Tales from tales: analisando o risco em previsões

J. Renato Leripio

29 de julho de 2019

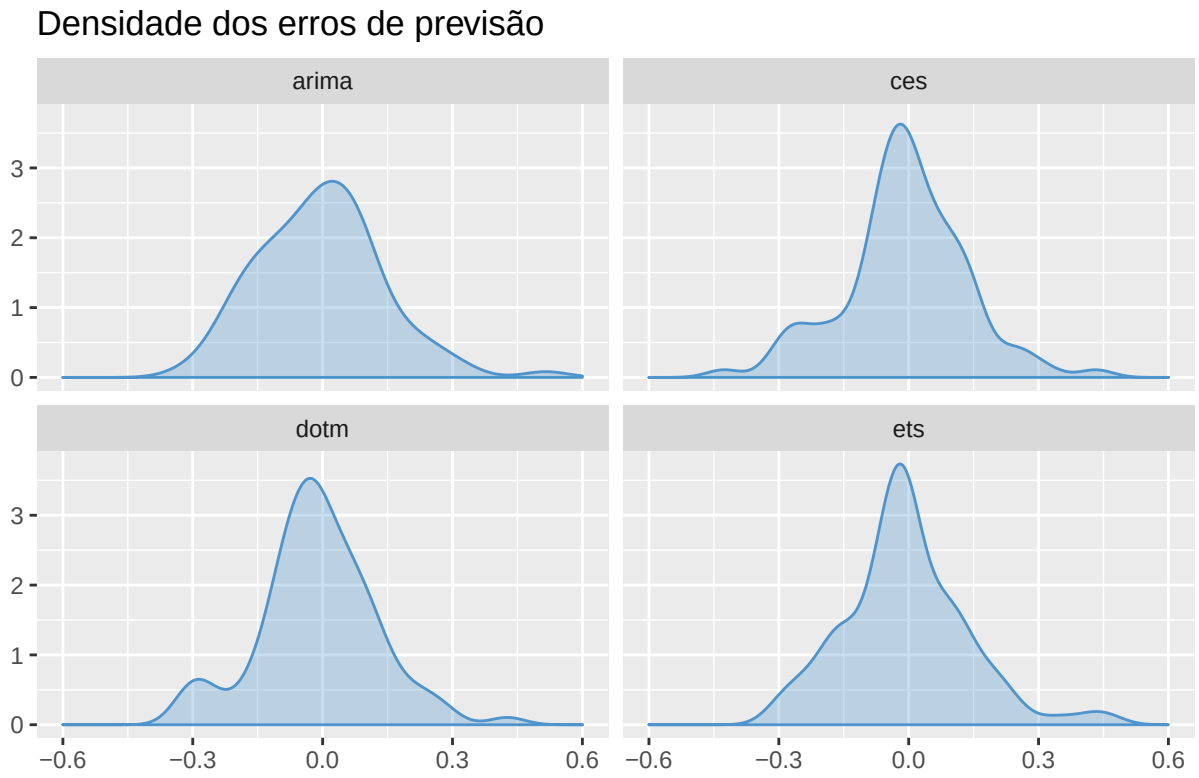
No post anterior, tratei de estratégias para combinar modelos de previsão a fim de obter melhores resultados. Melhor resultado, naquele contexto, significava apresentar menor **Root mean square forecast error (RMSFE)**.É muito comum – tanto na literatura como na prática – utilizar esta medida ou outras semelhantes que envolvam médias dos desvios quadráticos ou absolutos dos erros, por exemplo MSE, MSPE, MASE, MAPE, etc. **Sabe-se, contudo, que médias são muito sensíveis a valores extremos.** Portanto, um único valor extremo no conjunto de erros de previsão é capaz de elevar de maneira significativa aquelas estatísticas. Dito de outra forma, um modelo relativamente bom pode ser descartado porque apresentou uma única previsão ruim. Por esta razão, ganhou popularidade medidas que substituem a média pela mediana naquelas estatísticas anteriores. Os interessados em entender melhor as características destas medidas podem recorrer ao artigo de Hyndman e Koehler (2006): *"Another look at measures of forecast accuracy"*.

Neste ponto, eu gostaria de chamar atenção para uma coisa muito importante: **a estatística de acurácia utilizada para ranquear modelos é uma função-perda e, como tal, reflete o objetivo que se pretende alcançar.**Se o objetivo é reter o modelo que, em geral, não apresenta previsões muito distantes das realizações, as medidas consideradas até agora são razoáveis. Por outro lado, imagine que a previsão seja para o estoque de uma empresa ou para uma variável que define uma posição de investimento. **Nestes casos, o risco da previsão importa.**Em outras palavras, pode não fazer muito sentido escolher um modelo que, apesar de ter bom desempenho na média/mediana, apresenta maiores chances de valores extremos. No caso da empresa, tanto o excesso quanto a falta de estoque em níveis elevados pode comprometer as operações; igualmente catastrófico pode ser a realização muito abaixo/acima do previsto para uma variável-chave para o investidor.

Para usar um exemplo real, vamos considerar a mesma variável utilizada no post anterior: o núcleo do IPCA EX-3. A amostra vai de julho de 2006 a maio de 2019 e contém 155 observações. Os modelos utilizados foram ARIMA, ETS, CES e DOTM – todos vistos naquela ocasião – e os erros de previsão um passo à frente computados a partir de validação-cruzada com uma janela móvel de 60 observações – o que totalizou cerca de 95 pontos de erro para cada modelo. As densidades dos erros de previsão para cada modelo são apresentadas abaixo:

```
# 1. Carregar pacotes

library(tidyverse)
library(rbcb) # devtools::install_github("wilsonfreitas/rbcb")
library(forecast)
library(smooth)
library(forecTheta)
library(ggrepel)

# 2. Importar dados dados

dados <- rbcb::get_series(list("ipca_ex3" = 27839), end_date = "2019-05-01")

dados_ts <- ts(dados$ipca_ex3, start = c(2006,7), freq = 12)

# 4. Ajustar modelos

modelo_i <- list(
```

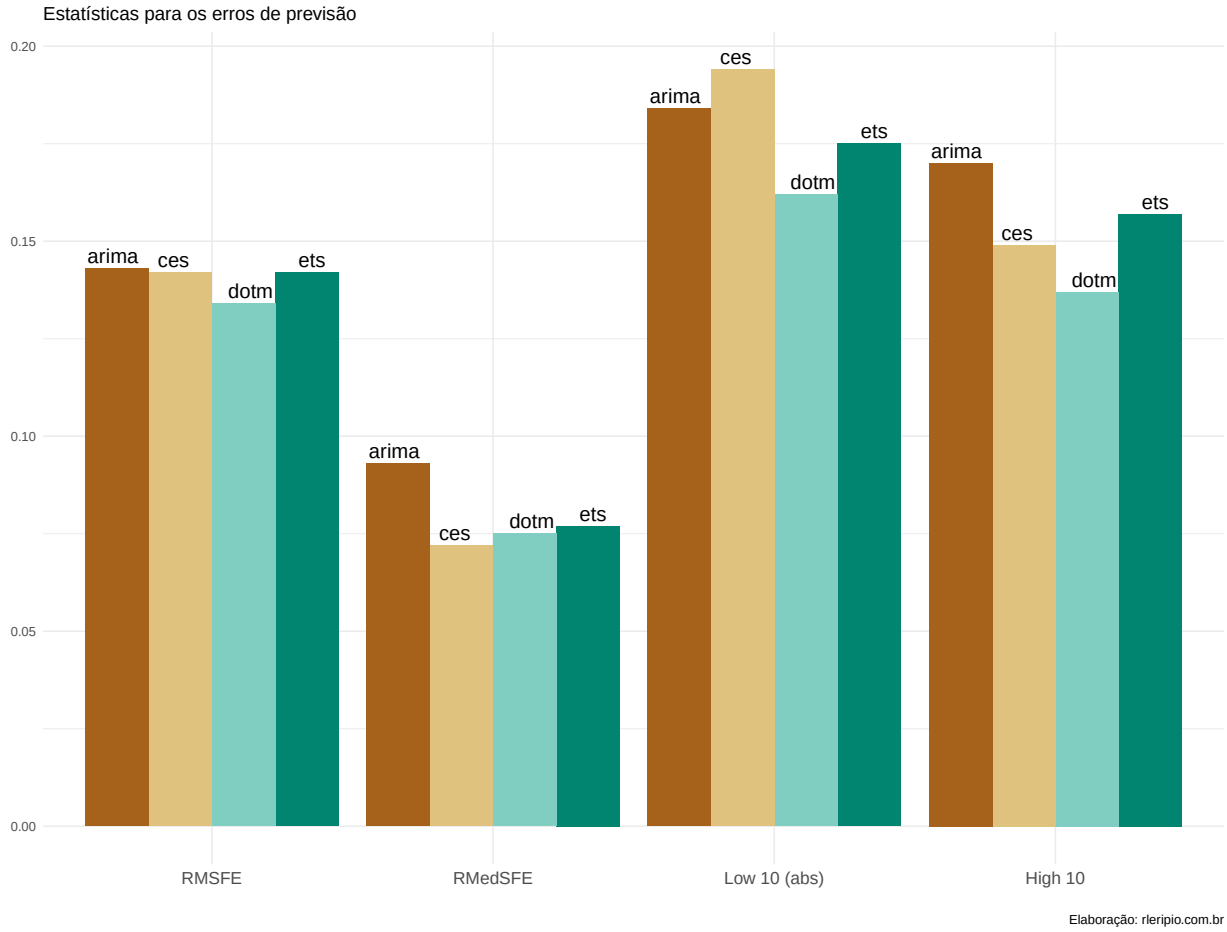
```
  "ets" = function(x,h) forecast(ets(x, lambda = "auto"), h = h),

  "ces" = function(x,h) forecast(smooth::auto.ces(x), h = h),

  "arima" = function(x,h) forecast(auto.arima(x), h = h),

  "dotm" = function(x,h) forecTheta::dotm(x, h = h)

)

# 5. Calcular os erros das previsões por validação-cruzada

fc_cv <- purrr::map2(.f = forecast::tsCV, .x = list(dados_ts), .y = modelo_i,
                     h = 1, window = 60)


fc_cv_aux <- fc_cv %>%

  magrittr::set_names(names(modelo_i)) %>%

  dplyr::bind_cols() %>%

  tidyr::drop_na() %>%

  rownames_to_column(var = "n") %>%

  dplyr::mutate(n = as.integer(n) %>% as.factor()) %>%

  tidyr::gather(key = modelo, value = erro, -n)

# 6. Gerar gráfico de densidade dos erros

fc_cv_aux %>%

  ggplot(aes(x = erro)) +

  geom_density(position = "stack", fill = "steelblue3", color = "steelblue3",
               alpha = 0.3) +

  facet_wrap(~ modelo) +

  labs(title = "Densidade dos erros de previsão",
       x = "", y = "", caption = "Elaboração: rleripio.com.br") +

  xlim(-0.6, 0.6)
```



De modo geral, todas as distribuições apresentam maior ocorrência em torno do zero, o que sugere que medidas que computam a tendência central devem ter desempenho mais ou menos parecido. Porém, cabe notar que a cauda das distribuições têm formatos bem diferentes: os erros do modelo ETS têm a cauda da direita maior que a do modelo DOTM e CES, por exemplo. É justamente neste aspecto que reside a ideia de risco: **a probabilidade de eventos extremos é maior ou menor de acordo com a área destas caudas**. Para traduzir isso de forma mais objetiva, o gráfico abaixo traz três medidas: duas de tendência central – Root mean square forecast error (RMSFE) e Root median square forecast error (RMedSFE) – e uma de risco: os limites da área com probabilidade de 10% à direita e à esquerda – esta última em valor absoluto para ficar mais fácil de visualizar com as demais.

```
# 7. Calcular as estatísticas de acurácia

fc_cv_acc <- fc_cv_aux %>%

  dplyr::group_by(modelo) %>%

  dplyr::summarise(RMSFE = erro %>% `^`(2) %>% mean() %>% sqrt(),
                   RMedSFE = erro %>% `^`(2) %>% median() %>% sqrt(),
                   "High 10" = quantile(erro, probs = 0.90),
                   "Low 10 (abs)" = quantile(erro, probs = 0.10)) %>%

  dplyr::mutate_at(vars(-modelo), ~ round(., 3))

# 8. Gerar gráfico

fc_cv_acc %>%
```

```
tidyr::gather(key = medida, value = valor, -modelo) %>%

dplyr::mutate(medida = factor(medida, levels = c("RMSFE", "RMedSFE",
                                                 "Low 10 (abs)", "High 10"))) %>%

ggplot(aes(x = medida, y = abs(valor), fill = modelo)) +

geom_col(position = position_dodge()) +

scale_fill_brewer(type = "div", palette = 1) +

theme_minimal() +

labs(x = "", y = "", fill = "",
     title = "Estatísticas para os erros de previsão",
     caption = "Elaboração: rleripio.com.br") +

geom_text_repel(aes(label = modelo), position = position_dodge(width = 1),
                angle = 0, size = 5, hjust = -0., vjust = 0.7) +

theme(legend.position = "none",
      axis.text.x = element_text(size = 12))
```



O primeiro ponto a notar é que existe diferença na classificação quando consideramos a média ou a mediana.

Pelo RMSFE, a escolha seria pelo DOTM, ao passo que pelo RMedSFE o modelo escolhido seria o CES. Por outro lado, se quiséssemos reduzir as chances de valores extremos – tanto para cima quanto para baixo – o DOTM seria a escolha inequívoca. Fica claro, portanto, a relevância de utilizar medidas aderentes aos objetivos da previsão. Adicionalmente, é sempre uma boa ideia comparar medidas alternativas para cada objetivo.

Por fim, uma questão interessante que se coloca é: **não é possível ter uma única medida capaz de sumarizar tanto a acurácia como o risco em um modelo de previsão?** A resposta parece ser positiva. No artigo *Tales from tails: On the empirical distributions of forecasting errors and their implication to risk*, os autores propõem uma medida chamada Risk measure (RM). A ideia é relativamente simples: aplica-se uma transformação do tipo Box-Cox sobre a distribuição dos erros de previsão a fim de normalizá-los e em seguida calcula-se a medida que é a soma da média com o desvio-padrão da distribuição transformada. Ao fim, a transformação é revertida. A tabela abaixo computa uma versão da medida RM para os erros de previsão dos modelos (transformados via Box-Cox) em conjunto com o p-valor associado ao teste Shapiro de normalidade.

```
fc_cv_aux_bc <- fc_cv_aux %>%

  dplyr::group_by(modelo) %>%

  dplyr::mutate(erro_bc = forecast::BoxCox(erro, lambda = "auto"))

fc_cv_rm <- fc_cv_aux_bc %>%

  dplyr::summarise(MAE = mean(abs(erro_bc)),
                   SD = sd(erro_bc),
                   RM = InvBoxCox(MAE + SD, lambda = forecast::BoxCox.lambda(erro)),
                   shapiro_bc = shapiro.test(erro_bc)$p.value) %>%

  dplyr::mutate_at(vars(-modelo), ~ round(., 2)) %>%

  dplyr::arrange(RM)

knitr::kable(fc_cv_rm)
```

modelo	MAE	SD	RM	shapiro_bc
dotm	0.94	0.11	2.03	0.02
ces	1.31	0.26	2.79	0.93
arima	1.41	0.32	3.09	0.06
ets	3.02	1.23	13.48	0.00

O problema é que nem sempre a distribuição transformada é normal. O modelo DOTM apresentou a menor RM, porém os resultados não são significativos uma vez que somente a distribuição dos erros do modelo CES – e com alguma "boa vontade" a do ARIMA – se aproximou de uma distribuição normal. **De fato, em termos de p-valor, a distribuição original dos erros apresentou resultados melhores para o teste de normalidade: respectivamente 0.14, 0.15, 0.11 e 0.02.** Ainda assim, o ideal seria ter p-valores maiores para dar mais segurança.

Ficou alguma dúvida ou tem sugestões? Entre em contato!

Os códigos dos exercícios encontram-se disponíveis no repositório do blog no github.

Siga nossa página RLeripio – Economia e Data Science no Facebook e fique sabendo de todas as nossas publicações!

Aviso legal: Todo o conteúdo desta página é de responsabilidade pessoal do autor e não expressa a visão da instituição a qual o autor tem vínculo profissional.