Universidade de Brasília
Departamento de Estatística

Previsão do Número de Acessos Diários a Páginas do Wikipedia

Pedro Vianna Alves da Silva

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2022

Pedro Vianna Alves da Silva

Previsão do Número de Acessos Diários a Páginas do Wikipedia

Orientador(a): Prof. Dr. José Augusto Fiorucci

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2022

Agradecimentos

Primeiramente, gostaria de agradecer à minha família por todo o apoio que eles me deram durante a minha graduação. A liberdade para eu fazer o que eu quisesse e me deixasse feliz nesse período foi fundamental para chegar neste momento. Queria agradecer também aos professores da Universidade de Brasília que fizeram parte de todo este processo, em especial ao Prof. José Augusto Fiorucci, com quem pude trabalhar de maneira mais próxima neste trabalho, e a Prof. Maria Teresa Leão Costa, por ter me guiado durante meu período como monitor.

Resumo

O seguinte estudo tem como objetivo realizar previsões do número de acessos diários a páginas do Wikipedia, utilizando estruturas hierárquicas e modelos da família ARIMA e TBATS. Em uma análise inicial de validação cruzada, utilizando apenas um nível total na hierarquia, indicou-se que os modelos ARIMA seriam mais adequados para a previsão deste conjunto de séries temporais. Apesar disso, nas previsões finais, observou-se o contrário, ou seja, os modelos TBATS obtiveram uma melhor acurácia em relação ao ARIMA, tanto quando utilizado apenas um nível total na hierarquia quanto quando as séries foram agrupadas segundo seu projeto. Nesse aspecto, o fato de os modelos com a estrutura com apenas o nível total ter obtido resultados um pouco melhores chamou atenção, pois esperava-se que a divisão dos projetos pudesse melhorar as previsões.

Palavras-chaves: séries temporais, modelagem preditiva, previsão, competição, janela deslizante, TBATS, ARIMA.

Lista de Tabelas

Lista de Figuras

Sumário

1 Introdução

As séries temporais estão presentes no dia a dia de diversas pessoas. Elas servem para observar o progresso de certas estatísticas no decorrer do tempo. Para um gerente de supermercado, podem servir para analisar a evolução das vendas de sua loja; para um economista, podem ser utilizadas para observar o aumento do preço de algum produto durante uma crise econômica; para um epidemiologista, avaliar a taxa de contágio de uma certa doença infecciosa. Inúmeras outras aplicações podem ser listadas. Apesar desta observação de dados históricos ser muito interessante e trazer diversas informações relevantes para quem as consome, as previsões se sobressaem como um interesse especial quando fala-se de séries temporais.

O entendimento de uma pequena parte do futuro, baseado apenas em padrões observados no passado, permite que diversos indivíduos, sejam eles governantes ou empresários, possam tomar decisões visando aproveitar oportunidades ou evitar catástrofes. O grande avanço tecnológico e o consequente aumento do poder computacional nos últimos anos permitiram que diversos modelos e algoritmos de previsão fossem criados. Estes visam utilizar as tecnologias atuais para realizar as previsões mais precisas e de forma rápida e simples.

O grande número de dados que são gerados nos dias de hoje permite também que modelos preditivos possam ser testados em larga escala. Dessa forma, análises comparativas do desempenho desses modelos podem ser feitas de forma mais robusta. Pensando nisso, este trabalho visa comparar a previsão de modelos preditivos em dados de acessos diários de 145.063 artigos do *website* Wikipedia, em que os dados estão disponíveis na competição *Web Traffic Time Series Forecasting*, organizada pelo Kaggle.

Tal banco de dados envolve diversos fatores não triviais de serem analisados e considerados em qualquer modelagem preditiva, como a manipulação de um grande conjunto de dados formado por múltiplas séries temporais altamente correlacionadas bem como a presença de múltiplas sazonalidades e hierarquia. Essas características permitem testar os modelos preditivos em um cenário adverso, em que a dificuldade de se obter boas previsões é mais elevada.

Neste estudo serão considerados dois modelos distintos, ambos paramétricos (ARIMA e TBATS). Estes serão aplicados em dois contextos diferentes, utilizando-se estruturas hierárquicas. No primeiro, as séries serão agrupadas em apenas um nível, que corresponderá ao total de acessos diários às páginas presentes no conjunto de dados. No segundo,

elas serão agrupadas em relação ao projeto que fazem parte, utilizando esta informação como um estrato.

2 Metodologia

2.1 Conjunto de dados

O conjunto de dados em estudo faz parte da competição *Web Traffic Time Series Forecasting*, organizada pelo Kaggle com apoio do Google, e foi realizada em 2017. Dos arquivos disponibilizados, dois serão utilizados neste trabalho.

O principal banco de dados é formado por 145.063 observações, em que cada uma destas corresponde a uma combinação de um artigo do Wikipedia, o projeto em que este artigo está inserido, a forma de acesso e o agente utilizado. As colunas deste banco de dados são:

- **Page:** combinação do nome do artigo, projeto, forma de acesso e tipo de agente;
- O restante das colunas corresponde ao número de acessos diário, no período de 01/07/2015 a 10/09/2017; cada uma destas datas está representada em uma coluna, gerando assim 803 colunas com os dados de acessos.

Seguem, abaixo, as primeiras 6 linhas e colunas deste arquivo como exemplo.

Page	**X2015.07.01**	**X2015.07.02**	**X2015.07.03**	**X2015.07.04**	**X2015.07.05**
2NE1_zh.wikipedia.org_all-access_spider	18	11	5	13	14
2PM_zh.wikipedia.org_all-access_spider	11	14	15	18	11
3C_zh.wikipedia.org_all-access_spider	1	0	1	1	0
4minute_zh.wikipedia.org_all-access_spider	35	13	10	94	4
52_Hz_I_Love_You_zh.wikipedia.org_all-access_spider	*NA*	*NA*	*NA*	*NA*	*NA*
5566_zh.wikipedia.org_all-access_spider	12	7	4	5	20

Tabela 1: Primeiras 6 linhas e colunas do conjunto de dados das séries temporais

Este banco de dados apresenta valores ausentes, representados no quadro acima com o valor *NA*. Esses valores podem representar que o artigo não teve nenhum acesso no dia especificado, ou que os organizadores da competição não conseguiram obter os dados referentes a esses dias.

O envio das previsões para a competição deve ser feito em um arquivo *.csv* com duas colunas: uma identificadora, composta por uma chave específica para cada combinação da variável Page com a data da previsão, e uma com o número de acessos previstos. Para obter as chaves, a competição disponibilizou um arquivo com cada uma delas. As seis primeiras linhas desse arquivo seguem abaixo.

Page	Id
007_ja.wikipedia.org_all-access_all-agents_2017-09-13	0b293039387a
007_ja.wikipedia.org_all-access_all-agents_2017-09-14	7114389dd824
007_ja.wikipedia.org_all-access_all-agents_2017-09-15	057b02ff1f09
007_ja.wikipedia.org_all-access_all-agents_2017-09-16	bd2aca21caa3
007_ja.wikipedia.org_all-access_all-agents_2017-09-17	c0effb42cdd5
007_ja.wikipedia.org_all-access_all-agents_2017-09-18	4ccd369adefc

Tabela 2: Primeiras 6 linhas e colunas do conjunto de dados das chaves das previsões

Assim, este arquivo será utilizado para conectar as previsões com as suas chaves identificadoras, para que a competição valide estas previsões e gere a pontuação final dos modelos utilizados neste estudo.

2.2 Ferramentas computacionais

A preparação e análise dos dados foi feita utilizando o *software* **RStudio**, em sua versão 4.1.1. Além disso, foram utilizados algumas funções dos pacotes **vroom** (importação dos dados), **tidyverse** (preparação e visualização), **forecast** (modelagem e previsão) e **tseries** (análise de resíduos).

2.3 Janela deslizante

A área da análise preditiva consiste em buscar os melhores modelos para explicar e prever dados de um certo assunto. A avaliação dos modelos pode ser feita de diversas formas, mas a validação cruzada é um dos métodos mais utilizados neste quesito. Ela consiste em separar o conjunto de dados em duas partes, uma que será utilizada para definir o modelo, que geralmente consiste de 80% do total dos dados, e outra para validação deste, calculando assim métricas de avaliação para quantificar a qualidade do ajuste.

Um dos métodos específicos de validação cruzada corresponde à utilização de janelas deslizantes. Neste conceito, utilizam-se diferentes janelas de treinamento e de teste, que podem variar de tamanho. Dessa forma, cada horizonte de previsão é avaliado múltiplas vezes. Esse sistema de repetição de treinamento e validação em diferentes partes do conjunto de dados gera resultados mais robustos, ajudando na escolha do melhor modelo.

A Figura 1 apresenta uma representação gráfica do processo de validação cruzada utilizando janelas deslizantes, mostrando como o banco de dados é dividido e como o horizonte de previsão se comporta à medida que o método é aplicado.

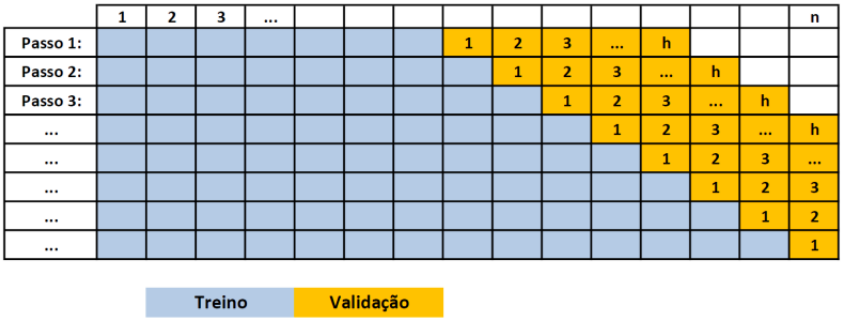

Figura 1: Demonstração do mecanismo das janelas deslizantes

FONTE: Notas de aula do Prof. Dr. José Augusto Fiorucci

Para avaliar as previsões deste processo, será utilizado o Erro Absoluto Médio (MAE) de cada horizonte de previsão. Esta métrica é definida por

$$MAE_i = \frac{1}{n_i} \sum_{j=1}^{n_i} |y_{ij} - \hat{y}_{ij}|,$$

em que:

- n_i: número de previsões realizadas para o i-ésimo horizonte de previsão, em que $i = 1, \ldots, h$;
- y_{ij}: valor real para o j-ésimo valor do i-ésimo horizonte de previsão;
- $\hat{y}_{ij}$: valor previsto para o j-ésimo valor do i-ésimo horizonte de previsão.

Portanto, ao final da aplicação da validação cruzada por janelas deslizantes, será possível estimar o desempenho preditivo dos modelos ajustados para diferentes horizontes de previsão, e notar como este comportamento varia à medida que este número aumenta.

2.4 Modelos ARIMA

Os modelos ARIMA estão detalhados em Morettin e Toloi (2019).

2.4.1 Operadores

Os operadores de diferença e retardo serão utilizados nesta seção para definir os modelos da família ARIMA. Esses, portanto, serão descritos abaixo.

- **Operador de diferença (∇):** é definidor por $\nabla x_t = x_t - x_{t-1}$;
  - $2^{\underline{a}}$ ordem: $\nabla^2 x_t = \nabla(\nabla x_t) = \nabla(x_t - x_{t-1}) = \nabla x_t - \nabla x_{t-1} = x_t - 2x_{t-1} + x_{t-2}$;
- **Operador de retardo (B):** é definido por $B\, x_t = x_{t-1}$;
  - $2^{\underline{a}}$ ordem: $B^2\, x_t = B(B\, x_t) = B\, x_{t-1} = x_{t-2}$;
  - p-ésima ordem: $B^p\, x_t = x_{t-p}$.

2.4.2 Modelo Autorregressivo Médias Móveis (ARMA)

Um modelo ARMA(p,q) é um processo estacionário do tipo

$$x_t = \alpha + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q},$$

em que $\alpha = (1 - \phi_1 - \cdots - \phi_p)\mu$ é o intercepto, $\mu = E[x_t]$ é a média incondicional da série, $\{\varepsilon_t,\ t = 1, 2, \dots\}$ é um ruído branco com média zero e variância σ^2 e $\phi_1, \dots, \phi_p,\ \theta_1, \dots, \theta_q$ são constantes.

Sem perda de generalidade, pode-se considerar $\alpha = 0$. Assim, ele pode ser reescrito utilizando o operador de retardo, tomando a forma

$$\Phi_p(B) x_t = \Theta_q(B) \varepsilon_t,$$

em que $\Phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ é o polinômio autorregressivo e $\Theta_q(B) = 1 + \theta_1 B + \theta_2 B^2 - \cdots + \theta_q B^q$ é o polinômio de média móvel.

As previsões pontuais para um modelo ARMA(p,q) são dadas por

$$\hat{x}_{n+h|n} = \alpha + \phi_1 x^*_{n+h-1} + \cdots + \phi_p x^*_{n+h-p} + \theta_1 \varepsilon^*_{n+h-1} + \cdots + \theta_q \varepsilon^*_{n+h-q},$$

em que:

$$x^*_j = \begin{cases} x_j, & j \leq n \\ \hat{x}_{j|n}, & j > n \end{cases}$$

e

$$\varepsilon^*_j = \begin{cases} \varepsilon_j, & j \leq n \\ 0, & j > n. \end{cases}$$

Assim, as previsões pontuais dos modelos ARMA podem ser calculadas de forma recursiva, para $h = 1, 2, 3, \dots$.

2.4.3 Modelo ARMA Integrado (ARIMA)

Como citado acima, o modelo ARMA é um processo para séries estacionárias. Apesar disso, grande parte das séries temporais são não-estacionárias. O modelo ARMA Integrado é uma forma de solucionar tal problema.

Uma série temporal não-estacionária x_t tende a se tornar estacionária após algumas diferenças. Portanto, seja $w_t = \nabla^d x_t$ uma série estacionária. O modelo ARIMA(p,d,q) pode ser escrito como um modelo ARMA(p,q) para a série w_t, ou seja,

$$\Phi_p(B)w_t = \Theta_q(B)\varepsilon_t.$$

Como $w_t = \nabla^d x_t = (1-B)^d x_t$, temos que

$$\Phi_p(B)(1-B)^d x_t = \Theta_q(B)\varepsilon_t.$$

As previsões pontuais para o modelo ARIMA são feitas como na seção anterior para a série estacionária w_t, e então estes resultados são utilizados para obter previsões para a série original x_t, utilizando o inverso da diferença.

2.4.4 Modelo ARIMA Sazonal (SARIMA)

O modelo ARIMA Sazonal é o modelo mais completo da família ARIMA que será apresentado neste trabalho. Ele contempla os dois modelos descritos anteriormente, além de incluir a modelagem da componente de sazonalidade, caso a série temporal em estudo a apresente.

Ele é definido por SARIMA(p,d,q)x(P,D,Q) e segue a forma:

$$\Phi_P(B^s)\Phi_p(B)\nabla_s^D\nabla^d x_t = \Theta_Q(B^s)\Theta_q(B)\varepsilon_t.$$

O modelo SARIMA é a forma geral dos modelos da família, ou seja, os modelos apresentados anteriormente estão englobados na forma deste último. Assim, se P, D e Q, os índices referentes à componente de sazonalidade, forem iguais a zero, obtém-se um modelo ARIMA.

A previsão pontual é dada da mesma forma que para os modelos anteriores. Primeiro, aplicam-se as diferenças simples e sazonais, gerando uma série estacionária. Com as previsões para a série diferenciada em mão, por meio de inversões chega-se nas

previsões da série original.

2.4.5 Estimação dos modelos ARIMA

O processo de estimação dos modelos descritos nesta seção são semelhantes. Por ser uma estimação para séries estacionárias, a descrição que se seguirá é aplicável diretamente ao modelo ARMA; para os modelos ARIMA e SARIMA, este processo é aplicável nas séries geradas após a execução das diferenças propostas por estes modelos.

Sejam $x_1, x_2, \ldots, x_n$ os valores observados de uma série temporal. A função densidade de probabilidade conjunta pode ser escrita como

$$f(x_1, \ldots, x_n) = f(x_1)f(x_2|x_1)f(x_3|x_2, x_1)\cdots f(x_n|x_{n-1}, \ldots, x_1) = f(x_1)\prod_{t=2}^{n} f_{t-1}(x_t),$$

em que $f_{t-1}(x_t) = f(x_t|x_{t-1}, \ldots, x_1)$.

Se as k primeiras observações forem consideradas necessárias apenas para inicializar o modelo, a equação acima pode ser escrita da seguinte forma:

$$f(x_1, \ldots, x_n) = f(x_1, \ldots, x_k)\prod_{t=k+1}^{n} f_{t-1}(x_t).$$

A função de verossimilhança condicional, que será utilizada por $f(x_1, \ldots, x_k)$ ser desconhecida, segue abaixo.

$$\mathcal{L}(\beta) = f(x_{k+1}, \ldots, x_n|x_1, \ldots, x_k) = \prod_{t=k+1}^{n} f_{t-1}(x_t),$$

em que $\boldsymbol{\beta}$ é o vetor com todos os parâmetros do modelo.

Como x_t é um modelo ARMA(p,q), já que os modelos ARIMA e SARIMA após as diferenças simples e sazonais se tornam um modelo deste tipo, temos que:

$$\mu_{t|t-1} = E[x_t|x_{t-1}, \ldots, x_1] = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + \theta_1\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q};$$
$$\sigma^2_{t|t-1} = Var[x_t|x_1, \ldots, x_{t-1}] = Var[\varepsilon_t] = \sigma^2.$$

Assim, $x_{t|t-1} \sim N(\mu_{t|t-1}; \sigma^2)$, e portanto

$$\mathcal{L}(\beta, \sigma^2) = \prod_{t=k+1}^{n} f_{t-1}(x_t) = \prod_{t=k+1}^{n} (2\pi\sigma^2)^{-1/2} \, exp\left\{ -\frac{(x_t - \mu_{t|t-1})^2}{2\sigma^2} \right\}$$
$$= (2\pi\sigma^2)^{-(n-k-1)/2} \, exp\left\{ -\frac{\sum_{t=k+1}^{n} (x_t - \mu_{t|t-1})^2}{2\sigma^2} \right\}.$$

Aplicando o logaritmo na função de verossimilhança condicional, temos

$$\ell(\beta, \sigma^2) \propto \frac{-(n-k-1)}{2} \, log(\sigma^2) + -\frac{1}{2\sigma^2} \sum_{t=k+1}^{n} (x_t - \mu_{t|t-1})^2,$$

obtendo assim o estimador de máxima verossimilhança para σ^2, que é

$$\hat{\sigma}^2 = \frac{\sum_{t=k+1}^{n} (x_t - \mu_{t|t-1})^2}{n-k-1}.$$

Substituindo σ^2 por $\hat{\sigma}^2$ em $\ell(\beta, \sigma^2)$, chega-se em

$$\ell(\beta) \propto \frac{-(n-k-1)}{2} \, log \left[\sum_{t=k+1}^{n} (x_t - \mu_{t|t-1})^2 \right].$$

Por fim, o vetor de parâmetros β é estimado visando maximizar $\ell(\beta)$, e assim, minimizar $\sum_{t=k+1}^{n} (x_t - \mu_{t|t-1})^2$.

2.5 TBATS

Uma vez que a série de interesse é formada por observações com frequência diária, isso possibilita a existência de múltiplas sazonalidades com ciclos não inteiros, geralmente chamados de ciclos complexos. Esse tipo de sazonalidade não é levada em consideração por modelos tradicionais como o SARIMA e o Holt-Winters, mas pode ser adequadamente modelados combinando esses modelos com séries harmônicas (HYNDMAN; ATHANASOPOULOS, 2018). O modelo TBATS integra essa funcionalidade e consiste em umas das abordagens mais promissoras para modelar esse tipo de série temporal. Ele foi proposto e descrito em Livera, Hyndman e Snyder (2010).

O nome do modelo se dá pelos seguintes componentes:

- **T:** termos **T**rigonometricos para sazonalidade.
- **B:** transformação de **B**ox-Cox para heterogeneidade;

- **A:** modelo **A**RMA para erros;
- **T:** **T**endência via alisamento exponencial (ETS);
- **S:** **S**azonalidade multipla ou para períodos não inteiros.

A utilização de séries de Fourier na componente de sazonalidade permite que essas sazonalidades complexas sejam englobadas por este modelo.

As suas componentes são dadas por:

$$
\begin{aligned}
&y_t = \text{série original no tempo } t \\
&y_t^{(\omega)} = \begin{cases} (y_t^{(\omega)} - 1)/\omega, & \text{se } \omega \neq 0 \\ \log\, y_t & \text{se } \omega = 0 \end{cases} && \text{(Transformação de Box-Cox)} \\
&y_t^{(\omega)} = \ell_{t-1} + \phi b_{t-1} + \textstyle\sum_{i=1}^{T} s_{t-m_i}^{(i)} + d_t && \text{(Alisamento exponencial)} \\
&\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha d_t \\
&b_t = (1-\phi)b + \phi b_{t-1} + \beta d_t \\
&s_t^{(i)} = \textstyle\sum_{j=1}^{k_i} s_{j,t}^{(i)} && \text{(Sazonalidade por termos trigonométricos)} \\
&s_{j,t}^{(i)} = s_{j,t-1}^{(i)} cos\lambda_j^{(i)} + s_{j,t-1}^{*(i)} sin\lambda_j^{(i)} + \gamma_1^{(i)} d_t \\
&s_{j,t}^{*(i)} = -s_{j,t-1} sin\lambda_j^{(i)} + s_{j,t-1}^{*(i)} cos\lambda_j^{(i)} + \gamma_2^{(i)} d_t \\
&d_t = \textstyle\sum_{i=1}^{p} \phi_i d_{t-1} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t && \text{(Modelo ARMA)}
\end{aligned}
$$

Neste modelo, $\gamma_1^{(i)}$ e $\gamma_2^{(i)}$ representam parâmetros de alisamento, $s_{j,t}^{(i)}$ o *level* estocástico da i-ésima componente sazonal, $s_{j,t}^{*(i)}$ o crescimento estocástico no *level* da i-ésima componente sazonal e k_i o número de harmônicas necessárias na i-ésima componente. Além disso, é necessário estimar $2(k_1 + k_2 + \cdots + k_T)$ valores sazonais iniciais, um número menor do que o necessário em outros modelos. Esta estimação é feita pelo método de máxima verossimilhança.

Para se identificar de forma mais simples um modelo da classe TBATS, este é representado da seguinte forma: TBATS$(\omega, \phi, p, q, \{m_1, k_1\}, \{m_2, k_2\}, \ldots, \{m_T, k_t\})$, em que m_i se refere ao período do i-ésimo ciclo sazonal da série em questão.

2.6 Análise de resíduos

Abaixo, serão definidos os métodos utilizados para realizar a análise residual deste trabalho.

2.6.1 Função de Autocorrelação (FAC)

Sejam $x_1, x_2, \ldots, x_n$ valores observados de um processo estacionário. Assim, temos que $\bar{x} = \left(\sum_{i=1}^{n} x_i\right)/n$ é um estimador da média $\mu = E[x_t]$, já que esta é constante. A autocovariância é constante entre pontos com a mesma defasagem; portanto, a autocovariância amostral de ordem h, dada por

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(x_t - \bar{x})(x_{t+h} - \bar{x}),$$

é um estimador de $\gamma(h) = cov(x_t, x_{t-h})$.

Dessa maneira, a autocorrelação amostral de ordem h é definida como

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}$$

e é um estimador para $\rho(h) = cor(x_t, x_{t-h})$. A função $\hat{\rho}(h)$ é conhecida como **função de autocorrelação**.

Para a análise de resíduos, um correlograma pode ser construído. Considere que $x_1, \ldots, x_n$ foram obtidos de um processo independente e identicamente distribuídos com variância finita. Brockwell e Davis (1991) mostram que

$$\hat{\rho}(h) \sim^{a} N(0,\ 1/n),$$

para qualquer $h = 1, 2, 3, \ldots$.

O correlograma costuma representar, além dos valores de $\hat{\rho}$, o intervalo $\pm 2/\sqrt{n}$, que são os limites onde usualmente o valor de $\hat{\rho}$ é considerado significativamente diferente de 0.

2.6.2 Função de Autocorrelação Parcial (FACP)

Considere um modelo ARMA(p,0). Este é definido por

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \varepsilon_t.$$

O gráfico FACP representa graficamente o valor de ϕ_p, para $p = 1, 2, 3, \ldots$. Na análise de resíduos, ele ajuda a observar alguma dependência com valores passados, o que não é o ideal de ser observado, pela suposição de independência destes dados.

2.6.3 Teste KPSS

O teste KPSS é um teste que averigua se a série de resíduos resultante de determinado modelo é estacionária. As hipóteses deste teste estão logo abaixo.

$$\begin{cases} H_0 : \text{A série é estacionária;} \\ H_1 : \text{A série é não estacionária.} \end{cases}$$

Este teste foi proposto em Kwiatkowski et al. (1992), e encontra-se implementado no RStudio pela função *kpss.test()*, do pacote tseries.

2.6.4 Teste de Ljung-Box

Este é um teste de independência proposto e descrito em Ljung e Box (1978). Seu objetivo é concluir se os resíduos de determinado modelo são independentes entre si. As hipóteses são definidas abaixo.

$$\begin{cases} H_0 : \rho(1) = \rho(2) = \cdots = \rho(m) = 0 \\ H_1 : \text{Pelo menos um destes valores é diferente de zero.} \end{cases}$$

A estatística do teste é

$$Q = n(n+2)\sum_{h=1}^{m} \hat{\rho}(h)^2/(n-h).$$

Normalmente, define-se $m \approx 15$; neste trabalho, $m = 15$ em todas as aplicações. Sob a hipótese nula, pode-se demonstrar que $Q \sim^a \chi^2_m$. Assim, rejeita-se H_0 quando $Q > \chi^2_{m,1-\alpha}$.

2.6.5 Teste de Shapiro-Wilk

O teste de Shapiro-Wilk é utilizado para observar se um conjunto de dados é normalmente distribuído. Sejam $x_1, x_2, \ldots, x_n$ uma amostra aleatória e $F(x)$ sua distribuição desconhecida. As hipóteses deste teste são descritas abaixo.

$$\begin{cases} H_0 : F(x) \text{ é uma distribuição normal com parâmetros não especificados;} \\ H_1 : F(x) \text{ não é uma distribuição normal.} \end{cases}$$

Sejam os postos das observações amostrais definidos como

$$x^{(1)} \leq x^{(2)} \leq \cdots \leq x^{(n)}.$$

A estatística do teste então é dada por

$$T_3 = \frac{1}{D}\left[\sum_{i=1}^{k} a_i (x^{(n-i+1)} - x^{(i)})\right]^2,$$

em que:

- $D = \sum_{i=1}^{n}(x_i - \bar{x})^2$;
- $a_1, \ldots, a_k$ são coeficientes tabelados;
- $k \sim n/2$.

Este teste é descrito em Conover (1999). Os coeficientes $a_1, \ldots, a_k$ e a distribuição da estatística do teste estão tabelados neste mesmo livro. Entretanto, este método é definido para amostras pequenas; a função *shapiro.test()*, utilizada neste trabalho, realiza uma adaptação deste teste, em que grandes amostras podem ser testadas. Esta adaptação está descrita em Royston (1982).

2.7 Séries Temporais Hierárquicas

As séries temporais, dependendo de sua natureza, podem ser segmentada em diversas séries menores, em que os segmentos criados são mutualmente exclusivos. Também podem ser agrupadas e formar uma seleção menor de séries para serem analisadas. Esta estruturação é o que chama-se de hierarquia em séries temporais, e é descrita em Hyndman e Athanasopoulos (2018).

Isso mostra as diversas formas com que pode-se trabalhar em um mesmo conjunto de séries temporais. Tome, por exemplo, uma série diária das vendas de uma loja de materiais esportivos; pode-se dividir a série do total de vendas da loja pela marca dos produtos que esta loja comercializa; se o objetivo são informações mais específicas, pode-se criar mais níveis, agora com relação a qual tipo de produto de cada marca as vendas estão se referindo, sejam eles tênis, roupas, acessórios, entre outros.

Segue, abaixo, um grafo exemplificando o esquema hierárquico.

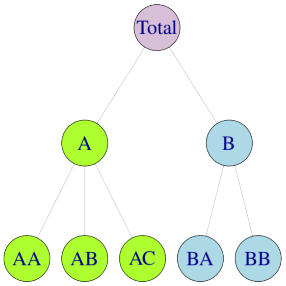

Figura 2: Exemplo de estrutura hierárquica para séries temporais

FONTE: Extraído de Hyndman e Athanasopoulos (2018)

Considerando a estrutura da Figura 2, seja y_t como o valor da série Total no tempo t, e $y_{j,t}$ como o valor do nó j no tempo t. Assim, temos que

$$y_t = y_{AA,t} + y_{AB,t} + y_{AC,t} + y_{BA,t} + y_{BB,t}.$$

Além disso, temos que $y_{A,t} = y_{AA,t} + y_{AB,t} + y_{AC,t}$ e $y_{B,t} = y_{BA,t} + y_{BB,t}$. Dessa forma, pode-se definir $y_t = y_{A,t} + y_{B,t}$.

Neste trabalho em questão, por se tratar de um grande número de séries e de estas serem bem detalhadas, níveis mais ao topo da estrutura hierárquica serão utilizados, ou seja, as séries serão agrupadas, para observar como este tipo de estrutura se comporta no quesito de previsões em um cenário competitivo.

2.7.1 A abordagem *Top-down*

A abordagem *Top-down* para séries hierárquicas consiste em realizar previsões para a série agregada em níveis mais altos e depois destrinchar essas previsões para as séries que formaram tais grupos.

Para exemplificar, será utilizado o nível mais alto, em que todas as séries temporais foram somadas, sendo chamada assim de série total. Portanto, sejam $p_1, \ldots, p_m$ as proporções que cada um dos m níveis desta série total representem das previsões realizadas. Assim, para o exemplo de estrutura na Figura 2, tem-se que

$$\tilde{y}_{AA,t} = p_1\hat{y}_t, \;\; \tilde{y}_{AB,t} = p_2\hat{y}_t, \;\; \tilde{y}_{AC,t} = p_3\hat{y}_t, \;\; \tilde{y}_{BA,t} = p_4\hat{y}_t \text{ e } \tilde{y}_{BB,t} = p_5\hat{y}_t.$$

Para definir o valor dos pesos, pode-se utilizar a proporção histórica média (global

ou recente). Esta é dada por

$$p_j = \frac{1}{T} \sum_{t=1}^{T} \frac{y_{j,t}}{y_t},$$

em que $j = 1, \ldots, m$ e T denota o número de observações dentro da janela considerada. Portanto, p_j representa a proporção histórica média que a série j representa da série total no período observado.

No caso em que alguns níveis são criados, a proporção histórica média será aplicada individualmente para cada um dos grupos, já que estes são mutualmente exclusivos, como dito anteriormente.

2.8 Métrica de avaliação das previsões

A métrica que será utilizada para avaliar as previsões obtidas será a **SMAPE**, sigla para Symmetric Mean Absolute Percentage Error. Esta é dada por

$$SMAPE = \frac{100\%}{n} \sum_{t=1}^{n} \frac{|F_t - A_t|}{(|A_t| + |F_t|)/2},$$

em que:

- t: tempo da previsão;
- n: número de previsões realizadas;
- F_t: valor previsto para o tempo t;
- A_t: valor real no tempo t.

3 Resultados

3.1 As séries temporais em estudo

Considerando que o número de séries temporais em estudo é muito grande, a aplicação individual de modelos mais tradicionais se torna inviável, principalmente pela demanda computacional que esta tarefa exigiria. A natureza similar destas séries temporais permite que estas formem estruturas hierárquicas, ou seja, sejam agrupadas em níveis maiores para que sejam analisadas de forma mais simples.

Neste trabalho, dois cenários serão estudados: a junção de todas as séries temporais do conjunto de dados em apenas um nível, gerado pela soma de todas as observações presentes, que corresponderá aos acessos totais diários ao Wikipedia; o outro cenário consiste em agrupar as séries temporais de acordo com o projeto a qual cada artigo faz parte, tendo assim nove séries temporais. De forma simplificada, o projeto é uma parte do link do artigo, e pode fazer parte de páginas comuns do Wikipedia ou de páginas secundárias, que possuem outro domínio. Na Tabela 3, eles estão listados e detalhados.

Projeto	Descrição
de.wikipedia.org	Artigos em alemão
en.wikipedia.org	Artigos em inglês
es.wikipedia.org	Artigos em espanhol
fr.wikipedia.org	Artigos em francês
ja.wikipedia.org	Artigos em japonês
ru.wikipedia.org	Artigos em russo
zh.wikipedia.org	Artigos na família de línguas chinesas
commons.wikimedia.org	Repositório central de imagens
www.mediawiki.org	Páginas do software livre que é utilizado pelo Wikipedia

Tabela 3: Descrição dos projetos presentes no conjunto de dados

Com estes agrupamentos, é viável criar representações gráficas para estas séries temporais e observar seu comportamento histórico. Estas análises seguem nas sub-seções abaixo.

3.1.1 Acessos diários a páginas do Wikipedia (Série total)

A Figura 3 representa o primeiro cenário descrito na seção anterior: o agrupamento em apenas um nível mais alto, contendo todos os artigos presentes no conjunto de dados. Nela, observa-se a presença de tendência, primeiramente de forma crescente, mas se transformando em decrescente na parte final do período analisado. O canto inferior direito do gráfico parece indicar o início de uma nova tendência positiva, mas a falta de continuidade dos dados não permite afirmar com toda certeza se este é o caso. Além disso, o fato de a série ser diária torna a presença de sazonalidade provável. O gráfico visualmente parece apresentar ciclos sazonais e, pela quantidade de dados que estão disponíveis, pode-se considerar tanto a sazonalidade semanal quanto a mensal.

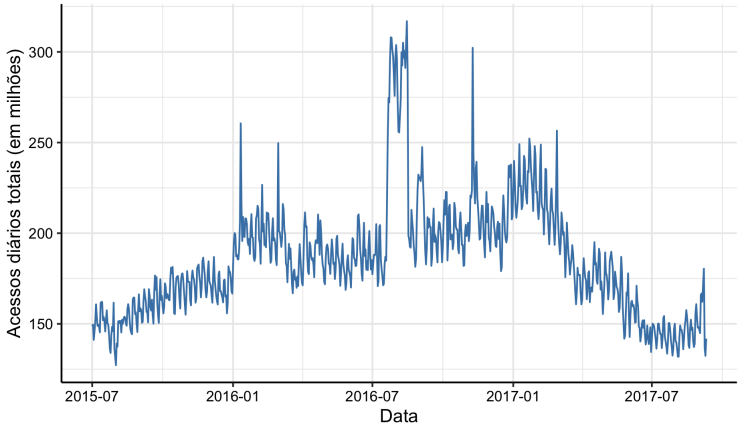

Figura 3: Número de acessos diários a artigos do Wikipedia

3.1.2 Acessos diários por projeto

Cada projeto possui um comportamento singular em relação aos acessos totais. A série referente a artigos em espanhol, por exemplo, tem um formato muito diferente do restante dos projetos; não parece apresentar tendência, apenas ciclos sazonais, como pode-se observar na Figura 4.

Espera-se que os projetos sirvam como estratos, ou seja, as séries que formam os projetos tenham um comportamento homogêneo dentro de seus grupos, mas heterogêneo em comparação com os outros oito projetos.

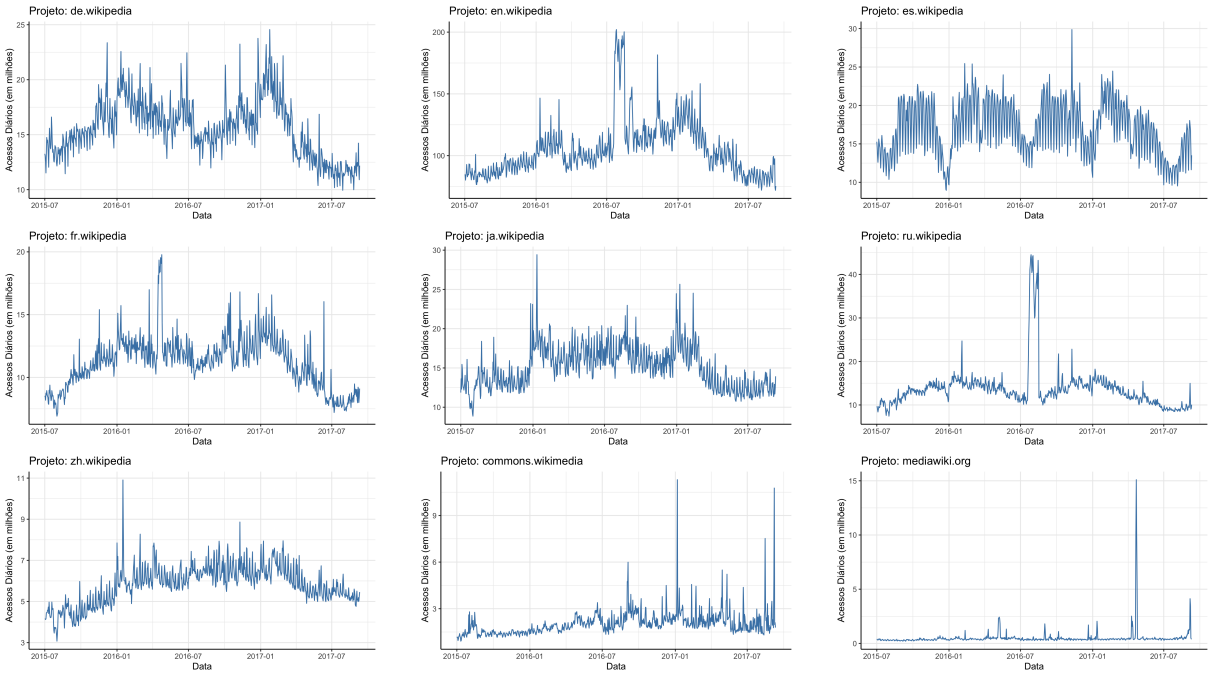

Figura 4: Acessos diários referentes a cada projeto

3.2 Análise inicial de desempenho dos modelos

Buscando um entendimento inicial do comportamento preditivo dos modelos selecionados para fazer parte do estudo, a utilização da validação cruzada pelo método de janelas deslizantes torna-se muito útil.

Para esta estapa, será considerada a série temporal descrita na Seção 3.1.1. Considerando as características descritas na mesma, o modelo TBATS foi selecionado para modelar e realizar as previsões necessárias, já que este consegue englobar de forma satisfatória a sazonalidade múltipla e complexa presente nesta série temporal. Apesar disso, a tendência mais provável de estar presente neste conjunto de dados é a semanal. Assim, um modelo SARIMA também pode ser aplicado para servir de comparação, já que este é um modelo bem mais simples que o TBATS.

Uma janela de treinamento inicial composta de 500 dias e um horizonte de previsão de 64 dias foram utilizados. A qualidade do ajuste foi medida utilizando o Erro Absoluto Médio. Os valores desta métrica para cada horizonte de previsão e modelo foram representados na Figura 5.

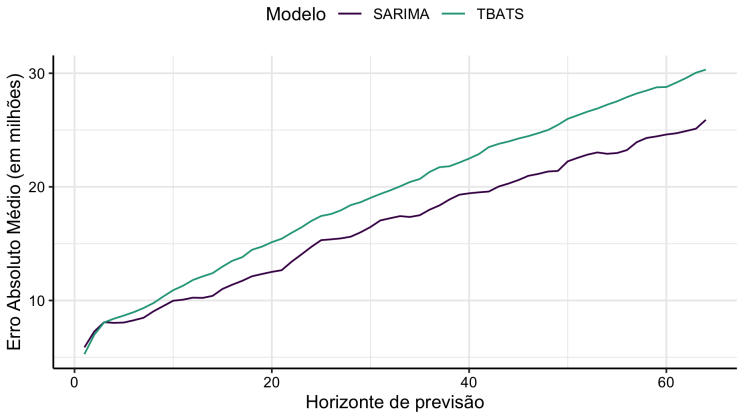

Figura 5: Erro Absoluto Médio para diferentes horizontes de previsão dos modelos TBATS e SARIMA

O gráfico mostra que o modelo SARIMA apresenta um Erro Absoluto Médio menor que os obtidos pelo modelo TBATS para praticamente todos os horizontes de previsão testados. Isso indica que este modelo tende a ter previsões melhores para a série total.

Vale ressaltar que o MAE não é a métrica que será utilizada para avaliar as previsões finais deste trabalho e esta análise foi realizada baseada em previsões das série agrupadas, e não separando individualmente para cada série em estudo, portanto esta avaliação é considerada apenas uma prévia e um indicativo da qualidade de desempenho destes modelos.

3.3 Análise preditiva sobre as séries agrupadas em acessos diários totais

Nesta seção, os modelos começaram a ser aplicados para gerar previsões para as séries temporais em estudo. Primeiramente, será estudado o caso mais geral, em que todas as 145.063 séries são agrupadas em um mesmo nível hierárquico.

Como referido na Seção 3.2, as previsões serão obtidas tanto por um modelo TBATS quanto por um modelo da família ARIMA. Dessa forma, os modelos específicos ajustados para representar a série de acessos totais foram o TBATS(0, {0,0}, 0,85, {<7, 3>, <30.44, 5>}) e o SARIMA(1,0,0)x$(2,1,0)_{[7]}$. Estes foram obtidos de forma automática, apenas especificando os períodos sazonais observados na análise gráfica da Seção 3.1.1, por meio das funções *auto.arima()* e *tbats()*, do pacote *forecast*. O ajuste em relação à série histórica deste dois casos segue na figura abaixo.

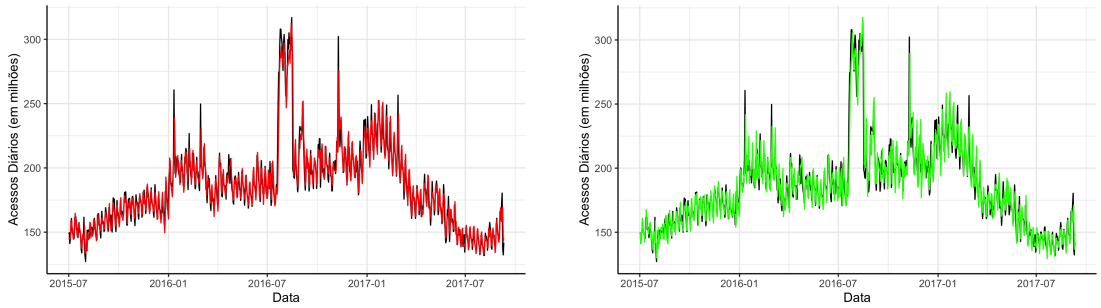

Figura 6: Ajuste dos modelos TBATS e SARIMA, respectivamente, sobre os artigos agregados em um nível

Visualmente, os dois modelos se ajustaram bem aos valores passados da série. Alguns padrões do comportamento apresentado parecem ter sido captados pelos modelos. Além da observação gráfica, a análise residual pode trazer informações relevantes sobre esses ajustes.

A Figura 7 traz indicativos de que os resíduos gerados pelo modelo TBATS formam uma série estacionária e que são independentes, como indicado pelos gráficos da FAC e da FACP. A normalidade não parece ser atendida, já que estes dados apresentam caldas bem pesadas. Apesar disso, esta será considerada para realizar as estimativas dos parâmetros dos modelos, já que o objetivo final deste trabalho são previsões pontuais. Para confirmar estes indicativos, testes estatísticos foram aplicados. O teste KPSS de estacionariedade não rejeita a hipótese destes dados serem estacionários; o teste de Ljung-Box para a independência dos resíduos também encontrou um alto p-valor, ou seja, a hipótese nula de independência também não é rejeitada; já a normalidade é rejeitada, como o gráfico de quantis indicou. Estes resultados estão descritos na Tabela 4.

Em relação aos resíduos gerados pelo modelo SARIMA, o primeiro gráfico da Figura 8 parece indicar a estacionariedade destes dados. Esta parece ser a única suposição que é atendida, já que o gráfico da FAC parece indicar uma correlação destes dados, principalmente nos lags sazonais. O mesmo comportamento é visto pela FACP, o que tende a indicar que estes dados não são independentes. A normalidade não parece ser atendida também, pelas caldas pesadas observadas na figura. Os mesmos testes foram realizados, confirmando os resultados descritos com a análise gráfica. Os resultados destes testes estão representados na Tabela 5.

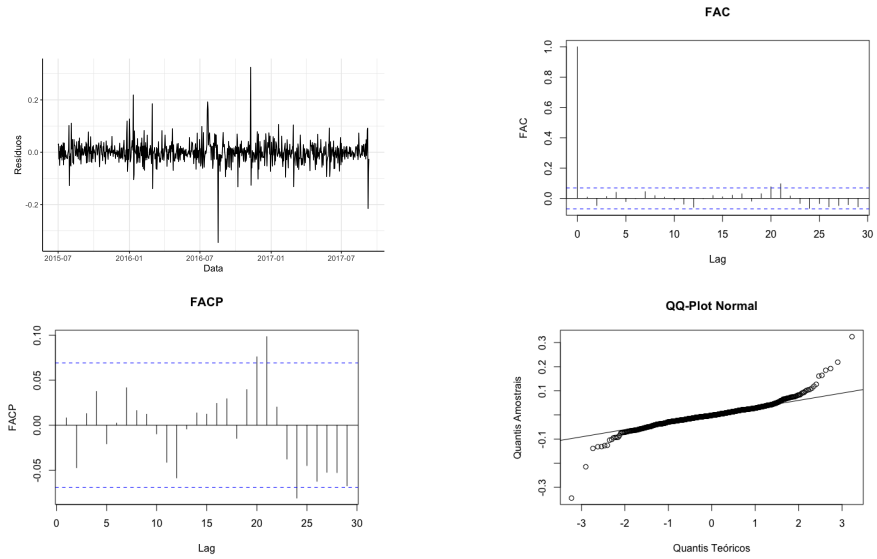

Figura 7: Análise gráfica dos resíduos do modelo TBATS

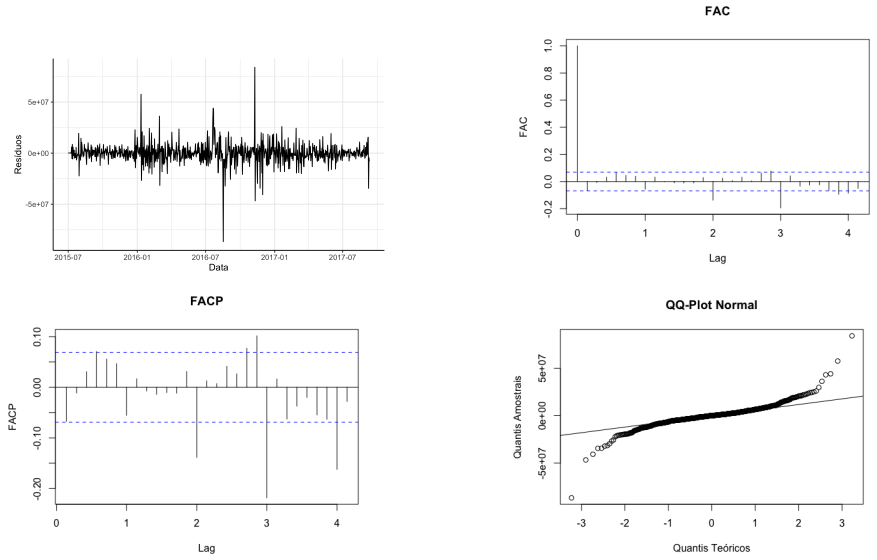

Figura 8: Análise gráfica dos resíduos do modelo SARIMA

Teste	p-valor
KPSS	$> 0{,}1$
Ljung-Box	$> 0{,}5$
Shapiro-Wilk	$< 0{,}001$

Tabela 4: Testes da análise de resíduos - TBATS

Teste	p-valor
KPSS	$> 0{,}1$
Ljung-Box	$< 0{,}001$
Shapiro-Wilk	$< 0{,}001$

Tabela 5: Testes da análise de resíduos - SARIMA

Ainda que importante para avaliar a adequação dos modelos, o ajuste histórico não é algo determinante para a qualidade das previsões pontuais, já que a captura de comportamentos específicos nos dados passados não significa que as especificidades futuras

conseguirão ser replicadas.

O objetivo final deste trabalho engloba a previsão dos acessos de 62 dias futuros, mas foi necessário primeiramente gerar a previsão para 64 dias, já que existe um intervalo de dois dias entre a data final dos acessos e a primeira data escolhida para se obter previsões. Estas previsões extras foram descartadas, já que simplesmente não fazem parte do objetivo final e só foram necessárias para obter as que serão de fato utilizadas.

Com a utilização de uma estrutura hierárquica para simplificar a modelagem destes dados, as previsões obtidas neste primeiro momento não são o objetivo final do estudo. Estas precisam ser divididas para cada uma das 145.063 páginas analisadas. A proporção histórica média será utilizada nesta etapa. Esta técnica, originalmente, consideraria o período de tempo total do estudo para obter qual a proporção que cada artigo representa nos dados históricos do total de acessos. Devido a presença de valores nulos no banco de dados e da mudança de comportamento que as séries individuais podem ter em um período tão longo, foram considerados apenas os últimos 60 dias dos dados históricos para calcular as proporções individuais, para que estas reflitam um comportamento mais recente destas séries.

Após a partição das previsões, estas foram enviadas para a avaliação automática gerada pelo site da competição. Nesta etapa, a SMAPE é calculada utilizando as mais de 8,9 milhões de previsões que foram formadas. Os resultados desta avaliação podem ser observados na Tabela 6.

Modelo	SMAPE
TBATS	47,862
SARIMA	48,374

Tabela 6: SMAPE das previsões das séries agrupadas em apenas um nível

Ambos os modelos tiveram um desempenho semelhante com relação às previsões. O modelo TBATS selecionado conseguiu um valor de SMAPE apenas 0,5 menor que o obtido pelo modelo SARIMA, apesar de a análise inicial utilizando janelas deslizantes ter indicado um melhor desempenho do modelo SARIMA.

3.4 Análise preditiva sobre as séries agrupadas por projeto

Na seção anterior, as séries foram somadas para formar uma série de acessos totais e, dessa forma, tornar o processo de previsão mais simples. Apesar disso, cada série

temporal presente no banco de dados tem suas características individuais, e agrupá-las em apenas um conjunto pode não ser a melhor maneira de se fazer isto.

Pode-se pensar que exista alguma relação entre as séries temporais que fazem parte dos mesmos projetos, como um estrato. Novamente, a estrutura hierárquica será utilizada, mas em um nível inferior em relação ao utilizado na Seção 3.3. Serão, portanto, utilizadas as séries descritas na Seção 3.1.2.

Outra vez, os modelos TBATS e SARIMA foram aplicados, considerando os mesmos ciclos sazonais utilizados na Seção 3.3. Portanto, nove modelos TBATS e nove modelos SARIMA foram gerados ao final deste processo. Assim, as previsões foram geradas para cada um dos modelos e novamente desagrupadas utilizando a proporção histórica média dos últimos 60 dias. Como cada página presente no banco de dados pertence a apenas um dos projetos, a proporção histórica de cada artigo foi calculada em relação ao projeto que este faz parte, e não em relação a proporção que estes representam para o total de acessos geral.

Assim, as previsões foram enviadas para avaliação, o que resultou em valores da SMAPE de 49,120 e 50,487 para as previsões utilizando-se o TBATS e o SARIMA, respectivamente. Este resultado é descrito na Tabela 7.

Modelo	SMAPE
TBATS	49,120
SARIMA	50,487

Tabela 7: Métrica de avaliação dos modelos criados utilizando os projetos

O modelo TBATS mais uma vez trouxe previsões um pouco mais próximas dos valores reais do que as obtidas pelo modelo SARIMA. Com a utilização das informações em relação aos projetos das páginas, esta diferença entre os dois modelos foi até um pouco maior que a observada quando este material era ignorado. Apesar disso, chama atenção o fato de este dado adicional não ter trazido melhora para o processo de previsão; vale até ressaltar uma pequena piora, já que os valores da SMAPE foram maiores. Isso pode indicar que, apesar de cada projeto possuir um comportamento singular, como observado na Figura 4, talvez este não generalize bem o comportamento das séries que os compõem.

3.5 Comparação com resultados da competição

Como explicado anteriormente, os dados deste trabalho se referem a uma competição de previsões (*Web Traffic Time Series Forecasting*). Apesar de ter ocorrido em 2017, ainda pode-se avaliar previsões realizadas após seu encerramento. Além disso, é possível ver os resultados dos participantes que enviaram suas previsões enquanto esta ainda estava ativa.

Modelo	SMAPE
TBATS - Nível total	47,862
ARIMA - Nível total	48,374
TBATS - Níveis por projeto	49,120
ARIMA - Níveis por projeto	50,487
Modelo do 1º colocado	35,481
Modelo do 2º colocado	36,785
Modelo do 3º colocado	36,853
⋮	⋮
Modelo do 109º colocado	41,197

Tabela 8: Resumo dos resultados obtidos e dos medalhistas da competição

O Kaggle, site que sediou a competição, possui um sistema de premiação para os participantes. São três tipos de medalhas, sendo elas de ouro, prata e bronze. Como a competição teve a participação de 1.095 times, a regra da distribuição de medalhas é a seguinte: o time que ficou no Top 12 é premiado com a medalha de ouro; se ficou no Top 50, recebe a de prata; por fim, se finalizou no Top 100, medalha de bronze. No final, 109 times foram premiados, provavelmente por empates que podem ter ocorrido.

Os modelos criados neste trabalho tiveram acurácia inferior que os obtidos pelos medalhistas, como pode-se observar na Tabela 8. Isso de certa forma era esperado, já que nenhuma covariável foi utilizada, a estrutura hierárquica foi aplicada em todos os modelos e também não ouve nenhum teste de outra forma mais inteligente de desagrupar as previsões, já que tanto a escolha de utilizar a proporção histórica média e a de se utilizar os últimos 60 dias para tal foram feitas de forma arbitrária. Considerando isso, os resultados foram positivos para o objetivo deste trabalho.

3.5.1 Modelos dos competidores e próximos passos

Ao final das competições de modelagem e previsão, é muito comum os participantes que obtiveram bons resultados compartilharem como estes foram obtidos. Na competição em estudo não foi diferente.

Entre os participantes que se classificaram nas dez primeiras colocações e que publicaram seus modelos, foi observada, majoritariamente, a utilização de modelos de redes neurais, tanto recorrentes quanto convolucionais. Além disso, o algoritmo XGBoost foi bem mencionado nas discussões encontradas no fórum da competição, podendo ser considerado como uma opção com bons resultados. Observando exclusivamente a solução construída pelo vencedor da competição, as redes neurais recorrentes foram utilizadas, em um modelo *seq2seq*; os ciclos sazonais trimestrais e anuais foram utilizados, ciclos estes diferentes dos utilizados neste trabalho. Além disso, covariáveis foram utilizadas, como a popularidade da página e o país a qual esta pertence.

Assim, encontram-se ideias de como evoluir o trabalho aqui descrito, seja adaptando os modelos já estruturados anteriormente para englobar algumas dessas ideias, como a utilização de outros ciclos sazonais, seja estudando e implementando novas classes de modelos preditivos, como os modelos de *Machine Learning* que foram citados acima.

4 Conclusão

O objetivo deste trabalho era realizar previsões sobre o número de acessos de 62 dias para páginas do Wikipedia, utilizando modelos mais tradicionais da área de análise preditiva de séries temporais, e avaliá-las por meio da Symmetric Mean Absolute Percentage Error (SMAPE).

Apesar de em uma análise inicial, utilizando a validação cruzada, os modelos da família ARIMA possuírem uma acurácia maior que o modelo TBATS, quando estes foram utilizados para realizar a previsão final, os resultados obtidos foram o oposto, ou seja, os modelos TBATS conseguiram prever melhor os dados nas duas estruturas hierárquicas estudadas. Esta diferença pode-se dar pelo fato da análise inicial ter utilizado apenas os dados agrupados, enquanto a avaliação final das previsões é realizada após o desmembramento das previsões para as séries individuais, ou também pela diferente métrica utilizada para avaliar os dois casos.

Com relação às previsões finais, o fato de que resultados melhores foram obtidos quando todas as séries foram agrupadas em apenas um nível hierárquico chamou atenção, já que se esperava que uma análise mais detalhada e uma separação por projetos pudesse trazer uma melhora para este estudo.

Por fim, em comparação com os demais times participantes da competição, nota-se que os modelos utilizados neste trabalho não foram tão competitivos, talvez pela escolha de estruturas hierárquicas que simplifiquem e generalizem demais as individualidades de cada uma das séries em estudo, além da escolha arbitrária para o desagrupamento das previsões. Ainda assim, considerando o objetivo deste trabalho, os resultados obtidos foram positivos.

Referências

BROCKWELL, P. J.; DAVIS, R. A. *Time Series: Theory and Methods*. 2ª. ed. Nova Iorque: Springer-Verlag, 1991.

CONOVER, W. J. *Practical Nonparametric Statistics*. 3ª. ed. [S.l.]: John Wiley and Sons, Inc, 1999.

HYNDMAN, R. J.; ATHANASOPOULOS, G. *Forecasting: Principles and Practice*. Melbourne, Australia: OTexts, 2018.

HYNDMAN, R. J.; ATHANASOPOULOS, G. *Forecasting: Principles and Practice*. 3ª. ed. Melbourne, Australia: OTexts, 2021.

KWIATKOWSKI, D. et al. Testing the null hypothesis of stacionarity against the alternative of a unit root. Journal of Econometrics, 1992.

LIVERA, A. M. D.; HYNDMAN, R. J.; SNYDER, R. D. Forecasting time series with complex seasonal patterns using exponential smoothing. Journal of the American Statistical Association, 2010.

LJUNG, G. M.; BOX, G. E. P. On a measure of lack of fit in time series models. Biometrika, 1978.

MORETTIN, P. A.; TOLOI, C. M. C. *Análise de Séries Temporais*. [S.l.]: Blucher, 2019.

ROYSTON, J. P. An extension of Shapiro and Wilk's w test for normality to large samples. *Journal of the Royal Statistical Society*, 1982.