Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Um Estudo Sobre Modelos Preditivos para o Número de Acidentes em Rodovias Federais

Igor de Oliveira Barros Faluhelyi

Orientador: Prof. Dr. José Augusto Fiorucci

Brasília
Agosto 2023

Igor de Oliveira Barros Faluhelyi

Um Estudo Sobre Modelos Preditivos para o Número de Acidentes em Rodovias Federais

Orientador:
Prof. Dr. José Augusto Fiorucci

Relatório apresentado para o Departamento de Estatística, Instituto de Ciências Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Brasília
2023

Este trabalho é dedicado às pessoas esforçadas que perseveram em seus objetivos, superando obstáculos e buscando constantemente o melhor de si.

“If you find that you’re spending almost all your time on theory,
start turning some attention to practical things;
it will improve your theories.
If you find that you’re spending almost all your time on practice,
start turning some attention to theoretical things;
it will improve your practice.”
(Donald Knuth)

Resumo

A partir dos dados abertos da Polícia Rodoviária Federal, em que são documentados, entre outros, acidentes em rodovias Federais, este trabalho tem por objetivo avançar na modelagem em séries temporais, afim de trazer previsões para a série diária do número de acidentes em diferentes níveis de agregação dentro de uma estrutura hierárquica e agrupada dos dados.

Como metodologia, são colocados os modelos sNAIVE, ARIMA ou ARIMA sazonal, a Regressão Dinâmica e sua variação, conhecida como Regressão Dinâmica Harmônica e o modelo TBATS, além de métodos para desagregar previsões e fazer validação cruzada, afim de fazer seleção de modelos.

Os resultados apontam para melhores previsões pelos modelos que captam as múltiplas sazonalidades da série diária do número de acidentes, isto é, a Regressão dinâmica Harmônica e o TBATS. Foram adicionadas variáveis explanatórias indicadoras ao modelo ARIMA e isso melhorou a capacidade preditiva do modelo nessa situação.

Como conclusão, pode-se citar o êxito to trabalho em cumprir com seus objetivos. Foram entregues previsões para a série do número de acidentes em rodovias Federais para cada rodovia abordada no banco (208 séries), para cada Estado brasileiro (27 séries), e, ainda, para cada região no Brasil (5 séries) - contabilizando 240 séries temporais. Com isso, pode-se elencar rodovias, ou Estados, destaques quanto ao número previsto de acidentes, tornando possível uma abordagem de forma preventiva (não somente remediativa) no âmbito de políticas públicas afim de diminuir os acidentes no Brasil.

Palavras-chave: Análise de Séries Temporais. Previsão. Séries temporais hierárquicas. Séries temporais agrupadas.

Abstract

This undergraduate final project focuses on analyzing open data provided by the PRF from Brazil, specifically regarding accidents on Federal highways. The primary aim of this study is to apply and compare existing time series models to improve the prediction accuracy of daily accident rates. The models are applied to different levels of data aggregation within a hierarchical and grouped structure.

The methodologies employed in this project encompass various well-established models, including sNAIVE, ARIMA, seasonal ARIMA, Dynamic Regression, Harmonic Dynamic Regression, and the TBATS model. By comparing their performance, the study identifies the most effective models for accurately predicting the daily number of accidents.

The results indicate that models capable of capturing the multiple seasonal patterns inherent in the daily accident rates, such as Harmonic Dynamic Regression and TBATS, outperform the other models. Moreover, the inclusion of explanatory variables in the ARIMA model significantly improves its predictive capabilities in this specific context.

This research contributes to the understanding of accident patterns on Federal highways and provides valuable insights for enhancing accident rate predictions. The findings have practical implications for traffic management and public safety, enabling authorities to allocate resources more efficiently and reduce the occurrence of accidents.

Keywords: Time Series Analysis. Forecasting. Hierarchical time series. Grouped time series.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AR	Modelo Auto Regressivo	19
ARIMA	Modelo misto integrado	26
BR	Rodovia Federal	37
DHR	Regressão dinâmica harmônica	29
DR	Regressão dinâmica	29
EUA	Estados Unidos da América	15
LAI	Lei de Acesso à Informação	15
MA	Modelo de Médias Móveis	19
MOFC	*Makridakis Open Forecasting Center*	15
PRF	Polícia Rodoviária Federal	15
RAE	*Regression with ARIMA errors*	29
SARIMA	Modelo ARIMA Sazonal	26
UF	Unidade da Federação	16

Lista de símbolos

Sumário

1 Introdução

Com o avanço contínuo da tecnologia e dos sistemas de informação, a coleta e análise de dados tornaram-se essenciais para aprimorar a eficiência na prestação de serviços públicos à sociedade. No contexto da Polícia Rodoviária Federal (PRF), apresentado em seu observatório de dados[1], o uso de dados na formulação de políticas públicas tem se mostrado indispensável para promover uma gestão eficiente.

Ademais, desde 2007, a PRF tem disponibilizado à sociedade um conjunto abrangente de dados abertos relacionados a acidentes e infrações de trânsito, consoante à elaboração do plano de dados abertos da PRF[2]. Essa iniciativa está em conformidade com a Lei de Acesso à Informação (LAI) e os compromissos assumidos pelo Brasil no âmbito do Plano de Ação Nacional de Governo Aberto.

São utilizados os dados abertos da PRF com intuito de observar o número de acidentes em rodovias no Brasil ao longo do tempo e aprofundar o estudo em Análise de Séries Temporais, tendo em vista modelagem para previsão de séries inseridas em estruturas hierárquicas e agrupadas, bem como métodos para agregar ou desagregar tais previsões em seus diferentes níveis.

A abordagem para a série do número de acidentes será respeitando, em um primeiro momento, a estrutura hierárquica de região para Unidade da Federação, isto é, o acidentes ocorre em uma UF que pertence unicamente à alguma região do Brasil, e, em um segundo momento, a estrutura agrupada desde a Unidade da Federação até a Rodovia, em que o acidente ocorre, por definição, em alguma parte da rodovia que está dentro (geograficamente) de alguma UF brasileira.

A construção de agrupamentos e hierarquias em séries temporais será abordada com mais detalhes no Capítulo 3. Neste momento, é feita uma contextualização por meio do texto destacado abaixo:

> **Séries temporais hierárquicas e agrupadas**
>
> De acordo com Wickramasuriya, Athanasopoulos e Rob J Hyndman (2015, seção 1), grande parte das séries temporais podem ser agregadas ou desagregadas com respeito à uma estrutura de restrições hierárquicas quanto à localidade ou categorias. Por exemplo, as vendas de uma empresa multinacional podem ser desagregadas em uma hierarquia geográfica de Estados, regiões e lojas. Muitas vezes o contexto da empresa exige previsões

[1] https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos/observatorio-de-dados-da-prf

[2] https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos/dados-abertos-da-prf

de vendas totais, nacionais, regionais ou até mesmo vendas para uma loja específica, e essas previsões devem ser somadas adequadamente em toda a hierarquia.

A empresa também pode produzir produtos dentro de uma hierarquia, isto é, eles são divididos em grupos e subgrupos de produtos. Nesse caso, a série de vendas do produto em cada hierarquia, própria, geralmente resulta em uma grande coleção de séries temporais individuais. Quando essas séries são observadas no nível de loja ou Estado ou região é preciso restringir a agregação visto que o mesmo produto pode ser vendido em diferentes lojas. Uma grande coleção de séries temporais com restrições de agregação é chamada de série temporal agrupada (HYNDMAN, R. J.; LEE; WANG, 2016).

Em 2020, *The Makridakis Open Forecasting Center (MOFC) at the University of Nicosia's Institute for the Future (IFF)*[3] concluiu com sucesso a competição M5[4]. Competição, essa, de previsão de séries temporais que compõe a sequência de competições chamadas *Makridakis (or M) Competitions*, que existem desde 1982 com a realização da M1. Em particular, a *M5 Forecasting - Accuracy* foi realizada através da plataforma Kaggle[5] e atraiu quase 6000 participantes ao redor do mundo com premiação total de 50 mil dólares.

A COMPETIÇÃO M5

Organizada pelo MOFC, ela traz como objetivo avançar a teoria e prática de previsão, identificando modelos que fornecem as previsões para cada uma das 43204 séries temporais da competição. O cerne da M5 foi fazer com que o competidor fizesse, o mais precisamente possível, previsões para as vendas diárias de unidades de produtos de varejo específicos que são vendidos, pela Walmart nos EUA, dentro de uma estrutura **hierárquica** de Estados, lojas e departamentos. Isso, pois, a venda do produto ocorre em uma loja, presente em algum Estado (nos EUA), e dentro da loja o produto é agrupado em departamentos.

Assim como observado na mencionada competição, a motivação aqui reside em realizar previsões para séries diárias, inseridas em uma estrutura hierárquica, utilizando os conjuntos de dados disponibilizados pela PRF. Neste contexto, este trabalho considera a possibilidade de impor restrições na série de acidentes, levando em conta a região, o Estado ou a rodovia onde o acidente ocorreu. Paralelamente, no que diz respeito às séries de vendas na competição M5, as restrições podem ser aplicadas às lojas e aos departamentos associados à comercialização do produto.

O propósito deste documento é avançar na utilização da modelagem de séries temporais, utilizando os dados abertos fornecidos pela PRF, para fins de previsão. Especificamente,

[3] https://www.unic.ac.cy/unic-launches-makridakis-open-forecasting-center/
[4] https://mofc.unic.ac.cy/m5-competition/
[5] https://www.kaggle.com/c/m5-forecasting-accuracy/

o enfoque recai sobre séries diárias que estão inseridas em uma estrutura hierárquica ou agrupada.

Neste trabalho, o número de acidentes em rodovias no Brasil será analisado dentro de uma hierarquia que engloba regiões e Estados. Isto é, os acidentes ocorrem nos Estados, que por sua vez estão vinculados à regiões específicas do Brasil. Além disso, também será considerada uma estrutura agrupada que engloba Unidades da Federação e rodovias. Em tal estrutura, os acidentes são registrados em segmentos das rodovias, os quais, por definição, estão dentro de determinada UF no Brasil.

Dessa forma, o objetivo central deste trabalho é fornecer previsões para o número de acidentes nos próximos dias, tanto para os Unidades da Federação quanto para cada rodovia documentada nos dados abertos da PRF. Isso possibilitará identificar quais Estados e rodovias apresentam previsões destacadas no que se refere ao número de acidentes esperados. Com essas informações em mãos, será viável adotar uma abordagem preventiva, em vez de somente remediativa, no âmbito de políticas públicas, visando a redução dos acidentes no Brasil.

Após a introdução, o Capítulo 2 apresenta o referencial teórico, em que é detalhada definições fundamentais e modelos relacionados à análise de séries temporais. O Capítulo 3, intitulado Metodologia, é construído sobre esse referencial e explora os modelos e métodos a serem aplicados.

Antes de discutir os resultados, o Capítulo 4 conduzirá uma análise exploratória dos dados. Não apenas apresentaremos a evolução do número de acidentes rodoviários no Brasil ao longo dos anos, em diferentes níveis, mas também destacaremos o banco de dados usado, variáveis de interesse e informações adicionais extraídas dos dados.

Os resultados, com detalhes acerca das previsões para o número de acidentes em vários níveis pelo Brasil, serão expostos no Capítulo 5. Este capítulo é dividido em três grandes partes. A primeira aborda o *top-down*, na qual os modelos são ajustados no nível hierárquico regional para prever os acidentes nas Unidades da Federação, com base na desagregação mostrada na subseção 3.8.1. A segunda parte inclui previsões para cada rodovia, seguindo o mesmo método, mas iniciando no nível da UF. Na terceira parte, comparamos as duas abordagens para prever o número de acidentes nas Unidades da Federação, a primeira parte utiliza o método *top-down*, enquanto a segunda realiza previsões diretamente a partir dos modelos já ajustados no nível de UF.

Por fim, as ideias são organizadas e resumidas em formato de conclusão no Capítulo 6, em que também é abordada as limitações do trabalho e possibilidade de pesquisas futuras.

2 Referencial teórico

Serão tratados, nesta seção, de forma geral, o conceito de Operadores, Estacionariedade e dos modelos de séries temporais básicos, como Autoregressivo, Médias Móveis e o modelo $ARMA$.

Esta seção assume relevância no texto, estabelecendo uma base sólida para a compreensão das sutilezas e complexidades inerentes à análise de séries temporais. Além disso, ela serve como alicerça fundamental para o desenvolvimento do Capítulo 3, dedicado à exposição metodológica.

O embasamento desta seção é firmemente influenciado por Fiorucci (2021) e Morettin e Toloi (2018) que enriquecem a compreensão das metodologias empregadas na análise de séries temporais. Será fornecido um panorama abrangente das ferramentas, conceitos e modelos essenciais para a análise de séries temporais e exploraremos desde as definições básicas.

2.1 Operadores

2.1.1 Operador diferença

Definição 2.1. Dado processo $\{x_t,\ t = 1, 2, ...\}$, o operador diferença (∇) é definido como:

$$\nabla x_t = x_t - x_{t-1}.$$

Com isso, segue que na segunda ordem:

$$\nabla^2 x_t = \nabla(\nabla x_t) = x_t - 2x_{t-1} + x_{t-2}.$$

2.1.2 Operador de retardo

Definição 2.2. Dado processo $\{x_t,\ t = 1, 2, ...\}$, o operador de retardo (B) é definido como:

$$Bx_t = x_{t-1}.$$

Com isso, segue que na segunda ordem:

$$B^2 x_t = B(Bx_t) = x_{t-2}.$$

Como propriedades deste operador $\nabla x_t = (1-B)x_t$ e $\nabla^2 x_t = (1-B)^2 x_t$.

2.2 Estacionariedade

Comumente dois tipos de Estacionariedade são tratados na literatura, a Estacionariedade Forte e Fraca. A Estacionariedade Forte segue para um processo estritamente Estacionário, enquanto que a Estacionariedade Fraca será referenciada como Estacionariedade.

Um processo é estritamente Estacionário se a distribuição conjunta de $\{x_{t_1}, x_{t_2}, \dots, x_{t_k}\}$ é a mesma de $\{x_{t_1+h}, x_{t_2+h}, \dots, x_{t_k+h}\}$ para quaisquer índices $t_1, \dots, t_k$ e qualquer defasagem $h = 0, 1, 2, \dots$.

Dizemos que o processo com variância finita é Estacionário (fracamente Estacionário), se:

1. A função média é constante, isto é, $\mu_t = E[x_t] = \mu, \forall t =, 1, 2, \dots$
2. A função de autocovariâncias, $\gamma(s,t)$, depende de s e t apenas a partir da diferença $|s-t|$

Se o processo é Estacionário, então $Var[x_t]$ e $E[{x_t}^2]$ são constantes.

2.3 Modelo Autoregressivo

Conhecidos como AR(p):

Definição 2.3. Um processo estacionário $\{x_t,\ t = 1, 2,\ \dots\}$ é Autoregressivo de ordem p se:

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

Em que $\mu, \phi_1, \phi_2, \dots, \phi_p$ são constantes e $\{\varepsilon_t\}$ é Ruído Branco, ou seja:

$$\varepsilon_t \sim^{iid} N(0, \sigma^2). \tag{2.1}$$

Como observação, pode-se assumir $\mu = 0$ sem perda de generalidade. Basta considerar o processo para $x_t^* = x_t - \mu$.

O processo pode ser escrito em função do operador de retardo:

$$x_t = \phi_1 B x_t + \phi_2 B^2 x_t + ... + \phi_p B^p x_t + \varepsilon_t \iff$$
$$(1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p) x_t = \varepsilon_t \iff \Phi_p(B) x_t = \varepsilon_t.$$

Em que:

$$\Phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p. \tag{2.2}$$

É chamado de polinômio autoregressivo de ordem p.

Teorema 2.1. O processo AR(p) é estacionário se as raízes do polinômio característico

$$\Phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p.$$

estão fora do círculo unitário, isto é, em módulo as raízes são maiores que 1.

Demonstração. Veja, Box, Jenkins & Reinsed (1994) □

2.4 Modelo de Médias Móveis

Conhecido como MA(q):

Definição 2.4. Um processo MA(q) é tal que:

$$x_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta_q \varepsilon_{t-q}.$$

Em que $\theta_1, \theta_2, ..., \theta_p$ são constantes e $\{\varepsilon_t\}$ é Ruído Branco, vide Equação 2.1.

O modelo MA(q) pode ser escrito em função dos operadores de retardo:

$$x_t = \varepsilon_t + \theta_1 B \varepsilon_t + ... + \theta_q B^q \varepsilon_t \iff$$
$$x_t = (1 + \theta_1 B + ... + \theta_1 B^q) \varepsilon_t \iff x_t = \Theta(B) \varepsilon_t.$$

Em que:

$$\Theta(B) = (1 + \theta_1 B + ... + \theta_1 B^q). \tag{2.3}$$

É chamado de polinômio de médias móveis de ordem q.

É importante ressaltar que processos MA(q) são sempre estacionários e existe uma condição de inversibilidade para eles. Isto é, o modelo MA(q) é inversível se as raízes do polinômio de médias móveis $\Theta(B) = (1 + \theta_1 B + ... + \theta_1 B^q)$ estão fora do círculo unitário.

A condição de inversibilidade implica que esse modelo pode ser escrito como um $AR(\infty)$.

Como observação, considere o modelo AR(1): $x_t = \phi_1 x_{t-1} + \varepsilon_t$, com $|\phi_1| < 1$. Note que:

$$\begin{aligned} x_t = \phi_1 x_{t-1} + \varepsilon_t &\iff \\ x_t = \phi_1(\phi_1 x_{t-2} + \varepsilon_{t-1}) + \varepsilon_t &\iff \\ x_t = \phi_1^2 x_{t-2} + \phi_1 \varepsilon_{t-1} + \varepsilon_t &\iff \\ x_t = ... \Rightarrow x_t = \phi_1^k x_{t-k} + \sum_{i=0}^{k-1} (\phi^i \varepsilon_{t-1}). \end{aligned}$$

Assim, no limite $k \to \infty$, temos:

$$x_t = \sum_{i=0}^{\infty} (\phi^i \varepsilon_{t-1}).$$

O que corresponde ao modelo de médias móveis de ordem infinita, $MA(\infty)$.

Generalizando o resultado anterior, tem-se que qualquer modelo AR(p) estacionário pode ser escrito como um processo $MA(\infty)$. Suponha um processo estacionário $AR(p)$, $\Phi(B)x_t = \varepsilon_t$. A representação desse processo como $MA(\infty)$ é dada por:

$$x_t = \Psi(B)\varepsilon_t.$$

Em que $\Psi(B) = (1 + \psi_1 B + +\psi_2 B^2 + \psi_3 B^3 + ...)$, sendo ψ_1,ψ_2, ... os coeficientes.

Logo:

$$\Phi(B)^{-1} \varepsilon_t = \Psi(B)\varepsilon_t.$$

E, portanto:

$$\begin{aligned} 1 = \Phi(B)\Psi(B) &\iff \\ 1 = (1 - \phi_1 B - ... - \phi_p B^p)(1 + \psi_1 B + \psi_2 B^2 + \psi_3 B^3 + ...) &\iff \end{aligned} \tag{2.4}$$

Ao desenvolver a Equação 2.4 e agrupa-la em B,B^2,B^3, ..., obtém-se:

$$(\psi_1 - \phi_1)B + (\psi_2 - \phi_1\psi_1 - \phi_2)B^2 + (\psi_3 - \phi_1\psi_2 - \phi_2\psi_1 - \phi_3)B^3 + ... = 0.$$

Assim, é possível obter os coeficientes MA de forma recursiva:

$$\begin{aligned}
\psi_1 &= \phi_1 \\
\psi_2 &= \phi_1\psi_1 + \phi_2 \\
\psi_3 &= \phi_1\psi_2 + \phi_2\psi_1 + \phi_3 \\
\psi_4 &= \phi_1\psi_3 + \phi_2\psi_2 + \phi_3\psi_1 + \phi_4 \\
&\vdots \\
\psi_i &= \sum_{j=1}^{i} \phi_j\psi_{i-j}.
\end{aligned}$$

Com $\psi_0 = 1$ e $\phi_j = 0$ para $j > p$.

2.5 Modelo ARMA

Conhecido como, modelo misto, ARMA(p,q), ele combina os modelos Autoregressivos e Médias Móveis:

Definição 2.5. Um processo ARMA(p,q) é um processo Estacionário tal que:

$$\mu = 0 \Rightarrow x_t = \phi_1 x_{t-1} + ... + \phi_p x_{t-p} + \varepsilon_t + \theta_1\varepsilon_{t-1} + ... + \theta_q\varepsilon_{t-q}.$$

ou

$$\mu \neq 0 \Rightarrow x_t - \mu = \phi_1(x_{t-1} - \mu) + ... + \phi_p(x_{t-p} - \mu) + \varepsilon_t + \theta_1\varepsilon_{t-1} + ... + \theta_q\varepsilon_{t-q} \iff$$
$$x_t = \alpha + \phi_1 x_{t-1} + ... + \phi_p x_{t-p} + \varepsilon_t + \theta_1\varepsilon_{t-1} + ... + \theta_q\varepsilon_{t-q}.$$

Em que $\mu, \phi_1, \phi_2, \ ... \ , \phi_p, \theta_1, \theta_2, \ ... \ , \theta_p$ são constantes, $\{\varepsilon_t\}$ é Ruído Branco, vide Equação 2.1, e $\alpha = (1 - \phi_1 - ... - \phi_p)\mu$ é o intercepto.

Os modelos ARMA(p,q) podem ser escritos em função dos operadores de retardo:

$$\Phi(B)x_t = \Theta(B)\varepsilon_t. \tag{2.5}$$

Em que $\Phi(B)$ é o polinômio autoregressivo, visto na Equação 2.2, e $\Theta(B)$ é o polinômio de média móvel, vista na Equação 2.3.

O processo ARMA(p,q) é estacionário se as raízes de $\Phi(B)$ estão fora do círculo unitário e ele é inversível se as raízes de $\Theta(B)$ estão fora do círculo unitário.

3 Metodologia

Nesta seção será exposto os procedimentos e abordagens adotados para obter os resultados e suas devidas interpretações. Nesta etapa, serão apresentados os métodos, modelos e técnicas empregados efetivamente no trabalho.

Exploraremos o conceito de Séries Temporais Hierárquicas, Séries Temporais Agrupadas, o modelo *ARIMA*, a Regressão Dinâmica, Regressão Dinâmica Harmônica, modelos de alisamento exponencial, a abordagem *Top-down*, medidas de acurácia e, por fim, técnicas para validação cruzada. A metodologia busca não apenas abordar as especificidades técnicas, mas também ressaltar a coerência e rigor que sustentam todo o processo de análise de séries temporais neste estudo.

3.1 Séries temporais hierárquicas

De acordo com a tratativa em Rob J. Hyndman, Lee e Wang (2016), a estrutura hierárquica é definida pela desagregação apenas por um atributo e ao varrer os níveis da hierarquia de baixo para cima, ou de cima para baixo, as séries (fixado nível) estarão sempre contidas (estritamente) entre si.

Como exemplo, considere os acidentes em rodovias no Brasil. Se tomarmos a estrutura de desagregação de Região para UF, ao varrer os níveis, tanto de cima para baixo quanto de baixo para cima, as séries vão conter e estar contidas, respectivamente, entre si. Veja a Figura 1, que ilustra o exemplo.

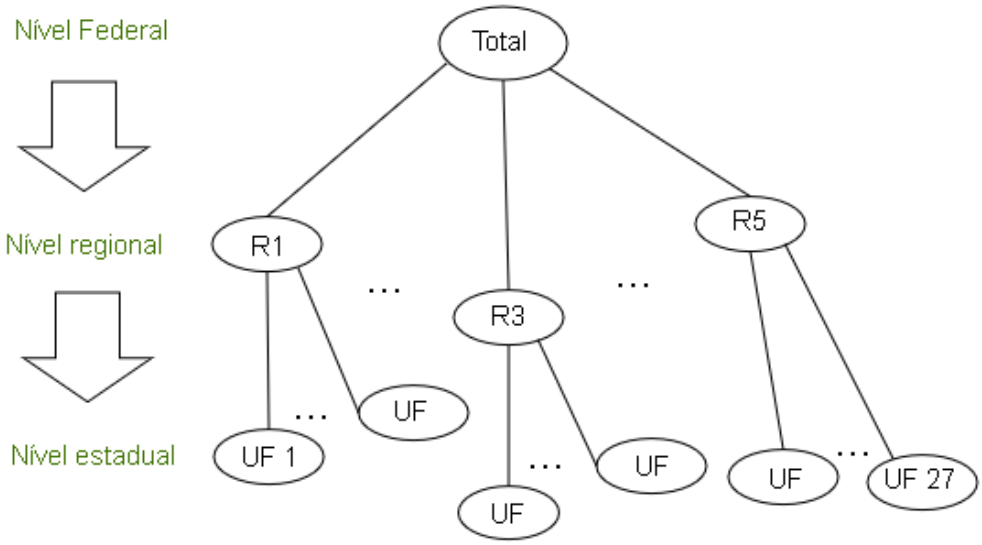

Figura 1 – Estrutura hierárquica no trabalho

Veja a Figura 2, em que existe apenas um (1) atributo na desagregação - que pode tomar valor A, B ou C - o que define estrutura hierárquica.

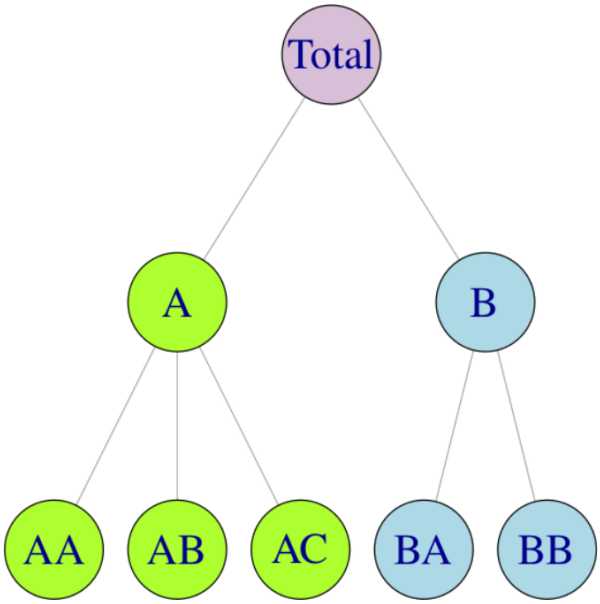

Figura 2 – Estrutura hierárquica geral

Fonte: Retirada de Rob J Hyndman e Athanasopoulos (2018, seção 10.1)

3.2 Séries temporais agrupadas

Ademais, em outros casos a desagregação pode seguir múltiplos atributos. Assim, são definidas as estruturas agrupadas. Ou seja, a desagregação é dada por diferentes hierarquias (uma para cada atributo), que quando consideradas separadamente podem formar estrutura hierárquica, porém quando consideradas conjuntamente a hierarquia se perde necessariamente, visto que ao varrer os níveis dessa estrutura (considerando os múltiplos atributos simultaneamente) as séries nem sempre contém ou estão contidas entre si (estritamente).

Como exemplo, considere os acidentes em rodovias no Brasil. Se tomarmos a estrutura de desagregação Região para UF, e UF para rodovia (Rodovia Federal ou BR), tem-se dois atributos - o primeiro geográfico e o segundo de natureza. Para esse exemplo, ao varrer os níveis da estrutura de cima para baixo, o acidente ocorre em alguma região do Brasil, que por sua vez ocorreu em algum Estado contido (estritamente) na região, e, por definição, ocorreu em alguma rodovia que cruza vários estados e possivelmente também mais de uma região. A mesma coisa acontece ao varrer os níveis de baixo para cima, o acidente ocorre, por definição em uma rodovia, que está dentro de diversos Estados no Brasil e possivelmente também cruza mais de uma região. Veja a Figura 3, que ilustra o exemplo.

Veja a Figura 4, em que existem múltiplos (dois) atributos na desagregação - o primeiro, que toma valore X ou Y e o segundo que toma valor A, B ou C - o que implica na

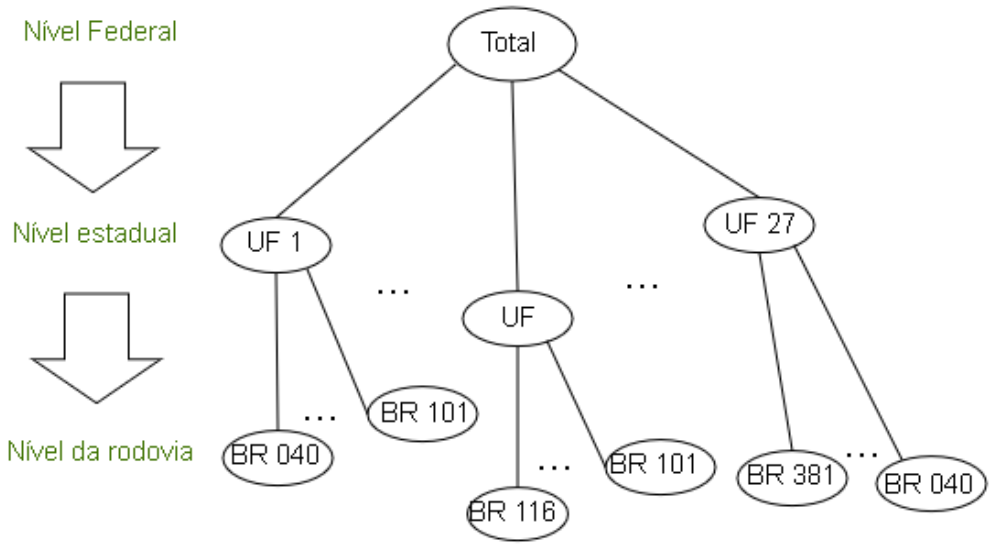

Figura 3 – Estrutura agrupada no trabalho

estrutura agrupada.

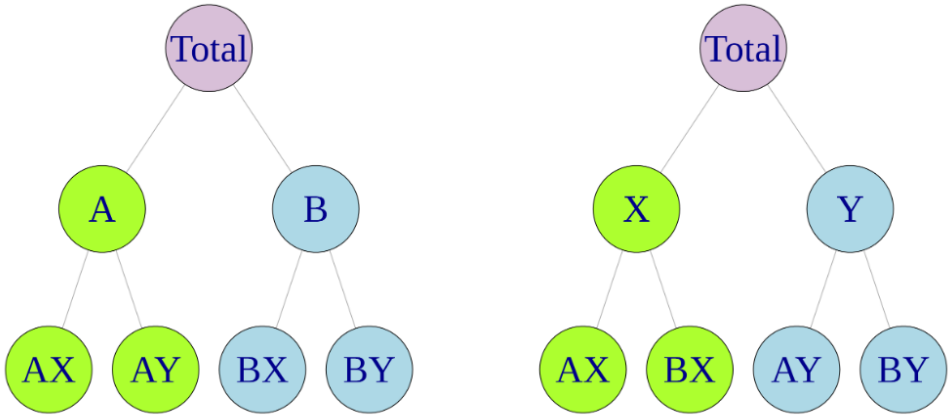

Figura 4 – Estrutura agrupada geral

Fonte: Retirada de Rob J Hyndman e Athanasopoulos (2018, seção 10.2)

3.3 Modelo ARIMA

Foi visto na seção 2.5 que modelos ARMA só podem ser aplicados para séries estacionárias. No entanto, a maior parte das séries não são estacionárias. Modelos autoregressivos integrados de médias móveis, $ARIMA(p,d,q)$, consistem em aplicar o modelo $ARMA(p,q)$ na d-ésima diferença da série.

Seja $\{x_t\}$ um processo não estacionário. Em geral, após algumas diferenças a série formada por $w_t = \nabla^d x_t$ se torna estacionária. Assim, o modelo $ARIMA(p,d,q)$ pode ser escrito como o modelo $ARMA(p,q)$ aplicada à série $\{w_t\}$, como visto na Equação 2.5, isto é:

$$\Phi_p(B)w_t = \Theta_q(B)\varepsilon_t. \tag{3.1}$$

Em que $\nabla x_t = (1-B)x_t \Rightarrow w_t = \nabla^d x_t = (1-B)^d x_t$.

Então, o modelo $ARIMA(p,d,q)$, pode ser escrito como:

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

Em que $\Phi(B)$ é o polinômio autoregressivo, visto na Equação 2.2, e $\Theta(B)$ é o polinômio de média móvel, vista na Equação 2.3.

Para mais detalhes, veja Rob J Hyndman e Athanasopoulos (2018, seção 8.5)

3.4 Modelo ARIMA Sazonal

Seja s o número de observações por ciclo sazonal. Por exemplo, para um ciclo sazonal anual em uma série mensal, $s = 12$. Em suma, o ciclo sazonal, para esse exemplo, da série implica que a obervação no mês de janeiro deste ano depende da observação de janeiro do ano passado (janeiro do ano passado depende de janeiro do ano retrasado, assim por diante), da mesma forma que fevereiro deste ano depende de fevereiro do ano passado, assim por diante.

Definição 3.1. O operador diferença sazonal (∇_s) é definido como:

$$\nabla_s x_t = x_t - x_{t-s}.$$

Para ordens maiores, basta aplicar o operador de forma recursiva, por exemplo:

$$\nabla_s^2 x_t = \nabla_s(x_t - x_{t-s}) = \nabla_s x_t - \nabla_s x_{t-s} = x_t - 2x_{t-s} + x_{t-2s}.$$

Em termos do operador de retardo:

$$\nabla_s x_t = (1-B^s)x_t \Rightarrow \nabla_s^D x_t = (1-B^s)^D x_t.$$

O modelo $SARIMA(p,d,q)x(P,D,Q)$ é escrito como:

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

Em que $\Phi(B^s)$ é o polinômio autoregressivo sazonal e $\Theta(B^s)$ é o polinômio de médias móveis sazonal escritos como:

$$\Phi(B^s) = 1 - \varphi_1 B^s - \varphi_2 B^{2s} - ... - \varphi_P B^{Ps}$$
$$\Theta(B^s) = 1 + \vartheta_1 B^s - \vartheta_2 B^{2s} - ... - \vartheta_Q B^{Qs}.$$

Ou seja, o modelo SARIMA generaliza todos os modelos da família ARIMA.

Para mais detalhes, veja Rob J Hyndman e Athanasopoulos (2018, seção 8.9)

3.5 Regressão dinâmica

De acordo com o prof. Robin John Hyndman[1], não existe uma única forma de adicionar variáveis explanatórias para o modelo ARIMA, visto na seção 3.3 e sua argumentação completa segue na página `https://robjhyndman.com/hyndsight/arimax/` de seu *site*. Por simplicidade, considere o modelo ARIMA não sazonal e a demanda por adicionar apenas uma variável explicativa. Ainda, assuma um processo estacionário.

Assim, adota-se o modelo $ARMA(p,q)$, escrito para a série $y_1, ..., y_n$, na forma:

$$y_t = \phi_1 y_{t-1} + ... + \phi_p y_{t-p} - \theta_1 z_{t-1} - ... - \theta_q z_{t-q} + z_t. \tag{3.2}$$

Em que o processo z_t é um ruído branco, visto em Equação 2.1.

3.5.1 Modelo ARMAX

Apresentado contexto, o modelo ARMAX simplesmente adiciona variáveis ao lado direito na Equação 3.2. Isto é:

$$y_t = \beta x_t \phi_1 y_{t-1} + ... + \phi_p y_{t-p} - \theta_1 z_{t-1} - ... - \theta_q z_{t-q} + z_t.$$

Em que x_t é a covariável no tempo t e β seu coeficiente. Embora isso pareça simples, uma desvantagem é que o coeficiente de covariável é difícil de interpretar.

O modelo ARMAX pode ser escrito em função dos operadores de retardo:

$$\Phi(B)y_t = \beta x_t + \Theta(B)z_t \iff y_t = \frac{\beta x_t}{\Phi(B)} + \frac{\Theta(B)z_t}{\Phi(B)}.$$

Em que $\Phi(B) = 1 - \phi_1 B - ... - \phi_p B^p$ e $\Theta(B) = 1 - \theta_1 B - ... - \theta_q B^q$. Note como os coeficientes da parte autoregressiva se confundem com as variáveis explicativas e o termo de erro.

[1] `https://robjhyndman.com/hyndsight/arimax/`

3.5.2 *Regression with ARMA errors*

O modelo é escrito na forma:

$$y_t = \beta x_t + \eta_t$$
$$\eta_t = \phi_1 \eta_{t-1} + ... + \phi_p \eta_{t-p} - \theta_1 \eta_{t-1} - ... - \theta_q \eta_{t-q} + z_t.$$

Nesse caso, o coeficiente de regressão tem sua interpretação usual. O modelo pode ser escrito em função dos operadores de retardo:

$$y_t = \beta x_t + \frac{\Theta(B) z_t}{\Phi(B)}.$$

3.5.2.1 *Regression with ARIMA errors*

Considere o modelo de regressão, com forma:

$$y_t = \beta_0 + \beta_1 x_{1,t} + ... + \beta_k x_{k,t} + \varepsilon_t.$$

Em que y_t é uma função linear das k variáveis preditoras $(x_{1,t}, \ ... \ , x_{k,t})$ e ε_t é assumido como ruido branco.

Aqui, será considerado que ε_t pode possuir autocorrelação. Para dar ênfase, troca-se ε_t por η_t, em que se assume que $\{\eta_t\}$ segue um modelo ARIMA. Se $\{\eta_t\}$ segue ARIMA(1,1,1), segue:

$$y_t = \beta_0 + \beta_1 x_{1,t} + ... + \beta_k x_{k,t} + \eta_t \ ,$$

$$(1 - \phi_1 B)(1 - B)\eta_t = (1 + \theta_1 B)\varepsilon_t.$$

Em que ε_t é um ruido branco.

Note que o modelo, nesse caso *Regression with ARIMA errors*, tem dois termos de erro - o erro da regressão, denotado por η_t, e o erro do modelo ARIMA, denotado por ε_t. Somente o erro do modelo ARIMA é assumido como sendo um ruido branco.

Para mais detalhes, veja Rob J Hyndman e Athanasopoulos (2018, capítulo 9).

3.5.3 *Transfer function models*

Tanto o modelo ARMAX, quanto a *Regression with ARMA errors* podem ser considerados como casos especiais dos *Transfer function models*, popularizados por Box e Jenkins,

como escreve o prof. Robin John Hyndman em sua página[2]:

$$y_t = \frac{\beta(B)}{\upsilon(B)} + \frac{\Theta(B)z_t}{\Phi(B)}.$$

Isso, permite a contabilização de *lagged and decaying effects* para variáveis explanatórias, via operadores $\beta(B)$ e $\upsilon(B)$. São estes, considerados *dynamic regression models*, para mais detalhes, veja Pankratz (1992).

3.6 Regressão dinâmica harmônica

A Regressão dinâmica harmônica, ou *Dynamic harmonic regression* (DHR), baseia-se no princípio de que uma combinação de funções senos e cossenos pode aproximar qualquer função periódica:

$$y_t = \beta_0 + \sum_{k=1}^{K}[\alpha_k s_k(t) + \gamma_k c_k(t)] + \varepsilon_t \ ,$$

Em que $s_k(t) = sin(\frac{2\pi kt}{m})$, $c_k(t) = cos(\frac{2\pi kt}{m})$, K é número de harmônicos necessários para aproximação, m é o tamanho do ciclo sazonal, fixado harmônico, α_k e γ_k são coeficientes da regressão e ϵ_t é modelado como um processo ARIMA não sazonal.

Para mais detalhes, veja Rob J Hyndman e Athanasopoulos (2018, capítulo 9).

3.7 Modelos de alisamento exponencial para dados sazonais

Em Livera, Rob J. Hyndman e Snyder (2011), os autores descrevem no artigo o uso de Modelos de alisamento exponencial em diferentes conjuntos de dados, com múltipla e complexa sazonalidade. Ao final, constatou-se que os modelos trigonométricos obtiveram melhor performance quando comparados com os modelo BATS. Além do comparativo conforme aplicação, o artigo traz a definiçãoe conceito, principalmente, do modelo de Holt-Winters, BATS e TBATS, além do procedimento para estimação e seleção de modelos.

3.7.1 Modelo de Holt-Winters

Esse modelo é comumente retratado como o mais famoso da família de alisamento exponencial.

[2] https://robjhyndman.com/hyndsight/arimax/

Considere m o tamanho do ciclo sazonal e o processo $\{s_t\}$ a componente sazonal. Por exemplo, para uma série mensal, seu ciclo sazonal anual possui $m = 12$.

A construção do modelo assume que a previsão um (1) passo a frente é formada pela soma da tendência $(\ell_t + b_t)$ e da sazonalidade(s_t). Assim, o **método** aditivo de Holt-Winters segue:

$$\begin{aligned} \hat{y}_t &= \ell_{t-1} + b_{t-1} + s_{t-m} \\ \ell_t &= \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1}) \\ b_t &= \beta^* (\ell_t - \ell_{t-1}) + (1-\beta^*) b_{t-1} \\ s_t &= \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1-\gamma) s_{t-m}. \end{aligned}$$

Em que $\alpha \in (0,1)$, $\beta^* \in (0,1)$ e $\gamma \in (0,1)$ são parâmetros de alisamento, também, $\ell_0 \in \mathbb{R}$, $b_0 \in \mathbb{R}$ e $(s_{-m+1}, s_{-m+1}, \dots, s_{-1}, s_0) \in \mathbb{R}^m$ são parâmetros de inicialização.

De acordo com Taylor (2003) a versão linear do método de Holt-Winters pode ser estendida para incorporar uma segunda componente sazonal:

$$\begin{aligned} y_t &= \ell_{t-1} + b_{t-1} + s_t^{(1)} + s_t^{(2)} + d_t \\ \ell_t &= \ell_{t-1} + b_{t-1} + \alpha d_t \\ b_t &= b_{t-1} + \beta d_t \\ s_t^{(1)} &= s_{t-m_1}^{(1)} + \gamma_1 d_t \\ s_t^{(2)} &= s_{t-m_2}^{(2)} + \gamma_2 d_t. \end{aligned} \tag{3.3}$$

Em que m_1 e m_2 são os tamanhos dos ciclos sazonais 1 e 2, respectivamente, e d_t é um ruído branco como na Equação 2.1. As componentes ℓ_t e b_t representam o *level* e a tendência da série no tempo t, respectivamente. Os coeficientes α, β, γ_1 e γ_2 são chamados parâmetros de alisamento. Por fim, $\ell_0, b_0, \{s_{t-m_1}^{(1)}, \dots, s_0^{(1)}\}$ e $\{s_{t-m_2}^{(2)}, \dots, s_0^{(2)}\}$ são parâmetros de inicialização.

3.7.2 Modelo BATS

Como visto em Livera, Rob J. Hyndman e Snyder (2011), pode-se estender o modelo visto pela Equação 3.3 afim de incluir a transformação de Box-Cox (BOX; COX, 1964), *ARMA errors* e T ciclos sazonais, isto é:

$$
\begin{aligned}
y_t^{(\omega)} &= \begin{cases} \frac{y_t^{\omega}-1}{\omega}; & \omega \neq 0 \\ log(y_t); & \text{otherwise} \end{cases} \\
y_t^{(\omega)} &= \ell_{t-1} + \phi b_{t-1} + (\sum_{i=1}^{T} s_{t-m_i}^{(i)}) + d_t \\
\ell_t &= \ell_{t-1} + \phi b_{t-1} + \alpha d_t \\
b_t &= (1-\phi)b + \phi b_{t-1} + \beta d_t \\
s_t^{(i)} &= s_{t-m_i}^{(i)} + \gamma_i d_t \\
d_t &= (\sum_{i=1}^{p} \varphi_i d_{t-i}) + (\sum_{i=1}^{q} \theta_i \varepsilon_{t-i}) + \varepsilon_t.
\end{aligned}
\tag{3.4}
$$

A notação $y_t^{(\omega)}$ é usada para representar a transformação Box-Cox realizada na série observada com parâmetro ω. Além disso, $(m_1, m_2, \ldots, m_T)$ denota os respectivos tamanhos dos T ciclos sazonais, ℓ_t é o *local level* no tempo t, b é a tendência de longo prazo, b_t é a tendência de curto prazo no tempo t, $s_t^{(i)}$ representa a i-ésima componente sazonal no tempo t, d_t denota um processo $ARMA(p,q)$ e ε_t é um Ruído Branco, como visto na Equação 2.1. Os parâmetros de alisamento ou *smoothing parameters* são dados por α, β e γ_i para $i = 1, \ldots, T$.

A identificação BATS é um acrônimo para partes chave do modelo, como a transformação ***B**ox-Cox, **A**RMA errors, **T**rend, e suas múltiplas componentes **S**azonais*. O modelo também é visto como BATS$(\omega, \phi, p, q, m_1, m_2, \ldots, m_T)$ para indicar o parâmetro de Box-Cox, Amortecimento ou *dumping*, ARMA(p,q) e os períodos sazonais $(m_1, m_2, \ldots, m_T)$.

3.7.3 Modelo TBATS

Também vistos como *trigonometric seasonal models*, pois suas componentes sazonais seguem uma representação trigonométrica baseada na série de fourier (WEST; HARRISON, 1997; HARVEY, 1990):

$$
s_t^{(i)} = \sum_{j=1}^{k_i} s_{j,t}^{(i)}.
$$

Em que:

$$
s_{j,t}^{(i)} = s_{j,t-1}^{(i)} cos(\lambda_j^{(i)}) + s_{j,t-1}^{*(i)} sin(\lambda_j^{(i)}) + \gamma_1^{(i)} d_t.
$$

$$
s_{j,t}^{*(i)} = -s_{j,t-1} sin(\lambda_j^{(i)}) + s_{j,t-1}^{*(i)} cos(\lambda_j^{(i)}) + \gamma_2^{(i)} d_t.
$$

Em que $\gamma_1^{(i)}$ e $\gamma_2^{(i)}$ são parâmetros de alisamento e $\lambda_j^{(i)} = 2\pi j/m_i$. O *stochastic level* do i-ésimo componente sazonal é $s_{j,t}^{(i)}$ e o crescimento estocástico, no nível do i-ésimo componente sazonal, é representado pelo $s_{j,t}^{*(i)}$. O número de harmônicos necessários para o i-ésimo componente sazonal é denotado por k_i. De acordo com Livera, Rob J. Hyndman e Snyder (2011), existe equivalência com a abordagem de indexar o número de harmônicos para a i-ésima componente sazonal em função do m_i. Para m_i ímpar, $k_i = \frac{m_i-1}{2}$ e para m_i par, $k_i = \frac{m_i}{2}$. De acordo com Livera, Rob J. Hyndman e Snyder (2011) a maioria dos componentes sazonais exigirão menos harmônicos, reduzindo assim o número de parâmetros para estimar. Uma representação determinística dos componentes sazonais pode ser obtida definindo os parâmetros de suavização iguais a zero.

O modelo TBATS(ω,ϕ, $p, q, \{m_1,k_1\}, \{m_2,k_2\}$, ... ,$\{m_T,k_T\}$), possui **T** inicial com conotação trigonométrica, e é obtido quando trocamos a componente sazonal $s_t^{(i)}$, pela sua formulação sazonal trigonométrica, e a forma de escrever $y_t^{(\omega)}$, para $y_t^{(\omega)} = \ell_{t-1} + \phi b_{t-1} + (\sum_{i=1}^{T} s_{t-1}^{(i)}) + d_t$, na Equação 3.4.

3.8 Abordagem *Top-down*

Consoante à estrutura hierárquica vista na seção 3.1, a estratégia, aqui, é basicamente gerar as previsões para a série no nível mais agregado e, em seguida, desagrega-las na hierarquia.

Considere $p_1, \ldots, p_m$ como proporções para desagregação. Elas ditam como as previsões no nível mais agregado serão distribuídas para obtenção as previsões para cada série no nível menos agrupado. Por exemplo, para a Figura 2 são dadas:

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

Ou, usando a notação matricial:

$$\tilde{\mathbf{b_t}} = \mathbf{p}\tilde{y}_t.$$

Uma vez que as previsões, com horizonte h, no nível menos agregado são feitas, elas podem ser agregadas para gerar previsões para o resto das séries dentro da estrutura:

$$\tilde{\mathbf{y_h}} = \mathbf{Sp}\tilde{y}_t.$$

Os dois métodos para **top-down** mais comuns especificam as proporções baseado nas proporções históricas pesquisadas nos dados históricos. Eles performaram bem no estudo de Gross e Sohl (1990).

3.8.1 Proporções das médias históricas

Conhecido como *top-down Gross-Sohl method F*, para $j = 1, \dots, m$, cada proporção p_j captura o valor histórico médio da série no nível menos agregado em relação ao valor médio da série no nível mais agregado.

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

3.8.2 Proporções médias históricas

Conhecido como *top-down Gross-Sohl method A*, para $j = 1, \dots, m$, cada proporção p_j reflete a média das proporções históricas entre os valores observados na série menos agregada, varrendo os tempos $t = 1, \dots, T$, em relação à série mais agregada.

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

3.9 Acurácia das previsões

3.9.1 Média dos erros absolutos

Conhecido como MAE, ou erro absoluto médio:

$$MAE = \frac{1}{h} \sum_{i=1}^{h} |y_{n+i} - \hat{y}_{n+i|n}|.$$

Nota-se que a MAE não é livre de escala.

3.9.2 *Symmetric mean absolute percentage*

Conhecida como sMAPE, ela é uma métrica invariante por escala:

$$sMAPE = \frac{100}{h} \sum_{i=1}^{h} \frac{|y_{n+i} - \hat{y}_{n+i|n}|}{\frac{|y_{n+i}| + |\hat{y}_{n+i|n}|}{2}}.$$

Note que a sMAPE favorece erros positivos.

3.9.3 *Mean Absolute Scaled Error*

Conhecida como MASE, proposta em Rob J. Hyndman e Koehler (2006), ela tem forma:

$$q_t = \frac{|y_{n+i} - \hat{y}_{n+i|n}|}{\frac{1}{n-1}\sum_{i=2}^{n}|y_i - y_{i-1}|}.$$

$$MASE = mean(|q_t|).$$

3.10 Validação cruzada

A ideia básica é avaliar o modelo fora do conjunto de dados de ajuste. Ou seja, é considerado um conjunto de treinamento, porção dos dados para ajuste dos modelos, e outro de teste, parte dos dados para avaliar os modelos.

3.10.1 Avaliação por janela fixa

Aqui, o conjunto de dados é separado em duas partes. O modelo é ajustado na parte de parte de treinamento e a qualidade da previsão é avaliada na parte de validação. Um grande ponto negativo dessa abordagem é que os resultados dependem muito da escolha da janela de validação e, com ela, há bastante chance de encontrar resultados diferentes se outra janela for escolhida.

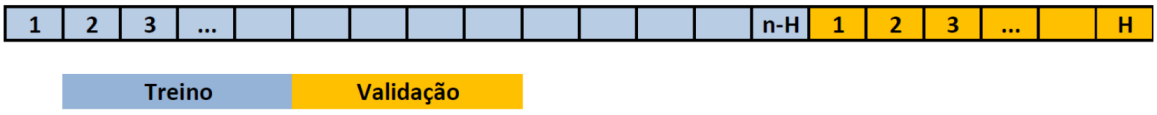

Figura 5 – Janela fixa

Fonte: Fiorucci (2021)

3.10.2 Avaliação por janela deslizante

Para essa abordagem, são consideradas múltiplas janelas de treinamento e teste e cada horizonte de previsão é avaliado múltiplas vezes. Por isso os resultados encontrados tendem a ser mais robustos.

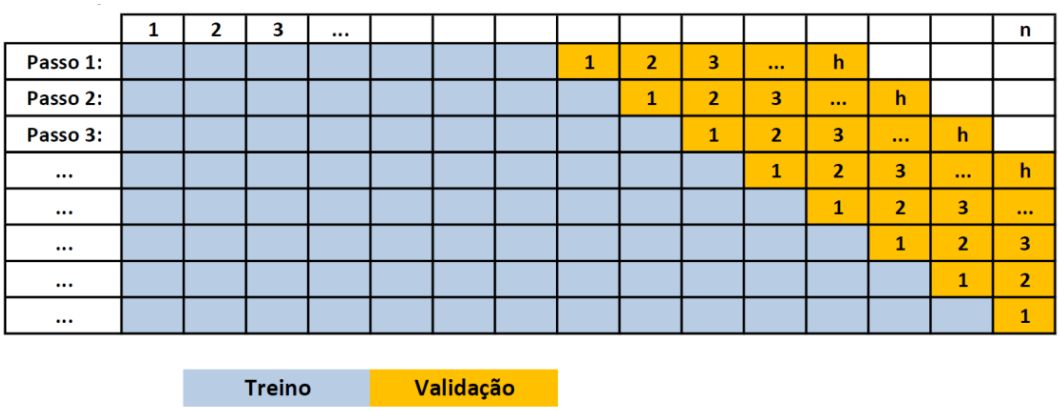

Figura 6 – Janela deslizante

Fonte: Fiorucci (2021)

4 Análise exploratória dos dados

A partir do *web scraping* escrito em *python*, vide seção A.1, foi feito um tratamento às planilhas disponibilizadas pela pela PRF desde 2007[1] (dados abertos da PRF) afim de concatena-las em uma única planilha. Com pouco mais de duas milhões de linhas e 37 colunas (variáveis), o trabalho seguiu completamente a partir deste banco de dados unificado - que aborda apenas acidentes em rodovias federais. Cada linha do banco representa uma ocorrência com *id* que o referencia unicamente. As 37 colunas do banco trazem informações com respeito à identificação da ocorrência (pelo *id* único), à data, horário e local do acidente (para o local são informadas latitude e longitude, a respectiva quilometragem (km) da BR, a BR, município e Estado). Além disso, não somente é informado o local, mas é, ainda, detalhado o local - é documentado o traçado[2] e sentido da via, além do tipo[3] da pista. O acidente também é classificado[4] e detalhado[5]. Veja a Figura 7 com o *print* das primeiras linhas.

	id	data_inversa	dia_semana	horario	uf	br	km	municipio	causa_acidente	tipo_acidente	...
0	10.0	2007-06-11	Segunda	15:30:00	MG	381	623.2	OLIVEIRA	Falta de atenção	Colisão frontal	...
1	10.0	2007-06-11	Segunda	15:30:00	MG	381	623.2	OLIVEIRA	Falta de atenção	Colisão frontal	...
2	1032898.0	2007-08-13	Segunda	14:25:00	MG	40	585.5	ITABIRITO	Outras	Saída de Pista	...
3	1051130.0	2007-02-12	Segunda	02:10:00	MA	135	11	SAO LUIS	Animais na Pista	Atropelamento de animal	...
4	1066824.0	2007-11-20	Terça	05:30:00	CE	222	30.8	CAUCAIA	Defeito mecânico em veículo	Capotamento	...

5 rows × 37 columns

Figura 7 – Retrato do banco de dados para o trabalho

Fonte: PRF com extração prória

Foi feito um extensivo tratamento aos dados para chegar na série temporal diária para o número de acidentes. Para cada nível dentro da hierarquia - Federal, Regional, Estadual e

[1] https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos/dados-abertos-da-prf

[2] Cruzamento, Curva, Desvio Temporário, Interseção de vias, Não Informado, Ponte, Reta, Retorno Regulamentado, Rotatória, Túnel ou Viaduto

[3] Dupla, Múltipla ou Simples

[4] Com Vítimas Fatais, Com Vítimas Feridas, Ignorado ou Sem Vítimas

[5] Atropelamento de Animal, Atropelamento de Pedestre, Atropelamento de animal, Atropelamento de pessoa, Capotamento, Colisão Transversal, Colisão com bicicleta, Colisão com objeto, Colisão com objeto em movimento, Colisão com objeto estático, Colisão com objeto fixo, Colisão com objeto móvel, Colisão frontal, Colisão lateral, Colisão lateral mesmo sentido, Colisão lateral sentido oposto, Colisão transversal, Colisão traseira, Danos Eventuais, Danos eventuais, Derramamento de Carga, Derramamento de carga, Engavetamento, Eventos atípicos, Incêndio, Queda de motocicleta / bicicleta / veículo, Queda de ocupante de veículo, Saída de Pista, Saída de leito carroçável ou Tombamento

ainda no nível da BR - foram feitos tratamentos diferentes e detalhados usando a linguagem de programação *Pyhton*. As séries podem ser visualizadas na Figura 8, que mostra os acidentes ao longo tempo das 5 regiões, Estados e rodovias federais com maior número de acidentes, desde 2007.

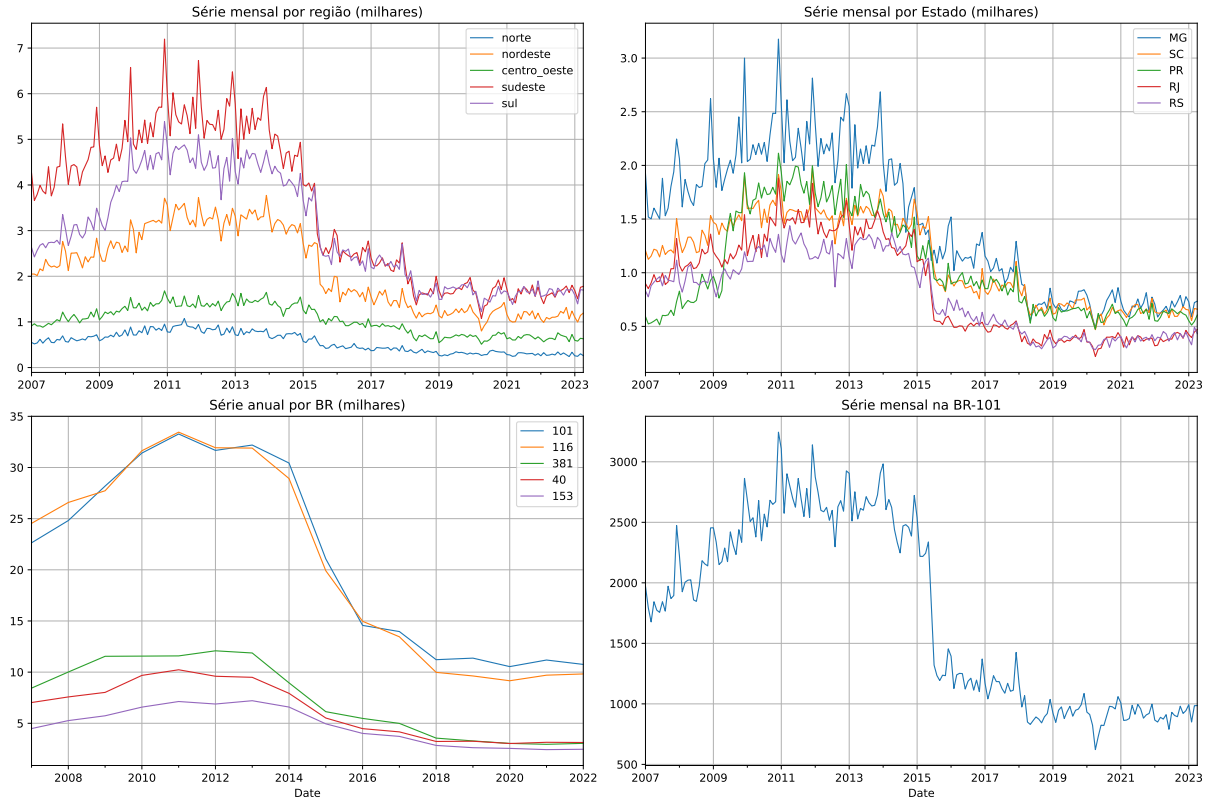

Figura 8 – Acidentes em diferentes níveis, ao longo dos anos, no Brasil

Fonte: Autoria própria

Nota-se que as Unidades Federativas do sul e sudeste do país se destacam com respeito ao número de acidentes, enquanto que no nível estadual, MG é destaque até o início de 2015, após esse período os Estados quase se equiparam.

De acordo com Rob J. Hyndman, Lee e Wang (2016), as séries mais desagregadas geralmente têm um alto grau de volatilidade, enquanto a série temporal mais agregada é geralmente suave e menos barulhenta. Veja que para o nível mais desagregado possível, na BR-101, é observado esse comportamento - mais evidenciado, ainda na Figura 9, em que é mostrado a série no nível mais agregado possível com frequência diária.

Para séries temporais mensais com sazonalidade, é esperado ciclo sazonal anual, como visto na seção 3.4. Entretanto, para toda a análise dentro deste trabalho, mostrada no Capítulo 5, trataremos com séries temporais diárias. É esperado que, para essas séries, haja mais de um único ciclo sazonal, como por exemplo um ciclo semanal, outro mensal ou até mesmo anual. Veja que ao filtrarmos a tendência e diferenciarmos a série resultante afim de

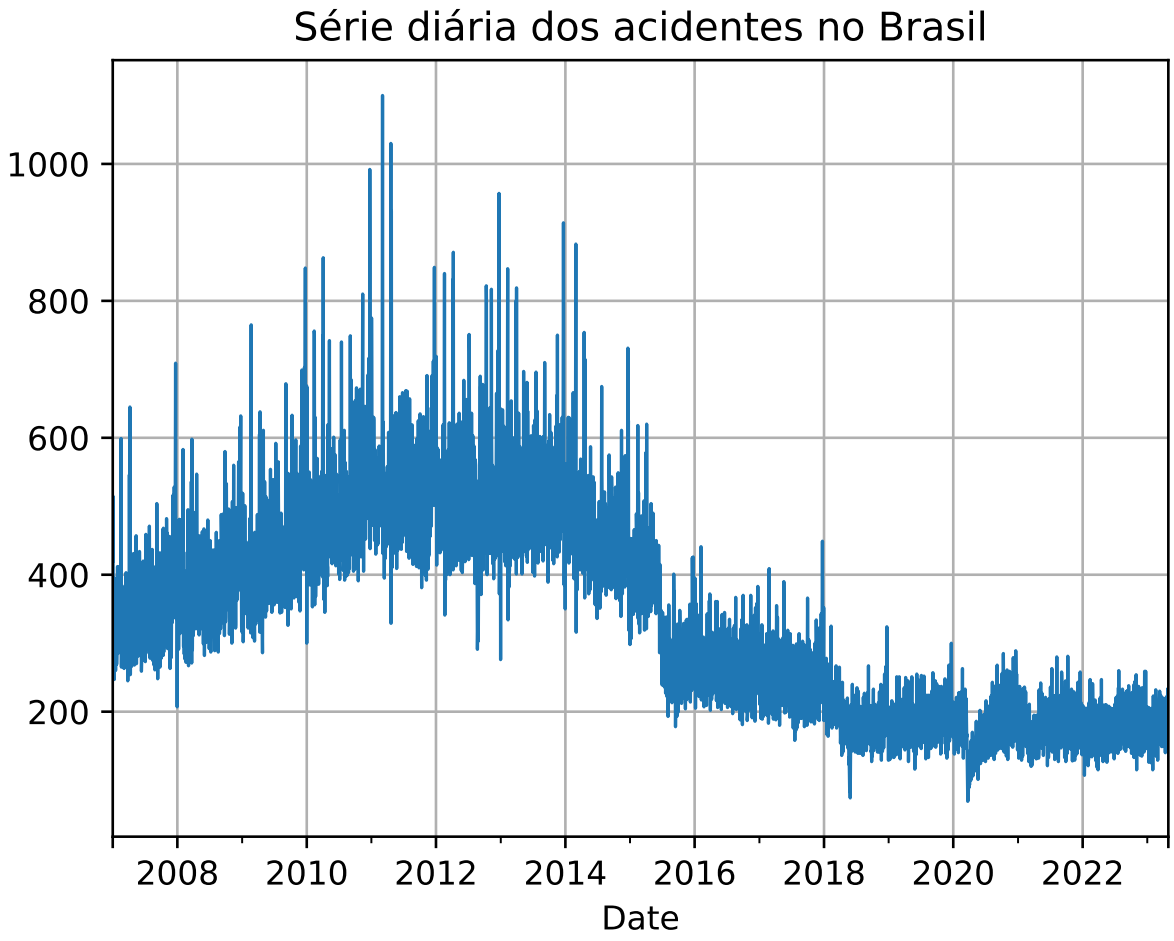

Figura 9 – Acidentes em rodovias com frequência diária no Brasil

Fonte: Autoria própria

filtrar um dos três possíveis ciclos[6], ainda assim sobra sazonalidade para a série do número de acidentes no Brasil.

Assim, faz-se o seguinte procedimento para chegarmos na Figura 10 e na Figura 11:

a) Ajusta-se a tendência da série de acidentes

b) Com a tendência ajustada, filtra-se um dos três possíveis ciclos sazonais aplicando $\nabla_{s_i} x_t$, sendo i o ciclo com sazonalidade s_i.

Com respeito à Figura 10, é ajustada sazonalidade anual e ainda é observada sazonalidade semanal, pela autocorrelação significativa no sétimo *lag*, ciclicamente (no gráfico com título *Autocorrelation* para sazonalidade anual filtrada). Além disso, quando filtrada sazonalidade semanal, é possível concluir que a sazonalidade mensal não é evidente, isto é, o número de acidentes não segue um padrão de repetição mês à mês. Portanto, o número de acidentes não é sensível quanto ao início, meio ou final do mês, mas sim quanto aos dias da semana (comportamento diferente entre o meio da semana e o final ou início).

[6] ciclo sazonal semanal, mensal ou anual

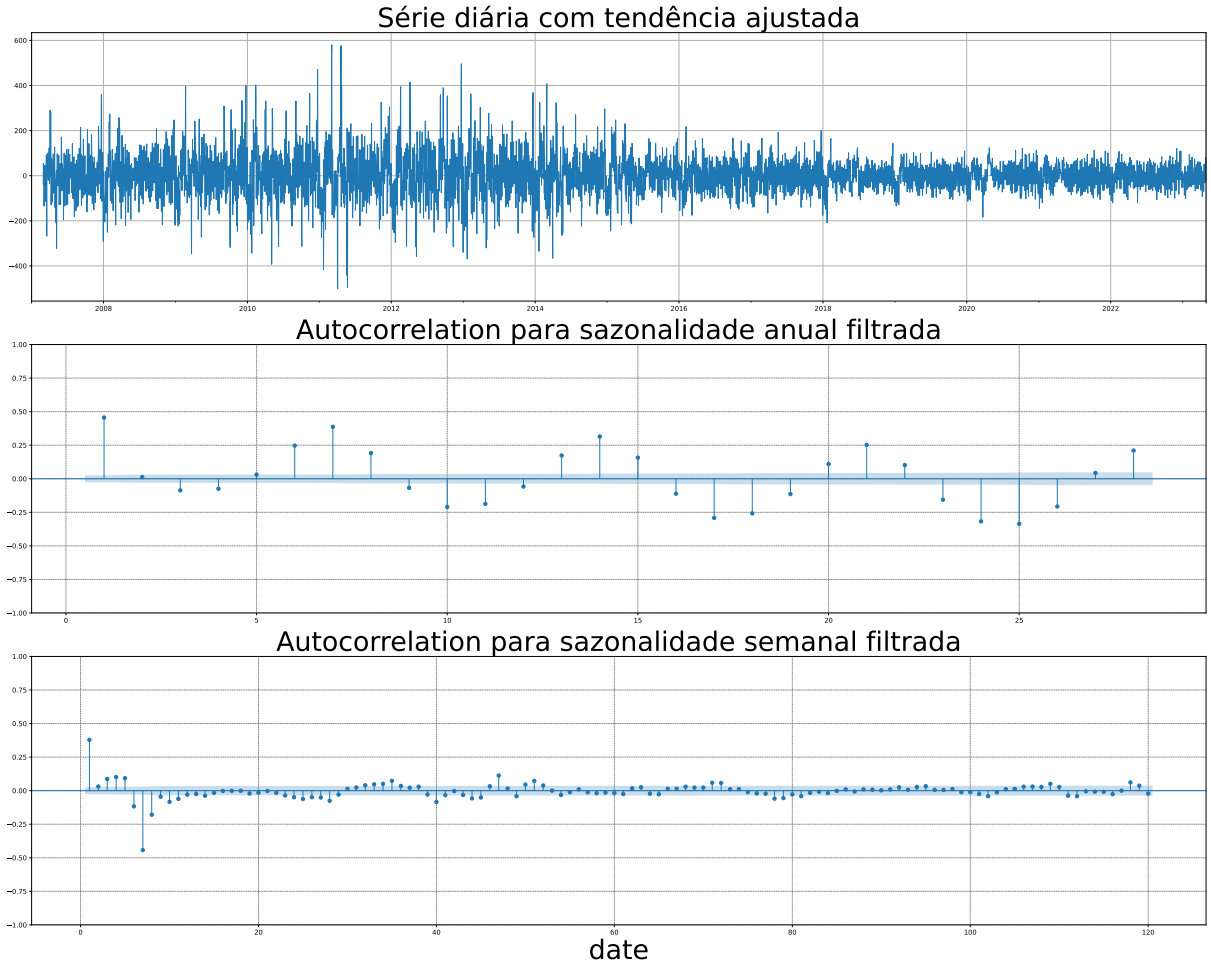

Figura 10 – Ciclo sazonal semanal e mensal para a série do número de acidentes no Brasil

Fonte: Autoria própria

Já na Figura 11, que mostra série mensal do número de acidentes, é notado ciclo sazonal anual evidente, caracterizado pela autocorrelação significativa no décimo segundo *lag*, ciclicamente. É importante ressaltar que claramente o padrão sazonal varia com o passar dos anos. Esse comportamento é evidenciado, tanto pelo gráfico da série diária sem tendência na Figura 10, quanto, também, pela intensidade da autocorrelação nos *lags* múltiplos de 12 na Figura 11. Em outras palavras, os dados apresentam diferentes padrões de variação ao longo do tempo.

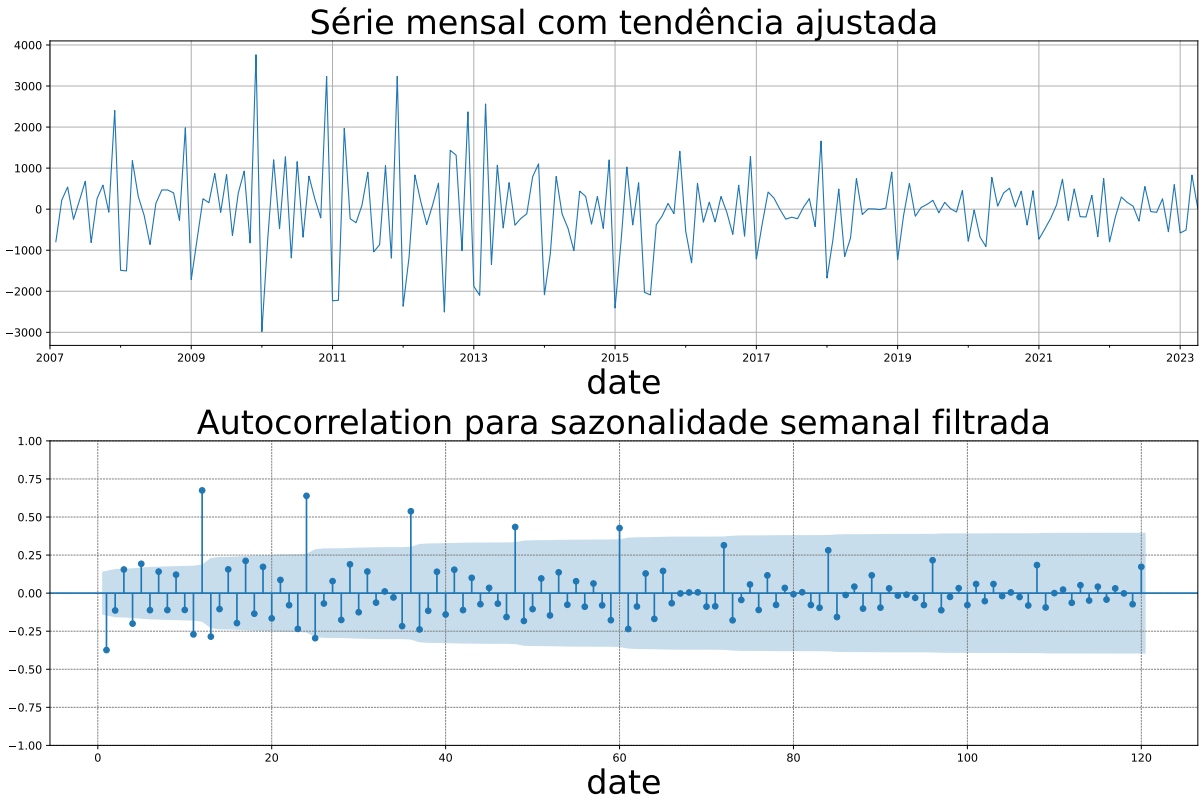

Figura 11 – Ciclo sazonal anual para a série do número de acidentes no Brasil

Fonte: Autoria própria

5 Resultados

Em conformidade com o objetivo deste trabalho, apresentado no Capítulo 1, será utilizada modelagem em séries temporais afim de fazer previsão com respeito às séries diárias de acidentes exploradas no Capítulo 4. Os resultados serão dispostos em três grandes partes.

A **Parte 1 - Ajuste no nível Regional com desagregação para UF** será desenvolvida respeitando a estrutura hierárquica de região para UF - isto é, o acidente ocorre em algum Estado que está inserido unicamente em uma região do Brasil. Nesta parte, então, serão ajustas modelos[1] de séries temporais no nível regional e serão feitas as respectivas previsões. Após isso, haverá a desagregação pelo *top-down*, descrito na subseção 3.8.1, para o nível Estadual, afim de obtermos, para cada UF no Brasil, as respectivas previsões para o número de acidentes em rodovias.

Como nota, segue que os modelos[2] serão ajustados conforme implementação em *Python* do pacote *Forecast* no R (HYNDMAN, R. J.; KHANDAKAR, 2008; HYNDMAN, R. et al., 2023). A implementação está documentada e centralizada em *Nixtla*[3]. A motivação da utilização deste meio é pelo fato da recomendação do Prof. Robin John Hyndman[4] em que é citada implementação como mais destaque dentro da linguagem. Isso, pois, é consideravelmente mais eficiente (computacionalmente), *optimized by using numba*[5], que outras alternativas.

Para a **Parte 2 - Ajuste no nível de UF com desagregação para BR** será considerada estrutura de desagregação respeitando dois atributos, um geográfico e outro de natureza, como vista na seção 3.2. Mais especificamente, será considerada estrutura de desagregação conforme à UF, em que o acidente ocorre, afim de chegar ao nível de rodovia. Assim, será possível prever, para cada rodovia seu respectivo número de acidentes nos próximos dias.

Ao final, para **Parte 3 - Comparativa**, tem-se o objetivo de comparar as previsões feitas nas duas partes anteriores no nível de UF. Isto é, será comparada as previsões feitas pelos modelos ajustados diretamente para a UF (**Parte 2 - Ajuste no nível de UF com desagregação para BR**), com as previsões feitas através do *Top-down* da **Parte 1 - Ajuste no nível Regional com desagregação para UF**.

Espera-se chegar em previsões para o número de acidentes (número bruto) nos próximos dias, não somente para cada UF (ao total 27), mas também para todas Rodovias

[1] sNAIVE, AutoARIMA, DHR, RAE e TBATS
[2] sNAIVE, AutoARIMA, DHR, RAE e TBATS
[3] https://nixtla.github.io/statsforecast/
[4] https://robjhyndman.com/hyndsight/python_time_series.html
[5] https://numba.readthedocs.io/en/stable/index.html

Federais documentadas nos dados abertos da PRF (ao total 208). Assim, será possível munir políticas públicas com possíveis abordagens preventivas, não somente remediativas, visando a redução do número de acidentes no Brasil

5.1 Discussão

Como vimos na Figura 10 e na Figura 11, a série diária do número de acidentes não possui um único ciclo sazonal. Na verdade foi observado que ela possui um ciclo semanal e anual conjuntamente. Antes de expor os resultados do trabalho propriamente, faz-se o questionamento do que é esperado com respeito ao ajuste e eficácia dos modelos sNAIVE, ARIMA, DHR, RAE e TBATS. Ou seja, é esperado que modelos que naturalmente não captam múltiplas sazonalidades percam em desempenho quando comparados com modelos que captam múltiplos ciclos sazonais, como os modelos vistos na seção 3.7 ou na seção 3.6.

Modelos ARIMA (SARIMA) e sNAIVE são restritos quanto a captação da sazonalidade. Isto é, eles somente vão conseguir tratar uma sazonalidade, em específico, pelo que foi visto na Figura 10, vamos utilizar modelos ARIMA com $m = 7$ (ciclo semanal). Afim de melhorar os resultados, dadas limitações do modelo, iremos adicionar ao ARIMA três variáveis indicadoras, a primeira de feriado, a segunda de feriado longo e uma terceira que ajuda o modelo a captar algum efeito referente ao feriado longo em dias nas extremidades. Todas as variáveis foram mensuradas externamente em relação ao banco de dados original, fazendo uso das informações referentes aos calendários estaduais, considerando feriados para cada unidade federativa (UF), e o calendário Nacional para considerar feriados federais.

O feriado longo será referenciado como um dia que consta como feriado e é uma sexta-feira. Já os dias que são afetados pela ocorrência de feriado longo são sextas-feiras quando segunda-feira é feriado, ou, também, nas quintas-feiras quando sexta-feira é feriado. Enquanto que em dias de feriado o cidadão médio tende a usar mais rodovias, ou em dias de feriado longo, em que o mesmo cidadão tende a viajar, os dias com efeito de feriado longo ajudam as pessoas que demandam viajar, visto que elas podem aproveitar a janela de oportunidade pós jornada de trabalho.

Então, espera-se que haja, pelas variáveis indicadoras, influência positiva no número de acidentes quando há marcação no dia do calendário. É importante indicar, ao modelo, informações quanto ao calendário, pois, naturalmente o número de acidentes depende do número de pessoas dirigindo nas rodovias e esse número sofre com a indicação de feriados, feriados longos ou ainda dias com efeito de feriado longo. Essas variáveis serão consideradas para o ajuste dos modelos RAE e DHR.

A avaliação por janela deslizante, vide Figura 12, serve para termos alguma ideia do que nos espera nos resultados. O método, como ilustrado na Figura 6, seguiu uma janela de 4 em 4 anos, que percorreu desde jan/07 até final de jan/23, com horizonte para previsão de 12

meses. As séries dos números (brutos) de acidentes no nível mais agregado são tratadas como diárias no ajuste e previsão, porém reindexadas com frequência mensal para suavização da curva da MAE na Figura 12. Ou seja, a avaliação no horizonte $h = 1$ seria considerando previsões para o número (bruto) de acidentes 1 mês a frente, assim por diante.

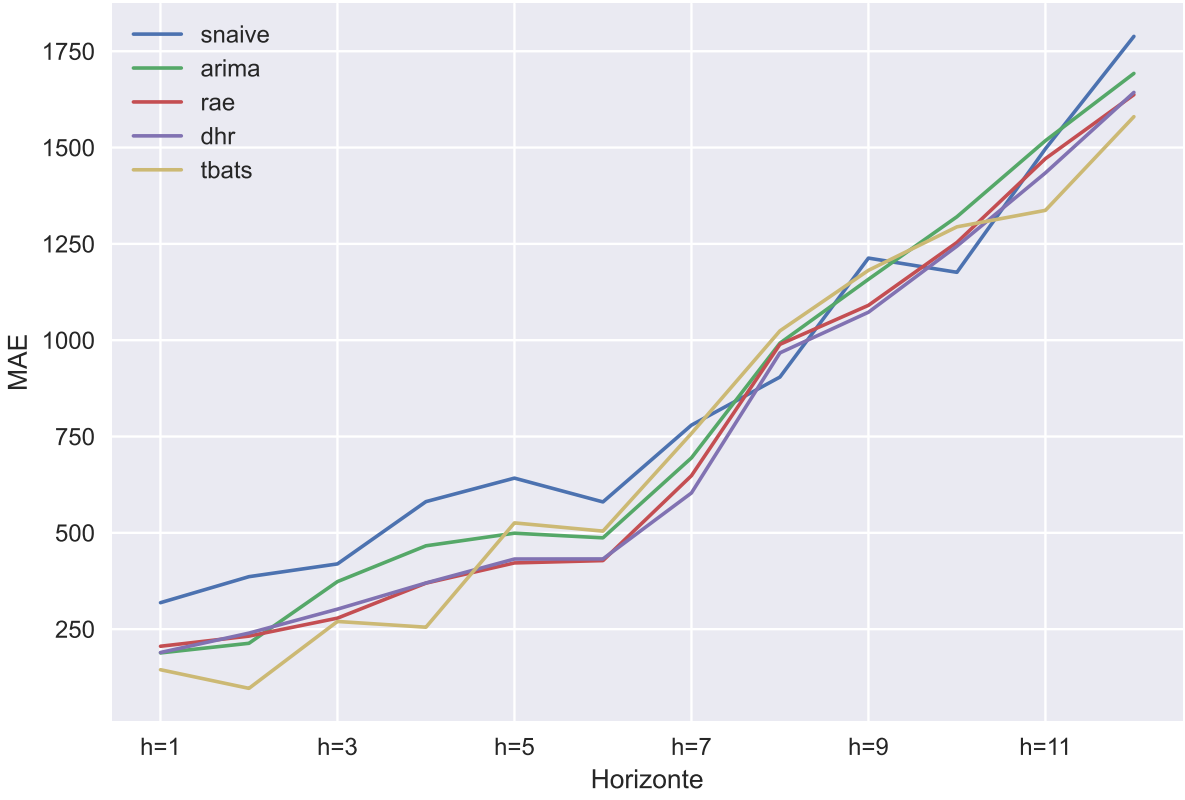

Figura 12 – Avaliação por janela deslizante no nível Federal

Fonte: Autoria própria

De acordo com a Figura 12, a curva que representa o modelo com melhores resultados com respeito à MAE, é tal que ela é dominada pelas outras, ou seja, fica a baixo das outras. Como resultado, pode-se concluir que os modelos mais sofisticados tem um ganho considerável no curto prazo (previsões até o sexto mês adiante), porém no longo prazo (após o sétimo mês) os modelos se equiparam com o sNAIVE. É importante ressaltar que as previsões à rigor estão sendo feitas para a série com frequência diária, ou seja, seis meses a frente representa, de fato, um horizonte de previsão bem grande com a aproximadamente $h = 6 * 30 = 180$. Para este trabalho, nos resultados, serão consideradas previsões apenas para os próximos 28 dias.

NOTA PARA VALIDAÇÃO CRUZADA NESTE TRABALHO

Tanto para a **parte 1 - de Região para UF**, quanto para **Parte 2 - Ajuste no nível de UF com desagregação para BR**, a validação cruzada será feita pela avaliação por janela deslizante, percorrendo desde jan/07 até final de jan/23, com janela de 4 em 4 anos e

horizonte de previsão de 28 dias. Sempre considerando séries diárias e as três variáveis explanatórias varrendo seus índices para comparar os modelos sNAIVE, ARIMA, RAE, DHR e TBATS

5.2 Parte 1 - Ajuste no nível Regional com desagregação para UF

Para a **Parte 1 - Ajuste no nível Regional com desagregação para UF**, serão ajustados os modelos para o nível regional e haverá a desagregação para o nível de Estado (UF). A estratégia é fazer uma validação cruzada para escolher o melhor modelo, fixada região, (com respeito à algum critério), entre os modelos sNAIVE, AutoARIMA, DHR, RAE e TBATS. Ao fazer para todas regiões, bastaria desagregar as respectivas previsões (para o número bruto de acidentes) por *top-down* para chegar no nível Estadual.

Tabela 1 – Resumo da validação cruzada com respeito à MAE pela abordagem na Parte 1

Região	TBATS	DHR	RAE	ARIMA	sNAIVE
NORTE	**3.569**	4.283	4.146	4.303	4.812
NORDESTE	9.382	**9.128**	9.681	9.728	11.071
CENTRO_OESTE	**5.406**	6.251	6.580	6.585	6.696
SUDESTE	**19.964**	21.877	21.974	22.023	25.411
SUL	15.057	**15.042**	15.668	15.880	17.125

Fonte: Autoria propria

Nota: RAE significa *Regression with ARIMA errors*

Ou seja, para cada região, selecionou-se o modelo mais efetivo (com menor MAE) dentro da avaliação. Então é feita desagregação, e, ao final, chega-se às previsões no nível estadual. Veja Tabela 1 com o resumo do processo.

O ajuste é feito da seguinte forma, conforme o objetivo do trabalho, ajusta-se o modelo de jan/2007 até o final de jan/2023, afim de prever, para os próximos 28 dias (fev/2023) os respectivos números de acidentes. Assim, sendo possível elencar Estados com destaque no número de acidentes. Veja Tabela 2 com o resumo do processo.

5.3 Parte 2 - Ajuste no nível de UF com desagregação para BR

Para a **Parte 2 - Ajuste no nível de UF com desagregação para BR**, serão ajustados os modelos para o nível de Estado e haverá a desagregação para o nível de rodovia (BR). A

Tabela 2 – Resumo das previsões pela abordagem na Parte 1

UF	Previsão fev/2023	Realizado fev/2023
MG	620.893	591
SC	568.089	554
PR	546.749	512
RS	365.466	328
RJ	352.216	396
SP	325.335	339
BA	263.019	254
GO	246.1	234

Fonte: Autoria propria

estratégia é fazer uma validação cruzada para escolher o melhor modelo, fixada UF, (com respeito à algum critério), entre os modelos sNAIVE, AutoARIMA, DHR, RAE e TBATS. Ao fazer para todos Estados, bastaria desagregar as respectivas previsões por *top-down* para chegar no nível mais desagregado possível, nível de rodovia.

Ou seja, para cada UF, selecionou-se o modelo que trouxe as melhores previsões dentro da avaliação. Então é feita desagregação, pelo método visto na subseção 3.8.1, para chegar nas previsões no nível de rodovia, após o ajuste do modelo escolhido à UF. Veja Tabela 3 com o resumo do processo.

O ajuste é feito da seguinte forma, conforme o objetivo do trabalho, ajusta-se o modelo de jan/2007 até o final de jan/2023, afim de prever, para os próximos 28 dias (fev/2023) os respectivos números de acidentes. Assim, será possível elencar rodovias com destaque no número de acidentes. Veja Tabela 4 com o resumo do processo.

5.4 Parte 3 - Comparativa

Ao final, para **Parte 3 - Comparativa**, tem-se o objetivo de comparar as previsões feitas nas duas partes anteriores no nível de UF. Isto é, será comparada as previsões feitas pelos modelos ajustados diretamente para a UF (**Parte 2 - Ajuste no nível de UF com desagregação para BR**), com as previsões feitas através do *Top-down* da **Parte 1 - Ajuste no nível Regional com desagregação para UF**. Faz-se o comparativo através do sMAPE e da MASE, vistos na seção 3.9, pela Tabela 5.

Os resultados observados conversam com Rob J. Hyndman, Lee e Wang (2016), em que é dito que as séries mais desagregadas geralmente têm um alto grau de volatilidade - portanto, são difíceis para prever - enquanto que a série temporal mais agregada é geralmente suave e menos barulhenta - portanto, mais fácil de prever. Veja Figura 13.

Tabela 3 – Resumo da validação cruzada com respeito à MAE pela abordagem na Parte 2

UF	TBATS	DHR	RAE	AutoARIMA	sNAIVE
AC	**0.723**	0.750	0.758	0.756	0.92
AL	**1.462**	1.519	1.557	1.547	1.678
AM	**0.431**	0.476	0.504	0.691	0.580
AP	0.573	**0.562**	0.613	0.6182	0.652
BA	4.129	3.958	**3.945**	4.0156	5.0625
CE	2.570	2.528	**2.526**	2.528	3.768
DF	1.40	**1.362**	1.381	1.392	1.687
ES	4.069	**3.729**	3.924	3.743	5.089
GO	*3.77*	3.878	4.048	4.024	4.589
MA	1.89	*1.871*	1.932	1.909	2.741
MG	14.370	13.325	**11.760**	11.873	17.20
MS	2.054	**2.014**	2.071	2.077	2.580
MT	2.269	**2.231**	2.357	2.325	3.446
PA	2.070	2.248	**2.200**	2.207	3.383
PB	2.196	**2.127**	2.235	2.326	3.285
PE	**3.229**	3.533	3.618	3.530	4.848
PI	**1.990**	2.011	2.076	2.053	2.3125
PR	7.779	7.468	**7.379**	7.985	8.294
RJ	7.244	6.775	**6.528**	6.545	7.785
RN	2.172	**2.125**	2.261	2.241	2.651
RO	**1.857**	2.136	2.129	2.124	2.348
RR	0.615	**0.592**	0.645	0.689	0.714
RS	5.385	**5.346**	5.478	5.648	5.76
SC	7.253	**6.454**	7.305	7.282	8.830
SE	1.488	**1.446**	1.481	1.507	1.633
SP	5.451	**5.240**	5.406	5.533	7.080
TO	**1.132**	1.167	1.186	1.133	1.660

Fonte: Autoria própria

Nota: RAE significa *Regression with ARIMA errors*

Tabela 4 – Resumo das previsões pela abordagem na Parte 2

BR	Previsão fev/2023	Realizado fev/2023
101	845.683	908
116	758.109	710
40	249.921	241
381	243.474	226
153	184.904	172

Fonte: Autoria própria

Tabela 5 – Comparação das duas abordagens nas Partes 1 e 2

Abordagem	MASE	sMAPE (%)
Ajuste e previsão diretamente	0.426	11.11
Ajuste e previsão pelo *top-down*	**0.394**	**10.24**

Fonte: Autoria propria

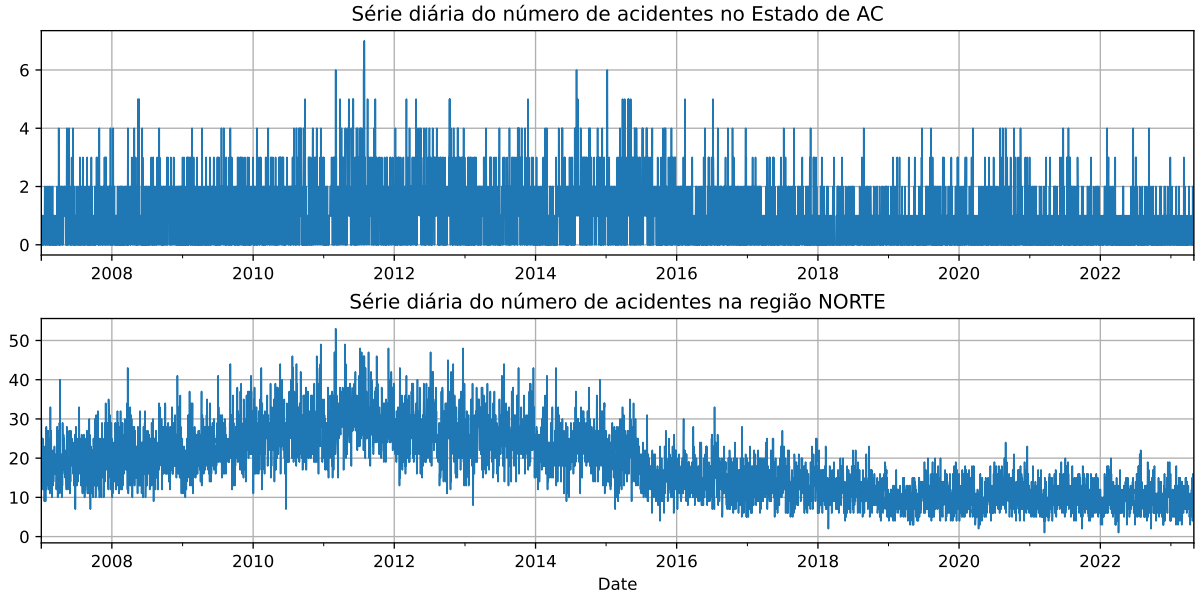

Figura 13 – Grau de volatilidade em diferentes níveis Hierárquicos

Fonte: Autoria própria

6 Conclusões

Ao final do estudo, pode-se afirmar que os modelos que captam múltiplas sazonalidades, como a DHR ou TBATS se sobressaíram em detrimento do ARIMA ou sNAIVE. Ao longo do trabalho, mas principalmente na Capítulo 4, foi evidenciado que as séries em questão possuíam ciclos sazonais semanais e anuais. Portanto o motivo pelo qual o modelo TBATS, dentro do contexto de Livera, Rob J. Hyndman e Snyder (2011), e a DHR, que utiliza o princípio de que uma combinação de funções senos e cossenos pode aproximar qualquer função periódica, ganharam destaque, foi, muito pelas frequência diárias das séries e suas múltiplas sazonalidades.

Também foi observada melhora nos resultados a partir da adição de variáveis explanatórias ao modelo ARIMA, com a *Regression with ARIMA errors* em detrimento do ARIMA usual, visto na Figura 12. Concluímos o trabalho com resultados que estão de acordo com os objetivos propostos no Capítulo 1.

6.1 Possibilidade de pesquisas futuras

Após a análise exploratória dos dados, no Capítulo 4, ficam nítidos os diferentes padrões de variação ao longo do tempo para o número de acidentes, visto que as estradas, desde 2007, sofreram melhorias e ganharam qualidade (os dados mostram exatamente isso, com sua média móvel caindo ao longo dos anos). Neste caso como a amplitude sazonal varia com o tempo, a transformação de Box-Cox (BOX; COX, 1964) pode ter um grande impacto para os modelos com séries harmônicas.

Além disso, também poderia ter sido feito um estudo para selecionar variáveis, com possíveis transformações, considerando *lags* (PANKRATZ, 1992) e amortizações. Como nota, dada robustez dos dados abertos da PRF, a seleção de variáveis ganha ainda mais importância.

Referências

BOX, G. E. P.; COX, D. R. An Analysis of Transformations. **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 26, n. 2, p. 211–243, 1964. DOI: https://doi.org/10.1111/j.2517-6161.1964.tb00553.x. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1964.tb00553.x. Disponível em: <https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1964.tb00553.x>. Citado nas pp. 30, 48.

FIORUCCI, J. A. Análise de Séries Temporais. **Youtube**, 2021. Disponível em: <https://www.youtube.com/playlist?list=PLNcS6_7VaJ3lWCs6cCXBAOAJVOItrACIY>. Acesso em: 16 ago. 2023. Citado nas pp. 17, 34, 35.

GROSS, C. W.; SOHL, J. E. Disaggregation methods to expedite product line forecasting. **Journal of Forecasting**, v. 9, n. 3, p. 233–254, 1990. DOI: https://doi.org/10.1002/for.3980090304. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/for.3980090304. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/for.3980090304>. Citado na p. 33.

HARVEY, A. C. **Forecasting, Structural Time Series Models and the Kalman Filter**. Cambridge University Press, 1990. DOI: 10.1017/CBO9781107049994. Citado na p. 31.

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**. 2023. R package version 8.21. Disponível em: <https://pkg.robjhyndman.com/forecast/>. Citado na p. 41.

HYNDMAN, R. J.; ATHANASOPOULOS, G. **Forecasting: Principles and Practice**. 2. ed. Australia: Otexts, 2018. Disponível em: <https://otexts.com/fpp2/>. Acesso em: 1 jun. 2022. Citado nas pp. 24–29.

HYNDMAN, R. J.; KHANDAKAR, Y. Automatic time series forecasting: the forecast package for R. **Journal of Statistical Software**, v. 26, n. 3, p. 1–22, 2008. DOI: 10.18637/jss.v027.i03. Citado na p. 41.

HYNDMAN, R. J.; KOEHLER, A. B. Another look at measures of forecast accuracy. **International Journal of Forecasting**, v. 22, n. 4, p. 679–688, 2006. ISSN 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2006.03.001. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0169207006000239>. Citado na p. 34.

HYNDMAN, R. J.; LEE, A. J.; WANG, E. Fast computation of reconciled forecasts for hierarchical and grouped time series. **Computational Statistics and Data Analysis**, v. 97, p. 16–32, 2016. ISSN 0167-9473. DOI: https://doi.org/10.1016/j.csda.2015.11.007. Disponível em: <https://www.sciencedirect.com/science/article/pii/S016794731500290X>. Citado nas pp. 15, 23, 37, 45.

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**, Taylor e Francis, v. 106, n. 496, p. 1513–1527, 2011. DOI: 10.1198/jasa.2011.tm09771. eprint: https://doi.org/10.1198/jasa.2011.tm09771. Disponível em: <https://doi.org/10.1198/jasa.2011.tm09771>. Citado nas pp. 29, 30, 32, 48.

MORETTIN, P.; TOLOI, C. **Análise de séries temporais: modelos lineares univariados**. BLUCHER., 2018. ISBN 9788521213529. Disponível em: <https://books.google.com.br/books?id=UwC5DwAAQBAJ>. Citado na p. 17.

PANKRATZ, A. **Forecasting with Dynamic Regression Models**. 1. ed. Australia: Wiley-Interscience, 1992. ISBN 978-1853105845. Disponível em: <https://www.amazon.com.br/dp/0471615285?geniuslink=true>. Acesso em: 11 jul. 2023. Citado nas pp. 29, 48.

TAYLOR, J. W. Short-term electricity demand forecasting using double seasonal exponential smoothing. **Journal of the Operational Research Society**, Taylor e Francis, v. 54, n. 8, p. 799–805, 2003. DOI: 10.1057/palgrave.jors.2601589. eprint: https://doi.org/10.1057/palgrave.jors.2601589. Disponível em: <https://doi.org/10.1057/palgrave.jors.2601589>. Citado na p. 30.

WEST, M.; HARRISON, J. **Bayesian Forecasting and Dynamic Models (2nd Ed.)** Berlin, Heidelberg: Springer-Verlag, 1997. ISBN 0387947256. Citado na p. 31.

WICKRAMASURIYA, S. L.; ATHANASOPOULOS, G.; HYNDMAN, R. J. Forecasting hierarchical and grouped time series through trace minimization. **The Bulletin of the Center for Children's Books**, 2015. Citado na p. 14.

Apêndices

APÊNDICE A – Códigos de programação

A.1 Códigos *Python* para buscar os dados

Código A.1 – Web scraping

```
1
2  ### .py para alcançar, via web scraping,\
3  ### os dados, da PRF, que serao utilizados no TCC
4
5  from bs4 import BeautifulSoup
6  import requests
7  from pyunpack import Archive
8  import zipfile
9
10 ##############################
11 ### Get links for download ###
12 ##############################
13 #INPUT_PATH = "C:/Users/u00378/Desktop/tcc_est_unb"
14 INPUT_PATH = "C:/Users/Igor/Desktop/TCC"
15 url = 'https://www.gov.br/prf/pt-br/acesso-a-informacao/\
16     dados-abertos/dados-abertos-acidentes'
17
18 agent = "Mozilla/5.0 (Windows NT 10.0; Windows; x64)
      AppleWebKit/537.36 (KHTML, like Gecko) Chrome/103.0.5060.114
      Safari/537.36"
19 # Making a GET request
20 # , headers={"User-Agent": agent} in r = requests.get(url)
21 r = requests.get(url, headers={"User-Agent": agent})
22
23 # check status code for response received
24 # success code - 200
25 print(f"Acesso ao site liberado para web scraping" if
      r.status_code == 200 else f"Acesso negado ao site para web
      scraping")
26
27 # Parsing the HTML
28 soup = BeautifulSoup(r.content, 'html.parser') #[2484:2569]
29 s = soup.find_all('a', class_= 'external-link')
30
31 links = []
32 for i in range(len(s)):
33     links.append(s[i]['href'])
34
35 links = links[4:22]
```

```
36 links.remove('https://arquivos.prf.gov.\
37              br/arquivos/index.php/s/n1T3lymvIdDOzzb')
38
39 ids = [i[32:65] for i in links]
40 urls = [f'https://drive.google.com/u/0/uc?id={i}&export=download'
      for i in ids]
41
42 ############################
43 ### Download .ZIP files ###
44 ############################
45 print('BEGINNING OF DOWNLOADS...')
46 name = 2023
47 for i in range(len(urls)):
48     url = urls[i]
49     response = requests.get(url, stream=True)
50     if response.status_code == 200:
51         with open(f'{INPUT_PATH}/dados/zips/{name}.zip', 'wb') as
              file:
52             file.write(response.content)
53         print(f'Arquivo baixado com sucesso: {name}: {url}')
54     else:
55         print(f'Falha ao baixar arquivo. Código de resposta:
              {response.status_code}')
56         print(f'Erro no download de: {name} {url}')
57         print("DOWNLOAD FAILED")
58         break
59     name = name -1
60
61 print('END OF DOWNLOADS')
62
63 ##############################
64 ### Extracting .ZIP Archives ###
65 ##############################
66 print('BEGINNING OF EXTRACTION...')
67 name = 2023
68 for i in range(len(urls)):
69     try:
70         #with
              zipfile.ZipFile(f'{INPUT_PATH}/dados/zips/{name}.zip',
              'r') as zip_ref:
71         #    zip_ref.extractall(f'{INPUT_PATH}/dados')
72         Archive(f'{INPUT_PATH}/dados/zips/{name}.zip')\
73             .extractall(f'{INPUT_PATH}/dados')
74         print(f'O arquivo {name} foi extraído com sucesso.')
75     except:
76         print(f'O arquivo {name} não foi encontrado.')
77     name = name -1
78
79 print('END OF EXTRACTION')
```

Código A.2 – Data wrangling

```
### DATA WRANGLING

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import os
import sys

np.set_printoptions(threshold=sys.maxsize)

#INPUT_PATH = "C:/Users/u00378/Desktop/tcc_est_unb"
INPUT_PATH = "C:/Users/Igor/Desktop/TCC"
year = 2007
df = pd.read_csv(f"{INPUT_PATH}/dados/datatran{year}.csv",\
                 encoding='latin1', on_bad_lines='skip', sep =';',\
                 dtype={'br':'object','km':'object'},
                    na_values='(null)')
for i in range(1, 17):
        year +=1
        df1 =
           pd.read_csv(f"{INPUT_PATH}/dados/datatran{year}.csv",\
                          encoding='latin1', on_bad_lines='skip',
                             sep =';',\
                          dtype={'br':'object','km':'object'},
                             na_values='(null)')
        #df1.columns=df.columns
        df = pd.concat([df,df1], ignore_index=True)

df['data_inversa'] = pd.to_datetime(df['data_inversa'], format =
   'mixed', dayfirst=True)

df.to_pickle(f'{INPUT_PATH}/dados/tcc_data.pkl')
print("Dados salvos com sucesso!")
```