DIEGO CARDOSO

A concentração de ozônio em uma microescala na cidade de São Paulo: análise exploratória de dados e modelagem

São Paulo
2024

DIEGO CARDOSO

A concentração de ozônio em uma microescala na cidade de São Paulo: análise exploratória de dados e modelagem

Versão Corrigida

Dissertação apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do Título de Mestre em Ciências

São Paulo
2024

DIEGO CARDOSO

A concentração de ozônio em uma microescala na cidade de São Paulo: análise exploratória de dados e modelagem

Versão Corrigida

Dissertação apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do Título de Mestre em Ciências

Área de Concentração: Sistemas Eletrônicos

Orientadora: Profa. Dra. Maria D. Miranda

São Paulo
2024

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, 19 de Junho de 2024

Assinatura do autor: Diego Cardoso

Assinatura do orientador: Maria D. Miranda.

Catalogação-na-publicação

Cardoso, Diego
A concentração de ozônio em uma microescala na cidade de São Paulo: análise exploratória de dados e modelagem / D. Cardoso -- versão corr. -- São Paulo, 2024.
118 p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Sistemas Eletrônicos.

1.processamento estatístico de sinais 2.ozônio troposférico 3.aprendizado de máquina I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Sistemas Eletrônicos II.t.

AGRADECIMENTOS

Algumas pessoas tiveram um papel fundamental para que este trabalho se concretizasse.

À minha orientadora, Profa. Maria D. Miranda, pela orientação rigorosa, extrema dedicação e esforço constante para a realização deste trabalho. Suas críticas e sugestões tiveram uma importância fundamental. Agradeço também pela paciência, confiança, amizade e preocupação com a minha carreira.

À minha irmã Camila, por toda a sua ajuda e inspiração.

Ao meu companheiro Victor, por todo o seu apoio, paciência e companheirismo.

Um agradecimento especial aos meus pais, Josiene e Marcos, pela grande ajuda ao longo dos anos e pela preocupação quanto a minha formação acadêmica e a minha realização profissional.

RESUMO

O ozônio troposférico é um poluente secundário e a sua concentração é função de complexas interações entre as condições climáticas e os poluentes primários. Estudos comprovam que, em níveis relativamente elevados e de modo persistente, o ozônio causa efeitos nocivos à saúde humana e ao ecossistema terrestre. Nos últimos anos, tem-se observado uma tendência de aumento desse poluente. Diante dessa problemática, prever a evolução temporal da concentração de ozônio pode ser extremamente útil à sociedade. No entanto, a capacidade de quantificar mudanças futuras desse poluente em escala global não é uma tarefa evidente. Recentemente, modelos estatísticos não lineares baseados no aprendizado de máquina têm sido propostos com esse fim. Nesse contexto, o objetivo deste trabalho é verificar a influência das variáveis climáticas e de poluição na tendência da concentração de ozônio ao longo dos anos de 2010 a 2023. Especificamente, considera-se a predição a curto prazo e em microescala na cidade de São Paulo. Os dados utilizados são provenientes da Companhia Ambiental do Estado de São Paulo (CETESB) e do Instituto Nacional de Meteorologia (INMET). Na predição a curto prazo, consideram-se o modelo de regressão multivariada baseada no método dos mínimos quadrados, bem como modelos baseados em árvores de decisão. A validação dos resultados para a escolha do melhor modelo é baseada em dados. Toma-se o melhor modelo e avalia-se a significância das suas variáveis considerando a sazonalidade em dois conjuntos de anos, 2010 a 2014 e 2019 a 2023. Assim, a partir da significância das variáveis, destaca-se como as concentrações de poluentes primários e as condições climáticas estão associadas ao aumento da tendência da concentração de ozônio nos conjuntos de anos considerados.

Palavras-chave: processamento estatístico de sinais, ozônio troposférico, aprendizado de máquina.

ABSTRACT

Tropospheric ozone is a secondary pollutant and its concentration is a function of interactions between climatic conditions and primary pollutants. Studies have shown that, at relatively high levels and persistently, ozone causes countervailing effects on human health and the Earth's ecosystem. In recent years, there has been an upward trend in this pollutant. Faced with this problem, predicting the temporal evolution of ozone concentration can be extremely useful for society. However, the ability to quantify future changes in this pollutant on a global scale is not an obvious task. Recently, nonlinear statistical models based on machine learning have been proposed for this purpose. In this context, the goal of this work is to verify the influence of climate change and pollution in the trend of ozone concentration over the years 2010 to 2023. Specifically, short-term and microscale forecasting in the city of São Paulo is considered. The data used comes from the Environmental Company of the State of São Paulo (CETESB) and the National Institute of Meteorology (INMET). In short-term forecasting, it is calculated the multivariate regression model based on the least squares method, as well as models based on decision trees. Validation of results to choose the best model is based on data. The best model is taken and the significance of its variables is evaluated considering seasonality in two sets of years, 2010 to 2014 and 2019 to 2023. Thus, based on the significance of the variables, it stands out how the concentrations of primary pollutants and climatic conditions are associated with the increasing trend in ozone concentration in the considered years.

Keywords: statistical signal processing, tropospheric ozone, machine learning.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BDMEP	Banco de Dados Meteorológicos
CART	*Classification And Regression Trees*
CETESB	Companhia Ambiental do Estado de São Paulo
CO	Monóxido de Carbono
INMET	Instituto Nacional de Meteorologia
LASSO	*Least Absolute Shrinkage and Selection Operator*
MAE	*Mean Absolute Error*
ML	*Machine Learning*
MP_{10}	Partículas Inaláveis
MSE	*Mean Square Error*
NO	Monóxido de Nitrogênio
NO_2	Dióxido de Nitrogênio
OLS	*Ordinary Least Squares*
OMS	Organização Mundial da Saúde
O_3	Ozônio Troposférico
PRE	Precipitação
PRESS	Pressão Atmosférica
QUALAR	Sistema de Informações da Qualidade do Ar
RADG	Radiação Solar Global
RF	*Random Forest*
RMSP	Região Metropolitana de São Paulo
TEMP	Temperatura do Ar
UR	Umidade Relativa do Ar
UTC	Tempo Universal Coordenado
VV	Velocidade do Vento

SUMÁRIO

1 INTRODUÇÃO

Os séculos XX e XXI têm sido marcados por um conjunto de mudanças tecnológicas, políticas e sociais. Essas mudanças estão transformando a sociedade e, principalmente, os padrões sociais de consumo. Nas próximas décadas, a população mundial tende a crescer (RITCHIE et al., 2023). Se os atuais padrões de consumo não forem alterados, o crescimento do consumo de combustíveis fósseis e de outros recursos naturais continuará contribuindo para a degradação do meio ambiente (HORTA et al., 2023; GABRIC, 2023; LELIEVELD et al., 2023). A atual condição de poluição do ar, principalmente em centros urbanos, é uma das consequências decorrentes dessas intensas transformações que vêm ocorrendo.

Segundo um levantamento da Organização Mundial da Saúde (OMS), estima-se que 90% da humanidade respira ar poluído (WHO, 2018). Em decorrência dessa poluição, sete milhões de pessoas morrem anualmente no mundo (WHO, 2018). Além disso, projeções indicam que a poluição do ar será a maior causa ambiental responsável pela mortalidade de crianças recém-nascidas até 2050 (OECD, 2012). Em termos de degradação ambiental, a poluição do ar é responsável por dois grandes problemas: o efeito estufa, relacionado ao aumento da temperatura média do planeta, e a destruição da camada de ozônio, aumentando a incidência de radiação solar nociva aos seres vivos.

O ozônio (O_3) troposférico é um dos mais importantes poluentes atmosféricos em termos de efeitos nocivos à saúde humana e ao ecossistema terrestre (FOWLER et al., 2008). Trata-se de um poluente secundário, produzido por meio das reações químicas que ocorrem na atmosfera entre poluentes primários. O ozônio apresenta um importante papel na química atmosférica, ele afeta o clima e o clima o afeta. As reações fotoquímicas responsáveis pela formação do ozônio troposférico estão bem documentadas na literatura (STOCKWELL et al., 1997; ATKINSON, 2000; FOWLER et al., 2008). Em particular, as concentrações desse poluente dependem muito das condições meteorológicas locais, como velocidade do vento, temperatura, cobertura de nuvens, etc. (FOWLER et al., 2008). Tais fatores climáticos não apenas possuem influência direta na formação de ozônio troposférico, como também são afetados a partir da presença desse poluente (STOCKWELL et al., 1997; FOWLER et al., 2008). Um esquema simplificado das interações entre o clima, ecossistemas e o ozônio troposférico é representado na Figura 1. Nessa figura as linhas sólidas grossas indicam alguns processos que estão relativamente bem compreendidos e que são representados em modelos climáticos/químicos para predições do ozônio. As linhas sólidas finas representam processos que estão parcialmente compreendidos e que por conta disso são representados por modelos aproximados. As linhas tracejadas representam alguns processos que estão emergindo como importantes, mas que ainda não são incluídos nas projeções dos modelos devido à sua complexidade (FOWLER et al., 2008).

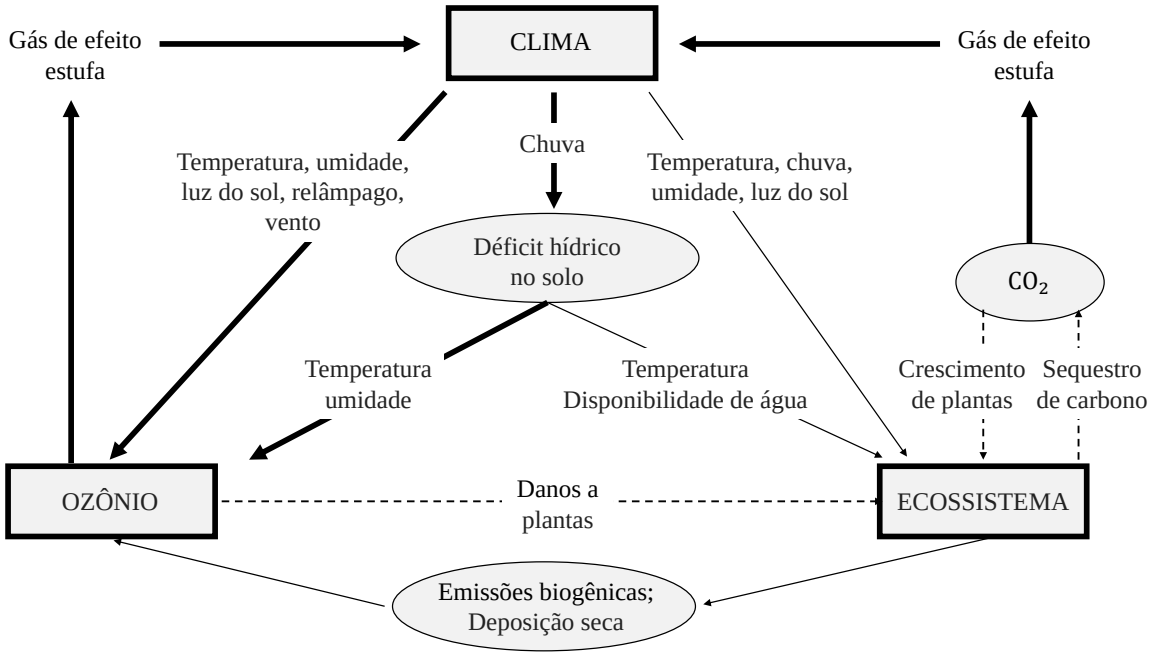

Figura 1 – Esquema das interações entre o clima, ecossistemas e o ozônio troposférico. Figura adaptada de (FOWLER et al., 2008).

Visando conter e mitigar os impactos negativos do ozônio sobre os ecossistemas terrestres e a população humana, a Organização Mundial da Saúde (OMS) publicou, em 2021, diretrizes globais de qualidade do ar com intuito específico de limitar a concentração máxima de ozônio presente no ar. A meta recomendada para a máxima concentração média em oito horas diárias de exposição a esse poluente não deve ultrapassar 100 $\mu g/m^3$ para garantir uma proteção razoável à saúde humana. De acordo com a CETESB, o aumento da concentração de ozônio pode causar o agravamento dos sintomas como tosse seca, cansaço, ardor nos olhos, nariz e garganta e ainda apresentar falta de ar ou respiração ofegante. Esses efeitos são ainda mais acentuados nos grupos mais sensíveis como crianças, idosos e pessoas com problemas cardiovasculares (CETESB, 2023). Apesar disso, a qualidade do ar em grandes centros urbanos não atende, em algumas épocas do ano, aos padrões recomendados pela OMS (CETESB, 2023; ORGANIZATION, 2021).

Diante da problemática ambiental relativa à presença em excesso de ozônio na atmosfera, estudos que permitem modelar a evolução temporal da concentração desse poluente podem ser extremamente úteis à sociedade. A partir de tais modelos, pode ser possível identificar e avaliar as principais causas do aumento de ozônio na atmosfera e, com isso, prever com antecedência soluções e estratégias de mitigação. Diferentes modelos têm sido desenvolvidos e utilizados para fazer a predição da concentração de ozônio ao longo do tempo. No entanto, a capacidade de quantificar mudanças futuras desse poluente em escala global é demasiadamente complexa. Na maioria dos estudos existentes, as interações entre variáveis meteorológicas e indicadores de poluição atmosférica, são utilizadas para a previsão da concentração de ozônio. Alguns desses estudos utilizam metodologias estatísticas usuais,

como por exemplo, os modelos de regressão linear (LIU et al., 2013; VERMA et al., 2015; ALLU et al., 2020). Apesar de amplamente utilizados devido à sua fácil implementação e interpretabilidade, esses modelos são obtidos a partir de fortes suposições que nem sempre são coerentes com a realidade (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Recentemente, uma nova classe de modelos estatísticos, conhecidos como métodos de Aprendizado de Máquina (do inglês, *Machine Learning* - ML), têm sido propostos para o estudo do ozônio. Esses modelos vêm sendo utilizados devido à sua capacidade de lidar com problemas complexos, como é o caso das interações espaço-temporais entre as concentrações de ozônio e as demais variáveis meteorológicas e indicadores de poluição. Entre os modelos de ML, destacam-se os que envolvem redes neurais artificiais (SOUSA et al., 2007; CHATTOPADHYAY; MIDYA; CHATTOPADHYAY, 2019) e aqueles baseados em árvores de decisão (REN; MI; GEORGOPOULOS, 2020).

1.1 OBJETIVOS DOS TRABALHO

Os objetivos deste trabalho são elencados a seguir.

- Analisar a concentração do O_3 em microescala[1] e a curto prazo na cidade de São Paulo. Considera-se os dados horários disponibilizados pela Companhia Ambiental do Estado de São Paulo (CETESB) e pelo Instituto Nacional de Meteorologia (INMET). A partir da análise dos dados, escolhe-se o cenário em que a tendência de aumento da concentração de O_3 é mais notável ao longo dos anos analisados.

- Investigar as soluções de modelos para a predição da concentração de ozônio a curto prazo em cenários onde se pode melhor observar a maior tendência de aumento da concentração do O_3. Aplica-se o modelo de regressão linear múltipla baseado no método dos mínimos quadrados ordinários (do inglês, *Ordinary Least Squares* (OLS)), o método de seleção e regularização de variáveis LASSO (acrônimo de *Least Absoluete Shrinkage and Selection Operator*), os modelos baseados em árvores de decisão, o CART (acrônimo de *Classification And Regression Trees*) e o modelo de Florestas Aleatórias (do inglês, *Random Forest* (RF)).

- Evidenciar as variáveis de maior significância nos modelos e verificar se a significância das variáveis se alteram em cada modelo ao longo do tempo.

1.2 ORGANIZAÇÃO DO TRABALHO

Esta dissertação está estruturada em 5 capítulos, cujos conteúdos são resumidos a seguir.

[1] Microescala: considera-se que as dispersões dos poluentes ficam relativamente constantes em torno de 100 metros da distância do ponto em que é feita a medição

- No Capítulo 2, apresenta-se um detalhamento das estações meteorológicas consideradas neste trabalho, as variáveis escolhidas, a sistemática empregada para a obtenção e organização dos dados e uma análise preliminar acerca dos dados obtidos.
- No Capítulo 3, apresenta-se a análise exploratória dos dados em que são verificados os diferentes cenários onde a tendência da concentração de O_3 é mais notável. Observam-se:
  - a evolução temporal das séries da temperatura, radiação solar global e da concentração de O_3;
  - a evolução temporal e horária das variáveis conforme a sazonalidade do ano;
  - as covariâncias das demais variáveis ao longo do tempo;
  - as correlações cruzadas entre as variáveis consideradas e a concentração de O_3 usando o Coeficiente de Correlação Linear de Pearson.
- No Capítulo 4, consideram-se modelos para estimar a concentração de O_3 a curto prazo no cenário em que a tendência de aumento da sua concentração é mais notável. A validação dos modelos é feita baseada em dados dos anos 2010 a 2023. Nessa etapa, o objetivo é evidenciar as variáveis de maior significância em cada modelo e verificar como a significância das variáveis se alteram quando se compara os resultados dos anos de 2010 a 2014 com os anos de 2019 a 2023.
- No Capítulo 5, apresenta-se a conclusão do trabalho e algumas sugestões para a sua continuidade.

2 EXTRAÇÃO E ORGANIZAÇÃO DOS DADOS

No Brasil, dados meteorológicos e de poluição do ar são geralmente coletados e disponibilizados por meio de Órgãos Públicos do Meio Ambiente. O acesso a esses dados pode ser feito através do portal de cada órgão. No entanto, a obtenção desses dados nem sempre é uma tarefa simples, especialmente quando os dados não são estruturados com o objetivo do fácil acesso do usuário. Além disso, raramente os dados estarão formatados e prontos para análise, sendo, portanto, necessário uma etapa de pré-processamento.

Para a realização deste trabalho, são utilizados dados públicos obtidos através da Companhia Ambiental do Estado de São Paulo (CETESB) e do Instituto Nacional de Meteorologia (INMET). A CETESB disponibiliza, somente para o estado de São Paulo, dados relacionados à concentração de poluentes do ar e, adicionalmente, alguns dados climáticos. Sua base de dados também está disponível na internet (QUALAR, 2023). O INMET disponibiliza dados relacionados ao clima de todo o Brasil (INMET, 2023). Suas bases de dados contam com diferentes medidas de variáveis climáticas e que podem ser acessadas pela internet (BDMEP, 2023).

O presente capítulo é estruturado da seguinte forma: Na Seção 2.1 é apresentado um detalhamento sobre as estações meteorológicas consideradas neste trabalho e as variáveis escolhidas. Na Seção 2.2 é apresentada a sistemática empregada para a obtenção dos dados. Na Seção 2.3 é demonstrada a forma de organização dos dados. Por fim, na Seção 2.4 é apresentada uma análise preliminar acerca dos dados obtidos.

2.1 ESTAÇÕES METEOROLÓGICAS E DADOS COLETADOS

Em estudos envolvendo a poluição do ar, dados de gases poluentes são coletados muitas vezes por meio de experimentos laboratoriais ou através de instrumentos de medição posicionados em espaços de grande movimento, como em grandes avenidas, túneis, parques, próximos a fábricas e outros locais de interesse. Da mesma forma, é comum a instalação de estações automáticas de monitoramento que medem periodicamente diversos parâmetros dessa natureza. No Brasil, apenas dez estados brasileiros e o Distrito Federal possuem redes de monitoramento da qualidade do ar, segundo dados do Instituto de Energia e Meio Ambiente (IEMA, 2022). Esses estados são o Ceará, Pernambuco, Bahia, Goiás, Minas Gerais, Espírito Santo, Rio de Janeiro, São Paulo, Paraná e Rio Grande do Sul, além do Distrito Federal.

2.1.1 COMPANHIA AMBIENTAL DO ESTADO DE SÃO PAULO (CETESB)

A CETESB é a agência do Governo do Estado de São Paulo encarregada pelo controle, fiscalização, monitoramento e licenciamento de atividades geradoras de poluição,

com a preocupação fundamental de preservar e recuperar a qualidade das águas, do ar e do solo (CETESB, 2023). Além disso, a CETESB possui uma rede de monitoramento da qualidade do ar no estado que conta com estações de monitoramento manuais e automáticas. Essa rede possui 61 estações automáticas fixas, duas estações automáticas móveis e 22 pontos de monitoramento manual, distribuídos ao longo do estado. Segundo dados do Relatório de Qualidade do Ar no Estado de São Paulo (CETESB, 2022), na Região Metropolitana de São Paulo (RMSP) estão localizadas 30 estações de monitoramento.

As estações automáticas são caracterizadas por sua capacidade de processar na forma de médias horárias, no próprio local e em tempo real, as amostragens realizadas com intervalos de cinco segundos. Cada média horária é representada por uma amostra de 3/4 das medidas válidas na hora. Essas médias são transmitidas para a central de telemetria da CETESB e armazenadas em um servidor de banco de dados dedicado, onde passam por um processo de validação técnica periódica e, posteriormente, são disponibilizadas de hora em hora para o público (CETESB, 2022).

Para acesso aos dados coletados por suas estações, o Sistema de Informações da Qualidade do Ar - QUALAR, possibilita a consulta dos dados por meio do endereço eletrônico da CETESB. Esse sistema, *on-line* e de livre domínio público, encontra-se disponível no endereço <https://cetesb.sp.gov.br/ar/qualar/> (QUALAR, 2023). Além de disponibilizar dados relativos a poluição do ar, a CETESB também disponibiliza alguns dados climáticos em suas estações de monitoramento. Apesar disso, bases mais consolidadas contendo dados climáticos podem ser obtidas através do Instituto Nacional de Meteorologia (INMET, 2023).

2.1.2 INSTITUTO NACIONAL DE METEOROLOGIA (INMET)

O INMET é um órgão do Ministério da Agricultura e Pecuária, criado em 1909 com a missão de prover informações meteorológicas através de monitoramento, análise e previsão do tempo e clima em superfície. Além disso, este órgão é responsável por coordenar processos de pesquisa aplicada para fornecer informações adequadas em situações diversas, como no caso de desastres naturais como inundações e secas extremas que afetam, limitam ou interferem nas atividades cotidianas dos brasileiros (INMET, 2023).

A rede de monitoramento de dados climáticos do INMET conta com 170 estações manuais e 408 estações automáticas que estão operantes e distribuídas ao longo de todo o território brasileiro. A rede de estações meteorológicas automáticas utiliza o que há de mais moderno internacionalmente. O INMET disponibiliza por meio do seu portal a consulta de diferentes medidas de variáveis climáticas disponibilizadas de forma horária, diária e mensal. Os dados podem ser acessados através do Banco de Dados Meteorológicos do INMET - BDMEP, disponível no endereço <https://bdmep.inmet.gov.br/> (BDMEP, 2023).

2.1.3 DISTRIBUIÇÃO ESPACIAL DAS ESTAÇÕES PRÓXIMAS DA CETESB E DO INMET

A partir da Figura 2 é possível observar a distribuição espacial das estações consideradas neste trabalho. Em vermelho, está localizada a estação **Parque D. Pedro II** (Latitude −23,5448°; Longitude −46,6276°) da CETESB. Na cor azul, está localizada a estação **Mirante de Santana** (Latitude −23,4962°; Longitude −46,6200°) do INMET.

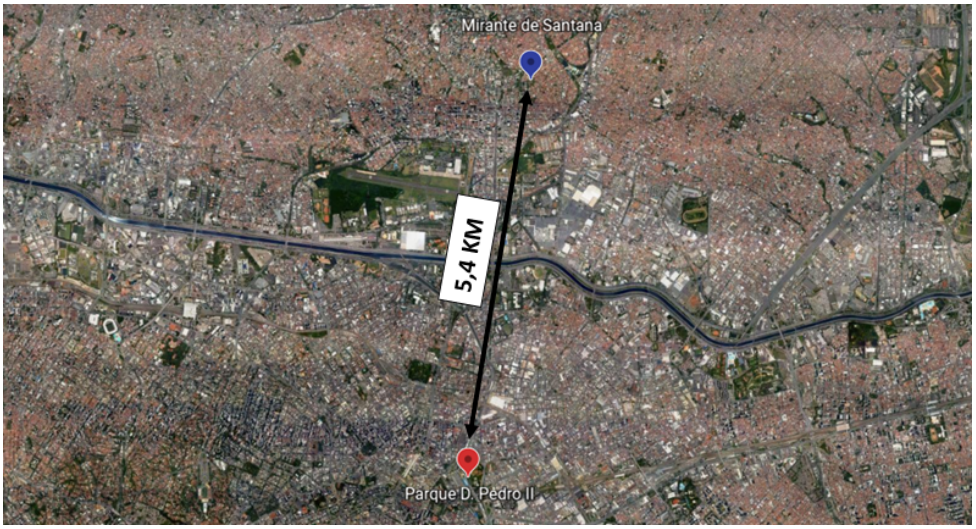

Figura 2 – Localização da estação Parque D. Pedro II da CETESB (em vermelho) e Mirante de Santana do INMET (em azul) (Via Google Earth).

A estação Parque D. Pedro II fica localizada no centro da cidade de São Paulo, em uma região de topografia plana e próxima a uma via com tráfego intenso de veículos leves e pesados. Para essa estação, são utilizados apenas os dados relativos a indicadores de poluição do ar. A representatividade espacial dos poluentes medidos nessa estação é definida como sendo de microescala, isto é, as medições realizadas nessa estação representam concentrações abrangendo áreas de dimensão de poucos metros até 100 metros (CETESB, 2022; QUALAR, 2023). Apesar da estação Parque D. Pedro II também monitorar alguns dados de variáveis climáticas, a consistência desses dados em termos de disponibilidade é relativamente baixa no intervalo de tempo analisado. Por conta disso, para o desenvolvimento deste trabalho são considerados os dados meteorológicos extraídos da estação Mirante de Santana. Essa é a principal estação meteorológica de São Paulo e está localização a uma distância de aproximadamente 5,4 km da estação Parque D. Pedro II. Observa-se que, na distância entre as duas estações, os dados climáticos não apresentam variações significativas em 5,4 km. Porém, os dados relacionados à poluição do ar, sim. Então, de fato, a região que vai ser considerada é a do Parque D. Pedro II para avaliar a evolução das concentrações de ozônio ao longo dos anos.

Cabe mencionar que cada uma das variáveis monitoradas apresenta um método ou instrumento de medição específico, de acordo com padrões estabelecidos. A Tabela 1 apresenta algumas características como a sigla, unidade de medida, o método ou instru-

mento de medição e a precisão dos dados obtidos (CETESB, 2022; INMET, 2023; FURG, 2023; MMA, 2019). A partir da observação da Tabela 1 pode-se notar que para cada variável escolhida, existe um erro de medição associado ao método ou instrumento de medição empregado. Se esses dados são usados para obtenção de um modelo, a precisão do modelo obtido dependerá da precisão dos dados utilizados. Determinar a precisão de um modelo com base na precisão dos dados a partir dos quais ele é obtido não é uma tarefa evidente, principalmente para modelos complexos e com não linearidades. Mesmo assim, deve-se levar em conta uma margem de segurança nas conclusões obtidas com os modelos desenvolvidos a partir dos dados.

Tabela 1 – Características das variáveis climáticas e dos indicadores de poluentes do ar.

Estação	Variável	Sigla	Unidade	Método / Instrumento	Precisão
Parque D. Pedro II (CETESB)	Monóxido de carbono	CO	ppm	Infravermelho não dispersivo	$\leq 1\%$
	Ozônio	O_3	$\mu g/m^3$	Ultravioleta	$\leq 2\%$
	Partículas inaláveis	MP_{10}	$\mu g/m^3$	Radiação beta	$\leq 10\%$
	Dióxido de nitrogênio	NO_2	$\mu g/m^3$	Quimiluminescência	$\leq 6\%$
	Monóxido de nitrogênio	NO	$\mu g/m^3$	Quimiluminescência	$\leq 6\%$
Mirante de Santana (INMET)	Radiação solar global	RADG	kJ/m^2	Piranômetro	$\pm 1\%$
	Temperatura	TEMP	°C	Termômetro	$\pm 0{,}2\%$
	Umidade relativa	UR	%	Higrômetro	$\pm 2\%$
	Precipitação	PRE	mm	Pluviômetro	$\pm 3\%$
	Pressão atmosférica	PRESS	hPa	Barômetro	$\pm 0{,}25\%$
	Velocidade do vento	VV	m/s	Anemômetro	$\pm 0{,}17\%$

2.2 RASPAGEM WEB

O acesso à base de dados do INMET é relativamente fácil, basta utilizar o sistema de consulta do banco de dados meteorológicos deste órgão e solicitar os dados desejados (BDMEP, 2023). Já o acesso às bases de dados da CETESB através do sistema QUALAR funciona bem para pequenas consultas, mas para a extração de um grande volume de dados algumas dificuldades são encontradas:

1. O sistema QUALAR apresenta algumas limitações a respeito da quantidade de variáveis selecionadas e ao tamanho do período da série de dados desejado. As variáveis são limitadas a três por consulta. Já as séries de dados são limitadas ao período máximo de um ano por consulta.

2. O processo para extração dos dados pode ser exaustivo ou impraticável, principalmente quando se pretende extrair uma série de dados contendo um período muito longo.

3. Como os dados são impressos em tela, a extração de uma série de dados contendo um período muito longo exige uma certa quantidade de memória RAM.

Nesse sentido, faz-se necessário o uso da técnica de *web scraping* (do português, raspagem de dados web) para facilitar o acesso aos dados. Em trabalhos anteriores onde

foram utilizados os dados da CETESB, autores desenvolveram uma biblioteca em linguagem R para extração dos dados horários do sistema QUALAR. A biblioteca desenvolvida chama-se *Rpollution* e está disponível através do GitHub[1] público. Após a instalação da biblioteca o processo se resume ao uso da função **scraper_cetesb()**, que apresenta entre outros argumentos, os seguintes parâmetros:

- *parameter* (int): ID da variável de interesse (exemplo: 63 - Ozônio);
- *station* (int): ID da estação de interesse (exemplo: 72 - Parque D. Pedro II);
- *start* (string): data início (exemplo: "01/03/2007");
- *end* (string): data fim (exemplo: "28/02/2008");
- *login* (string): login utilizado no sistema Qualar;
- *password* (string): senha do sistema Qualar.

A partir dessa função é possível extrair os dados horários de um determinado parâmetro, para uma determinada estação e ao longo de um determinado período. Vale mencionar que a identificação dos ID's que indicam as variáveis e as estações podem ser obtidos a partir da execução dos objetos: **cetesb_param_ids** e **cetesb_station_ids**.

Utilizando-se a função **scraper_cetesb()**, o sistema para extração dos dados e sua organização em forma de tabelas é realizado. A estrutura de dados utilizada foi o *Data Frame* da linguagem R, que são objetos de dados genéricos em R utilizados para armazenar os dados de forma tabular. Os *Data Frames* podem também ser interpretados como matrizes onde cada coluna de uma mesma matriz pode conter dados de diferentes tipos, como por exemplo, dados numéricos e alfanuméricos. O *Data Frame* é então inicializado com colunas para a data, hora, dia, mês e ano dos dados. Em seguida, as demais colunas são preenchidas com as variáveis de interesse de forma iterativa, a partir da utilização da função **scraper_cetesb()**.

Após a conclusão do processo de extração dos dados da estação Parque D. Pedro II da CETESB, as informações coletadas foram concatenadas com os dados obtidos da estação Mirante de Santana do INMET. Na Tabela 2 está ilustrado um trecho do *Data Frame* final com os nomes dos dados de interesse que serão utilizados nas análises subsequentes. A dimensão do *Data Frame* é 148.920 linhas (24 horas × 365 dias × 17 anos) por 16 colunas. As colunas de 6 a 10 são relativas aos dados extraídos da CETESB e as colunas de 11 a 16 aos dados do INMET.

[1] <https://github.com/openvironment/Rpollution>

Tabela 2 – *Data Frame* final contendo os dados de interesse extraídos da CETESB e do INMET.

Parâmetros de referência					**Variáveis da CETESB**					**Variáveis do INMET**					
DATA	**HORA**	**DIA**	**MÊS**	**ANO**	**O3**	**MP10**	**NO**	**NO2**	**CO**	**PRESS**	**RADG**	**TEMP**	**UR**	**VV**	**PRE**
01/03/2007	01:00:00	1	01	2007	20	23	Na	Na	0	925,6	Na	19,6	93	3,4	0
01/03/2007	02:00:00	1	01	2007	16	39	1	19	Na	924,9	Na	19,2	94	3,4	0
01/03/2007	03:00:00	1	01	2007	11	28	2	21	Na	924,5	Na	19	96	2,8	0,2
01/03/2007	04:00:00	1	01	2007	14	18	1	17	Na	924,2	Na	18,5	96	3,3	0
01/03/2007	05:00:00	1	01	2007	19	5	1	13	Na	924,7	Na	18,1	95	3,4	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
28/02/2024	19:00:00	28	02	2023	23	33	9	68	0,7	924,3	9,9	23,9	90	0,7	0
28/02/2024	20:00:00	28	02	2023	25	9	6	49	0,5	924,9	1,6	23,8	90	1,8	0
28/02/2024	21:00:00	28	02	2023	29	14	3	33	0,3	925,6	1,2	23,5	89	1,5	0
28/02/2024	22:00:00	28	02	2023	31	5	3	32	0,3	926	Na	23,5	86	0,5	0
28/02/2024	23:00:00	28	02	2023	23	64	9	46	0,4	926,2	Na	23,3	87	0,1	0
28/02/2024	00:00:00	28	02	2023	21	11	11	48	0,4	925,9	Na	23,3	87	0	0

148.920 linhas x 16 colunas

2.3 ORGANIZAÇÃO DOS DADOS

A partir do sistema de extração utilizado, os dados brutos podem então ser estruturados de acordo com as análises a serem empregadas sobre eles. Para avaliar as propriedades estatísticas dos dados, como por exemplo a variação da temperatura ou a concentração de algum poluente ao longo das horas do dia, tal organização é ainda mais importante. Deste modo, cada conjunto de variáveis foi organizado de forma a compor uma matriz. Um esquema da estrutura de organização das matrizes é ilustrado na Figura 3. A dimensão de cada matriz é (N_d, N_h) em que N_d representa o número de dias e N_h representa o número de horas em um dia. A matriz indicada como ano de 2007 contém os dados horários entre 1 de março de 2007 até 28 de fevereiro de 2008 e assim sucessivamente até a matriz do ano de 2023, que contém os dados horários de 1 de março de 2023 até 28 de fevereiro de 2024. No caso dos anos bissextos, o dia 29 de fevereiro foi desconsiderado de modo que N_d é igual para todos os anos. Portanto, cada matriz possui dimensão 365×24. O número total de matrizes N_a é 17, dado que são considerados os dados entre os anos de 2007 (01/03/2007 - 28/02/2008) a 2023 (01/03/2023 - 28/02/2024).

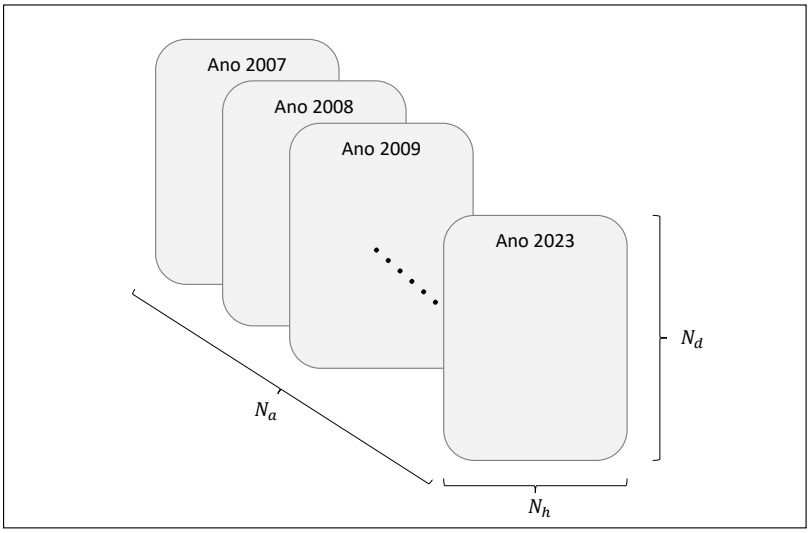

Figura 3 – Estrutura de organização dos dados, sendo que as indicações do ano, do dia e da hora, são denotadas por N_a, N_d e N_h.

As variações sazonais presentes nos dados meteorológicos e de poluição do ar são características importantes e que devem ser consideradas no processo de organização e análise dos dados. Por esse motivo, os dados foram organizados de forma a agrupar quatro períodos de três meses para dividir as estações do ano. Considera-se as estações do ano agrupas em meses inteiros conforme a classificação meteorológica: outono (março, abril e maio); inverno (junho, julho e agosto); primavera (setembro, outubro e novembro) e verão (dezembro, janeiro e fevereiro). Esse esquema de organização dos dados leva em consideração o calendário das estações meteorológicas que é baseado no ciclo anual da temperatura (NOAA, 2024). O esquema da estrutura de organização das estações do ano está representado na Figura 4.

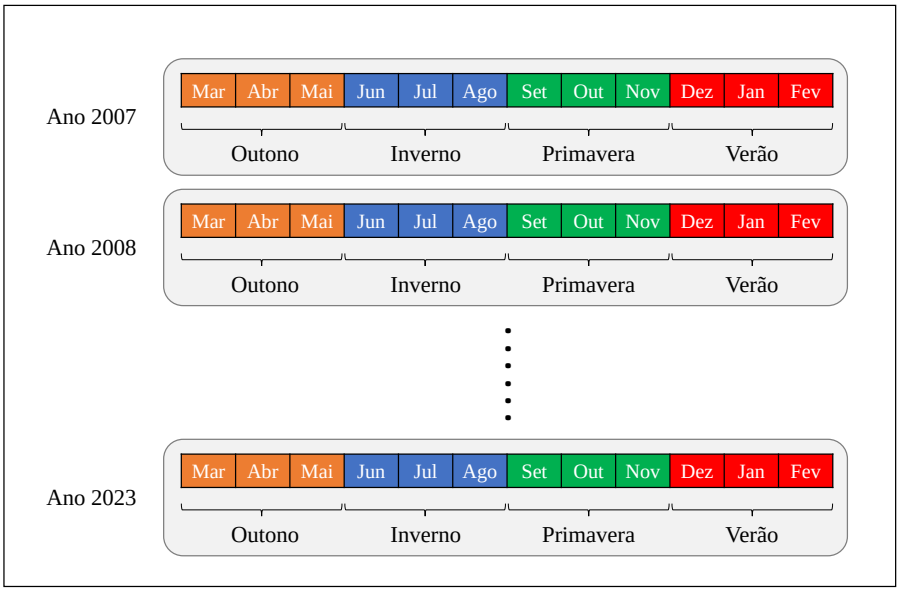

Figura 4 – Estrutura de organização das estações do ano.

Vale mencionar que os dados coletados pelo INMET estão no padrão internacional de Tempo Universal Coordenado (UTC). A fim de padronização com os dados da CETESB, utilizou-se o padrão UTC-3 que é a diferença de fuso horário que subtrai três horas do UTC.

2.4 ANÁLISE DE DADOS FALTANTES

Por muitos motivos, dados de conjuntos de variáveis meteorológicas e de indicadores de poluentes do ar nem sempre são completos ou válidos. Isso pode ocorrer devido a problemas relacionados ao mau funcionamento do equipamento, ou até mesmo por falta de energia ou necessidade de manutenção periódica do equipamento. Dados faltantes (do inglês, *missing data*) são uma característica marcante nos dados selecionados e aparecem de várias formas no conjunto analisado: em curtos períodos (como em algumas horas), em

longos períodos (como em alguns dias) ou intermitentes (podendo parar e voltar diversas vezes).

Para analisar a presença e a intensidade de dados faltantes, fez-se o cálculo da proporção de dados falantes ao longo dos anos para cada uma das variáveis selecionadas. Na Tabela 3 são apresentados os resultados deste cálculo considerando-se como referência a série de dados horária completa contendo 8.760 pontos por ano (24 horas × 365 dias). Em relação as variáveis meteorológicas, com exceção da radiação solar global (RADG), todos as demais variáveis apresentam baixa proporção de dados faltantes. Já em relação aos dados de indicadores de poluentes do ar, observa-se uma maior proporção de dados faltantes. Destaca-se a grande proporção de dados faltantes para o ano de 2008, com valores acima de 29% para todos os indicadores de poluentes monitorados pela CETESB. É possível notar ainda que não existem informações para o monóxido de carbono (CO) nos primeiros anos analisados.

Tabela 3 – Proporção de dados faltantes para cada uma das variáveis medidas na estação Parque D. Pedro II (CETESB) e Mirante de Santana (INMET). Valores considerando as medidas de 24 horas para os anos 2007 a 2023.

Ano	CETESB - Parque D. Pedro II					INMET - Mirante de Santana					
	O_3	MP_{10}	NO	NO_2	CO	PRESS	RADG	TEMP	UR	VV	PRE
2007	5%	1%	6%	6%	100%	0%	47%	0%	0%	0%	0%
2008	45%	33%	29%	29%	100%	1%	47%	2%	2%	1%	1%
2009	16%	17%	14%	14%	91%	0%	46%	1%	0%	0%	0%
2010	9%	12%	7%	7%	10%	0%	46%	0%	0%	0%	0%
2011	11%	6%	13%	13%	9%	0%	47%	0%	0%	0%	0%
2012	11%	6%	11%	11%	13%	0%	46%	0%	0%	0%	0%
2013	21%	13%	18%	18%	21%	1%	46%	1%	1%	1%	1%
2014	14%	9%	11%	11%	28%	2%	48%	2%	2%	2%	2%
2015	8%	5%	10%	10%	8%	1%	47%	1%	1%	1%	1%
2016	12%	5%	8%	8%	11%	0%	48%	0%	0%	0%	0%
2017	16%	11%	14%	14%	15%	1%	47%	1%	1%	1%	1%
2018	10%	5%	12%	12%	7%	0%	48%	0%	0%	0%	0%
2019	8%	17%	4%	4%	5%	0%	47%	0%	0%	0%	0%
2020	14%	11%	28%	28%	11%	0%	47%	0%	0%	0%	0%
2021	10%	10%	6%	6%	6%	0%	48%	0%	0%	0%	0%
2022	11%	17%	6%	6%	10%	0%	48%	0%	0%	0%	0%
2023	13%	3%	13%	13%	8%	3%	47%	3%	3%	3%	7%

A proporção de dados faltantes varia de acordo com as horas do dia por diferentes razões. No caso da estação Parque D. Pedro II, não são observados dados às 6h da manhã pois esse é o horário em que o equipamento sofre manutenção. Já no caso da estação Mirante de Santana, a alta proporção de dados faltantes para a radiação solar global (RADG) é explicada porque essa variável só é medida nas horas de ocorrência da incidência de luz solar (período das 6 às 19 horas). Dessa forma, é interessante avaliar a proporção de dados faltantes em horários específicos do dia. A partir da Tabela 4 é possível observar a proporção de dados faltantes para o período da tarde, especificamente das 12 às 17 horas. Destaca-se que há uma menor proporção de dados faltantes para a maioria das variáveis

selecionadas neste período do dia.

Tabela 4 – Proporção de dados faltantes para cada uma das variáveis medidas na estação Parque D. Pedro II (CETESB) e Mirante de Santana (INMET). Valores considerando as medidas horárias das 12 às 17 horas para os anos 2007 a 2023.

Ano	CETESB - Parque D. Pedro II					INMET - Mirante de Santana					
	O_3	MP_{10}	NO	NO_2	CO	PRESS	RADG	TEMP	UR	VV	PRE
2007	1%	1%	3%	3%	100%	0%	0%	0%	0%	0%	0%
2008	43%	37%	27%	27%	100%	1%	2%	2%	2%	1%	1%
2009	12%	17%	11%	11%	91%	0%	0%	1%	0%	0%	0%
2010	5%	12%	3%	3%	7%	0%	0%	0%	0%	0%	0%
2011	8%	7%	10%	10%	6%	0%	1%	0%	0%	0%	0%
2012	6%	5%	5%	5%	8%	0%	1%	0%	0%	0%	0%
2013	17%	11%	14%	14%	18%	0%	1%	0%	0%	0%	0%
2014	9%	9%	7%	7%	25%	2%	2%	2%	2%	2%	2%
2015	4%	5%	6%	6%	4%	1%	1%	1%	1%	1%	1%
2016	8%	4%	6%	6%	10%	0%	6%	0%	0%	0%	0%
2017	12%	11%	13%	13%	14%	1%	5%	1%	1%	1%	1%
2018	6%	6%	13%	13%	7%	0%	4%	0%	0%	0%	0%
2019	2%	17%	2%	2%	3%	0%	4%	0%	0%	0%	0%
2020	6%	10%	25%	25%	7%	0%	5%	0%	0%	0%	0%
2021	2%	10%	2%	2%	2%	0%	4%	0%	0%	0%	0%
2022	3%	18%	2%	2%	6%	0%	5%	0%	0%	0%	0%
2023	4%	2%	9%	9%	3%	1%	2%	1%	1%	1%	6%

A presença de dados faltantes é um problema comum em séries de dados meteorológicos e de indicadores de poluentes do ar e influencia os tipos de análises possíveis a partir de uma série. Dessa forma, é imprescindível avaliar a quantidade ou proporção de dados faltantes em relação aos dados efetivamente existentes para assegurar que as análises feitas e suas respectivas conclusões sejam consistentes com o problema analisado.

2.5 COMPARATIVO ENTRE OS DADOS METEOROLÓGICOS DA CETESB E DO INMET

Conforme apresentando anteriormente, a estação Parque D. Pedro II também disponibiliza alguns dados de medidas de variáveis meteorológicas. No entanto, esses dados não possuem um histórico muito longo de medições. Na Tabela 5 apresenta-se um comparativo acerca da proporção de dados faltantes das variáveis climáticas para as estações Parque D. Pedro II e Mirante de Santana. Nota-se que entre os anos de 2007 e 2015 não são observados nenhum dado para as variáveis da estação Parque D. Pedro II. Em meados do ano de 2016 esses dados começam a ser disponibilizados e apenas a partir de 2018 são observados dados com uma baixa proporção de dados faltantes ($< 6\%$).

Tabela 5 – Proporção de dados faltantes para cada uma das variáveis meteorológicas monitorados na estação Parque D. Pedro II (CETESB) e Mirante de Santana (INMET). Valores considerando as medidas horárias das 12 às 17 horas para os anos 2007 a 2023.

Ano	CETESB - Parque D. Pedro II					INMET - Mirante de Santana				
	PRESS	RADG	TEMP	UR	VV	PRESS	RADG	TEMP	UR	VV
2007	100%	100%	100%	100%	100%	0%	0%	0%	0%	0%
2008	100%	100%	100%	100%	100%	1%	2%	2%	2%	1%
2009	100%	100%	100%	100%	100%	0%	0%	1%	0%	0%
2010	100%	100%	100%	100%	100%	0%	0%	0%	0%	0%
2011	100%	100%	100%	100%	100%	0%	1%	0%	0%	0%
2012	100%	100%	100%	100%	100%	0%	1%	0%	0%	0%
2013	100%	100%	100%	100%	100%	0%	1%	0%	0%	0%
2014	100%	100%	100%	100%	100%	2%	2%	2%	2%	2%
2015	100%	100%	100%	100%	100%	1%	1%	1%	1%	1%
2016	51%	51%	51%	51%	85%	0%	6%	0%	0%	0%
2017	12%	12%	12%	12%	11%	1%	5%	1%	1%	1%
2018	6%	5%	5%	5%	5%	0%	4%	0%	0%	0%
2019	2%	2%	2%	2%	2%	0%	4%	0%	0%	0%
2020	2%	1%	2%	1%	1%	0%	5%	0%	0%	0%
2021	5%	5%	5%	5%	5%	0%	4%	0%	0%	0%
2022	2%	2%	2%	2%	2%	0%	5%	0%	0%	0%
2023	2%	2%	2%	1%	1%	1%	2%	1%	1%	1%

Um dos principais problemas relacionados a valores faltantes é a dificuldade da análise e consequentemente na precisão de modelos que forem construídos a partir dos mesmos. Diferentes estratégias podem ser utilizadas para tratar dos dados faltantes. No caso de análise de dados meteorológicos em que o ponto de coleta possui intervalos muito longos de dados faltantes, os dados da estação mais próxima cuja distância seja inferior a 8 km podem ser utilizados. Já no caso em que as estações mais próximas estiverem a uma distância superior a 8 km, os valores médios das estações circundantes podem ser utilizados para preencher os dados faltantes (STEKHOVEN; BüHLMANN, 2011). No presente trabalho, optou-se por utilizar os dados meteorológicos da estação Mirante de Santana, pois a distância física entre as duas estações é de apenas 5,4 km.

Para demonstrar que os dados meteorológicos não apresentam variações significativas dentro de 5,4 km, os dados da temperatura podem ser comparados. A Figura 5 apresenta um comparativo entre os dados da temperatura na estação Parque D. Pedro II e Mirante de Santana para o período entre 2018 e 2023. A partir da Figura 5 (a) é possível observar a dispersão das médias diárias da temperatura para as duas estações. A maioria dos pontos não estão sobre a reta (vermelha) de 45 graus. Entretanto, a dispersão da maioria dos valores em torno da reta é relativamente pequena. Nota-se que o coeficiente de Correlação Linear de Pearson calculado é de ≈ 1. A Figura 5 (b) apresenta um comparativo acerca da temperatura média, considerando o total de anos e dias, ao longo das 24 horas para cada uma das estações. É possível observar que os valores são muito próximos. As demais variáveis climáticas também foram avaliadas considerando os trechos de dados contendo observações para ambas as estação e não foram constatadas diferenças significativas.

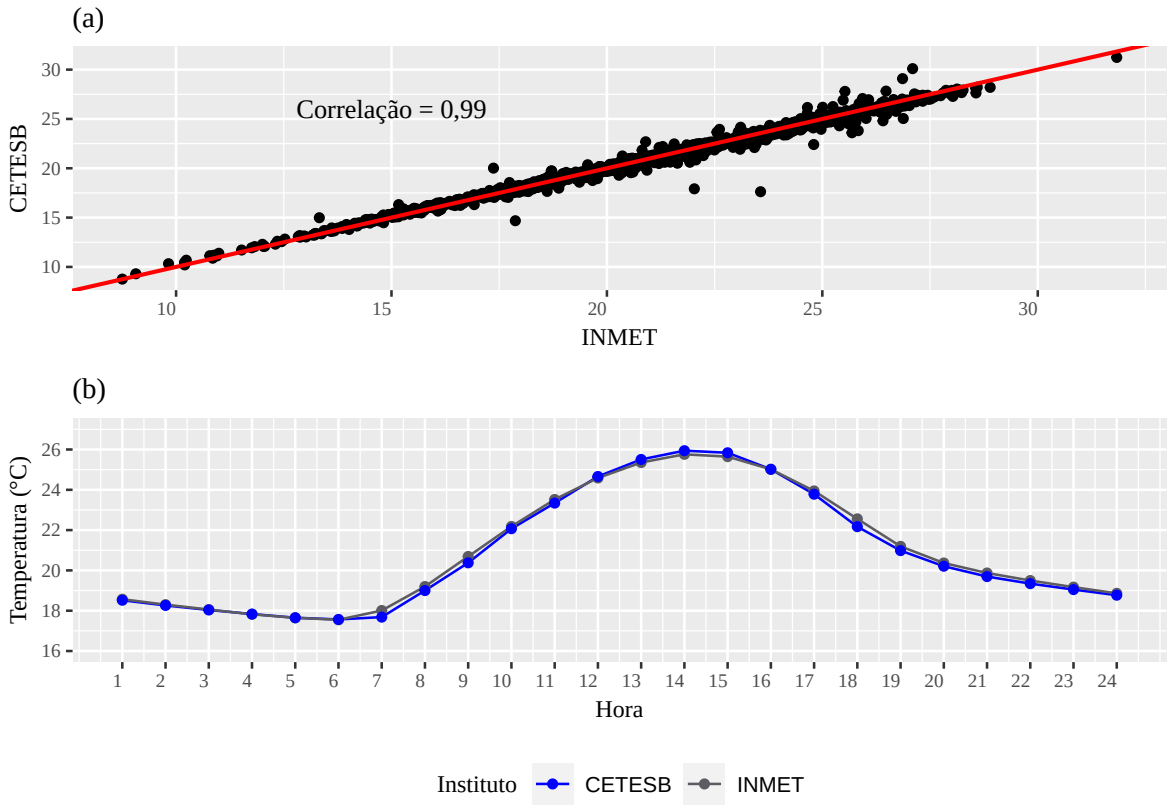

Figura 5 – Comparativo entre os dados da temperatura na estação Parque D. Pedro II e Mirante de Santana para o período de 2018 a 2023. No painel (a) apresenta-se o gráfico de dispersão da média diária da temperatura. No painel (b) apresenta-se a média horária da temperatura.

Vale mencionar que os dados da temperatura coletados pela CETESB e pelo INMET apresentam diferentes métodos de referência. Os dados da CETESB utilizam como referência uma média horária de 3/4 das medidas coletadas a cada cinco segundos. Já os dados coletados pelo INMET utilizam como referência o valor máximo coletado na hora.

2.6 CONCLUSÃO

A coleta e a organização dos dados são duas etapas cruciais na realização de qualquer trabalho envolvendo análise de dados. Os dados utilizados nesse trabalho são provenientes da CETESB e do INMET, dois órgãos públicos com relevância nacional e internacional nos temas que atuam. Em geral, os dados coletados por esses dois institutos podem ser obtidos através do portal de cada órgão. Para diminuir o tempo e o esforço gasto na etapa de coleta dos dados, avaliou-se necessário utilizar a técnica de *web scraping* para obtenção dos dados da CETESB. Após o processo de extração dos dados da CETESB e do INMET, as informações obtidas foram então armazenadas na estrutura de *data frames* da linguagem R e posteriormente organizadas em matrizes, facilitando a análise das mesmas.

A estação Parque D. Pedro II da CETESB foi escolhida para a análise da evolução dos níveis de O_3 ao longo dos anos. Essa estação fica localizada no centro da cidade de São

Paulo, em uma região de topografia plana e próxima a uma via com tráfego intenso de veículos leves e pesados. Além do ozônio, são considerados outros poluentes monitorados nessa estação, como o monóxido de carbono (CO), partículas inaláveis (MP_{10}), monóxido de nitrogênio (NO) e dióxido de nitrogênio (NO_2). A escala de representatividade espacial desses poluentes é definida como de microescala, pois representa áreas com dimensão de até 100 metros.

Através da análise de dados faltantes, observou-se que esse é um problema marcante nos dados extraídos da estação Parque D. Pedro II da CETESB. Esse problema aparece de diferentes formas ao longo da série analisada, como em períodos curtos, longos e até mesmo intermitentes. Por não dispor de uma série muito longa de dados meteorológicos, são considerados os dados da estação Mirante de Santana do INMET, localizada a uma distância de aproximadamente 5,4 km. Observa-se que, na distância entre as duas estações, os dados climáticos não apresentam variações significativas. Além disso, observou-se que em algumas horas do dia a ocorrência de dados faltantes é maior, seja porque o equipamento sofre manutenção periódica ou pelo fato de não coletar dados de determinados horários. A despeito das causas que possam justificar esse problema, observou-se que no período da tarde a ocorrência de dados faltantes é menor, em especial no período das 12 às 17 horas. A avaliação da quantidade de dados faltantes é uma tarefa importante para poder realizar o planejamento da análise e assegurar a consistência das conclusões obtidas.

Para o desenvolvimento das análises subsequentes, consideram-se períodos de tempo específicos de acordo com a disponibilidade dos dados medidos e as análises a serem realizadas sobre eles. No caso da análise da temperatura, são utilizados os dados de 2007 a 2023 pois, conforme consta na Tabela 3, não há quantidade significativa de dados faltantes do INMET. Já para a análise do ozônio são considerados os dados de 2007 a 2023 (com exceção de 2008), devido à grande quantidade de dados faltantes conforme apresentado na Tabela 3. Posteriormente, quando todas as variáveis forem usadas nas análises são considerados os anos de 2010 a 2023 das 12 às 17 horas. Este é um período consecutivo de anos onde são observados dados para todas as variáveis selecionadas. Além disso, há uma menor proporção de dados faltantes nesses horários.

3 ANÁLISE EXPLORATÓRIA DOS DADOS

A análise exploratória preliminar de dados, feita no Capítulo 2, teve como objetivo observar intervalos de tempo em que há quantidade significativa de dados faltantes. Particularmente, observou-se que nos anos de 2010 a 2023 e no período das 12 às 17 horas, a quantidade relativa de dados faltantes da CETESB não é significativa e, portanto, espera-se extrair informações relevantes sobre o fenômeno em estudo. A análise exploratória de dados que é apresentada no Capítulo 3 tem o objetivo de avaliar as variáveis climáticas e indicadores de poluentes no ar. A fim de verificar as tendências dos comportamentos dessas variáveis, são calculadas e interpretadas as suas estatísticas ao longo dos dias, ao longo das estações do ano e ao longo das horas do dia. Além disso, é feita uma análise multivariada com base no Coeficiente de Correlação Linear de Pearson.

O presente capítulo é estruturado da seguinte forma: Na Seção 3.1 é apresentada a análise da série da temperatura do ar. Na Seção 3.2 é apresentada a análise da série do O_3. Na Seção 3.3 é apresentada a análise da série da radiação solar global. Por fim, na Seção 3.4 é apresentada a análise multivariada do conjunto de variáveis utilizadas nesse trabalho.

3.1 ANÁLISE DOS DADOS DA TEMPERATURA

Devido à aparente relação entre a concentração de O_3 e a temperatura, são consideradas as medidas horárias da temperatura do ar nos anos de 2007 a 2023. Inclui-se no caso da temperatura também os anos de 2007 a 2009 pois, conforme consta na Tabela 5, não há quantidade significativa de dados faltantes do INMET. A fim de verificar a tendência do seu comportamento são avaliadas as suas estatísticas ao longo dos dias, ao longo das estações do ano e ao longo das horas do dia.

3.1.1 OBSERVAÇÃO TEMPORAL DA TEMPERATURA

Uma primeira análise acerca dos dados da temperatura pode ser feita pela simples inspeção visual de trecho da série em intervalos específicos de tempo. Na Figura 6 apresenta-se a série horária da temperatura no mês de março de 2007 (Figura 6 (a)), ao longo do ano de 2007 (Figura 6 (b)) e ao longo do ano de 2023 (Figura 6 (c)).

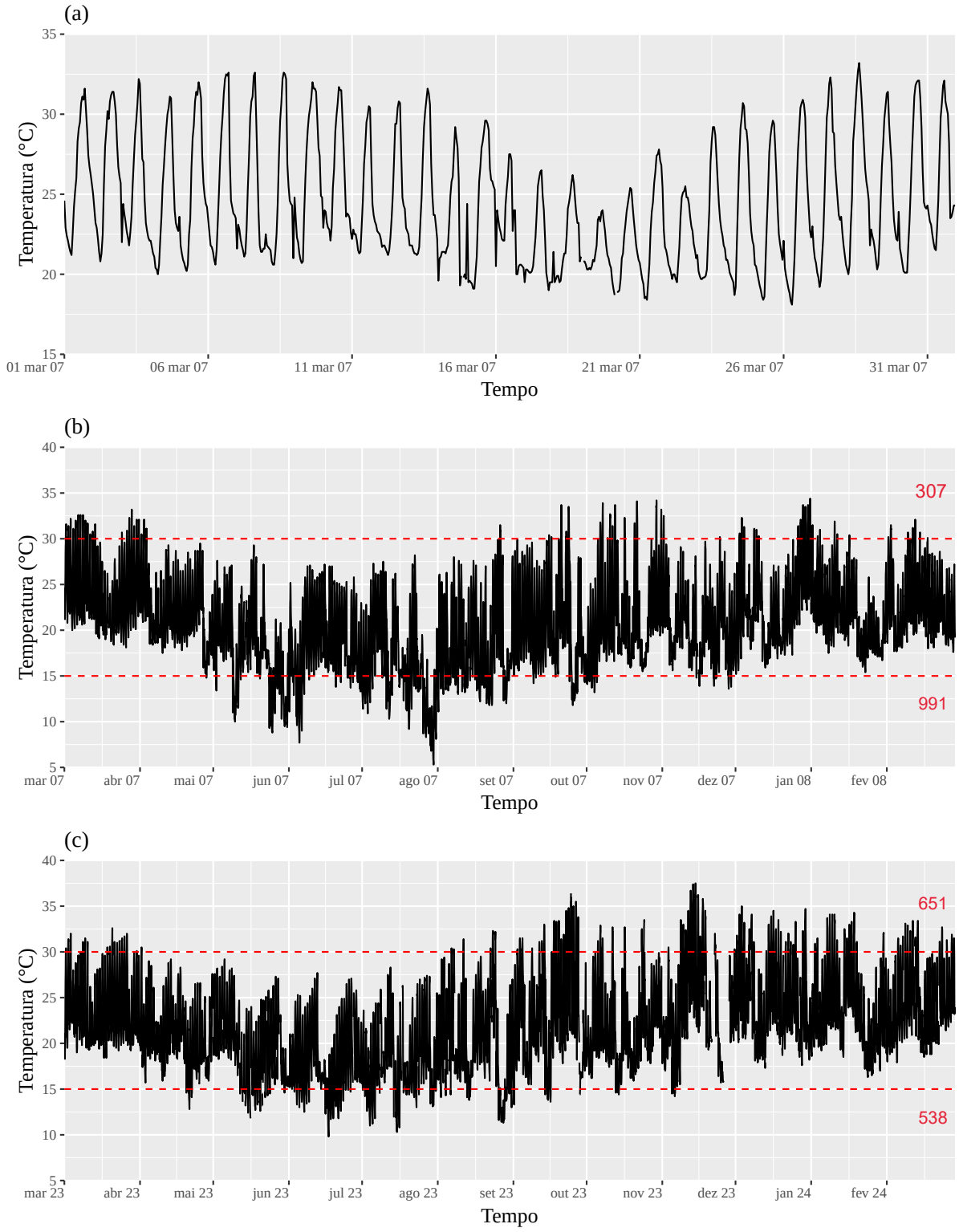

Figura 6 – Série horária da temperatura do ar para os períodos: (a) março/2007, (b) ano de 2007 e (c) ano de 2023. Uma linha horizontal vermelha foi utilizada como limiar para indicar o número de ocorrências em que a temperatura excedeu 30°C ou foi inferior a 15°C. Nos anos de 2007 e 2023, o limiar de 30°C foi ultrapassado 307 e 528 vezes, respectivamente, enquanto o limiar de 15°C foi ultrapassado 991 e 538 vezes, respectivamente.

Em relação a cada um dos gráficos da Figura 6 nota-se que:

◇ No gráfico da Figura 6 (a) está o trecho da série ao longo do mês de março de 2007. Nota-se um comportamento cíclico da série, com valores de máximos e mínimos iguais ao número de dias do mês. Além disso, é possível observar algumas variações relevantes entre os valores de máximos, mínimos e a amplitude térmica, que é a diferença entre a máxima e a mínima temperatura de um mesmo dia, ao longo dos dias.

◇ Nos gráficos das Figuras 6 (b) e (c) estão os trechos das séries ao longo dos anos de 2007 e de 2023, respectivamente. Analisando a série ao longo do período de um ano, nota-se que a variação da temperatura está associada à sazonalidade (primavera, verão, outono e inverno). Destaca-se que as maiores temperaturas são registradas em cada ano entre setembro a janeiro, enquanto que as menores temperaturas são registradas entre abril a agosto. Além disso, comparando as curvas das Figuras 6 (b) e (c) é possível notar que a temperatura apresenta variações sazonais entre os anos de 2007 e 2023. Destaca-se que o número de vezes em que a temperatura ultrapassou o valor de 30°C praticamente dobrou quando se compara os anos de 2007 com 2023, passando de 307 para 651 observações. Nesses mesmos anos, o número de vezes em que o valor da temperatura ficou inferior a 15°C diminui, passando de 991 observações em 2007 para 538 observações em 2023.

3.1.2 OBSERVAÇÃO DA MÉDIA MÓVEL

Devido às características intrínsecas da temperatura, nota-se que além das suas variações estatísticas devido à sazonalidade, há também variações ao longo das horas e ao longo dos anos. A seguir, a temperatura é analisada ao longo do tempo. Os processamentos efetuados sobre os dados horários da temperatura são brevemente descritos a seguir:

- **Temperatura máxima diária**:

  Maior valor diário obtido a partir das medidas horárias;

- **Temperatura média diária**:

  $$\frac{\text{Soma dos valores registrados ao longo do dia}}{\text{Número de registros efetuados}};$$

- **Temperatura mínima diária**:

  Menor valor diário obtido a partir das medidas horárias.

Após essas transformações, a série de dados passou de uma sequência de 148.920 pontos (24 horas × 365 dias × 17 anos) para três séries de 6.205 pontos (365 dias ×

17 anos), representadas pelas temperaturas máximas, médias e mínimas diária. A partir dos dados diários, foi calculada a média anual. Para isso, aplicou-se o cálculo da média aritmética para cada um dos anos das sequências de dados das temperaturas máximas, médias e mínimas. A partir dos dados anuais de 2007 a 2023, calculou-se uma média móvel com janelas de 5 anos para avaliar a evolução das diferentes temperaturas ao longo do tempo. Para cada uma das curvas das temperaturas máximas, médias e mínimas, o primeiro ponto representa a média dos valores anuais de 2007 a 2011, o segundo ponto representa a média dos anos de 2008 a 2012, e assim sucessivamente, até a média dos anos de 2019 a 2023. A média móvel de cinco anos foi utilizada de forma a atenuar as variações meteorológicas que são observadas de ano para ano. Os resultados estão ilustrados na Figura 7.

A evolução da média das médias móveis de cinco anos para as temperaturas máximas, médias e mínimas diárias estão ilustradas nas curvas da Figura 7 (a). A partir da observação da média móvel nota-se que a temperatura tem variado ao longo dos anos. Pode-se perceber uma leve tendência de aumento quando se compara os cinco primeiros anos (2007 a 2011) em relação aos últimos (2019 a 2023). A maior variação é observada para as máximas diárias, com um aumento de cerca de 0,90°C no período analisado. Enquanto que as temperaturas médias e mínimas diárias apresentam um aumento de 0,57°C e 0,32°C, respectivamente. A evolução da média móvel da amplitude térmica ao longo dos anos está ilustrada através do gráfico de barras da Figura 7 (b). Nota-se um aumento da amplitude térmica com o passar dos anos. O que decorre principalmente do maior aumento das temperaturas máximas diárias.

A partir das curvas da Figura 7, constata-se que há uma tendência de aumento não somente das diferentes médias de temperatura, mas também há uma tendência de aumento da amplitude térmica. O aumento mais marcante é observado na média da temperatura máxima que é de aproximadamente 0,9 graus Celsius. Sabe-se que o aumento da temperatura e também da amplitude térmica são preocupantes pois alteram os padrões climáticos e perturbam o equilíbrio da natureza (HANSEN et al., 2023). Devido a variabilidade dos valores da temperatura ao longo do tempo, pode-se considerar que a série temporal da temperatura não possui a propriedade da estacionariedade. Fato que torna sua análise não muito evidente. Assim, devido às características de não estacionariedade é conveniente tentar quantificar a tendência sob diferentes formas de avaliar os dados. A seguir como a variação da temperatura está associada também à sazonalidade, observa-se como a tendência se comporta ao longo das estações do ano e como cada estação se comporta ao longo dos anos.

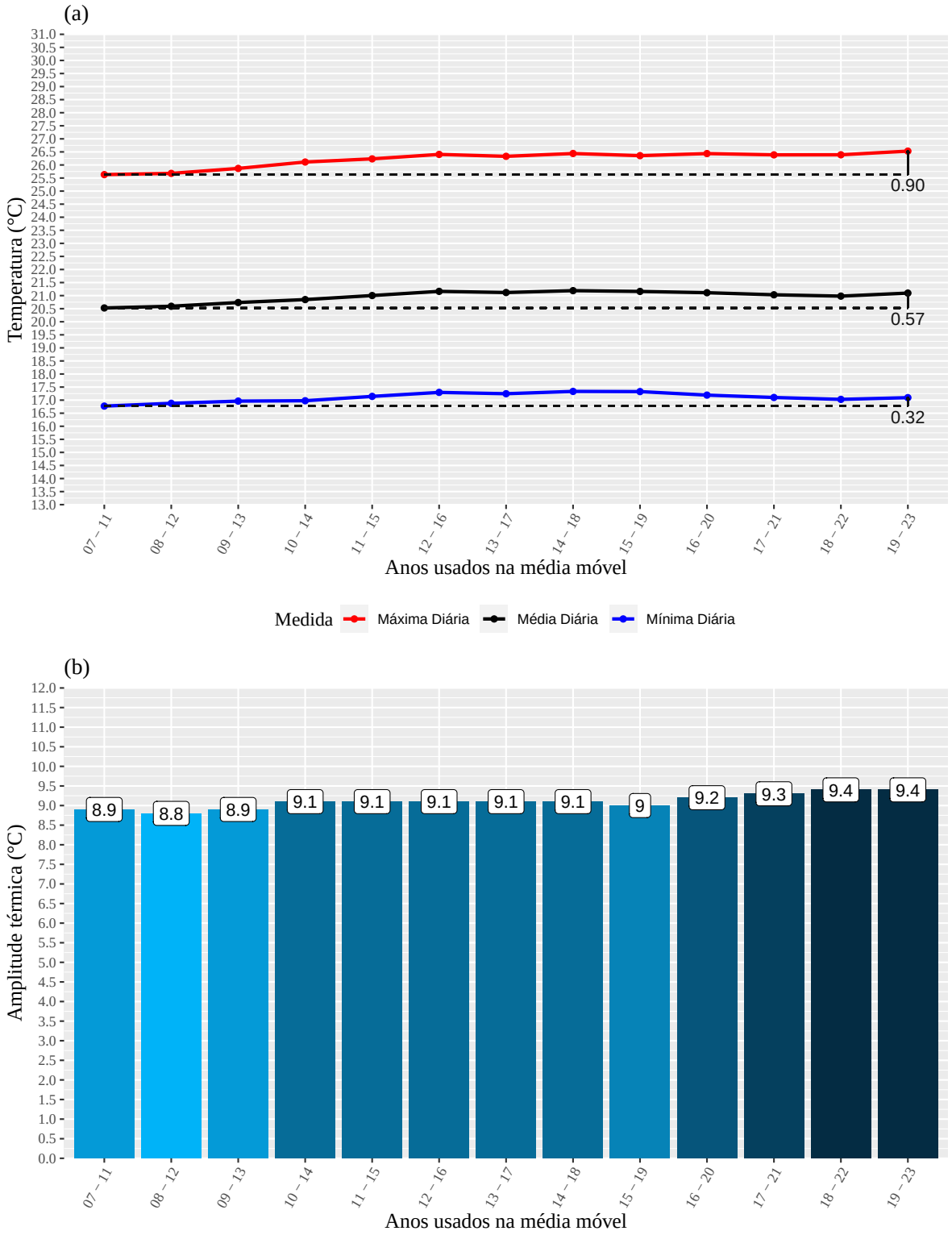

Figura 7 – Evolução da média móvel de 5 anos da temperatura considerando: (a) as temperaturas máximas, médias e mínimas diárias; (b) as amplitudes térmicas.

3.1.3 OBSERVAÇÃO SAZONAL

Visando quantificar como a temperatura está associada à sazonalidade inicialmente considerou-se todos os dados horários registrados para cada estação do ano para o período de 2007 a 2023. Na Figura 8 estão ilustrados os histogramas da temperatura para cada estação do ano. Como já esperado, os histogramas revelam que as maiores temperaturas estão concentradas em média no verão e as menores no inverno. Nota-se que as caudas dos histogramas são relativamente diferentes, mas uma interpretação direta a partir dos histogramas não é evidente.

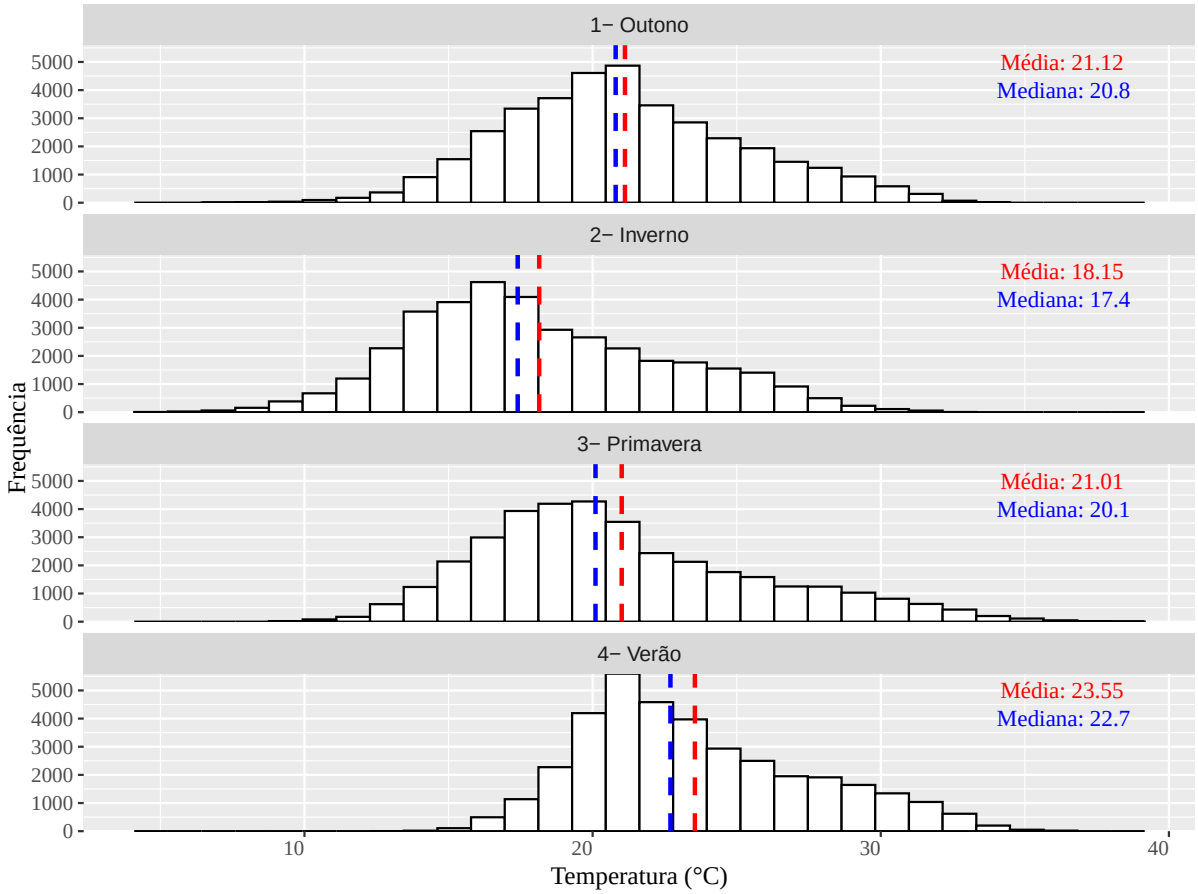

Figura 8 – Histogramas da temperatura por estação do ano. Valores considerando todos os dados horários registrados para os anos de 2007 a 2023. As retas pontilhadas azuis representam o valor da mediana e as retas pontilhadas vermelhas representam o valor da média.

Para uma melhor análise da distribuição dos dados considera-se a seguir, a representação com gráficos de violino (WILKE, 2019), (MONTGOMERY; RUNGER, 2021). Uma breve descrição sobre a construção desses gráficos é apresentada nas Seções A.4 e A.5 do Apêndice A. Para obter esses gráficos, conforme ilustrados na Figura 9, considerou-se todos os dados horários registrados para os anos de 2007 a 2023 agrupados por cada estação do ano.

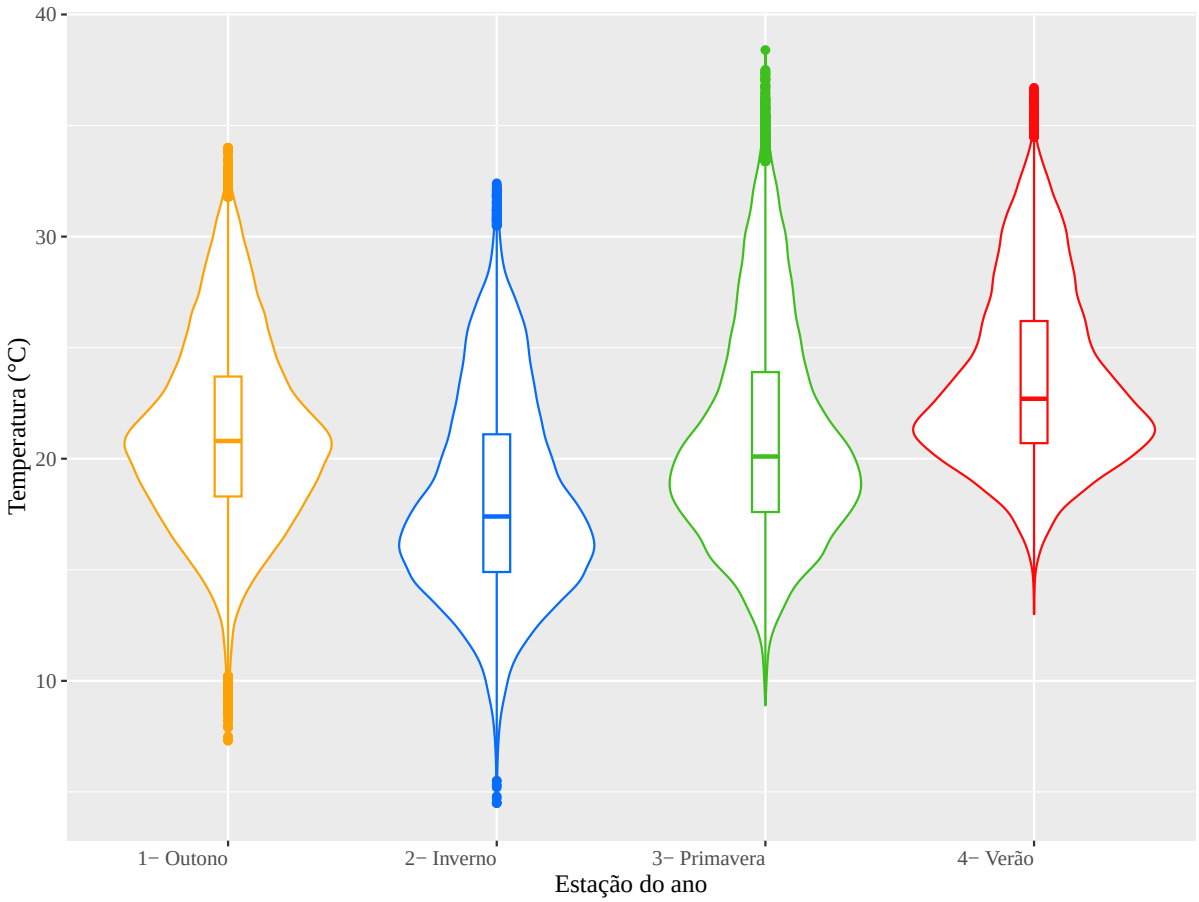

Figura 9 – Gráficos de violino da distribuição dos dados da temperatura para o outono, inverno, primavera e verão. Valores considerando todos os dados horários observados para os anos de 2007 a 2023.

Os gráficos de violino fornecem uma estimativa da densidade de probabilidade a partir do histograma suavizado com um kernel gaussiano. Porém, a suavização leva a uma distorção. A fim de reduzi-la nos extremos, considera-se a opção *trim* na função do violino. Essa opção limita os valores máximo e mínimo da estimativa da densidade de probabilidade aos valores do conjunto de dados. A partir desses gráficos, nota-se que os valores máximo e mínimo, mediana e a moda (valor em que a densidade estimada atinge valor máximo) são diferentes para cada estação do ano. As maiores temperaturas absolutas são registradas na primavera e no verão, sendo em ambos os casos maior que $\approx 35{,}0°C$, enquanto as menores temperaturas são registradas no inverno, com registros abaixo de $\approx 5{,}0°C$. As maiores modas ocorrem no verão e no outono. Nota-se que nas estações do inverno, primavera e verão, os dados da temperatura apresentam uma distribuição mais assimétrica à direita. Enquanto que no outono, os dados apresentam uma distribuição mais simétrica. Pode-se notar que na primavera ocorre uma maior quantidade de pontos extremos de valores mais altos da temperatura.

A seguir, verifica-se a tendência das variações de cada estação do ano ao longo dos anos. Considera-se as temperaturas máximas, médias e mínimas, conforme definidas na Seção 3.1.2, para cada estação do ano. Calcula-se as médias de cada uma dessas medidas de

temperatura e em seguida aplica-se a média móvel considerando janelas com comprimento de 5 anos ao longo de todos os anos analisados para cada estação do ano. Na Figura 10 estão ilustrados 3 gráficos que são relativos as médias das temperaturas máximas, médias e mínimas ao longo dos anos usados na média móvel. Cada gráfico contém 4 curvas relativas as estações do ano e uma quinta relativa à média anual (representada em preto) que é feita sem considerar à sazonalidade. A tendência do aumento das médias móveis fica mais evidente tomando uma reta que melhor se aproxima do conjunto de pontos de cada curva segundo o método dos mínimos quadrados. Como obter os coeficientes da equação da reta que melhor se aproxima de um conjunto de pontos segundo o método dos mínimos quadrados é brevemente revisto na Seção A.1 do Apêndice A. Uma descrição mais geral do método dos mínimos quadrados pode ser encontrada em (HAYKIN, 2009).

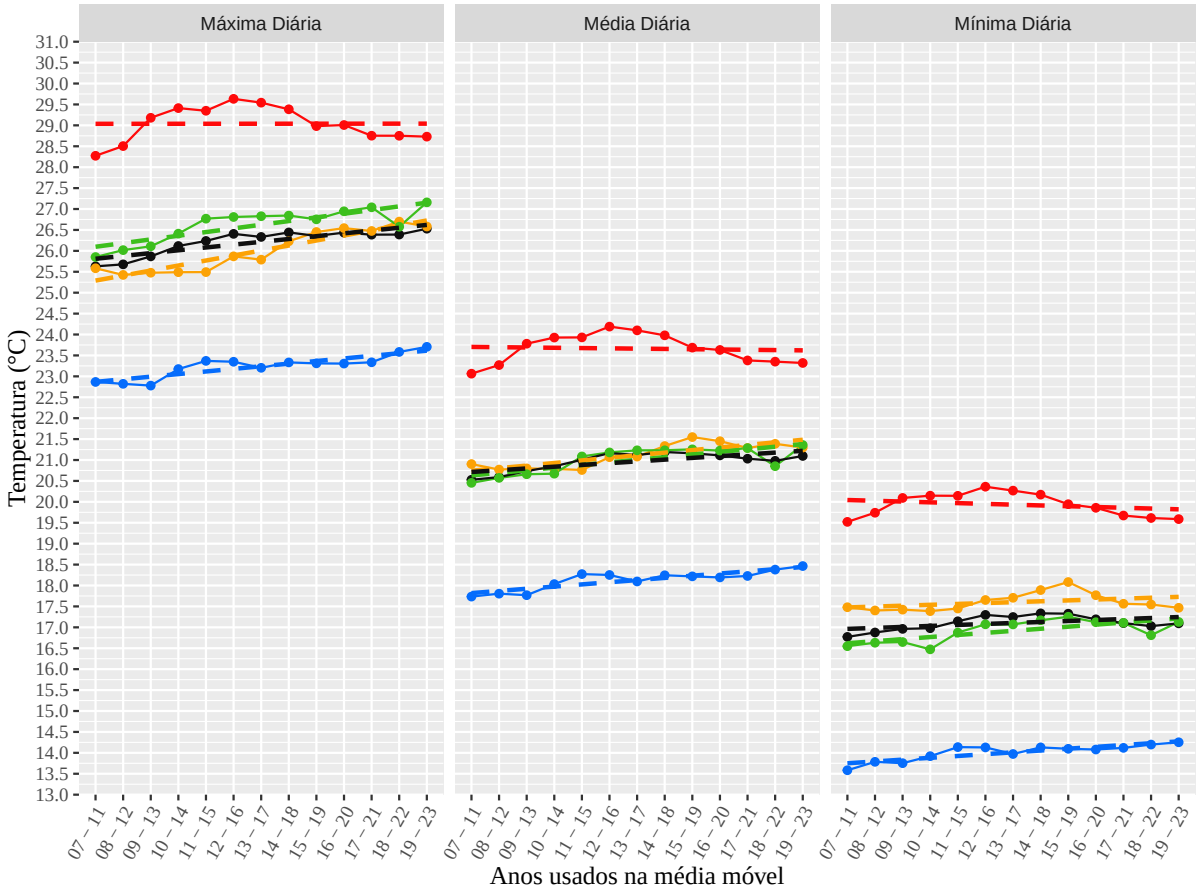

Figura 10 – Evolução da média móvel de 5 anos da temperatura máxima, média e mínima diária. As cores indicam o outono (amarelo), inverno (azul), primavera (verde) e verão (vermelho). Em cada um dos gráficos, a curva indicada em preto representa as médias sem considerar sazonalidade.

Observando as inclinações destas retas, nota-se que há uma tendência de aumento das médias da temperatura máxima, média e mínima diária ao longo dos anos. A tendência de aumento é mais marcante quando se observam os valores das máximas temperaturas diárias. No entanto, as inclinações são relativamente diferentes para cada estação do ano. Particularmente no outono e na primavera, as retas apresentam uma maior inclinação,

ou seja, é observada uma maior tendência de aumento da temperatura nessas estações. Em relação as máximas diárias, pode-se observar uma tendência de aumento de 1,4°C no outono e 1,0°C na primavera. No caso do inverno, observa-se uma tendência de aumento de 0,7°C. O verão é a única estação do ano onde não se observa uma tendência clara de aumento da temperatura. Destaca-se que a média anual feita sem considerar a sazonalidade (representada em preto) torna menos evidente a tendência de aumento se comparada ao outono e a primavera.

Na Tabela 6, apresenta-se a evolução da amplitude térmica para cada estação do ano considerando médias móveis de cinco anos. Nota-se um aumento da amplitude térmica para todas as estações com o passar dos anos. Os maiores aumentos são observados nas estações do outono (aumento de $\approx$ 1,0°C) e primavera (aumento de $\approx$ 0,7°C) ao longo dos anos considerados. Verifica-se que esse aumento acontece principalmente por conta do maior aumento das temperaturas máximas diárias, conforme evidenciado pela Figura 10 (b).

Tabela 6 – Evolução da amplitude térmica por estação do ano considerando médias móveis de cinco anos.

Anos usados na média móvel	Período Sazonal				
	Outono	Inverno	Primavera	Verão	Anual
07 - 11	8,1	9,3	9,3	8,8	8,8
08 - 12	8,0	9,0	9,4	8,8	8,8
09 - 13	8,1	9,0	9,5	9,1	8,9
10 - 14	8,1	9,3	9,9	9,3	9,1
11 - 15	8,0	9,2	9,9	9,2	9,1
12 - 16	8,2	9,2	9,7	9,3	9,1
13 - 17	8,1	9,2	9,8	9,3	9,1
14 - 18	8,3	9,2	9,7	9,2	9,1
15 - 19	8,4	9,2	9,5	9,0	9,0
16 - 20	8,8	9,2	9,8	9,2	9,2
17 - 21	8,9	9,2	9,9	9,1	9,3
18 - 22	9,2	9,4	9,8	9,1	9,4
19 - 23	9,1	9,5	10,0	9,1	9,4
Diferença (19-23 *versus* 07-11)	1,0	0,2	0,7	0,3	0,6

3.1.4 OBSERVAÇÃO HORÁRIA

Considera-se os dados agrupados ainda conforme as estações do ano, porém as médias feitas para cada hora do dia ao longo dos anos. Os 4 gráficos das médias horárias da temperatura em função das horas do dia ao longo dos anos são ilustradas na Figura 11. Como considerou-se os anos de 2007 a 2023, cada gráfico contém 17 curvas. A partir das 17×4 curvas nota-se que:

- Há uma variação significativa da média da temperatura conforme as hora do dia.
- O valor médio de cada hora vai depender do ano e também da estação do ano.

- Os valores máximos observados acontecem na primavera e no verão entre 13h e 15h.
- Os valores mínimos observados acontecem no inverno entre 3h e 8h.
- Em cada conjunto de 17 curvas por estação do ano, nota-se que há valores horários diferentes, porém, não é possível evidenciar se há alguma tendência ao longo dos anos.

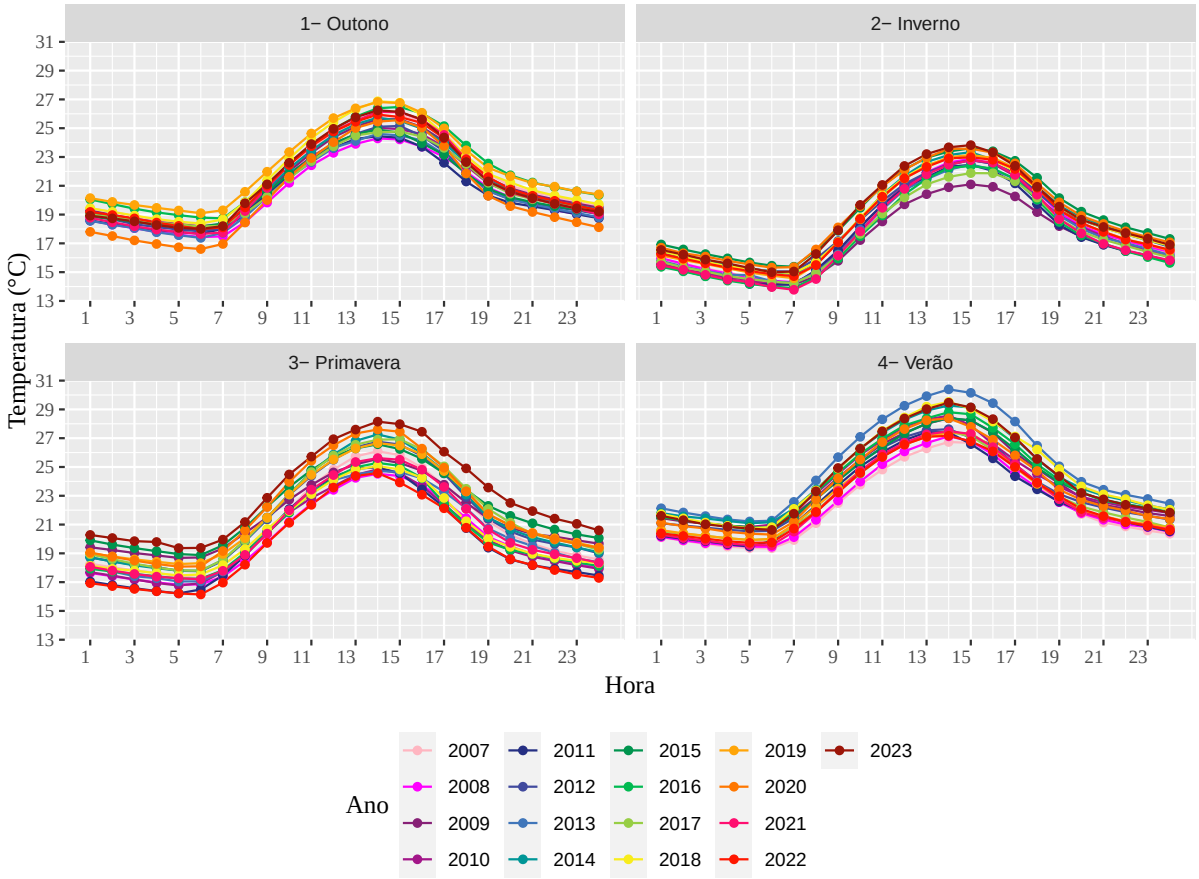

Figura 11 – Representação da média da temperatura ao longo das horas do dia sendo (1) outono, (2) inverno, (3) primavera e (4) verão. Total são 17 curvas, considerando os dados horários observados entre os anos de 2007 a 2023.

Com objetivo de atenuar as variações das médias horárias observadas a cada ano, calcula-se o valor da média horária da temperatura em conjuntos específicos de anos. Os valores são apresentados na Figura 12. Especificamente, foram considerados os valores médios de temperatura por hora nos períodos de 2007 a 2011, 2012 a 2018 e 2019 a 2023. As curvas mostram que no outono, primavera e inverno, entre 11h e 19h, há uma tendência de aumento das médias horárias de temperatura ao longo dos anos, especialmente à tarde. Às 14h, por exemplo, a temperatura média de 2019 a 2023 aumentou aproximadamente 1,3°C na primavera, 1,1°C no outono e 0,9°C no inverno, em comparação com 2007 a 2011. No verão, não se observa uma tendência clara de aumento da temperatura.

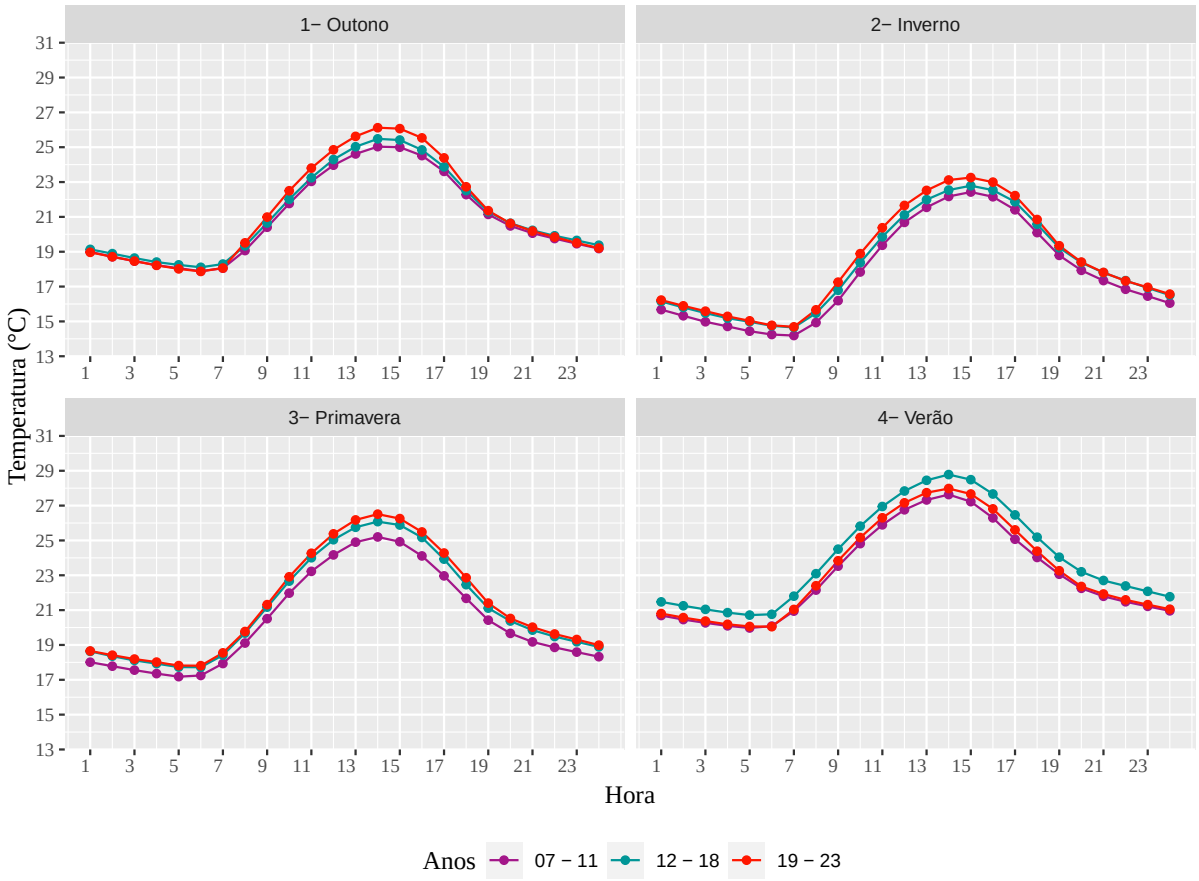

Figura 12 – Representação da média da temperatura ao longo das horas dos dia, agrupadas em conjuntos de anos e conforme a estação do ano.

3.2 ANÁLISE DOS DADOS DO OZÔNIO (O_3)

A concentração de O_3 pode ser prejudicial aos ecossistemas e à saúde humana quando ultrapassa certos limites. Em 2021, a Organização Mundial da Saúde (OMS) publicou diretrizes globais de qualidade do ar, recomendando que a máxima concentração média em oito horas diárias não ultrapasse 100 $\mu g/m^3$ para garantir uma proteção adequada à saúde humana (ORGANIZATION, 2021).

De forma similar aos resultados apresentados para a série da temperatura do ar, são avaliados nesta seção a tendência do comportamento da concentração de O_3 e as suas estatísticas ao longo dos dias, ao longo das estações do ano e ao longo das horas do dia. Com base na análise de dados faltantes (Seção 2.4), considera-se os dados horários da concentração de O_3 dos anos 2007 e 2009 a 2023 (excluindo os dados de 2008).

3.2.1 OBSERVAÇÃO TEMPORAL

Uma primeira análise acerca dos dados da concentração de O_3 pode ser feita pela inspeção visual de alguns trechos da série, conforme apresentado na Figura 13.

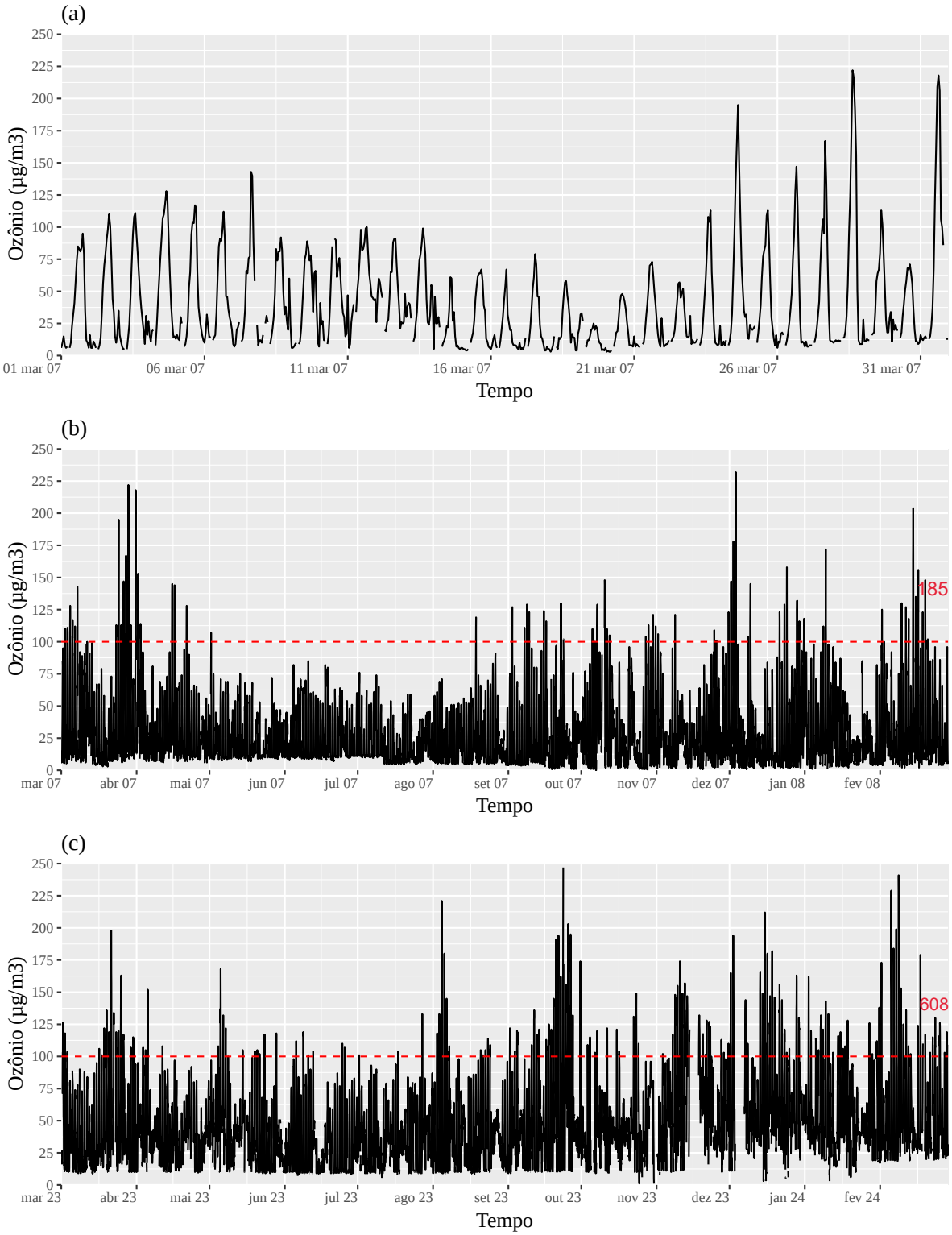

Figura 13 – Série horária da concentração de O_3 para os períodos: (a) março/2007, (b) ano de 2007 e (c) ano de 2023. A linha horizontal vermelha indica o limiar de 100 $\mu g/m^3$, destacando o número de vezes que a concentração de O_3 excedeu esse valor. Nos anos de 2007 e 2023, esse limiar foi ultrapassado 185 e 608 vezes, respectivamente.

A partir dos gráficos nota-se que:

◇ No gráfico da Figura 13 (a) está o trecho da concentração de O_3 ao longo do mês

de março de 2007. Nota-se que o seu comportamento varia significativamente ao longo dos dias. Os valores mínimos e máximos são iguais aos dias do mês, porém há variações significativa desses valores ao longo dos dias. Além disso, há picos de concentração relativamente elevados em determinadas horas.

- ◇ Nos gráficos das Figuras 13 (b) e (c) estão os trechos das séries ao longo dos anos de 2007 e de 2023, respectivamente. Nota-se que as maiores concentrações de O_3 estão associadas à sazonalidade. Nos meses mais quentes do ano, observam-se as maiores concentrações. Além disso, é possível notar uma grande variação entre as concentrações ao longo dos anos. Enquanto que no ano de 2007 são observados 185 pontos acima do limiar de 100 $\mu g/m^3$, no ano de 2023 esse valor mais que triplica, sendo observados 608 pontos acima do mesmo limiar.

- ◇ Pode-se ressaltar ainda a presença de dados faltantes. Essa característica fica evidente através da Figura 13 (a), visto que são notados alguns "buracos" em alguns intervalos dos dias. Nota-se que os intervalos com ausência de dados não necessariamente possuem a mesma duração, ou seja, o número de dados ausentes a cada intervalo não é sempre o mesmo.

3.2.2 OBSERVAÇÃO DA MÉDIA MÓVEL HORÁRIA

Com o intuito de apresentar como as concentrações de O_3 evoluíram ao longo dos anos, considerou-se os valores diários observados para esse poluente. Para isso, seguindo as normas estabelecidas em estudos de qualidade do ar (CETESB, 2023), obtêm-se os valores diários de O_3 por meio do cálculo da maior média móvel de 8 horas do dia e horário da ocorrência. Especificamente, inicia-se por aplicar, a cada hora do dia, a média em 8 amostras consecutivas de dados; para a $i-$ésima hora do dia calcula-se

$$\bar{p}_i = \frac{p_i + p_{i-1} + \cdots + p_{i-7}}{8}.$$

Na primeira hora do dia, para obter $\bar{p}_1$, ao valor de p_1 são somados os valores de 7 amostras do dia anterior, e assim sucessivamente. Desse modo, na última hora do dia, para obter $\bar{p}_{24}$, são somados os valores de p_{17} a p_{24}. Os valores de p_{18} a p_{24} serão usados para calcular o $\bar{p}_1$ do dia posterior. Portanto, a cada dia, tem-se 24 pontos que correspondem a uma média móvel de 8 horas. Posteriormente, considera-se o maior valor dos 24 pontos para cada um dos dias.

Após o cálculo do valor máximo da média móvel da concentração de O_3, ou seja, $\max(\bar{p}_i)$ com $i = 1, 2, \ldots, 24$, a série com 140.160 amostras (24 horas $\times$ 365 dias $\times$ 16 anos) resultou em uma série 5.840 amostras (365 dias $\times$ 16 anos). Em seguida, fez-se uma média aritmética considerando todas as amostras de cada ano resultando, então, em apenas 16 amostras representando cada ano. Com base nas amostras dos anos de 2007 a 2023, excluindo apenas o ano de 2008, calculou-se uma média móvel com janelas de

comprimento de 5 anos. Assim, como foi feito na séria da temperatura, a média móvel de cinco anos foi utilizada com o objetivo de atenuar as variações meteorológicas que são observadas de ano para ano. No caso do O_3, o primeiro ponto da média móvel considerou os anos de 2007 a 2012 (excluindo o ano de 2008). O resultado dos valores médios da concentração de O_3 ao longo dos anos está ilustrado na curva da Figura 14. Constata-se um aumento do valor médio da concentração de O_3 de aproximadamente 14 $\mu g/m^3$ quando se compara a média dos valores observados nos anos de 2019 a 2023 em relação aos cinco primeiros anos da série.

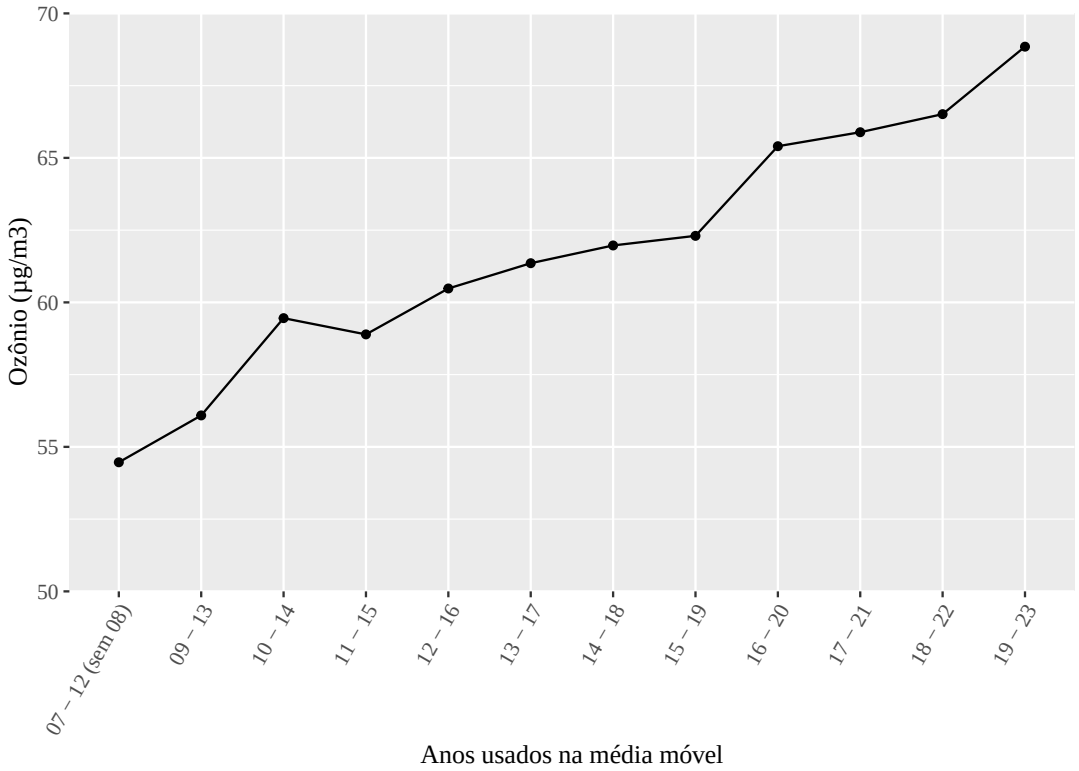

Figura 14 – Evolução da média móvel com janelas de 5 anos dos valores médios da concentração de O_3.

Devido à variabilidade dos valores de ozônio ao longo do tempo, pode-se considerar que a série temporal do ozônio também não possui a propriedade de estacionariedade. A seguir, em razão das características de não estacionariedade da série do O_3 e levando em consideração a relação entre o aumento da temperatura e a elevação da concentração do O_3, avalia-se a tendência do aumento da concentração do O_3 considerando os períodos de sazonalidade do ano.

3.2.3 OBSERVAÇÃO SAZONAL

Visando quantificar como a concentração de O_3 está associada à sazonalidade, inicialmente considerou-se todos os dados horários registrados para cada estação do ano ao longo dos anos de 2007 a 2023 (sem o ano de 2008). Na Figura 15 estão ilustrados os histogramas da concentração de O_3 para cada uma das estações. Em cada um dos

histogramas, destaca-se retas pontilhadas indicando os valores médio e da mediana. Nota-se que, independentemente do estação analisada, a concentração de O_3 apresenta uma distribuição com assimetria positiva ou à direita. Nota-se uma longa "cauda" dos pontos da distribuição à direita. As estações da primavera e do verão são os períodos do ano onde há uma incidência maior de altas concentrações de O_3. Avaliando o cumprimento das caudas das distribuições, pode-se notar que o histograma da primavera apresenta a cauda relativamente maior, com uma maior ocorrência de valores superiores a 100 $\mu g/m^3$. A estação do inverno uma cauda relativamente menor.

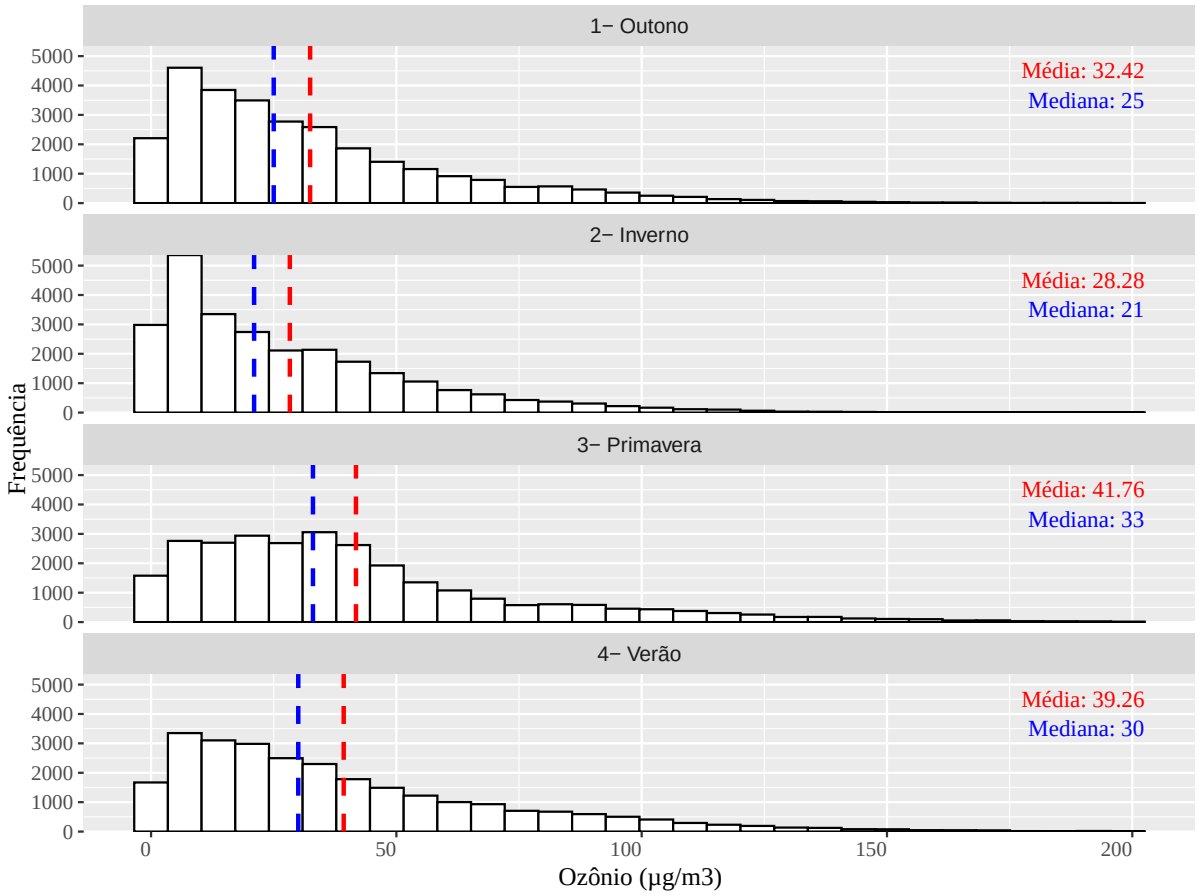

Figura 15 – Histogramas da concentração de O_3 ao longo das estações do ano. Valores considerando todos os dados horários registrados para os anos de 2007 a 2023 (sem o ano de 2008). As retas pontilhadas azuis representam o valor da mediana e as retas pontilhadas vermelhas representam o valor da média.

A fim de avaliar como os níveis de poluição do O_3 estão variando ao longo do ano, observou-se a quantidade de dias contendo concentrações de O_3 acima do limiar de 100 $\mu g/m^3$ para cada uma das estações. Na Tabela 7 apresenta-se a quantidade de dias em que a maior média móvel de 8 horas do dia e horário da ocorrência é superior a esse limiar para cada estação ao longo do período analisado. A partir da observação dos valores apresentados na Tabela 7 destacam-se:

- Os níveis de concentração de O_3 registrados na estação Parque D. Pedro II ultrapas-

sam os limites recomendados pela OMS em vários momentos da série. Ao longo dos anos analisados, observa-se 633 dias não consecutivos com níveis acima do limiar.

- Primavera e verão são as estações do ano que apresentam as maiores quantidades de dias com níveis de O_3 acima do padrão ideal. Essas duas estações concentram $\approx 80\%$ do total de dias em que os níveis de O_3 estão acima do limiar.
- As maiores ocorrências históricas de dias com níveis de O_3 acima do padrão ideal de qualidade são observados na primavera de 2020 (com 33 dias) e no verão de 2023 (com 30 dias).
- Nas estações com as temperaturas mais baixas do ano, observa-se que a maioria dos dias apresentam níveis de O_3 dentro dos limites de qualidade. O inverno é a estação que apresenta o melhor nível de qualidade do ar com relação ao O_3.

Tabela 7 – Quantidade de dias com concentrações de O_3 acima do limiar de qualidade do ar estabelecido pela OMS (100 $\mu g/m^3$).

Ano	Estação do ano				Total
	1 - Outono	2 - Inverno	3 - Primavera	4 - Verão	
2007	6	0	1	9	16
2009	3	0	5	6	14
2010	2	11	17	11	41
2011	1	1	17	13	32
2012	3	0	23	10	36
2013	4	2	18	29	53
2014	2	2	23	17	44
2015	5	5	12	11	33
2016	9	4	15	20	48
2017	3	3	24	12	42
2018	10	1	6	21	38
2019	6	1	22	8	37
2020	11	2	33	5	51
2021	10	5	20	17	52
2022	4	3	14	8	29
2023	10	3	24	30	67
Total	89	43	274	227	633

Para facilitar a avaliação da distribuição do O_3 considera-se a representação através dos gráficos de violino. Os resultados são apresentados na Figura 16. A partir da observação da Figura 16, nota-se uma grande quantidade de pontos discrepantes (*outliers*) em todas as estações do ano. Além disso, assim como observado nos dados da temperatura, nota-se que os valores máximo e mínimo, mediana e a moda das concentrações de ozônio são diferentes para cada estação do ano. As maiores ocorrências de valores acima de 100 $\mu g/m^3$ são registrados na primavera e no verão. Já no outono e inverno, nota-se uma concentração de O_3 relativamente menor, com maior faixa de valores próximos a zero. As maiores modas ocorrem no outono e no inverno, porém são para valores relativamente baixos.

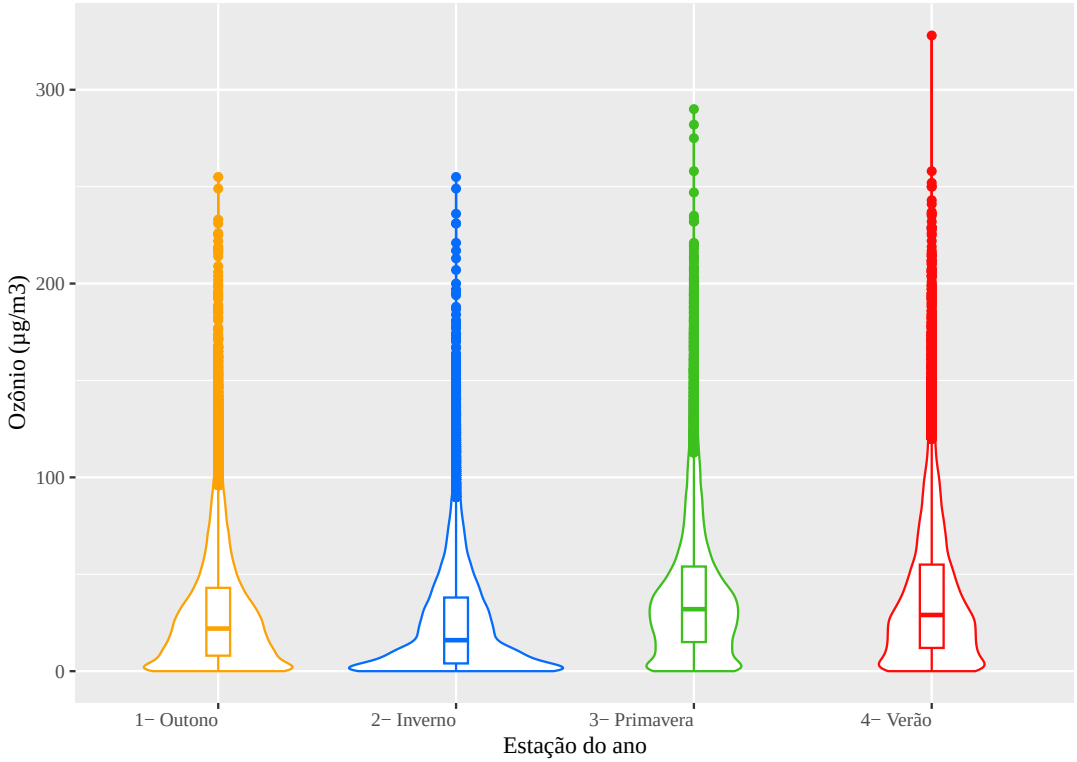

Figura 16 – Gráficos de violino da distribuição dos dados da concentração de ozônio para o outono, inverno, primavera e verão. Valores considerando todos os dados horários observados para os anos de 2007 a 2023 (excluindo os dados de 2008).

A partir dos valores observados ao longo das estações do ano, fez-se o cálculo da média móvel conforme apresentado na Seção 3.2.2. Especificamente, na Figura 17 apresenta-se a evolução das médias móveis com janela de 5 anos das concentrações diárias de O_3 para cada estação do ano. A avaliação da tendência de aumento das concentrações de O_3 fica mais evidente tomando as retas que melhor se ajustam ao conjunto de pontos. As retas apresentadas na Figura 17 são as retas que melhor se ajustam ao conjunto de 12 amostras segundo o método dos mínimos quadrados. As inclinações das retas evidenciam que há uma tendência de aumento do O_3 em todas as estações do ano. A reta com maior inclinação representa o outono com um aumento na média da concentração de O_3 de 21 $\mu g/m^3$ entre os anos de 2007 a 2023. O inverno apresenta a segunda maior tendência de aumento, de cerca de 16 $\mu g/m^3$, seguido pela primavera com 10 $\mu g/m^3$ e o verão com 6 $\mu g/m^3$. A reta que representa a média anual, ou seja, não considerando a sazonalidade, releva uma tendência de aumento da concentração deste poluente de 13 $\mu g/m^3$ ao longo do período analisado.

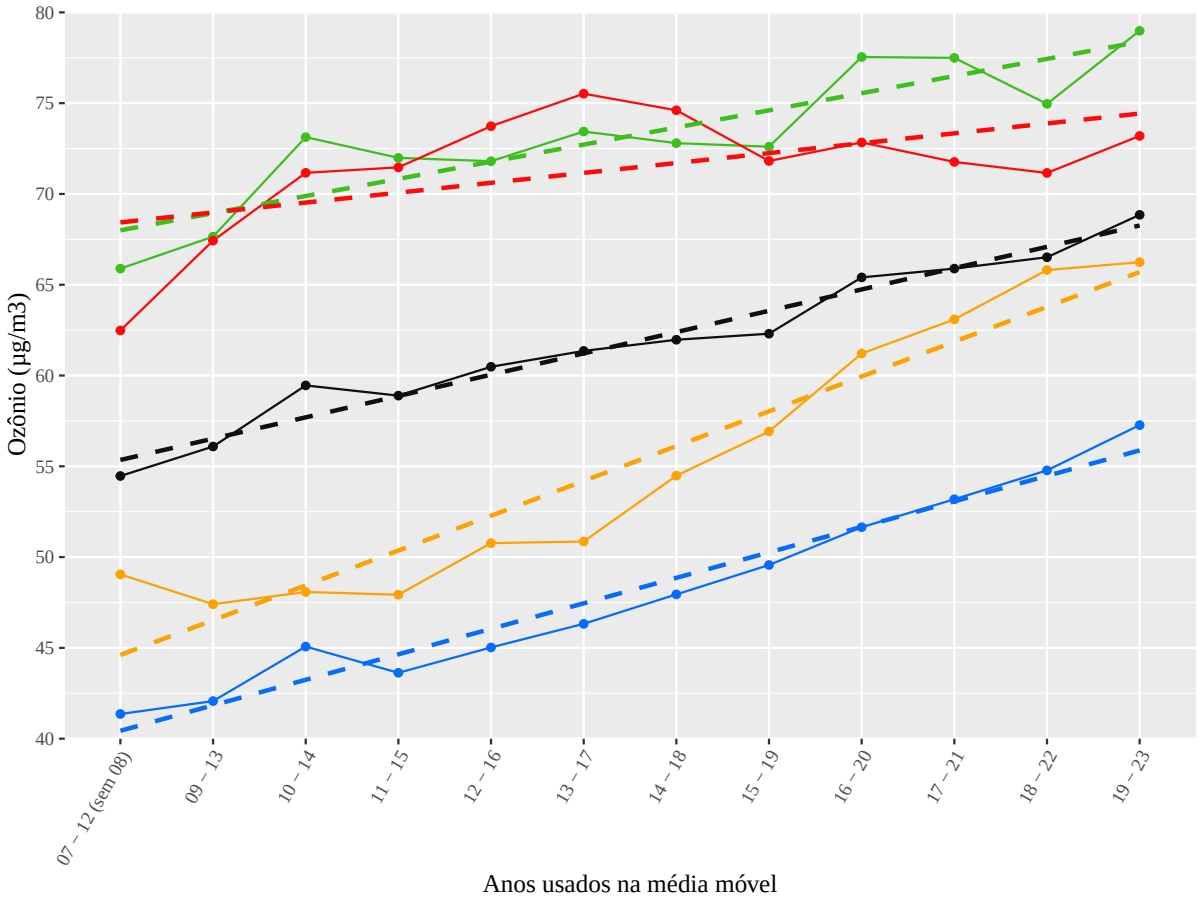

Figura 17 – Evolução da média móvel de 5 anos da concentração de O_3. As cores indicam o outono (amarelo), inverno (azul), primavera (verde) e verão (vermelho). Em cada um dos gráficos, a curva indicada em preto representa as médias sem considerar sazonalidade.

3.2.4 OBSERVAÇÃO HORÁRIA

A seguir agrupam-se as amostras das concentrações de O_3 para cada hora do dia conforme as estações ano para todos os anos entre 2007 e 2023 (com exceção de 2008). Inicialmente, para cada ano fez-se o cálculo da média aritmética das concentrações de O_3 para cada hora do dia. O resultado da concentração média de O_3 ao longo das horas do dia é apresentado na Figura 18. Em cada um dos 4 gráficos são apresentadas 16 curvas, uma curva para cada ano. Como há dados faltantes às 6 horas da manhã, pois esse é o horário em que o equipamento sofre manutenção, nas curvas não há valores indicados nesse horário.

A partir da observação das curvas dos gráficos da Figura 18 destacam-se que:

- A concentração de O_3 apresenta um comportamento cíclico ao longo do dia semelhante ao da temperatura.
- Entre o período da noite e o começo da manhã (por volta das 20 às 9 horas) a concentração de O_3 apresenta um comportamento aproximadamente constante, registrando as menores concentrações do dia.

- Ao longo da manhã, pode-se observar que ocorre um aumento da concentração de O_3. As maiores concentrações de O_3 são observadas ao longo das horas em que ocorre as maiores temperaturas do dia, em geral, o pico acontece no período entre 12 e 17 horas.
- Posteriormente, no fim da tarde e início da noite, as concentrações de ozônio voltam a cair para níveis mais baixos.
- Pode-se mencionar ainda que existe uma grande variação entre as médias horárias de ano para ano.

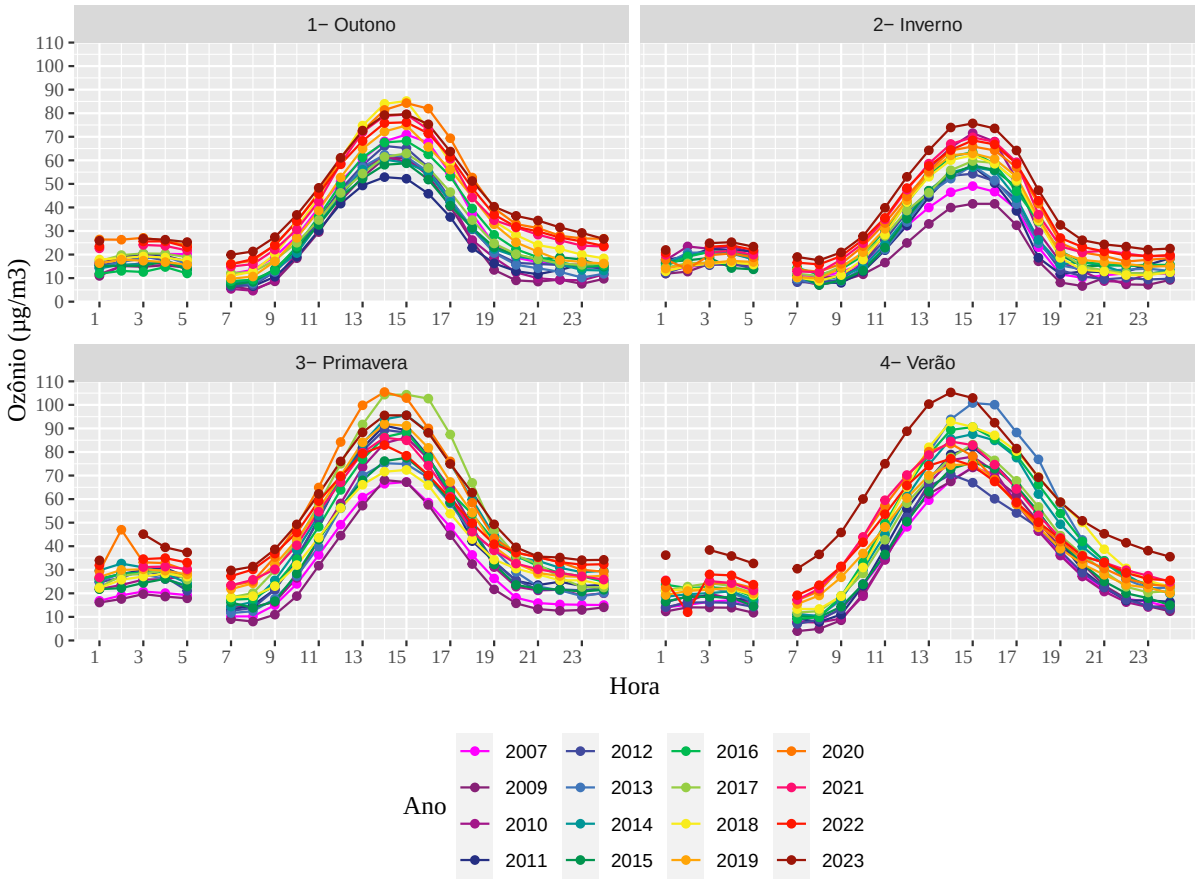

Figura 18 – Representação da média da concentração de O_3 ao longo das horas do dia sendo (1) outono, (2) inverno, (3) primavera e (4) verão. Total são 16 curvas, considerando os dados horários observados entre os anos de 2007 a 2023 (excluindo o ano de 2008).

Visando melhor quantificar a evolução da concentração de O_3 para cada hora do dia ao longo dos anos, considerou-se os dados apresentados Figura 18. Porém, a cada hora aplicou-se uma média considerando um conjunto de anos. Na Figura 19 é apresentado o resultado dessa nova média. No lugar de 16 curvas, tem-se apenas 3 curvas para cada estação do ano. Cada uma das curvas representa a média de uma dada hora na janela de anos de 2007 a 2012 (sem 2008), 2013 a 2018 e de 2019 a 2023. Nota-se que no outono, primavera e inverno, entre 11h e 19h, há uma tendência de aumento das médias horárias das

concentrações de O_3 ao longo dos anos. Especificamente às 14h, a média da concentração de O_3 nos anos de 2019 a 2023 apresenta um aumento em relação aos anos de 2007 a 2012 (sem 2008) de aproximadamente 16 $\mu g/m^3$ no outono, 15 $\mu g/m^3$ no inverno e cerca de 13 $\mu g/m^3$ na primavera. No verão apesar de não ser possível notar uma tendência de aumento na metade/final da tarde, às 14h observa-se um aumento de 13 $\mu g/m^3$.

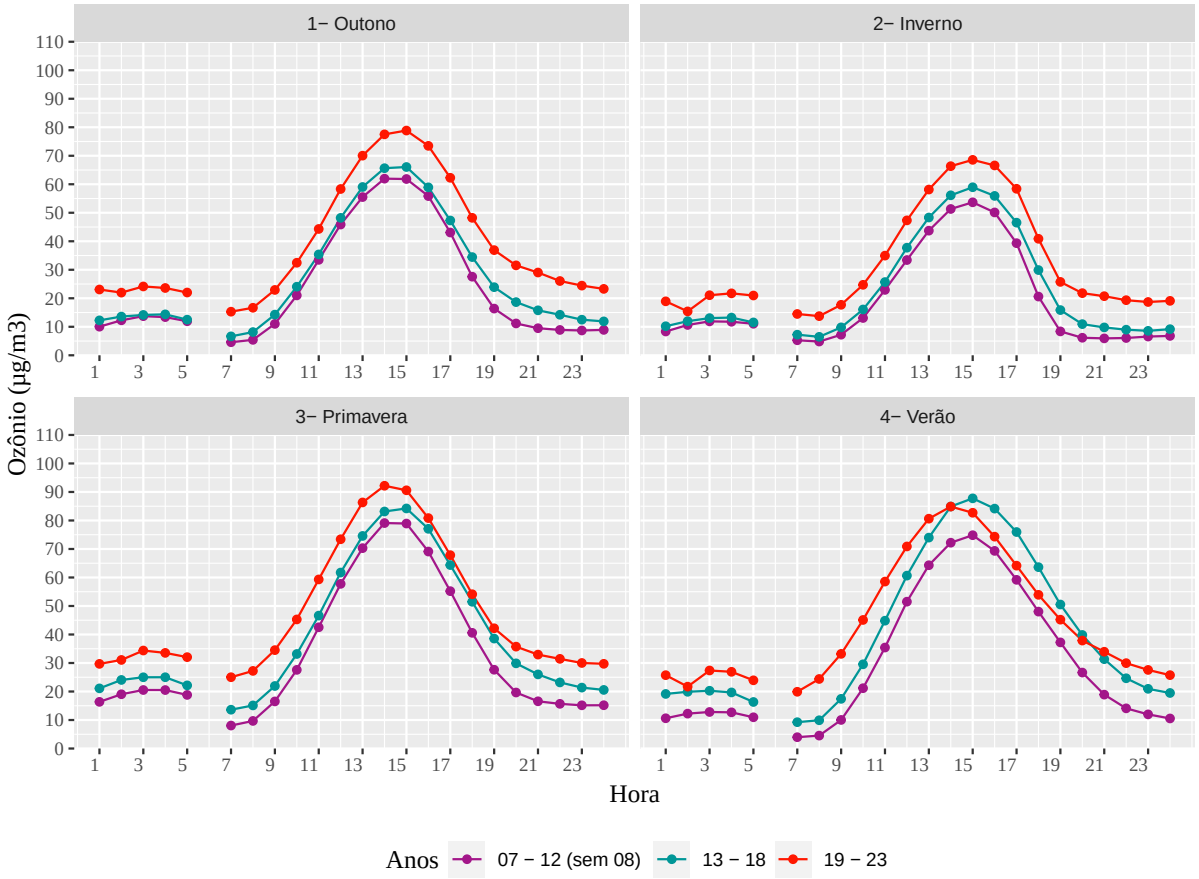

Figura 19 – Representação da média da concentração de O_3 ao longo das horas do dia, agrupadas em conjuntos de anos e conforme a estação do ano.

3.3 ANÁLISE DOS DADOS DA RADIAÇÃO SOLAR GLOBAL

A radiação solar global é a principal fonte primaria de energia de onde derivam quase todas as outras formas de energia, sendo responsável por diversos processos naturais que ocorrem na superfície terrestre. Ao aquecer a superfície, a radiação solar influencia diversos fenômenos de caráter meteorológicos como a evaporação das águas dos rios e oceanos e, consequentemente, a formação de nuvens e a ocorrência de precipitação. Além disso, a radiação solar também influencia em outros processos ambientais relacionados à temperatura do ar e do solo, afetando a troca de calor por meio da evaporação e transpiração (SPOKAS, 2009). Devido à sua importância, são analisados os dados da radiação solar global para avaliar a sua relação com as concentrações de O_3. Na Figura 20 apresenta-se a série horária da radiação solar global.

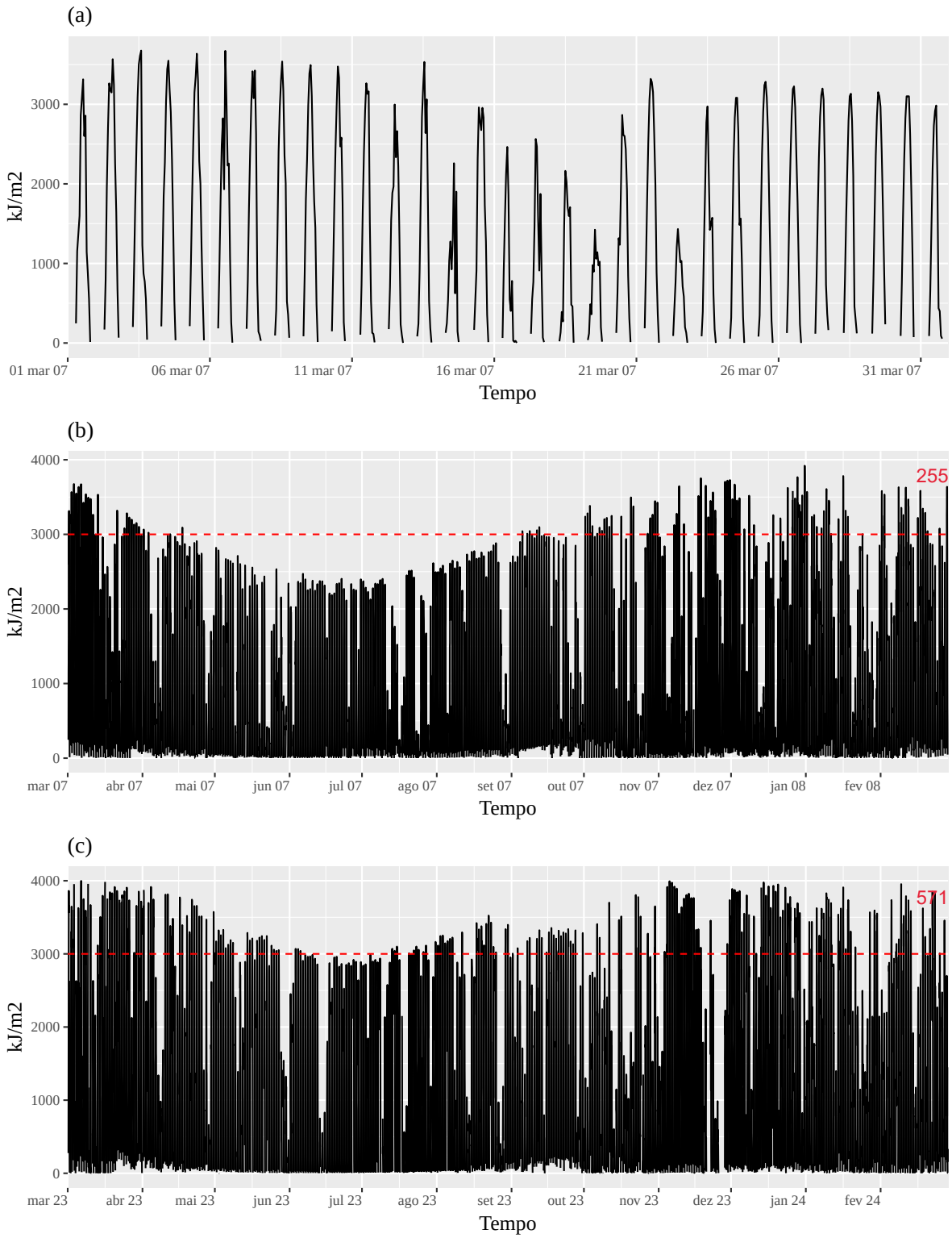

Figura 20 – Série horária da radiação solar global para os períodos: (a) março/2007, (b) ano de 2007 e (c) ano de 2023. A linha horizontal vermelha indica o limiar de 3.000 KJ/m^2, destacando o número de vezes que a radiação solar global excedeu esse valor. Nos anos de 2007 e 2023, esse limiar foi ultrapassado 255 e 571 vezes, respectivamente.

A partir da observação dos gráficos da Figura 20 destacam-se que:

◇ No gráfico da Figura 20 (a) está o trecho da radiação solar global ao longo do mês

de março de 2007. Nota-se que o seu comportamento varia significativamente ao longo dos dias. Os valores mínimos e máximos são iguais aos dias do mês, porém há variações significativa desses valores ao longo dos dias.

◇ Nos gráficos das Figuras 20 (b) e (c) estão os trechos das séries ao longo dos anos de 2007 e de 2023, respectivamente. Analisando a série ao longo do período de um ano, pode-se notar que a variação da radiação solar global está associada a sazonalidade (primavera, verão, outono e inverno). Destaca-se que a maior incidência de radiação solar global é registrada no início e fim de cada ano (novembro a abril), enquanto que a menor incidência é registrada no meio do ano (maio a agosto). Além disso, comparando as curvas das Figuras 20 (b) e (c) é possível notar que a radiação apresenta grandes variações ao longo dos anos. Nota-se que nos meses entre o final da primavera e meados do outono, quando ocorrerem os maiores picos de temperaturas, a RADG ultrapassa os 3.000 KJ/m^2. Considerando então esse valor como limiar. Nota-se que, enquanto no ano de 2007 são observados 255 pontos acima de 3.000 kJ/m^2, no ano de 2023 esse valor praticamente dobra, sendo observados 571 pontos acima do mesmo limiar.

- Pode-se ressaltar ainda a presença de dados faltantes. Essa característica fica evidente através da Figura 20 (a), visto que a radiação solar global só é medida nas horas em que ocorrem maior incidência de luz solar.

Níveis de radiação solar global acima de 3.000 kJ/m^2 indicam uma quantidade substancial de energia solar atingindo a superfície terrestre, o que tem implicações significativas para a geração de energia solar, o clima, a saúde humana e o meio ambiente. Esses níveis podem aumentar a produção de energia solar, mas também elevar os riscos de danos à saúde, como queimaduras solares e câncer de pele, e causar estresse térmico em plantas e humanos (GUEYMARD, 2004).

Para uma melhor análise a respeito da distribuição da radiação solar global, considera-se a seguir a representação dos dados através do gráfico de violino. Os resultados são apresentados na Figura 21. Para a construção desses gráficos, considerou-se todos os dados horários observados entre os anos de 2007 e 2023 agrupados por cada estação do ano. Assim como observado nos dados da temperatura e do O_3, os valores máximo e mínimo, mediana e a moda são diferentes para cada estação do ano. A maior incidência de faixa de valores relativamente altos da radiação solar global, acima dos $\approx$ 4.000 kJ/m^2, ocorre na primavera e no verão. As modas acontecem para valores relativamente baixos de radiação solar global, porém, as maiores modas ocorrem no outono e no inverno. Os dados não estão concentrados em torno da mediana e a há faixas de valores relativamente altos acima da mediana, notadamente no verão e na primavera. No inverno a incidência de valores acima da mediana fica significativamente concentrada em torno de 2.000 a 3.000 kJ/m^2.

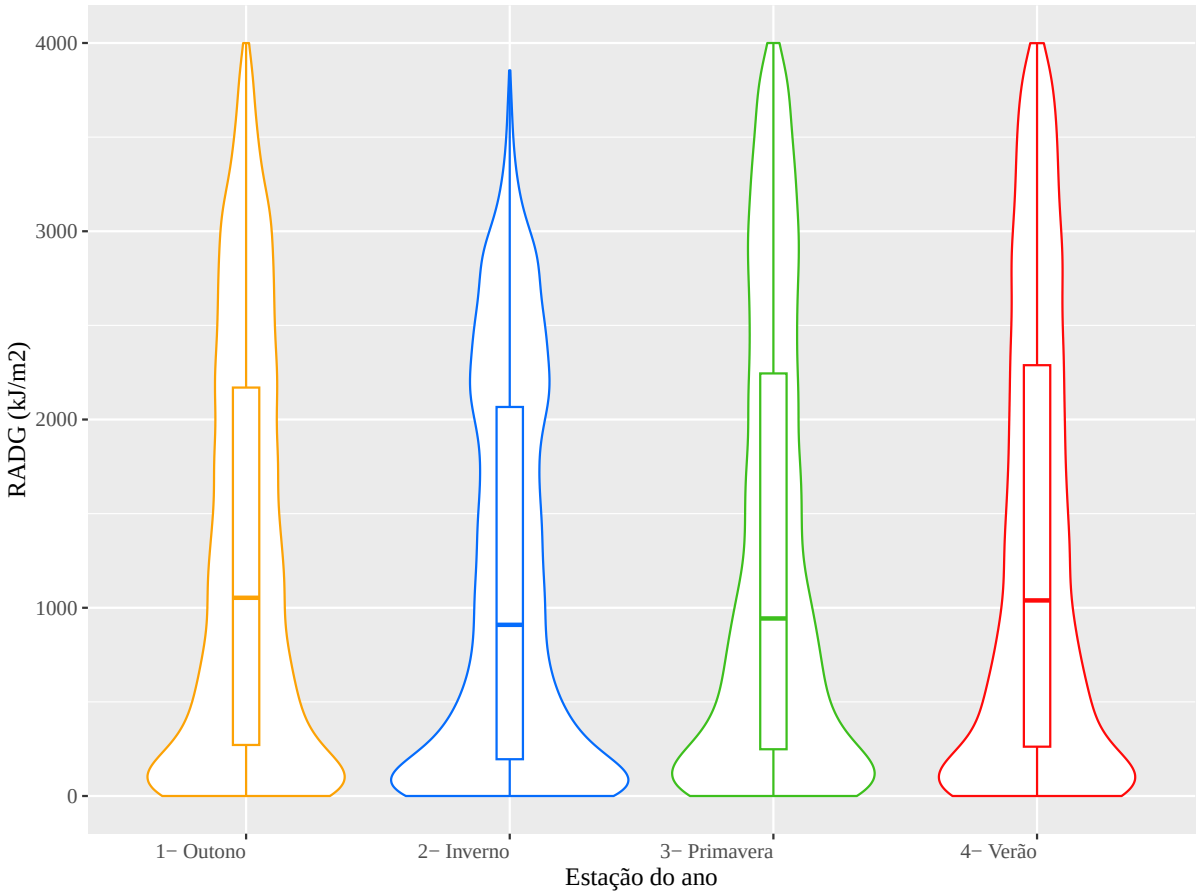

Figura 21 – Gráficos de violino da distribuição dos dados da incidência da radiação solar para o outono, inverno, primavera e verão. Valores considerando todos os dados horários observados para os anos de 2007 a 2023.

A fim de avaliar a evolução média da incidência da radiação solar global para cada hora do dia ao longo dos anos, calcula-se a o valor médio de cada média horária em conjuntos de anos (Figura 22). Em específico, cada curva representa a média de uma dada hora na janela de anos de 2007 a 2011, 2012 a 2018 e de 2019 a 2023. A partir da observação das curvas, é possível notar que na estação do outono, inverno e na primavera, há uma tendência de aumento na incidência da radiação solar global em todas as horas do dia, onde os dados são observados, com o passar dos anos. A maior tendência de aumento da radiação solar global é observada para a estação do outono, seguida pelo inverno e posteriormente primavera. Apenas na estação do verão, não se observa uma tendência clara de aumento da radiação solar global com o passar dos anos. Especificamente às 12h, a média da incidência da radiação solar global nos anos de 2019 a 2023 apresenta um aumento em relação aos anos de 2007 a 2011 de aproximadamente 605 kJ/m^2 no outono, 486 kJ/m^2 no inverno, 137 kJ/m^2 na primavera e 44 kJ/m^2 no verão.

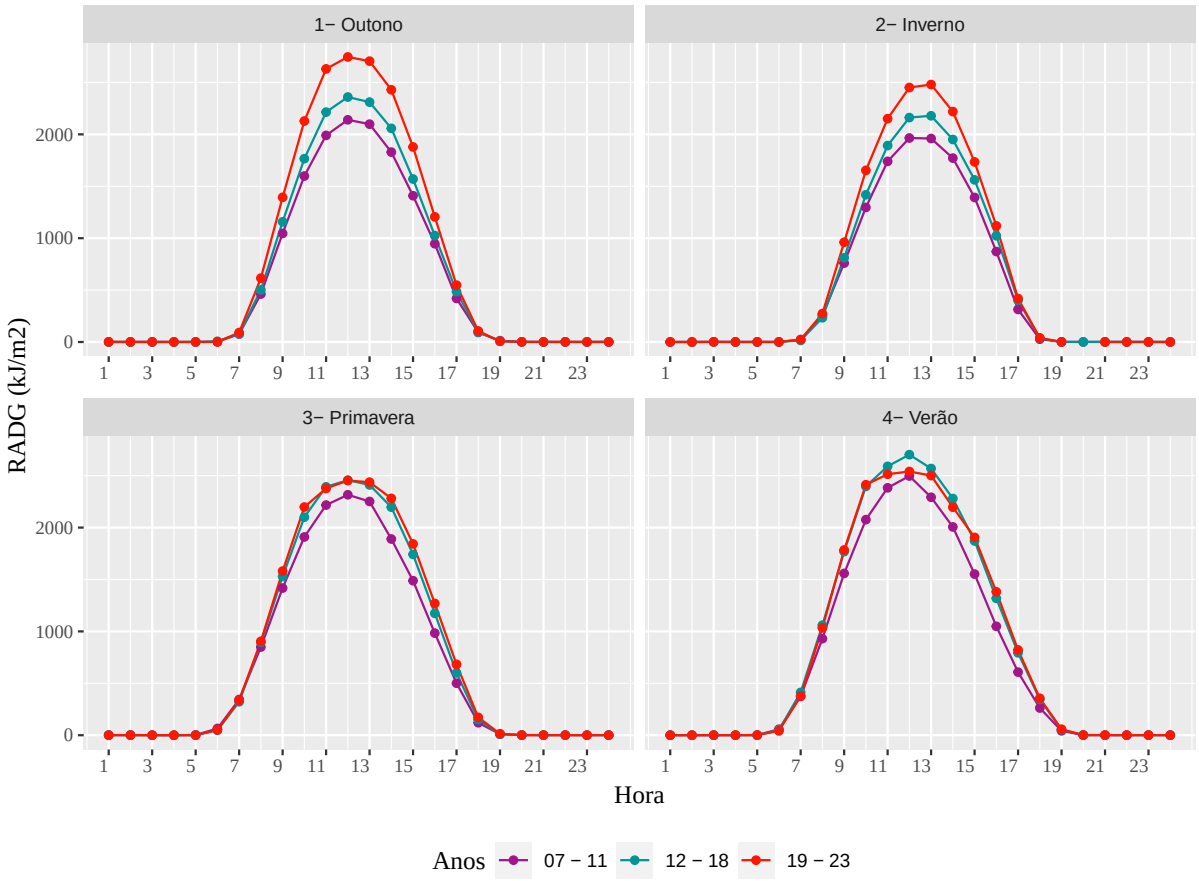

Figura 22 – Representação das curvas médias da incidência da radiação solar global ao longo das horas, agrupadas em conjuntos de anos e conforme a estação do ano.

A partir dos valores observados ao longo das estações do ano, fez-se o cálculo da média móvel considerando os valores diários da radiação solar global. O valor diário da radiação solar é representado pelo maior valor diário obtido a partir das medidas horários do dia. O cálculo é similar ao feito na Seção 3.1.2 para o cálculo da série da temperatura máxima diária. Na Figura 23 apresenta-se a evolução das médias móveis com janela de comprimento de 5 anos da incidência de radiação solar global para cada estação do ano. A tendência de aumento das médias móveis fica mais evidente tomando uma reta que melhor se aproxime do conjunto de pontos de cada curva segundo o método dos mínimos quadrados. As inclinações das retas evidenciam que há uma tendência de aumento para todas as estações. Nota-se uma tendência de aumento de aproximadamente 900 kJ/m^2 no outono, 600 kJ/m^2 no inverno, 460 kJ/m^2 na primavera e 300 kJ/m^2 no verão. A reta representada pela média anual, ou seja, que não considera à sazonalidade, releva uma tendência de aumento da incidência de radiação solar global de 550 kJ/m^2 ao longo do período analisado. Similarmente aos dados da temperatura e do O_3, nota-se a propriedade da não-estacionariedade da série.

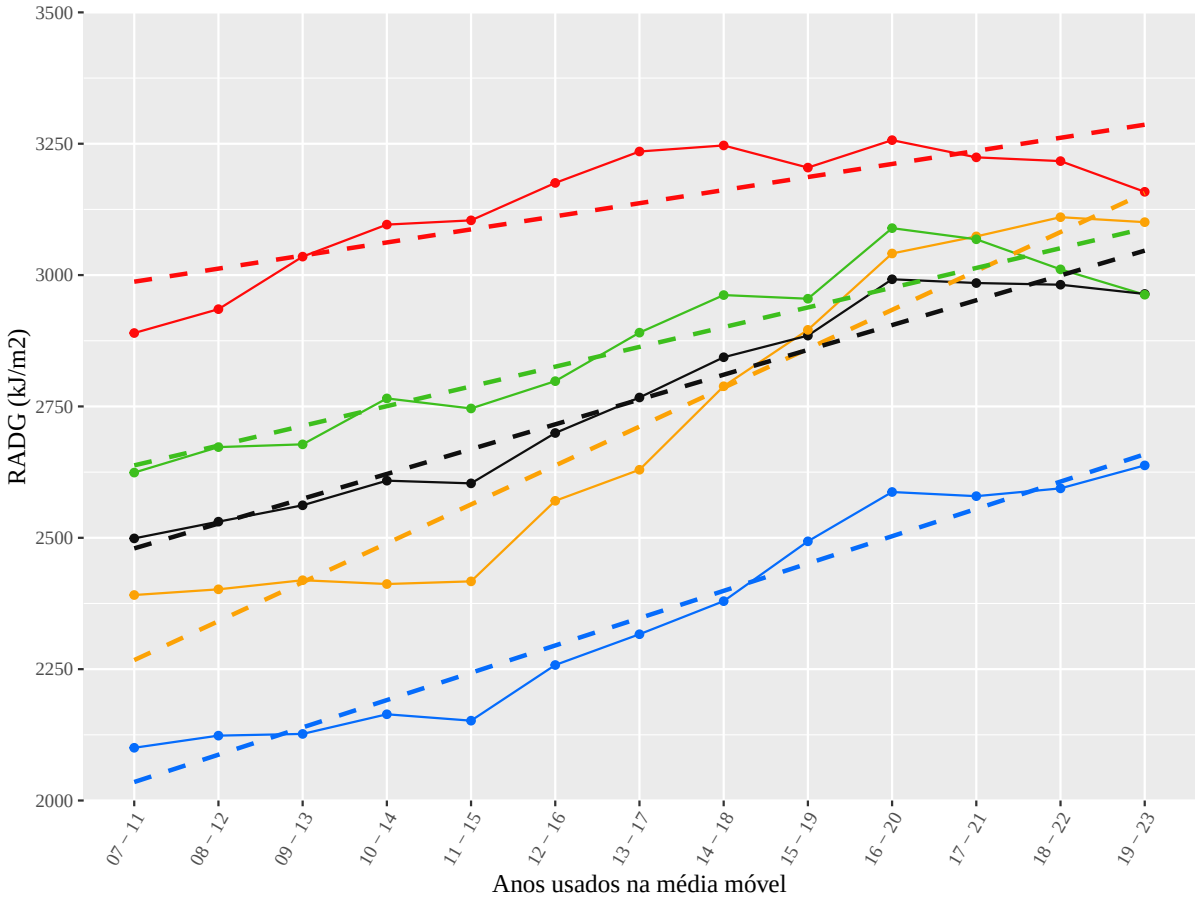

Figura 23 – Evolução da média móvel de 5 anos da incidência da radiação solar global. As cores indicam o outono (amarelo), inverno (azul), primavera (verde) e verão (vermelho). Em cada um dos gráficos, a curva indicada em preto representa as médias sem considerar sazonalidade.

3.4 OBSERVAÇÃO CONJUNTA DAS DEMAIS VARIÁVEIS

Além dos dados da temperatura, ozônio e radiação solar global, a CETESB e o INMET também disponibilizam outros dados relativos ao clima e a concentração de poluentes do ar. Especificamente, são considerados nessa seção a relação entre a concentração de ozônio e as demais variáveis relativas aos poluentes e às condições climáticas.

Sobre as variáveis relativas às condições climáticas:

◇ umidade relativa do ar (UR; %), velocidade do vento (VV; m/s), pressão atmosférica (PRESS; mB) e precipitação (PRE; mm).

Sobre as variáveis relativas aos poluentes:

◇ monóxido de carbono (CO; ppm), partículas inaláveis (MP_{10}; $\mu g/m^3$), dióxido de nitrogênio (NO_2; $\mu g/m^3$), monóxido de nitrogênio (NO; $\mu g/m^3$).

Além do O_3, os demais poluentes atmosféricos considerados nesse trabalho também apresentam índices de referência para medir a qualidade do ar. A Tabela 8 apresenta o padrão de qualidade do ar para os poluentes CO, O_3, MP_{10} e NO_2. Essa qualificação do ar está vinculada à norma legal (Resolução CONAMA n° 491/2018) e independe do padrão de qualidade/meta intermediária em vigor, visto que está associada aos efeitos à saúde humana (CETESB, 2023). Individualmente, cada poluente apresenta diferentes efeitos sobre a saúde da população para faixas de concentração distintas, identificados por estudos epidemiológicos desenvolvidos dentro e fora do país. Apesar disso, os índices diferentes de bons são prejudiciais a saúde humana e ao meio ambiente com maior ou menor escala.

Tabela 8 – Padrões de qualidade do ar para o CO, O_3, MP_{10} e NO_2.

Poluentes	Período de Referência	Padrões de Qualidade do Ar				
		Boa	Moderada	Ruim	Muito Ruim	Péssima
CO	8 horas[1]	0 - 9	>9 – 11	>11 – 13	>13 – 15	>15
MP_{10}	24 horas	0 - 50	>50 – 100	>100 – 150	>150 – 250	>250
NO_2	1 hora[2]	0 - 200	>200 – 240	>240 – 320	>320 – 1130	>1130
O_3	8 horas[1]	0 - 100	>100 – 130	>130 – 160	>160 – 200	>200

1 - máxima média móvel obtida no dia
2 - média horária

A seguir, são observadas as médias horárias das demais variáveis consideradas nesse trabalho. Para isso, são avaliados apenas os dados observados entre os anos de 2010 a 2023, pois é o período consecutivo de anos onde são observados dados com poucas informações faltantes para todas as variáveis.

3.4.1 OBSERVAÇÃO HORÁRIA

A fim de verificar o comportamento médio das variáveis ao longo das horas do dia conforme as estações do ano, fez-se o cálculo da média horária dos dados observados entre os anos de 2010 a 2023. Na Figura 24 é apresentado o resultado dessa média. Nota-se que todas as variáveis apresentam em média uma variação ao longo das horas do dia e ao longo das estações do ano. Nos horários entre 12h e 17h, indicados nos gráficos com retas verticais pontilhadas, observa-se a maior concentração de O_3.

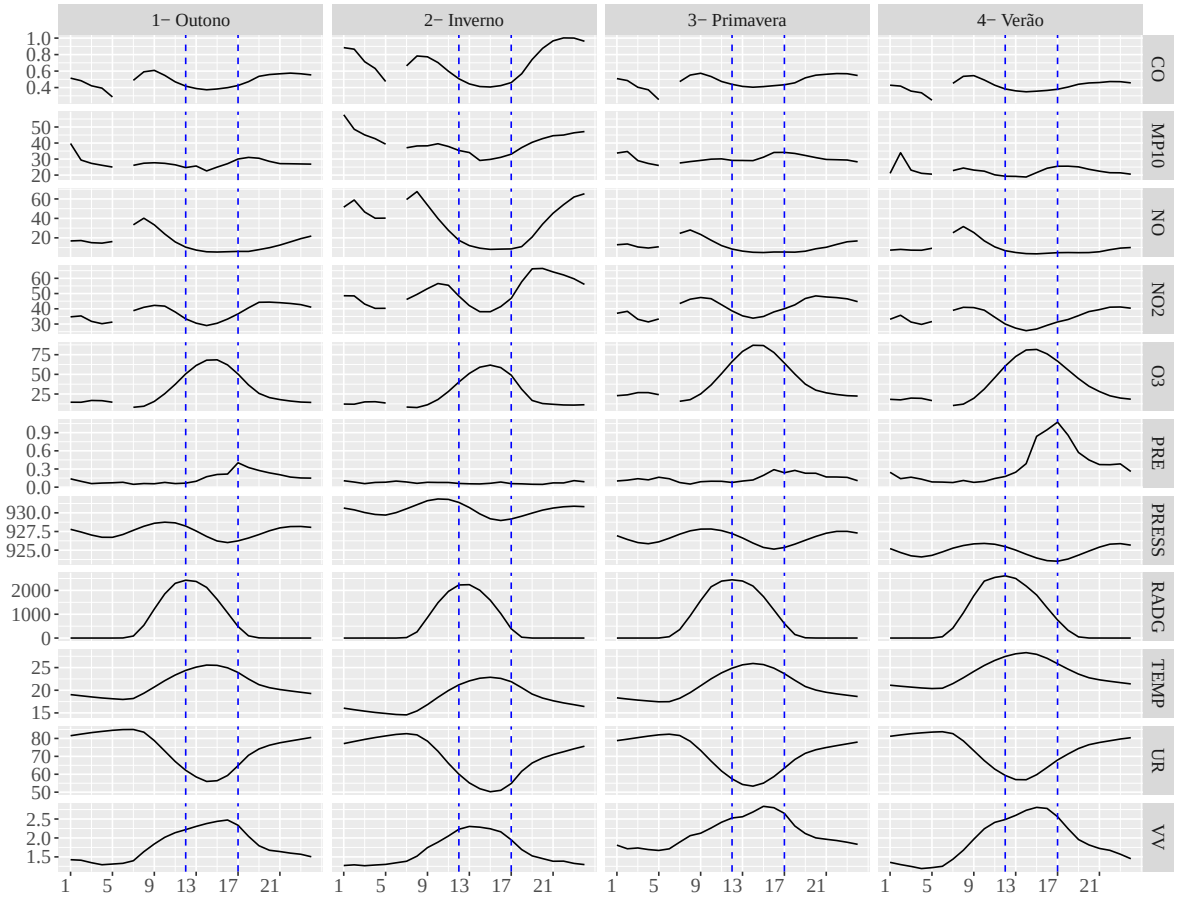

Figura 24 – Representação da média das variáveis para cada hora do dia no (1) outono, (2) inverno, (3) primavera e (4) verão. Cada curva é obtida com uma média dos dados observados entre os anos de 2010 a 2023.

Utiliza-se o recorte do trecho das séries considerando apenas as horas das 12h às 17h para o desenvolvimento das análises seguintes.

3.4.2 OBSERVAÇÃO SAZONAL

Para avaliar como as diferentes variáveis se comportam ao longo do ano, fez-se o cálculo do valor médio de cada uma das variáveis considerando as estações do ano. Esse cálculo foi feito considerando os primeiros 5 anos (2010 a 2014) e os últimos 5 anos (2019 a 2023) para avaliar o efeito da tendência. O cálculo do valor médio foi feito considerando apenas os horários das 12 às 17 horas (Tabela 9). A partir da observação dos valores representes na Tabela 9 destacam-se:

- Nota-se uma redução considerável na média dos poluentes (MP_{10}, NO, NO_2 e CO) quando se compara o período de 2010 a 2014 com o período de 2019 a 2023 para todas as estações do ano. Além disso, as maiores concentrações destes poluentes ocorrem na estação do inverno.

- Nota-se um aumento considerável do O_3 com o aumento da radiação solar global. Em especial, para a estação do outono. Nessa estação, a média da concentração de ozônio nos anos de 2019 a 2023 cresceu $\approx 34\%$ em relação a média dos anos de 2010 a 2014. Em relação a radiação solar global, esse aumento foi de $\approx 31\%$.

- A temperatura também apresenta um aumento considerável para as estações do outono, inverno e primavera. Por outro lado, a umidade relativa reduz consideravelmente com o aumento da temperatura.

Tabela 9 – Valor médio das variáveis por estação do ano para o período de 2010 a 2014 e 2019 a 2023 e o seu % de variação. Valores considerando os dados horários das 12 às 17 horas.

Variáveis	1 - Outono			2 - Inverno		
	2010 a 2014	2019 a 2023	% Variação	2010 a 2014	2019 a 2023	% Variação
O_3	52,5	70,1	34%	49,2	60,9	24%
MP_{10}	28	23,1	-17%	35,1	31,3	-11%
NO	9,5	4,1	-57%	15,1	6	-60%
NO_2	37,1	26,8	-28%	51,5	33,7	-34%
CO	0,6	0,2	-58%	0,6	0,3	-46%
PRESS	926,7	926,7	0%	929,7	929,8	0%
RADG	1448,7	1899,1	31%	1410,3	1736,3	23%
TEMP	24,3	25,4	4%	22	22,6	3%
UR	62,5	56,1	-10%	55,8	51,6	-7%
VV	2,5	2,1	-19%	2,3	2	-16%
PRE	0,2	0,2	17%	0,1	0	-64%
Variáveis	3 - Primavera			4 - Verão		
	2010 a 2014	2019 a 2023	% Variação	2010 a 2014	2019 a 2023	% Variação
O_3	76	81,9	8%	72,6	76,3	5%
MP_{10}	33,3	31,3	-6%	23	20,1	-12%
NO	7,9	4,5	-43%	6	3,1	-47%
NO_2	43,6	30,9	-29%	33,1	23	-30%
CO	0,6	0,3	-47%	0,5	0,2	-57%
PRESS	926	925,8	0%	924,1	924,2	0%
RADG	1665,2	1785,8	7%	1764,1	1859	5%
TEMP	24,9	25,7	3%	27,7	27,2	-2%
UR	56,8	56,3	-1%	60,9	61,7	1%
VV	2,7	2,6	-6%	2,7	2,5	-8%
PRE	0,2	0,2	26%	0,7	0,5	-18%

Além da observação dos valores médios, pode-se analisar a variabilidade dos parâmetros ao longo das estações do ano. O gráfico de *boxplot* foi utilizado com essa finalidade. Todos os valores foram normalizados pela Fórmula Min-Max (MONTGOMERY; RUNGER, 2021) para melhor entendimento e comparação dos resultados obtidos pelo estudo dos *boxplot*. Na Figura 25 é apresentado o resultado desta construção. A partir dos *boxplot* é possível observar a variabilidade dos dados durante os 14 anos e para cada estação ano. Observando os *boxplot* da Figura 25 nota-se que:

- Os poluentes MP_{10}, NO, NO_2 e CO apresentam pequena amplitude e concentrados nos valores mais baixos. Apesar disso, pode-se notar uma grande ocorrência de pontos

extremos. Além disso, a maior ocorrência dos pontos extremos está concentrada nas estações do inverno e primavera.

- A RADG e a TEMP apresentam maior variabilidade na estação da primavera, período sazonal de maior ocorrência do O_3.
- Pode-se destacar que nas estações mais quentes do ano (primavera e o verão) e com maior ocorrência de O_3, há uma maior variabilidade dos dados climáticos (especialmente RADG, TEMP e UR), quando comparados com as estações mais frias (inverno e outono).

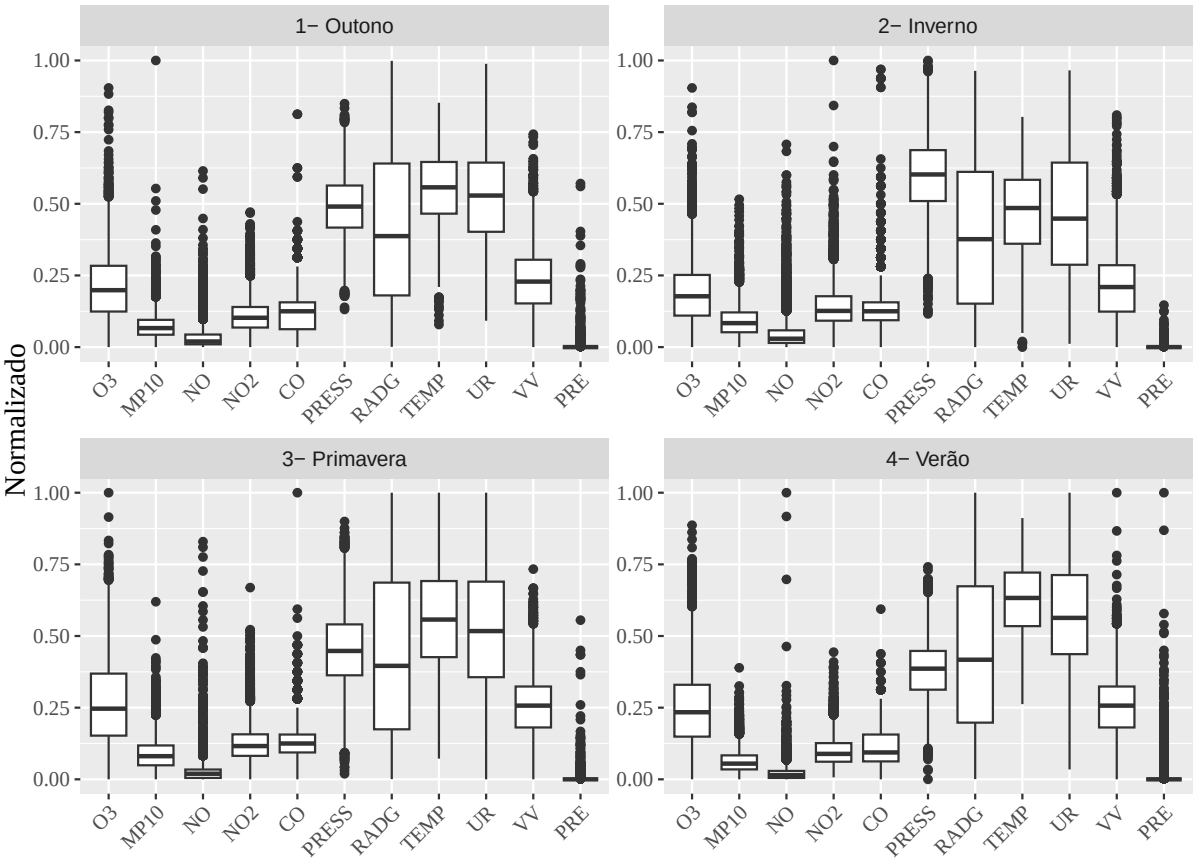

Figura 25 – *Boxplot* das variáveis normalizadas. Valores considerando os dados das 12 às 17 horas dos anos de 2010 a 2023.

3.4.3 OBSERVAÇÃO DA MÉDIA MÓVEL

Com o intuito de apresentar como as variáveis evoluíram ao longo dos anos, calculou-se a média móvel. Inicialmente, aplicou-se o cálculo da média aritmética para cada um dos anos das sequências de dados observados no período das 12 às 17 horas para cada estação do ano. A partir dos dados anuais, calculou-se uma média móvel com janelas de comprimento de 5 anos para avaliar a evolução das variáveis ao longo dos anos. O resultado

é apresentado na Figura 26. A cada gráfico está associada a reta que melhor de ajusta ao conjunto de 10 pontos, segundo o método dos mínimos quadrados.

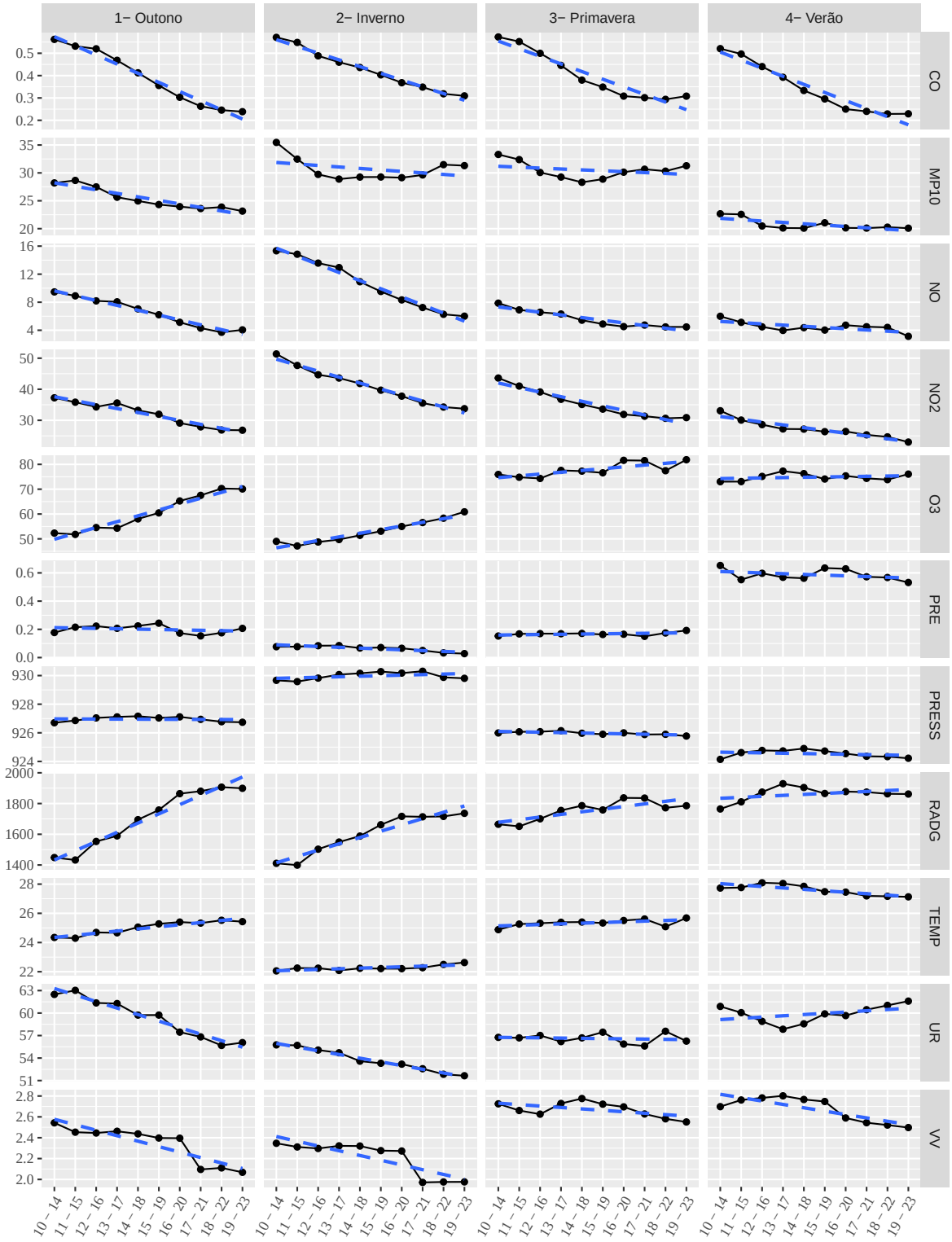

Figura 26 – Evolução da média móvel com janelas de 5 anos das variáveis em estudo.

Em relação a cada um dos gráficos da Figura 26 destacam-se:

- As inclinações das retas de tendência da médias da variáveis são relativamente diferentes para cada estação do ano. Há uma tendência de queda das médias dos poluentes CO, MP_{10}, NO e NO_2 ao longo dos anos. Mas a concentração média do O_3 é o único poluente em que se observa uma tendência de aumento ao longo dos anos. Essa tendência é relativamente mais significativa no outono e no inverno e menos significativa no verão.

- A RADG apresenta uma tendência de aumento ao longo dos anos para todas as estações. Particularmente no outono, a reta ajustada apresenta maior inclinação.

- A TEMP apresenta tendência de aumento para as estações do outono, inverno e primavera. Assim como observado na série da RADG, o outono é a estação com maior tendência de aumento nos valores da temperatura, notadamente pela maior inclinação da reta. No caso da estação do verão, não é possível notar um aumento da temperatura assim como foi observado nas demais estações. Inclusive, nota-se uma tendência relativamente pequena de queda da temperatura.

- A VV apresenta uma tendência de queda ao longo dos anos para todas as estações. Apesar disso, essa tendência é menor nas estações da primavera e verão.

- A UR apresenta uma tendência de queda para as estações do outono e inverno. Já no verão, observa-se um aumento da UR com o passar dos anos. Na primavera não é possível observar nenhuma tendência.

- A PRESS e a PRE não apresentam nenhum comportamento claro de tendência. As retas ajustadas apresentam inclinações relativamente pequenas ao longo dos anos.

3.4.4 COEFICIENTE DE CORRELAÇÃO LINEAR DE PEARSON

A partir das análises anteriores, tem-se uma primeira noção dos parâmetros possivelmente correlacionados com o O_3. Entre eles a temperatura do ar e a radiação solar global, por exemplo. A próxima análise é entender como as variáveis estão correlacionadas com o ozônio através do cálculo do coeficiente de Correlação Linear de Pearson (PEARSON, 1896), (MONTGOMERY; RUNGER, 2021). Esse coeficiente é uma medida utilizada para medir o quanto a ocorrência de duas variáveis aleatórias (v.a.) estão linearmente correlacionadas. Em outras palavras, mede o grau da correlação linear entre duas variáveis. O cálculo do coeficiente de Correlação Linear de Pearson é brevemente revisto na Seção A.2 do Apêndice A. Na Tabela 10 apresentam-se os valores dos coeficientes de Correlação Linear de Pearson do O_3 em função das demais variáveis consideradas nesse trabalho.

Tabela 10 – Coeficiente de Correlação de Pearson do O_3 em função das demais variáveis. Os valores dos dados são das 12 às 17 horas por estação do ano calculados para o período de anos de 2010 a 2014 e 2019 a 2023.

Variáveis	Dados dos anos (2010 a 2014)				Dados dos anos (2019 a 2023)			
	Outono	Inverno	Primavera	Verão	Outono	Inverno	Primavera	Verão
MP_{10}	0,15	0,17	0,32	0,26	0,22	0,34	0,39	0,35
NO	-0,6	-0,55	-0,54	-0,59	-0,34	-0,5	-0,25	-0,24
NO_2	0,05	-0,08	-0,03	-0,11	-0,16	-0,19	-0,13	-0,05
CO	-0,09	-0,19	-0,02	-0,1	-0,07	-0,13	0,04	0,04
PRESS	-0,24	-0,22	-0,18	-0,02	-0,16	-0,24	-0,34	0,01
RADG	0,45	0,37	0,44	0,26	0,34	0,29	0,54	0,47
TEMP	0,71	0,7	0,75	0,64	0,68	0,68	0,82	0,75
UR	-0,65	-0,72	-0,7	-0,53	-0,55	-0,72	-0,78	-0,68
VV	-0,08	0	-0,1	0,1	0,06	-0,04	0,03	0,01
PRE	-0,1	-0,17	-0,03	-0,01	-0,05	-0,14	-0,07	-0,12

A partir da observação dos valores da Tabela 10 destacam-se:

- As variáveis que apresentam as maiores correlações com o O_3 são a TEMP, UR, RADG, NO e MP_{10}.
- As variáveis NO_2, CO, PRESS, VV e PRE apresentam correlações muito baixas com o O_3.
- Avaliando a correlação das variáveis com o O_3 ao longo das estações do ano, pode-se notar que na primavera, período de maior concentração do O_3, há uma forte correlação deste poluente com a RADG e o MP_{10}. Já no inverno, estação marcada pela menor concentração de O_3, percebe-se uma menor correlação do O_3 com a RADG e o MP_{10}, porém uma maior correlação com o NO.
- Os valores dos coeficientes de correlação para os anos de 2010 a 2014 se mantiveram parecidos com os anos de 2019 a 2023. Porém, é importante ressaltar algumas mudanças como o aumento (em valor absoluto) da correlação do O_3 com o MP_{10} para todas as estações do ano e com a RADG, TEMP e UR para as estações da primavera e verão.

Para complementar os resultados do Coeficiente de Correlação Linear de Pearson, são apresentados os gráficos de dispersão do O_3 em função das demais variáveis. Esses gráficos estão apresentados nas figuras 27 e 28 para cada uma das estações do ano considerando os dados dos conjuntos de anos de 2010 a 2014 e 2019 a 2023, respectivamente. Nota-se que as variáveis não apresentam uma relação linear com o O_3. Pode-se destacar que até mesmo as variáveis que apresentam altos valores do coeficiente de Correlação Linear de Pearson, como no caso da temperatura (TEMP) e da umidade relativa do ar (UR), a relação entre elas e o O_3 não é linear. Além disso, ao longo de cada conjunto de anos, há uma variabilidade "subjetiva"das relações entre o O_3 e as demais variáveis.

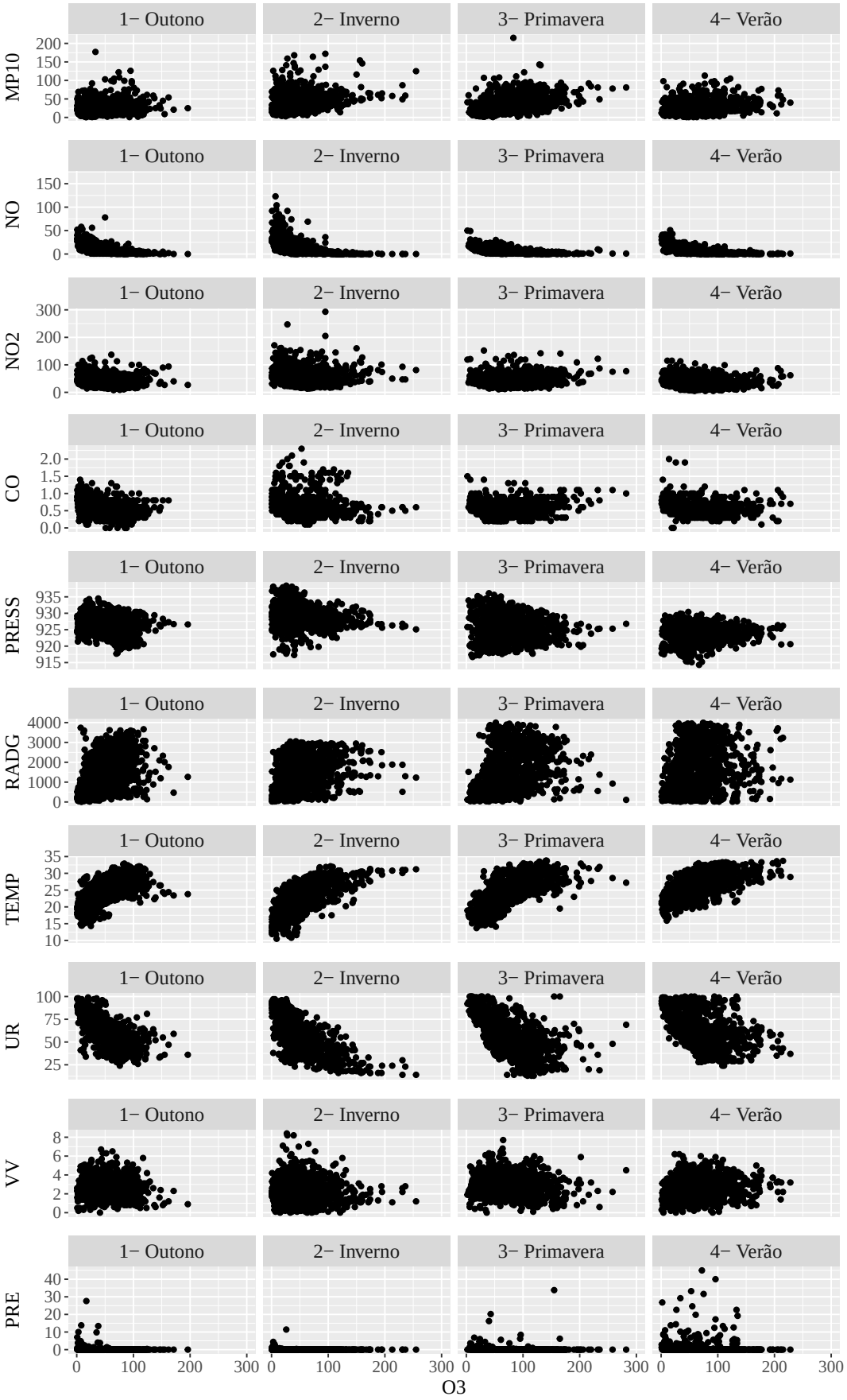

Figura 27 – Gráficos de dispersão do O_3 em função das demais variáveis. Os valores dos dados são das 12 às 17 horas por estação do ano calculados para o período de anos de 2010 a 2014.

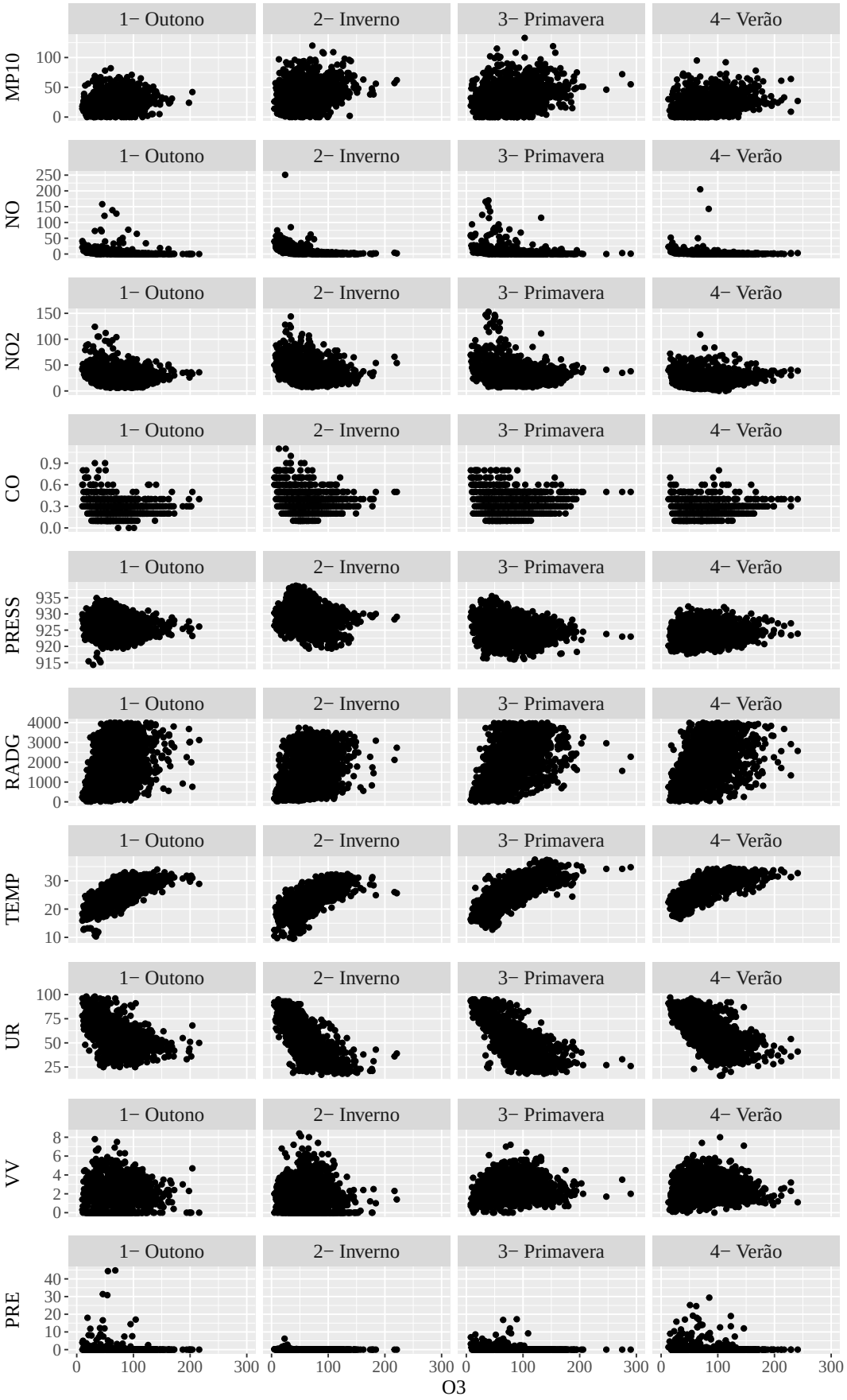

Figura 28 – Gráficos de dispersão do O_3 em função das demais variáveis. Os valores dos dados são das 12 às 17 horas por estação do ano calculados para o período de anos de 2019 a 2023.

3.5 CONCLUSÃO

A partir da análise exploratória dos dados, foi possível compreender algumas características sobre as séries de dados consideradas nesse trabalho. Pode-se observar que o O_3 varia ao longo do dia de forma cíclica e também ao longo das estações do ano. Além disso, características da série do O_3 como a periodicidade e a não-estacionariedade puderam ser evidenciadas. Uma análise sobre a avaliação da tendência da concentração de O_3 foi realizada em diferentes cenários. Os cenários de avaliação levam em consideração diferentes formas de estatísticas. Por exemplo, medidas de valores máximos, médios e mínimos observados nos dias para todos os dias do ano, a média dos valores máximos do dia conforme a estação do ano ao longos dos anos, a média de cada uma das horas do dia para cada estação do ano, ao longo dos anos. Essas análises, mostraram as diferentes formas de se observar a tendência de aumento das concentrações de O_3 de acordo com diferentes estatísticas.

Em uma primeira análise para avaliar os possíveis parâmetros correlacionados com o aumento das concentrações de O_3, avaliou-se as séries de dados da temperatura do ar e da radiação solar global. A partir dessas análises, verificou-se a não-estacionariedade das séries e evidenciou-se também a tendência de aumento da temperatura do ar e da radiação solar global. As maiores tendências de aumento também foram observadas nas estações do ano onde constatou-se os maiores aumentos das concentrações de O_3.

- ◇ As curvas de média móvel das diferentes medidas de médias de temperatura revelam tendência de aumento ao longo dos anos. Constata-se um aumento de 0,90°C para as máximas diárias, conforme observado na Figura 7. Além disso, a partir dos resultados apresentados na Tabela 6, é possível notar uma tendência de aumento da amplitude térmica de 0,6°C, quando não se considera a sazonalidade. Avaliando a evolução da amplitude térmica ao longo das estações do ano, verifica-se uma tendência de aumento para todas as estações. Os maiores aumentos são observados nas estações do outono (aumento de 1,0°C) e primavera (aumento de 0,7°C).

- ◇ A tendência de aumento da temperatura foi melhor observada considerando a média dos seus valores máximos diários para cada estação do ano. Nota-se uma tendência de aumento de 1,4°C no outono, 1,0°C na primavera e 0,7°C no inverno, conforme observado na Figura 10. O verão é a única estação do ano que não apresenta uma tendência de aumento.

- ◇ Em relação as médias horárias do dia, também foi possível observar um aumento da temperatura. Especificamente às 14h, observa-se um aumento de 1,3°C na primavera, 1,1°C no outono e cerca de 0,9°C no inverno. Não foi possível notar uma tendência de aumento nos valores da temperatura para a estação do verão, conforme observado na Figura 12.

- ◇ A incidência de radiação solar global apresenta um aumento de 550 kJ/m^2 ao longo dos anos considerados, conforme observado na Figura 23.
- ◇ A tendência de aumento da radiação solar global também foi melhor observada considerando cada estação do ano. Nota-se uma tendência de aumento na média da incidência da radiação solar global de aproximadamente 900 kJ/m^2 no outono, 600 kJ/m^2 no inverno, 460 kJ/m^2 na primavera e 300 kJ/m^2 no verão, conforme observado na Figura 23.
- ◇ A radiação solar também apresenta diferentes tendências de aumento de acordo com as horas do dia. Especificamente às 12h, a média da incidência da radiação solar global aumentou 605 kJ/m^2 no outono, 486 kJ/m^2 no inverno, 137 kJ/m^2 na primavera e 44 kJ/m^2 no verão, conforme observado na Figura 22.
- ◇ Nota-se uma tendência de aumento de 14 μg/m^3 nas concentrações de O_3 ao longo dos anos, conforme observado na Figura 14.
- ◇ As análises evidenciam variações estatísticas distintas dos dados conforme as estações do ano, conforme apresentado na Figura 16. As maiores concentrações de ozônio são registradas na primavera e no verão, estações com as maiores temperaturas do ano. Nessas estações, é comum a ocorrência de altas concentrações de O_3, com valores acima de 100 μg/m^3.
- ◇ Em relação aos períodos sazonais do ano, observou-se uma tendência de aumento da concentração de O_3 de 21 μg/m^3 no outono, 16 μg/m^3 no inverno, 10 μg/m^3 na primavera e 6 μg/m^3 no verão, conforme observado na Figura 17.
- ◇ A tendência de aumento do O_3 também varia conforme as horas do dia. Especificamente às 14h, observou-se um aumento de 16 μg/m^3 no outono, 15 μg/m^3 no inverno e cerca de 13 μg/m^3 na primavera e verão, conforme observado na Figura 19.

A partir da análise das demais variáveis considerando os dados medidos apenas nos horários das 12 às 17 horas, pode-se notar diferentes comportamentos. Identificou-se uma diminuição na concentração média dos poluentes CO, MP_{10}, NO e NO_2 para todas as estações do ano. Possivelmente associado ao aumento da radiação solar global e consequentemente da temperatura no ar, verificou-se uma queda da umidade relativa do ar nas estações do outono e inverno. Apesar disso, tal comportamento não é notado nas estações da primavera e verão. A média da velocidade do vento também apresenta tendência de diminuição ao longo do tempo, apesar dessa tendência de diminuição ser maior notada nas estações do outono e inverno. Além disso, verificou-se que a pressão atmosféricas e a precipitação não apresentam tendência de aumento, exceto a pressão que apresenta uma pequena tendência de aumento no inverno.

A partir dos valores do coeficiente de Correlação Linear de Pearson, pode-se notar que as variáveis que apresentam maior correlação positiva com o O_3 são a temperatura e a radiação solar global, ambas apresentando tendência de aumento ao longo do tempo. Já as variáveis que apresentam maior correlação negativa com o O_3 são a umidade relativa do ar e o NO. No outono e no inverno, estações com maior tendência de aumento do ozônio, observa-se uma tendência de diminuição da umidade relativa do ar, tornando o tempo mais quente e seco nessas estações. Já o NO apresenta tendência de diminuição ao longo do tempo.

Para se obter uma compreensão maior acerca de como as variáveis estão associadas com a tendência de aumento da concentração de O_3, nas análises subsequentes são considerados modelos para a predição da concentração de ozônio a curto prazo. Os modelos são avaliados cenários que evidenciam a maior tendência de aumento de O_3.

4 A ESTIMAÇÃO DA CONCENTRAÇÃO DE O_3 A CURTO PRAZO

Neste capítulo, o objetivo inicialmente é avaliar modelos para estimar a concentração de O_3. Considera-se como entrada dos modelos as variáveis climáticas e de poluição do ar. Os parâmetros dos modelos são obtidos em conjuntos distintos de anos. Comparam-se os conjuntos de parâmetros dos modelos em cada conjunto de anos para avaliar a significância das variáveis na tendência de aumento da concentração de O_3. Cabe destacar que o objetivo não é obter um modelo para prever a concentração do O_3 a longo prazo, mas sim, obter um modelo preciso o suficiente que possibilite verificar a influência das variáveis de entrada na concentração O_3 em conjuntos distintos de anos. Assim, a estimativa da concentração do O_3 é feita a curto prazo, ou seja, faz-se a estimativa da concentração do O_3 na hora t com todas as variáveis de entrada do modelo na hora $t-1$. Os resultados observados no Capítulo 3 sustentam a validade das conclusões obtidas nesse capítulo.

O presente capítulo é estruturado da seguinte forma: na Seção 4.1 é apresentado o cenário das simulações. Na Seção 4.2 é apresentado brevemente o método dos mínimos quadrados ordinários e os resultados computacionais relativos à sua aplicação ao conjunto de dados. Na Seção 4.3 são apresentados os modelos baseados em árvores de decisão e também os resultados computacionais relativos à sua aplicação ao conjunto de dados. Os dados considerados nas simulações computacionais estão agrupados em cada estação do ano separadamente para os anos de 2010 a 2014 e para os anos de 2019 a 2023. Compara-se a importância relativa das variáveis em cada modelo para cada conjunto de 5 anos.

4.1 O CENÁRIO DAS SIMULAÇÕES

A Figura 29 representa a formulação de um problema de regressão. Sejam os valores de uma variável resposta desejada representada pelo vetor $\mathbf{y}$ e uma matriz de dados $\mathcal{X}_0$. A matriz $\mathcal{X}_0$ aplicada à entrada do *Sistema* representa o conhecimento implícito sobre todos os dados necessários para se obter a resposta desejada $\mathbf{y}$. Porém, na maioria das vezes, não se sabe qual é o melhor *modelo* que representa o sistema para se obter a resposta desejada, não se possui todos as variáveis que resultam na resposta desejada ou não se possuem informações suficientes sobre as variáveis. Na prática à entrada do modelo é aplicada parte das informações conhecidas da matriz $\mathcal{X}_0$, representada por $\mathcal{X}$. O modelo é utilizado para estimar a resposta desejada. Considera-se, então, um sinal de erro, definido como a diferença entre a resposta desejada e a resposta estimada observada na saída do modelo. Uma função do sinal do erro é utilizada para aprimorar a estimativa da resposta desejada e avaliar o desempenho preditivo do modelo segundo algum critério de otimização.

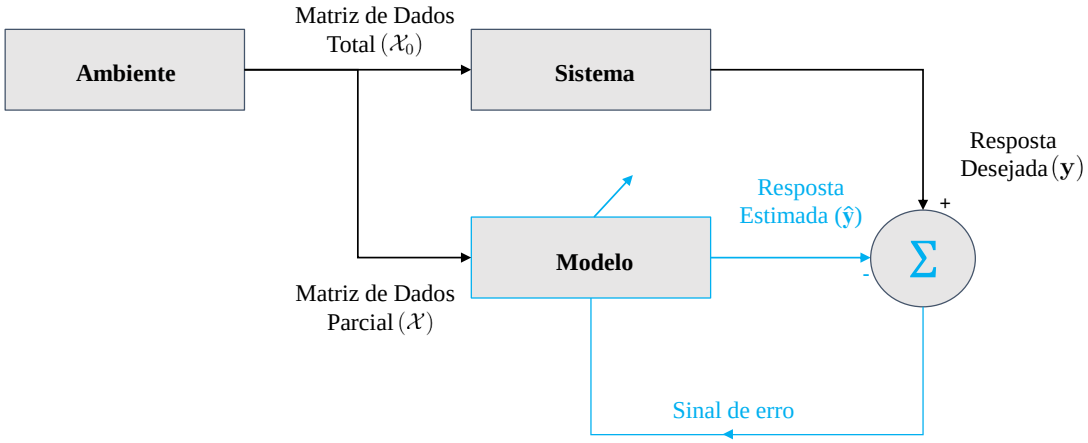

Figura 29 – Diagrama de blocos simplificado para o procedimento de avaliação computacional dos modelos simulados.

4.1.1 O VETOR DE RESPOSTA DESEJADA E A MATRIZ DE DADOS

Para descrever os dados usados nas simulações computacionais considera-se o vetor denotado como

$$\mathbf{x}_\ell(k,s) = \begin{bmatrix} x_\ell(11,k,s) \\ x_\ell(12,k,s) \\ x_\ell(13,k,s) \\ x_\ell(14,k,s) \\ x_\ell(15,k,s) \\ x_\ell(16,k,s) \end{bmatrix} \tag{4.1}$$

Esse vetor de dimensão 6×1, possui o valor da variável ℓ, no dia k, em um determinado ano, na estação s e nos horários das 11h às 16h. Os índices ℓ, s, k assumem os seguintes valores:

- $\ell = 1,2,\ldots 11$ indicam as variáveis descritas na Tabela 1, ou seja, O_3, MP_{10}, NO, NO_2, CO, RADG, TEMP, UR, VV), PRESS e PRE, respectivamente.
- $s = 1$ (outono), 2 (inverno), 3 (primavera) ou 4 (verão) indicam as estações do ano.
- $k = 1, 2, \cdots, K$ em que K é o número de dias e depende do s considerado. No outono (março, abril, maio) e no inverno (junho, julho, agosto) $K = 92$ dias, primavera (setembro, outubro, novembro) $K = 91$ e no verão (dezembro, janeiro, fevereiro) $K = 90$. Por conveniência, se o ano considerado é bissexto o dia adicional é desprezado.

Os dados usados na entrada do modelo para cada ano são representados com quatro matrizes do tipo

$$\mathcal{X}_s = \begin{bmatrix} \mathbf{x}_1(1,s) & \mathbf{x}_2(1,s) & \mathbf{x}_3(1,s) & \cdots & \mathbf{x}_{11}(1,s) \\ \mathbf{x}_1(2,s) & \mathbf{x}_2(2,s) & \mathbf{x}_3(2,s) & \cdots & \mathbf{x}_{11}(2,s) \\ \mathbf{x}_1(3,s) & \mathbf{x}_2(3,s) & \mathbf{x}_3(3,s) & \cdots & \mathbf{x}_{11}(3,s) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_1(K-2,s) & \mathbf{x}_2(K-2,s) & \mathbf{x}_3(K-2,s) & \cdots & \mathbf{x}_{11}(K-2,s) \\ \mathbf{x}_1(K-1,s) & \mathbf{x}_2(K-1,s) & \mathbf{x}_3(K-1,s) & \cdots & \mathbf{x}_{11}(K-1,s) \\ \mathbf{x}_1(K,s) & \mathbf{x}_2(K,s) & \mathbf{x}_3(K,s) & \cdots & \mathbf{x}_{11}(K,s) \end{bmatrix}, \tag{4.2}$$

A dimensão da matriz $\mathcal{X}_s$ é $N \times p$. Note que $N = 6\,K$ e vai depender da estação meteorológica considerada. O número de colunas é $p = 11$ e representa o número de variáveis consideradas no modelo.

A variável resposta é a concentração de ozônio (O_3) e é definida a cada ano, para todos os dias da estação meteorológica s, nos horários entre 12h às 17h, conforme descrito pelo vetor

$$\mathbf{y}_s = \begin{bmatrix} \bar{\mathbf{x}}_1(1,s) \\ \bar{\mathbf{x}}_1(2,s) \\ \vdots \\ \bar{\mathbf{x}}_1(K-1,s) \\ \bar{\mathbf{x}}_1(K,s) \end{bmatrix} \tag{4.3}$$

sendo

$$\bar{\mathbf{x}}_1(k,s) = \begin{bmatrix} x_1(12,k,s) \\ x_1(13,k,s) \\ x_1(14,k,s) \\ x_1(15,k,s) \\ x_1(16,k,s) \\ x_1(17,k,s) \end{bmatrix}, \tag{4.4}$$

para $k = 1,2,\ldots K$. A dimensão do vetor $\mathbf{y}_s$ é $N = 6\,K$.

4.1.2 MÉTRICAS DE AVALIAÇÃO

O primeiro passo para construir bons modelos é criar um critério para medir o desempenho da predição da variável resposta $\mathbf{y}$. Em um contexto de regressão, é usual considerar como função custo o erro quadrático médio,

$$\texttt{EQM} = \mathbb{E}[(\mathbf{y} - \hat{\mathbf{y}})^2], \tag{4.5}$$

em que $\mathbb{E}[.]$ denota a operação do valor esperado e $\hat{\mathbf{y}}$ é a estimativa da variável resposta obtida a partir de algum modelo. Essa estimativa pode ser feita de forma linear ou

não-linear. O critério é obter a estimativa da variável resposta de modo a minimizar o EQM.

Na prática, tem-se disponível medidas coletadas ao longo do tempo. As medidas de erros constituem operações fundamentais para poder quantificar a diferença entre os valores reais observados $(y_1, y_2, \cdots, y_N)$ e os valores preditos pelo modelo $(\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_N)$, em que N é o número de observações. Os principais estimadores utilizados são (IZBICKI; SANTOS, 2020), (ROSS, 2021):

a) **Erro Quadrático Médio (*Mean Square Error* - MSE)**. No caso, a média do erro quadrático é feita da seguinte forma

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2. \tag{4.6}$$

Apesar de amplamente utilizada, essa métrica é sensível na presença de *outliers*. Uma característica do MSE é que os erros da predição são elevados ao quadrado antes de ter a média calculada. Dessa forma, se houver um *outlier* no conjunto de dados, seu peso será maior para o cálculo do MSE e, consequentemente, afetará sua métrica deixando-a maior.

b) **Erro Médio Absoluto (*Mean Absolute Error* - MAE)** é uma medida representada pela média dos erros das diferenças absolutas entre o valor real e o valor predito, em que todas as diferenças individuais apresentam o mesmo peso, ou seja,

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|. \tag{4.7}$$

Por ser menos afetada por pontos *outliers*, é considerada uma métrica precisa e robusta para medir a habilidade de um modelo em reproduzir a realidade.

c) **Coeficiente de Determinação (R^2)** é uma medida que fornece o percentual da variabilidade dos dados que é reproduzida pelos modelos. Essa medida é definida como

$$R^2 = 1 - \frac{1}{\mathcal{V}}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \tag{4.8}$$

em que

$$\mathcal{V} = \sum_{i=1}^{N}(y_i - \bar{y})^2 \quad \text{e} \quad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i.$$

Hipoteticamente, se o modelo é perfeito o `EQM` $= 0$. Assim, em condições práticas, considera-se que quando menor forem os valores fornecidos pelas Equações 4.6 e 4.7, melhor será o modelo. Já, o coeficiente de determinação varia entre $0 \leq R^2 \leq 1$, portanto, quanto mais próximo de um (zero) for o seu valor, melhor (pior) é a qualidade do modelo.

4.1.3 LINGUAGENS COMPUTACIONAIS E PACOTES

As simulações computacionais foram desenvolvidas na linguagem de programação estatística e gráfica conhecida como R. Além de ter o código aberto, o R possui facilidades na manipulação, análise e visualização de dados e uma biblioteca com amplo conjunto de pacotes. Especificamente nas avaliações realizadas nesse trabalho considerou-se os pacotes conforme especificados na Tabela 11. O OLS, acrônimo de *Ordinary Least Squares*, é um método usado para estimar os parâmetros de um modelo de regressão linear minimizando a soma dos resíduos quadrados. O LASSO, acrônimo de *Least Absolute Shrinkage and Selection Operator*, é uma extensão do OLS que adiciona um termo de penalidade para estimar os parâmetros do modelo de regressão linear. O CART e o RF, acrônimos de *Classification and Regression Tree* e de *Random Forest*, são algoritmos de aprendizado de máquina comumente utilizados em problemas de regressão. As simulações computacionais foram implementadas em um processador Samsung Intel(R) Core(TM) i5-10210U CPU @ 1.60GHz 2.11 GHz, com memória ram de 24GB, ssd 256GB e placa gráfica MX110.

Tabela 11 – Lista de pacotes, originários do R, utilizados neste trabalho.

Método / Ferramenta	Pacote	Referência
OLS	{caret}	(KUHN, 2008)
LASSO	{glmnet}	(FRIEDMAN; TIBSHIRANI; HASTIE, 2010)
CART	{party}	(ZEILEIS; HOTHORN; HORNIK, 2008)
RF	{caret}	(KUHN, 2008)
Cálculo de importância de variáveis para os modelos desenvolvidos	{vip}	(GREENWELL; BOEHMKE, 2020)
Construção gráfica	{ggplot2}	(WICKHAM, 2016)

4.2 MÉTODO DOS MÍNIMOS QUADRADOS ORDINÁRIOS (OLS)

Seja o diagrama de blocos da Figura 29. Considerando o modelo linear, a saída do modelo pode ser representada de forma matricial como

$$\mathbf{y}_s = \mathbf{X}_s\boldsymbol{\theta} + \boldsymbol{\epsilon}_s \tag{4.9}$$

- No caso, a matriz de dados $\mathbf{X}_s$ aplicada a entrada do modelo é definida como

$$\mathbf{X}_s = \begin{bmatrix} \mathbf{1} & \mathcal{X}_s \end{bmatrix} \tag{4.10}$$

em que $\mathcal{X}_s$ é a matriz conforme definida na Equação 4.2 e $\mathbf{1}$ é um vetor coluna com N uns. Nota-se que todas as colunas de $\mathcal{X}_s$ contém as variáveis entre 11 e 16 horas. A matriz $\mathbf{X}_s$ possui dimensão $N \times (p+1)$, sendo $N = 6\,K$, K o número de dias considerados e $p + 1 = 12$.

- O vetor $\mathbf{y}$ conforme definido na Equação 4.3, é a variável resposta, ou seja, o O_3 entre 12 e 17 horas, com dimensão $N \times 1$.
- O vetor de resíduos $\boldsymbol{\epsilon}$ representa o erro entre a resposta desejada e a estimada, dimensão $N \times 1$.
- O vetor $\boldsymbol{\theta}$ contém os parâmetros que combinados linearmente com as colunas da matriz de dados fornecem a estimativa da resposta desejada, dimensão $(p+1) \times 1$;

No método OLS, o objetivo é estimar o vetor de parâmetros $\hat{\boldsymbol{\theta}}$ que minimiza a norma euclidiana, do vetor de resíduos $\boldsymbol{\epsilon}$ ou seja,

$$\hat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}} \|(\mathbf{y} - \mathbf{X}_s\boldsymbol{\theta})\|^2. \tag{4.11}$$

Para facilitar, define-se a função $f(\boldsymbol{\theta})$ como

$$f(\boldsymbol{\theta}) \triangleq (\mathbf{y} - \mathbf{X}_s\boldsymbol{\theta})^t(\mathbf{y} - \mathbf{X}_s\boldsymbol{\theta}), \tag{4.12}$$

em que $(^t)$ denota o transposto de uma matriz. Note que essa equação quando se considera a divisão por N é um caso particular da Equação 4.6. Em outras palavras, $f(\boldsymbol{\theta}) = N \times MSE$, no caso em que $\hat{y}_i$ é a estimativa da resposta desejada com o método OLS. Desenvolvendo a Equação 4.12, resulta

$$f(\boldsymbol{\theta}) = \mathbf{y}^t\mathbf{y} - \mathbf{y}^t\mathbf{X}_s\boldsymbol{\theta} - \boldsymbol{\theta}^t\mathbf{X}_s^t\mathbf{y} + \boldsymbol{\theta}^t\mathbf{X}_s^t\mathbf{X}_s\boldsymbol{\theta}. \tag{4.13}$$

Note que $(\mathbf{y}^t\mathbf{X}_s\boldsymbol{\theta}) = (\boldsymbol{\theta}^t\mathbf{X}_s^t\mathbf{y})^t$. Como o produto matricial é um número real, o seu transposto é igual a ele próprio, dessa forma , pode-se escrever $\mathbf{y}^t\mathbf{X}_s\boldsymbol{\theta} = \boldsymbol{\theta}^t\mathbf{X}_s^t\mathbf{y}$. Portanto, a Equação 4.13 pode ser escrita por

$$f(\boldsymbol{\theta}) = \mathbf{y}^t\mathbf{y} - 2\boldsymbol{\theta}^t\mathbf{X}_s^t\mathbf{y} + \boldsymbol{\theta}^t\mathbf{X}_s^t\mathbf{X}_s\boldsymbol{\theta}. \tag{4.14}$$

Para calcular $\boldsymbol{\theta}$ que minimiza a expressão, pode-se calcular o gradiente e igualar a zero, isto é,

$$\begin{aligned}
\nabla f(\boldsymbol{\theta}) &= 0 \\
\nabla f(\boldsymbol{\theta}) &= -2\mathbf{X}_s^t\mathbf{y} + \mathbf{X}_s^t\mathbf{X}_s\boldsymbol{\theta} + \boldsymbol{\theta}^t\mathbf{X}_s^t\mathbf{X}_s \\
\nabla f(\boldsymbol{\theta}) &= -2\mathbf{X}_s^t\mathbf{y} + 2\mathbf{X}_s^t\mathbf{X}_s\boldsymbol{\theta} \\
\nabla f(\boldsymbol{\theta}) &= 0 \Rightarrow \mathbf{X}_s^t\mathbf{X}_s\boldsymbol{\theta} = \mathbf{X}_s^t\mathbf{y}.
\end{aligned} \tag{4.15}$$

Quando $\mathbf{X}_s^t\mathbf{X}_s$ admite inversa, o estimador $\hat{\boldsymbol{\theta}}$ é obtido como

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}_s^t\mathbf{X}_s)^{-1}\mathbf{X}_s^t\mathbf{y}. \tag{4.16}$$

A partir do vetor de parâmetros estimado, a variável resposta pode ser calculada a partir de um conjunto de dados amostrais segundo a expressão

$$\hat{\mathbf{y}} = \mathbf{X}_s\hat{\boldsymbol{\theta}}. \tag{4.17}$$

TESTES PARA A SIGNIFICÂNCIA E A ESCOLHA DOS COEFICIENTES DO MODELO

As hipóteses para testar a significância de qualquer coeficiente individual do modelo, como $\hat{\theta}_j$, são (MONTGOMERY; RUNGER, 2021)

$$H_0 : \hat{\theta}_j = 0 \text{ ou } H_1 : \hat{\theta}_j \neq 0.$$

A estatística utilizada para a avaliação do teste de hipóteses é definida por

$$t^* = \frac{\hat{\theta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \sim t_{n-2}, \tag{4.18}$$

em que representa uma distribuição t de student com $n-2$ graus de liberdade, C_{jj} é o elemento da diagonal de $[\mathbf{X}^t\mathbf{X}]^{-1}$ e a estimativa de $\hat{\sigma}^2$ é dada por

$$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \mathbf{X}_s\hat{\boldsymbol{\theta}})^t(\mathbf{y} - \mathbf{X}_s\hat{\boldsymbol{\theta}}). \tag{4.19}$$

A hipótese H_0 será rejeitada, a um nível de significância α, se $|t^*| \geq t_{n-2;\alpha/2}$ tabulado da distribuição t de student. É prática comum considerar como nível de significância um dos valores de 1%, 5% ou 10%.

SELEÇÃO DE VARIÁVEIS E REGULARIZAÇÃO EM REGRESSÃO

O método de seleção e regularização de variáveis conhecido como LASSO (do inglês, *Least Absolute Shrinkage and Selection Operator*) é usado em estatística e aprendizado de máquina, para aumentar a precisão da predição e a interpretabilidade de modelos estatísticos. Essa técnica é usualmente aplicada para selecionar as variáveis de um modelo de regressão. Matematicamente, consiste em um modelo linear com um termo de regularização adicionado. A função custo a ser minimizada pode ser escrita na forma Lagrangiana (HASTIE; TIBSHIRANI; FRIEDMAN, 2009) como

$$\boldsymbol{\theta}^{LASSO} := \arg\min_{\boldsymbol{\theta}} \left((\mathbf{y} - \mathbf{X}_s\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}_s\boldsymbol{\theta}) + \lambda \sum_{i=1}^{p} |\theta_i| \right). \tag{4.20}$$

A regularização pelo LASSO produz um modelo esparso, isto é, um modelo que tem a maior parte de seus parâmetros igual a zero. A constante $\lambda \geq 0$ controla a proporção de parâmetros zerados. À medida que λ cresce, mais parâmetros do modelo são zerados. A determinação do valor de λ é geralmente feita através de validação cruzada (HASTIE; TIBSHIRANI; WAINWRIGHT, 2015).

Originalmente o LASSO foi usado para modelos que usam o estimador de mínimos quadrados ordinários, porém, o termo de penalização pode ser estendido a outros modelos estatísticos, como modelos lineares generalizados e modelos de risco proporcional. Uma abordagem detalhada do LASSO e como obter o valor de λ através do método de validação cruzada podem ser encontrados em (HASTIE; TIBSHIRANI; WAINWRIGHT, 2015).

4.2.1 RESULTADOS DAS SIMULAÇÕES BASEADAS NO MÉTODO OLS

Considerou-se dois conjuntos de dados, um para os anos de 2010 a 2014 e outro para os anos de 2019 a 2023. Em cada um desses conjuntos foram considerados quatros subconjuntos relativos ao outono, inverno, primavera e verão. Assim no total foram calculados 4 vetores de coeficientes, para cada conjunto de anos. Em uma primeira etapa o método LASSO foi utilizado para seleção de variáveis. A partir da aplicação do LASSO, buscou-se o valor de λ que fornece o modelo mais regularizado, de modo que o erro de validação cruzada (utilizando *k-fold* = 10) esteja dentro de um erro padrão mínimo. Com base no valor de λ, selecionou-se as variáveis para serem descartadas (com coeficientes nulos). Esses valores foram obtidos através da utilização do pacote `GLMNET` do Software *R*. Esses resultados estão resumidos na Tabela 12. Nota-se que dependendo da janela de anos e da estação, algumas variáveis foram descartadas e outras não.

Tabela 12 – Seleção de variáveis por meio da aplicação do método LASSO.

Estação	Anos	Lambda	Coeficientes zerados
Outono	2010 a 2014	0,0741605	TEMP, MP_{10}
	2019 a 2023	0,0794949	NO, NO_2
Inverno	2010 a 2014	0,0606558	PRESS, PRE
	2019 a 2023	0,0366557	-
Primavera	2010 a 2014	0,2526542	NO, PRESS
	2019 a 2023	0,0661460	-
Verão	2010 a 2014	0,0672888	-
	2019 a 2023	0,1023526	TEMP

Em seguida, calculou-se o desempenho do modelo LASSO e um comparativo com o modelo completo contendo todas as variáveis por meio do método dos OLS. O desempenho do modelo foi avaliado a partir da utilização das medidas MSE, MAE e R^2. A Tabela 13 resume os resultados dos modelos ajustados a cada subconjunto de dados.

Tabela 13 – Avaliação dos modelos LASSO e OLS.

Estação	Modelo	Dados dos anos (2010 a 2014)			Dados dos anos (2019 a 2023)		
		MSE	MAE	R^2(%)	MSE	MAE	R^2(%)
Outono	OLS	137,36	8,55	87%	112,42	7,62	86%
	LASSO	137,48	8,55	87%	112,49	7,62	86%
Inverno	OLS	139,55	8,06	86%	85,90	6,72	89%
	LASSO	139,67	8,07	86%	85,94	6,72	89%
Primavera	OLS	276,50	11,77	85%	219,25	10,91	87%
	LASSO	277,21	11,78	85%	219,34	10,92	87%
Verão	OLS	215,73	10,47	86%	193,10	10,02	86%
	LASSO	215,79	10,46	86%	193,26	10,01	86%

A partir da observação dos valores da Tabela 13 destacam-se:

- Quando se compara os valores das métricas de desempenho, obtidas com o OLS contendo todas as variáveis e com o LASSO excluindo as variáveis com coeficientes descartados, nota-se um pequeno aumento nos valores do MSE e do MAE no caso do LASSO. Porém, os valores do R^2 são os mesmos para ambos.

- Os menores valores de MSE e MAE são observados no inverno e os maiores são observados na primavera e no verão. Nota-se que esses valores variam conforme a janela de anos. Observa-se uma aparente tendência de diminuição do MSE e do MAE quando se compara os valores da janela de anos de 2019 a 2023 com os anos de 2010 a 2014.

- Observa-se que os valores dos coeficiente de determinação R^2, acima de $\approx 85\%$, revelam que tanto o OLS como o LASSO explicam a maior parte da variabilidade dos dados em todos os cenários. Porém, os melhores resultados são observados para o inverno nos anos de 2019 a 2023, com $R^2 \approx 89\%$.

Na Tabela 14 estão resumidos os valores dos coeficientes do modelo OLS, a estatística t e o $p-valor$ do teste t para significância. A estatística t é utilizada para medir o grau de importância das variáveis para o modelo conforme definido na equação (4.18). Foi considerado o valor $\alpha = 5\%$ na aplicação dos testes de hipóteses. A partir da observação dos valores da Tabela 14 nota-se que a importância das variáveis para o modelo varia de acordo com a estação do ano e o período analisado. Em relação à significância das variáveis cabem os seguintes destaques:

- A concentração de O_3 apresenta os maiores valores da estatística t em valor absoluto para todas as estações do ano para ambos os conjuntos de anos. No outono, inverno e primavera, entre as janelas de dados de 2010 a 2014 e 2019 a 2023 há um aumento da importância da concentração do O_3 da hora (t-1) na estimativa da concentração da hora (t).

- A estatística t da RADG apresenta os maiores valores após o O_3. Entretanto, nota-se que dos anos de 2010 a 2014 para os anos 2019 a 2023, a importância da RADG na estimativa da concentração da hora (t) diminui no outono e na primavera.

- A estatística t da VV, após o O_3 e a RADG, apresenta os maiores valores. Porém, essa posição de importância da VV acontece apenas na primavera e no verão. No outono a importância da VV nos anos de 2010 a 2014 dá lugar para o CO nos anos de 2019 a 2023. No inverno, a estatística t revela a importância do NO_2 nos anos de 2010 a 2014 e do NO nos anos de 2019 a 2023.

- Pode-se notar que algumas variáveis não são significativas para o modelo, porém a relação dessas variáveis é função da estação do ano e das janelas de dados (2010 a

2014 e 2019 a 2023). Especificamente o NO e o NO_2 são não significativas no verão. Além disso, no verão a PRE e a TEMP também são não significativas no verão. Já no inverno a PRE e a PRESS são não significativas.

Tabela 14 – Resumo dos valores dos coeficientes do modelo OLS, a estatística t e o $p-valor$ do teste t.

Estação	Variáveis	Dados dos anos (2010 a 2014)			Dados dos anos (2019 a 2023)		
		Coeficiente	Valor |t|	P-valor (teste t)	Coeficiente	Valor |t|	P-valor (teste t)
Outono	CO	5,397	2,788	<0,05	18,600	5,032	<0,05
	MP_{10}	-0,001	0,053	**0,958**	-0,037	1,699	**0,090**
	NO	-0,091	2,350	<0,05	0,010	0,269	**0,788**
	NO_2	0,061	3,389	<0,05	-0,006	0,164	**0,870**
	O_3	0,789	53,289	<0,05	0,818	61,898	<0,05
	PRE	-0,684	2,376	<0,05	-0,621	4,952	**<0,05**
	PRESS	0,117	0,975	**0,330**	0,291	2,830	<0,05
	RADG	0,011	30,145	<0,05	0,008	26,314	<0,05
	TEMP	-0,005	0,038	**0,970**	0,146	1,440	**0,150**
	UR	0,093	3,405	<0,05	0,024	0,962	**0,336**
	VV	-2,647	10,720	<0,05	-0,799	4,084	<0,05
Inverno	CO	1,002	0,608	**0,543**	11,770	5,050	<0,05
	MP_{10}	-0,026	1,544	**0,123**	0,046	3,196	<0,05
	NO	-0,101	4,057	<0,05	-0,211	7,364	<0,05
	NO_2	0,103	5,870	**<0,05**	0,082	3,851	<0,05
	O_3	0,800	52,444	<0,05	0,820	65,797	<0,05
	PRE	-0,050	0,057	**0,955**	0,873	1,859	**0,063**
	PRESS	-0,007	0,082	**0,934**	0,096	1,365	**0,172**
	RADG	0,011	26,367	<0,05	0,007	27,186	<0,05
	TEMP	0,349	3,005	<0,05	0,335	3,920	<0,05
	UR	0,117	4,320	<0,05	0,075	3,404	<0,05
	VV	-1,017	4,300	<0,05	-0,367	2,468	<0,05
Primavera	CO	4,810	1,682	**0,093**	22,540	5,751	<0,05
	MP_{10}	-0,051	2,457	<0,05	-0,098	4,705	<0,05
	NO	-0,012	0,163	**0,871**	-0,008	0,179	**0,858**
	NO_2	0,074	2,329	<0,05	-0,021	0,545	**0,586**
	O_3	0,744	45,642	<0,05	0,717	46,506	<0,05
	PRE	-0,491	1,890	**0,059**	-0,597	2,300	<0,05
	PRESS	0,163	1,235	**0,217**	0,390	3,218	<0,05
	RADG	0,010	21,508	<0,05	0,008	18,797	<0,05
	TEMP	0,530	2,974	<0,05	0,428	2,525	<0,05
	UR	-0,063	1,590	**0,112**	-0,147	3,568	<0,05
	VV	-2,953	10,012	<0,05	-3,303	10,004	<0,05
Verão	CO	11,770	3,167	<0,05	20,900	3,697	<0,05
	MP_{10}	-0,124	4,379	<0,05	-0,071	2,546	<0,05
	NO	0,049	0,685	**0,493**	-0,085	1,001	**0,317**
	NO_2	0,102	3,021	**<0,05**	0,057	1,199	**0,231**
	O_3	0,812	51,933	<0,05	0,758	46,780	<0,05
	PRE	-0,325	2,128	**<0,05**	-0,019	0,094	**0,925**
	PRESS	0,317	2,260	<0,05	0,695	4,953	<0,05
	RADG	0,007	14,266	<0,05	0,007	14,656	<0,05
	TEMP	0,398	1,949	**0,051**	-0,161	0,806	**0,420**
	UR	-0,220	4,977	<0,05	-0,234	4,523	<0,05
	VV	-2,516	7,765	<0,05	-3,850	10,537	<0,05

A partir dos resultados das simulações para estimar a concentração de O_3 da hora (t) com o modelo OLS, ficou evidente que a variável mais significativa é a concentração de

O_3 da hora (t-1). A fim de se ter uma noção qualitativa desse resultado, considerou-se um modelo com apenas a concentração do O_3 da hora (t-1) como variável preditora. Além disso, foram considerados outros cenários. Em um desses cenários, o O_3 foi excluído como variável preditora. Nesse caso, obteve-se o pior MSE em comparação com os casos em que cada uma das outras variáveis preditoras foi excluída. Verificou-se com exaustivas simulações que a inclusão gradativa das demais variáveis, mesmo as menos significativas, não pioram a capacidade preditiva do modelo. Para ilustrar esses resultados, a dispersão entre os valores preditos e os observados são apresentados nos gráficos da Figura 30. Nos gráficos de dispersão da Figura 30 (a), são exibidos os valores preditos na hora (t) apenas com os valores da concentração do O_3 da hora (t-1). Nos gráficos de dispersão da Figura 30 (b), são exibidos os valores preditos na hora (t) com todas as 11 as variáveis preditoras. Esses resultados revelam que apesar das demais variáveis apresentarem uma significância relativamente menor para o modelo, elas contribuem positivamente para a diminuição do MSE. Assim, visando a obtenção de um modelo que explique as concentrações do O_3 da forma mais realista possível, considera-se todas as 11 variáveis nas simulações a seguir.

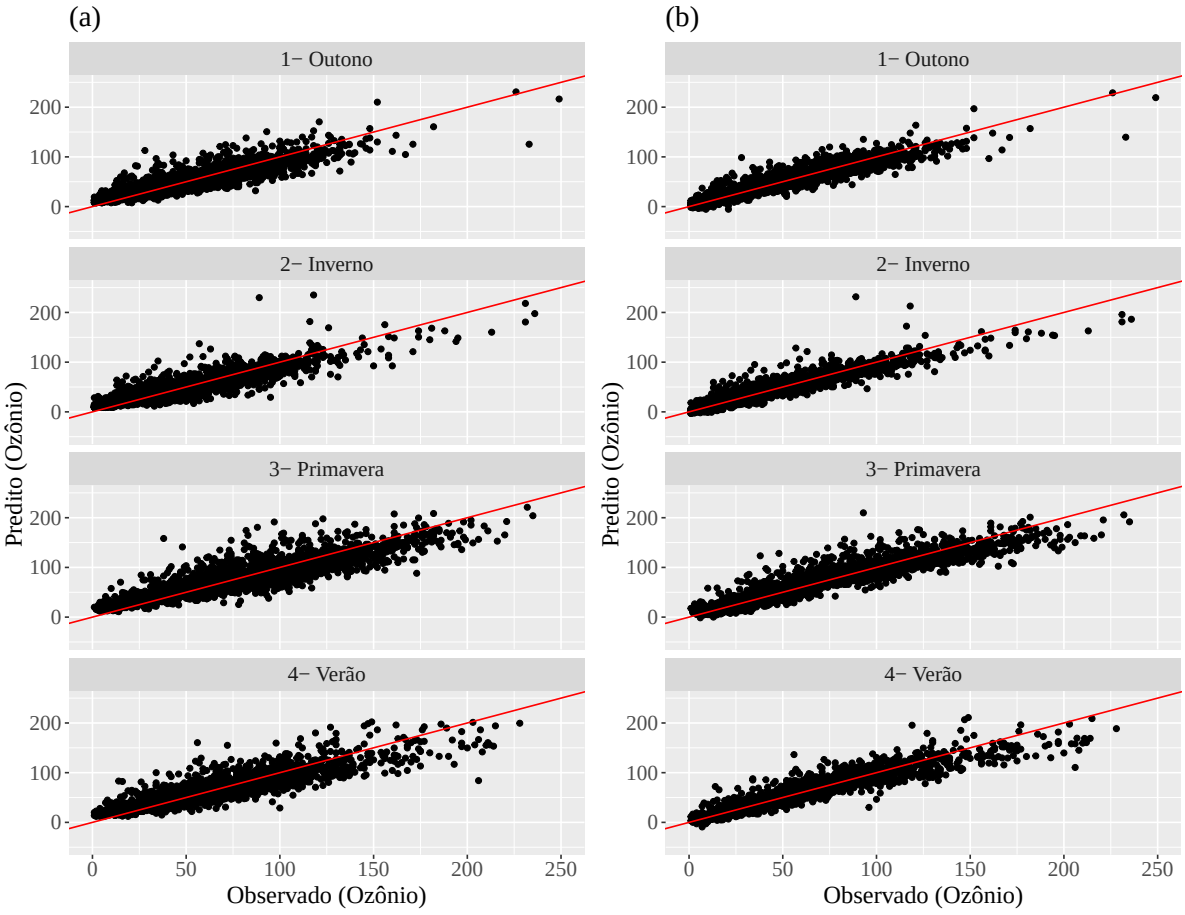

Figura 30 – Gráficos de dispersão dos valores preditos da concentração de O_3 com o modelo OLS em função dos valores observados: (a) valor predito apenas com o O_3 da hora (t-1) como variável preditora e (b) valores preditos com todas as variáveis preditoras da hora (t-1). Dados dos anos de 2010 a 2014. A reta indicada em vermelho é o caso ideal quando $x = y$.

A fim de avaliar o modelo OLS com todas as 11 variáveis preditoras ao longo do tempo considerou-se as janelas nos anos (2010 a 2014) e (2019 a 2023). Os gráficos de dispersão resultantes estão ilustrados na Figura 31.

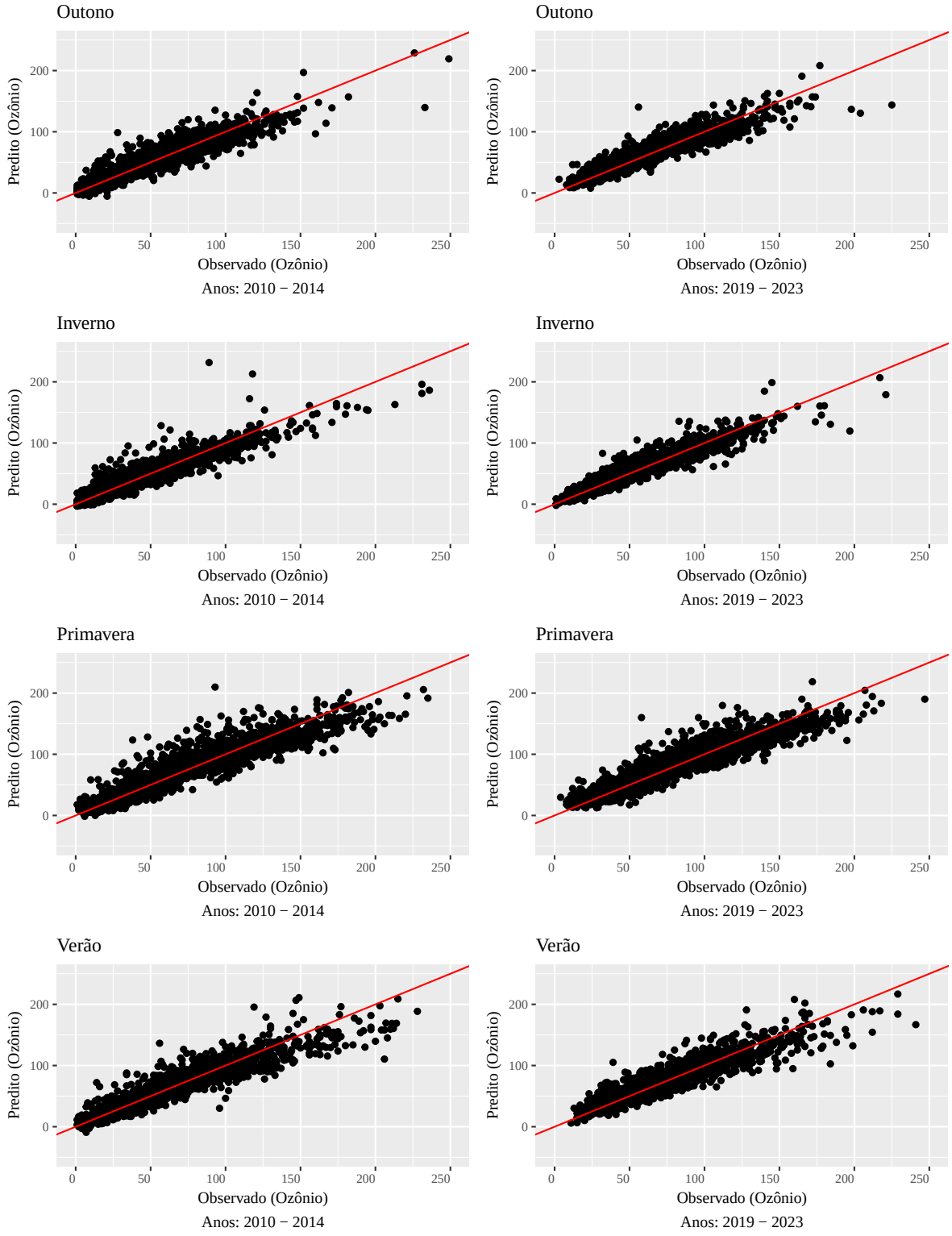

Figura 31 – Gráficos de dispersão dos valores do O_3 preditos com o modelo OLS em função dos valores observados. Cada gráfico representa uma estação anual e um conjunto de anos. A reta indicada em vermelho é o caso ideal quando $x = y$.

Os pontos acima da reta vermelha representam as observações superestimadas pelo modelo, isto é, aquelas cujo valor predito é maior que o valor observado. Os pontos abaixo da curva representam os valores subestimados, o que sugere um efeito mais significativo de não linearidades. Pode-se observar que o modelo subestima valores de concentração de O_3 acima de 100 $\mu g/m^3$, isto é notado principalmente na estação da primavera e verão. Nessas estações, é possível observar entre as janelas dos anos (2010 a 2014) e (2019 a 2023) uma variabilidade relativamente grande em torno da reta para valores acima de 100 $\mu g/m^3$. No outono, entre a primeira e a segunda janela de anos, nota-se um aumento na ocorrência dos pontos relativamente altos. Enquanto nas estações do inverno, primavera e verão nota-se uma menor ocorrência de aumento desses pontos. Esses valores estão associados com a diminuição dos valores de MSE observados na Tabela 13.

A fim de facilitar a visualização da importância das variáveis de forma quantitativa, fez-se gráficos de barras conforme ilustrados na Figura 32. Nos quatro gráficos de barras da primeira coluna estão indicados os resultados da Correlação de Pearson e nos quatro gráficos da segunda coluna estão os resultados obtidos com o método OLS. A altura das barras, na primeira coluna é o valor da correlação de Pearson. Na segunda coluna a altura das barras é o valor da estatística t (valor $|t|$ da Tabela 14). As barras indicadas em laranja e azul representam os valores nos conjuntos de anos de 2010 a 2014 e 2019 a 2023, respectivamente. Nota-se que, tanto a correlação de Pearson, como a importância das variáveis calculadas por meio do método OLS variam para cada conjunto de anos e também sofrem alterações conforme as estações do ano. Entretanto, apesar da correlação de Pearson e o método OLS suporem relações lineares, essas alterações são notadamente distintas. Enquanto para o modelo OLS a RADG é a variável com maior significância, depois do O_3 da hora (t-1), na correlação de Pearson a temperatura, a umidade relativa e o NO apresentam coeficientes de correlação lineares maiores que a RADG.

Os resultados obtidos por meio do OLS sugerem relações entre as variáveis para explicar as concentrações de ozônio. No entanto, como foi visto nos resultados apresentados na Tabela 13 e na Figura 31 o desempenho do modelo não é satisfatório conforme revelado pelos valores das estatísticas de erros MSE e MAE. Em decorrência disso, são avaliadas outras classes de modelos, que consigam explicar as concentrações de ozônio da forma mais realista possível para poder representar de fato a importância das variáveis na representação do aumento das concentrações de ozônio. Para isso, a seguir são avaliados os modelos não lineares baseados em árvores de decisão.

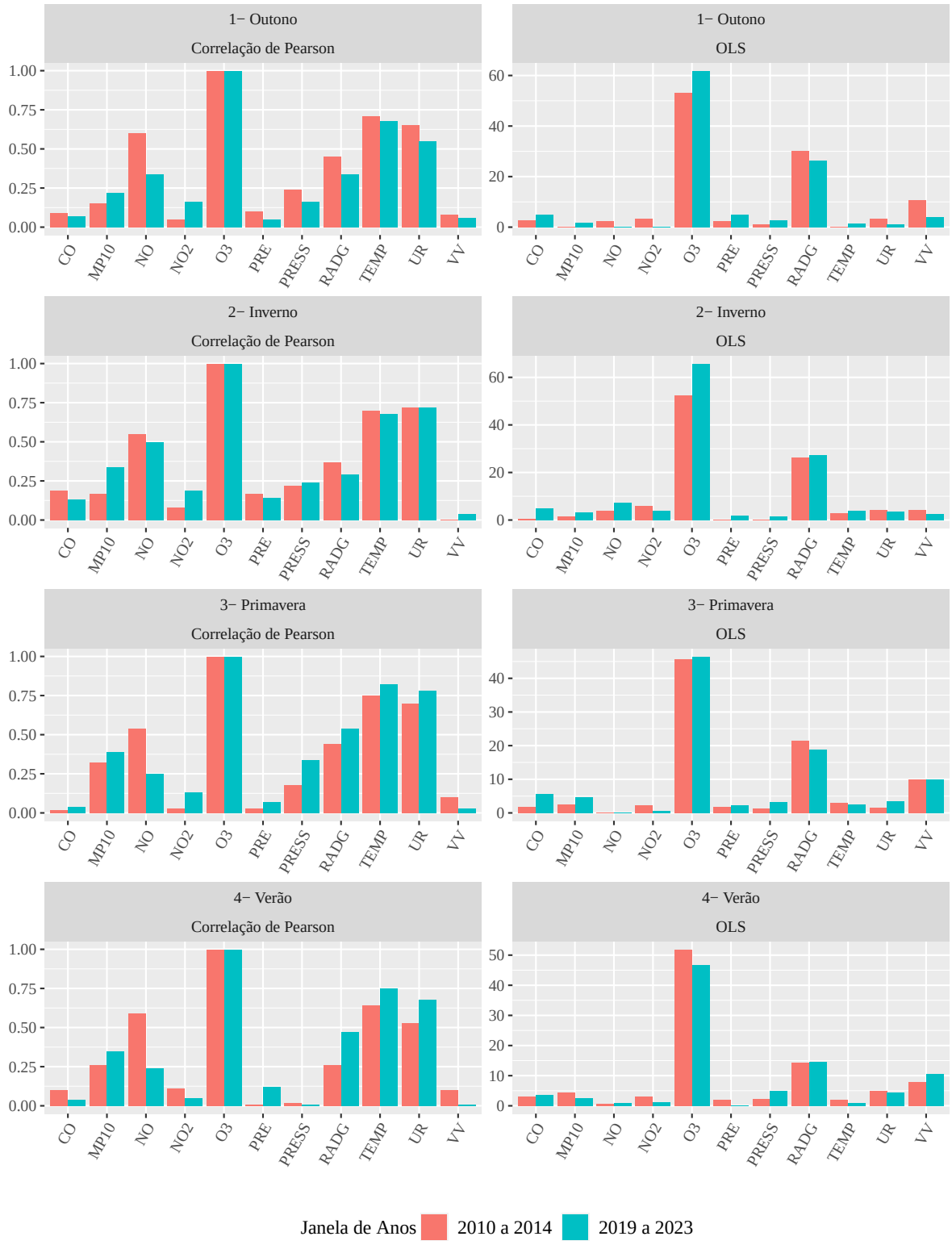

Figura 32 – Comparativo entre o Coeficiente de Correlação de Pearson e a significância das variáveis no modelo OLS. Valores considerando as janelas de anos de 2010-2014 e 2019-2023.

4.3 MÉTODOS BASEADOS EM ÁRVORES DE DECISÃO

O CART (BREIMAN et al., 1984), acrônimo de *Classification And Regression Trees*, é um método de aprendizado de máquina baseado em árvores de decisão que é utilizado em problemas de classificação e regressão. As árvores são utilizadas para classificação quando a variável resposta é categórica, enquanto que nos problemas de regressão a variável resposta é numérica. No caso do presente trabalho, o O_3 é a variável resposta e assume valores numéricos, tratando-se, portanto, de um problema de regressão. Neste caso, aborda-se nessa Seção um descritivo dos métodos baseados em árvores de decisão para o contexto de regressão.

Uma árvore de decisão se assemelha a um fluxograma em que são estabelecidos nós de decisão de forma ordenada. O nó inicial é o nó raiz. Em cada nó é determinada uma condição que avalia através de um teste lógico qual será o próximo nó descendente ou filho. A depender de a condição ser satisfeita ou não, divide-se exatamente em dois nós filhos: o nó esquerdo e o nó direito. Quando os dados satisfazem o teste lógico do nó intermediário seguem para o nó esquerdo e quando não satisfazem seguem para o nó direito. O procedimento é repetido até que um nó sem ramificação seja alcançado. Os nós sem ramificações, também denominados como nós terminais, são os resultados finais da árvore. A partir de um conjunto de dados, a construção de uma árvore é feita em três etapas:

1. Definição da melhor partição em cada nó.
2. Decisão de quando declarar um nó terminal ou optar por continuar particionando.
3. Predição a partir dos valores obtidos a cada nó terminal. Nessa etapa, a predição resulta do cálculo da média dos valores obtidos em cada nó terminal.

O objetivo do CART é obter uma função preditora para a resposta de interesse a partir dos dados de entrada, $\hat{y} = T(\mathbf{x})$, em que T representa a árvore construída a partir do conjunto de dados de entrada $\mathbf{x}$. Denota-se o conjunto de valores de entrada $\mathbf{x} = (x_1, \cdots, x_p)'$ de forma que $\mathbf{x}$ é uma realização do vetor de variáveis aleatórias $\mathbf{X} = (X_1, \cdots, X_p)'$. O vetor y é a variável resposta definida como uma realização da variável Y. Considera-se o conjunto de dados D constituído por n amostras de pares $(\mathbf{x}_i, y_i)$, em que $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip}), i = 1, \cdots, n$, são os valores observados de entrada e a respectiva variável resposta (BREIMAN et al., 1984; SOUZA, 2021).

Para exemplificar o processo de construção de uma árvore de decisão, considere o conjunto de dados $\mathcal{D} = \{(\mathbf{x}_i, y_i), |\mathcal{D}| = n\}$, para $p = 2$, e que $(X_1, X_2, Y) \in \mathbb{R}$. Os pares $(\mathbf{x}_i, y_i)$ são realizações das variáveis X_1, X_2, Y e $|\mathcal{D}| = n$ é o número de observações de $\mathcal{D}$. A árvore T é construída a partir de $\mathbf{x} = (x_1, x_2)$. Na Figura 33 apresenta-se uma ilustração hipotética de uma árvore obtida a partir deste conjunto de dados. A construção

da árvore ilustrada na Figura 33 (a) começa com todos os dados em $\mathcal{D}$. O teste $x_1 \leq v_1$ define uma partição do conjunto de dados $\mathcal{D}$ em dois subconjuntos: Subconjunto $\mathcal{D}_e$: sim, para o teste $x_1 \leq v_1$ (ramificação à esquerda) e subconjunto $\mathcal{D}_d$: não, para o teste $x_1 \leq v_1$, isto é, $x_1 > v_1$ (ramificação à direita), sendo $\mathcal{D}_e \cap \mathcal{D}_d = \oslash$. Os dados $\mathcal{D}_e$ alocados à esquerda são particionados novamente pelo teste $x_2 \leq v_2$. Os dados $\mathcal{D}_d$ alocados à direita são particionados novamente pelo teste $x_1 \leq v_3$. Os nós são particionados sucessivamente até se chegar aos nós terminais. Cada nó terminal define uma região R_j correspondente, $j = 1, \cdots, 5$. Em cada um dos nós terminais calcula-se o valor da predição $\hat{y}_j$ para as observações alocados no determinado nó, representada por $\tilde{T}_j$. O cálculo da predição é dado pela média das observações de cada nó. Na Figura 33 (b) representa-se o diagrama com as regiões delimitadas pelos nós terminais da árvore na Figura 33 (a).

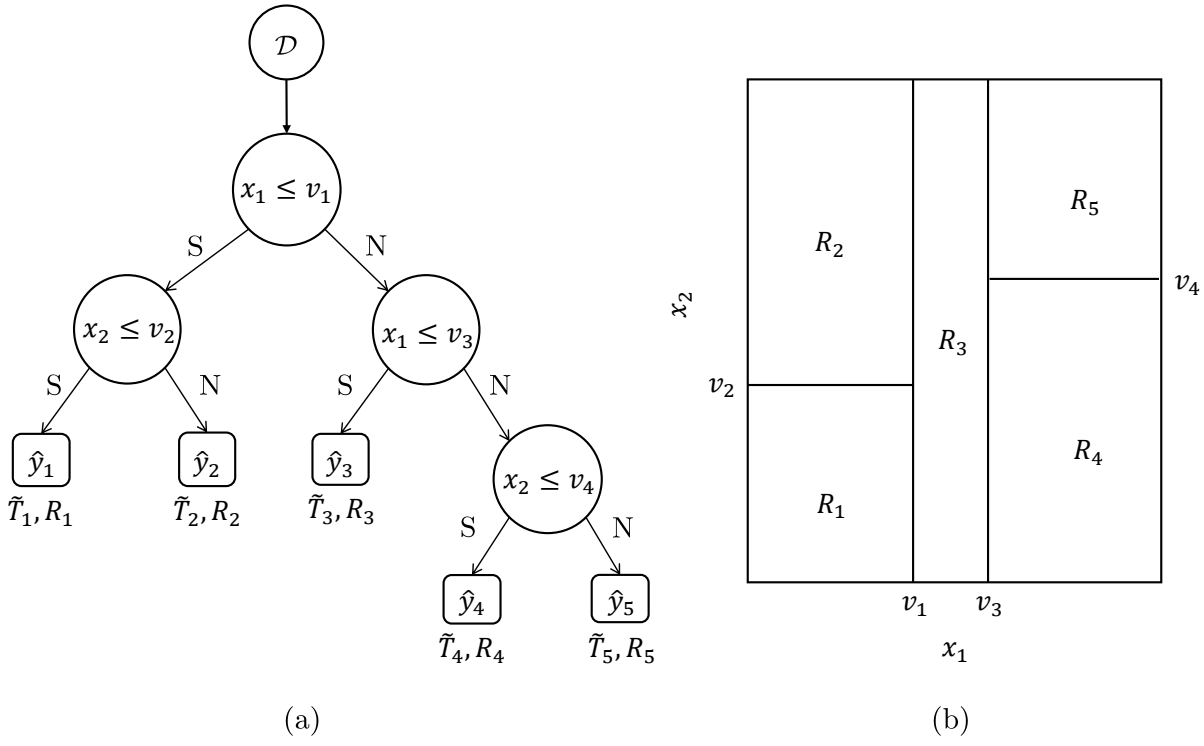

Figura 33 – (a) Respectiva árvore construída a partir de um conjunto de dados $\mathcal{D}$. (b) Partição do espaço bidimensional das variáveis preditoras por divisão binária recursiva como usado no CART. Figura inspirada de (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

A escolha da melhor partição de um nó, dado o conjunto de valores observados de cada variável aleatória, é feita através do cálculo de uma **medida de impureza** (JAMES et al., 2014). Em árvores de decisão para regressão, a medida de impureza utilizada é a variância. A variância mede a dispersão dos valores em relação à média, e é utilizada para determinar a homogeneidade dos valores dentro de um nó. Quando uma divisão resulta em nós com menor variância, significa que os valores dentro desses nós são mais semelhantes entre si, o que indica uma melhor qualidade da divisão. Durante a construção da árvore

de decisão, a melhor divisão é aquela que maximiza a redução da variância, resultando em nós filhos que são mais homogêneos em relação aos valores das variáveis dependentes. Assim, a variância serve como uma medida de impureza que guia o processo de divisão, ajudando a criar uma árvore que minimiza os erros de predição.

A seleção das partições é baseada na abordagem de divisão binária recursiva (*recursive binary splitting*) que é de cima para baixo e gananciosa (*top-down and freedy*) (JAMES et al., 2014). É de cima para baixo, porque o ponto de partida para a construção da árvore é o nó raiz, com todo o conjunto de aprendizado $\mathcal{D}$. Para cada variável preditora, são percorridos todos os valores possíveis procurando entre todas as partições binárias a melhor partição que maximiza o decréscimo de impureza. É recursiva porque o processo se repete a cada novo nó até que todos os nós sejam declarados nós terminais.

De posse de um critério para selecionar a melhor partição de um nó, o próximo passo é determinar quando fixar um nó como terminal e parar de particionar. As principais abordagens utilizadas para declarar um nó terminal são (BREIMAN et al., 1984):

1. Quando as variáveis de entrada x_j são localmente constantes. Nesta situação, não é possível dividir um nó em dois subconjuntos não vazios.

2. Quando um nó contiver menos do que um número mínimo de observações pré-estabelecido.

3. Quando a profundidade de uma árvore, em termos da quantidade de nós intermediários, for maior ou igual a um limiar fixo pré-estabelecido.

4. Quando o decréscimo de impureza for menor do que um limiar fixo pré-estabelecido.

O CART foi primeiramente apresentado por (BREIMAN et al., 1984) e fornece uma base para importantes algoritmos de aprendizado de máquina, como o *Bagging*, *Random Forest* e *XGBoost*. A seguir apresenta-se um exemplo detalhado de construção de uma árvore CART, incluindo as definições das regiões de tomada de decisão.

EXEMPLO DETALHADO DE CONSTRUÇÃO DE UMA ÁRVORE CART

Considera-se nesse exemplo que os dados estão disponíveis na matriz $\mathcal{X}_s$ e no vetor de variável resposta $\mathbf{y}_s$. A primeira coluna de $\mathcal{X}_s$, aqui representada como $\mathcal{X}_s(:,1)$, contém os valores da temperatura (TEMP) e a segunda coluna, $\mathcal{X}_s(:,2)$, contém os valores do monóxido de carbono (CO). O vetor de resposta deseja $\mathbf{y}_s$ contém os valores da concentração do O_3. Por simplicidade, considera-se que se tem apenas 10 valores hipotéticos de cada variável. Esses valores, indicados na Tabela 15 representam as $n = 10$ observações horárias das variáveis TEMP e CO. Essas variáveis estão indicadas no instante (t-1) em relação as variáveis da resposta deseja O_3, que estão no instante t.

Tabela 15 – Conjunto de dados hipotéticos.

i	$\mathcal{X}_s(:,1)$	$\mathcal{X}_s(:,2)$	$\mathbf{y}_s$
1	18,9	1,1	85
2	19,0	0,7	77
3	20,4	0,8	99
4	19,9	0,7	103
5	20,0	0,7	87
6	20,1	0,8	71
7	20,1	0,8	54
8	20,0	1,2	102
9	22,5	0,6	108
10	22,6	0,8	89

Os seguintes passos são usados para a construção da árvore:

1. No nó raiz t_0, considera-se todas as observações do conjunto de dados D.

2. No caso de árvores de regressão, a medida de impureza utilizada no processo de partição é a variância. Inicialmente, calcula-se a média do conjunto de valores da resposta desejada. Assim, o valor médio dos elementos de todos os elementos do vetor $\mathbf{y}_s$ é

$$\bar{y}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i = 87{,}5$$

Em seguida calcula-se a variância de $\mathbf{y}_s$ (grau de impureza),

$$i(t_0) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_0)^2 = 251{,}65$$

3. Para especificar os pontos de corte v de cada uma das variáveis, inicialmente considera-se os valores ordenados de cada uma das colunas da matriz de dados, excluindo os valores repetidos. Note que no caso da primeira coluna da matriz de dados tem-se $x_1(6) = x_1(7)$ e $x_1(5) = x_1(8)$. Excluindo os valores repetidos e ordenando os elementos restantes tem-se

$$\mathbf{x}_1 = \mathcal{X}_s(:,1) \Longrightarrow (18{,}9; 19{,}0; 19{,}9; 20{,}0; 20{,}1; 20{,}4; 22{,}5; 22{,}6)$$

No cálculo dos pontos de corte, aplica-se a média entre dois pontos sucessivos dos valores ordenados, ou seja,

$$\begin{aligned}
v(1) &= (18{,}9 + 19{,}0)/2 = 18{,}95\\
v(2) &= (19{,}0 + 19{,}9)/2 = 19{,}45\\
v(3) &= (19{,}9 + 20{,}0)/2 = 19{,}95\\
v(4) &= (20{,}0 + 20{,}1)/2 = 20{,}05\\
v(5) &= (20{,}1 + 20{,}4)/2 = 20{,}25\\
v(6) &= (20{,}4 + 22{,}5)/2 = 21{,}45\\
v(7) &= (22{,}5 + 22{,}6)/2 = 22{,}55.
\end{aligned}$$

4. Para cada ponto de corte, particiona-se o nó raiz t_0 em nós filhos t_1 e t_2. Em seguida calcula-se, para cada um deles, o valor da medida de impureza.

   Como exemplo, considere $v(4) = 20{,}05$ que define as duas partições, uma com todos os elementos do vetor $\mathbf{x}_1$ que são $\leq 20{,}05$ e outra com todos os elementos de $\mathbf{x}_1$ que são $> 20{,}05$. Essas partições são alocadas em t_1 e t_2 juntamente com os respectivos elementos do vetor $\mathbf{y}_s$. Especificamente em t_1 e t_2 os pares $(x_1(i), y_i)$ alocados são

   $$D_1 = (18{,}9; 85), (19{,}0; 77), (19{,}9; 103), (20{,}0; 87), (20{,}0; 102)$$

   e

   $$D_2 = (20{,}1; 71), (20{,}1; 54), (20{,}4; 99), (22{,}5; 108), (22{,}6; 89),$$

   respectivamente. Nota-se que para cada D_1 e D_2 tem-se 5 pares alocados. Calcula-se os valores médios, para os subconjuntos de elementos do vetor de resposta desejada associados a D_1 e D_2, ou seja,

   $$\bar{y}_1 = \frac{85 + 77 + \cdots + 102}{5} = 90{,}8$$

   e

   $$\bar{y}_2 = \frac{71 + 54 + \cdots + 89}{5} = 84{,}2.$$

   Para cada um dos conjuntos com 5 elementos, calcula-se as medidas de impureza nos nós $i(t_1)$ e $i(t_2)$, ou seja,

   $$i(t_1) = \frac{[(85 - \bar{y}_1)^2 + (77 - \bar{y}_1)^2 + \cdots + (102 - \bar{y}_1)^2]}{5} = 102{,}56$$

   e

   $$i(t_2) = \frac{[(71 - \bar{y}_2)^2 + (54 - \bar{y}_2)^2 + \cdots + (89 - \bar{y}_2)^2]}{5} = 378{,}96.$$

   Destaca-se que a impureza inicial é $i(t_0) = 251{,}65$, e com a partição $x_1 \leq 20{,}05$ tem-se $i(t_1) = 102{,}56$ e $i(t_2) = 378{,}96$. Assim, resulta que, para a partição $x_1 \leq 20{,}05$, o decréscimo de impureza vale

   $$\Delta I(x_1 \leq 20{,}05, t_0) = i(t_0) - \frac{5}{10} \times i(t_1) - \frac{5}{10} \times i(t_2) = 10{,}89.$$

5. O cálculo do decréscimo de impureza é feito de forma recursiva para todos os pontos de corte v das variáveis x_1 e x_2. As Tabelas 16 e 17 especificam os valores de todos os pontos de corte possíveis para as variáveis x_1 e x_2 e os correspondentes decréscimos de impurezas. A partição que resulta no maior decréscimo de impureza é $x_1 \leq 20{,}25$, cujo valor é 53,44. Assim, o ponto de corte $x_1 \leq 20{,}25$ é definido como nó raiz t_0.

Tabela 16 – Ilustração CART - Partição $\mathcal{X}_s(:,1)$.

v	18,95	19,45	19,95	20,05	20,25	21,45	22,55
$i(t_1)$	0,00	16,00	118,22	102,56	257,35	254,19	279,33
$i(t_2)$	278,84	297,36	308,41	378,96	60,22	90,25	0,00
$\Delta I(x_1 \leq v, t_0))$	0,69	10,56	0,30	10,89	53,44	30,25	0,25

Tabela 17 – Ilustração CART - Partição $\mathcal{X}_s(:,2)$.

v	0,65	0,75	0,95	1,15
$i(t_1)$	0,00	153,69	285,25	253,65
$i(t_2)$	227,73	273,56	72,25	0,00
$\Delta I(x_2 \leq v, t_0))$	46,69	26,04	9,00	23,36

6. O resultado do particionamento do nó raiz em dois nós filhos, t_1 e t_2, é apresentado na Figura 34.

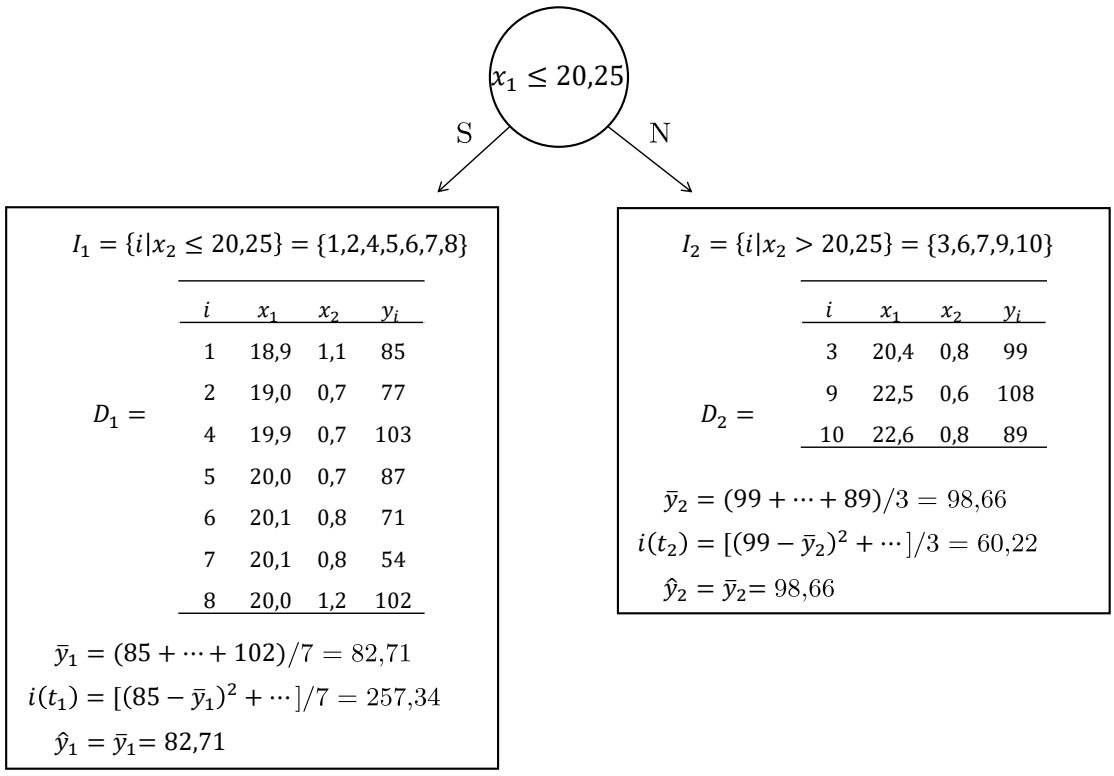

Figura 34 – Ilustração CART: Particionamento do nó raiz e os agrupamentos de dados considerados nas estimativas $\hat{y}_1$ e $\hat{y}_2$.

Nota-se que, $i(t_1) = 257{,}34$ e $i(t_2) = 60{,}22$. Assim, o erro interno do nó terminal no conjunto de dados é obtido por

$$Erro = \frac{7 \times i(t_1) + 3 \times i(t_2)}{10} = 198{,}204.$$

Esse erro está associado a qualidade da estimativa $\hat{y}$. O erro de predição é obtido a partir da minimização do erro em cada nó terminal, isto é, quanto menor o erro dentro de

cada nó terminal, menor será o erro de predição $\hat{y}$ da árvore $T(\mathbf{x})$. Além disso, quanto mais significativa a variável, menor será o erro do nó terminal e por consequência menor será o erro da predição de $\hat{y}$. A escolha da variável mais importante é feita por meio na maximização do decréscimo de impureza. Dessa forma escolhemos a variável que maximiza essa função e que está associada ao menor erro do nó terminal.

Destaca-se que o valor da impureza inicial $i(t_0)$ subtraído do erro é o valor obtido no decréscimo de impureza, isto é

$$\begin{aligned} i(t_0) - Erro &= \Delta I(x_1 \leq 20{,}25, t_0) \\ 251{,}65 - 198{,}204 &= 53{,}44. \end{aligned}$$

O processo de construção da árvore é feito de forma recursiva, particionando os nós t_1 e t_2 em novos nós filhos até que todos os nós sejam declarados terminais. Uma vez que a árvore foi construída, pode-se verificar quais são as variáveis mais importantes. No caso do exemplo, como a variável x_1 é usada apenas uma única vez na raiz da árvore, sua importância é dada por

$$Imp(x_1) = \Delta I(x_1 \leq 20{,}25, t_0) = 53{,}44.$$

No caso de uma árvore mais profunda, com maiores ramificações, a medida de impureza de uma dada variável X_l é definida como a soma dos decréscimos de impureza gerados em cada nó interno de uma árvore, ou seja,

$$Imp(x_l) = \sum_{j=i}^{J} \Delta I(s^l, t_j), t_j \in T. \tag{4.21}$$

sendo s^l todos os cortes selecionados da variável X_l para cada nó interno t da árvore T. Os critérios de parada, definidos em termos de parâmetros (hiperparâmetros) são responsáveis por controlar a complexidade das árvores (BREIMAN et al., 1984). Dependendo da sua escolha, árvores muito profundas ou árvores muito simples, com poucos nós intermediários, podem ser obtidas. Em particular, árvores muito profundas tendem a aprender padrões altamente irregulares: elas se ajustam bem ao conjunto de treinamento $\mathcal{D}$, mas apresentam pouco poder de generalização para novos conjuntos de dados. Em outras palavras, essas árvores apresentam baixo viés e alta variância (comumente conhecido como *"trade-off"* viés-variância (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; JAMES et al., 2014)). A escolha dos valores dos hiperparâmetros pode ser feita pelo usuário ou através de procedimentos computacionais conhecidos como *tuning* (JAMES et al., 2014). Esses procedimentos podem ser computacionalmente custosos, porém, usualmente resultam em modelos com melhor capacidade preditiva. Vale ressaltar que uma característica interessante das árvores de decisão é que elas são extremamente interpretáveis. No entanto, costumam apresentar baixo poder preditivo quando comparadas com outros modelos mais complexos.

RANDOM FOREST

Para contornar alguns problemas apresentados nas árvores de decisão, foram desenvolvidos alguns métodos denominados por *ensembles*. Esses métodos combinam - de diferentes formas - várias árvores, aumentando consideravelmente o seu poder preditivo, às custas de perda da interpretabilidade. Entre essas técnicas, destaca-se o modelo de Florestas Aleatórias, do inglês, *Random Forest* (RF). O RF é uma adaptação do CART e consiste em calcular a média de várias árvores de decisão, treinadas a partir de diferentes amostras do mesmo conjunto de treinamento com o objetivo de reduzir a variância e aumentar o seu poder preditivo (BREIMAN, 2001; HASTIE; TIBSHIRANI; FRIEDMAN, 2009; JAMES et al., 2014).

O modelo RF pode ser construído a partir dos seguintes passos:

1. Utilizando a técnica de *bootstrap*, são geradas B amostras com reposição de $\mathcal{D}$ formando um conjunto de réplicas $\{\mathcal{D}_1, \cdots, \mathcal{D}_B\}$.

2. Para cada umas das instância $\mathcal{D}_b, b = 1, \cdots, B$, uma árvore T_b é construída repetindo recursivamente as seguintes etapas para cada nó terminal da árvore, até que o tamanho mínimo do nó n_{min} seja atingido:

   (a) São sorteados m dos p preditos, $m < p$.

   (b) Escolha a melhor variável/ponto de partição entre m.

   (c) Divida o nó em dois nós filhos.

3. Resultando na sequência de árvores $\{T_1, \cdots, T_B\}$.

A predição final para um vetor de entrada $\mathbf{x}$ genérico é dada pela média dos previsores, isto é,

$$T(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} T_b(\mathbf{x}). \tag{4.22}$$

Geralmente em problemas de regressão com p variáveis preditoras, utiliza-se para m uma aproximação do valor de $p/3$ (arredondado para baixo). Além disso, alguns autores sugerem que $B \approx 100$ costuma ser suficiente para gerar bons resultados (HASTIE; TIBSHIRANI; FRIEDMAN, 2009), (JAMES et al., 2014). No entanto, os melhores valores para esses parâmetros dependerão do problema e podem ser tratados na otimização dos hiperparâmetros por meio da validação cruzada.

A extensão do cálculo da importância das variáveis para o modelo RF é feita de maneira análoga ao CART. Especificamente, para o modelo RF com B árvores $\{T_1, \cdots, T_B\}$,

a medida de importância de uma dada variável X_l é dada pela média da soma dos decréscimos de todas as árvores da floresta, ou seja

$$Imp(X_l) \quad = \quad \frac{1}{B}\sum_{b=1}^{B}\sum_{j=i}^{J}\Delta I(s^l, t_{bj}). \tag{4.23}$$

4.3.1 RESULTADOS DAS SIMULAÇÕES BASEADAS EM ÁRVORES

Na implementação dos algoritmos baseados em árvores de decisão, CART e RF, os dados são organizados em uma tabela bidimensional contendo o vetor de resposta desejada e a matriz de dados, isto é

$$\mathbf{DF}_s = \left[\begin{array}{cc} \mathcal{X}_s & \mathbf{y}_s \end{array} \right]. \tag{4.24}$$

Como feito na implementação do OLS, os dados são agrupados em dois conjuntos de anos, um conjunto contendo os dados de 2010 a 2014 e outro com os dados de 2019 a 2023. Além disso, para cada conjunto de anos os dados são considerados separadamente para cada uma das estações meteorológica. Assim, no total foram consideradas 8 tabelas bidimensionais. A partir dos dados de cada tabela buscou-se obter o melhor conjunto de parâmetros para a predição das concentrações de ozônio.

Na implementação dos algoritmos baseados em árvores de decisão, é necessária a definição de alguns parâmetros. A escolha desses parâmetros influencia diretamente na performance do modelo resultante, pois controlam o processo de aprendizagem. Importante notar que, um parâmetro cujo valor é usado para controlar o processo de aprendizagem é batizado de hiperparâmetro. O ajuste de parâmetros (*Hyperparameter Tuning*) é feito com técnicas de otimização. Nas simulações computacionais com o RF, usualmente são necessários os seguintes parâmetros (PROBST; WRIGHT; BOULESTEIX, 2019; BARTZ et al., 2023):

1. *mtry*: Número de variáveis candidatas sorteadas em cada divisão. Valores mais baixos desse parâmetro levam a árvores mais diversas e menos correlacionadas, proporcionando melhor estabilidade na agregação. O intervalo possível é $[1, 2, \cdots, p]$, sendo p o número de variáveis do conjunto de dados. O padrão para problemas de regressão é utilizar como valor $p/3$. Neste trabalho, têm-se o valor de $p = 11$.

2. *ntree*: Número de árvores na floresta. A rigor, pode-se usar qualquer valor no intervalo $[1, 2, \cdots, +\infty]$. Usualmente, os valores máximos utilizados ficam entre 500 a 1.000 árvores. Valores mais altos tendem a melhorar a performance do algoritmo em termos de qualidade da estimativa, porém, aumentam o custo computacional.

3. *nodesize*: Número mínimo de observações em um nó terminal. Diminuir este parâmetro aumentará a profundidade das árvores, o que significa que mais divisões são realizadas até os nós terminais. O intervalo possível é $[1, 2, \cdots, +\infty]$. Usualmente,

em problemas de regressão, utiliza-se como padrão o valor de 5 observações em um nó terminal.

No caso das simulações com o algoritmo CART, considerou-se o parâmetro *nodesize*. O procedimento computacional e as rotinas utilizadas no treinamento dos algoritmos CART e RF estão descritos no Apêndice C. Os valores usados na implementação CART e no RF, após o processo de otimização, estão indicados na Tabela 18.

Tabela 18 – Parâmetros usados na implementação do RF e CART após o processo de otimização.

Estação	Anos	Modelo RF			Modelo CART
		mtry	*nodesize*	*ntree*	*nodesize*
Outono	2010 a 2014	4	2	600	2
	2019 a 2023	4	1	700	1
Inverno	2010 a 2014	5	3	700	2
	2019 a 2023	4	3	600	2
Primavera	2010 a 2014	4	1	800	3
	2019 a 2023	5	3	800	3
Verão	2010 a 2014	3	2	500	2
	2019 a 2023	4	1	500	4

Os algoritmos CART e do RF foram implementados, considerando os parâmetros da Tabela 18. As avaliações em termos das medidas MSE, MAE e R^2 estão indicados na Tabela 19. Esses resultados de desempenho são também comparados com o OLS e o LASSO.

Tabela 19 – Avaliação dos modelos LASSO, OLS, CART e RF.

Estação	Modelo	Dados dos anos (2010 - 2014)			Dados dos anos (2019 - 2023)		
		MSE	MAE	R^2(%)	MSE	MAE	R^2(%)
Outono	OLS	137,36	8,55	87%	112,42	7,62	86%
	LASSO	137,48	8,55	87%	112,49	7,62	86%
	CART	124,73	7,89	88%	83,34	6,54	90%
	RF	**26,98**	**3,64**	**98%**	**22,47**	**3,26**	**98%**
Inverno	OLS	139,55	8,06	86%	85,9	6,72	89%
	LASSO	139,67	8,07	86%	85,94	6,72	89%
	CART	92,07	6,60	91%	61,03	5,65	92%
	RF	**23,27**	**3,23**	**98%**	**15,76**	**2,72**	**98%**
Primavera	OLS	276,5	11,77	85%	219,25	10,91	87%
	LASSO	277,21	11,78	85%	219,34	10,92	87%
	CART	235,86	10,63	88%	137,25	8,37	92%
	RF	**52,73**	**4,85**	**97%**	**36,17**	**4,21**	**98%**
Verão	OLS	215,73	10,47	86%	193,1	10,02	86%
	LASSO	215,79	10,46	86%	193,26	10,01	86%
	CART	156,50	8,77	90%	154,67	8,82	89%
	RF	**39,77**	**4,42**	**98%**	**38,82**	**4,37**	**97%**

A partir da observação dos valores da Tabela 19 destacam-se:

- Os modelos não-lineares CART e RF apresentam um desempenho melhor para modelar a concentração de ozônio que os modelos OLS e LASSO. Destaca-se que o RF apresenta um desempenho significativamente melhor em todos os cenários considerados e com o R^2 acima de $\approx 98\%$, consegue explicar melhor a variabilidade dos dados que os demais métodos considerados.
- Os menores valores de MSE e MAE para os modelos CART e RF são observados no outono e no inverno e os maiores são observados na primavera e no verão.
- Comparando os valores de desempenho obtidos pode-se conjecturar que as métricas de avaliação estão sujeitas as condições climáticas das estações dos anos e não estão diretamente relacionadas aos intervalos de tempo dos conjuntos de anos considerados.

A fim de obter uma avaliação qualitativa da comparação dos modelos CART e RF, fez-se os gráficos de dispersão entre os valores preditos e observados utilizando o conjunto de dados de 2010 a 2014. Esses gráficos estão apresentados na Figura 35. Nas Figuras 35 (a) e (b) apresentam-se os resultados do CART e do RF, respectivamente.

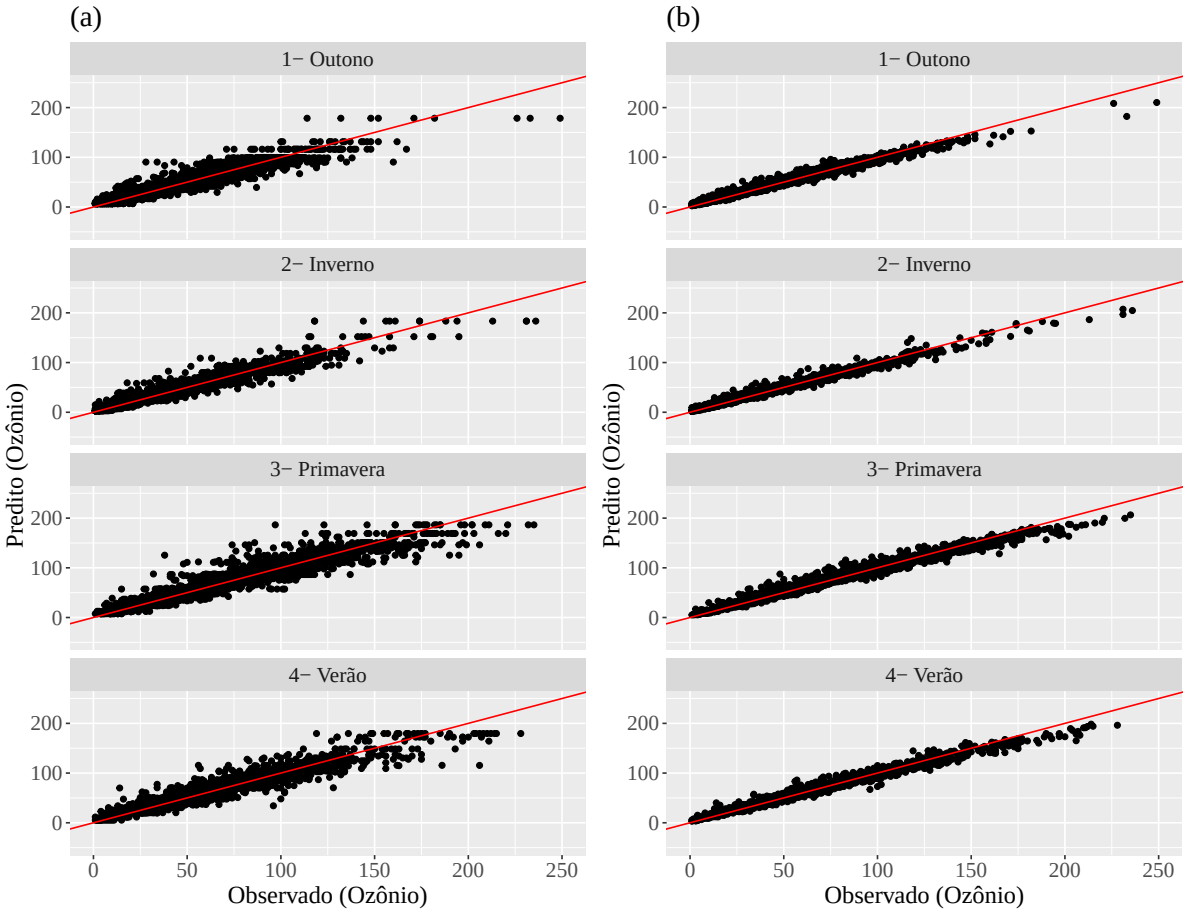

Figura 35 – Gráficos de dispersão dos valores preditos do ozônio em função dos valores observados: (a) Modelo CART e (b) Modelo RF. Valores considerando os dados dos anos de 2010 a 2014. A reta indicada em vermelho é o caso ideal quando $x = y$.

A partir da análise da Figura 35, observa-se que os valores preditos pelo modelo RF estão mais próximos da reta vermelha $x = y$ em comparação com o modelo CART. Isso indica que o modelo RF possui uma precisão maior na predição das concentrações de O_3. Mesmo para os pontos mais extremos, onde são observados os maiores valores de O_3, o modelo RF apresenta um desempenho superior ao do modelo CART, apesar de ainda subestimar esses valores. Este comportamento sugere que o modelo RF consegue capturar melhor a relação entre as variáveis, resultando em previsões mais próximas dos valores observados, o que não é tão evidente no modelo CART.

Dando continuidade à análise qualitativa, foram realizadas simulações com o modelo RF considerando os conjuntos de anos de 2010 a 2014 e de 2019 a 2023. Em cada conjunto, os dados foram analisados separadamente para as estações do outono, inverno, primavera e verão. Os gráficos de dispersão entre os valores observados e preditos em cada cenário estão ilustrados na Figura 36. Observa-se que, em todos os cenários considerados, os valores preditos e os observados se mantêm próximos da reta vermelha ($x = y$), indicando um bom desempenho do modelo RF. Isso é válido até mesmo para os pontos mais extremos, onde são registrados os maiores valores de O_3, o que demonstra a capacidade do modelo em lidar com valores atípicos e altos. Apesar dos resultados promissores, esses gráficos não evidenciam de maneira clara a influência específica das variáveis independentes na concentração de O_3 e como essa influência pode variar ao longo do tempo. Para uma análise mais detalhada, seria necessário investigar a importância das variáveis dentro do modelo RF, bem como realizar uma decomposição temporal para entender como as relações entre as variáveis e as concentrações de O_3 mudam entre as diferentes estações e períodos analisados.

A fim de facilitar a visualização da importância quantitativa das variáveis na predição da concentração do O_3, considera-se os mesmos cenários dos gráficos de dispersão, porém agora serão evidenciados os valores da importância das variáveis. No modelo OLS, conforme as implementações do caso linear, a importância das variáveis é representada pela estatística t. No modelo RF, a importância das variáveis definida conforme Equação 4.23, é calculada com o auxílio do pacote VIP (GREENWELL; BOEHMKE, 2020). O pacote VIP (*Variable Importance Plots*) do software *R* é uma ferramenta utilizada para avaliar e visualizar a importância das variáveis em modelos preditivos. Ele permite calcular a relevância de cada variável para uma variedade de modelos de *machine learning*, como florestas aleatórias, redes neurais, entre outros. Além disso, o VIP oferece métodos diversificados, incluindo importância por permutação (BREIMAN, 2001) e *Shapley Values* (LUNDBERG; LEE, 2017), e é compatível com diversos pacotes do *R*. A visualização da importância das variáveis, por meio de gráficos intuitivos, facilita a interpretação e comunicação dos resultados, tornando o pacote uma valiosa adição para análises preditivas e a compreensão dos fatores que influenciam as predições dos modelos.

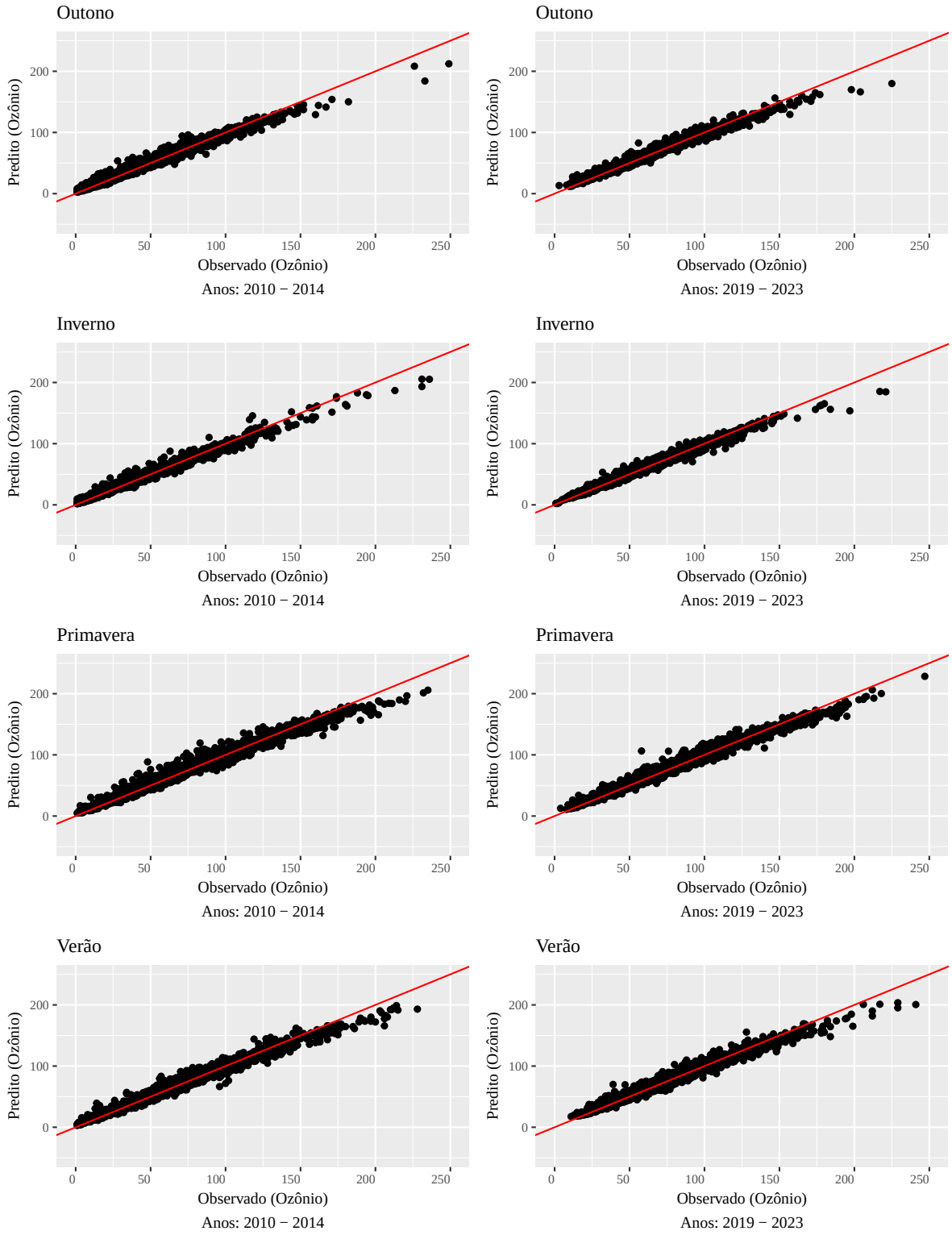

Figura 36 – Gráfico de dispersão dos valores do ozônio preditos pelo modelo RF em função dos valores observados. A reta indicada em vermelho é o caso ideal quando $x = y$.

Para melhor visualizar o desempenho preditivo do modelo RF, apresenta-se na Figura 37 um trecho da série para a estação do outono para o conjunto de anos de 2019 a 2023. A linha vermelha pontilhada indica os valores preditos pelo modelo, enquanto a linha contínua preta representa os valores observados. A partir da Figura 37, nota-se que o modelo RF consegue capturar bem a tendência geral dos valores observados de O_3 ao longo do tempo. As linhas predita e observada apresentam uma concordância razoável, com o modelo acompanhando as flutuações e variações na concentração de O_3. No entanto, há alguns períodos em que o modelo subestima ou superestima os valores observados, especialmente em picos de alta concentração de O_3. Apesar disso, os resultados mostram que o modelo RF apresenta um desempenho robusto.

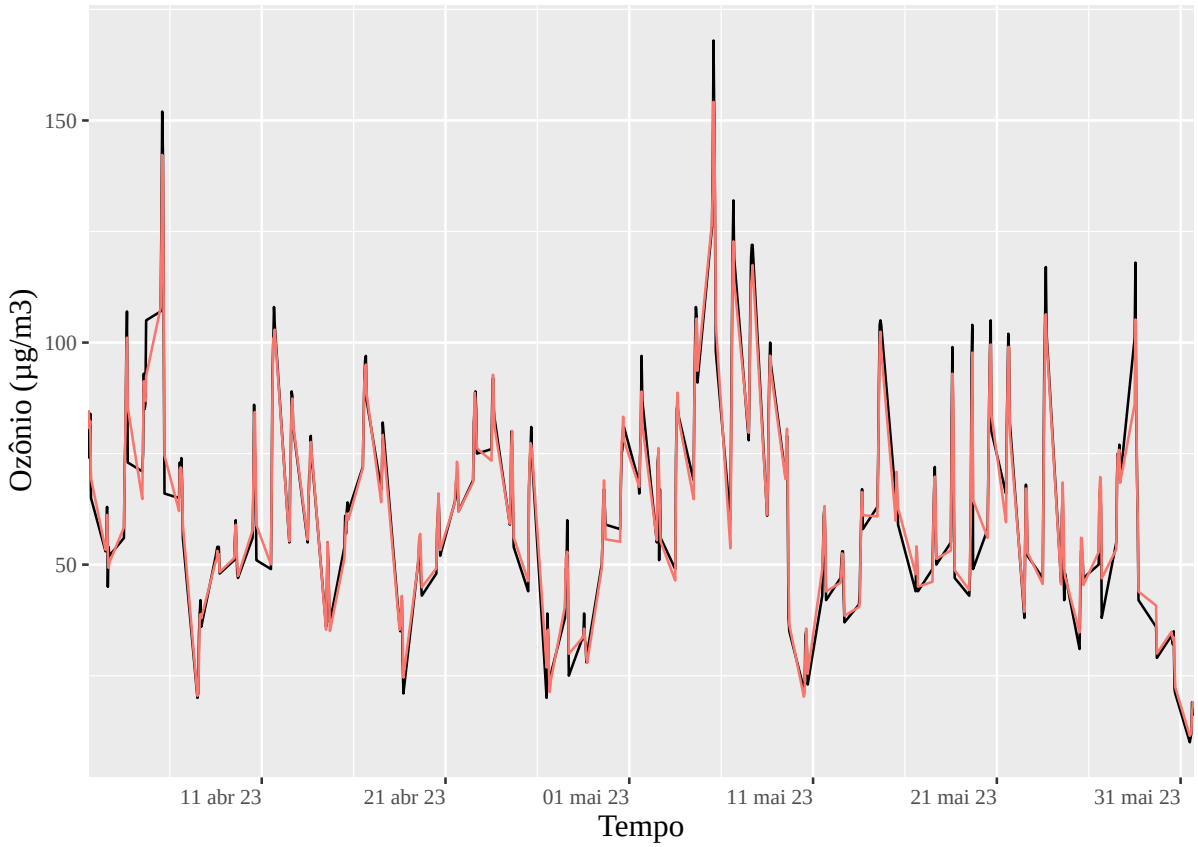

Figura 37 – Trecho da série temporal dos valores preditos pelo modelo RF versus os valores observados das concentrações de O_3 para o mês do outono do conjunto de anos de 2019 a 2023. A linha vermelha pontilhada indica os valores preditos pelo modelo, enquanto a linha contínua preta representa os valores observados.

Na Figura 38 apresenta-se os gráficos de barras contendo a importância das variáveis para os modelos OLS e RF. A altura da barra em cada caso é proporcional à importância da variável. Nos quatro gráficos de barras da primeira e da segunda colunas estão indicadas a importância de cada variável no OLS e no RF, respectivamente. As barras indicadas em laranja e azul representam os valores nos conjuntos de anos de 2010 a 2014 e 2019 a 2023, respectivamente. A partir da observação da Figura 38 destacam-se:

- A importância das variáveis calculadas em ambos os modelos varia significativamente para cada conjunto de anos (2019 a 2023 e 2010 a 2014) e também sofre alterações conforme as estações do ano. Entretanto, essas variações são notadamente distintas conforme o modelo.

- No modelo OLS, a importância das variáveis está concentrada de forma desproporcional em poucas variáveis. Especificamente, a concentração de ozônio (O_3) da hora anterior (t-1) e a radiação solar (RADG) são as variáveis com maior significância. Observa-se que, do primeiro para o segundo conjunto de anos, a significância da concentração de O_3 da hora anterior (t-1) diminui no verão, mas aumenta nas demais estações do ano. Em relação à radiação solar (RADG), há uma pequena queda na sua importância durante o outono e a primavera ao longo dos anos, enquanto sua importância se mantém relativamente constante durante o inverno e o verão. Adicionalmente, ao comparar os dois conjuntos de anos, observa-se um aumento significativo na importância do monóxido de carbono (CO) durante o outono, inverno e primavera. No entanto, no verão, esse aumento na significância do CO não é tão pronunciado.

- No modelo RF, a radiação solar (RADG) e a concentração de ozônio (O_3) da hora anterior (t-1) possuem importâncias equivalentes no outono, na primavera e no inverno. No entanto, no verão, a significância da RADG diminui substancialmente em comparação com a concentração de O_3 em (t-1), que se torna a variável mais importante. Ao longo dos anos, observa-se uma evolução na importância da concentração de O_3 da hora anterior (t-1). Essa importância aumenta consideravelmente durante o outono e o inverno, sugerindo que as concentrações anteriores de ozônio são indicadores cada vez mais relevantes para prever as concentrações futuras nessas estações. Em contraste, na primavera e no verão, a importância do O_3 em (t-1) diminui, o que pode ser atribuído à maior variabilidade e influência de outras variáveis durante esses períodos. Um efeito semelhante é observado para a RADG. No outono e no inverno, a RADG mantém uma importância significativa, mas essa importância diminui durante a primavera e o verão. Essa variação pode ser explicada pela maior influência da radiação solar nas reações fotoquímicas que produzem ozônio durante os meses mais quentes, o que torna outras variáveis, como a temperatura, também muito relevantes.

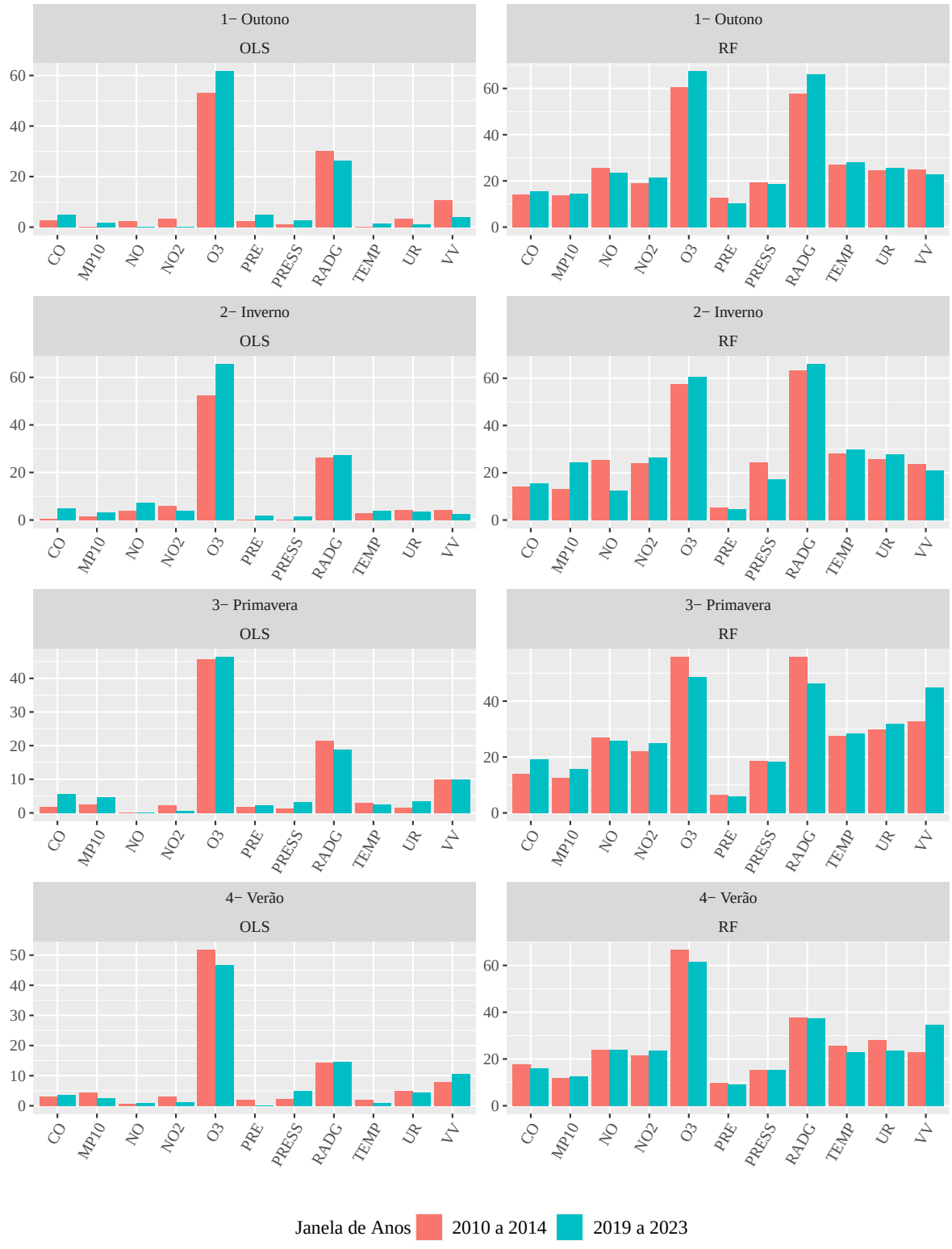

Figura 38 – Resumo da importância das variáveis paras os modelos OLS e RF. Valores de importância considerando as janelas de anos de 2010 a 2014 e 2019 a 2023: (1) outono, (2) inverno, (3) primavera e (4) verão.

De forma a complementar as observações feitas a partir da observação da Figura 38, os valores numéricos da importância das variáveis e a sua variação ao longo do tempo para o modelo RF são apresentados na Tabela 20. Adicionalmente, são incluídos os valores

médios, conforme apresentado na Tabela 20.

Tabela 20 – Valores numéricos da importância das variáveis em diferentes cenários para o modelo RF.

Variável	2010 a 2014		2019 a 2023		% Variação		Ordem de Importância[1]
	Valor Médio	Importância RF	Valor Médio	Importância RF	Valor Médio	Importância RF	
1- Outono							
CO	0,6	14,2	0,2	15,4	-58%	8%	(9)
MP_{10}	28,0	13,9	23,1	14,5	-17%	4%	(10)
NO	9,5	25,6	4,1	23,4	-57%	-9%	(5)
NO_2	37,1	19,1	26,8	21,5	-28%	13%	(7)
O_3	52,5	60,5	70,1	67,7	34%	12%	(1)
PRE	0,2	12,6	0,2	10,4	17%	-17%	(11)
PRESS	926,7	19,3	926,7	18,7	0%	-3%	(8)
RADG	1448,7	57,9	1899,1	66,3	31%	15%	(2)
TEMP	24,3	27,1	25,4	28,1	4%	4%	(3)
UR	62,5	24,6	56,1	25,8	-10%	5%	(4)
VV	2,5	24,9	2,1	22,8	-19%	-8%	(6)
2- Inverno							
CO	0,6	14,1	0,3	15,6	-46%	11%	(9)
MP_{10}	35,1	13,0	31,3	24,5	-11%	88%	(6)
NO	15,1	25,5	6,0	12,3	-60%	-52%	(10)
NO_2	51,5	24,1	33,7	26,4	-34%	10%	(5)
O_3	49,2	57,4	60,9	60,6	24%	6%	(2)
PRE	0,1	5,2	0,0	4,7	-64%	-10%	(11)
PRESS	929,7	24,3	929,8	17,1	0%	-30%	(8)
RADG	1410,3	63,3	1736,3	66,2	23%	5%	(1)
TEMP	22,0	28,2	22,6	29,7	3%	5%	(3)
UR	55,8	25,9	51,6	27,8	-7%	7%	(4)
VV	2,3	23,6	2,0	20,9	-16%	-11%	(7)
3- Primavera							
CO	0,6	13,9	0,3	19,3	-47%	39%	(8)
MP_{10}	33,3	12,6	31,3	15,8	-6%	25%	(10)
NO	7,9	26,9	4,5	25,9	-43%	-4%	(6)
NO_2	43,6	22,2	30,9	24,9	-29%	12%	(7)
O_3	76,0	56,0	81,9	48,6	8%	-13%	(1)
PRE	0,2	6,5	0,2	5,8	26%	-10%	(11)
PRESS	926,0	18,5	925,8	18,4	0%	-1%	(9)
RADG	1665,2	55,8	1785,8	46,4	7%	-17%	(2)
TEMP	24,9	27,6	25,7	28,3	3%	3%	(5)
UR	56,8	29,8	56,3	32,0	-1%	7%	(4)
VV	2,7	32,7	2,6	44,9	-6%	37%	(3)
4- Verão							
CO	0,5	17,6	0,2	16,0	-57%	-9%	(8)
MP_{10}	23,0	12,0	20,1	12,5	-12%	4%	(10)
NO	6,0	24,1	3,1	23,9	-47%	-1%	(4)
NO_2	33,1	21,6	23,0	23,5	-30%	9%	(6)
O_3	72,6	66,9	76,3	61,7	5%	-8%	(1)
PRE	0,7	9,8	0,5	9,1	-18%	-7%	(11)
PRESS	924,1	15,4	924,2	15,2	0%	-1%	(9)
RADG	1764,1	37,9	1859,0	37,5	5%	-1%	(2)
TEMP	27,7	25,7	27,2	23,1	-2%	-10%	(7)
UR	60,9	28,0	61,7	23,7	1%	-15%	(5)
VV	2,7	23,0	2,5	34,5	-8%	50%	(3)

[1]Ordem de importância das variáveis considerando os dados de 2019 a 2023

Sobre as variáveis relativas aos poluentes no modelo RF cabe notar:

- ◇ Apesar da diminuição do valor médio dos poluentes CO, MP_{10}, NO e NO_2 ao longo dos anos de 2019 a 2023 versus 2010 a 2014, as importâncias de alguns desses poluentes sofrem algumas mudanças, apresentando aumento (diminuição) de acordo com a estação do ano.
- ◇ Aumento da importância dos poluentes CO (+8% no outono, +11% no inverno e +39% na primavera) e NO_2 (+13% no outono, +10% no inverno e +12% na primavera).
- ◇ Aumento da importância do poluente MP_{10} (+88% no inverno e +25% na primavera).
- ◇ O NO apresenta uma diminuição de sua importância para todas as estações.
- ◇ No verão, onde são observadas as menores tendências de aumento do O_3, as variáveis relativas aos poluentes apresentam pouca variação quando se compara os valores de importância dos anos de 2010 a 2014 com os anos de 2019 a 2023, diferentemente do que se observa nas demais estações do ano.

Sobre as variáveis relativas às condições climáticas no modelo RF cabe notar:

- ◇ A radiação solar global apresenta um comportamento distinto entre as estações. No outono e inverno, a radiação solar global apresenta um aumento de sua importância de +15% e +5%, respectivamente. Já no caso da primavera, observa-se uma diminuição da sua importância de −17%. O verão apresenta pouca varição.
- ◇ No outono, inverno e primavera, observa-se um pequeno aumento da importância da temperatura e da umidade relativa quando se compara os resultados dos modelos de 2019 a 2023 versus 2010 a 2014. Já no verão, nota-se uma pequena diminuição da importância da temperatura e da umidade relativa.
- ◇ A velocidade do vento também apresenta um comportamento distinto entre as estações. No outono e inverno, nota-se que a velocidade do vento apresenta uma diminuição da sua importância para o modelo de −8% e −11%, respectivamente. Já no caso da primavera e verão, nota-se um aumento da sua importância de +37% e +50%, respectivamente.

4.4 CONCLUSÃO

Neste capítulo foram avaliados diferentes modelos para estimar a concentração de O_3 a curto prazo. Considerou-se como entrada dos modelos as variáveis climáticas e de

poluição do ar em conjuntos distintos de anos. O modelo RF obteve os melhores resultados, apresentando os menores valores do erro quadrático médio. A partir dos resultados das simulações do RF pode-se concluir que:

- ◇ O outono e o inverno são as estações do ano que apresentam maior tendência de aumento nas concentrações de O_3. Observa-se um aumento de 17,6 $\mu g/m^3$ (+34%) no outono e de 11,7 $\mu g/m^3$ (+24%) no inverno, quando se compara os valores médios dos anos de 2019 a 2023 em relação aos anos de 2010 a 2014. Nessas estações, nota-se um aumento da importância de algumas variáveis climáticas como a radiação solar global, a umidade relativa do ar e a temperatura para o modelo. Além disso, essas variáveis aparecem como as variáveis mais importantes para o modelo. A temperatura e a radiação solar global apresentam aumento da sua importância para o modelo, assim como o valor médio dessas variáveis também cresce ao longo do tempo. A umidade relativa do ar apresenta um comportamento contrário, ela aumenta a sua importância para o modelo, mas observa-se uma diminuição da média da umidade ao longo dos anos, ou seja, o outono e o inverno estão mais secos. Os poluentes CO, MP_{10} e o NO_2 apresentam um aumento da sua importância, apesar de apresentarem diminuição dos seus valores médios. Os resultados observados sugerem que algumas variações das variáveis climáticas, como o aumento da incidência de radiação solar global, o aumento da temperatura e a diminuição da umidade relativa do ar, associadas as concentrações de poluentes como o CO, MP_{10} e NO_2 apresentam uma relação com a tendência de aumento das concentração de O_3 nas estações do outono e inverno.

- ◇ A primavera e o verão são as estações do ano que apresentam menor tendência de aumento nas concentrações de O_3. Observa-se um aumento de 5,9 $\mu g/m^3$ (+8%) na primavera e de apenas 3,7 $\mu g/m^3$ (+5%) no verão, quando se compara os valores médios dos anos de 2019 a 2023 em relação aos anos de 2010 a 2014. Em relação às variáveis climáticas, nota-se que no verão apenas a velocidade do vento apresenta um aumento de sua importância. Já na primavera, a temperatura, a umidade relativa e a velocidade do vento apresentam aumento da sua importância para o modelo. Em relação aos poluentes do ar, nota-se um aumento da importância do CO, MP_{10} e o NO_2 na primavera e do MP_{10} e o NO_2 no verão, apesar de apresentarem diminuição dos seus valores médios. A partir dos resultados observados, pode-se concluir que a menor tendência de aumento nas concentrações de O_3 nas estações da primavera e verão, está relacionada com a menor variação das variáveis climáticas, em especial da radiação solar global e da temperatura.

5 CONCLUSÃO

O trabalho trata da análise da concentração de O_3 medidos na cidade de São Paulo no período de 2007 a 2023. O O_3 é um dos mais importantes poluentes atmosféricos em termos de efeitos nocivos à saúde humana e ao ecossistema terrestre. Além disso, o O_3 apresenta um importante papel na química atmosférica, ele afeta o clima e o clima o afeta. Os níveis de O_3 monitorados em grandes centro urbanos, como no caso de São Paulo, não atendem, em algumas épocas do ano, aos padrões recomendados de qualidade do ar. Diante da problemática ambiental relativa à presença em excesso de O_3 na atmosfera, estudos que permitem modelar a evolução temporal da concentração desse poluente podem ser extremamente úteis à sociedade, porém não é uma tarefa evidente.

Tendo isso em vista, nesse trabalho o enfoque foi dado para analisar a concentração de O_3 em microescala e a curto prazo na cidade de São Paulo. Durante o trabalho, buscou-se compreender o problema, investigar as soluções de modelos para a predição da concentração de O_3 a curto prazo em cenários onde se pode melhor observar a maior tendência de aumento deste poluente ao longo dos anos. Isso tudo com o intuito de possibilitar evidenciar as variáveis de maior significância nos modelos e verificar se a significância das variáveis se alteram em cada modelo ao longo do tempo.

No Capítulo 2, apresentou-se um detalhamento das estações meteorológicas consideradas no trabalho e as variáveis escolhidas. A estação Parque D. Pedro II da CETESB foi escolhida para a análise da evolução do níveis de O_3 ao longo dos anos. Além do O_3, foram consideradas outras variáveis relacionadas à concentração de poluentes do ar medidas nessa estação. Dados complementares, monitorados pelo INMET, foram utilizados com o intuito de associar as tendências observadas das concentrações de O_3 às principais variáveis climáticas capazes de controlar o seu processo de formação e de remoção da atmosfera. Através da análise de dados faltantes, constatou-se que esse é um problema marcante nos dados da CETESB. Por conta disso, considerou-se períodos de tempo específicos de acordo com a disponibilidade dos dados medidos e as análises que foram empregues sobre eles.

No Capítulo 3, fez-se a análise exploratória dos dados a partir da qual foi possível observar uma tendência de aumento na concentração de O_3 ao longo do tempo. Além disso, foi possível verificar os diferentes cenários em que essa tendência de aumento é mais notável. As análises evidenciaram variações estatísticas distintas dos dados conforme as estações do ano e as horas do dia. As maiores tendências de aumento foram observadas no outono e inverno. Em relação às horas do dia, notou-se uma tendência maior de aumento nos horários das 12 às 17 horas, período do dia em que são observadas as maiores concentrações horárias deste poluente. Além disso, constatou-se que algumas variáveis climáticas fortemente correlacionadas com o O_3 também apresentam variações ao longo do tempo. Em especial, mostrou-se variações distintas da radiação solar global e da temperatura conforme as

estações do ano e as horas do dia.

No Capítulo 4, aplicaram-se os métodos OLS, LASSO, CART e RF para estimar a concentração de O_3 a curto prazo. Usualmente, ao trabalhar com algoritmos de aprendizado de máquina, há duas etapas: a etapa de treinamento e a etapa de teste. Entretanto, no caso deste trabalho, que visa obter um modelo preciso o suficiente que possibilite verificar a influência das variáveis de entrada na concentração O_3 em conjuntos distintos de anos, utilizou-se apenas a etapa de treinamento. Para a realização das simulações computacionais foram considerados os dados das 12 às 17 horas, agrupados em cada estação do ano, separadamente para conjunto de anos de 2010 a 2014 e de 2019 a 2023. Os resultados mostram que suposições como linearidade, feitas por alguns modelos usais, podem ser muito restritivas para predizer com muita precisão as concentrações de O_3 a curto prazo. O modelo RF obteve os melhores resultados para a predição do O_3 segundo as métricas de avaliação dos erros. A partir da análise das variáveis de maior significância para o modelo RF, destacam-se que as maiores variações nas variáveis climáticas são observadas no outono e no inverno. Em especial a radiação solar global e a temperatura. Além disso, como observado nas simulações, quando as variáveis climáticas interagem com os poluentes CO, MP_{10} e NO_2, nota-se uma maior tendência de aumento da concentração de O_3.

Vislumbra-se na continuidade deste trabalho considerar os aspectos descritos a seguir.

- Verificar a significância das variáveis considerando as 24 horas do dia.
- Ampliar geograficamente a validade do procedimento aplicado nesse trabalho e comparar os resultados de regiões com características diferentes. Especificamente, para realizar esse passo é necessário ter ausência significativa de dados faltantes. Além disso, é necessário conhecer os dados de outras estações de coleta de dados da Região Metropolitana de São Paulo e do Estado em um mesmo intervalo de tempo e com poucos dados faltantes.
- Supondo que existem dados disponíveis em vários pontos de coleta de dados. Repetir o mesmo procedimento feito na dissertação para cada ponto de coleta. Posteriormente, então, levar em conta técnicas de processamento distribuído e observar a influência de forma conjunta das condições meteorológicas e de poluentes de várias estações de aquisição de dados na concentração de O_3.
- Realizar uma avaliação de causalidade entre as variáveis climáticas e de poluição do ar com as concentrações de O_3 e identificar as relações de causa e efeito entre esses fatores. O objetivo é determinar como diferentes condições climáticas (como temperatura, umidade e radiação solar) e níveis de poluentes atmosféricos influenciam as concentrações de O_3. Esse tipo de análise pode envolver técnicas estatísticas e de

modelagem para discernir a direção e a magnitude dessas influências, ajudando a compreender os mecanismos subjacentes que regulam a formação e variação do O_3.

REFERÊNCIAS BIBLIOGRÁFICAS

ALLU, S. et al. Seasonal ground level ozone prediction using multiple linear regression (mlr) model. *Modeling Earth Systems and Environment*, 2020.

ATKINSON, R. Atmospheric chemistry of vocs and nox. *Atmospheric Environment*, v. 34, n. 12, p. 2063 – 2101, 2000. ISSN 1352-2310. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1352231099004604>.

BARTZ, E. et al. (Ed.). *Hyperparameter Tuning for Machine and Deep Learning with R - A Practical Guide*. Springer, 2023. ISBN 978-981-19-5170-1. Disponível em: <https://doi.org/10.1007/978-981-19-5170-1>.

BDMEP. *Banco de Dados Meteorológicos do INMET*. 2023. Disponível em: <https://bdmep.inmet.gov.br/>.

BREIMAN, L. Random forests. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001.

BREIMAN, L. et al. *Classification and Regression Trees*. [S.l.]: Chapman and Hall/CRC, 1984.

CETESB. *Relatório de Qualidade do Ar no Estado de São Paulo*. 2022. Disponível em: <https://cetesb.sp.gov.br/ar/publicacoes-relatorios/>.

CETESB. *Companhia Ambiental do Estado de São Paulo*. 2023. Disponível em: <https://cetesb.sp.gov.br/>.

CHATTOPADHYAY, G.; MIDYA, S. K.; CHATTOPADHYAY, S. Mlp based predictive model for surface ozone concentration over an urban area in the gangetic west bengal during pre-monsoon season. *Journal of Atmospheric and Solar-Terrestrial Physics*, v. 184, p. 57 – 62, 2019. ISSN 1364-6826. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1364682618305509>.

FOWLER, D. et al. *Ground-level ozone in the 21st century: future trends, impacts and policy implications*. Rs1276. United Kingdom: The Royal Society, 2008. v. 15/08. (Royal Society Policy Document 15/08, v. 15/08). ISBN 978-0-85403-713-1.

FRIEDMAN, J.; TIBSHIRANI, R.; HASTIE, T. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, v. 33, n. 1, p. 1–22, 2010.

FURG. *Banco de Dados Meteorológicos*. 2023. Disponível em: <http://www.meteorologia.furg.br/bancodedados.html>.

GABRIC, A. J. The climate change crisis: A review of its causes and possible responses. *Atmosphere*, v. 14, n. 7, 2023. ISSN 2073-4433. Disponível em: <https://www.mdpi.com/2073-4433/14/7/1081>.

GREENWELL, B. M.; BOEHMKE, B. C. Variable importance plots—an introduction to the vip package. *The R Journal*, v. 12, n. 1, p. 343–366, 2020. Disponível em: <https://doi.org/10.32614/RJ-2020-013>.

GUEYMARD, C. A. The sun's total and spectral irradiance for solar energy applications and solar radiation models. *Solar Energy*, v. 76, n. 4, p. 423–453, 2004. ISSN 0038-092X. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0038092X03003967>.

HANSEN, J. E. et al. Global warming in the pipeline. *Oxford Open Climate Change*, v. 3, n. 1, p. kgad008, 11 2023. ISSN 2634-4068. Disponível em: <https://doi.org/10.1093/oxfclm/kgad008>.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The elements of statistical learning: data mining, inference and prediction*. 2. ed. Springer, 2009. Disponível em: <http://www-stat.stanford.edu/~tibs/ElemStatLearn/>.

HASTIE, T.; TIBSHIRANI, R.; WAINWRIGHT, M. *Statistical Learning with Sparsity: The Lasso and Generalizations*. [S.l.]: Chapman & Hall/CRC, 2015. ISBN 1498712169.

HAYKIN, S. *Neural Networks and Learning Machines*. [S.l.]: Prentice Hall, 2009. (Neural networks and learning machines, v. 10). ISBN 9780131471399.

HORTA, P. et al. Brazil fosters fossil fuel exploitation despite climate crises and the environmental vulnerabilities. *Marine Policy*, v. 148, p. 105423, 2023. ISSN 0308-597X. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0308597X22004705>.

IEMA. *Instituto de Energia e Meio Ambiente*. 2022. Disponível em: <https://energiaeambiente.org.br>.

INMET. *Instituto Nacional de Meteorologia*. 2023. Disponível em: <https://portal.inmet.gov.br/>.

IZBICKI, R.; SANTOS, T. M. dos. *Aprendizado de máquina: uma abordagem estatística*. [S.l.: s.n.], 2020. ISBN 978-65-00-02410-4.

JAMES, G. et al. *An Introduction to Statistical Learning: With Applications in R*. [S.l.]: Springer Publishing Company, Incorporated, 2014. ISBN 1461471370, 9781461471370.

KUHN, M. Building predictive models in r using the caret package. *Journal of Statistical Software*, v. 28, n. 5, p. 1–26, 2008. Disponível em: <https://www.jstatsoft.org/index.php/jss/article/view/v028i05>.

LELIEVELD, J. et al. Air pollution deaths attributable to fossil fuels: observational and modelling study. *BMJ*, BMJ Publishing Group Ltd, v. 383, 2023. Disponível em: <https://www.bmj.com/content/383/bmj-2023-077784>.

LIU, P.-W. G. et al. Establishing multiple regression models for ozone sensitivity analysis to temperature variation in taiwan. *Atmospheric Environment*, v. 79, p. 225 – 235, 2013. ISSN 1352-2310. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1352231013004561>.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2017. (NIPS'17), p. 4768–4777. ISBN 9781510860964.

MMA. *Guia Técnico para o Monitoramento e avaliação da qualidade do ar*. 2019. Disponível em: <https://www.gov.br/mma/pt-br/centrais-de-conteudo/mma-guia-tecnico-qualidade-do-ar-pdf>.

MONTGOMERY, D.; RUNGER, G. *Estatística Aplicada e Probabilidade para Engenheiros*. 7. ed. ed. [S.l.]: GEN LTC, 2021. ISBN 9788521637332.

NOAA. *Meteorological Versus Astronomical Seasons*. 2024. Disponível em: <https://www.ncei.noaa.gov/news/meteorological-versus-astronomical-seasonsl>.

OECD. *OECD Environmental Outlook to 2050*. [S.l.]: Organisation for Economic Co-operation and Development, 2012. (Technical documents).

ORGANIZATION, W. H. Publications. *WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide: executive summary*. [S.l.]: World Health Organization, 2021. 10 p. p.

PEARSON, K. Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity, and Panmixia. *Philosophical Transactions of the Royal Society of London Series A*, v. 187, p. 253–318, Jan 1896.

PROBST, P.; WRIGHT, M. N.; BOULESTEIX, A. Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery*, Wiley, v. 9, n. 3, jan. 2019. ISSN 1942-4795. Disponível em: <http://dx.doi.org/10.1002/widm.1301>.

QUALAR. *Sistemas de informações da qualidade do ar*. 2023. Disponível em: <https://cetesb.sp.gov.br/ar/qualar/>.

REN, X.; MI, Z.; GEORGOPOULOS, P. G. Comparison of machine learning and land use regression for fine scale spatiotemporal estimation of ambient air pollution: Modeling ozone concentrations across the contiguous united states. *Environment International*, v. 142, p. 105827, 2020. ISSN 0160-4120. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0160412020317827>.

RITCHIE, H. et al. Population growth. *Our World in Data*, 2023. Https://ourworldindata.org/population-growth.

ROSS, S. M. *Introduction to Probability and Statistics for Engineers and Scientists*. Sixth edition. [S.l.]: Academic Press, 2021. ISBN 978-0-12-824346-6.

SOUSA, S. et al. Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. *Environmental Modelling and Software*, v. 22, p. 97–103, 2007.

SOUZA, C. B. de. *Árvores de Decisão: A Evolução do CART ao BART*. Dissertação (Mestrado) — Universidade de São Paulo, 2021.

SPOKAS, K. Estimating hourly incoming solar radiation from limited meteorological data. *Weed Science*, v. 54, p. 182–189, 01 2009.

STEKHOVEN, D. J.; BüHLMANN, P. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, v. 28, n. 1, p. 112–118, 10 2011. ISSN 1367-4803. Disponível em: <https://doi.org/10.1093/bioinformatics/btr597>.

STOCKWELL, W. et al. Ozone formation, destruction and exposure in europe and the united states. 01 1997.

VERMA, N. et al. Prediction of ground level ozone concentration in ambient air using multiple regression analysis. *Journal of Chemical, Biological and Physical Sciences*, v. 55, p. 3685–3696, 09 2015.

WHO. *Air pollution and child health: prescribing clean air: summary*. [S.l.]: World Health Organization, 2018. (Technical documents).

WICKHAM, H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. Disponível em: <https://ggplot2.tidyverse.org>.

WILKE, C. *Fundamentals of Data Visualization: A primer on making informative and compelling figures*. [S.l.]: O'Reilly Media, Inc., 2019. ISBN 1492031089.

ZEILEIS, A.; HOTHORN, T.; HORNIK, K. Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, v. 17, n. 2, p. 492–514, 2008.

APÊNDICE A – CONCEITOS BÁSICOS

A.1 AJUSTE DE PONTOS A UMA RETA COM O MÉTODO DOS MÍNIMOS QUADRADOS

Deseja-se ajustar a reta $y = ax + b$ aos pontos do conjunto

$$\{(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n)\}.$$

Somente no caso em que esses pontos são colineares, a reta passa exatamente por todos os n pontos e os coeficientes a e b satisfazem $y_i = ax_i + b$, para $i = 1,\ \ldots,\ n$. Evidentemente, neste caso $y_i - ax_i - b = 0$. Se os pontos não são colineares o sistema de equações representado por $y_i = ax_i + b$ para $i = 1,\ \ldots,\ n$, é inconsistente, ou seja, é impossível determinar os coeficientes a e b que satisfaçam $y_i - ax_i - b = 0$. Visando encontrar a melhor solução para o caso inconsistente, representa-se a diferença y_i e $ax_i + b$ como $e_i = y_i - b - x_i a$, para $i = 1,\ \ldots,\ n$. Assim, o método dos mínimos quadrados tem como objetivo obter os coeficientes a e b de modo a assegurar a menor soma dos quadrados dos erros representados na função

$$\mathbb{J} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2, \tag{A.1}$$

Derivando $\mathbb{J}$ em relação à cada aos coeficientes a e b, obtém-se

$$\begin{array}{rcccc} \dfrac{\partial \sum_{i=1}^{n} e_i^2}{\partial b} & = & 2\sum_{i=1}^{n} e_i \dfrac{\partial e_i}{\partial b} & = -2\sum_{i=1}^{n} e_i x_i & = \sum_{i=1}^{n} (y_i - b - x_i a) x_i \\ \dfrac{\partial \sum_{i=1}^{n} e_i^2}{\partial a} & = & 2\sum_{i=1}^{n} e_i \dfrac{\partial e_i}{\partial a} & = -2\sum_{i=1}^{n} e_i & = \sum_{i=1}^{n} (y_i - b - x_i a) \end{array} . \tag{A.2}$$

Igualando cada equação de (A.2) a zero tem-se

$$\begin{array}{c} \sum_{i=1}^{n} (y_i - b^* - x_i a^*) x_i = 0 \\ \sum_{i=1}^{n} (y_i - b^* - x_i a^*) = 0 \end{array} \tag{A.3}$$

Portanto, os coeficientes a e b que satisfazem (A.3), denotados como a^* e b^*, representam a solução dos mínimos quadrados. Após manipulações algébricas essa solução pode ser expressa como

$$\begin{bmatrix} b^* \\ a^* \end{bmatrix} = \left(\sum_{i=1}^{n} \begin{bmatrix} 1 & x_i \\ x_i & x_i^2 \end{bmatrix} \right)^{-1} \left(\sum_{i=1}^{n} \begin{bmatrix} 1 \\ x_i \end{bmatrix} y_i \right). \tag{A.4}$$

A.2 CORRELAÇÃO

O coeficiente de Correlação Linear de Pearson (PEARSON, 1896), (MONTGOMERY; RUNGER, 2021) é uma medida utilizada para medir o grau e o sinal da correlação linear entre duas variáveis. Considere (X,Y) um par de variáveis aleatórias distribuídas conjuntamente, calcula-se o coeficiente de correlação de Pearson segundo a fórmula

$$\rho = \frac{Cov(X,Y)}{\sqrt{Cov(X,X)}\sqrt{Cov(Y,Y)}}, \tag{A.5}$$

também na forma

$$\rho = \frac{\sum_{i=i}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=i}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=i}^{n}(y_i - \bar{y})^2}}, \tag{A.6}$$

em que n é igual ao número de elementos na amostra, x_i é o i-ésimo elemento da variável X e $\bar{x} = \sum_{i=1}^{n} x_i/n$ o seu valor médio (analogamente para y_i e $\bar{y}$).

O Coeficiente de Correlação Linear de Pearson é uma medida adimensional que varia no intervalo $-1 \leq \rho \leq +1$. A vantagem de ser adimensional está no fato de que o seu valor não é afetado pelas unidades de medidas adotadas. Para $\rho = -1$ temos o caso de uma correlação linear negativa perfeita e $\rho = +1$ temos o caso de uma correlação linear positiva perfeita, e quando zero, indicando que não existe uma relação linear entre as variáveis.

A.3 HISTOGRAMA

O histograma, também conhecido como distribuição de frequências, é uma representação gráfica em barras (retângulos) de um conjunto de dados previamente tabulados e divididos em classes (também denominadas como "*bins*"). A base de cada retângulo representa uma determinada classe e a sua altura representa a frequência absoluta dos valores da classe presentes no conjunto de dados (WILKE, 2019), (MONTGOMERY; RUNGER, 2021).

Este tipo de gráfico permite avaliar a distribuição dos dados e identificar características que devem ser levadas em consideração no processo de modelagem. Na Figura 39 é apresentado o histograma da concentração de ozônio na estação Parque D. Pedro II. A partir do histograma pode-se observar uma certa assimetria da distribuição da variável em estudo, com a distribuição alongada para a direita (também denominada positivamente assimétrica).

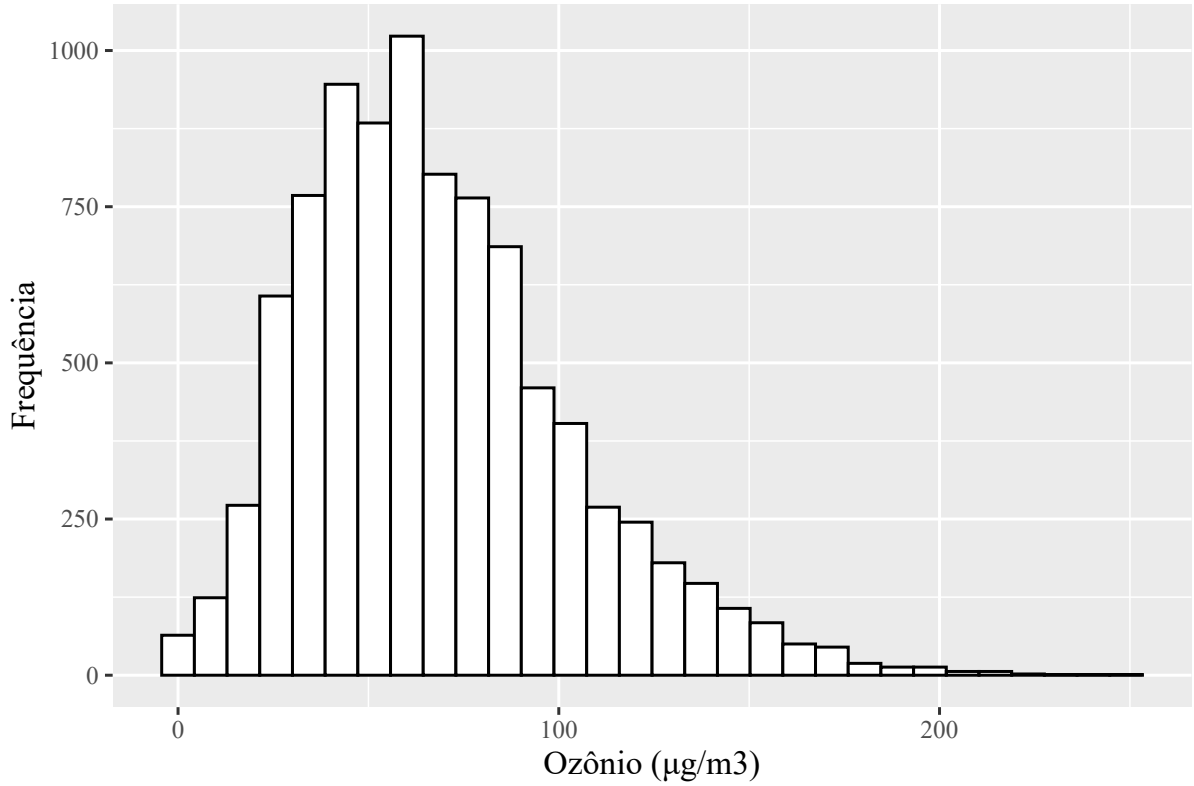

Figura 39 – Histograma da concentração de ozônio observada na estação Parque D. Pedro II no período da tarde (12h00 às 17h00). Valores observados para o período de 2019 a 2023.

A.4 BOXPLOT

O boxplot, também conhecido como diagrama de caixa, é um método tradicional e comumente usado para visualizar distribuições (WILKE, 2019), (MONTGOMERY; RUNGER, 2021). O boxplot divide os dados em quartis e os visualiza de maneira padronizada, conforme ilustrado na Figura 40. Um boxplot é composto pelos seguintes elementos:

- A linha que divide a caixa em 2 partes representa a mediana dos dados;
- As extremidades da caixa mostram os quartis superior (Q_3) e inferior (Q_1) que representam, respectivamente, 75% e 25% das observações;
- A diferença entre os quartis Q_3 e Q_1 é chamada de intervalo interquartil (IQR);
- A linha extrema mostra $Q_3 + 1{,}5(\text{IQR})$ a $Q_1 - 1{,}5(\text{IQR})$ (representa o valor máximo e mínimo excluindo os outliers);
- Pontos (ou outros marcadores) além da linha extrema mostram potenciais outliers (valores discrepantes).

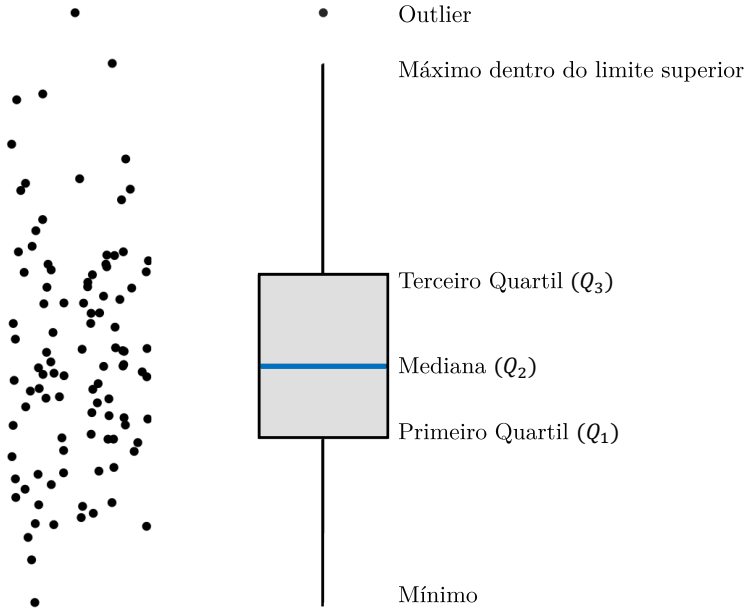

Figura 40 – Anatomia de um boxplot. É mostrada uma nuvem de pontos (à esquerda) e o correspondente boxplot (à direita). Figura adaptada de (WILKE, 2019).

A.5 VIOLIN PLOT

Um gráfico de violino (do inglês, *violin plot*) é um método de representação de dados numéricos. É semelhante a um boxplot, com a adição de um gráfico de densidade do kernel giradas em 90 graus e depois espelhadas em cada lado (WILKE, 2019). Este tipo de gráfico é mais informativo que um boxplot pois ele mostra a distribuição completa dos dados. A diferença é particularmente útil quando a distribuição de dados é multimodal (mais de um pico), podendo avaliar sua posição e amplitude relativa.

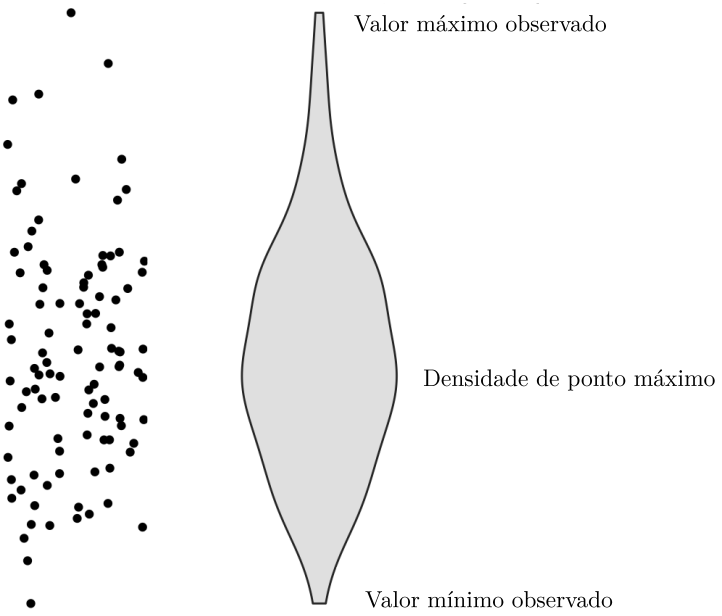

Figura 41 – Anatomia de um gráfico de violino. É mostrada uma nuvem de pontos (à esquerda) e o correspondente gráfico de violino (à direita). Figura adaptada de (WILKE, 2019).

APÊNDICE B – VALIDAÇÃO CRUZADA PARA ESTIMAÇÃO DO TERMO DE PENALIZAÇÃO DO LASSO

O método consiste em criar novas amostras artificiais a partir da base de dados original para, através de divisões aleatórias, fazer a estimação da performance do modelo em cada uma das amostras sob diferentes valores de λ e encontrar aquele valor que maximiza a performance do modelo.

Na validação cruzada *k-fold* a base de dados é dividida aleatoriamente em $k > 1$ partes com praticamente a mesma quantidade de observações e sem intersecção. Um destes k grupos é fixado como teste, para validar a performance do modelo, e os demais $k-1$ grupos são usados para estimação do modelo. Primeiro devemos construir o modelo usando o conjunto de treinamento, os $k-1$ grupos que restaram, para diferentes valores de λ. O grupo que ficou para teste é usado para medir o desempenho do modelo, por exemplo, estimar o Erro Quadrático Médio (MSE) de predição. O processo é então repetido k vezes até que todas as k partes tenham participado tanto do treinamento quanto da validação do modelo (HASTIE; TIBSHIRANI; FRIEDMAN, 2009), (JAMES et al., 2014). Apresenta-se na Figura 42 uma esquematização do processo de validação cruzada *k-fold*, para k= 10.

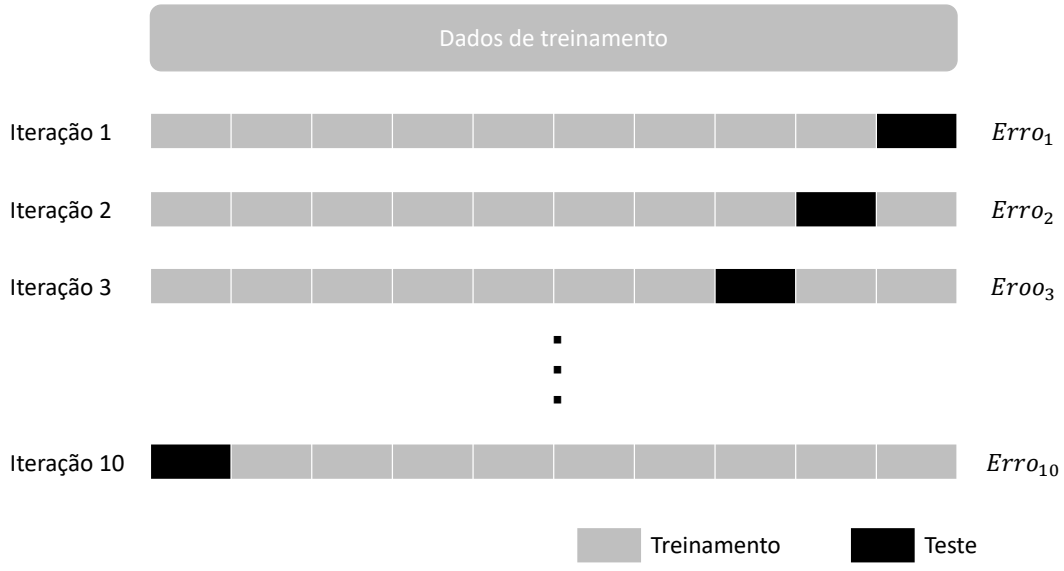

Figura 42 – Processo de validação cruzada *k-fold*, para k= 10.

Finalmente, para cada valor de λ dentro do intervalo considerado teremos k estimativas do MSE. Após isso, a média do MSE é calculada para cada valor de λ e o gráfico da curva de erro de validação cruzada é construído. Esse gráfico apresenta os valores de λ no eixo x e a média do MSE no eixo y. Por meio desse gráfico é possível observar qual o valor do parâmetro de penalização que maximiza a performance do modelo, isto é, qual o valor de λ que minimiza o MSE.

Na Figura 43 é apresentada a curva de erro da validação cruzada construída através do pacote `GLMNET` (FRIEDMAN; TIBSHIRANI; HASTIE, 2010) (no pacote `GLMNET` é plotado o logaritmo de λ). Este exemplo foi construído a partir da base de dados da estação do outono, considerando os dados dos anos de 2010 a 2014. No gráfico existem dois pontos destacados como melhor λ. O primeiro representa aquele do qual provém o menor MSE e o outro fornece o modelo mais regularizado (com menos variáveis), que apresenta MSE menor que o do modelo com menor MSE mais um erro padrão.

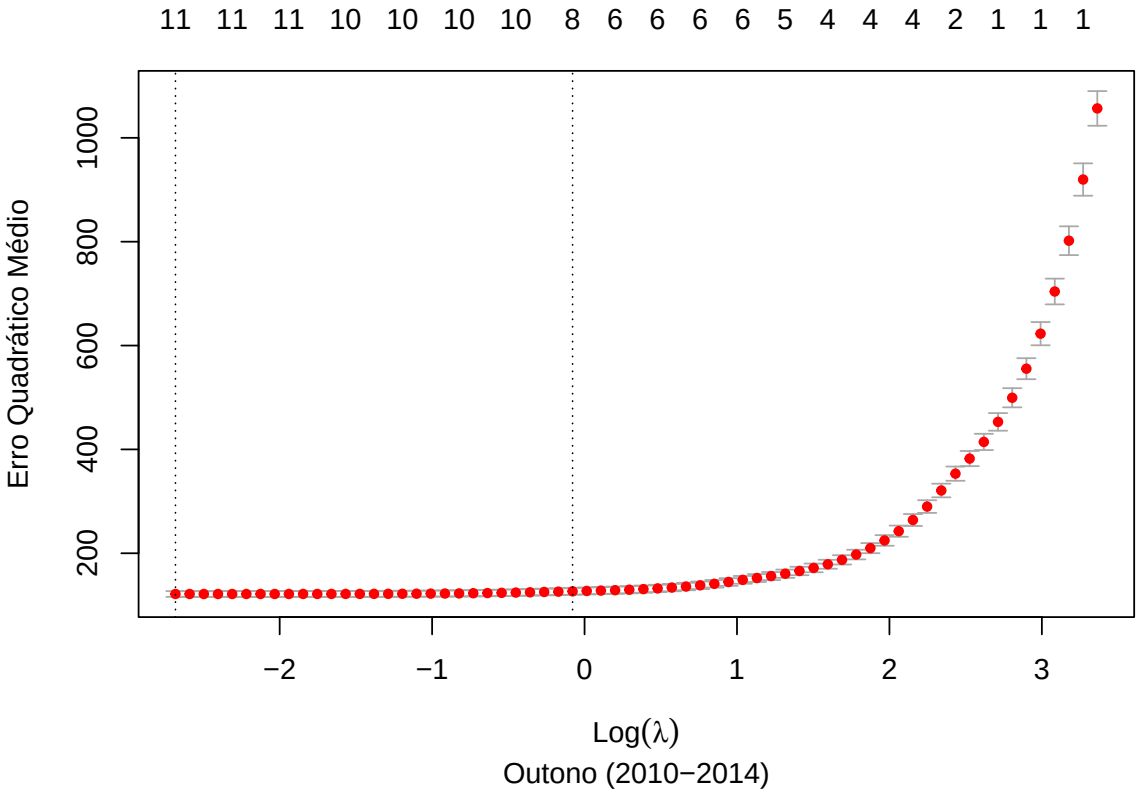

Figura 43 – Gráfico da curva de erro da validação cruzada feita através do pacote `GLMNET`.

APÊNDICE C – ROTINAS UTILIZADAS PARA CONSTRUÇÃO DOS MODELOS CART E RF

CONSTRUÇÃO DO MODELO CART

Na construção do modelo CART, deve-se escolher como um dos parâmetros de entrada do modelo o número mínimo de observações em cada nó. Esse parâmetro é utilizado como um critério de parada na construção da árvore. A escolha adequada desse parâmetro pode ser feita através do método *Grid Search*, que corresponde a uma pesquisa exaustiva por meio de um subconjunto especificado manualmente onde são avaliados os valores do erro do modelo associados a escolha de cada um dos parâmetros definidos. Para isso, utilizou-se a função `expand.grid()` e definiu-se todos os valores para serem avaliados.

```
1        grid_simples <- expand.grid(nodesize = c(1,2,3,4,5,6,7,8,9,10))
```

A construção do modelo CART foi realizada utilizando o pacote `RPART` do software R. Para automatizar o processo de modelagem, configurou-se um *for loop* para iterar o processo de construção de cada árvore ao número de observações definidas no *grid*. A função `rpart` treina o modelo com os dados contidos em cada um dos *data frames*. Para a aplicação da função são definidos os seguintes argumentos:

- `CONTROL`: argumento que controla a lista dos parâmetros avaliados no modelo.
- `METHOD`: argumento que controla o tipo de árvore (classificação ou regressão). Deve-se definir o argumento `METHOD = "anova"` para árvores de regressão.
- `DATA`: conjunto de dados utilizado. São os *data frames* ($\mathbf{DF}_s$).
- `FORMULA`: especificação da variável resposta e do conjunto de variáveis preditoras.

```
1        lista_cart <- list()
2
3        for (i in 1:nrow(grid_simples)) {
4
5                minsplit <- grid_simples$nodesize[i]
6
7                models[[i]] <- rpart(
8                formula = Y ~ O3 + MP10 + NO + NO2 + CO + RADG + TEMP + UR +
9                VV + PRESS + PRE,
10               data     = DF1,
11               method   = "anova",
12               control = list(minsplit = minsplit)
13               )
14       }
```

Como saída da função, obtêm-se uma lista de todos os modelos possíveis, com os valores preditos da variável desejada para cada um dos hiperparâmetros avaliados. A partir disso, calculou-se o valor do MSE para cada um dos hiperparâmetros e selecionou-se o hiperparâmetro que gera o menor MSE. A partir dessa escolha, ajustou-se o modelo final CART para cada um dos $\mathbf{DF}_s$. A importância das variáveis para os modelos foi calculada através do pacote `VIP` utilizando-se a função `vi`.

CONSTRUÇÃO DO MODELO RF

Para a construção do modelo RF utilizou-se o pacote `CARET` do software *R*. Esse pacote funciona como uma interface utilizada em diversos outros pacotes para construção de modelos de aprendizado de máquina. As principais funções utilizadas nesse pacote são: `train()` e `trainControl()`.

1. A função `train()` é responsável pelo treinamento do modelo. Nela, são indicados os dados utilizados para o treinamento, a variável resposta e as variáveis explicativas, o método a ser utilizado (no caso o método utilizado é o "rf" de *Random Forest*), e os parâmetros do método que serão testados.

2. A função `trainControl()` é utilizada para indicar a utilização de validação cruzada, a quantidade de subconjuntos (*k-folds*) que serão utilizados e a quantidade de repetições realizadas.

Com o objetivo de investigar a melhor combinação de parâmetros que leva a melhores resultados na predição das concentrações de ozônio, utilizou-se o método *Grid Search* e definiu-se a lista de valores avaliados para o modelo.

```
1        grid_completo <- expand.grid(mtry = c(1,2,3,4,5,6,7,8,9,10,11),
2        ntree = c(300,400,500,600,700,800,900,1000),
3        nodesize = c(1,2,3,4,5,6,7,8,9,10))
```

O *grid* definido possui 880 diferentes combinações que serão avaliados utilizando a técnica de validação cruzada (*k-fold* = 10). No processo de validação cruzada, o conjunto de dados é dividido em 10 partes iguais e em cada umas dessas partes todos os parâmetros são avaliados. É escolhido o conjunto de parâmetros que melhor performa, segundo alguma função objetivo, o maior número de vezes dentro das 10 partições. Para isso, define-se a função `trainControl()`. Nela, é indicado que será realizada uma validação cruzada com repetições (`method = "repeatedcv"`), na qual serão considerados 10 subconjuntos em cada uma das três validações cruzadas. O argumento `seeds` define a lista de sementes aleatórias utilizadas nas iterações do processo de validação cruzada de forma a possibilitar a reprodutibilidade dos resultados obtidos. Os resultados finais que são apresentados nesta seção utilizam a `seeds = 104369`.

```
fit_control <- trainControl(method = "repeatedcv",
number = 10,
repeats = 3,
seeds = seeds)
```

Finalmente, define-se a função que efetivamente treina o modelo para cada combinação de parâmetros escolhidos. Dessa forma, essa função recebe como o argumento cada um dos parâmetros do *grid* e os dados de entrada.

```
rf_cv <- function(mtry_val, ntree_val, nodesize_val) {

        mtry_temp <- data.frame(.mtry = mtry_val)

        set.seed(123456)

        rf_fit <- train(Y ~ O3 + MP10 + NO + NO2 + CO + RADG +
                            TEMP + UR + VV + PRESS + PRE,
        data = DF1,
        method = "rf",
        trControl = fit_control,
        tuneGrid = mtry_temp,
        ntree = ntree_val,
        nodesize = nodesize_val)

        resultados <- rf_fit$results %>%
        mutate(ntree = ntree_val,
        nodesize = nodesize_val) %>%
        select(mtry, ntree, nodesize, MAE)

        return(resultados)
}
```

Por conta do grande número de combinações dos parâmetros avaliados, o processo de construção do RF foi feito de forma paralelizada para melhor performance de tempo computacional. Na abordagem paralelizada, vários núcleos do computador trabalham simultaneamente, isto é, cada um deles fica responsável por um conjunto de combinações. Para isso, utilizou-se o pacote `FURRR` e a função `future_pmap_dfr`.

```
furrr_df <- future_pmap_dfr(grid_completo, rf_cv)
```

Como saída da função, obtêm-se uma lista com a combinação de todos os hiperparâmetros e os valores já calculados do MAE. A partir dessa lista, escolheu-se os hiperparâmetros que geram o menor MAE e ajustou-se os modelos finais do RF para cada um dos $\mathbf{DF}_s$. A importância das variáveis para os modelos foi calculada através do pacote `VIP` utilizando-se a função `vi`.