UNIVERSIDADE FEDERAL DO PARANÁ

AMAURI CUNHA SOARES

OTIMIZAÇÃO DA QUANTIDADE DE PERITOS CRIMINAIS NAS REGIONAIS DO PARANÁ UTILIZANDO ANÁLISE DE DADOS E ALGORITMOS PREDITIVOS DE SÉRIES TEMPORAIS

CURITIBA
2024

AMAURI CUNHA SOARES

OTIMIZAÇÃO DA QUANTIDADE DE PERITOS CRIMINAIS NAS REGIONAIS DO PARANÁ UTILIZANDO ANÁLISE DE DADOS E ALGORITMOS PREDITIVOS DE SÉRIES TEMPORAIS

Dissertação apresentada ao curso de Pós-Graduação em Métodos Numéricos em Engenharia, Setor de Ciências Exatas e Setor de Tecnologia, da Universidade Federal do Paraná, como requisito parcial à obtenção do título de Mestre em Métodos Numéricos em Engenharia.

Orientador: Prof. Dr. José Eduardo Pécora Junior

CURITIBA
2024

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
UNIVERSIDADE FEDERAL DO PARANÁ
SISTEMA DE BIBLIOTECAS – BIBLIOTECA DE CIÊNCIA E TECNOLOGIA

Soares, Amauri Cunha
Otimização da quantidade de peritos criminais nas regionais do Paraná utilizando análise de dados e algoritmos preditivos de séries temporais / Amauri Cunha Soares. – Curitiba, 2023.
1 recurso on-line : PDF.

Dissertação (Mestrado) - Universidade Federal do Paraná, Setor de Ciências Exatas, Programa de Pós-Graduação em Métodos Numéricos em Engenharia.

Orientador: José Eduardo Pécora Junior

1. Análise de séries temporais - Processamento de dados. 2. Pedição. 3. Algoritmos. 4. Peritos. 5. Segurança Pública (PR). I. Universidade Federal do Paraná. II. Programa de Pós-Graduação em Métodos Numéricos em Engenharia. III. Pécora Junior, José Eduardo. IV. Título.

Bibliotecário: Elias Barbosa da Silva CRB-9/1894

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação MÉTODOS NUMÉRICOS EM ENGENHARIA da Universidade Federal do Paraná foram convocados para realizar a arguição da Dissertação de Mestrado de **AMAURI CUNHA SOARES** intitulada: **OTIMIZACAO DA QUANTIDADE DE PERITOS CRIMINAIS NAS REGIONAIS DO PARANA UTILIZANDO ANALISE DE DADOS E ALGORITMOS PREDITIVOS DE SERIES TEMPORAIS**, sob orientação do Prof. Dr. JOSÉ EDUARDO PÉCORA JUNIOR, que após terem inquirido o aluno e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de mestre está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

Curitiba, 23 de Dezembro de 2023.

Assinatura Eletrônica
09/01/2024 17:10:46.0
JOSÉ EDUARDO PÉCORA JUNIOR
Presidente da Banca Examinadora

Assinatura Eletrônica
12/01/2024 12:28:41.0
LEONARDO SILVA DE LIMA
Avaliador Externo (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
09/01/2024 09:51:39.0
CASSIUS TADEU SCARPIN
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

RESUMO

O trabalho tem como objetivo a análise temporal preditiva dos atendimentos realizados pela Polícia Científica do Paraná no período de 2014 a 2023. A necessidade advém da busca pela compreensão da evolução dos atendimentos ao longo do tempo. Almeja-se orientar a alocação eficiente de recursos humanos e formulação de políticas públicas na área de segurança. Os dados utilizados sobre as mortes violentas provêm da Polícia Científica do Paraná. Uma comparação entre a biblioteca Prophet e Holt-Winters foi feita para saber qual dois era o mais adequado neste estudo.

Palavras-chave: Data science; Modelagem preditiva; Prophet; Séries temporais; Segurança Pública.

ABSTRACT

The objective of the work is the predictive temporal analysis of the services provided by the Scientific Police of Paraná in the period from 2014 to 2023. The need arises from the search for understanding the evolution of services over time. The aim is to guide the efficient allocation of human resources and the formulation of public policies in the area of security. The data used on violent deaths comes from the Scientific Police of Paraná. A comparison between the Prophet and Holt-Winters libraries was made to find out which two were most suitable for this study.

Keywords: Data science; Predictive modeling; Prophet; Time series; Public security.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

RN	Redes neurais;
APE	Erro percentual absoluto;
LSTM	Long Short-Term Memory;
WS	Manchas solares de Wolf;
CLD	Dados de linces canadenses;
BPUSD	Câmbio de libras esterlinas e dólares americanos;
ARIMA	Modelo autorregressivo integrado de médias móveis;
MSE	Erro médio quadrático;
MAD	Desvio absoluto médio;
RMSE	Raiz do erro médio quadrático;
STS	Séries temporais estruturais;
MAPE	Erro percentual absoluto médio;
GM	Modelo Grey;
MSPE	Erro percentual quadrático médio;
BPNN	Rede neural com retropropagação;
CNN	Convolutional Neural Network;
CLSTM	Método híbrido com RN convolucional (CNN) e LSTM;
MAE	Erro Médio Absoluto;
DNN	Rede neural profunda;
MFNN	RN com multi-filtros;
KELM	Kernel extreme learning machine;
nRMSE	RMSE normalizado;
nMAE	Erro absoluto médio normalizado;
R	Coeficiente de correlação;
ARDL-ECM	Modelo de correção de erro com atraso distribuído autorregressivo

AGRADECIMENTOS

Agradeço ao meu professor orientador pela paciência e compreensão durante esta jornada da minha vida. Também agradeço a todos envolvidos direta e indiretamente que me deram apoio durante a elaboração e revisão desta dissertação. Em especial, agradeço aos membros PPGMNE, que contribuíram a me tornar o ser humano que sou hoje.

“Os relatos espúrios que enganam os ingênuos são acessíveis.
As abordagens céticas são muito mais difíceis de encontrar.
O ceticismo não vende bem.”
(Carl Sagan)

SUMÁRIO

1 INTRODUÇÃO

A gestão pública é encarregada de buscar soluções para atender as demandas sociais de forma racional e eficiente, por meio da implantação de políticas públicas. A análise de dados e a utilização de processos analíticos podem auxiliar na compreensão dos padrões de ocorrência de crimes em uma cidade, permitindo o uso inteligente dos recursos do Estado em prol do desenvolvimento socioeconômico e da qualidade de vida da população.

Assim, esta pesquisa se justifica pela necessidade de atender às demandas sociais no campo da segurança pública, acompanhando as tendências de mortes violentas (MV) no Paraná.

1.1 A POLÍCIA CIENTÍFICA DO PARANÁ

Nesta seção, apresenta-se um resumo da linha do tempo da Polícia Científica do Paraná (PCP), com o objetivo de situar a instituição dentro da organização administrativa do Estado, como mostrado na Tabela 1.

Tabela 1 – Leis e decretos.

Ano	Evento
1892	A Lei Orgânica nº 15, de 21 de maio, institui a Repartição Central de Polícia, com o cargo de Médico da Polícia, subordinados à Secretaria de Negócios do Interior e Justiça.
1929	Regulamento Geral da Polícia Civil do Estado do Paraná; alterando o nome do Serviço de Medicina Legal para Departamento de Medicina Legal
1935	Decreto Estadual nº 790, de 16 de maio, criação do Instituto de Criminalística; A Lei nº 26, de 21 de outubro, organizou as três Secretarias de Estado, dentre elas a do Interior e Justiça, que obteve a responsabilidade sobre os serviços de Justiça Pública; Polícia Civile Militar
1942	O Decreto-lei nº 41, de 22 de junho, redefiniu a denominação desta Secretaria, passando a denominar-se Secretaria do Interior, Justiça e Segurança Pública
1962	Lei nº 4615, de 9 de julho, criação da Secretaria de Segurança Pública e institui o Instituto de Polícia Técnica e Instituto Médico Legal.
1974	Foi instituído o Estatuto da Polícia Civil, foram criadas as carreiras de Químico Legal e de Toxicologista.
2001	Emenda Constitucional Estadual nº 10, de 24 de outubro, desvinculação da Polícia Civil e criação da Polícia Científica do Paraná
2019	Lei n° 13.964/2019, de 24 de dezembro, Pacote Anti-crime proposto pelo Ministro Sérgio Moro

Fonte: O Autor.

A Polícia Científica do Paraná (PCP) é o órgão central de perícia oficial de natureza criminal e está subordinada à Secretaria de Estado de Segurança Pública do Paraná (SESP PR). Ela é composta pelo Instituto de Criminalística (IC) e Institutos Médico-Legal (IML). O Instituto de Criminalística é um órgão de natureza técnico-científica, cujas atividades são descritas na legislação penal vigente. Suas principais

funções incluem o exame pericial e a emissão de Laudo Pericial, realizados por Peritos Oficiais. O objetivo dessas ações é descrever e/ou apontar o modus operandi de um fato delituoso ou em suspeita de delito, fornecendo informações relevantes para as investigações criminais.

Recentemente, a estrutura do Instituto de Criminalística passou por uma significativa expansão, indo de 10 sedes para 18 unidades entre os anos de 2019 a 2021, majoritariamente devido ao ex-ministro da Justiça Sérgio Moro, que neste período propôs o Pacote Anticrime (Lei n° 13.964/2019) e a destinação de recursos para a Segurança Pública. Essas unidades estão localizadas nas cidades de Londrina, Maringá, Foz do Iguaçu, Cascavel, Guarapuava, Umuarama, Ponta Grossa, Paranaguá, Francisco Beltrão, Jacarezinho, União da Vitória, Paranavaí, Pato Branco, Toledo, Telêmaco Borba, Apucarana, Campo Mourão e Curitiba. O Instituto também participa de operações anuais, deslocando efetivos para apoiar investigações e atender demandas específicas.

Os Peritos Oficiais são servidores públicos que desempenham duas funções distintas na Polícia Científica: Perito Criminal e Médico Legista. Sua atuação é fundamental na área da perícia criminal, sendo responsáveis por uma variedade de exames e análises em diferentes áreas de conhecimento. Entre as principais áreas de atuação dos peritos oficiais no Paraná, de acordo com a PCP, estão os exames de acidente de trânsito, anatomopatológico, antropologia forense, audiovisual, balística, clínica médica, computação forense, contábil, crimes contra o meio ambiente, crimes contra o patrimônio, crimes contra a pessoa, desenho técnico, documentoscopia, engenharia legal, exame merceológico, genética molecular forense, hipnose forense, identificação veicular, informação, informação Técnica, nada Consta - DPVAT, necrópsia, odontologia legal, projetos e pareceres técnicos, química legal e toxicologia forense.

A atuação desses profissionais é de extrema importância para o sistema de justiça, pois suas análises são imparciais e fundamentadas cientificamente, auxiliando na tomada de decisões judiciais.. O Instituto Médico Legal é o responsável por realizar exames na área de Medicina Legal. Ele é encarregado de perícias em cadáveres, partes de corpos, ossadas completas e em pessoas, além de conduzir exames laboratoriais em patologia, toxicologia, química legal e sexologia forense.

As atividades do IML são essenciais para o esclarecimento de óbitos, identificação de vítimas, investigação de mortes violentas e análises relacionadas à medicina legal.

A Polícia Científica do Paraná desempenha um papel fundamental no âmbito da perícia criminal, fornecendo suporte técnico e científico para a investigação de crimes e a análise de situações complexas. A atuação dos peritos oficiais, bem como a estruturação do Instituto de Criminalística e do Instituto Médico Legal, são elementos essenciais para a promoção da justiça e a garantia da segurança pública no Estado do

Paraná.

A questão de pesquisa que orienta este estudo é: Como ferramentas de análise computacional podem ser utilizadas para prever a tendência das mortes violentas e dimensionar o efetivo necessário para atendimento da demanda a fim de orientar a gestão pública?

Este trabalho responderá esta questão para o caso específico da Polícia Científica do Paraná, utilizando previsões a partir de dados sobre mortes violentas para determinar a quantidade de peritos atuando na instituição.

1.2 OBJETIVO GERAL

Estimar o número de peritos necessário para atender com qualidade a carga de trabalho demandada nos próximos dez anos.

1.3 OBJETIVOS ESPECÍFICOS

a) Descrever uma análise histórica e geográfica dos dados referentes à mortes violentas disponibilizados pela PCP;

b) Tratar os dados obtidos para aplicar análises estatísticas, modelagens e heurísticas;

c) Aplicar diferentes métodos de modelagem e previsão aos dados analisados;

d) Determinar entre os métodos Prophet e Holt-Winter qual é o mais adequado para realizar previsões com a base de dados obtida;

e) Prever o número de peritos necessários para manter os níveis de qualidade e carga de trabalho na PCP para os próximos dez anos.

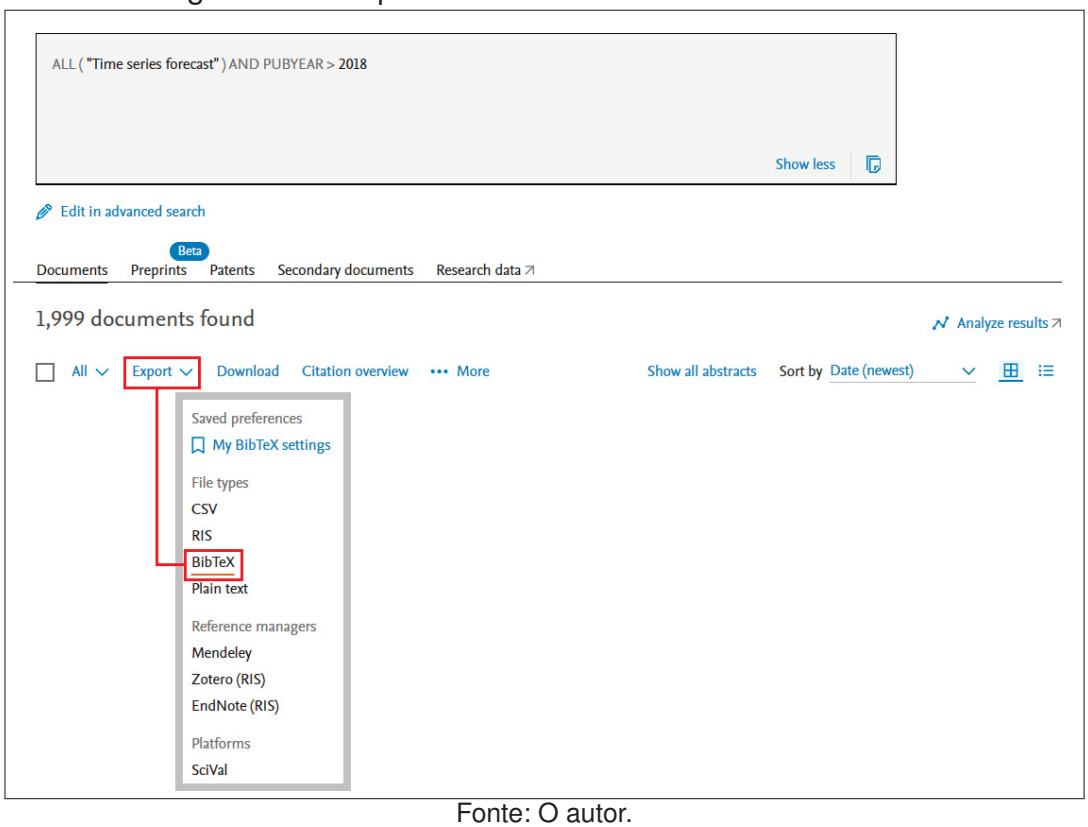
2 REVISÃO DE LITERATURA

Iniciou-se a revisão a partir de uma pesquisa bibliométrica, utilizando ferramentas de análise de dados para identificar os principais autores, documentos e fontes relevantes para a área de interesse do trabalho. Também foi necessário estabelecer uma revisão bibliográfica dos artigos relacionados aos métodos Prophet e outros mais convencionais, como de suavização exponencial e ARIMA. Por fim, foi realizada uma revisão dos artigos identificados na primeira etapa, buscando resumir os principais métodos e ferramentas utilizadas pelos autores.

2.1 PESQUISA BIBLIOMÉTRICA

Utilizou-se a biblioteca *bibliometrix* em R para avaliar quantitativamente a produção acadêmica relacionada a previsão com séries temporais, tendo em vista avaliar quais métodos vem sendo mais utilizados e objetivamente avaliar quais autores, livros e revistas são mais relevantes na análise de séries temporais. Para isso, iniciou-se na base de dados Scopus a partir de uma pesquisa avançada utilizando as palavras-chave "time series forecast” a serem buscadas em todos os campos (ou seja, no título do artigo, no título da fonte, língua, autor, editor, afiliação, resumo, palavras-chave, referências, etc.). Ao mesmo tempo, limitou-se a pesquisa aos artigos publicados a partir de 2019. Assim, foi utilizada a pesquisa por “ALL(“time series forecast”) AND PUBYEAR AFT 2019”. Obtiveram-se 1999 ocorrências (Figura 1), entre artigos, papers de conferências, artigos de revisão, livros (na íntegra ou capítulos), entre outros. Em seguida, os resultados foram exportados para o formato BibTeX de referências em um arquivo “scopus.bib”, selecionando-se a opção para exportar todos os metadados disponíveis.

Figura 1 – Pesquisa utilizando a biblioteca *bibliometrix*



Fonte: O autor.

Posteriormente, recorreu-se ao passo-a-passo dado por Aria e Cuccurullo (2017) na documentação do bibliometrix para importar, tratar e visualizar os dados referentes aos 1999 documentos pesquisados. Iniciou-se pela importação dos dados para um dataframe bibliográfico:

```
file <- "~/R/scopus.bib"
M <- convert2df
(file = file, dbsource = "scopus", format = "bibtex")
```

Na sequência, realizou-se a análise, gerando um objeto da classe bibliometrix "results" contendo estatísticas básicas como contagem de autores e artigos, distribuições de frequências, índices de colaboração entre países, etc. A função summary foi utilizada para agrupar as principais informações.

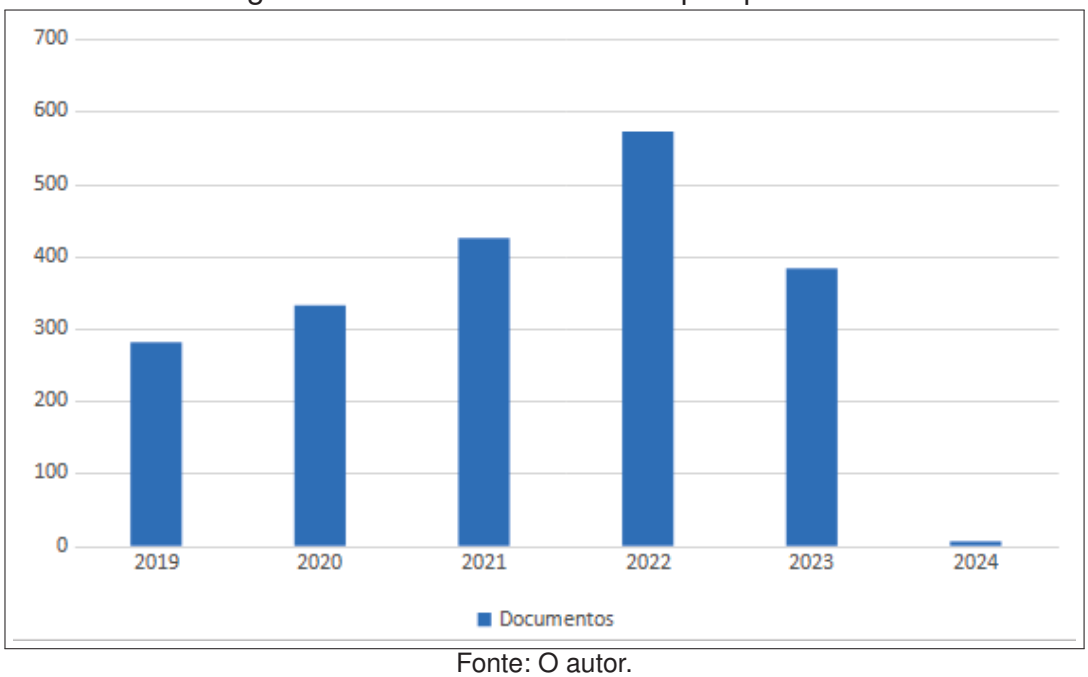
```
results <- biblioAnalysis(M, sep = ";")
options(width=100)
S <- summary(object = results, k = 10, pause = FALSE)
```

Outras funções foram utilizadas para gerar gráficos e analisar os dados, que serão explorados em detalhe na Subseção Resultados.

2.1.1 Resultados

Os documentos pesquisados foram publicados de 2019 a 2023, e alguns ainda serão publicados em 2024 mas já constam na base de dados da Scopus, conforme o gráfico da Figura 2. Alguns deles podem ter sido repetidos ou terem dados incompletos na busca, por isso totalizaram 1980 documentos (19 a menos que os dados de entrada).

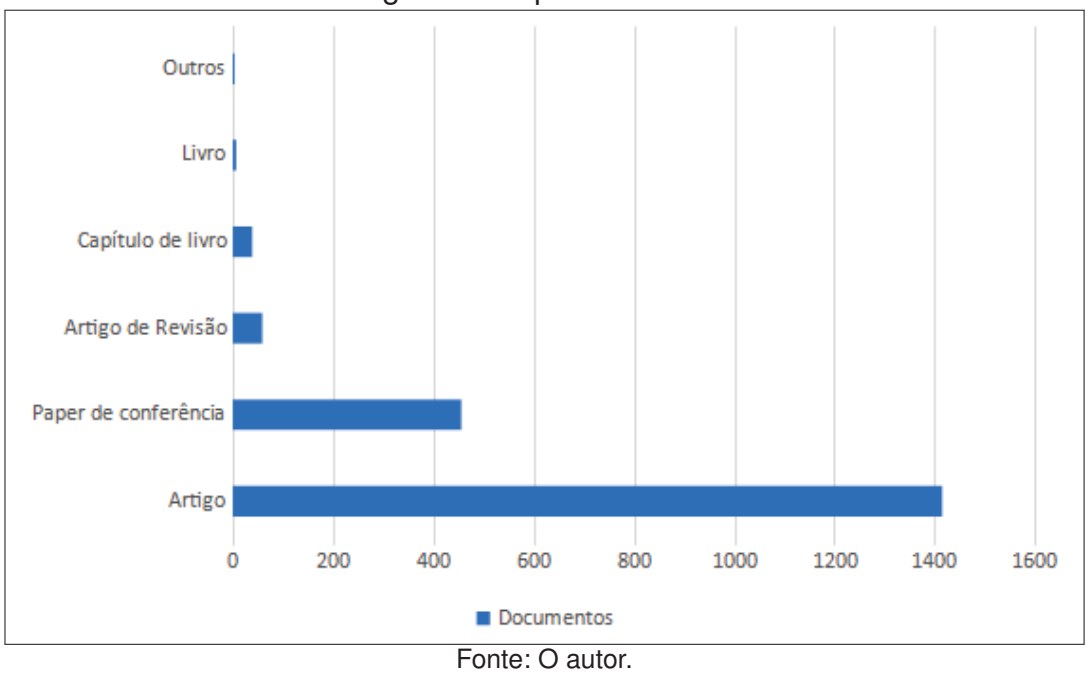
Figura 2 – Ano dos documentos pesquisados.



Fonte: O autor.

Estes documentos provêm de 1025 fontes (revistas, livros, conferências etc.), e possuem uma média de citações de 10,36 por documento. Eles estão estratificados conforme mostrado na Figura 3, com a maior parte sendo artigos e *papers* de conferência.

Figura 3 – Tipos de fontes.

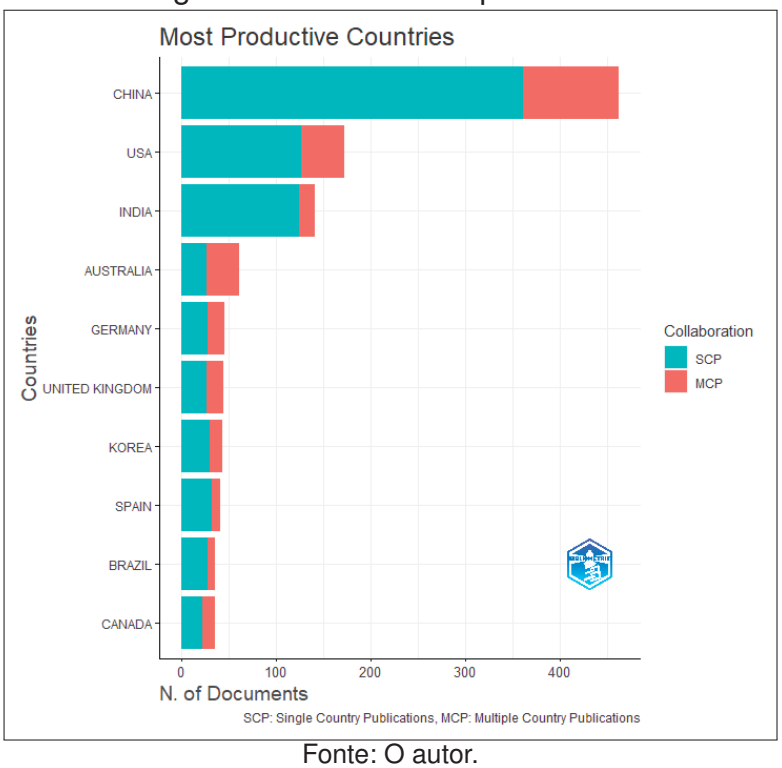


Fonte: O autor.

Foram identificados 5370 autores diferentes, sendo apenas 128 autores de artigos de autor único. A média é de 3,73 autores por artigo, sendo que 25,61% dos artigos de múltiplo autor são de cooperações internacionais.

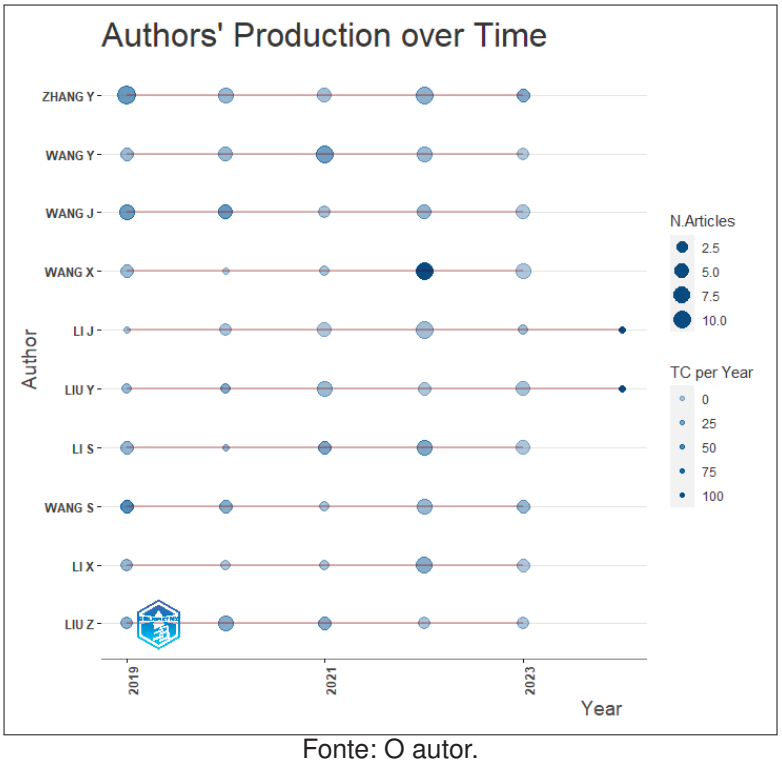
Os países mais produtivos foram China, Estados Unidos e Índia, com destaque para o Brasil na nona posição, conforme mostra o Gráfico da Figura 4. Na Figura 5, observa-se os autores mais produtivos e suas produções ao longo dos anos. Os países tiveram interações significativas, e a representação gráfica destas interações pode ser observada na Figura 6, onde cada país é uma bolha (com o diâmetro proporcional ao número de artigos publicados) e as linhas representam a conexão entre autores de diferentes países dentro de um mesmo artigo.

Figura 4 – Países mais produtivos.



Fonte: O autor.

Figura 5 – Autores mais produtivos.



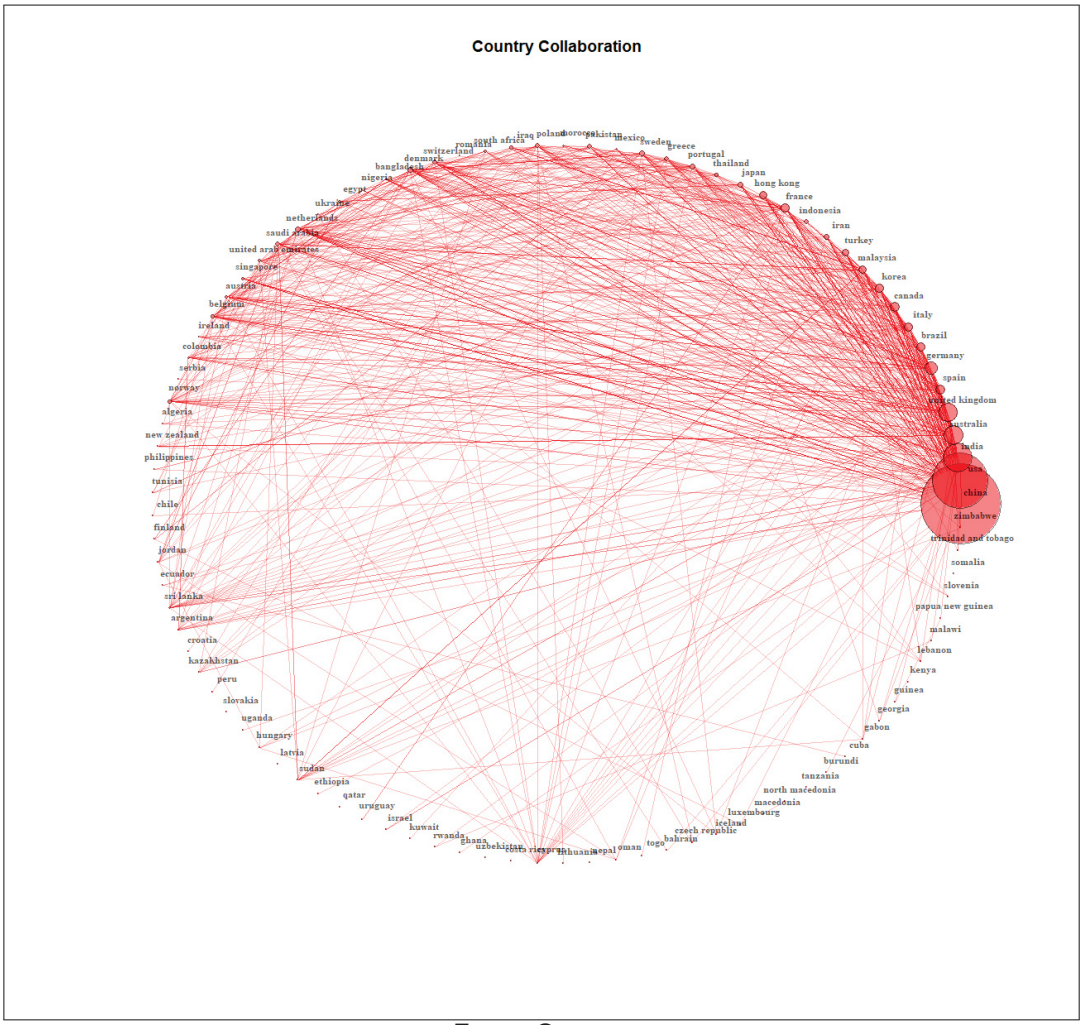
Fonte: O autor.

Os artigos mais citados dentre os pesquisados são quatro artigos de revisão e seis artigos científicos. Quatro deles são relacionados a energia solar, três sobre

turismo, um sobre preço de ações, um sobre economia verde e uma revisão teórica e prática sobre previsões. Estes artigos serão explorados na revisão bibliográfica.

Por outro lado, os documentos mais citados pelos artigos pesquisados foram três livros e sete artigos. Um dos livros trata previsões de modo mais generalizado, enquanto os outros dois focam em *deep learning* e análise de séries temporais, respectivamente. Os artigos foram publicados de 1996 a 2018 e tratam desde redes neurais até previsão de necessidades energéticas de países. Estes livros e artigos também serão explorados com maior detalhamento na revisão bibliográfica.

Figura 6 – Interações entre autores de diferente países



Fonte: O autor.

As fontes mais relevantes, considerando o número de artigos publicados em cada uma, são dadas na Tabela 5. Há um destaque nas áreas de energia e sustentabilidade, que se mostram mais uma vez relevantes.

Por fim, as palavras-chave mais relevantes foram: Forecasting, Machine Learning, Deep Learning, Time Series, Covid-19, Time Series Forecast, Time Series Forecasting, LSTM, ARIMA e Neural Networks. Observa-se o grande interesse em aprendizado de máquina, redes neurais e inteligência artificial para obter previsões

Tabela 2 – Principais revistas científicas identificadas na bibliometria.

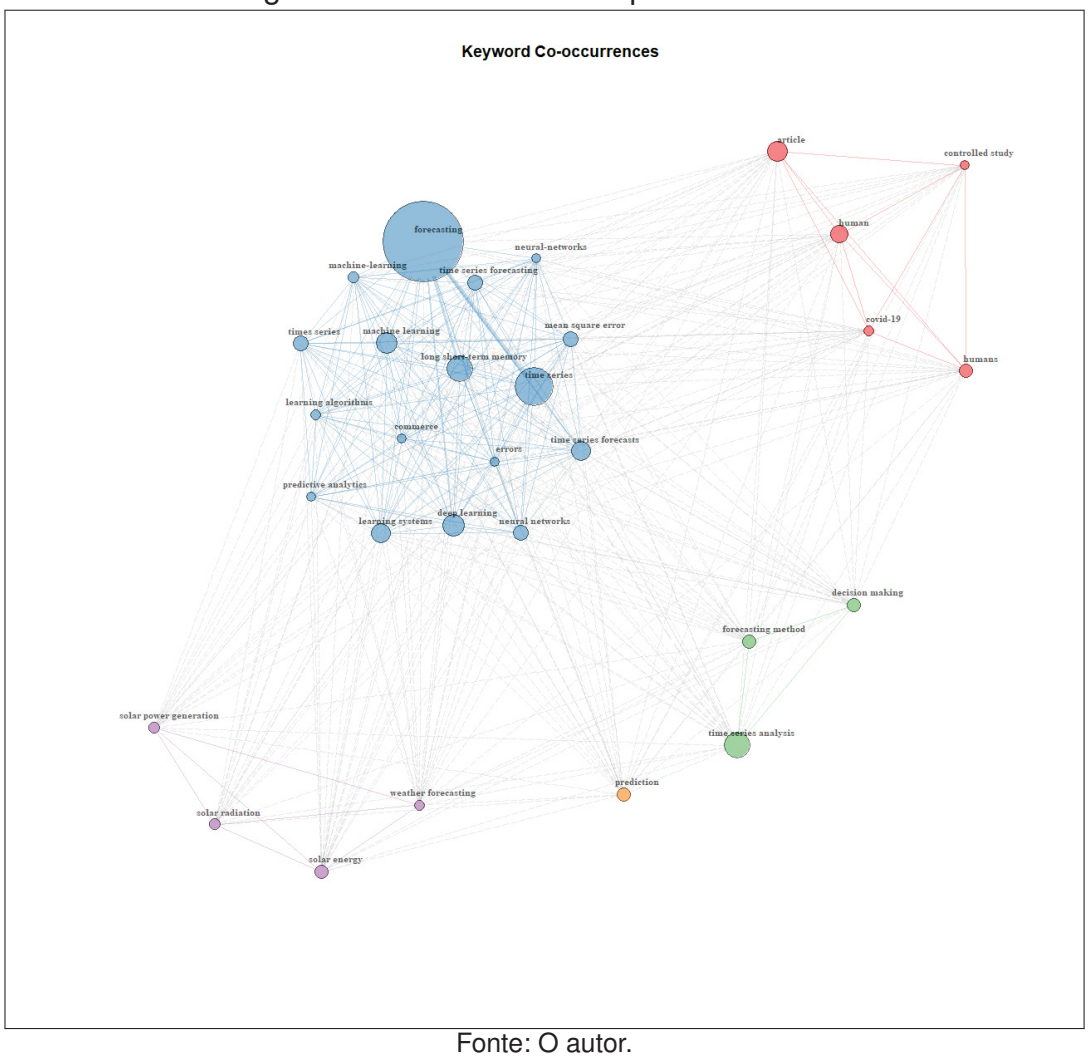
Fonte	**Nº de artigos**
Energies	52
IEEE Access	29
Sustainability (Switzerland)	27
Expert Systems with Applications	24
International Journal of Forecasting	24
Lecture Notes in Computer Science	24
Energy	22
ACM International Conference Proceeding Series	20
Applied Energy	20
International Journal of Environmental Research and Public Health	18

Fonte: O autor.

com séries temporais, sendo que LSTM (Long Short-Term Memory) é um dos métodos utilizados em redes neurais para aprendizado de máquina. Outro método muito recorrente é o ARIMA (modelo autorregressivo integrado de médias móveis). O assunto mais destacado na análise de palavras-chave foi Covid-19, o que é esperado visto que a janela temporal analisada foi exatamente a partir do início da pandemia, em 2019.

As palavras-chave foram estruturadas em um gráfico de rede (Figura 7), onde se pode observar cinco aglomerados principais. O primeiro e maior aglomerado (azul) reúne palavras-chave correlacionadas a previsões e inteligência artificial. Observa-se que estes métodos são os mais usuais atualmente. Outro aglomerado importante, em vermelho, está relacionado a estudos de saúde e ao Covid-19. Em roxo, as palavras-chave relacionadas a previsão do tempo e energia solar. Já em verde, métodos de previsão e tomada de decisões. Em laranja, a palavra-chave "*prediction*", que é quase um sinônimo de *forecast* (predição/previsão).

Figura 7 – Ocorrências das palavras-chave.



Fonte: O autor.

2.2 REVISÃO BIBLIOGRÁFICA

A partir da documentação do método *Prophet*, foram explorados os artigos fundamentais que embasaram as técnicas utilizadas pela biblioteca. Esta Seção está subdividida de acordo com o viés prioritário de cada artigo pesquisado. Inicialmente, explicita-se a teoria por trás da técnica Box-Jenkins e, depois, os métodos convencionais e suas aplicações.

2.2.1 Teoria

Uma técnica de modelagem autorregressiva muito popular é a chamada Box-Jenkins, descrita pelos autores inicialmente em 1970 em seu livro que teve sua quinta edição publicada em 2016 (BOX et al., 2015). Neste livro, os autores abordam a diferença entre processos determinísticos e estocásticos, e, dentre estes, os estacionários e não estacionários. Ou seja, os processos determinísticos seriam completamente explicados pelas variáveis de interesse, enquanto os estocásticos admitem que existem

flutuações e influências não previamente determinadas que levam a uma incerteza, que pode ser calculada levando a uma probabilidade de o valor futuro estar dentro de um limite determinado.

Modelos estacionários de processos estocásticos pressupõem que ao longo do tempo a média e variância são constantes, enquanto os não estacionários admitem variação. Os autores apontam os modelos autorregressivos como muito úteis na representação de certas séries temporais que ocorrem na prática. Eles explicam que neste modelo, o valor atual é expresso como um agregado finito linear de valores anteriores do processo mais um choque aleatório a_T. Através de manipulações algébricas, os autores mostram que essa série finita é equivalente a uma série infinita com $a_{T1}, a_{T2}, ...$ No caso dessa série ser convergente, o processo é estacionário; caso contrário, é não estacionário.

Os autores também definem, similarmente, os modelos de médias móveis, que seriam uma soma finita de parâmetros "peso" que multiplicam variáveis de choque aleatório anteriores a_{T-1}, a_{T-2}, ...; os autores esclarecem que o termo "média móvel" é um tanto errôneo, uma vez que a soma dos parâmetros "peso" não necessariamente é igual a unidade. Por fim, os autores unem estes modelos em um modelo ARMA (modelo autorregressivo de médias móveis), sendo ele dado pela soma dos parâmetros que multiplicam observações anteriores (autorregressivos) com os parâmetros peso multiplicando as variáveis de choque anteriores (médias móveis).

Os autores afirmam ainda que alguns processos não estacionários poderiam ser representados a partir de modelos ARIMA, ou seja, modelos mistos autorregressivos integrados de médias móveis, obtido teoricamente pela soma ou integração de processos ARMA estacionários. Hyndman (2002) explica que o método Box-Jenkins possui três estágios, acrescidos de um passo preliminar de preparo dos dados mais um passo posterior de previsão em si. Seriam eles:

1. Preparação dos dados, transformando-os com o objetivo de estabilizar a variância em séries em que esta varia com o nível; também é usual tomar a variação entre as leituras consecutivas da série (ou entre leituras separadas por um ano), processo chamado de differencing, o que, segundo Hyndman (2002), seria mais fácil de modelar que os dados originais;

2. Estimação dos parâmetros, selecionando os melhores valores para o ajuste dos dados;

3. Verificação do modelo, testando as suposições para garantir que o modelo é adequado. Caso contrário é necessário retornar ao passo 2;

4. Revisão, o objetivo final do processo, é facilmente computável uma vez que se tem um modelo selecionado, estimado e verificado.

2.2.2 Métodos e Aplicações

Hastie e Tibshirani (1987) propuseram aplicações para os modelos aditivos generalizados (GAMs) publicados por eles no ano anterior, indicando seu uso seja para análise de covariância não paramétrica ou regressão logística não paramétrica. Estes autores fundamentaram a base do conhecimento necessário para que Taylor e Letham (2017) pudessem desenvolver a biblioteca Prophet do Facebook, automatizando as previsões e permitindo a análise de um volume de dados muito maior de forma confiável. Os autores dessa ferramenta se basearam nas ideias de modelos de séries temporais estruturais, já apontados como tradicionais por Harvey e Peters (1990), em que o modelo se separa em componentes de tendência, sazonalidade, ciclos e uma componente irregular. O desafio em 1990 era definir métodos de computar os estimadores de verossimilhança máxima, de forma que alguns algoritmos foram testados para definir o mais eficiente computacionalmente.

Um método muito utilizado para modelar efeitos sazonais é o de suavização exponencial. O pacote forecast para R, proposto por Hyndman e Khandakar (2008), utiliza este método e o corresponde a um modelo de espaço de estados usando um framework proposto anteriormente (HYNDMAN et al., 2002). No trabalho de 2002, os autores utilizam o método automatizado e o comparam a soluções testadas nas competições de Makridakis (M-competition e M3-competition, vide Subseção 2.2.2.1), obtendo resultados comparáveis aos melhores métodos. Estas competições contribuíram fortemente para os avanços na área de previsões com séries temporais. Mais tarde, Livera, Hyndman e Snyder (2011) propõem um novo framework de espaço de estados incorporando transformações Box-Cox (algoritmos utilizados para normalizar dados cuja distribuição não seja normal), representações em séries de Fourier com coeficientes que variam com o tempo e correção de erro por ARMA (modelo autorregressivo de médias móveis).

2.2.2.1 Competições de Makridakis

As competições de Makridakis são uma série de competições que ocorrem desde 1982 para avaliar modelos de séries temporais e a acurácia dos métodos de previsão. A primeira competição, chamada M-competition, avaliou 15 métodos e 9 variações em 1001 séries temporais. Em 2000, a terceira edição (M3-competition) avaliou 24 métodos com dados de 3003 séries temporais. A sexta competição está em andamento, tendo seus resultados finais publicados até 2024, e está usando dados em tempo real de bolsas de valores para testar todos os mais importantes métodos de previsão, incluindo aprendizado de máquina, *deep learning* e métodos estatísticos.

Makridakis, Spiliotis e Assimakopoulos (2022) explicam que em suas primeiras edições, as competições concluíram que métodos mais complexos não necessaria-

mente traziam resultados melhores que modelos mais simplificados (na M-competition, o método de suavização exponencial superou o método Box-Jenkins, mais sofisticado, que na época era conhecido como “rei”). Os autores ainda complementam que a M2-competition, realizada em tempo real permitindo incorporar julgamentos pessoais dos analistas acerca dos dados e da economia, mostrou que a intervenção humana não melhorou as previsões estatísticas, ao contrário das expectativas. Ao longo do tempo, crescentemente se observou que a combinação de vários métodos gera melhores resultados, na média, em comparação a métodos isolados. Somente em 2018 (M4) que a balança se inverteu, tendo dois métodos complexos que combinavam estatística e aprendizado de máquina, superaram os métodos combinados mais simples. Na M5, em 2022, a dominância dos métodos de aprendizado de máquina se intensificou, com os 50 métodos mais bem posicionados pertencendo a esta classe.

2.3 RESULTADOS DA PESQUISA BIBLIOMÉTRICA

Conforme exposto na Seção 2.1, selecionou-se os artigos relacionados à análise de séries temporais que: a) foram mais citados dentro da pesquisa; e b) foram mais citados pelos artigos da pesquisa, incluindo aqueles que não estão dentro da pesquisa. Estes artigos científicos estão descritos com mais detalhes nas Subseções 2.3.1 e 2.3.2.

2.3.1 Artigos recentes mais relevantes

Os documentos mais citados dentre os selecionados, ou seja, aqueles que foram publicados nos últimos cinco anos, foram quatro artigos de revisão e seis artigos. Os quatro artigos de revisão envolveram previsões de demandas turísticas (SONG; QIU; PARK, 2019), o estado da arte em previsões relacionadas a produção de energia solar (AHMED et al., 2020), aprendizado de máquina e técnicas meta-heurísticas aplicadas a previsões de geração fotovoltaica (AKHTER et al., 2019), e um grande artigo de revisão sobre previsões em geral, explicitando teoria e prática e reunindo diversos modelos, técnicas, métodos e aplicações (PETROPOULOS et al., 2022).

Entre os sete outros artigos, há três na área de energia (GHIMIRE et al., 2019; HE et al., 2019; ZANG et al., 2020), sendo dois deles aplicados a previsão de radiação solar e um aplicado a expansão de energias renováveis em empresas chinesas. Os dois primeiros usam redes neurais com métodos híbridos, enquanto o último utiliza regressões com variáveis explicativas e modelagem com valor limítrofe (semelhante à modelagem com pontos de mudança explicada adiante, em Modelos Utilizados pelo Prophet). Dois outros artigos realizam previsões na área turística, sendo que Sun et al. (2019) usam aprendizado de máquina aplicado à dados de pesquisa em buscadores de internet para prever a chegada de turistas em Pequim; enquanto Zhang et al. (2021)

combinam métodos econométricos e qualitativos para projetar a recuperação turística em Hong Kong após a pandemia de Covid-19. Long, Lu e Cui (2019) aplicam redes neurais multi-filtros para prever movimentos no mercado de ações, obtendo resultados que, segundo os autores, são superiores a métodos tradicionais de aprendizado de máquinas, modelos estatísticos e redes neurais de estrutura singular.

2.3.2 Artigos históricos mais citados

Por outro lado, os artigos mais citados pelos documentos pesquisados, ou seja, de mais alto impacto historicamente, se dividiram em dois grandes grupos: primeiro, aqueles que tratam da metodologia (todos relacionados a redes neurais); segundo, aqueles que aplicam diversas metodologias a problemas de outras áreas (sendo a maioria absoluta relacionada à energia).

Hill, O'Connor e Remus (1996) comparam um método de previsão com base em redes neurais a seis métodos tradicionais estatísticos e gráficos utilizados na primeira competição de Makridakis (1982). Os autores concluíram que as redes neurais desempenharam melhor com dados trimestrais e mensais, enquanto empataram com os outros métodos com dados anuais. Isso se deu, segundo eles, pela boa capacidade das redes neurais de lidar com descontinuidades.

Hochreiter e Schmidhuber (1997) desenvolveram o método LSTM (long short-term memory, memória longa de curto prazo), tendo em vista a alta complexidade e longos tempos de computação com os métodos anteriores. Segundo os autores, o LSTM tem complexidade O(1) para cada iteração temporal, sendo então muito mais rápido e também mais eficiente, gerando bons resultados com mais recorrência do que os outros métodos a que foi comparado. Este artigo foi tão influente que o método LSTM se encontra, também, entre as dez palavras-chave mais utilizadas nos artigos pesquisados na bibliometria.

Zhang (2003) utiliza uma metodologia híbrida de redes neurais e o método ARIMA, combinando duas técnicas muito utilizadas para previsões com séries temporais. O autor explica que séries mais complexas de dados podem ser divididas em duas componentes, uma linear e uma não linear. O método consiste em dois passos, primeiro os dados são modelados com ARIMA para capturar toda componente linear; depois, utiliza-se o método de redes neurais para modelar os resíduos obtidos a partir do primeiro passo. Assim, a não linearidade seria modelada pela rede neural. O autor conclui, utilizando três conjuntos de dados distintos, que o método híbrido é superior aos dois métodos utilizados isoladamente.

O único artigo de metodologia aplicada com temática diferente de energia foi escrito por Lim e McAleer (2002), abordando a previsão de demanda turística na Austrália analisando a chegada de visitantes de Hong Kong, Malásia e Singapura em dados trimestrais. Os autores utilizaram modelagem ARIMA usando o método Box-

Jenkins, buscando tratar os dados quando eles se mostram não estacionários (como descrito no passo 1 da metodologia descrita por Hyndman (2002). Para selecionar o modelo mais adequado, os autores minimizaram o RMSE (root mean squared error, raiz do erro quadrático médio). Os autores identificaram alguns eventos isolados que alteraram significativamente o turismo em datas específicas, como a crise de preços de petróleo em 1979, a celebração bicentenária da colonização europeia de 1988, e a greve de pilotos aéreos australianos em 1989-1990. Para modelar estes efeitos, os autores usaram variáveis *dummy*.

Reikard (2009) utilizou dados de irradiação solar em séries de 5, 15, 30 e 60 minutos. O autor comparou os métodos de regressão logarítmica, ARIMA e o método de séries temporais estruturais (no entanto, o autor chama esse método de modelo de componentes não observadas, UCM), além de utilizar redes neurais e um método híbrido de regressões e redes neurais. Utilizando o método ARIMA com variáveis causais, o autor também define um método de funções de transferência. Ao todo, foram seis métodos submetidos a testes e comparados. O autor identificou que os melhores resultados em geral foram obtidos com o método ARIMA, embora houvesse algumas exceções. Em resoluções muito altas, um método de funções de transferência, utilizando dados de cobertura de nuvens, se sobressaiu. Em alguns casos, especialmente em resoluções de 5 minutos, os métodos de redes neurais e híbridos foram melhores. O autor comenta, no entanto, que o método ARIMA toma frações de segundo em um computador pessoal para ser processado, enquanto os métodos de UCM e redes neurais são muito mais lentos e tem resultados semelhantes ou piores que o ARIMA.

O trabalho de Wang, Li e Li (2018) também é relacionado à energia, mas do ponto de vista de demanda. Os autores utilizaram o modelo ou sistema Grey (GM), que utiliza poucos pontos de dados e equações diferenciais para estimar os pontos futuros. No entanto, os autores admitem que o GM foi melhorado ao longo dos anos e novas metodologias foram adicionadas. Primeiro, como o método utiliza apenas 5 a 10 pontos de dados, os dados históricos mais antigos normalmente não eram representados; para corrigir isso, os autores utilizam o método metabólico móvel, em que os dados são tomados cinco a cinco e retroalimentados ao modelo, de modo que os dados mais antigos serão utilizados, mas em próximas iterações os dados mais novos os substituem, dando um peso maior aos dados mais atuais. Outra melhoria foi a adição de não linearidades ao modelo. Por último, os autores propõem um modelo combinado com o método ARIMA. Assim, três modelos são testados: MGM (modelo Grey metabólico móvel), NMGM (modelo grey metabólico móvel não linear) e MGM-ARIMA (combinado). Os dados utilizados são o histórico anual de consumo de energia da China e da Índia. Os autores utilizam o indicador de erro MAPE (*mean absolute percent error*, erro percentual absoluto médio) para avaliar os modelos, chegando a

erros médios muito baixos, na ordem de 0,8 a 2,19%. O modelo NMGM teve a menor dispersão de erro, enquanto o modelo MGM-ARIMA chegou aos menores valores de erro.

Qing e Niu (2018) utilizaram redes neurais baseadas em LSTM para estimar a irradiação solar com base em dados meteorológicos de temperatura, ponto de orvalho, umidade, visibilidade, velocidade de ventos e sumário descritivo. Eles compararam o método ao algoritmo de retropropagação (BPNN), obtendo erros médios quadráticos mais baixos com LSTM.

Os artigos pesquisados (excluindo os livros e artigos de revisão) foram sumarizados na Tabela 3, evidenciando seus principais pontos.

Tabela 3 – Resumo dos artigos pesquisados

Dados	**Método(s)**	**Métrica(s)**	**Observação**	**Referência**
M-competition	RN	APE	Comparação aos métodos da M-competition	(HILL; O'CONNOR; REMUS, 1996)
N/A	LSTM	N/A	Descrição do método	(HOCHREITER; SCHMIDHUBER, 1997)
WS, CLD, BPUSD	Híbrido RN-ARIMA	MSE, MAD	Componentes linear e não-linear	(ZHANG, 2003)
Irradiação solar (5 a 60 min)	RL, ARIMA, STS, RN, Híbrido RN-ARIMA	MAPE	ARIMA é superior	(REIKARD, 2009)
Consumo energético anual da China e Índia	GM	MAPE, MSPE	Único com modelo Grey	(WANG; LI; LI, 2018)
Meteorológicos (1h)	LSTM, BPNN	RMSE	Estima incidência solar a partir de outras variáveis	(QING; NIU, 2018)
Radiação solar (30 min)	CLSTM	MAPE, MAE, RMSE	Compara com LSTM, CNN, DNN	(GHIMIRE et al., 2019)
Indicadores econômicos de 150 empresas chinesas	Estatístico Limítrofe	N/A	Tenta prever créditos verdes e investimentos em energia renovável	(HE et al., 2019)
Preços, volumes e quantidades de ações negociadas (1 min)	MFNN	Retorno total, Acurácia (%)	Compara com outros métodos	(LONG; LU; CUI, 2019)
Pesquisas efetuadas em buscadores de internet	KELM	nRMSE, MAPE	Diversos modelos testados e comparados	(SUN et al., 2019)
Parâmetros meteorológicos geolocalizados	Híbrido CNN-LSTM	MAE, nMAE, RMSE, nRMSE, R	Compara com os métodos individuais	(ZANG et al., 2020)
Dados turísticos trimestrais e dados econômicos	ARDL-ECM	N/A	Estima a recuperação turística pós-pandemia	(ZHANG et al., 2021)

Fonte: O autor.

Com isso, observou-se que os métodos que usam redes neurais são os mais utilizados atualmente, enquanto métodos consagrados incluem ARIMA e séries temporais estruturais. As métricas mais utilizadas foram MAPE e RMSE, incluindo algumas de suas variações como APE, MAE, nRMSE e MSE.

2.4 PESQUISA BIBLIOMÉTRICA ESPECÍFICA AO TEMA

O mesmo método de pesquisa da seção anterior porém os parâmetros foram alterados para "ALL ("violent deaths"OR "crime forecasting") AND PUBYEAR > 2018 AND PUBYEAR < 2024 AND (LIMIT-TO (SUBJAREA , "COMP") OR LIMIT-TO (SUBJAREA , "ENGI") OR LIMIT-TO (SUBJAREA , "MATH") OR LIMIT-TO (SUBJAREA , "DECI"))", ou seja, as palavras-chave passaram para *violent deaths* OU *crime forecasting* entres os anos de 2019 e 2023, limitados às subáreas de *computer science*, *engenharia*, *mathematics* e *decision science*. A busca retornou 666 itens cujos trabalhos mais citados, com relevância ao tema, são os artigos ciêntificos:

a) Wang et al. (2020) compararam o poder preditivo entre vários algoritmos de aprendizado de máquina Os resultados baseados apenas nos dados históricos de crimes sugerem que o modelo LSTM superou KNN, *random forest*, máquina de vetores de suporte, Bayes ingênuo e redes neurais convolucionais;

b) Zhang et al. (2020) construiram a rede de conscientização da situação do crime (*crime situation awareness network* - CSAN) como um novo modelo de previsão de referência por meio da integração de estruturas de codificadores automáticos variacionais e rede neural geradora de sequência baseada em contexto. Os experimentos finais demonstram que o CSAN supera principalmente outros algoritmos de previsão espaço-temporal comumente usados, como o Conv-LSTM, na previsão regional de frequência de crimes multitipos.

3 O MODELO PROPHET

No âmbito deste estudo, a análise de dados e a previsão desempenham um papel fundamental na compreensão das tendências temporais nas ocorrências de Mortes Violentas (MV). Uma abordagem que empregamos é o Modelo Prophet, uma ferramenta desenvolvida pelo *Facebook* (atual Meta) para previsão de séries temporais. Este modelo se baseia em princípios consolidados de modelagem e estatística, como modelos aditivos generalizados (GAMs) e algoritmos de minimização Quase-Newton, para capturar tendências lineares e não lineares, resultando em previsões precisas e detalhadas, associadas à possibilidade de automatização para lidar com um número cada vez maior de dados. O Modelo Prophet oferece diversas características distintas que o tornam adequado para a análise das ocorrências de MV ao longo do tempo:

a) Identificação Automática de Tendências: É capaz de identificar automaticamente as tendências presentes nos dados, eliminando a necessidade de configuração manual;

b) Modelagem de Sazonalidades: Ele incorpora sazonalidades diárias, semanais e anuais, adaptando-se às complexas variações sazonais nas ocorrências de MV;

c) Resistência a Dados Faltantes e *Outliers*: O modelo é robusto em relação a valores ausentes e dados anômalos, garantindo que as previsões sejam estáveis mesmo em situações desafiadoras;

d) Flexibilidade e Intervalos de Incerteza: Uma característica notável é a capacidade de especificar janelas de incerteza para as previsões, permitindo uma representação do nível de confiabilidade das variações possíveis.

Em seu artigo explicativo acerca da biblioteca Prophet para Python e R, Taylor e Letham (2018) criticam os métodos automatizados do pacote forecast em R, descrito e desenvolvido por Hyndman e Khandakar (2008). Utilizando métodos clássicos como o modelo ARIMA (auto-regressivo integrado de médias móveis), modelos de suavização exponencial e um modelo de passeio aleatório, a ferramenta em R não foi capaz de prever com precisão as tendências em momentos escolhidos ao longo do tempo na série temporal de Eventos do *Facebook*. Ou seja, numa série temporal que apresenta sazonalidades semanais e anuais, além de uma queda pronunciada nas semanas de Natal e Ano Novo, os métodos automatizados que existiam até então não forneciam uma boa resposta. Deste modo, os autores propõem um modelo de série temporal decomponível, descrito por Harvey e Peters (1990), com três componentes principais: tendência, sazonalidade e feriados. O modelo se descreve de acordo com a Eq. (3.1), onde $g(t)$ é a função tendência, que modela as mudanças não-periódicas nos valores

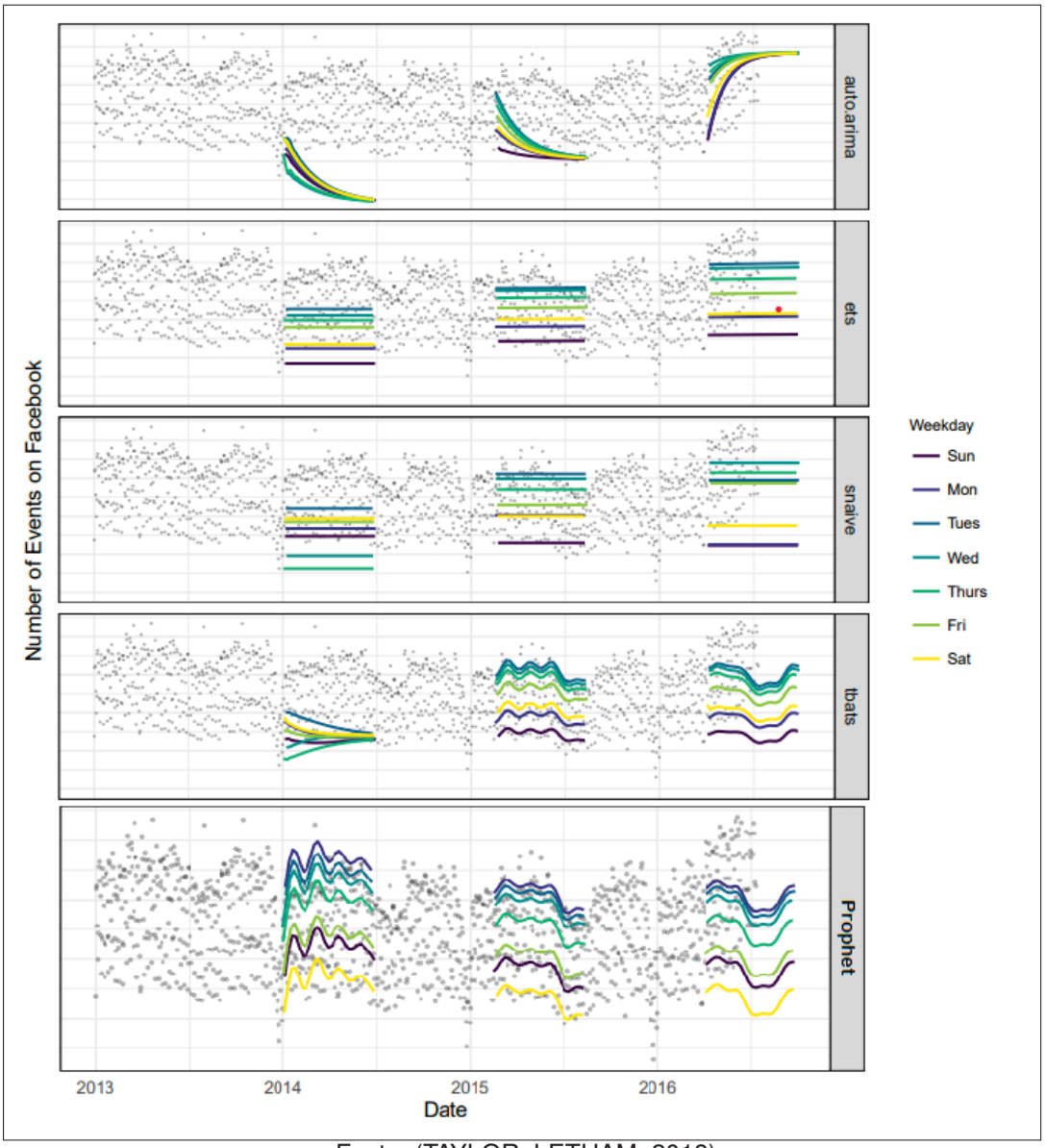
da série temporal, $s(t)$ modela as variações periódicas (sazonalidades), $h(t)$ modela os efeitos dos feriados e há, ainda, um termo ϵ_t que inclui o erro gerado a partir de influências cujo comportamento não foi modelado pelas outras funções.

$$y(t) = g(t) + s(t) + h(t) + \epsilon_1 \tag{3.1}$$

A ferramenta permite, também, a modelagem da sazonalidade de forma multiplicativa em vez de aditiva. Isso é necessário quando a intensidade dos efeitos da sazonalidade varia no tempo de forma semelhante à tendência. Neste caso, a abordagem aditiva não é capaz de modelar com precisão os efeitos da sazonalidade.

Com isso os autores foram capazes de melhorar as modelagens em relação aos métodos fornecidos pelo pacote *forecast*, como se pode observar a partir da Figura 8, onde se modelou previsões a partir dos dados coletados somente antes de determinados pontos.

Figura 8 – *Forecast* aprimorado.



Fonte: (TAYLOR; LETHAM, 2018)

Esta modelagem é semelhante, segundo os autores, ao modelo aditivo generalizado (GAM) proposto por Hastie e Tibshirani (1987).

3.1 GAM

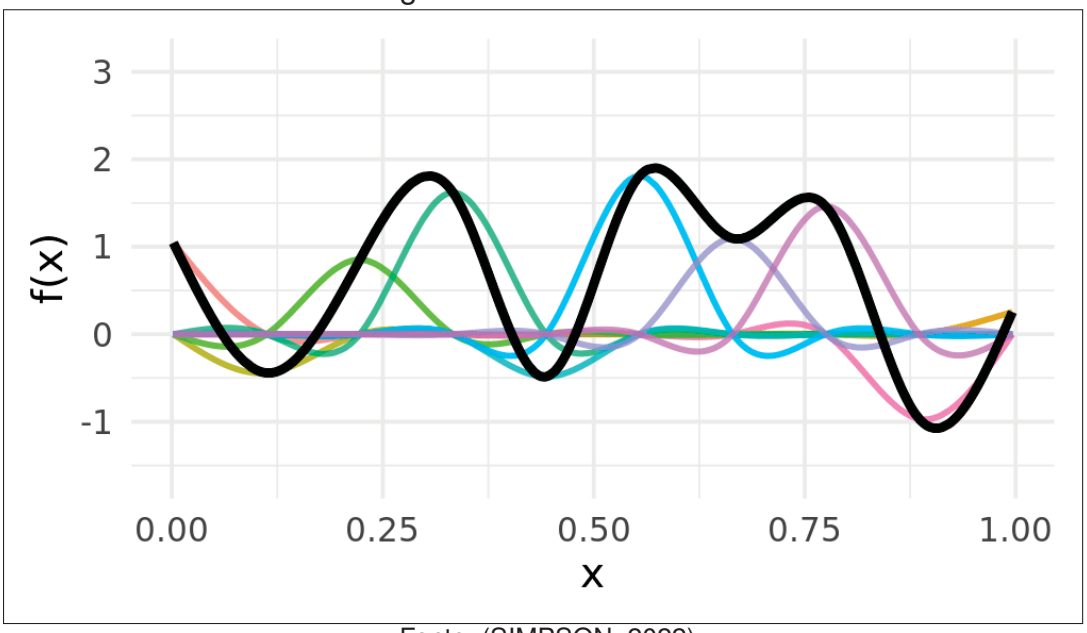
GAM é um modelo de regressão que, em vez de operar com a soma de termos lineares, como na regressão linear, utiliza a soma de funções suaves (WOOD, 2017). Ou seja, enquanto a média dos dados pode ser descrita como na Eq. (3.2), com β representando constantes que modelam os efeitos (lineares), em um GAM estas constantes são substituídas por funções suaves , como mostrado na Eq. (3.3).

$$y_i = \beta_0 + \sum_j \beta_j x_{ji} + \epsilon_i \tag{3.2}$$

$$y_i = \beta_0 + \sum_j s_j x_{ji} + \epsilon_i \tag{3.3}$$

As funções suaves são definidas pelo analista, mas usualmente são utilizadas splines devido a sua simplicidade computacional (BARROS, 2002). Estas *splines* são separadas em componentes simples chamadas funções de base, que somadas formam a função suave. O número de componentes é importante pois determina o número k de nós e, por consequência, determina a complexidade máxima que se pode atribuir à *splines*. Na Figura 1, é mostrada uma *splines* formada por k=10 componentes (linhas coloridas), cada um com um peso atribuído de modo a formar a curva *splines* (linha preta).

Figura 9 – Método GAM.



Fonte: (SIMPSON, 2022)

Os pesos das curvas são definidos de forma a maximizar a função logarítmica de verossimilhança penalizada. Esta penalização é aplicada para reduzir a possibilidade de sobreajuste, e pode ser ajustada por uma variável λ, conforme mostra a Eq. (3.4),

$$L_p(\beta) = L(\beta) - \frac{1}{2}\lambda\beta^T S\beta \tag{3.4}$$

onde $L_p(\beta)$ é a função logarítmica de verossimilhança penalizada, $L(\beta)$ é a função logarítmica de verossimilhança e $\beta^T S\beta$ é um termo que descreve a complexidade da curva. Ou seja, um λ grande irá penalizar mais (favorecendo curvas menos complexas) e um λ pequeno penaliza menos, favorecendo curvas mais complexas e podendo levar ao sobreajuste.

Nas bibliotecas que calculam GAMs, usualmente são estas as duas variáveis que podem ser escolhidas pelo analista para modificar os modelos obtidos: k e λ, sendo que o k escolhido será o máximo número de pontos e o algoritmo da biblioteca é

capaz de reduzir o número de nós para o mínimo necessário através da maximização da função $L_p(\beta)$.

3.2 MODELOS UTILIZADOS NO PROPHET

As funções utilizadas no *Prophet* para representar a tendência, a sazonalidade e feriados foram estabelecidas com base nas necessidades do *Facebook* e nos desafios inerentes a realizar previsões em escala, segundo Taylor e Letham (2018).

3.2.1 Tendência

Para representar a tendência, ou seja, os efeitos de longo prazo não submetidos a uma variação cíclica, utilizou-se dois modelos. O primeiro, um modelo de crescimento saturado, é baseado em um modelo de crescimento logístico como o apresentado na Eq. (3.5),

$$g(t) = \frac{C}{1 + exp(-k(t - m))} \tag{3.5}$$

com C representando a capacidade de suporte, k a taxa de crescimento e m um parâmetro de desvio.

Esta equação é similar ao crescimento populacional em ecossistemas naturais, onde a capacidade de suporte limita o crescimento a uma assíntota. No caso da modelagem do número de usuários de uma rede social, C seria, por exemplo, o número de pessoas com acesso à internet em determinada região.

No entanto, os autores identificaram dois fatores que não estão contemplados na Eq. (3.5), para os casos usuais encontrados no *Facebook*: a capacidade de suporte não é constante, mas variável com o tempo, tal qual o número de pessoas com acesso à internet é crescente; e a taxa de crescimento pode ser alterada ao longo do tempo, por exemplo com o lançamento de novos produtos. Deste modo, chega-se ao modelo de crescimento logístico por partes, em que a constante C é substituída por uma função $C(t)$ e a taxa de crescimento k passa a ser modificado por pontos de mudança de taxa, representados pelo vetor de ajustes de taxa $\boldsymbol{\delta}$ (onde $\delta \in \mathbb{R}^S$, com S sendo o número de pontos de mudança nos tempos, com $j = 1,...,S$), sendo que a taxa de crescimento é um valor de base k acrescido de todas as mudanças de taxa ocorridas até o tempo t. Isso é obtido a partir do vetor $\boldsymbol{a}$, definido por:

$$a_j(t) = \begin{cases} 1, & \text{se } t \geq S_j \\ 0, & \text{se } t < S_j \end{cases} \tag{3.6}$$

e deve-se ajustar o parâmetro de desvio m para conectar os pontos finais de cada segmento, utilizando para isso o termo $\boldsymbol{a}(t)^T\boldsymbol{\gamma}$.

Deste modo, é obtida a Eq. (3.7), com o modelo logístico de crescimento parte a parte. É importante destacar que o analista precisa conhecer bem os dados que estão sendo trabalhados para determinar corretamente a função de capacidade de suporte C(t). Os pontos de mudança podem ser detectados automaticamente pelo algoritmo ou determinados previamente pelo analista.

$$\begin{aligned} g(t) &= \frac{C(t)}{1 + exp(-(k + \boldsymbol{a}(t)^T\boldsymbol{\delta})(t - (m + \boldsymbol{a}(t)^T\boldsymbol{\gamma})))} \\ \text{onde: } \boldsymbol{\gamma} &\in \mathbb{R}^S, \text{com } j = 1,...,S \\ \gamma_j &= (s_j - m - \sum_{l<j} \gamma_l)\left(1 - \frac{k + \sum_{l<j}\delta_l}{k + \sum_{l\leq j}\delta_l}\right) \end{aligned} \tag{3.7}$$

O segundo modelo que pode ser utilizado é a tendência linear com pontos de mudança, usado para previsões em que não é prevista uma saturação, mostrado na Eq. (3.9), utilizando as mesmas variáveis já descritas.

$$\begin{aligned} g(t) &= (k + \boldsymbol{a}(t)^T\boldsymbol{\delta})t + (m + \boldsymbol{a}(t)^T\boldsymbol{\gamma}) \\ &\text{onde: } \gamma_j = s_j\delta_j \end{aligned} \tag{3.8}$$

Quando o modelo é utilizado para previsão, pontos de mudança futuros são inferidos a partir dos dados anteriores, de modo que a frequência de mudanças e suas intensidades correspondem àquelas observadas historicamente. Por meio de simulação de possíveis futuras tendências, é calculado o nível de incerteza da previsão.

3.2.2 Sazonalidade

O Prophet utiliza uma série de Fourier para representar os efeitos periódicos com base no trabalho de Harvey e Shephard (1993). Sendo P o período regular que se espera que a série tenha, pode-se aproximar os efeitos sazonais (suavizados) a partir da Eq. 3.9, que é uma série de Fourier padrão sem o termo do intercepto. Despreza-se esse termo em razão do modelo ser calculado simultaneamente com o componente de tendência, ou seja, somar uma constante a s(t) forçaria a subtração desta constante de g(t) e nenhuma nova informação seria obtida.

$$s(t) = \sum_{n=1}^{N}(a_n \cos\frac{2\pi nt}{P} + b_n \sin\frac{2\pi nt}{P}) \tag{3.9}$$

Ajustar a sazonalidade significa estimar os 2N parâmetros $\beta = [a_1,a_2,...,a_N,b_N]^T$, o que é feito a partir de uma matriz de vetores de sazonalidade para cada valor de t nos dados históricos e futuros. Por exemplo, para sazonalidade anual com dados diários, e N=10, tem-se:

$$X(t) = \left[\cos\frac{2\pi(1)t}{365{,}25},...,\sin\frac{2\pi(10)t}{365{,}25}\right] \tag{3.10}$$

e a componente sazonal é:

$$s(t) = X(t)\beta \tag{3.11}$$

com β sendo *a priori* uma distribuição normal de média zero e variância σ^2 na modelagem generativa Bayesiana.

Truncar a série em N tem um efeito de filtro passa-baixa, ou seja, efeitos cuja frequência seja muito alta não são capturados com N abaixo de certo limite. Logo, aumentar N permite ajustar padrões sazonais mais curtos (aumentando o risco de sobreajuste).

3.2.3 Feriados

Os feriados são tratados como uma lista personalizada fornecida pelo analista, sendo que os efeitos de dias subsequentes (como a semana de Carnaval) devem ser modelados como se cada dia fosse um feriado. Assume-se que os efeitos de cada feriado são independentes. De modo similar à sazonalidade, é gerada uma matriz $Z(t)$ de funções indicadoras representando se o tempo t está dentro de um feriado i, e um parâmetro κ_i para cada feriado. Assim, tem-se:

$$h(t) = Z(t)\kappa \tag{3.12}$$

Com κ sendo a priori, também, uma distribuição normal de média zero e variância v^2

3.2.4 Ajuste dos modelos

É utilizado o algoritmo L-BFGS (Broyden-Fletcher-Goldfarb-Shanno com memória limitada) em código Stan para chegar a uma estimativa máxima a posteriori. O algoritmo BFGS é um método de minimização Quase-Newton, onde se estima o inverso da matriz Hessiana ao invés de a calcular explicitamente, o que é muito mais pesado computacionalmente. A versão de memória limitada não guarda a matriz completa, mas apenas alguns vetores que representam a aproximação implicitamente.

Ao combinar as características de sazonalidade e de feriados em uma única matriz X e os indicadores de pontos de mudança em uma matriz A, o modelo da Eq. (4) pode ser expresso em algumas linhas de código Stan (Stan Development Team, 2022), como mostrado na Figura 10.

Figura 10 – Linhas de código Stan.

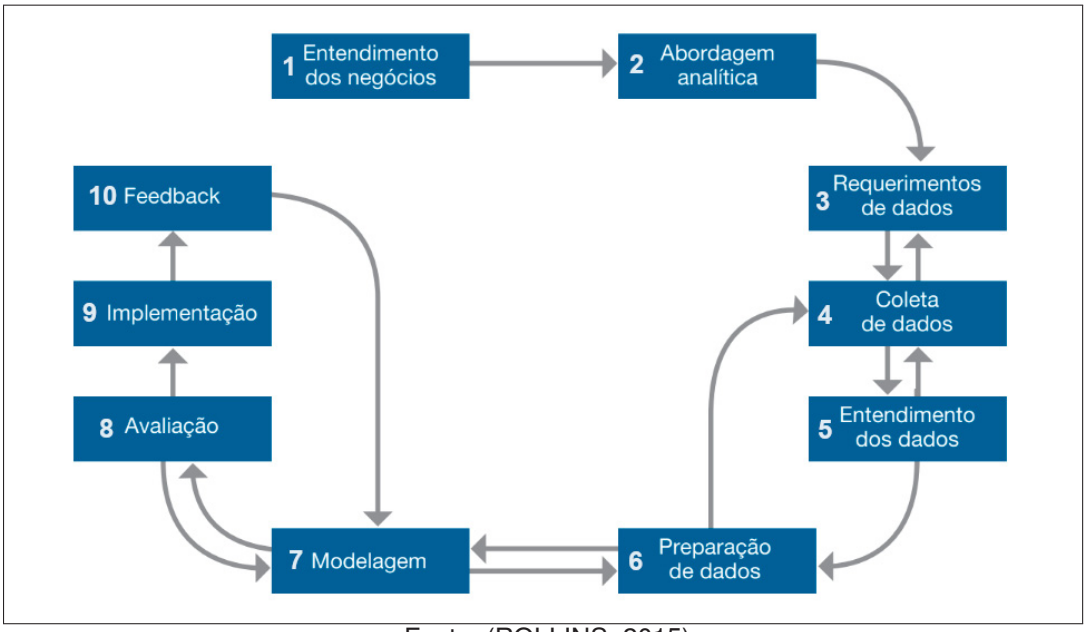
```
model {
  // Priors
  k ~ normal(0, 5);
  m ~ normal(0, 5);
  epsilon ~ normal(0, 0.5);
  delta ~ double_exponential(0, tau);
  beta ~ normal(0, sigma);

  // Logistic likelihood
  y ~ normal(C ./ (1 + exp(-(k + A * delta) .* (t - (m + A * gamma)))) +
             X * beta, epsilon);
  // Linear likelihood
  y ~ normal((k + A * delta) .* t + (m + A * gamma) + X * beta, sigma);
}
```

Fonte: (Stan Development Team, 2022)

4 METODOLOGIA

De modo a obter os dados, tratá-los e interpretá-los, utilizou-se uma metodologia semelhante àquela adotada pela IBM para Ciência de Dados (ROLLINS, 2015), mostrada esquematicamente na Figura 11.

Figura 11 – Metodologia de base para ciência de dados segundo a IBM



Fonte: (ROLLINS, 2015)

Antes do início deste projeto, foi analisado e entendido o modelo de atuação da Polícia Científica do Paraná dentro de seu contexto (Etapa 1), observando-se a necessidade da aplicação de ferramentas de apoio para a tomada de decisões gerenciais. Nesta etapa surge a situação-problema que deu origem a execução deste trabalho. Em seguida, inicia-se a Etapa 2, de abordagem analítica, em que se estabelece o estudo das ferramentas de análise que serão apresentadas nas próximas Seções. As ferramentas foram selecionadas de modo a contemplar métodos consagrados (já identificados na Revisão de Literatura) e um método inovador, buscando compará-los e estabelecer qual pode ser a ferramenta mais adequada para os dados em estudo. Os estágios 3, 4, 5 e 6 foram realizados em paralelo, ou seja, conforme se coletavam os dados eram realizadas análises preliminares, que davam subsídio para a necessidade da coleta de mais dados, estabelecendo-se um padrão de preparo e o entendimento destes dados. A modelagem foi realizada utilizando métodos de suavização exponencial e a ferramenta *Prophet*. Para a etapa de avaliação, separou-se o conjunto de dados em dois subconjuntos, um para ajuste dos modelos e outro para cálculo do erro, utilizando as métricas MAPE e RMSE. Por fim, o modelo escolhido foi utilizado para prever as necessidades de pessoal nos próximos dez anos.

4.1 OBTENÇÃO DA BASE DE DADOS (UM TUTORIAL)

Os dados empregados neste estudo foram adquiridos por meio do portal eletrônico da Polícia Científica do Paraná[1], disponibilizados em formato de relatório no Microsoft Power BI. O acesso aos dados foi realizado conforme o procedimento a seguir:

1. Acessou-se o portal eletrônico da Polícia Científica do Paraná (Figura 12);
2. Navegou-se até a seção Transparência no referido portal;
3. Dentro da seção de Transparência, foram selecionadas as opções Produtividade e Desempenho;
4. Dentro do relatório, foi selecionada a opção Exame pericial seccionado por natureza e regional, especificamente Natureza do exame V2 (Figura 13).

Figura 12 – Site da Polícia Científica do Paraná

Fonte: Site da PCP (Polícia Científica do Paraná).

[1] https://www.policiacientifica.pr.gov.br/Pagina/Produtividade-e-Desempenho

Figura 13 – Site da PCP/produtividade e desempenho

Fonte: Site da PCP.

Na página subsequente, foram identificados os exames de interesse para o escopo do estudo, considerando sua ocorrência no período de janeiro de 2014 a outubro de 2023.

Os dados pertinentes foram extraídos a partir da tabela apresentada no relatório, utilizando a funcionalidade de cópia disponível no ambiente Power BI (Figura 14). Os dados copiados foram então transferidos para um arquivo no formato LibreOffice Calc.

Posteriormente, esses dados foram convertidos em um arquivo no formato CSV, preparando-os para análises subsequentes.

Figura 14 – Cópia dos dados

Fonte: Site da PCP.

Este procedimento foi realizado para todas as informações de interesse deste trabalho.

4.2 TRATAMENTO E ANÁLISE DOS DADOS

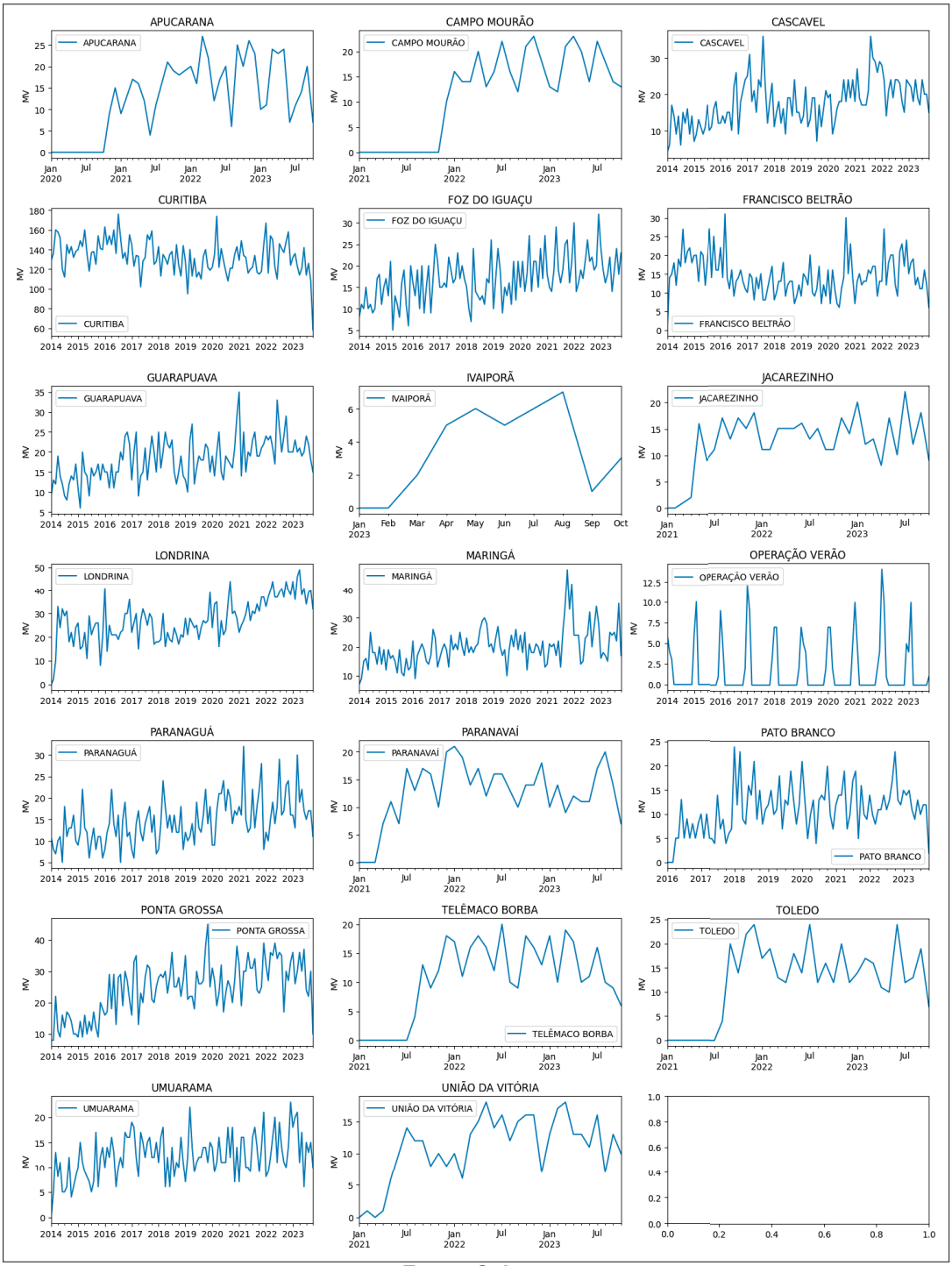
Após a coleta dos dados, reunidos no arquivo *MV.csv*, procedeu-se ao tratamento e visualização dos dados, bem como a utilização dos métodos consagrados de previsão com séries temporais e, posteriormente, a análise com o Prophet.

Os dados foram importados utilizando-se a classe dataframe da biblioteca pandas, uma ferramenta consagrada da análise de dados utilizando Python.

O tratamento de dados iniciou-se ao renomear as variáveis referentes aos meses, que se apresentavam na forma escrita por extenso, para variáveis numéricas de 1 a 12. Do mesmo modo, as colunas do dataframe *Ano* e *Mês* foram renomeadas, respectivamente, para *YEAR* e *MONTH*, de modo a facilmente criar uma coluna com o método *to_datetime()* da classe *dataframe*. Por fim, para utilizar o padrão exigido pela biblioteca *Prophet*, a coluna referente ao tempo (em *datetime* no formato AAAA-MM-DD, com o dia fixado em 1) foi nomeada *ds* e a coluna referente a resposta do sistema renomeada para *y*.

A seguir, procedeu-se a visualização dos dados das dezenove regionais, utilizando a biblioteca matplotlib, obtendo-se os gráficos que foram reunidos na Figura 15. O Anexo A mostra a primeira página do *MV.csv*.

Figura 15 – Visualização dos dados referentes as 19 regionais mais a Operação Verão, até outubro de 2023.



Fonte: O Autor.

Na sequência, realizou-se a avaliação das estatísticas anuais por regional, descrevendo quantos anos a unidade esteve ativa entre 2014 e outubro 2023, a média e desvio padrão no número de MV, bem como o mínimo, máximo e a distribuição em quartis. Com base nesses resultados foi possível gerar um gráfico do tipo *box-plot* para representar graficamente as estatísticas. Estes resultados serão explorados na Seção Resultados.

Tabela 4 – MV na PCP de 2014 até outubro de 2023.

YEAR	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Regional										
APUCARANA	0	0	0	0	0	0	24	175	234	151
CAMPO MOURÃO	0	0	0	0	0	0	0	10	205	170
CASCAVEL	136	144	198	267	187	182	221	286	258	202
CURITIBA	1661	1681	1779	1614	1560	1460	1569	1531	1673	1174
FOZ DO IGUAÇU	145	160	194	212	178	188	234	237	257	197
FRANCISCO BELTRÃO	201	223	184	145	137	145	170	160	216	134
GUARAPUAVA	152	167	208	208	227	214	226	257	281	202
IVAIPORÃ	0	0	0	0	0	0	0	0	0	35
JACAREZINHO	0	0	0	0	0	0	0	119	164	141
LONDRINA	241	244	303	299	245	314	347	366	467	399
MARINGÁ	187	171	215	233	274	242	229	309	289	214
OPERAÇÃO VERÃO	13	17	16	24	16	18	21	21	30	15
PARANAGUÁ	133	133	159	150	171	170	207	217	213	177
PARANAVAÍ	0	0	0	0	0	0	0	118	184	125
PATO BRANCO	0	0	63	86	170	150	143	152	157	113
PONTA GROSSA	153	159	276	297	329	335	309	354	381	281
TELÊMACO BORBA	0	0	0	0	0	0	0	56	176	126
TOLEDO	0	0	0	0	0	0	0	84	189	143
UMUARAMA	83	121	152	171	139	153	152	160	167	146
UNIÃO DA VITÓRIA	0	0	0	0	0	0	0	82	158	131

Fonte: O autor.

4.3 MODELAGEM E EXPERIMENTAÇÃO COMPUTACIONAL

Prosseguiu-se com a aplicação de métodos de suavização exponencial consagrados, como descritos por (HYNDMAN; ATHANASOPOULOS, 2021), e implementados pela biblioteca *statsmodels* em *Python*. Inicialmente, o método de suavização exponencial simples, que não considera tendência nem sazonalidade. Foi aplicado utilizando o método *fit* da classe *SimpleExpSmoothing*. Este método se ajusta bem aos dados, mas é inadequado para previsões (as saídas são constantes). Segundo, o método de suavização exponencial linear de Holt, que considera a tendência apenas. Foi aplicado utilizando o mesmo método *fit*, mas da classe *ExponentialSmoothing*. Assim como o método anterior, as previsões deste não são adequadas, pois as saídas são lineares. Por fim, ainda utilizando a classe *ExponentialSmoothing*, o método Holt-Winters adiciona a sazonalidade ao método anterior, permitindo que seja modelada de forma aditiva ou multiplicativa. Neste caso as previsões parecem ser mais adequadas, visto que a forma do gráfico para o futuro se assemelha as observações do passado.

Utilizando-se a biblioteca *prophet* para *Python*, modelou-se os dados utilizando o modo de sazonalidade multiplicativa e o modelo padrão de tendência, que é a tendência linear com pontos de mudança, dada na Eq. (3.9). Não se utilizou a modelagem de feriados, uma vez que os dados utilizados eram mensais. Em seguida, realizou-se a previsão para os próximos anos.

4.4 AVALIAÇÃO

Os modelos foram avaliados utilizando três métricas consagradas, MAE, MSE e RMSE, mapeadas na Revisão de Literatura como as técnicas muito utilizadas para avaliação de modelos.

4.5 APLICAÇÃO

O modelo obtido e avaliado foi então utilizado, em conjunto com uma divisão simples, para determinar o número projetado necessário de peritos para os próximos dez anos nas regionais. Calculou-se a razão do número de atendimentos em relação ao número de peritos. Com base nisso, foi estabelecida uma faixa razoável esperada deste quociente para cada perito. Assim, foi possível estimar uma faixa de valor, a partir da previsão do número de ocorrências, da quantidade de peritos necessária e suficiente para atender com qualidade a demanda projetada.

5 RESULTADOS

5.1 ANÁLISE DE DADOS

As análises estatísticas dos dados fornecem percepções valiosas sobre as ocorrências de MV ao longo dos anos no Paraná. A Tabela 5 apresenta a distribuição das ocorrências de MV por ano, classificadas em seis tipos distintos. A partir desses dados, também segmentados por regional, foi realizada a análise estatística.

Tabela 5 – Ocorrências com MV por ano.

NATUREZA DO EXAME	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023*
Atropelamento	151	169	179	180	169	182	166	209	202	154
Trânsito	585	559	698	690	660	680	679	973	1157	915
Local de morte	553	581	714	712	767	849	922	1457	1801	1400
Suicídio	311	364	445	517	583	637	665	748	942	734
Homicídio	1455	1416	1557	1440	1211	1015	1138	1023	1261	889
Ação Policial	28	122	140	153	195	185	237	255	336	184
Total	3083	3211	3733	3692	3585	3548	3807	4665	5699	4276

*somente até outubro.

Fonte: SESP PR.

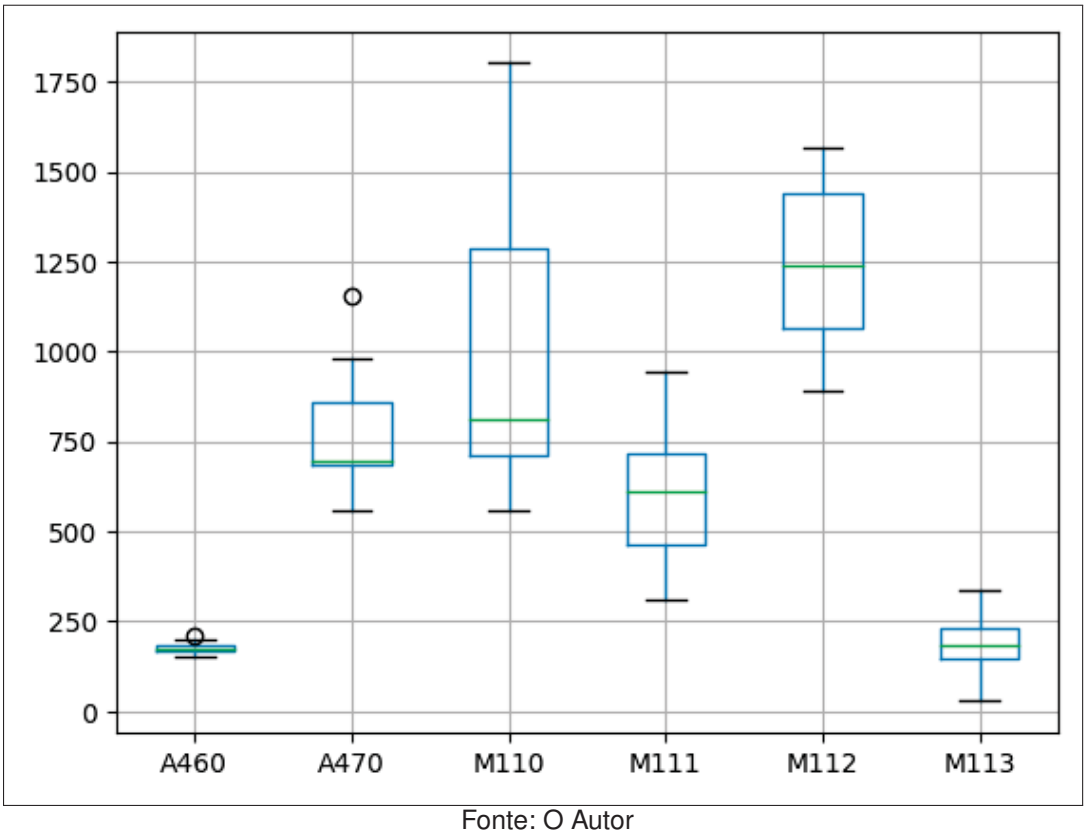
A média de ocorrências de MV por ano foi calculada como sendo de 3950,3 casos com desvio padrão de 768,85. A mediana, que representa o valor central da distribuição, foi encontrada como sendo de 3726,5 ocorrências por ano. O coeficiente de variação calculado foi de 19,46%. Foi traçado o gráfico *box-plot* dos dados acima expostos, dado na Figura 16, com a relação dos códigos dada pela Tabela 6.

Tabela 6 – Correspondência do código e descrição.

Código	Descrição
A460	EXAME EM LOCAL DE ATROPELAMENTO E MORTE
A470	EXAME EM LOCAL DE ACIDENTE DE TRANSITO E MORTE
M110	EXAME EM LOCAL DE MORTE
M111	EXAME EM LOCAL DE MORTE – SUICÍDIO
M112	EXAME EM LOCAL DE MORTE – HOMICÍDIO
M113	EXAME EM LOCAL DE MORTE – (HOMICÍDIO) DECORRENTE DE INTERVENÇÃO POLICIAL

Fonte: O Autor

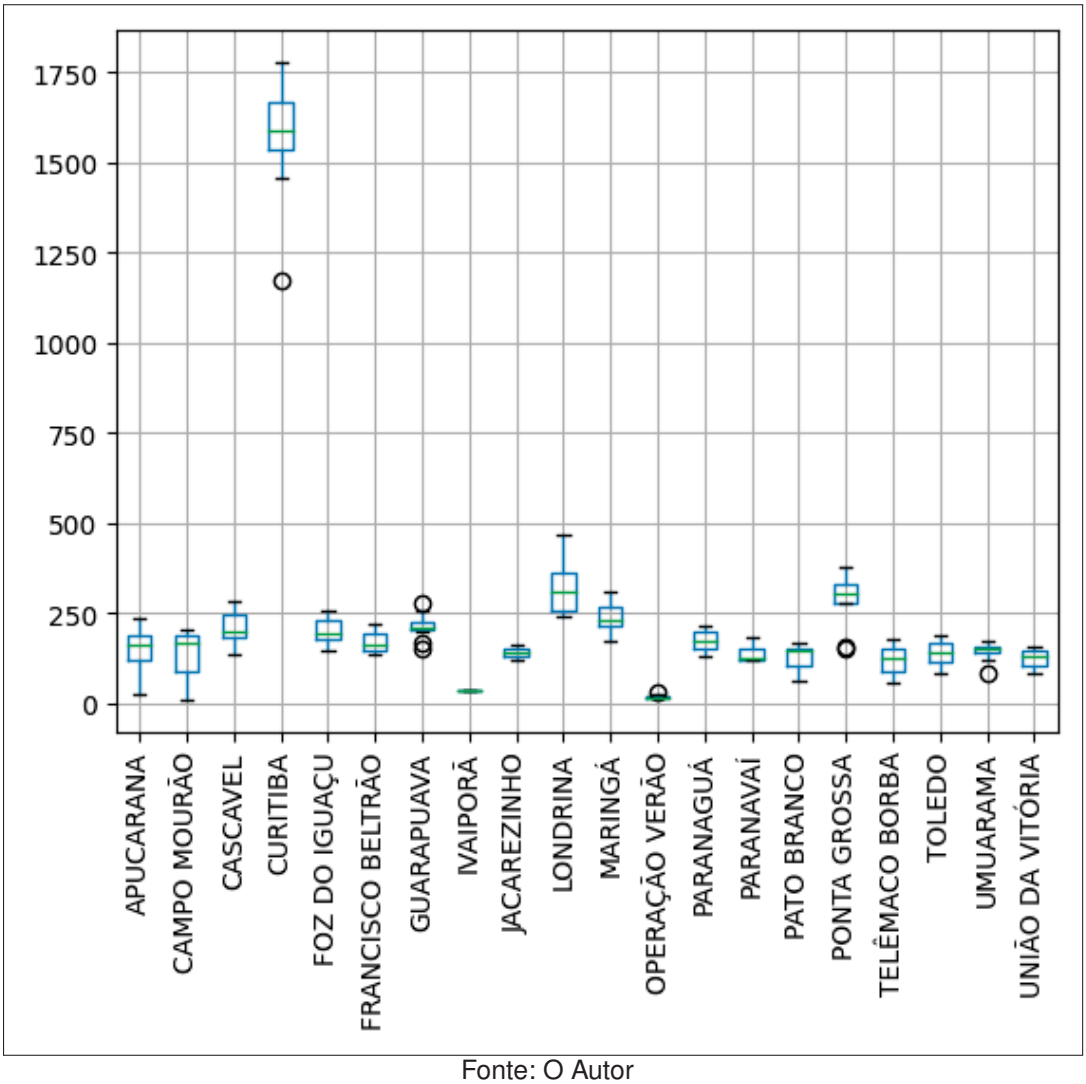
Figura 16 – Representação gráfica da distribuição anual de mortes violentas por tipo.



Fonte: O Autor

A distribuição segundo cada regional foi representada no gráfico da Figura 17, onde observa-se que a regional Curitiba concentra a maior parte dos casos, enquanto as unidades do interior apresentam números menores. A Operação Verão se dá apenas durante a alta temporada nos municípios que apresentam alta demanda turística durante a temporada, por isso apresenta o menor valor anual.

Figura 17 – Representação gráfica da distribuição anual de mortes violentas por regional.

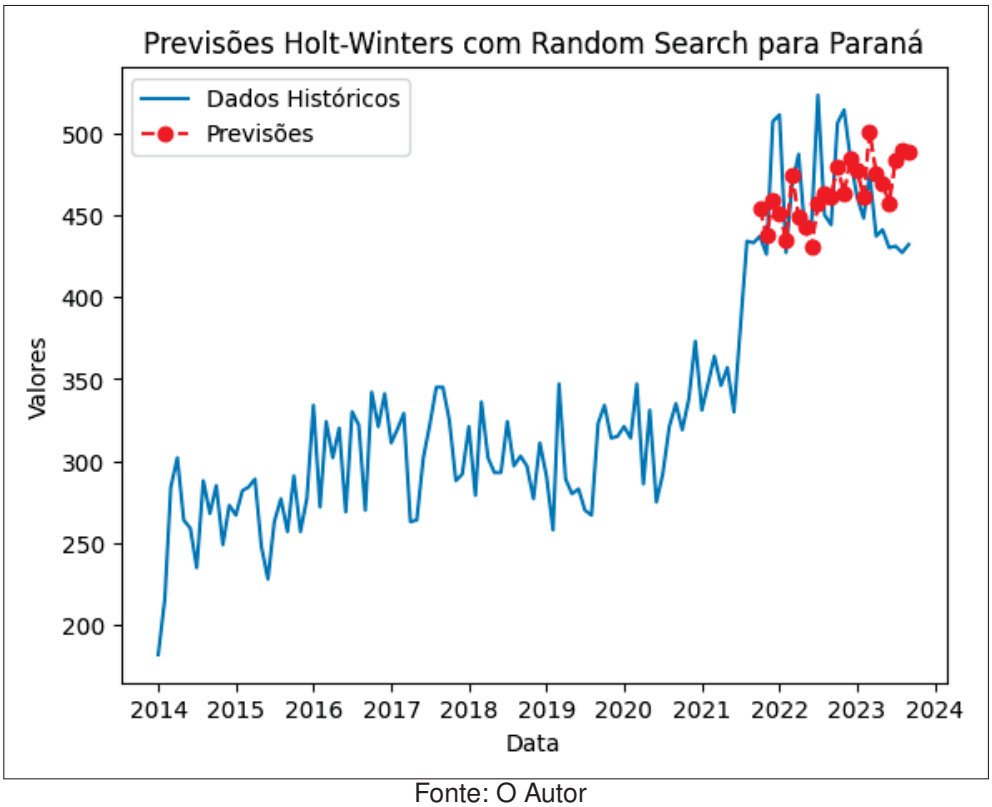


Fonte: O Autor

5.2 SUAVIZAÇÃO EXPONENCIAL

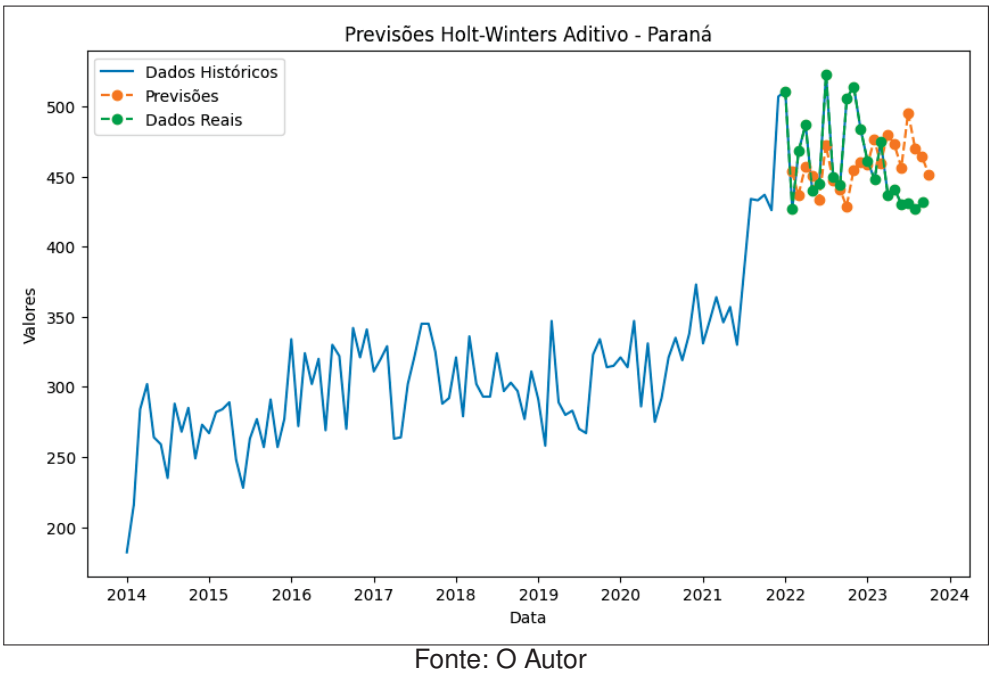
Realizou-se a modelagem por suavização exponencial utilizando-se o método Holt-Winters com tendência aditiva, sazonalidade aditiva e multiplicativa usando a função *ExponentialSmoothing* da biblioteca *statsmodels* para as regionais e todo o PR, com exceção de Ivaiporã, que é uma regional recente (segundo semestre de 2023) e não apresenta muitas ocorrências de MV e tampouco dados para sazonalidade anual. Para determinar os hiperparâmetros alfa, beta e gama referentes aos coeficientes de suavização no método de suavização exponencial triplo de Holt-Winters, optou-se por utilizar o ajuste automático inato da função, pois os testes de erro, como MSE e MAE, apresentaram melhores resultados para o ajuste automático do que determinando os coeficientes com *Random Search*(Figura 18), para o Paraná. As Figuras 19 - 30 mostram as previsões Holt-Winters com sazonalidade aditiva e multiplicativa para o Paraná e regionais.

Figura 18 – Previsões Holt-Winters com *Random Search* para o Paraná.
MSE: 1351.4064732121637.
MAE: 30.182587970778254.



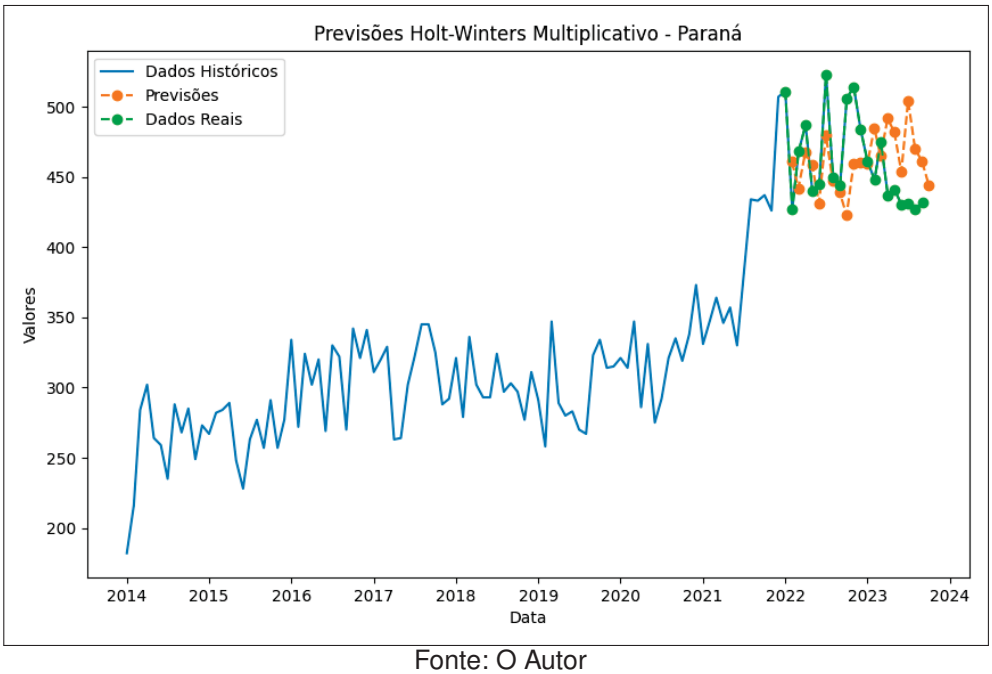
Fonte: O Autor

Figura 19 – Previsões Holt-Winters Aditivo - Paraná.
MSE: 1304.8729531482427.
MAE: 29.701641904196734.



Fonte: O Autor

Figura 20 – Previsões Holt-Winters Multiplicativo - Paraná.
MSE: 1357.5708533937463.
MAE: 30.74336564023505.



Fonte: O Autor

Figura 21 – Previsões Holt-Winters Aditivo - Regionais(parte 1).

Fonte: O Autor

Figura 22 – Previsões Holt-Winters Aditivo - Regionais(parte 2).

Fonte: O Autor

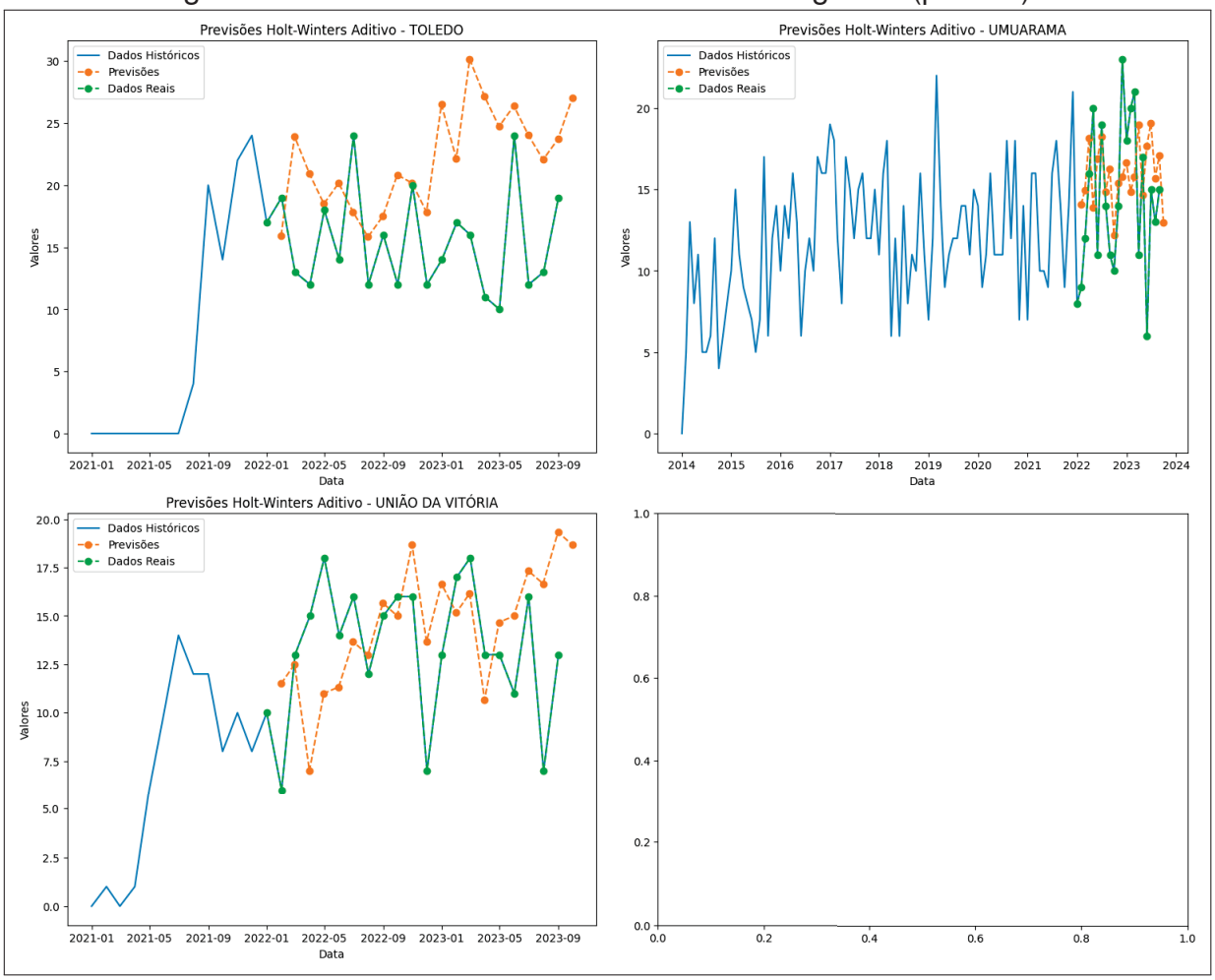
Figura 23 – Previsões Holt-Winters Aditivo - Regionais(parte 3).

Fonte: O Autor

Figura 24 – Previsões Holt-Winters Aditivo - Regionais(parte 4).

Fonte: O Autor

Figura 25 – Previsões Holt-Winters Aditivo - Regionais(parte 5).

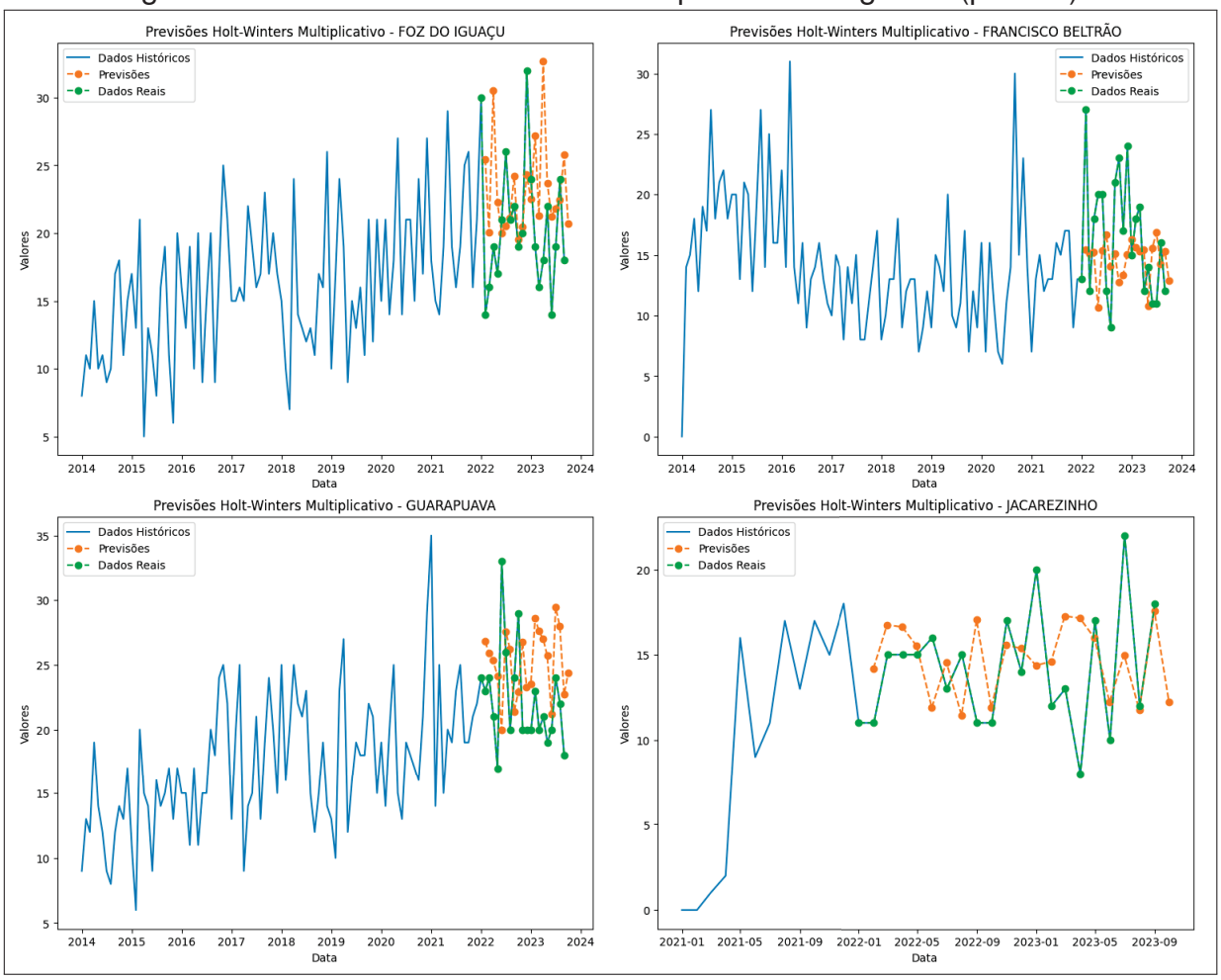


Fonte: O Autor

Figura 26 – Previsões Holt-Winters Multiplicativo - Regionais(parte 1).

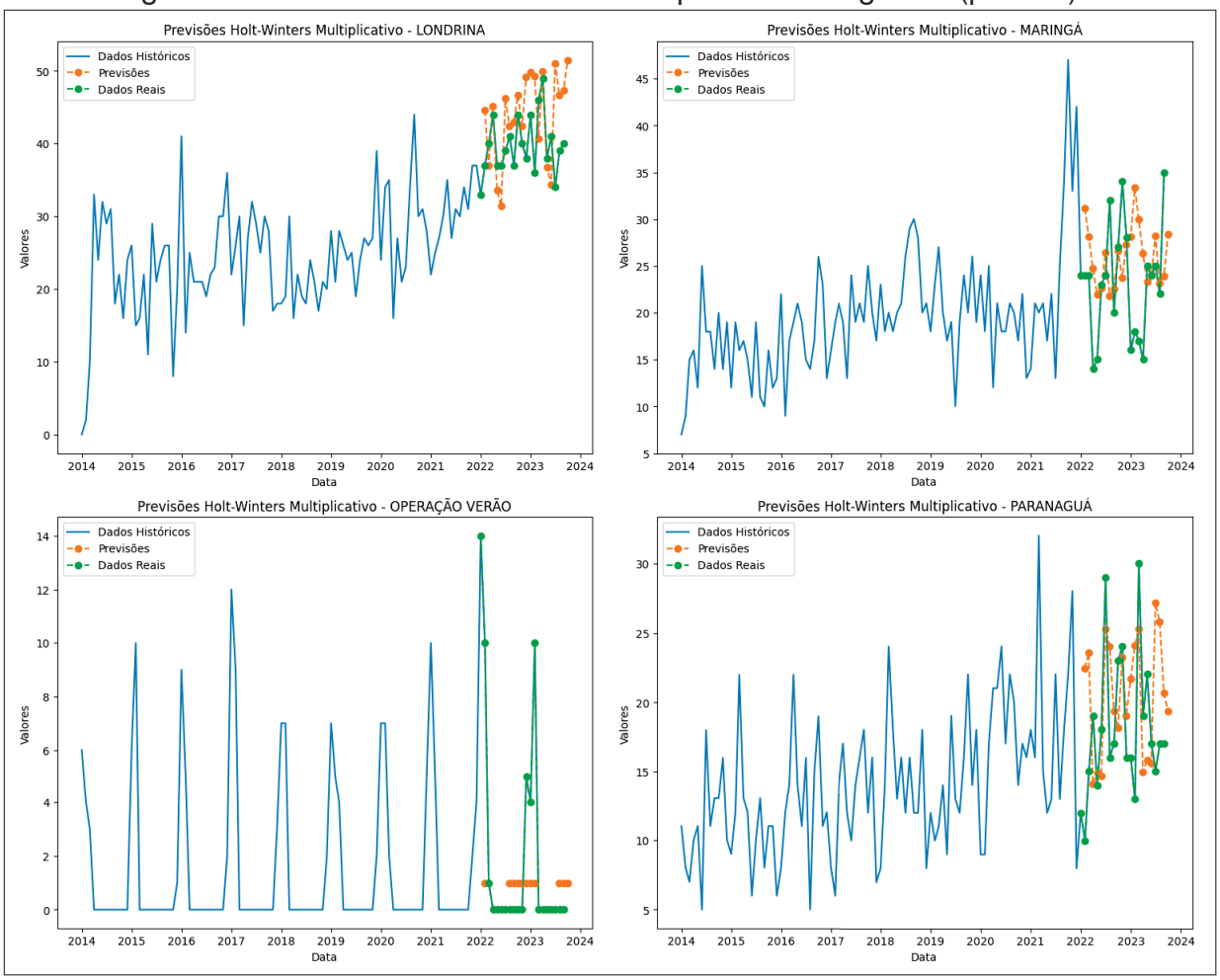
Fonte: O Autor

Figura 27 – Previsões Holt-Winters Multiplicativo - Regionais(parte 2).



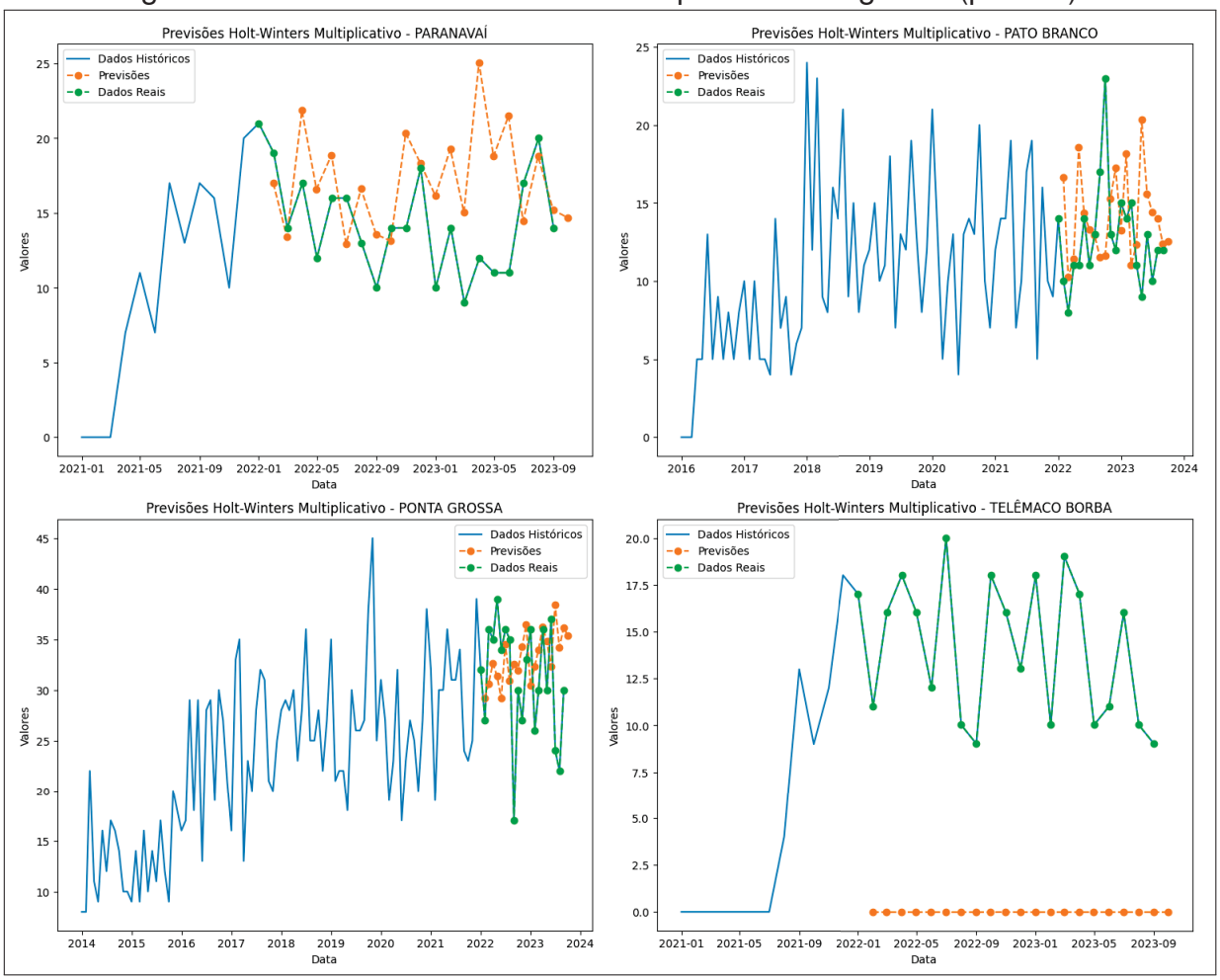
Fonte: O Autor

Figura 28 – Previsões Holt-Winters Multiplicativo - Regionais(parte 3).



Fonte: O Autor

Figura 29 – Previsões Holt-Winters Multiplicativo - Regionais(parte 4).



Fonte: O Autor

Figura 30 – Previsões Holt-Winters Multiplicativo - Regionais(parte 5).

Fonte: O Autor

Observa-se que na abordagem multiplicativa a previsão apresenta qualidade inferior quando comparada com a aditiva. O assunto será abordado novamente na próxima subseção.

5.2.1 Erro de previsão

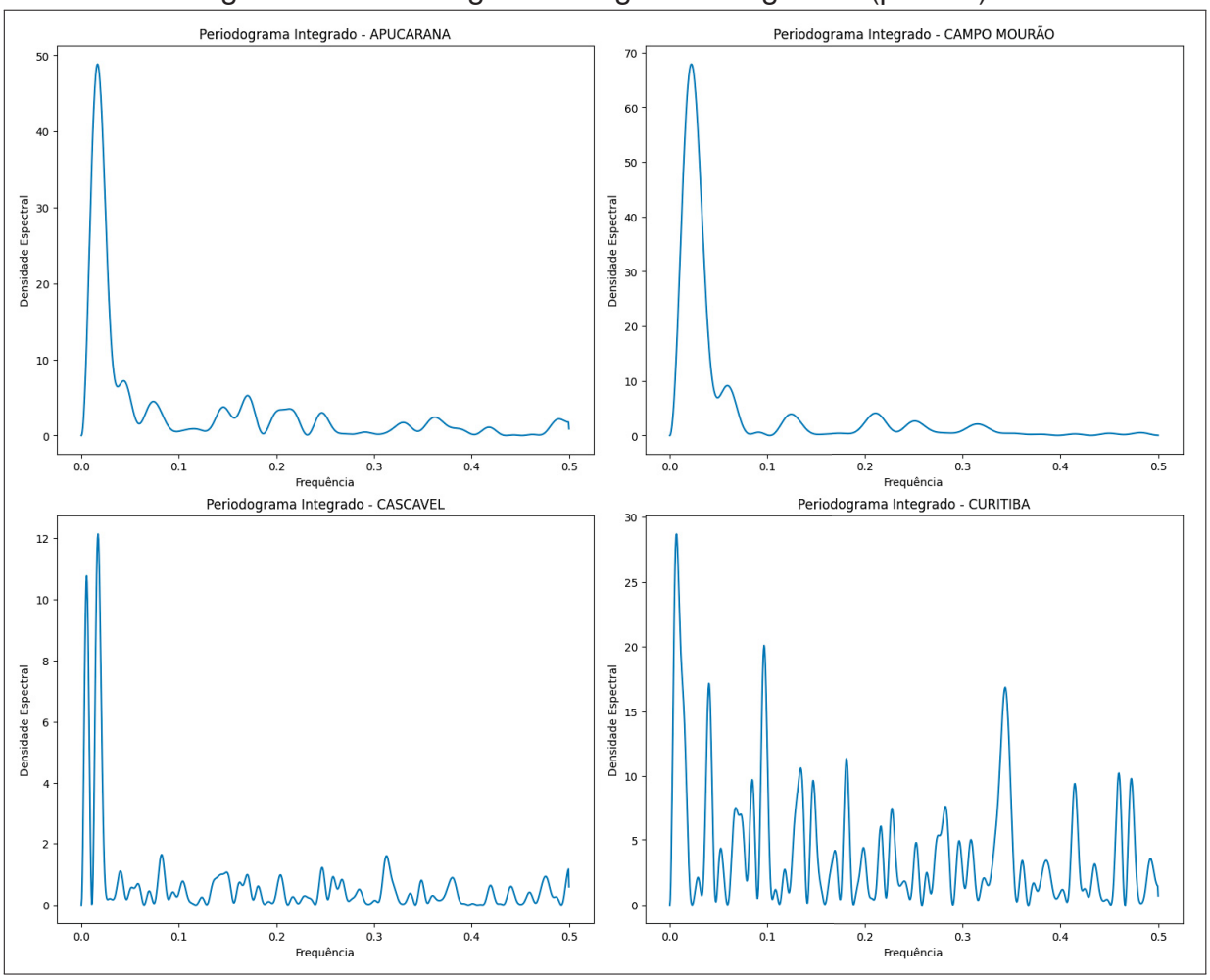
Os erros foram calculados usando os dados reais dos anos de 2022 e 2023 (parcial), comparando-se aos pontos previstos. As métricas MAE, RMSE e MSE foram calculadas. A Tabela 7 mostra os resultados.

Tabela 7 – Comparação das métricas entre Holt-Winters Aditivo e Multiplicativo nas Regionais.

	Aditivo			Multiplicativo		
Regional	**MAE**	**RMSE**	**MSE**	**MAE**	**RMSE**	**MSE**
Apucarana	7.429	8.632	74.514	9.557	11.222	125.951
Campo Mourão	9.581	10.655	113.522	-	-	-
Cascavel	3.695	4.620	21.342	4.222	5.403	29.195
Curitiba	13.925	17.307	299.547	14.087	17.493	306.033
Foz do Iguaçu	4.244	5.417	29.349	4.854	6.355	40.386
Francisco Beltrão	4.144	5.241	27.478	4.241	5.261	27.686
Guarapuava	3.210	3.826	14.644	3.710	4.444	19.751
Jacarezinho	6.563	7.957	63.324	3.848	4.596	21.130
Londrina	4.254	5.044	25.448	7.228	8.196	67.186
Maringá	6.728	8.092	65.484	5.810	7.126	50.786
Operação Verão	4.041	5.696	32.444	-	-	-
Paranaguá	5.365	6.369	40.572	6.149	7.525	56.637
Paranavaí	7.167	8.257	68.192	4.718	6.091	37.109
Pato Branco	2.873	3.477	12.095	3.383	4.239	17.971
Ponta Grossa	4.908	6.172	38.099	5.474	6.790	46.116
Telêmaco Borba	3.524	4.196	17.606	14.094	14.559	211.990
Toledo	7.674	8.870	78.679	15.475	15.973	255.162
Umuarama	4.066	4.935	24.363	4.176	5.142	26.443
União da Vitória	4.063	5.164	26.677	23.106	32.023	1025.527

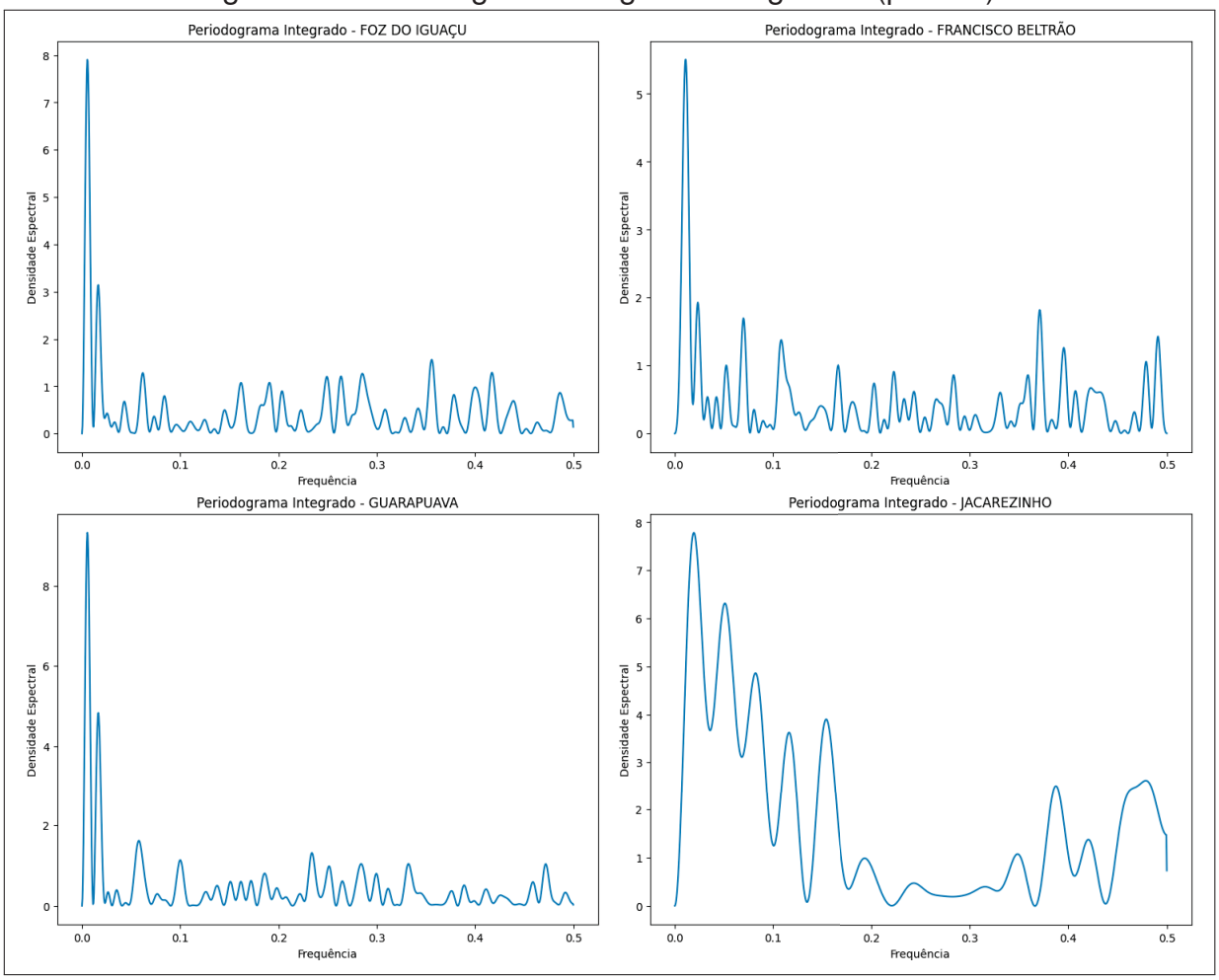
Nota-se pelas tabelas e gráficos que *Holt-Winters* com tedência aditiva mostra-se uma escolha melhor quando comparada com a multiplicativa, isto considerando que algumas regionais apresentam poucos dados (devido serem novas) e que os hiperparâmetros foram determinados automaticamente pela função inata. Os periodogramas integrados (Figuras 31 - 35) também corroboram que Holt-Winters Aditivo é a melhor escolha, pois os gráficos, em sua maioria, apresentam picos de amplitude uniforme e frequências uniformemente espaçadas.

Figura 31 – Periodograma integrado - Regionais (parte 1).



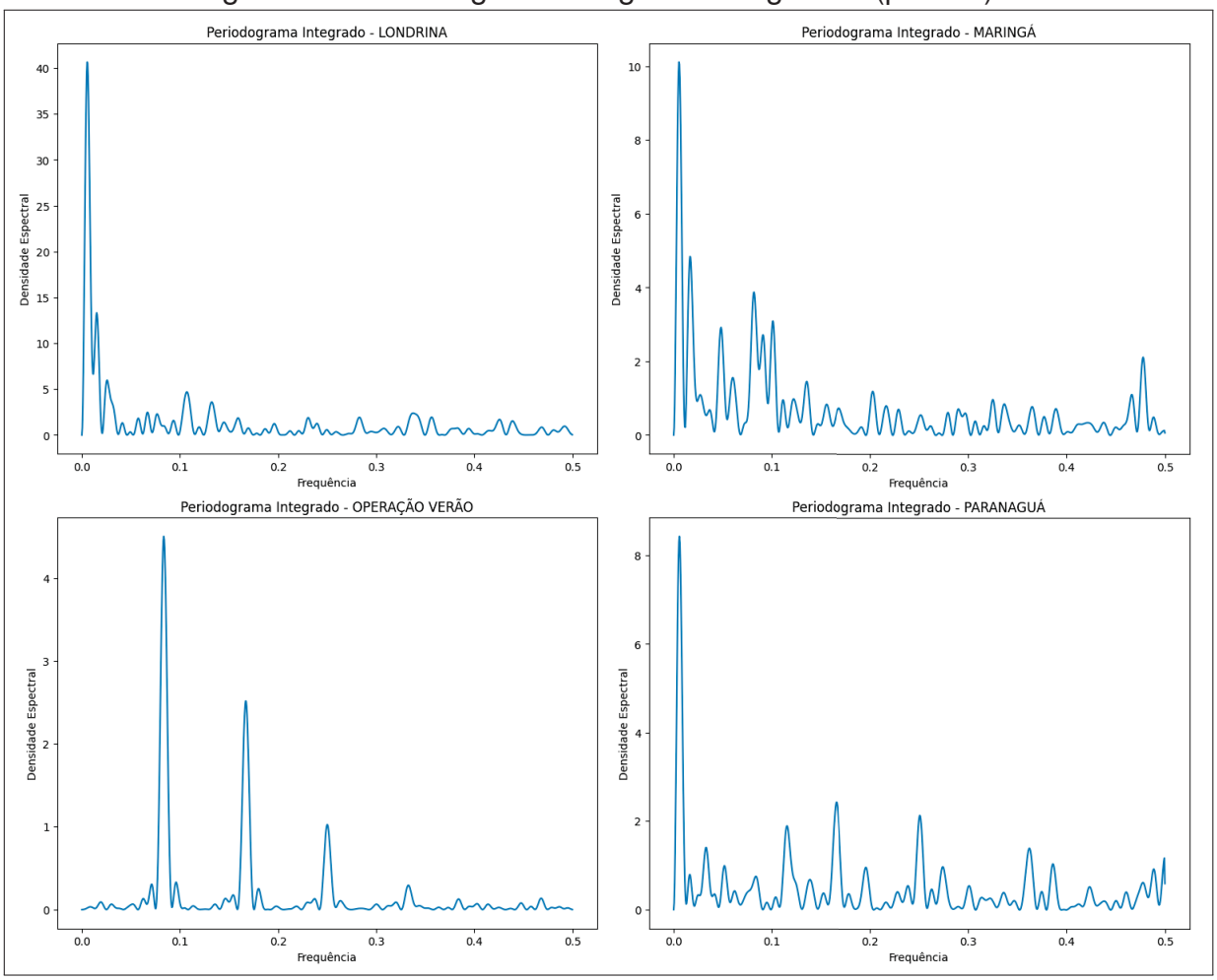
Fonte: O Autor

Figura 32 – Periodograma integrado - Regionais (parte 2).



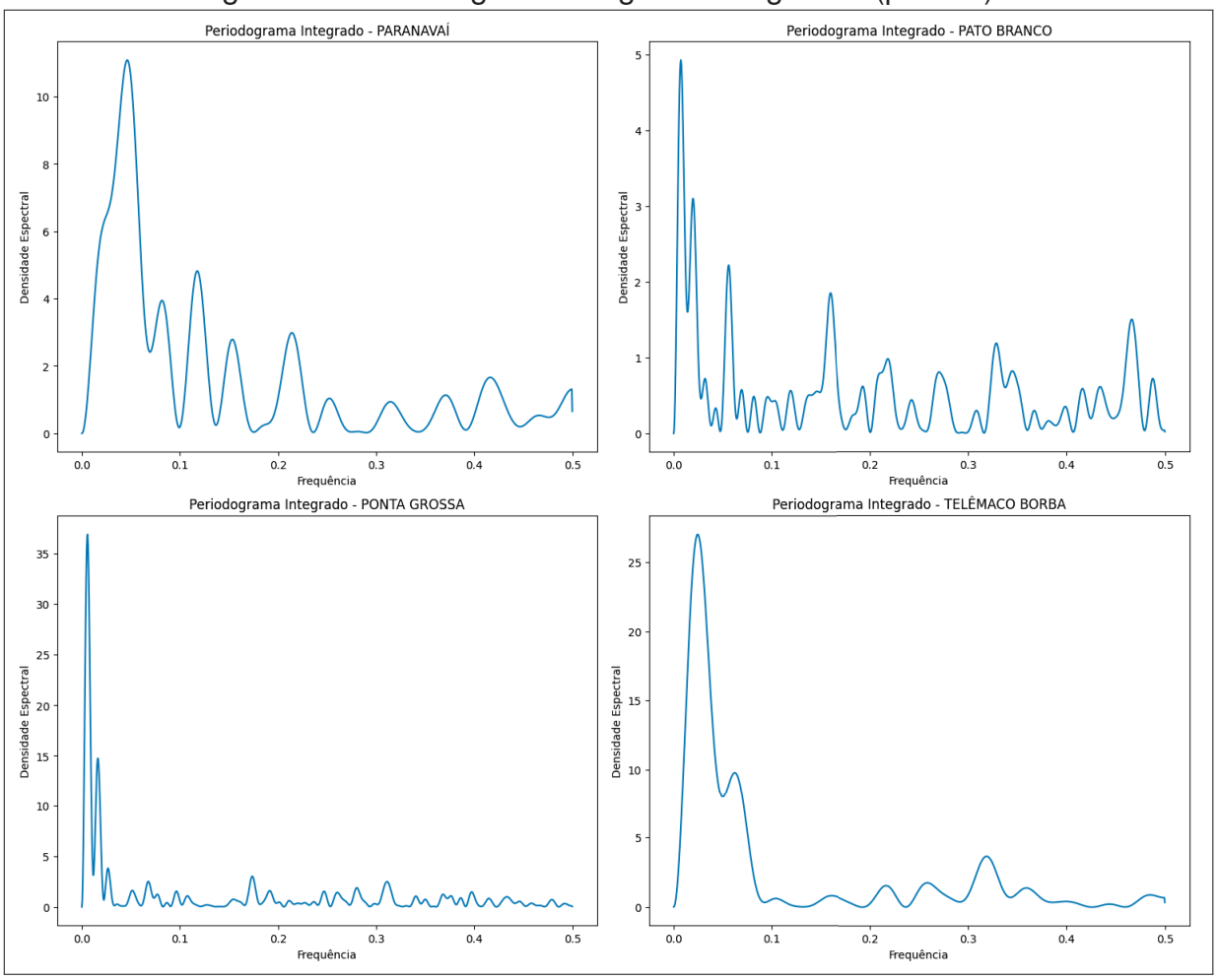
Fonte: O Autor

Figura 33 – Periodograma integrado - Regionais (parte 3).



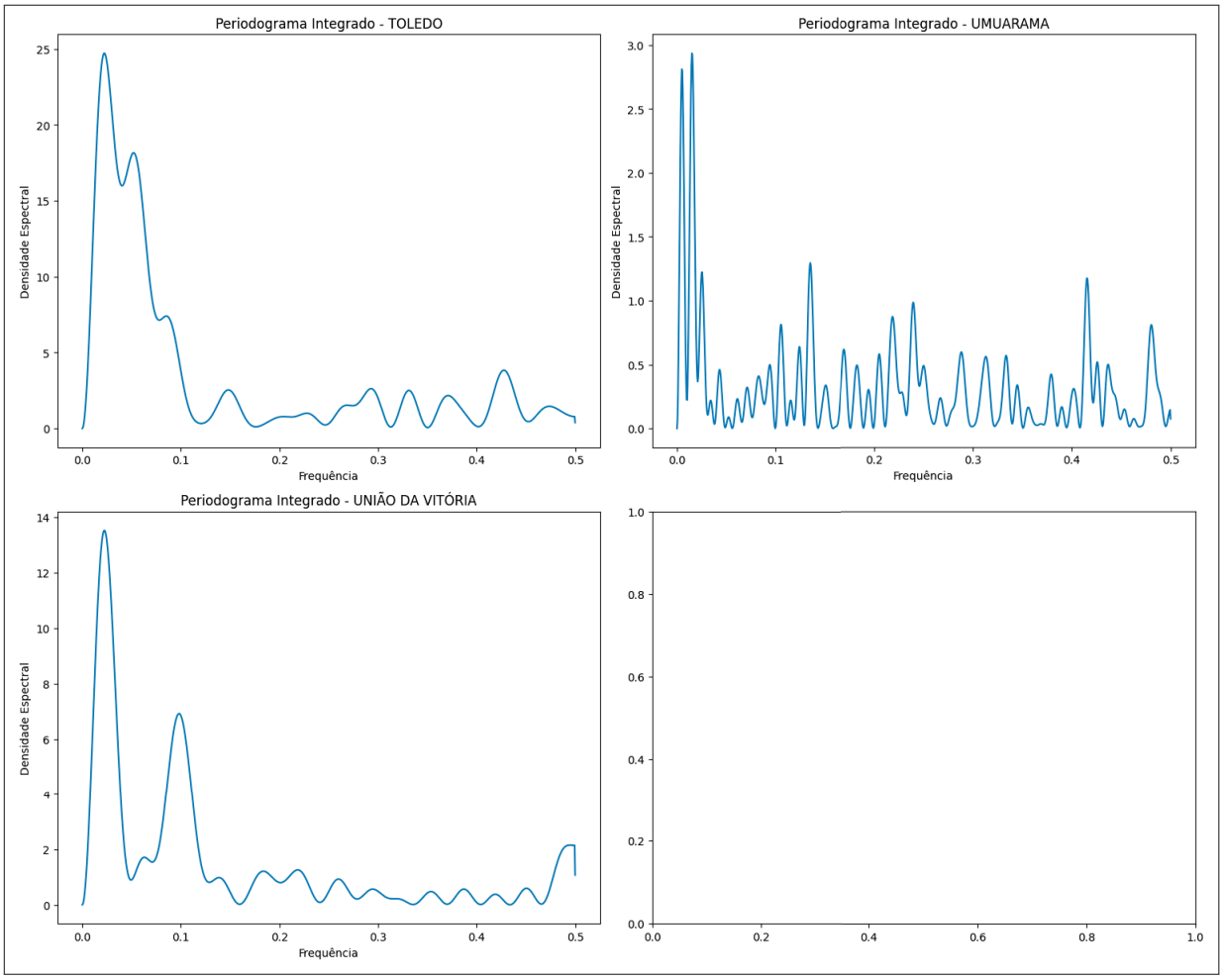
Fonte: O Autor

Figura 34 – Periodograma integrado - Regionais (parte 4).



Fonte: O Autor

Figura 35 – Periodograma integrado - Regionais (parte 5).

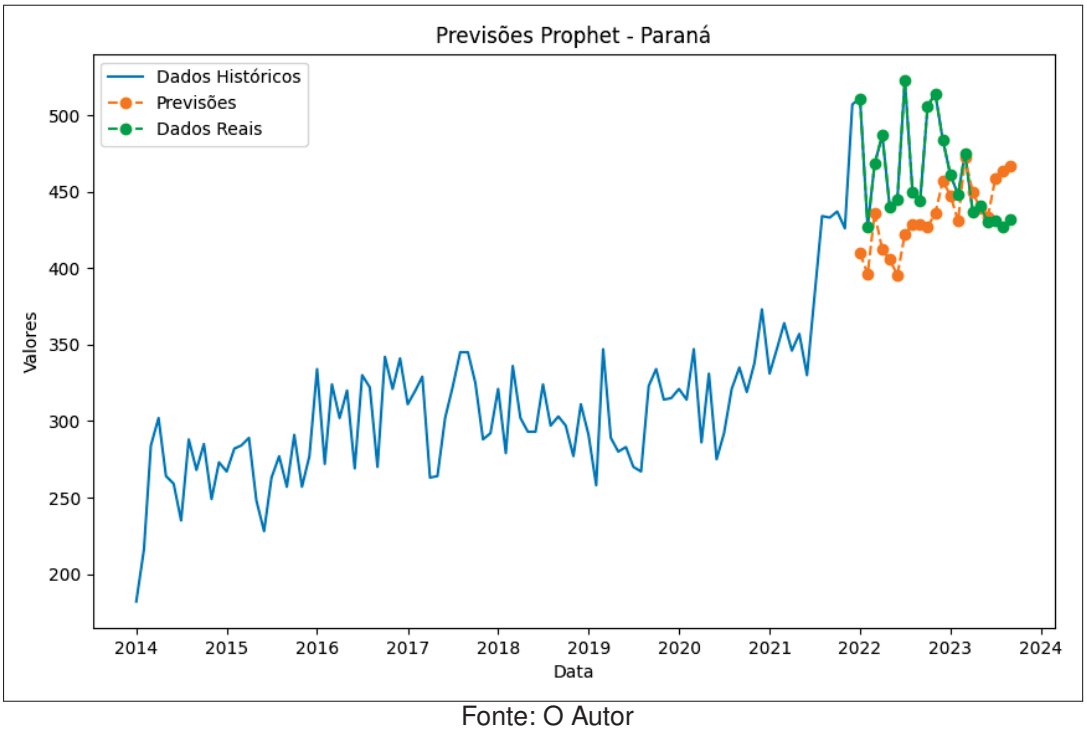


Fonte: O Autor

5.3 PROPHET

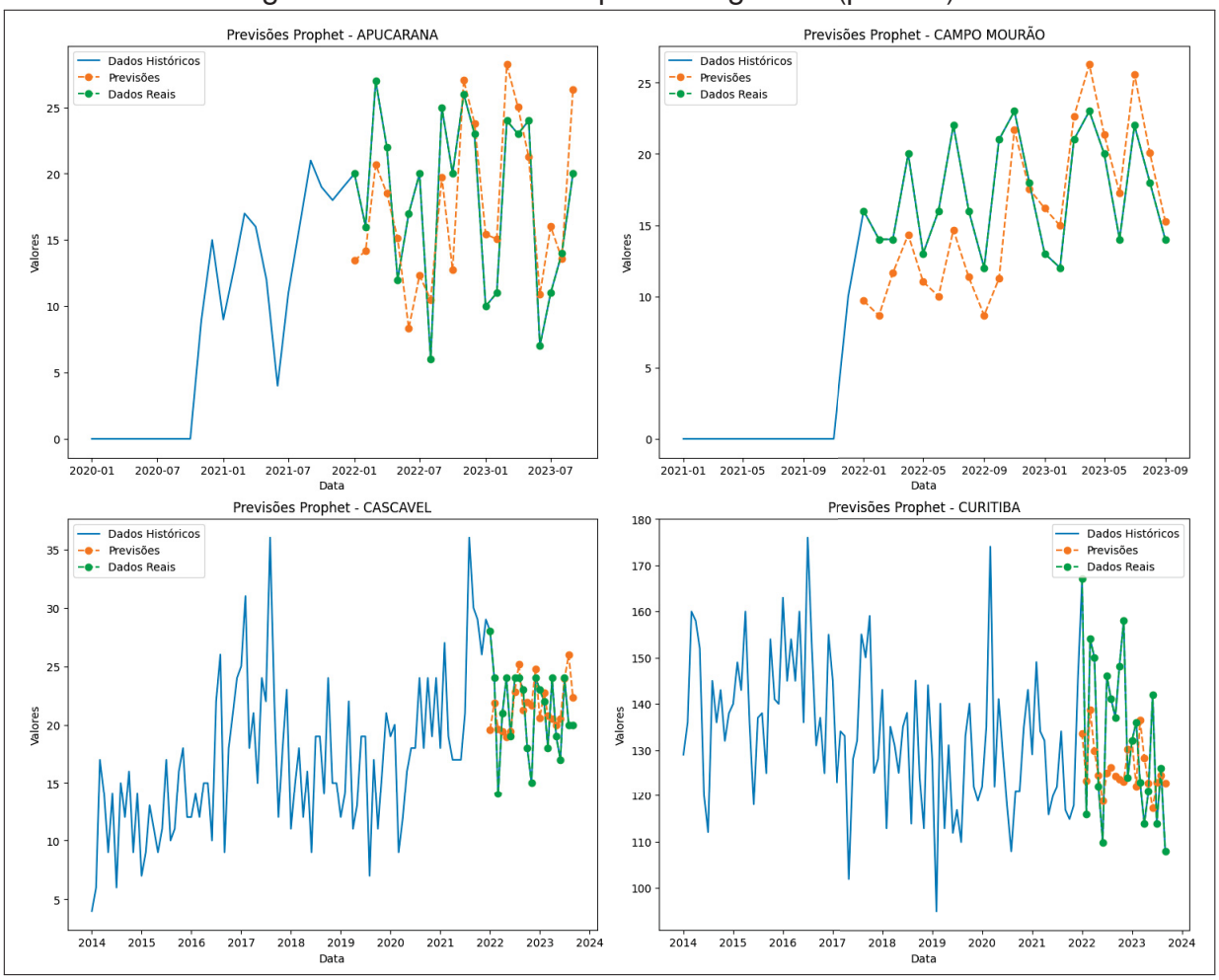
Na Figura 36, é possível observar uma representação visual dos dados históricos de atendimentos no Paraná ao longo dos anos com a previsão gerada pela modelagem *Prophet*. Como parâmetros para a função *Prophet()*, foi utilizada a sazonalidade aditiva, de acordo com o peridograma integrado da subseção anterior, intervalo de confiança de 0.95 e frequência sazonal 'WS', que significa que os dados estão registrados no início de cada mês, que no caso são todos, pois todas as ocorrências de cada mês foram condensadas no primeiro dia do mês no tratamento do *dataframe*. As Figuras 37 - 41 mostram os gráficos deste método para as regionais.

Figura 36 – Prvisões Prophet - Paraná.



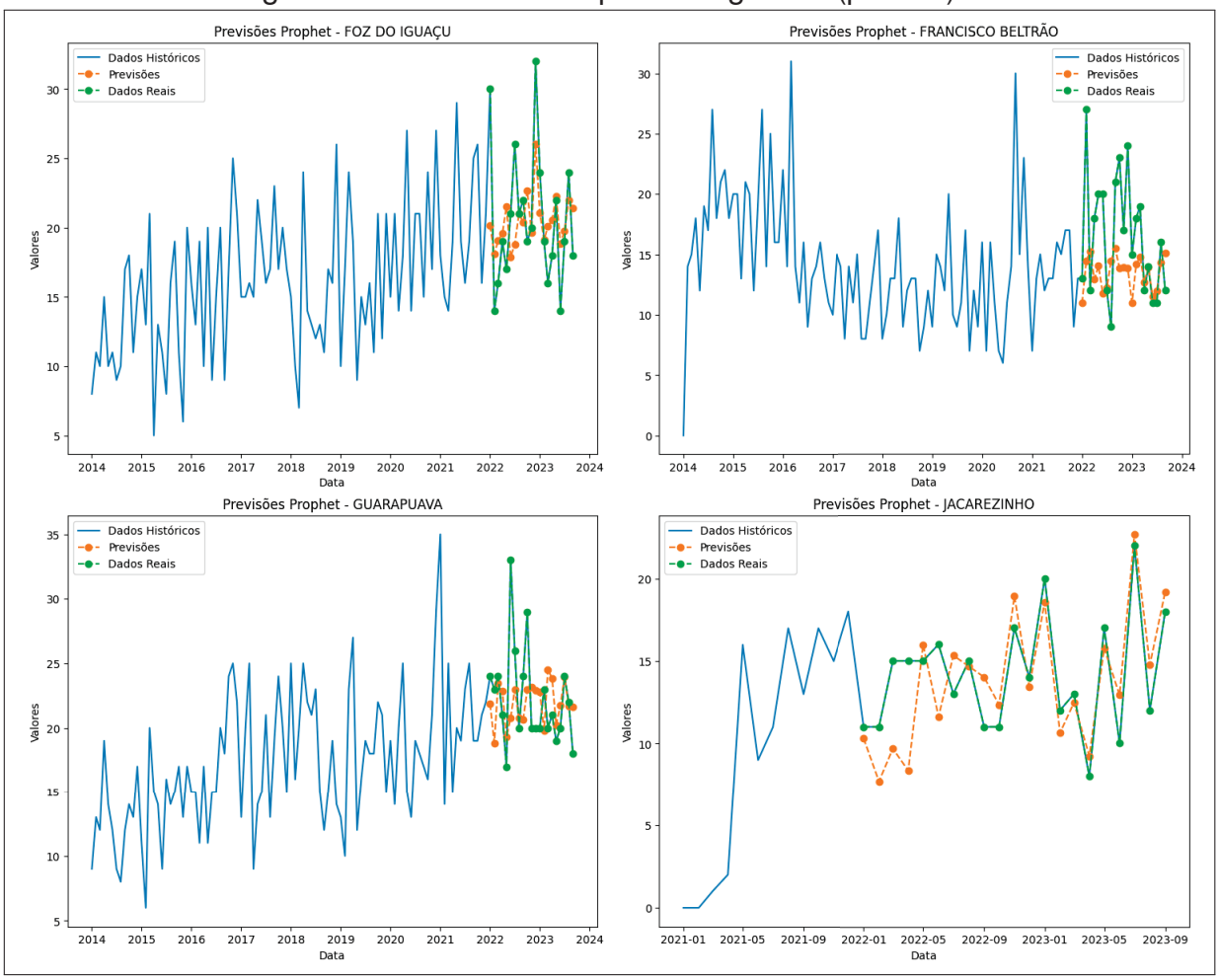
Fonte: O Autor

Figura 37 – Previsões Prophet - Regionais (parte 1).



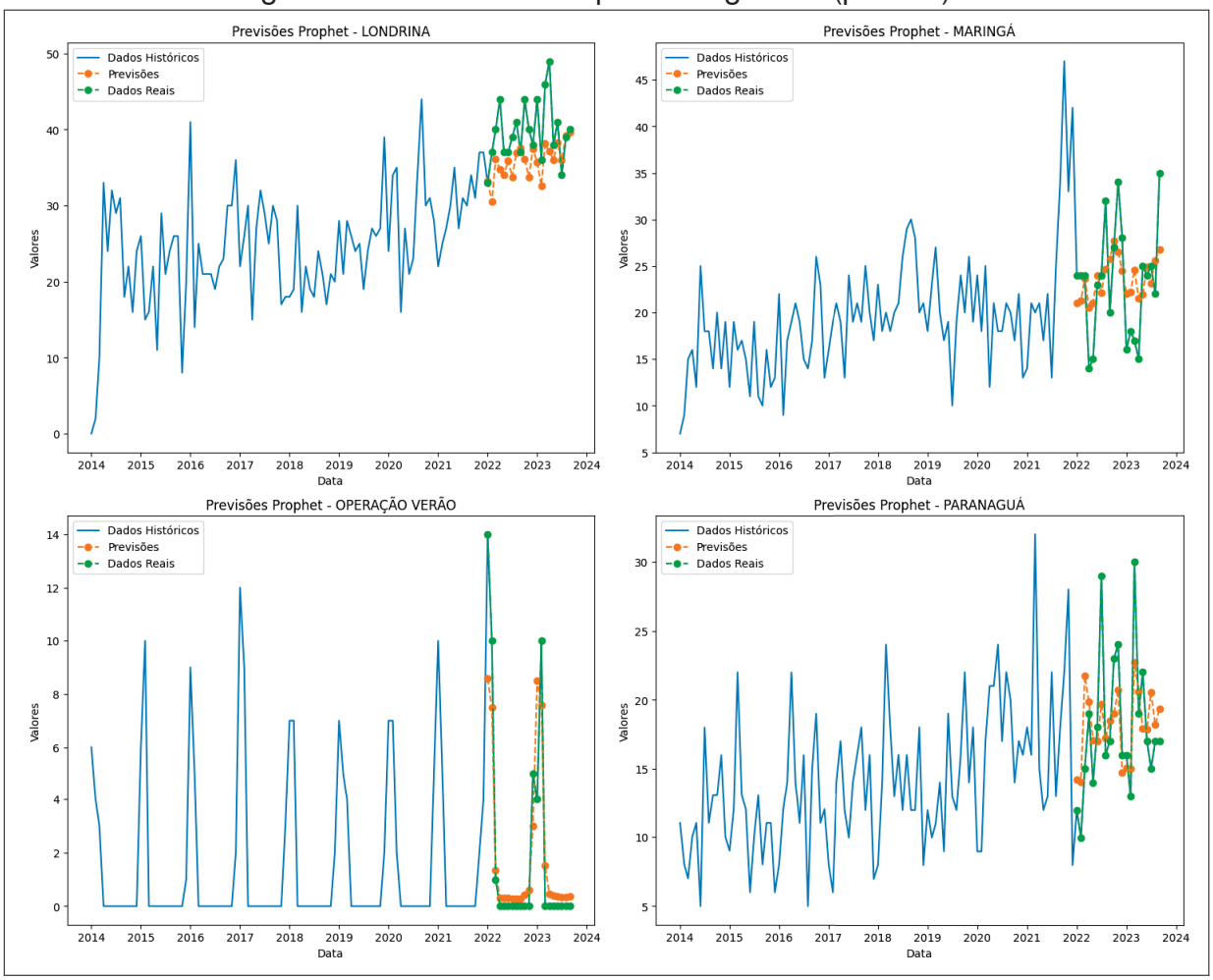
Fonte: O Autor

Figura 38 – Previsões Prophet - Regionais (parte 2).



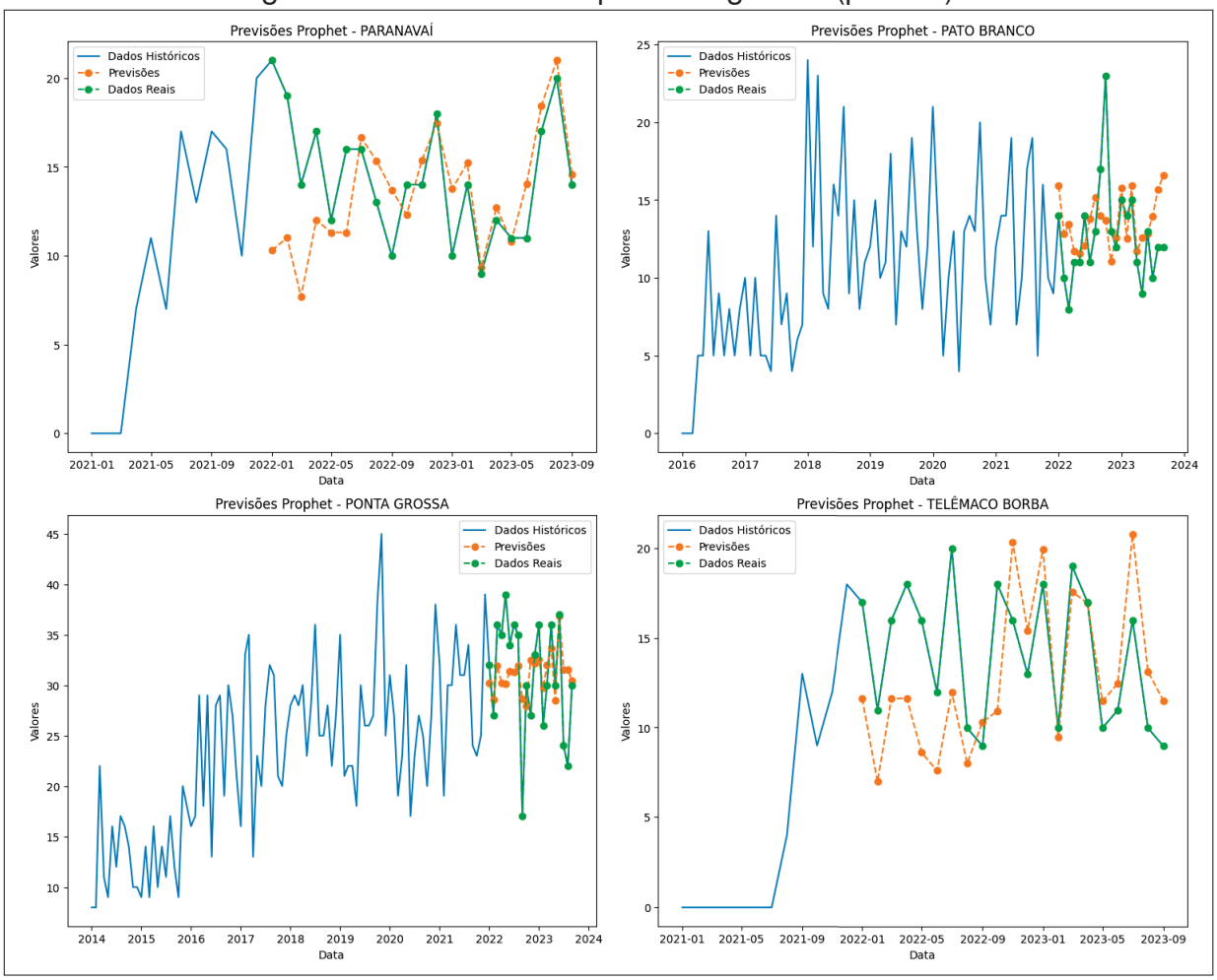
Fonte: O Autor

Figura 39 – Previsões Prophet - Regionais (parte 3).



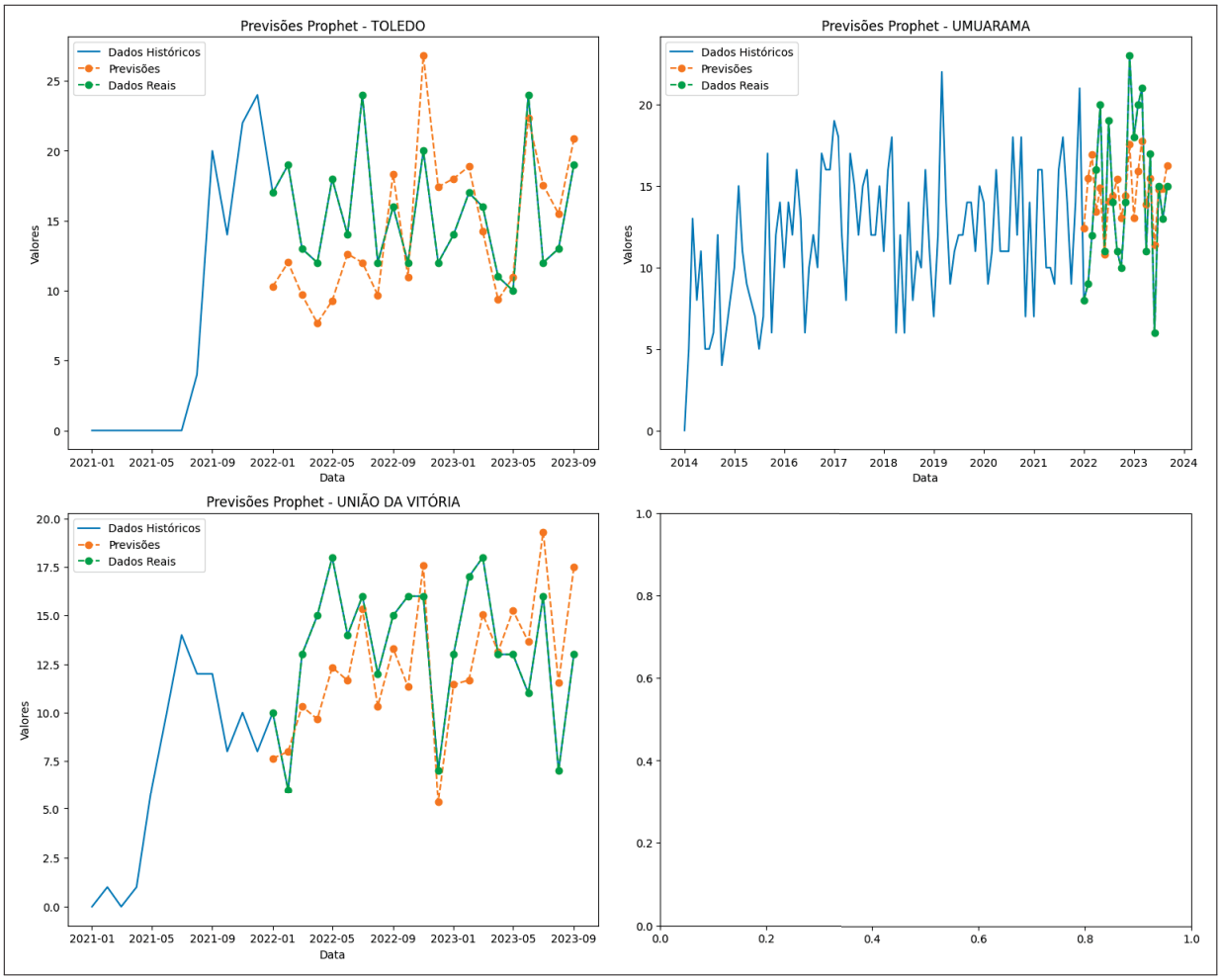
Fonte: O Autor

Figura 40 – Previsões Prophet - Regionais (parte 4).



Fonte: O Autor

Figura 41 – Previsões Prophet - Regionais (parte 5).



Fonte: O Autor

5.3.1 Erro de previsão

Os erros foram calculados usando os dados reais dos anos de 2022 e 2023 (parcial), assim como feito com *Holt-Winters*, comparando-se aos pontos previstos. As métricas MAE, RMSE e MSE foram calculadas. A Tabela 8 mostra os resultados.

Tabela 8 – Métricas da modelagem Prophet do Paraná e Regionais.

Regional	**MAE**	**RMSE**	**MSE**
Apucarana	4.316	4.881	23.824
Campo Mourão	3.674	4.328	18.732
Cascavel	2.908	3.672	13.484
Curitiba	14.028	16.980	288.332
Foz do Iguaçu	3.959	3.095	15.681
Francisco Beltrão	4.266	5.426	29.450
Guarapuava	2.990	3.909	15.283
Jacarezinho	2.105	2.672	7.143
Londrina	4.134	5.323	28.334
Maringá	4.208	4.905	24.060
Operação Verão	1.132	1.826	3.336
Paranaguá	3.063	3.858	14.889
Paraná	37.869	48.523	2354.565
Paranavaí	2.761	3.899	15.203
Pato Branco	2.521	3.269	10.691
Ponta Grossa	3.911	4.969	24.698
Telêmaco Borba	3.536	4.203	17.669
Toledo	3.957	4.890	23.915
Umuarama	3.220	3.764	14.169
União da Vitória	2.833	3.229	10.431

5.4 COMPARAÇÃO DOS MÉTODOS

A Tabela 9 mostra os cálculos da média, desvio padrão e mediana das métricas entre *Holt-Winters Aditivo* e *Prophet*.

Tabela 9 – Comparação das métricas entre Holt-Winters Aditivo e Multiplicativo nas Regionais.

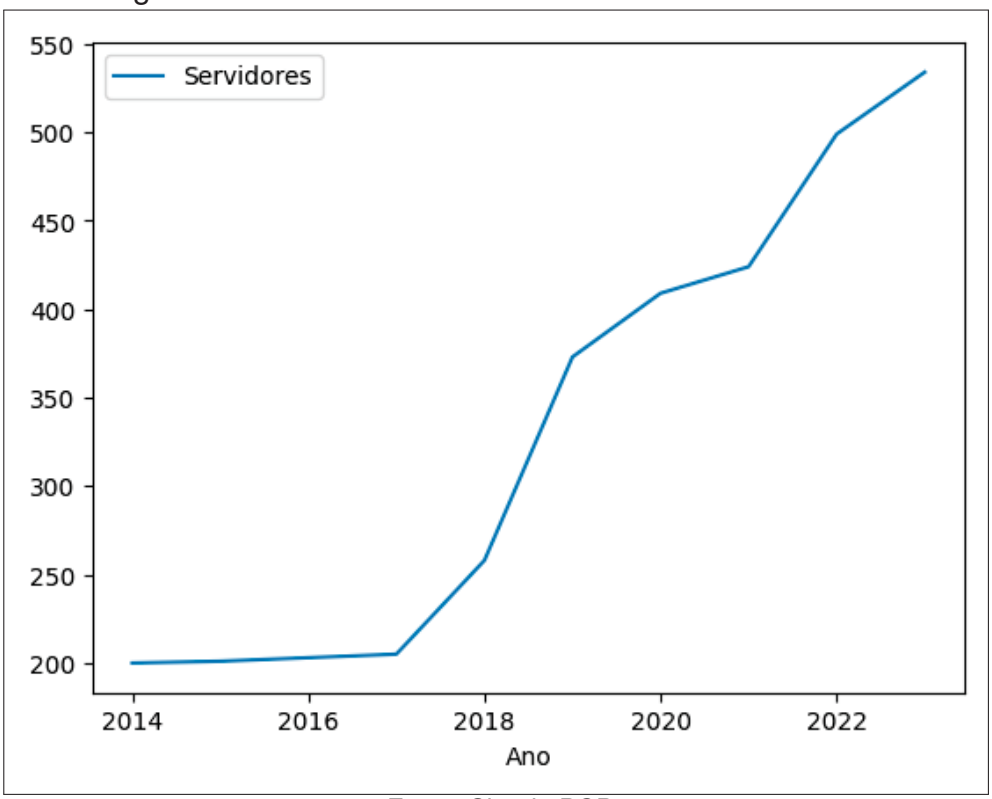
	MAE			**RMSE**		
	Média	**Desvio Padrão**	**Mediana**	**Média**	**Desvio Padrão**	**Mediana**
Holt-Winters Aditivo	6.857	5.987	4.580	8.243	7.281	5.933
Prophet	5.569	8.011	3.605	6.881	10.261	4.056

Nota-se que não há diferença significativa entre os resultados. Portanto, optou-se por aplicar o método *Prophet* na previsão do número de peritos, dado na Seção 5.5.

5.5 APLICAÇÃO DO MÉTODO PARA PREVER O NÚMERO DE PERITOS

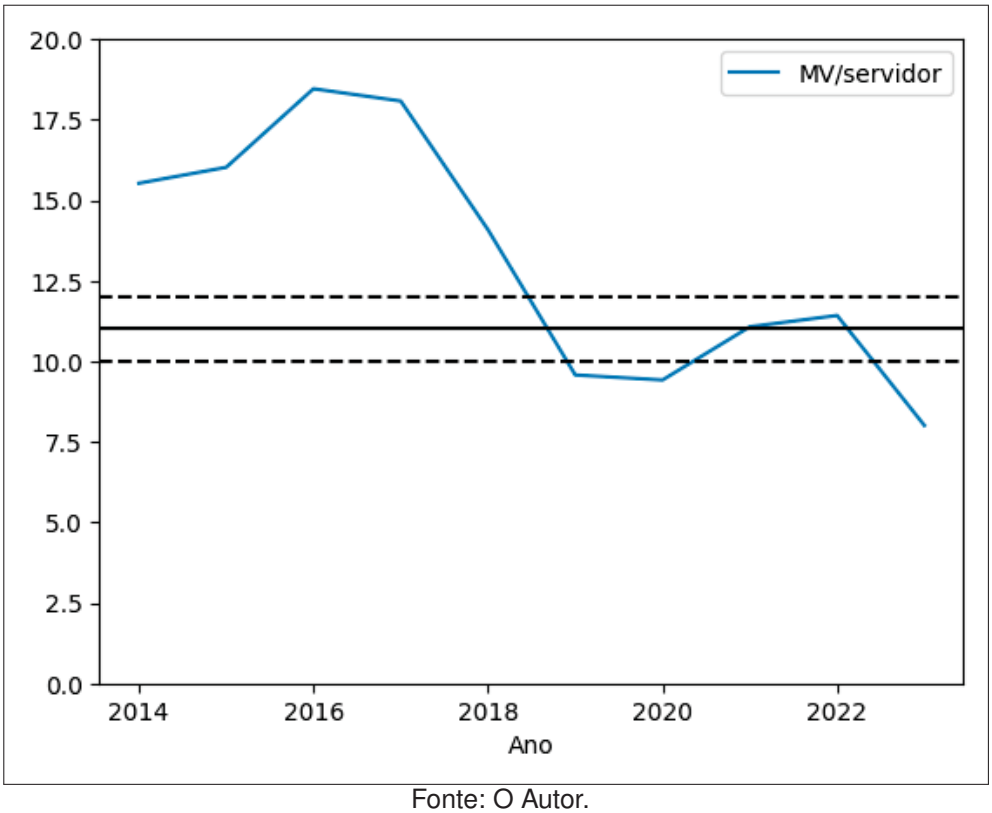
A partir do número histórico de servidores (Figura 42) e a quantidade de MV, calculou-se a razão MV/servidor. Essa razão variou ao longo do tempo segundo o gráfico da Figura 43. Em 09/02/2024, a quantidade de servidores atuando na Polícia Científica, segundo o Portal da Transparência do Governo do Paraná, era de 570 servidores, incluindo peritos, legistas e auxiliares. Entretando, dos 570 servidores, 301 são peritos oficias de 40h (peritos criminais), que podem estar ou não atuando na atividade-fim.

Figura 42 – Número de servidores ao final de cada ano.



Fonte: Site da PCP

Figura 43 – MV/servidor ao longo dos anos na PCP.



Fonte: O Autor.

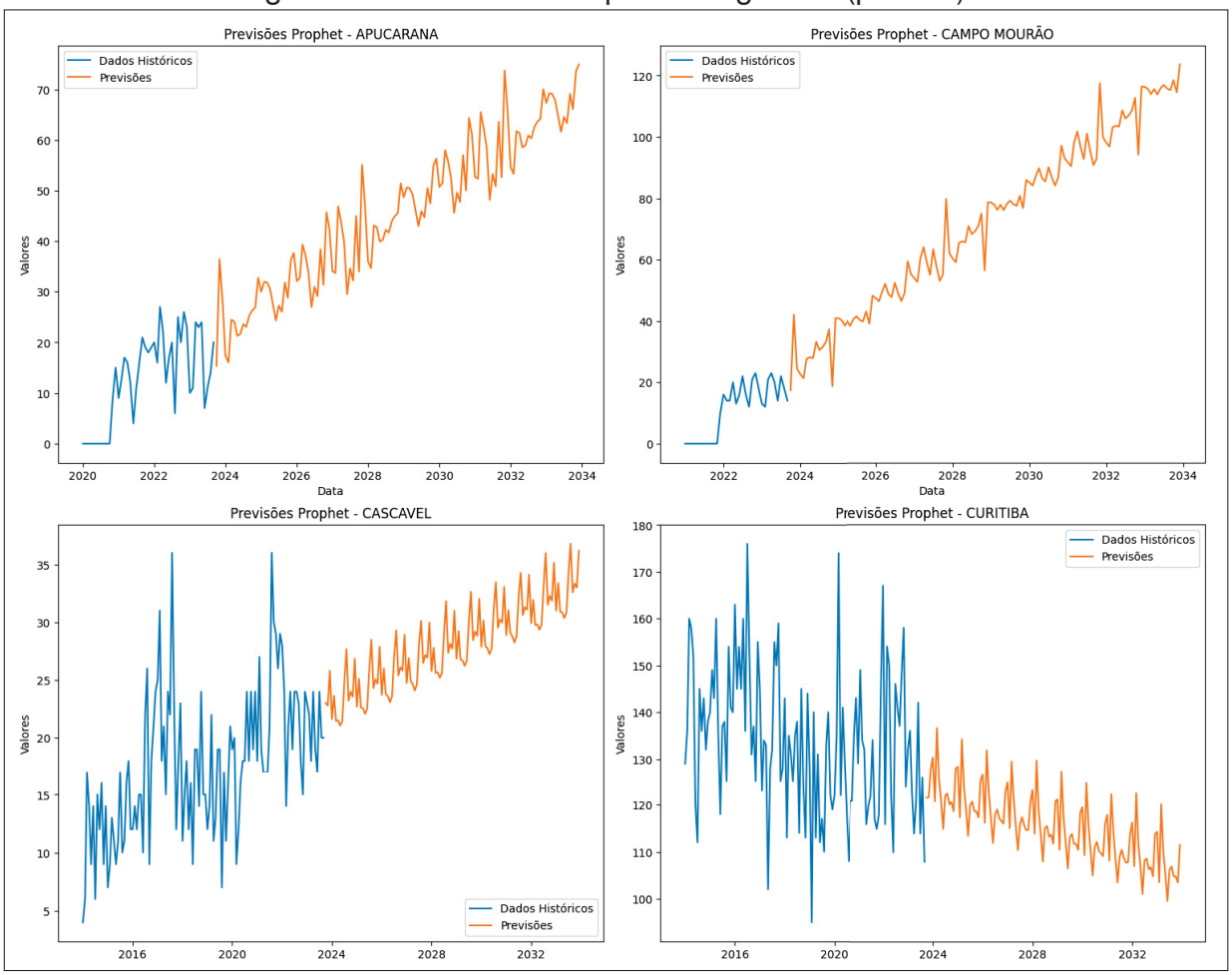
Foram usados os dados de 2019 a 2022 para obter uma estimativa estável de MV/servidor, pois em 2017 houve uma grande contratação de servidores e, conforme o gráfico, houve uma queda de MV/servidor naquele ano e voltando a estabilizar em 2019. A média de MV/servidor de 10,37 foi arrendondada para 11 e usou-se o desvio padrão calculado de 1,0. Com isso, tem-se uma faixa de MV/servidor no período de 2019 a 2022 entre 10 e 12, ou seja, valores de 11 ± 1. No final de 2022, segundo a PCP, eram 499 servidores atuando. Proporcionalmente, calculou-se que 52% sejam peritos criminais, ou seja, 264 peritos criminais.

Para determinar a projeção de peritos necessários nas regionais em 10 anos, verificou-se a média de MV em 2022 das regionais com pelo menos cinco anos de existência e calculou-se a razão desta média por 5 peritos, que é a quantidade mínima de peritos para cobrir a escala de uma regional, resultando em 61 MV/perito. Os cálculos foram feitos para que a quantidade peritos ao final de 2033 esteja nesta razão com a quantidade de MV. A Tabela 10 mostra os resultados da projeção da quantidade de peritos nas Regionais (com arredondamento para cima) e as Figuras 44 a 48 mostram as projeções de MV usando o *Prophet*. Para esta tabela, somente as regionais com pelo menos cinco anos de existência foram consideradas. Regionais muito novas, como mostram os gráficos, possuem uma previsão superestimada devido a ausência de dados.

Tabela 10 – Projeção da quantidade de peritos nas Regionais em 2033.

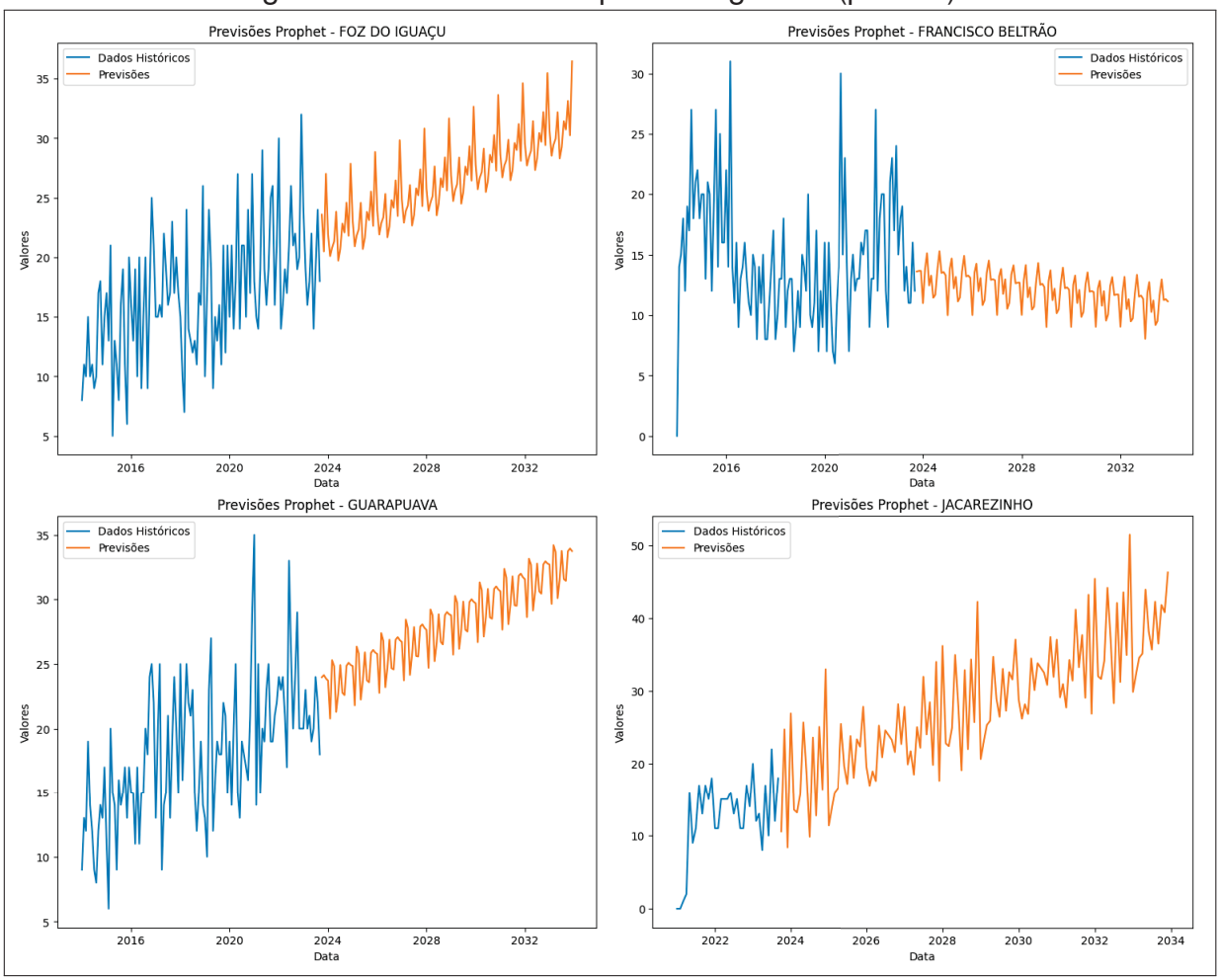
Regional	**MV**	**Nº de peritos**
Cascavel	393	7
Curitiba	1291	22
Foz do Iguaçu	370	7
Francisco Beltrão	130	5
Guarapuava	390	7
Londrina	710	12
Maringá	411	7
Operação Verão	34	5
Paranaguá	330	6
Pato Branco	260	5
Ponta Grossa	549	9
Umuarama	250	5

Figura 44 – Previsões Prophet - Regionais (parte 1).



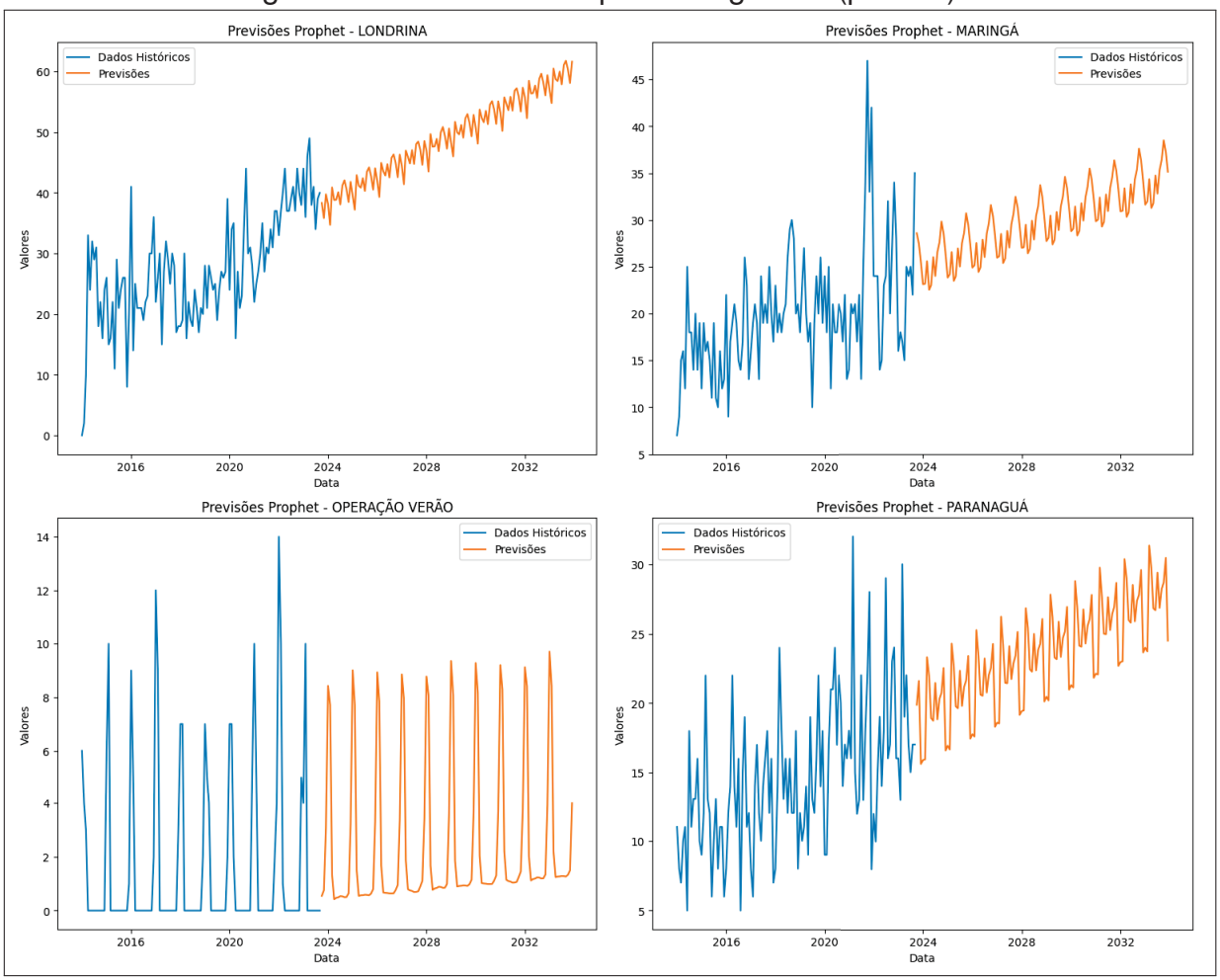
Fonte: O Autor

Figura 45 – Previsões Prophet - Regionais (parte 2).



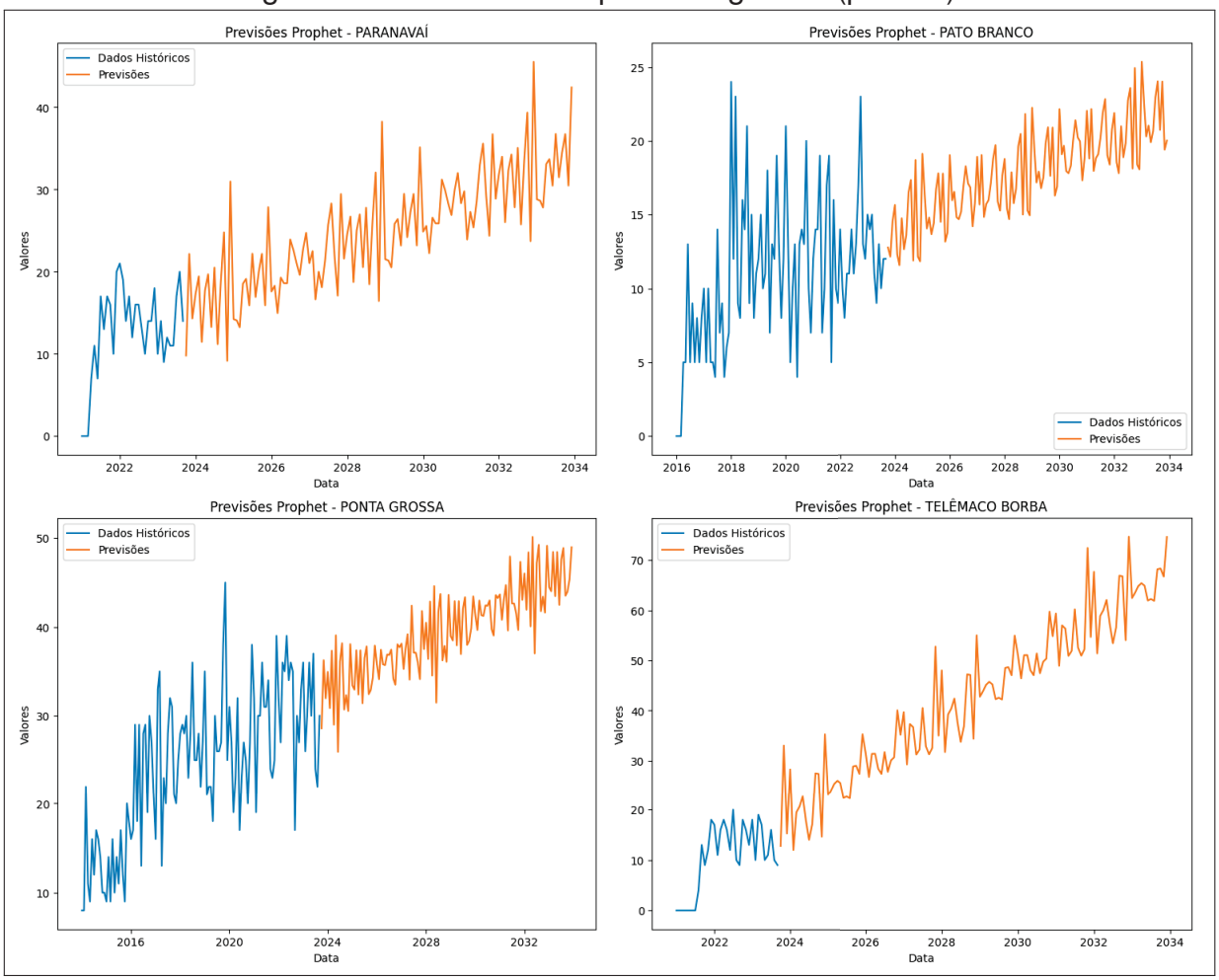
Fonte: O Autor

Figura 46 – Previsões Prophet - Regionais (parte 3).



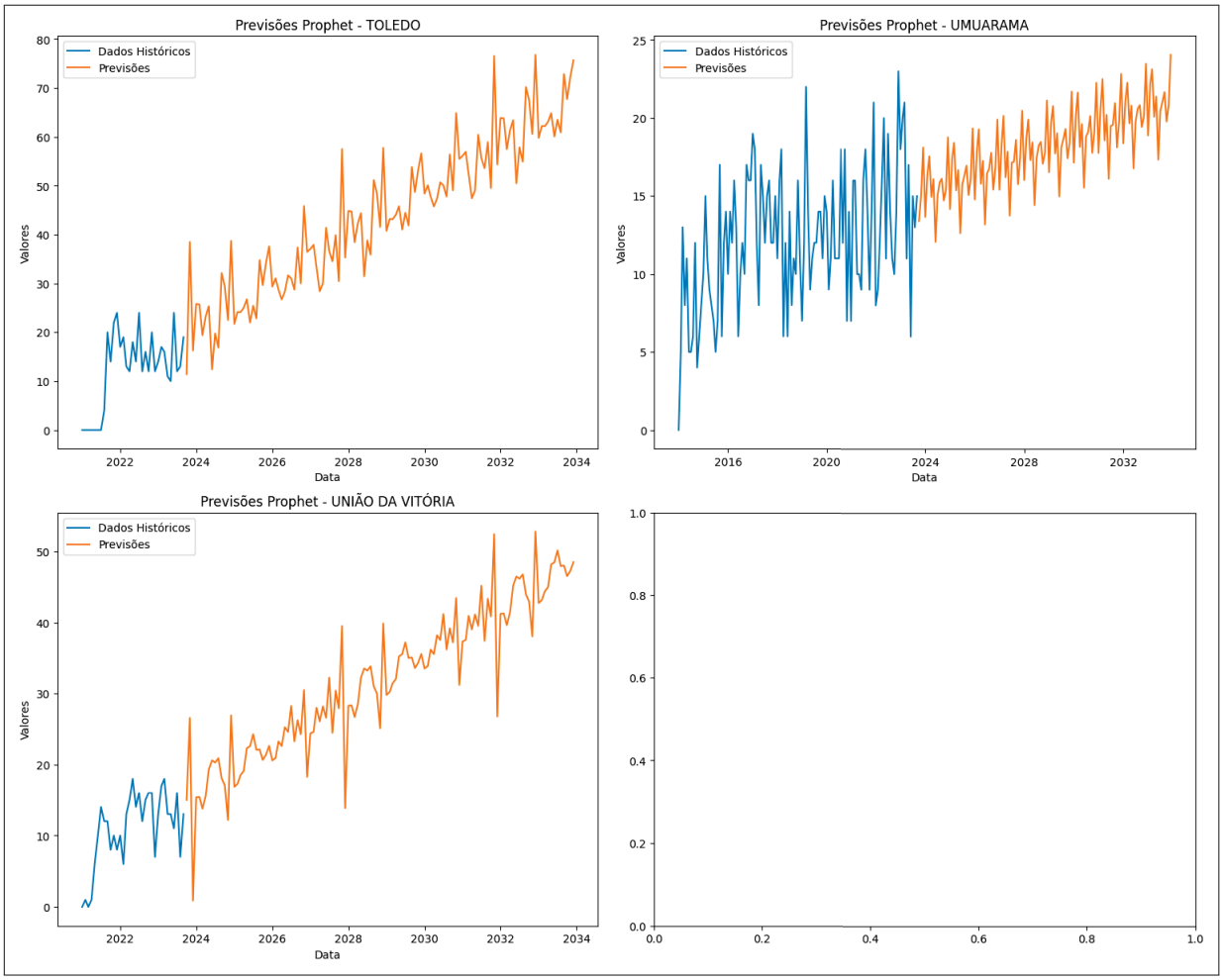
Fonte: O Autor

Figura 47 – Previsões Prophet - Regionais (parte 4).



Fonte: O Autor

Figura 48 – Previsões Prophet - Regionais (parte 5).



Fonte: O Autor

6 CONCLUSÃO

O trabalho adotou uma abordagem semelhante à metodologia de Ciência de Dados da IBM, como descrito por Rollins (2015), para obter, tratar e interpretar dados. Foram realizadas análises estatísticas das ocorrências de Mortes Violentas (MV) ao longo dos anos, incluindo médias, desvios padrão, e distribuição por regional. Métodos de suavização exponencial, como Holt-Winters, foram aplicados e comparados, sendo escolhido o modelo que melhor se ajustou aos dados. O processo incluiu a análise do modelo de atuação da Polícia Científica do Paraná, a identificação da situação-problema e a escolha de ferramentas analíticas, incluindo ferramenta Prophet para modelagem.

Salienta-se que a abordagem analisou somente as MV, sem estimar ou considerar as outras perícias e a atuação de seções específicas, como os laboratórios. Os resultados obtidos mostram uma expansão conservadora quanto ao número de peritos necessários em dez anos, pois apesar da relação de MV/perito ter descrescido, não é possível dizer se a média de 61 MV/perito em 2022 é a razão ideal. Pode ser que os peritos já estejam sobrecarregados ou subutilizados nesta data.

O trabalho reconhece algumas limitações, como a superestimação em regionais muito novas devido à falta de dados históricos. No entanto, as análises estatísticas e modelagens fornecem percepções valiosas sobre as ocorrências de MV e as necessidades de pessoal na Polícia Científica do Paraná.

REFERÊNCIAS

AHMED, R.; SREERAM, V.; MISHRA, Y.; ARIF, M. A review and evaluation of the state-of-the-art in pv solar power forecasting: Techniques and optimization. **Renewable and Sustainable Energy Reviews**, v. 124, p. 109792, 2020. ISSN 1364-0321. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1364032120300885>.

AKHTER, M. N.; MEKHILEF, S.; MOKHLIS, H.; SHAH, N. M. Review on forecasting of photovoltaic power generation based on machine learning and metaheuristic techniques. **IET Renewable Power Generation**, Wiley Online Library, v. 13, n. 7, p. 1009–1023, 2019.

ARIA, M.; CUCCURULLO, C. bibliometrix: An r-tool for comprehensive science mapping analysis. **Journal of informetrics**, Elsevier, v. 11, n. 4, p. 959–975, 2017.

BARROS, A. P. d. **Modelos aditivos generalizados com defasagens distribuídas**. Tese (Doutorado) — Universidade de São Paulo, 2002.

BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C.; LJUNG, G. M. **Time Series Analysis**: Forecasting and control. Hoboken, New Jersey: John Wiley and Sons Inc., 2015.

GHIMIRE, S.; DEO, R. C.; RAJ, N.; MI, J. Deep solar radiation forecasting with convolutional neural network and long short-term memory network algorithms. **Applied Energy**, v. 253, p. 113541, 2019. ISSN 0306-2619. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0306261919312152>.

HARVEY, A. C.; PETERS, S. Estimation procedures for structural time series models. **Journal of forecasting**, Wiley Online Library, v. 9, n. 2, p. 89–108, 1990.

HASTIE, T.; TIBSHIRANI, R. Generalized additive models: some applications. **Journal of the American Statistical Association**, Taylor & Francis, v. 82, n. 398, p. 371–386, 1987.

HE, L.; ZHANG, L.; ZHONG, Z.; WANG, D.; WANG, F. Green credit, renewable energy investment and green economy development: Empirical analysis based on 150 listed companies of china. **Journal of Cleaner Production**, v. 208, p. 363–372, 2019. ISSN 0959-6526. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0959652618331354>.

HILL, T.; O'CONNOR, M.; REMUS, W. Neural network models for time series forecasts. **Management science**, INFORMS, v. 42, n. 7, p. 1082–1092, 1996.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural computation**, MIT press, v. 9, n. 8, p. 1735–1780, 1997.

HYNDMAN, R. J. Box-jenkins modelling. In: DAELLENBACH, H.; DAELLENBACH, H.; FLOOD, R. (Ed.). **The Informed Student Guide to Management Science**. Thomson Learning, 2002, (Informed student guides). ISBN 9781861525420. Disponível em: <https://books.google.com.br/books?id=mU8kqbzrssoC>.

HYNDMAN, R. J.; ATHANASOPOULOS, G. **Forecasting: principles and practice**. 3. ed. Melbourne, Australia: OTexts, 2021. Disponível em: <https://otexts.com/fpp3/>.

HYNDMAN, R. J.; KHANDAKAR, Y. Automatic time series forecasting: the forecast package for r. **Journal of statistical software**, v. 27, p. 1–22, 2008.

HYNDMAN, R. J.; KOEHLER, A. B.; SNYDER, R. D.; GROSE, S. A state space framework for automatic forecasting using exponential smoothing methods. **International Journal of forecasting**, Elsevier, v. 18, n. 3, p. 439–454, 2002.

LIM, C.; MCALEER, M. Time series forecasts of international travel demand for australia. **Tourism Management**, Elsevier, v. 23, n. 4, p. 389–396, 2002.

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 & Francis, v. 106, n. 496, p. 1513–1527, 2011.

LONG, W.; LU, Z.; CUI, L. Deep learning-based feature engineering for stock price movement prediction. **Knowledge-Based Systems**, Elsevier, v. 164, p. 163–173, 2019.

MAKRIDAKIS, S.; SPILIOTIS, E.; ASSIMAKOPOULOS, V. M5 accuracy competition: Results, findings, and conclusions. **International Journal of Forecasting**, Elsevier, v. 38, n. 4, p. 1346–1364, 2022.

PETROPOULOS, F.; APILETTI, D.; ASSIMAKOPOULOS, V.; BABAI, M. Z.; BARROW, D. K.; BEN TAIEB, S.; BERGMEIR, C.; BESSA, R. J.; BIJAK, J.; BOYLAN, J. E.; BROWELL, J.; CARNEVALE, C.; CASTLE, J. L.; CIRILLO, P.; CLEMENTS, M. P.; CORDEIRO, C.; CYRINO OLIVEIRA, F. L.; DE BAETS, S.; DOKUMENTOV, A.; ELLISON, J.; FISZEDER, P.; FRANSES, P. H.; FRAZIER, D. T.; GILLILAND, M.; GÖNÜL, M. S.; GOODWIN, P.; GROSSI, L.; GRUSHKA-COCKAYNE, Y.; GUIDOLIN, M.; GUIDOLIN, M.; GUNTER, U.; GUO, X.; GUSEO, R.; HARVEY, N.; HENDRY, D. F.; HOLLYMAN, R.; JANUSCHOWSKI, T.; JEON, J.; JOSE, V. R. R.; KANG, Y.; KOEHLER, A. B.; KOLASSA, S.; KOURENTZES, N.; LEVA, S.; LI, F.; LITSIOU, K.; MAKRIDAKIS, S.; MARTIN, G. M.; MARTINEZ, A. B.; MEERAN, S.; MODIS, T.; NIKOLOPOULOS, K.; ÖNKAL, D.; PACCAGNINI, A.; PANAGIOTELIS, A.; PANAPAKIDIS, I.; PAVíA, J. M.; PEDIO, M.; PEDREGAL, D. J.; PINSON, P.; RAMOS, P.; RAPACH, D. E.; READE, J. J.; ROSTAMI-TABAR, B.; RUBASZEK, M.; SERMPINIS, G.; SHANG, H. L.; SPILIOTIS, E.; SYNTETOS, A. A.; TALAGALA, P. D.; TALAGALA, T. S.; TASHMAN, L.; THOMAKOS, D.; THORARINSDOTTIR, T.; TODINI, E.; TRAPERO ARENAS, J. R.; WANG, X.; WINKLER, R. L.; YUSUPOVA, A.; ZIEL, F. Forecasting: theory and practice. **International Journal of Forecasting**, v. 38, n. 3, p. 705–871, 2022. ISSN 0169-2070. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0169207021001758>.

QING, X.; NIU, Y. Hourly day-ahead solar irradiance prediction using weather forecasts by lstm. **Energy**, Elsevier, v. 148, p. 461–468, 2018.

REIKARD, G. Predicting solar radiation at high resolutions: A comparison of time series forecasts. **Solar energy**, Elsevier, v. 83, n. 3, p. 342–349, 2009.

ROLLINS, J. B. **Metodologia de Base para Ciência de Dados**. [S.l.]: IBM Corporation, 2015. Https://www.ibm.com/downloads/cas/B1WQ0GM2.

SIMPSON, G. **Statistical Methods Seminar Series**. [S.l.]: GitHub, 2022. <https://github.com/eco4cast/Statistical-Methods-Seminar-Series/tree/main/simpson-gams>.

SONG, H.; QIU, R. T.; PARK, J. A review of research on tourism demand forecasting: Launching the annals of tourism research curated collection on tourism demand forecasting. **Annals of Tourism Research**, v. 75, p. 338–362, 2019. ISSN 0160-7383. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0160738318301312>.

SUN, S.; WEI, Y.; TSUI, K.-L.; WANG, S. Forecasting tourist arrivals with machine learning and internet search index. **Tourism Management**, v. 70, p. 1–10, 2019. ISSN 0261-5177. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0261517718301572>.

TAYLOR, S. J.; LETHAM, B. Forecasting at scale. **The American Statistician**, Taylor & Francis, v. 72, n. 1, p. 37–45, 2018.

WANG, Q.; JIN, G.; ZHAO, X.; FENG, Y.; HUANG, J. Csan: A neural network benchmark model for crime forecasting in spatio-temporal scale. **Knowledge-Based Systems**, Elsevier, v. 189, p. 105120, 2020.

WANG, Q.; LI, S.; LI, R. Forecasting energy demand in china and india: Using single-linear, hybrid-linear, and non-linear time series forecast techniques. **Energy**, Elsevier, v. 161, p. 821–831, 2018.

WOOD, S. N. **Generalized additive models: an introduction with R**. [S.l.]: CRC press, 2017.

ZANG, H.; LIU, L.; SUN, L.; CHENG, L.; WEI, Z.; SUN, G. Short-term global horizontal irradiance forecasting based on a hybrid cnn-lstm model with spatiotemporal correlations. **Renewable Energy**, v. 160, p. 26–41, 2020. ISSN 0960-1481. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0960148120308557>.

ZHANG, G. P. Time series forecasting using a hybrid arima and neural network model. **Neurocomputing**, Elsevier, v. 50, p. 159–175, 2003.

ZHANG, H.; SONG, H.; WEN, L.; LIU, C. Forecasting tourism recovery amid covid-19. **Annals of Tourism Research**, Elsevier, v. 87, p. 103149, 2021.

ZHANG, X.; LIU, L.; XIAO, L.; JI, J. Comparison of machine learning algorithms for predicting crime hotspots. **IEEE access**, IEEE, v. 8, p. 181302–181310, 2020.

ANEXO A - AMOSTRA DE MV.CSV

Regional,Mês,Contagem de cod_rep,Ano,Natureza do Exame (cod)
CASCAVEL,janeiro,1,2014,A460 – EXAME EM LOCAL DE ATROPELAMENTO E MORTE
CASCAVEL,fevereiro,0,2014,A460 – EXAME EM LOCAL DE ATROPELAMENTO E MORTE
CASCAVEL,março,1,2014,A460 – EXAME EM LOCAL DE ATROPELAMENTO E MORTE
CASCAVEL,abril,0,2014,A460 – EXAME EM LOCAL DE ATROPELAMENTO E MORTE
CASCAVEL,maio,0,2014,A460 – EXAME EM LOCAL DE ATROPELAMENTO E MORTE
CASCAVEL,junho,1,2014,A460 – EXAME EM LOCAL DE ATROPELAMENTO E MORTE
CASCAVEL,julho,0,2014,A460 – EXAME EM LOCAL DE ATROPELAMENTO E MORTE
CASCAVEL,agosto,1,2014,A460 – EXAME EM LOCAL DE ATROPELAMENTO E MORTE
CASCAVEL,setembro,1,2014,A460 – EXAME EM LOCAL DE ATROPELAMENTO E MORTE
CASCAVEL,outubro,0,2014,A460 – EXAME EM LOCAL DE ATROPELAMENTO E MORTE
CASCAVEL,novembro,1,2014,A460 – EXAME EM LOCAL DE ATROPELAMENTO E MORTE
CASCAVEL,dezembro,1,2014,A460 – EXAME EM LOCAL DE ATROPELAMENTO E MORTE
CURITIBA,janeiro,4,2014,A460 – EXAME EM LOCAL DE ATROPELAMENTO E MORTE
CURITIBA,fevereiro,5,2014,A460 – EXAME EM LOCAL DE ATROPELAMENTO E MORTE
CURITIBA,março,9,2014,A460 – EXAME EM LOCAL DE ATROPELAMENTO E MORTE
CURITIBA,abril,6,2014,A460 – EXAME EM LOCAL DE ATROPELAMENTO E MORTE
CURITIBA,maio,8,2014,A460 – EXAME EM LOCAL DE ATROPELAMENTO E MORTE
CURITIBA,junho,3,2014,A460 – EXAME EM LOCAL DE ATROPELAMENTO E MORTE
CURITIBA,julho,5,2014,A460 – EXAME EM LOCAL DE ATROPELAMENTO E MORTE
CURITIBA,agosto,4,2014,A460 – EXAME EM LOCAL DE ATROPELAMENTO E MORTE
CURITIBA,setembro,4,2014,A460 – EXAME EM LOCAL DE ATROPELAMENTO E MORTE
CURITIBA,outubro,4,2014,A460 – EXAME EM LOCAL DE ATROPELAMENTO E MORTE
CURITIBA,novembro,9,2014,A460 – EXAME EM LOCAL DE ATROPELAMENTO E MORTE
CURITIBA,dezembro,8,2014,A460 – EXAME EM LOCAL DE ATROPELAMENTO E MORTE