de abril de 2019 a março de 2020.

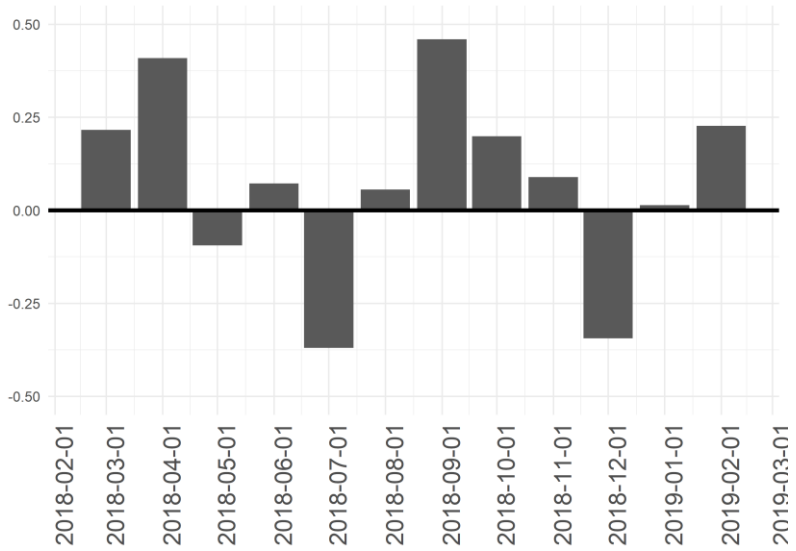
h	Point	Lower_0.95	Higher_0.95
1	19.8	18.8	20.5
2	19.7	18.6	20.5
3	19.4	18.3	20.3
4	18.9	17.9	19.7
5	18.7	17.5	19.5
6	18.4	17.3	19.4
7	18.4	17.3	19.5
8	18.3	17.2	19.5
9	18.2	17.1	19.3
10	18.6	17.5	19.9
11	19.2	18.1	20.6
12	19.8	18.6	21.3

Fonte: Dados do modelo
Elaboração: O Autor

Além das estatísticas agregadas, uma análise importante consiste em observar os erros das previsões. Em especial, isto pode indicar a presença de algum padrão que seja informativo das características das previsões como viés e risco de valores extremos. A Gráfico 4 traz os erros de previsão para a taxa de desemprego entre março de 2018 e fevereiro de 2019 a partir do modelo TBATS. Nota-se, no período em questão, que a magnitude dos erros individuais tem se situado abaixo de 0.5 p.p nos casos mais extremos. Por outro lado, há uma clara tendência das projeções em superestimar a taxa de desemprego.

Isto traz a possibilidade de implementar ferramentas adicionais que buscam incrementar a performance das previsões ao incorporar estimativas dos erros das próprias previsões, como é o caso da estratégia *rectify* (TAIEB E HYNDMAN, 2012). Mais especificamente, esta abordagem incorpora as previsões diretas (específicas para cada horizonte) dos erros de previsão para melhorar a performance preditiva dos modelos. Os resultados parecem funcionar bem para algumas séries sob certas especificações.

Gráfico 4 - Erro das projeções de março de 2018 a fevereiro de 2019



Fonte: Dados do modelo
Elaboração: O Autor

4. CONSIDERAÇÕES GERAIS

A tarefa de gerar previsões é bastante desafiadora, sobretudo quando as variáveis de interesse sofrem influência de contextos altamente voláteis, como ocorre nas ciências sociais. Nestes casos, é imprescindível tornar o trabalho mais robusto tanto no que diz respeito aos valores centrais quanto aos intervalos de confiança – medida de incerteza na realização dos dados. O algoritmo proposto parte deste princípio ao incorporar ferramentas mais robustas, tanto na escolha do modelo quanto no refinamento dos valores centrais e intervalos de confiança – por meio do método *bagging*.

Apesar dos avanços, alguns recursos com potencial para melhorar o desempenho das previsões devem ser adicionados. O primeiro deles – e talvez o mais importante – é a possibilidade de incorporar variáveis exógenas (covariáveis) com poder preditivo sobre a variável de interesse. Até o momento, os modelos incorporados realizam projeções com base na própria série e não consideram informações externas. Sob este aspecto, cabe notar que o modelo ARIMA é capaz de utilizar variáveis exógenas. Portanto, um caminho imediato seria incorporar um argumento adicional à função para utilizá-las no modelo ARIMA. Este é um passo importante, com enorme potencial, para melhorar as previsões.

Um outro recurso que pode ser considerado é a possibilidade de utilizar os próprios erros de previsão do modelo como informação adicional para aprimorar o desempenho dos modelos. Mais especificamente, a abordagem *rectify* utiliza previsões diretas (específicas para cada horizonte) dos erros de previsão para melhorar a performance preditiva dos modelos. Os resultados parecem funcionar bem para algumas séries sob certas especificações.

Por fim, vale ressaltar que o sucesso em projetar valores depende, em grande medida, da disponibilidade de dados de boa qualidade, bem como de relativa estabilidade no padrão gerador das observações. Neste sentido, além do esforço de aprimorar as ferramentas estatísticas, deve-se também buscar ampliar a disponibilidade de dados de boa qualidade.

.

REFERÊNCIAS BIBLIOGRÁFICAS

Breiman, L. **Bagging predictors, Machine Learning**, vol. 2, nº 24, pp. 126-140, 1996.

Castro, M. R. de; Gouvea, S. N.; Minella, A.; Santos, R. C.; Souza-Sobrinho, N. F. **Samba: Stochastic analytical model with a bayesian approach**. Central Bank of Brazil Working paper, abril de 2011.

Gurkaynak, Refet S; Kisacikoglu, Burçin; Rossi, B. **Do dsge models forecast more accurately out-of sample than var models?** VAR Models in Macroeconomics. New Developments and Applications: Essays in Honor of Christopher A. Sims (Advances in Econometrics), vol. 32, nº 239, pp. 27-79, 2013.

Hyndman, R; Athanasopoulos, G; Bergmeir, C; Caceres, G; Chhay, L; O'Hara-Wild, M; Petropoulos, F; Razbash, S; Wang, E; Yasmeen, F. **Forecast: Forecasting functions for time series and linear models**. 2019. R package version 8.5.

Hyndman, R; Athanasopoulos, G. **Forecasting: principles and practice**. OTexts: Melbourne, Australia, second ed., 2018.

Hyndman, R; Taieb, S. B. **Recursive and direct multi-step forecasting: the best of both worlds**. Department of Econometrics and Business Statistics. Monash University. Working paper, sep. 2012.

R Core Team, **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, Austria, 2018.

Spiliotis, E; Nikolopoulos, K; Assimakopoulos, V. **Tales from tails: On the empirical distributions of forecasting errors and their implication to risk**. International Journal of Forecasting, vol. 35, nº 2, pp. 687-698, 2019.

Talagala, G. A. Thiyanga S; Hyndman, Rob J. **Meta-learning how to forecast time series**. Department of Econometrics and Business Statistics. Monash University. Working paper, may 2018.

Revisão e copidesque
Eliane Menezes

Editoração Eletrônica
Maurício Suda

Companhia de Planejamento
do Distrito Federal - Codeplan

Setor de Administração Municipal
SAM, Bloco H, Setores Complementares
Ed. Sede Codeplan
CEP: 70620-080 - Brasília-DF
Fone: (0xx61) 3342-2222
www.codeplan.df.gov.br
codeplan@codeplan.df.gov.br

codeplan
COMPANHIA DE PLANEJAMENTO DO DISTRITO FEDERAL

Secretaria de
Economia do
Distrito Federal

Governo do Distrito Federal