ENZO REI FERREIRA
GABRIEL BOECHAT CELANI ZWIRMAN

ANÁLISE ESTATÍSTICA SOBRE A POLUIÇÃO DO AR NA CIDADE DO RIO DE JANEIRO

PROJETO DE GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO APRESENTADO AO DEPARTAMENTO DE ENGENHARIA INDUSTRIAL DA PUC-RIO, COMO PARTE DOS REQUISITOS PARA OBTENÇÃO DO TÍTULO DE ENGENHEIRO DE PRODUÇÃO

Orientadora: Soraida Aguilar

Departamento de Engenharia Industrial
Rio de Janeiro, 17 de novembro de 2023

RESUMO

A qualidade do ar é um fator de extrema importância a ser considerado nos dias atuais, sobretudo nos grandes centros urbanos, como é o caso da cidade do Rio de Janeiro, foco da pesquisa em questão. Alguns poluentes emitidos nesses centros são responsáveis pela geração de um grande número de doenças e complicações para os seres humanos. Dessa forma, este trabalho teve como objetivo analisar o comportamento dos poluentes PM10, PM2.5, NO2 e O3 ao longo de 9 anos em suas estações de medidas espalhadas pela cidade do Rio de Janeiro. Com isso, utilizou-se uma série de técnicas com a finalidade de constatar algum padrão de comportamento das medidas dos poluentes. Primeiramente implementou-se uma análise descritiva extraindo-se medidas de localização/posição e dispersão dos dados, bem como uma análise gráfica dos poluentes com o objetivo de analisar de forma tanto visual como numérica quais seriam os possíveis padrões de comportamento desses dados levando em conta tanto as estações de medidas, bem como os anos. Para isto implementou-se o ciclo de vida de um projeto de Ciência de Dados em que a modelagem foi efetuada através de uma análise ANOVA não paramétrica. Adicionalmente, foram aplicados testes estatísticos que forneceram um embasamento para que fosse possível concluir que os dados das medidas apresentam um comportamento consideravelmente distinto entre as estações de medidas. Esse fator pode ocorrer devido a diversas questões urbanas de localização das estações. Entretanto, quando os anos eram considerados como fator, notou-se uma similaridade no comportamento dos dados medidos, indicando que os anos tiveram pouca influência no padrão de medida dos dados.

PALAVRAS-CHAVE

Poluentes, PM10, PM2.5, NO2, O3, Estação, Ano.

ABSTRACT

The air quality is an extremely important factor to be considered nowadays, especially in large urban centers, such as the city of Rio de Janeiro, focus of this research. Some pollutants emitted in these centers are responsible for the generation of a large number of diseases and complications for humans. Thus, this study aimed to analyze the behavior of pollutants PM10, PM2.5, NO2, and O3 over a period of 9 years at measurement stations scattered throughout the city of Rio de Janeiro. Various techniques were used to identify any pattern in the behavior of pollutant measurements. Initially, a descriptive analysis was implemented, extracting measures of location/position and data dispersion, along with a graphical analysis of pollutants. This aimed to visually and numerically analyze possible patterns in the data, considering both measurement stations and years. Thus, the life cycle of a Data Science project was implemented in which the modeling was carried out through a non-parametric ANOVA analysis. Additionally, statistical tests were applied that provided a basis for concluding that the measurement data show considerably different behavior between the measurement stations. This difference may be due to various urban location-related issues of the stations. However, when the years were considered as a factor, a similarity in the behavior of the measured data was observed, indicating that the years had little influence on the pattern of data measurement.

KEY WORDS

Pollutants, PM10, PM2.5, NO2, O3, Station, Year.

SUMÁRIO

LISTA DE TABELAS

LISTA DE IMAGENS

1. INTRODUÇÃO

A qualidade do ar é um fator crítico e fundamental para a saúde pública e o bem-estar das comunidades urbanas em todo o mundo. Além disso, o tema está cada vez mais em pauta devido à conscientização sobre a preservação do meio ambiente. São diversos os estudos que comprovam e relacionam a poluição do ar com o agravamento de doenças nos seres humanos e com a degradação da natureza.

Fatores como o crescimento populacional ainda levam a crer que a qualidade do ar pode estar cada vez mais comprometida quando o fluxo de automóveis está cada vez mais intenso e a produção industrial cada vez mais acelerada para atender a crescente demanda, principalmente nos grandes centros urbanos. Dentre estes grandes centros, situa-se a cidade do Rio de Janeiro. Centro cultural da humanidade e com uma alta densidade demográfica – cerca de 367 habitantes por quilômetro quadrado em 2022 segundo o IBGE – o município será foco da pesquisa em questão e local em que a mesma foi realizada.

Nas últimas décadas, o Rio de Janeiro ainda tem testemunhado um notável crescimento em sua população e urbanização que segundo Loureiro (2023) e o Jornal G1, em 12 anos, o aumento foi de 0,40%, o que representa em números absolutos mais 64.595 pessoas em relação ao censo anterior, de 2010. No entanto, esse desenvolvimento tem trazido consigo um desafio crescente e premente: a poluição do ar. Estudos anteriores, como o de Samet e Krewski (2007), demonstraram os impactos adversos de exposições prolongadas à poluição do ar na saúde da população.

O Rio de Janeiro é conhecido por suas belezas naturais, praias deslumbrantes e uma geografia única, caracterizada por morros e baías. No entanto, segundo o INEA (2018), através do Relatório da qualidade do Ar do Rio de Janeiro, essa mesma topografia singular, aliada a uma frota considerável de veículos e atividades industriais, têm contribuído para a concentração de poluentes atmosféricos, como partículas em suspensão (PM2.5 e PM10), dióxido de nitrogênio (NO2) e ozônio troposférico (O3). Os congestionamentos nas vias urbanas, especialmente em horários de pico, exacerbam essa situação, afetando a qualidade do ar e, por conseguinte, a saúde da população

Além do tráfego, as atividades industriais na região da Baía de Guanabara desempenham um papel significativo na emissão de poluentes atmosféricos e resíduos químicos, afetando tanto a atmosfera quanto os ecossistemas aquáticos circundantes. Um estudo de Miranda et al. (2006) fornece uma análise da composição química das partículas em suspensão (PM2.5) na cidade do Rio de Janeiro, mostrando que o impacto ambiental dessas atividades industriais é uma preocupação crescente que requer uma análise aprofundada e baseada em dados.

Diante disso, este trabalho busca lançar luz sobre a complexa dinâmica da poluição atmosférica na cidade do Rio de Janeiro. Através da análise do dados de emissões de poluentes, pretendemos entender as variações nas concentrações de PM2.5, PM10, NO2 e O3 em estações de

medição distribuídas pela cidade e ao longo do tempo, bem como relacioná-las com fatores externos que possam contribuir para tal. Isto posto, o trabalho irá fornecer uma visão abrangente e fundamentada sobre a qualidade do ar na cidade, identificando a existência de áreas críticas de poluição assim como períodos nos quais podem existir diferenças no níveis de concentração destes poluentes. Com isso, esperamos contribuir para a conscientização sobre a poluição do ar na cidade do Rio de Janeiro e fornecer informações essenciais que possam orientar políticas públicas e práticas de gestão ambiental visando à melhoria da qualidade de vida dos habitantes e à preservação do meio ambiente.

O restante deste trabalho está disposto da seguinte forma: capítulo 2 apresenta o referencial teórico; o capítulo 3 apresenta a metodologia empregada e todas as análises que serão efetuadas; o capítulo 4 expõe os resultados obtidos através das pesquisas; o capítulo 5 encerra apresentando as conclusões que foram possíveis serem estabelecidas; assim, o capítulo 6 traz as referências bibliográficas as quais foram consultadas e utilizadas para a elaboração dessa pesquisa.

2. REFERENCIAL TEÓRICO

2.1 Poluentes Tratados

Os dados coletados são referentes a 4 poluentes: PM2.5; PM10; NO2; O3

2.1.1 PM2.5 e PM10

Do inglês “Particulate Matter”, os poluentes PM2.5 e PM10 significam materiais particulados e são emitidos em atividades poluidoras como a queima de combustíveis e a construção civil, sendo a parte numérica uma referência ao tamanho das partículas: PM2.5 são os materiais particulados com diâmetro inferior a 2,5 micrômetros e o PM10 aqueles com diâmetro entre 2.5 e 10 micrômetros (Lazzari, 2013). Tais partículas são responsáveis por uma série de problemas de saúde, sendo os sistemas respiratório e cardiovascular os mais afetados. A contaminação pelo PM2.5 e pelo PM10 tem sido associada a ataques de asma e ao crescente número de internações por problemas pulmonares, além da diminuição da capacidade respiratória, do desenvolvimento de asma e da redução da capacidade respiratória em crianças (Lazzari, 2013).

Além disso, o impacto do material particulado não é exclusividade da saúde humana, sendo o meio ambiente consideravelmente afetado. O impacto ambiental dos poluentes é observado: na redução da visibilidade por produção de neblina; no desequilíbrio dos ecossistemas marinhos e aquáticos principalmente pelo assentamento de metais que contém PM2.5 e PM10 e pelas chuvas ácidas; na dificultação da fotossíntese vegetal quando tais partículas assentadas sobre a superfície vegetal bloqueiam a absorção da luz solar; e, também, na fertilidade do solo, ao alterar a composição química do mesmo. Este último ainda representa um impacto econômico ao passo que afeta a produtividade agrícola e o rendimento das culturas (Sharma, 2021).

2.1.2 NO2

O Dióxido de Nitrogênio é um gás tóxico para pessoas e animais e a exposição de longa duração provoca danos sérios à saúde. Emitido de forma artifical em motores de combustão interna e usinas termelétricas e siderúrgicas, o NO2 pode provocar severos danos aos pulmões e a exposição contínua ainda pode causar diminuição permanente das funções pulmonares. Além dos malefícios à saúde humana, o NO2 pode reagir na atmosfera formando ácido nítrico, que, posteriormente, será o principal originador das chuvas ácidas. Eutrofização dos lagos e aumento do efeito estufa também estão entre os danos do NO2 ao meio ambiente.

2.1.3 O3

Outro fator importante é que o Dióxido de Nitrogênio (NO2) pode reagir e formar ozônio (O3). O ozônio é um oxidante muito forte, citotóxico (tóxico às células) e que, mesmo em baixas concentrações, pode causar agravamento de doenças respiratórias. Segundo Martins e Rodrigues (2001) o ozônio da troposfera é o principal poluente responsável, ainda, por danificar as plantas, tornando-as mais suscetíveis a estresses ambientais como seca e calor excessivo, bem como ao ataque de pragas, resultando em perdas consideráveis à produtividade da agropecuária. Pela Resolução n° 03 de 28/06/1990 do Conselho Nacional do Meio Ambiente (CONAMA), o padrão de qualidade do ar para as concentrações de O3 deve ser abaixo de 160 μg m-3 (Lazzari, 2013).

2.2 Ciclo de vida de projeto de ciencias de dados

O ciclo de vida de um projeto de ciências de dados, segundo Chapman et al. (2000), é uma parte essencial para a condução bem-sucedida de análises de dados. O processo de mineração de dados CRISP-DM (Cross-Industry Standard Process for Data Mining) oferece uma estrutura robusta para orientar as etapas de um projeto de ciências de dados. Abaixo estão as seis fases do CRISP-DM e destaca como elas moldam o ciclo de vida de um projeto de ciências de dados de acordo com Manresa (2020).

2.2.1 Compreensão do Negócio:

A primeira fase do CRISP-DM é entender o contexto de negócios e os objetivos da análise de dados. Isso envolve a colaboração com partes interessadas para definir claramente os problemas que estão sendo abordados e estabelecer métricas de sucesso.

2.2.2 Compreensão dos Dados:

Nesta etapa, os cientistas de dados exploram os dados disponíveis para o projeto. Isso inclui a coleta, limpeza e análise exploratória dos dados. A qualidade dos dados é avaliada, e os requisitos para preparação de dados são identificados.

2.2.3 Preparação de Dados:

A terceira fase se concentra na preparação de dados, envolvendo a transformação e integração de dados para criar conjuntos de dados adequados para modelagem. Técnicas como a seleção de atributos e tratamento de valores ausentes são aplicadas.

2.2.4 Modelagem:

Nesta etapa, modelos de análise de dados são desenvolvidos, ajustados e avaliados. Isso pode incluir técnicas de aprendizado de máquina, estatísticas ou outras abordagens analíticas, dependendo dos objetivos do projeto.

2.2.5 Avaliação:

A fase de avaliação visa determinar quão bem os modelos de dados funcionam na resolução do problema de negócios. Isso envolve a validação dos modelos em relação às métricas de desempenho e a análise de resultados Chapman et al. (2000).

2.2.6 Implementação:

A última fase aborda a implementação dos resultados do projeto no ambiente de produção. Isso pode incluir a integração dos modelos em sistemas existentes, a criação de APIs ou a implementação de soluções para uso prático.

O framework CRISP-DM fornece uma estrutura abrangente para o ciclo de vida de projetos de ciências de dados, garantindo que cada etapa seja abordada de forma sistemática. Isso ajuda a garantir que os projetos sejam bem-sucedidos e contribuam efetivamente para os objetivos de negócios.

2.3 Análise gráfica

A análise gráfica dos dados tratados em uma pesquisa estatística é a melhor maneira de apresentá-los de forma visual, o que facilita a interpretação desses conteúdos numéricos. De uma maneira mais geral, os gráficos podem ser classificados em 5 tipos básicos conforme a Figura 1 a seguir:

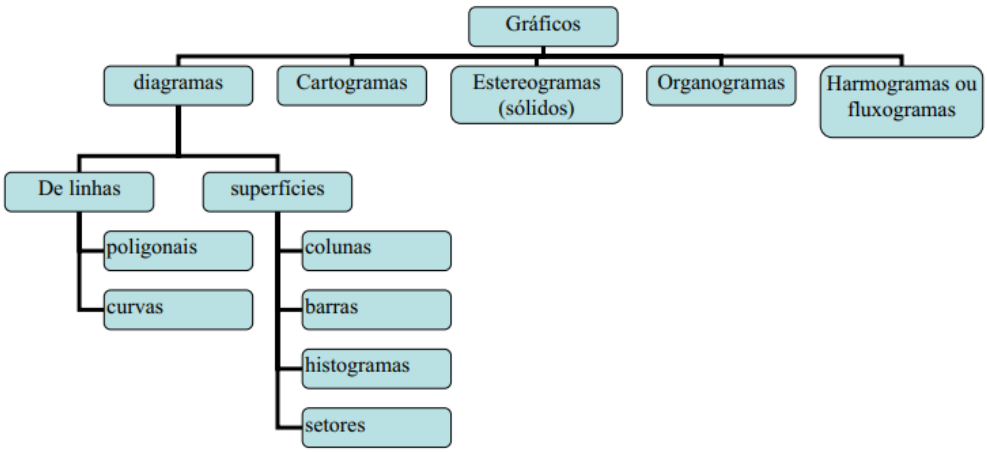

Figura 1 - Principais tipos de gráficos Fonte: Saad et al. (1990).

Conforme o tipo de análise desejada, um tipo de gráfico se torna mais adequado que outro. Os gráficos são comumente utilizados para facilitar a visualização da dependência de uma grandeza à outra, como a concentração de poluente ao longo do tempo ou a concentração de poluente por região. Os gráficos, quando construídos sob regras universais, são mais fáceis de serem interpretados.

2.3 Análise descritiva

A fase inicial do processo de análise dos dados coletados é a análise descritiva. Nessa etapa, empregamos técnicas de Estatística Descritiva para a organização, síntese e descrição dos aspectos significativos de um conjunto de características observadas. Isso viabiliza a identificação de padrões e tendências nessas características, bem como a comparação entre dois ou mais conjuntos. Algumas das principais ferramentas em Estatística Descritiva incluem medidas sumárias, tais como: medidas de localização (média, mediana, etc), medidas de dispersão (desvio padrão, amplitude, etc.), frequências e porcentagens, juntamente com instrumentos de visualização de dados, que abrangem uma ampla variedade de gráficos e tabelas.

A descrição dos dados também busca identificar anomalias e incongruências. As anomalias são representadas por valores de medição que se desviam da tendência geral do conjunto de dados, como uma queda no valor das ações devido a mudanças na situação econômica. Enquanto isso, as incongruências referem-se a valores de medição que foram registrados de maneira incorreta no banco de dados, sendo um exemplo muito comum as omissões de valores ou valores faltantes (Reis et al., 2002).

2.3.1 Medidas de localização

A medida de localização mais amplamente reconhecida e empregada para resumir dados é a média aritmética simples, comumente referida simplesmente como "média". Em termos simples, ela é obtida somando todos os valores presentes na base de dados e dividindo pelo número total de elementos.

$$\bar{x} = \frac{soma\ de\ todas\ as\ observações}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (1)$$

Sendo, n o tamanho da amostra; x_i o valor da i-ésima observação; $\sum_{i=1}^{n} x_i$ a soma de todas as observações; $\underline{x}$ o símbolo que representa a média aritmética simples.

Outras medidas de localização importantes são os quartis. Todos os quartis da amostra são definidos como médias ponderadas dos valores ordenados consecutivos. Os quartis da amostra podem ser obtidos através de diversas formas de interpolação. Neste trabalho será adotada a interpolação proposta por Hyndman e Fan (1996), tipo 7. Definindo o quartil, o valor do p fica fixado e temos:

$$p(n-1) = j + \gamma \quad (2)$$

Sendo, p pode assumir o valor de *25%, 50% e 75%*, representando os quartis 1, 2 e 3 respectivamente; n é o tamanho da amostra, e o resultado numérico resultante do lado esquerdo da Eq. (2) pode se desdobrar em $j + \gamma$, sendo j a parte inteira do resultado numérico e γ representando a parte decimal. Desta forma, o cálculo do quartil é dado por:

$$Q(p) = (1-\gamma)\, x_{(j)} + \gamma x_{(j+1)} \quad (3)$$

sendo x_j É a j-ésima estatística de ordem. Cabe destacar que a mediana de um conjunto de dados corresponde ao quartil 2.

2.3.2 Medidas de dispersão

A variância é uma medida de dispersão amplamente usada em estatística e probabilidade para avaliar o grau de variação dos dados em relação à média. Uma variância

baixa indica que os dados estão próximos à média, enquanto uma alta variância sugere que os dados estão mais espalhados, sendo útil na identificação de anomalias e na avaliação da precisão estatística. A seguir, a expressão que permite seu cálculo:

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1} \quad (4)$$

sendo x_i uma observação da amostra; $\bar{x}$ é a média aritmética do conjunto dos dados; e n é o total de dados na amostra.

Por outro lado, tem-se o desvio padrão, denotado por S, é a raiz quadrada da variância (Eq. (5)). Esta medida é utilizada para efeitos de comparação e análise da informação e pois preserva as mesmas unidades. A seguir, a expressão que permite calcular o desvio padrão:

$$S = \sqrt{S^2} \quad (5)$$

2.4 Análise de Variância (ANOVA)

A Análise de Variância, amplamente conhecida como ANOVA, é uma ferramenta essencial no campo da estatística, sendo uma coleção de modelos estatísticos poderosos que desempenham um papel crucial na compreensão das variações em um conjunto de dados. A principal premissa subjacente à ANOVA é a capacidade de particionar a variância amostral em diversos componentes, cada um deles associado a diferentes fatores ou variáveis que desempenham um papel fundamental em uma determinada aplicação, seja ela relacionada a um processo industrial, um produto em desenvolvimento ou um serviço prestado. A análise em questão requer um conjunto de pressupostos que devem ser atendidos de maneira integral para que a mesma possa ser posta em prática. Os tais pressupostos para a realização da ANOVA seriam, por sua vez, possuir um conjunto de dados livres de dados extremos (outliers), um conjunto de dados que siga uma distribuição normal e que preservem a homogeneidade da variância.

Ao realizar essa partição da variância, a ANOVA oferece uma janela para investigar e analisar minuciosamente como cada um desses fatores impacta a característica de interesse. Em outras palavras, ela nos permite desvendar o que está por trás das variações observadas nos dados, revelando quais variáveis são estatisticamente significativas e quais não são. Essa

capacidade de discernimento é essencial para tomadas de decisão informadas em uma ampla gama de contextos, desde a otimização de processos industriais até o aprimoramento de produtos e a melhoria na qualidade dos serviços.

Portanto, a ANOVA não é apenas uma técnica estatística, mas uma ferramenta analítica poderosa que oferece insights profundos sobre a influência e a interação dos fatores em um sistema, permitindo que os profissionais e pesquisadores ajam com base em dados sólidos e fundamentados. Através dela, podemos explorar as nuances do mundo estatístico e desvendar relações complexas que podem ser cruciais para o sucesso em diversas áreas de atuação. (Costa e Leo, 2020; Maçaira, 2022).

Considerando um cenário em que desejamos avaliar como o fator A, que possui *k* níveis fixos, afeta um processo, produto ou serviço específico. Nesse contexto, selecionamos uma amostra aleatória de *N* unidades experimentais de uma população de unidades experimentais. Cada unidade experimental serve como a unidade fundamental na qual os tratamentos são aplicados. Como exemplo:

Tabela 1 - Dados organizados de ANOVA para um único fator.

Nível	**Fator A**	**Observações**			**Somas**	**Médias**
1	y_{11}	$\boldsymbol{y}_{12}$	…	$\boldsymbol{y}_{1n1}$	$\boldsymbol{y}_{1.}$	$\underline{\boldsymbol{y}}_{1.}$
2	$\boldsymbol{y}_{21}$	$\boldsymbol{y}_{22}$	…	$\boldsymbol{y}_{2n2}$	$\boldsymbol{y}_{2.}$	$\underline{\boldsymbol{y}}_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
k	$\boldsymbol{y}_{k1}$	$\boldsymbol{y}_{k2}$	…	$\boldsymbol{y}_{knk}$	$\boldsymbol{y}_{k.}$	$\underline{\boldsymbol{y}}_{k.}$

Fonte: Costa e Leo, 2020; Maçaira (2022).

Para uma análise eficaz, é fundamental descrever os dados por meio de um modelo adequado, e um dos modelos mais simples é o modelo de efeitos, descrito abaixo:

$$\boldsymbol{y_{ij} = \mu + \alpha_i + \varepsilon_{ij}} \tag{6}$$

No qual:

- $j = 1, \ldots, n_i$ e $i = 1, 2, \ldots, k$
- μ representa a média geral dos dados e é um parâmetro que se aplica uniformemente a todos os tratamentos.

- α_i é o efeito que o nível *i* do fator provoca na variável resposta
- ε_{ij} é o erro aleatório experimental

O erro experimental (ε_{ij}) pode ser conceituado como a variabilidade resultante de outros fatores que afetam o processo, produto ou serviço, mas que não foram incluídos no experimento. Essa variável representa as flutuações não contabilizadas pelo modelo, decorrentes de diversas fontes não abordadas na pesquisa (Costa e Leo, 2020; Maçaira, 2022).

Com base nos dados, adotaremos a seguinte notação:

- $y_{i.} = \sum_{j=1}^{n_i} y_{ij}$: soma das observações do nível i do fator A (7)

- $\bar{y}_{i.} = \frac{\sum_{j=1}^{n_i} y_{ij}}{ni}$: média das observações do nível i do fator A (8)

- $y_{..} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}$: soma de todas as observações (9)

- $\bar{y}_{..} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}}{N}$: média geral das observações,
  sendo $N = \sum_{i=1}^{k} n_i$, total de observações (10)

Além disso, será assumido que o erro experimental consiste em variáveis aleatórias independentes e igualmente distribuídas, seguindo uma distribuição normal com média zero e variância σ^2. Em outras palavras, nossa suposição é a de que $\varepsilon_{ij} \sim N(0, \sigma^2)$. Deste modo, extraímos que y_{ij} também tem distribuição normal com média $\mu + \alpha_i$ e variância σ^2, para todo $j = 1, \cdots, n_i \ e \ i = 1, \cdots, k$ (Costa e Leo, 2020; Maçaira, 2022).

Na prática, nosso foco é avaliar o impacto do fator na resposta. Para isso, estamos interessados em avaliar como os diferentes níveis do fator influenciam a variável resposta. Em outras palavras, buscamos testar as seguintes hipóteses:

$$H_o: \alpha_1 = \ldots = \alpha_k = 0$$
$$H_1: \alpha_i \neq 0 \text{ (para algum } i = 1, \ldots, k) \quad (11)$$

2.5 Teste de Tukey

O Teste de Tukey, desenvolvido é uma técnica estatística amplamente utilizada para realizar comparações múltiplas entre médias de diferentes grupos ou tratamentos. Ele é frequentemente aplicado após a realização de uma Análise de Variância (ANOVA) quando se detectam diferenças significativas entre grupos. O Teste de Tukey tem o propósito de identificar quais grupos específicos são significativamente diferentes entre si, enquanto controla a taxa de erro global (Zar, 2014).

A base do Teste de Tukey é a comparação das diferenças entre as médias amostrais dos grupos e a estimativa do erro padrão dessas diferenças. O procedimento envolve o cálculo de uma estatística de teste denominada estatística Q, que é comparada a um valor crítico de acordo com o número de grupos e o nível de significância escolhido. A estatística Q é calculada da seguinte maneira:

$$Q = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{\frac{MSE}{n}}} \tag{12}$$

Onde:

- $\bar{x}_i \; e \; \bar{x}_j$ sao as médias amostrais dos grupos i e j, respectivamente
- MSE é o erro médio quadrático obtido a partir da ANOVA
- n é o tamanho da amostra

Os procedimentos do teste de Tukey podem ser listados como:

- Realizar uma ANOVA para determinar se há diferenças significativas entre os grupos.
- Calcular a estatística Q para todas as combinações possíveis de pares de grupos.
- Comparar a estatística Q calculada com o valor crítico da distribuição de Studentized Range (Q) para o nível de significância escolhido e o número total de grupos.

Se a estatística Q calculada for maior que o valor crítico, as médias dos grupos correspondentes são consideradas estatisticamente diferentes.

Uma das vantagens do Teste de Tukey é que ele controla a taxa de erro global, evitando que ocorram muitos falsos positivos (erros do tipo I) ao realizar várias comparações. Isso é alcançado ao ajustar o valor crítico com base no número de comparações sendo feitas.

2.6 Modelagem: Teste de Kruskal-Wallis

As análises não paramétricas apresentam a vantagem quando as suposições de normalidade e homogeneidade de variância são seriamente violadas, uma vez que não é apropriado confiar nos resultados de uma análise de variância tradicional, visto que a probabilidade de cometer um erro do Tipo I (rejeitar a hipótese nula quando ela é verdadeira) se distancia consideravelmente de α. Uma alternativa não paramétrica à ANOVA é o teste de Kruskal-Wallis, também conhecido como ANOVA não paramétrica. Esse teste é utilizado para comparar três ou mais conjuntos de dados independentes e determinar se há diferenças significativas entre pelo menos dois deles. No teste de Kruskal-Wallis, os valores numéricos são transformados em postos e reunidos em um único conjunto de dados. A comparação entre os grupos é realizada com base nas médias dos postos (posto médio) (Vieira, 2018).

O método começa com a definição das hipóteses como na Eq. (13), e posteriormente, atribui um posto a cada valor observado, com o menor valor recebendo o posto mais baixo e o maior valor o posto mais alto. Em seguida, os postos de cada conjunto de dados são somados. Se a hipótese nula for verdadeira e existirem apenas diferenças aleatórias entre os grupos, espera-se que os postos mais altos e mais baixos estejam distribuídos de maneira equilibrada entre os grupos. No entanto, se houver uma preponderância de postos mais altos ou mais baixos em qualquer grupo, isso provavelmente refletirá diferenças significativas devidas à variável independente.

$$\begin{aligned} &H_o: os\ postos\ médios\ são\ iguais \\ &H_1: os\ postos\ médios\ não\ são\ iguais \end{aligned} \tag{13}$$

A estatística do teste de Kruskal-Wallis é calculada através da seguinte Equação:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1) \tag{14}$$

Sendo:

k = Número de grupos;

N = Número total de medições experimentais;

R_i = Soma dos ranks de cada grupo;

n_i = Número de medidas em cada grupo;

H = Valor da estatística de Kruskal-Wallis.

A hipótese nula deve ser rejeitada se o valor observado da estatística H for superior ao valor crítico (teste unilateral à direita), para isso se busca esse valor crítico na tabela de Qui-Quadrado (χ^2), com ($k-1$) graus de liberdade, pois são k grupos, e $\alpha = 5\%$. Quando se rejeita a hipótese nula H_o no teste de Kruskal-Wallis, existe evidência de que pelo menos um dos grupos é diferente dos demais. Porém, não se tem a informação de qual ou quais são diferentes (Vieira, 2018).

2.7 Teste de Dunn

Análogo ao teste de Tukey, o teste de Dunn é empregado quando os pesquisadores não ficam satisfeitos com a conclusão de que as populações amostradas não são iguais ou de que os tratamentos estudados não têm, todos, o mesmo efeito. Eles querem saber onde estão as diferenças. Nesses casos, é muito usado o teste de Dunn.O teste de comparações múltiplas de Dunn é aplicado como um procedimento subsequente ao teste de Kruskal-Wallis, mas apenas quando o teste de K-W resulta em rejeição da hipótese nula (H_o). Devido a essa condição, ele é ocasionalmente referido como pós-teste de Dunn ou teste post-hoc de Dunn. O teste de Dunn compara postos médios de grupos, dois a dois. Não exige que os grupos tenham o mesmo número de observações, mas tem melhor aproximação quando as amostras são grandes.

$$\begin{aligned} &H_o: Os\ grupos\ em\ comparação\ têm\ os\ mesmos\ postos\ médios \\ &H_1: Os\ grupos\ em\ comparação\ não\ têm\ os\ mesmos\ postos\ médios \end{aligned} \tag{15}$$

Calcule os postos médios dos k grupos:

$$\bar{R_i} = \frac{\sum_{i=1}^{k} R_i}{n_i} \tag{16}$$

em que $i = 1,2,...,k$ indica o posto médio do i-ésimo grupo.

Calcule o erro padrão para cada par de postos médios:

$$SE = \sqrt{\frac{n(n+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \tag{17}$$

Calcule as estatística de teste para comparar as médias de postos, duas a duas:

$$q = \frac{\bar{R}_i - \bar{R}_j}{SE} \tag{18}$$

em que *i* e *j* indicam o *i*-ésimo e o *j*-ésimo grupo, *i = 1,2,...,k; j = 1,2,...,k* . Toda vez que o valor calculado de q for maior do que o valor crítico de *q* no nível estabelecido de significância e para *k* grupos, a diferença entre grupos é significante. Se você estiver usando um programa para computador, procure o *p*-valor associado a cada diferença de postos médios (Vieira, 2018)**.**

3. MATERIAIS E MÉTODOS

Neste capítulo serão apresentados os dados do estudo, assim como a metodologia abordada. Para isto, inicialmente foi realizada uma análise descritiva e exploratória analítica dos dados para entender o seu comportamento. A seguir, foram implementados os testes que permitem verificar os pressupostos para a aplicação da ANOVA, seja ela paramétrica ou não. É importante ressaltar que é adotada a estrutura do ciclo de vida de um projeto de Ciência de Dados.

3.1 Poluentes e informações

Para a elaboração do presente estudo foi utilizado um conjunto de dados públicos que advém do programa de monitoramento da qualidade do ar da Prefeitura da Cidade do Rio de Janeiro – (Data.Rio, 2023) do site Data Rio, publicada pela prefeitura do estado do Rio de Janeiro e inclui as estações fixas de qualidade do ar e a série histórica dos dados. O período de análise está compreendido entre 01/01/2012 e 31/12/2021 devido ao fato de esse periodo conter uma disponibilidade adequda de dados. Nele constam várias medições de alguns parâmetros meteorológicos e de um conjunto de poluentes em estações específicas localizadas na cidade do Rio de Janeiro, nas quais tais medições são apresentadas com uma frequência horária. Para efeitos de análise, transformamos os dados para uma frequência diária considerando a média das horas de cada dia. Esse conjunto de dados possui o total de 28.194 linhas que representam as medidas realizadas e 26 colunas que indicam algumas variáveis adicionais referentes a variáveis climatológicas.

O conjunto de dados filtrado apresenta ao todo 4 poluentes (PM10, PM2.5, NO2 e O3) e também 8 estações que encontram-se destacadas na Tabela 2 abaixo através do fornecimento dos códigos das estações e dos seus respectivos nomes.

Tabela 2 – Códigos e nomes das estações consideradas no estudo.

Códigos das estações	Nomes das estações
BG	Bangu
CG	Campo Grande
CA	Centro
AV	Copacabana
IR	Irajá
PG	Pedra de Guaratiba
SC	São Cristóvão
SP	Tijuca

Fonte: Autores (2023).

Para uma melhor visualização e compreensão espacial com relação às estações consideradas nesse estudo, elaborou-se um mapa da cidade do Rio de Janeiro onde destacou-se somente as estações utilizadas no estudo.

Figura 2 - Localização no mapa do Rio de Janeiro das estações consideradas. Fonte: Google Maps

Cabe destacar que nem todos os poluentes foram medidos em todas as estações Aqui o PM10 e o O3, foram medidos nas 8 estações consideradas nessa base de dados. Já o PM2.5 foi medido em apenas uma estação considerada (IR) (Irajá), e o poluente NO2 foi medido nas estações: BG (Bangu), CG (Campo Grande), IR (Irajá), SP (Tijuca).

3.2 Análise Descritiva

Para a etapa de análise descritiva, as ações que serão executadas consideraram as estações e os anos para cada um dos poluentes em destaque. Além disso, de forma a possibilitar uma inspeção visual dos dados, é realizada uma análise gráfica considerando gráficos de linhas e Boxplot. Para análise e modelagem dos dados o *software* R (R Core Team, 2023) e seu editor RStudio (RStudio Team, 2020) foram usados, mais especificamente, os seguintes pacotes: readr (Wickham et al., 2023) (para leitura dos dados), tidyverse (Wickham et al., 2019) (para limpeza e tratamento dos dados), ggplot2 (Wickham, 2016) (análise gráfica), RVAideMemoire (Hervé, 2023) e car (Fox, 2023) sendo estes dois últimos para aplicação de testes estatísticos de normalidade e variância.

Além da inspeção visual, foi possível realizar uma análise das medidas de localização/posição desses dados (média, mediana, mínimo, máximo, Quartil 1 e Quartil 3), bem como medidas de dispersão dos dados (variância, desvio padrão e amplitude, valores ausentes).

Tanto para os gráficos produzidos, como para as medidas extraídas foram levados em conta separadamente para cada poluente as estações de medição de cada um deles, bem como os anos dentro da faixa estabelecida anteriormente. A partir dessas imagens e medidas geradas, iniciou-se a observação e identificação da presença de valores faltantes e valores extremos para suas respectivas remoções.

3.3 Metodologia

A etapa de metodologia explica os passos e os caminhos que foram tomados a partir das análises e dos resultados das mesmas, sendo aplicada tanto para a análise regional que vai levar em conta somente as diferentes estações, bem como para a análise temporal que vai levar em conta somente cada um dos anos de medições dos dados. Em outra palavras, para o estudo regional são considerados como níveis do fator (poluente), a diferentes *estações*, já para o estudo temporal os níveis do fator corresponderão ao *anos*.

A primeira parte, consiste em realizar análises, pós tratamento de dados para detecção de outliers, análise de normalidade dos dados e da homogeneidade da variância para cada um dos poluentes. Para essas experimentações foram utilizados respectivamente o teste de Shapiro-Wilk e o teste de Levene e para a execução das mesmas foram utilizados os comandos *byf.shapiro* (Hervé, 2023)e *leveneTest* (Fox, 2023), respectivamente. As hipóteses definidas para o teste de Shapiro-Wilk são:

$$\begin{aligned} H_o &: Os\ dados\ entre\ grupos\ seguem\ uma\ distribuição\ normal \\ H_1 &: Os\ dados\ não\ seguem\ uma\ distribuição\ normal \end{aligned} \tag{19}$$

Para esse teste considerou-se um nível de significância $\alpha = 5\%$, sendo assim, utilizamos o p-valor de cada um dos grupos, nesse caso, cada uma das estações consideradas, para saber se a hipótese nula será ou não rejeitada. Caso o p-valor seja menor do que α, a hipótese nula será rejeitada, caso contrário, se p-valor for maior do que α, não há evidências estatísticas para rejeitar a hipótese nula. Ou seja, para os grupos que apresentarem p-valor menor do que α, não poderemos considerar que esses grupos sigam uma distribuição normal.

O teste de Levene, por sua vez, possui como suas hipóteses a hipótese nula H_o e a hipótese alternativa H_1 como é expressãdo a seguir:

$$\begin{aligned} H_o &: \textit{As variâncias dos grupos são homogêneas} \\ H_1 &: \textit{As variâncias dos grupos não são homogêneas} \end{aligned} \tag{20}$$

Esse teste também considera um nível de significância $\alpha = 5\%$, sendo assim, caso o p-valor seja menor ou igual a α, deve-se então rejeitar a hipótese nula, admitindo que as variâncias dos grupos não são homogêneas. Por outro lado, caso apresente um p-valor maior que α, deve-se aceitar a hipótese nula e admitir que as variâncias desses grupos são homogêneas.

A realização os dois testes anteriores: a normalidade dos dados e a homogeneidade das variâncias fornecem informações extremamente importantes para seguir com as futuras análises. Caso os dados sigam uma distribuição normal e também a variância dos grupos sejam homogêneas, deve-se partir para a aplicação do teste da análise de variância (ANOVA) de uma um fator. Por outro lado, quando as suposições de normalidade e homogeneidade de variâncias são violadas, não é apropriado confiar nos resultados de uma análise de variância tradicional, e alternativamente deve-se realizar a ANOVA não paramétrica, isto é, o teste de Kruskal-Wallis.

3.3.1 Análise Regional

A etapa de Análise regional contou com a aplicação de todos os passos e procedimentos que foram descritos na parte da metodologia. Sendo assim, levou em conta somente as diferentes estações nas quais os poluentes foram medidos e utilizou como sendo os níveis do fator a variável "*estação*".

Esta análise foi aplicada para todos os poluentes considerados com exceção do poluente PM2.5 que, devido ao fato de apresentar suas medições somente em uma estação (IR) (Irajá), não é possível efetuar as devidas análises e comparações, feriu também o pressuposto de possuir menos de três grupos para serem estudados. Essa etapa foi importante para avaliar o comportamento dos grupos com relação às estações de medida, para que houvesse possibilidade de extrair conclusões a respeito desse comportamento que os dados seguem.

3.3.2 Análise Temporal

A etapa de Análise Temporal foi necessária para obter uma visão com relação aos anos nos quais os poluentes foram medidos, e contou com a aplicação de todos os passos e procedimentos que foram descritos na parte da metodologia. Sendo assim, levou em conta somente os diferentes anos e utilizou como níveis do fator a variável "*ano*". Assim como na análise regional, análise temporal foi aplicada para todos os poluentes considerados no trabalho. Essa etapa foi importante para avaliar o comportamento dos grupos com relação ao longo do tempo dos diferentes poluentes, para que houvesse possibilidade de extrair conclusões a respeito desse comportamento que os dados seguem.

4. RESULTADOS E DISCUSSÃO

Neste capítulo foram apresentados todos os resultados obtidos a partir das análises executadas no capítulo 3, bem como as discussões necessárias para o entendimento dos resultados obtidos.

4.1 Análise descritiva

A primeira atividade executada foi a montagem dos gráficos de linha discriminados por estação ao longo do tempo, bem como os gráficos de Boxplot para cada um dos poluentes discriminados por suas estações, seguidos das tabelas com as medidas de localização/posição desses dados (média, mediana, etc), bem como medidas de dispersão dos dados (variância, desvio, etc). Vale ressaltar que o poluente PM2.5 foi medido unicamente na estação "IR" (Irajá), sendo assim, ele não integrou a parte da análise regional devido à falta de estações a serem usadas para as comparações.

Considerando o poluente PM10, os gráficos gerados estão representados nas Figuras 3, 4 e 5, indicando respectivamente a visualização de linhas e de Boxplot por estacão e por ano medidos em [$\mu g/m3$].

Figura 3 - Série temporal das medições do poluente PM10 por estação. Fonte: Autores (2023).

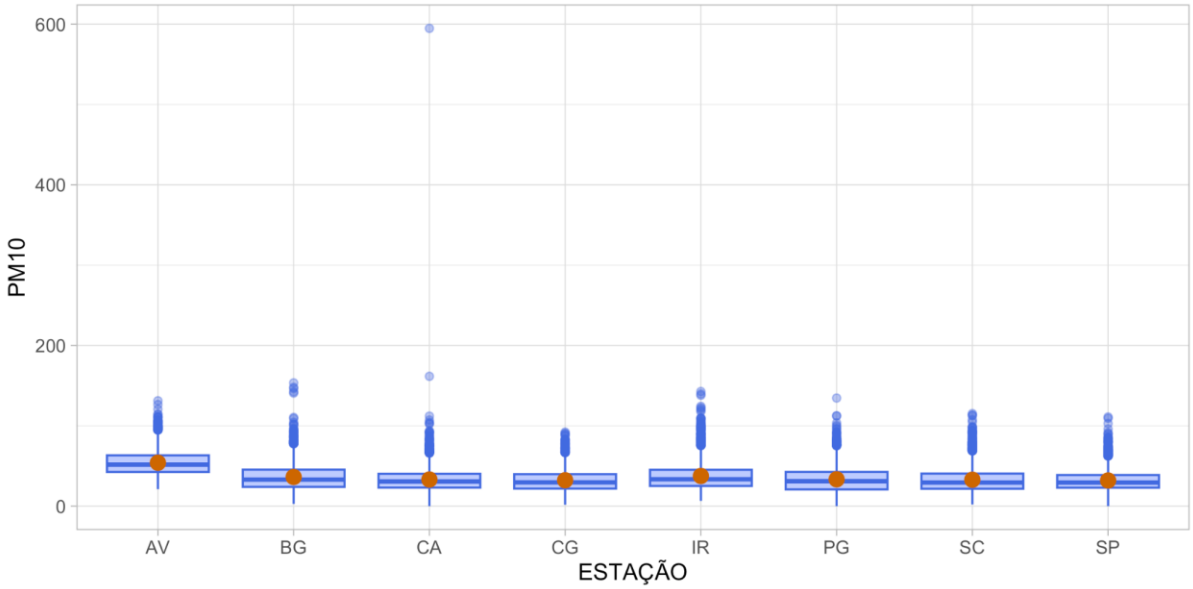

Figura 4 - Gráfico de Boxplot do poluente PM10 por estação. Fonte: Autores (2023).

Inicialmente é possível notar que o poluente PM10 é medido em todas as estações consideradas neste estudo. Ao analisar a série temporal, é possível inspeccionar visualmente que há a presença de dados faltantes na maioria das estações quando analisamos ao longo dos anos e também, a presença de dados extremos em alguns anos para todas as estações. Nota-se também por essa visão anual que os dados parecem ter um comportamento semelhante.

Olhando agora pela visão do Boxplot, percebemos a ocorrência de um outlier extremo específico para a estação "CA" (Centro) e também a presença de outliers em todas as estações consideradas.

A distribuição dos demais dados, considerando as médias, parece estar concentrada na faixa de 30 [$\mu g/m3$].

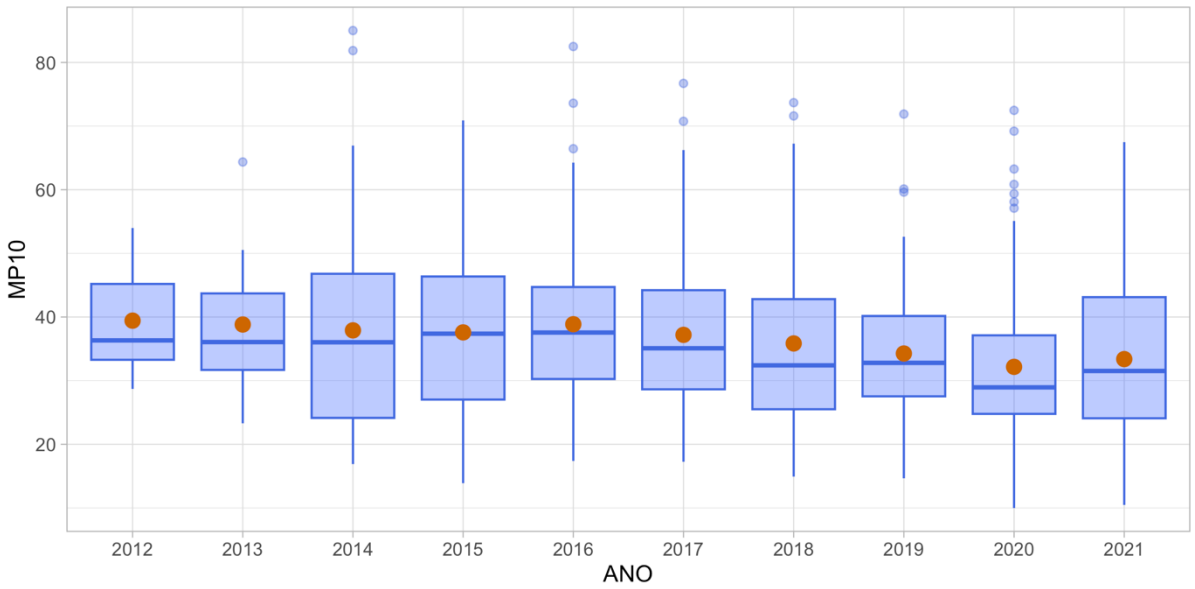

Figura 5 - Gráfico de Boxplot do poluente PM10 por ano. Fonte: Autores (2023).

Para o Boxplot de todos os anos do poluente PM10, observa-se uma distribuição das médias concentrada na faixa da de 30 [$\mu g/m3$] até 40 [$\mu g/m3$]. É possível também identificar a presença de valores extremos na maioria dos anos considerados e destacar uma maior dispersão dos dados no ano de 2014, 2015 e 2021. Além disso, a mediana do poluente por ano dimunui a partir de 2018, tendo sua maior queda em 2020, ma voltando a aumentar em 2021. Observa-se que o ano com menor valor mediano do poluente PM10 corresponde a 2020

As medições de localização/posição, assim como as medidas de dispersão dos dados de acordo com cada uma das estações nas quais o poluente foi medido estão expressas na Tabela 3.

Tabela 3 - Medidas de localização/posição e dispersão para o poluente PM10, por estação

Estatísticas descritivas	AV	BG	CA	CG	IR	PG	SC	SP
Média	54,05	36,33	33,14	32,24	37,58	33,42	32,88	31,95
Mediana	51,72	33.08	30,71	29,62	33,50	31,06	29,44	29,42
Variância	236,39	301,74	306,21	199,84	320,93	283,32	240,70	170,79
Desvio Padrão	15,37	17,37	17,50	14,14	17,91	16,83	15,51	13,07
Mínimo	21,15	2,83	0,00	1,79	6,62	0,00	2,17	0,00
Máximo	131,20	153,50	594,71	92,21	142,75	134,58	115,25	111,02
Amplitude	110,05	150,67	594,71	90,42	136,12	134,58	113,08	111,02
Q1	42,57	23,99	23,08	21,88	25,20	20,92	21,71	22,98
Q3	63,28	45,46	40,29	39,83	45,29	42,67	40,53	38,77
IQR (Q3-Q1)	20,71	21,47	17,21	17,95	20,09	21,75	18,82	15,79
Valores Ausentes	225	917	505,00	953,00	1025,00	989,00	630,00	353,00

Fonte: Autores (2023).

Os valores apresentados pela Tabela 3 confirmam as observações apontadas nas análises visuais do poluente PM10 nas estações de medida, com ressalva para a estação "AV" (Copacabana) que apresentou uma média acima da faixa considerada. É notável também que todas estações apresentam um desvio padrão em uma faixa comum e através da largura dos Boxplots, comprova-se a proximidade dos quartis entre as estações.

Destaca-se a grande quantidade de valores ausentes em todas as estações , bem como a presença de valores extremos. Sendo assim, percebe-se a necessidade de um tratamento para esses dados através da remoção desses registros para os valores ausentes bem como a remoção dos valores extremos, de maneira a não comprometer os testes de normalidade e homogeneidade das variâncias posteriores.

Considerando agora o poluente NO2, os gráficos de linhas e de Boxplot por estacão e por ano gerados são representados nas Figuras 6, 7 e 8, medidos em [$\mu g/m3$].

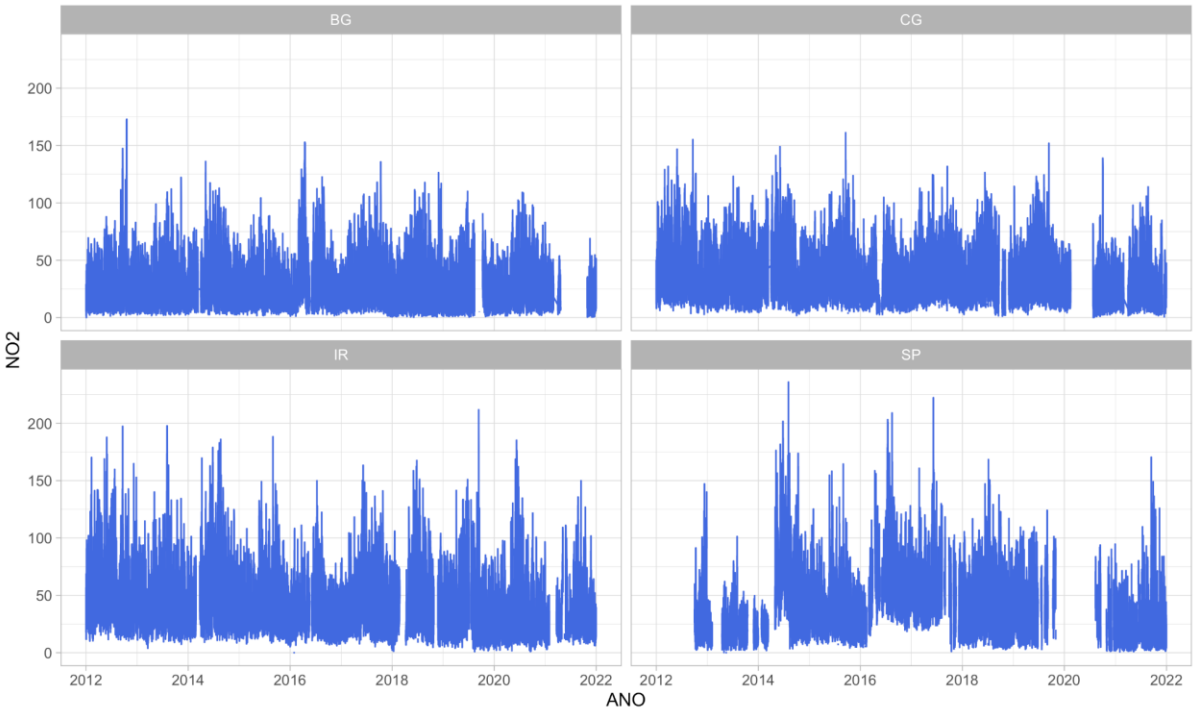

Figura 6 - Série temporal das medições do poluente NO2 por estação. Fonte: Autores (2023).

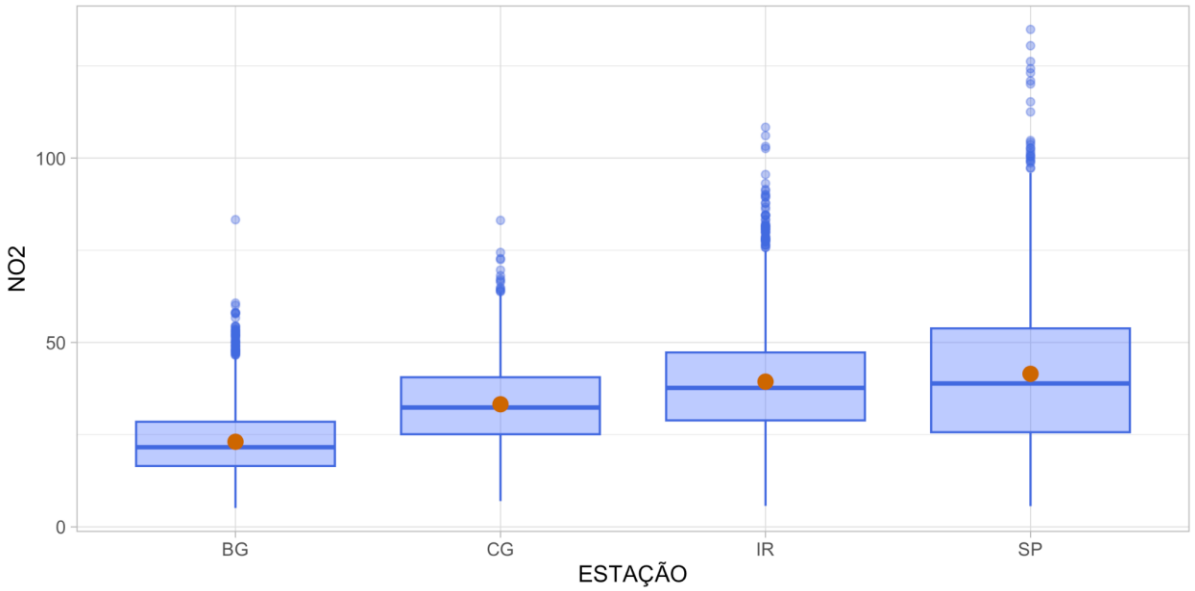

Figura 7 - Gráfico de Boxplot do poluente NO2 por estação. Fonte: Autores (2023).

Através da observação dos gráficos, é possível notar que o poluente NO2 é medido apenas nas estações: "BG" (Bangu), "CG" (Campo Grande), "IR" (Irajá) e "SP" (Tijuca). Pela série temporal percebe-se que há ausência de dados em todas estações, principalmente considerando o ano de 2020. Quando consideramos a visão por anos detecta-se a presença de

dados extremos em algumas faixas de tempo para todas as estações. Nota-se também por essa visão anual que os dados parecem ter um comportamento semelhante. Olhando agora pela visão do Boxplot, percebemos a ocorrência de outliers extremos específicos para a estação "SP" (Tijuca) e também a presença de outliers em todas as estações consideradas, bem como uma maior dispersão dos dados para as estações de "SP" (Tijuca) e "IR" (Irajá), ao observarmos a largura de Boxplots. A distribuição dos demais dados, considerando as médias parece estar concentrada em uma faixa entre 20 [$\mu g/m3$] até 45 [$\mu g/m3$].

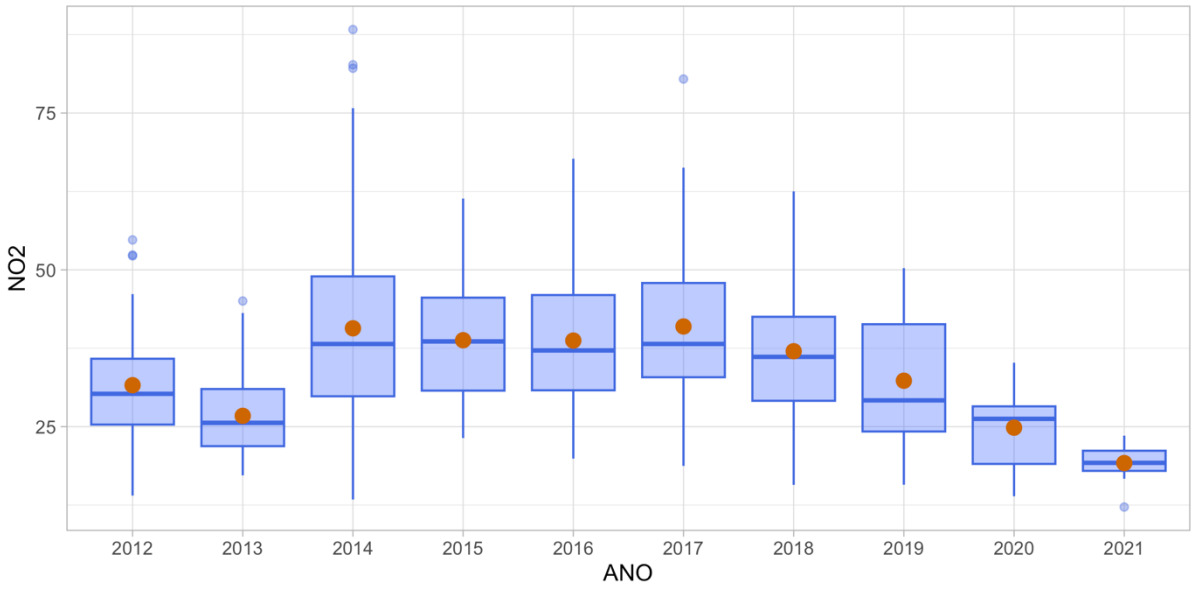

Figura 8 - Gráfico de Boxplot do poluente NO2 por ano. Fonte: Autores (2023).

Para o Boxplot de todos os anos do poluente NO2 é possível identificar a presença de valores extremos na maioria dos anos considerados. Os anos de 2014 e 2019 apresentam uma a maior dispersão dos dados, mas destaca-se 2021 por ter a menos variabilidade do poluente. Quando analisamos a médiana e a média (ponto vermelho), pode-se identificar a sua variabilidade ao longo dos anos, com valores abaixo de 20 [$\mu g/m3$] (ano 2021), valores proximos do poluente, na mediana, para os anos 2012, 2019; assim como para o conjunto de anos entre 2014 e 2018, inclusive.

Para um melhor entendimento quantitativo do comportamento desses dados, calculou-se as medida de localização/posição, assim como as medidas de dispersão dos dados de acordo com cada uma das estações nas quais o poluente foi medido estão expressas na Tabela 4.

Tabela 4 - Medidas de localização/posição e dispersão para o poluente NO2 por estação.

Estatísticas descritivas	BG	CG	IR	SP
Média	23,07	33,24	29,37	41,50
Mediana	21,59	32,36	37,68	38,88
Variância	87,09	128,95	211,35	422,53
Desvio Padrão	9,33	11,36	14,54	20,56
Mínimo	5,10	6,97	5,69	5,62
Máximo	83,29	83,11	108,35	134,90
Amplitude	78,19	76,14	102,66	129,28
Q1	16,50	25,11	28,86	25,66
Q3	28,50	40,57	47,29	53,82
IQR (Q3-Q1)	12,00	15,46	18,43	28,16
Valores Ausentes	1313	1640	1096	1934

Fonte: Autores (2023).

Através dos valores apresentados na Tabela 4, obtém-se uma confirmação das observações destacadas das análises visuais do poluente NO2 nas estações de medida. É notável também que todas estações apresentam um desvio padrão em uma faixas diferentes, e através das medidas de IQR comprova-se a maior largura dos Boxplots, principalmente para as estações "SP" (Tijuca) e "IR" (Irajá), que possuem uma distância maior dos quartis entre as estações.

Destaca-se a grande quantidade de valores ausentes em todas as estações , bem como a presença de valores extremos. Sendo assim, percebe-se a necessidade de um tratamento para esses dados através da remoção desses valores ausentes bem como a remoção dos valores extremos, de maneira a não comprometer os testes de normalidade e homogeneidade das variâncias posteriores.

Por fim, levando em conta o poluente O3, os gráficos gerados estão representados nas Figuras 9, 10 e 11 representado, respectivamente, a visualização do gráfico de linhas e o Boxplot por estacão e por ano, medidos em [*µg/m3*].

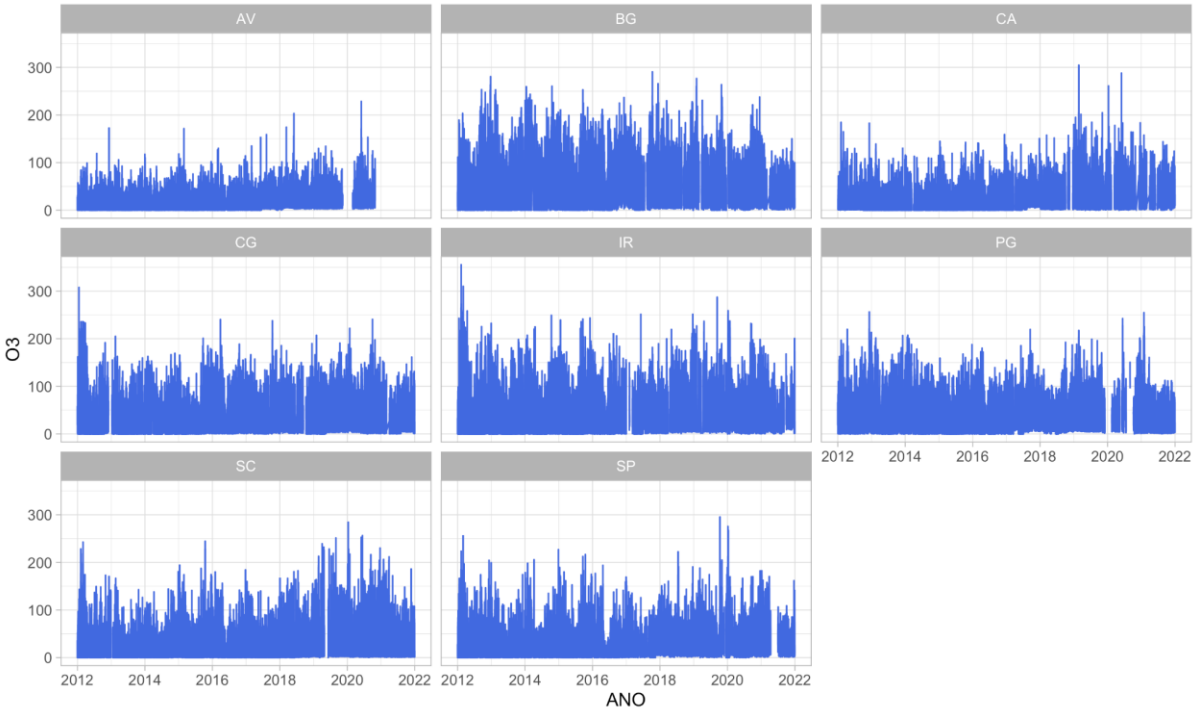

Figura 9 - Série temporal das medições do poluente O3 por estação. Fonte: Autores (2023).

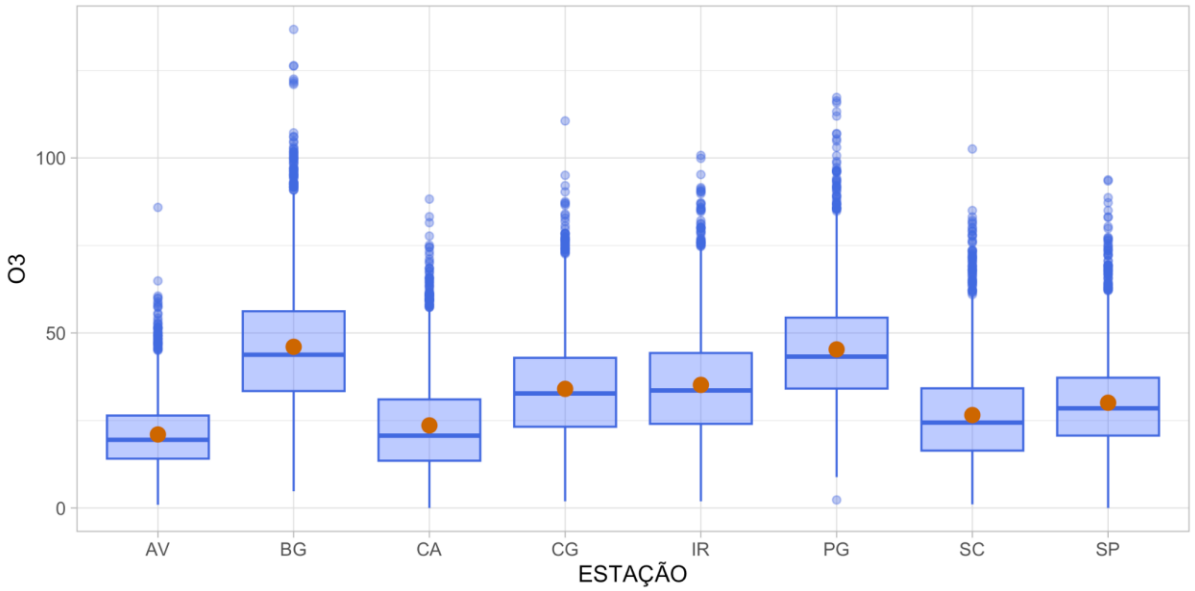

Figura 10 - Gráfico de Boxplot do poluente O3 por estação. Fonte: Autores (2023).

Para o O3 é possível notar que o poluente é medido em todas as estações consideradas neste estudo. Pela série temporal percebe-se que, similar aos outros poluentes há ausência de dados na maioria das estações, principalmente considerando o ano de 2020. Percebe-se que o comportamento dessas medições segue um padrão oscilatório ao longo do tempo. Quando consideramos a visão por anos detecta-se novamente a presença de valores extremos em algumas faixas de tempo para todas as estações. Nota-se também por essa visão anual que os dados parecem ter um comportamento semelhante. Olhando agora pela visão do Boxplot, percebemos a ocorrência de outliers extremos específicos para as estações "BG" (Bangu) e "PG" (Pedra de Guaratiba) e também a ocorrência de outliers normais em todas as estações consideradas.

A distribuição dos demais dados, considerando as médias, parece estar concentrada em uma faixa entre 20 [$\mu g/m3$] até 45 [$\mu g/m3$]. Percebe-se também um comportamento distinto entre os dados para as diferentes estações.

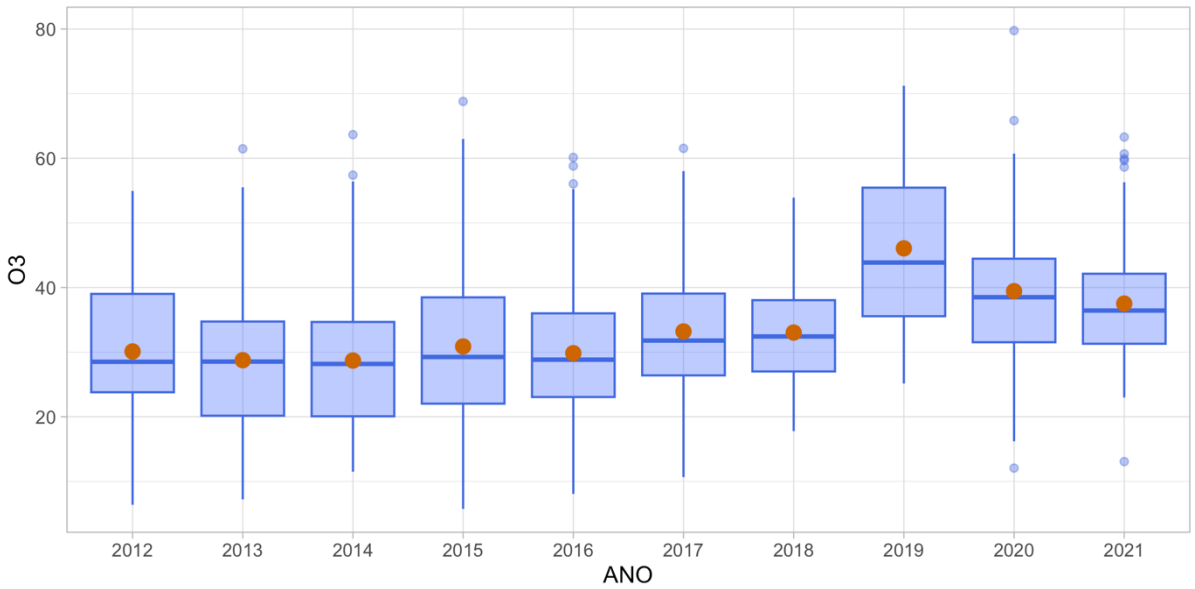

Figura 11 - Gráfico de Boxplot do poluente O3 por ano. Fonte: Autores (2023).

Para o Boxplot de todos os anos do poluente O3 é possível identificar a presença de valores extremos na maioria dos anos considerados, bem como observa-se uma distribuição das médias concentrada dentro da faixa da de 30[$\mu g/m3$] até 40 [$\mu g/m3$] para a maioria dos anos, com exceção de 2019 que, por sua vez, também apresenta uma maior dispersão dos dados. Cabe destacar que entre os anos 2012 e 2016 o comportamento mediano do poluente foi semelhante; mas a partir de 2017, evidenciou-se um aumento atingindo seu maior valor

mediano e médio também (ponto vermelho) em 2019, e caindo para 2020 e 2021, mas continuando em paratamares elevados se comparados com anos da década anterior.

Para um melhor entendimento quantitativo do comportamento desses dados, realizou-se uma análise numérica através das medidas de localização, assim como das medidas de dispersão dos dados de acordo com cada uma das estações nas quais o poluente foi medido (Tabela 5).

Tabela 5 - Medidas de localização/posição e dispersão para o poluente O3 por estação.

Estatísticas descritivas	**AV**	**BG**	**CA**	**CG**	**IR**	**PG**	**SC**	**SP**
Média	21,01	46,05	23,61	34,03	35,18	45,32	26,55	30,07
Mediana	19,49	43,78	20,68	32,73	33,58	43,25	24,40	28,49
Variância	94,11	311,07	180,04	210,84	234,38	245,52	190,59	175,25
Desvio Padrão	9,70	17,64	13,42	14,52	15,31	15,67	13,81	13,24
Mínimo	0,91	4,78	0,00	1,91	1,91	2,31	1,04	0,00
Máximo	85,87	136,72	88,25	110,58	100,73	117,29	102,57	93,75
Amplitude	84,97	131,94	88,25	108,67	98,82	114,98	101,53	93,75
Q1	14,09	33,42	13,51	23,21	24,03	34,13	16,40	20,69
Q3	26,42	56,21	31,03	42,91	44,29	54,39	34,20	37,23
IQR (Q3-Q1)	12,33	22,79	17,52	19,70	20,26	20,26	17,80	16,54
Valores Ausentes	286,00	531,00	505,00	536,00	764,00	659,00	470,00	611,00

Fonte: Autores (2023).

Ao observar os valores da Tabela 5, obtém-se uma confirmação das observações destacadas das análises visuais do poluente O3 nas estações de medida. É notável também que algumas estações apresentam um desvio padrão maior do que outras, como por exemplo "AV" (Copacabana). Os Boxplots evidenciam os diferentes valores da média (ponto vermelho), mediana, assim como dispersão ao comprar as estações.

Destaca-se a grande quantidade de valores ausentes em todas as estações , bem como a presença de valores extremos. Sendo assim, percebe-se a necessidade de um tratamento para

esses dados através da remoção dos valores ausentes bem como a remoção dos valores extremos, de maneira a não comprometer os testes de normalidade e homogeneidade das variâncias posteriores.

Após todas as análises visuais, bem como numéricas para os poluentes considerando as estações de medida de cada um deles, consegue-se perceber que provavelmente há uma diferença entre as estações em relação ao comportamento dos dados.

4.2 Análise Regional

Com a finalidade de lançar luz sobre o comportamento dos poluentes entre as estações, será efetuada uma análise regional através da abordagem ANOVA. Primeiramente, implementou-se a etapa de remoção dos valores extremos e ausentes que foram identificados a partir da etapa anterior de análise descritiva. Uma vez tratados, os dados encontraram-se prontos para serem utilizados na parte de aplicação dos testes. A análise regional separou novamente cada um dos poluentes para serem submetidos a testes que ajudaram na compreensão do comportamento desses dados analisados.

Para os poluentes PM10, NO2 e O3 foram realizados os testes de normalidade, a partir da aplicação do teste de Shapiro-Wilk, e os testes de homogeneidade das variâncias, a partir do teste de Levene, gerando assim os seguintes resultados expressos nas Tabelas 6 até 8, respectivamente.

Tabela 6 - Resultados dos testes de Shapiro-Wilk e Levene para o poliente PM10 por estação.

Estações	Valor W Teste de Shapiro-Wilk	P-Valor Teste de Shapiro-Wilk	Aceita Ho para Shapiro-Wilk?	Graus de Liberdade Teste de Levene	Valor F Teste de Levene	P-Valor Teste de Levene	Aceita Ho para Levene?
AV	0,9748	$2{,}2e^{-16}$	Não	7	60,391	$2{,}2e^{-16}$	Não
BG	0,9657	$2{,}2e^{-16}$	Não				
CA	0,9783	$2{,}2e^{-16}$	Não				
CG	0,9652	$2{,}2e^{-16}$	Não				
IR	0,9584	$2{,}2e^{-16}$	Não				
PG	0,9726	$2{,}2e^{-16}$	Não				
SC	0,9597	$2{,}2e^{-16}$	Não				
SP	0,9732	$2{,}2e^{-16}$	Não				

Fonte: Autores (2023).

Tabela 7 - Resultados dos testes de Shapiro-Wilk e Levene para NO2 por estação.

Estações	Valor W Teste de Shapiro-Wilk	P-Valor Teste de Shapiro-Wilk	Aceita Ho para Shapiro-Wilk?	Graus de Liberdade Teste de Levene	Valor F Teste de Levene	P-Valor Teste de Levene	Aceita Ho para Levene?
BG	0,9784	$2{,}2e^{-16}$	Não	3	436,14	$2{,}2e^{-16}$	Não
CG	0,9932	$9{,}298e^{-8}$	Não				
IR	0,9878	$1{,}258e^{-13}$	Não				
SP	0,9679	$2{,}2e^{-16}$	Não				

Fonte: Autores (2023).

Tabela 8 - Resultados dos testes de Shapiro-Wilk e Levene para O3 por estação.

Estações	Valor W Teste de Shapiro-Wilk	P-Valor Teste de Shapiro-Wilk	Aceita Ho para Shapiro-Wilk?	Graus de Liberdade Teste de Levene	Valor F Teste de Levene	P-Valor Teste de Levene	Aceita Ho para Levene?
AV	0,9792	$2{,}2e^{-16}$	Não	7	147,41	$2{,}2e^{-16}$	Não
BG	0,9858	$2{,}2e^{-16}$	Não				
CA	0,9584	$2{,}2e^{-16}$	Não				
CG	0,9866	$3{,}212e^{-16}$	Não				
IR	0,9850	$2{,}2e^{-16}$	Não				
PG	0,9859	$3{,}397e^{-16}$	Não				
SC	0,9747	$2{,}2e^{-16}$	Não				
SP	0,9857	$2{,}2e^{-16}$	Não				

Fonte: Autores (2023).

Os resultados das Tabelas 6 – 7 indicaram que para todos os poluentes a hipótese nula em todas as estações deve ser rejeitada, ou seja, os dados não seguem uma distribuição normal uma vez que o p-valor de todas as estações de todos os poluentes resultou em valores menores

do que o nível de significância $\alpha = 5\%$. Por sua vez, para o teste de Levene para todos os poluentes, devemos rejeitar também a hipótese nula uma vez que o p-valor de todos os testes de Levene resultou em um valor menor valor do que $\alpha = 5\%$ ou seja, a homogeneidade das variâncias não foram aceitas.

Sendo assim, deve-se então seguir para todos os poluentes com a aplicação da ANOVA não paramétrica, isto é aplicar o teste de Kruskal-Wallis visto que todos os pressupostos de normalidade e homogeneidade da variância dos dados foram violados. Desta forma, os resultados para todos os pololuentes se encontram na Tabela 9.

Tabela 9 - Resultados do teste de Kruskal-Wallis para os poluentes PM10, NO2 e O3 por estação.

Poluente	**Qui-Quadrado**	**Graus de Liberdade**	**P-VALOR**
PM10	4050,4	7	$2{,}2e^{-16}$
NO2	2049,8	3	$2{,}2e^{-16}$
O3	7083,7	7	$2{,}2e^{-16}$

Fonte: Autores (2023).

Esse teste foi o responsável por realizar uma comparação entre grupos "ANOVA de 1 via" para dados não normais e heterocedásticos. Sendo assim após a aplicação do mesmo ele revelou em todos os poluentes um p-valor menor do que o nível de significância $\alpha = 5\%$, isso significa que existe evidência estatística de que pelo menos um dos grupos é diferente dos demais. Porém, não se tem a informação de qual ou quais são diferentes.

Em virtude do exposto no parágrafo anterior, o próximo passo foi a realização do teste de post-hoc de Dunn para os três poluentes, que tem como objetivo encontrar quais são os grupos distintos, isto é, identificar quais são as estações que diferem entre si, para cada tipo de poluente. Os resultados da aplicação do post-hoc de Dunn para as estações dos poluentes PM10, NO2 e O3 estão contidos nas Tabelas 10 – 12.

A Tabela 10 destaca, para o poluente PM10, apenas os grupos comparados entre si que ***não rejeitaram*** a hipótese nula, ou seja, obtiveram um p-valor maior do que o nível de significância de 5%. Todos os demais grupos apresentaram um p-valor menor do que o nível de significância alpha = 5%, rejeitando a hipótese nula. Sendo assim, esses grupos em destaque são os que não possuem os mesmos postos médios. Nesse caso, dentre um total de 28 grupos, os 11 grupos listados destacam-se por terem um compartamento semelhante. Aqui percebe-se que a estação cujo comportamento é semelhante às outras é a "CA" (Centro), seguida de "CG"

(Campo Grande) e "PG" (Pedra de Guaratiba). Além disso a estações "BG" (Bangu) e "IR" (Irajá), são estatísticamente iguais no que respeita ao comportamento do poluente PM10.

Tabela 10 - Resultados do teste de post-hoc de Dunn que aceitam Ho para o poluente PM10 por estação.

Poluente	Grupo 1	Grupo 2	Estatística de teste	P-VALOR
PM10	BG	IR	1,06381108	1,00
PM10	CA	CG	-2,2143	$7{,}505e^{-1}$
PM10	CA	PG	0,08512	1,00
PM10	CA	SC	-2,2362	$7{,}094e^{-1}$
PM10	CA	SP	-2,7525	$1{,}656e^{-1}$
PM10	CG	PG	2,2095	$7{,}599e^{-1}$
PM10	CG	SC	0,0468	1,00
PM10	CG	SP	-0,4019	1,00
PM10	PG	SC	-2,2276	$7{,}254e^{-1}$
PM10	PG	SP	-2,7192	$1{,}832e^{-1}$
PM10	SC	SP	-0,4645	1,00

Fonte: Autores (2023).

Tabela 11 - Resultados do teste de post-hoc de Dunn que aceitam Ho para o poluente NO2 por estação.

Poluente	Grupo 1	Grupo 2	Estatística de teste	P-VALOR
NO2	IR	SP	-1,1242	1,00

Fonte: Autores (2023).

Na Tabela 11 dentre um total de 6 grupos, o único grupo que não rejeita a hipótese nula corresponde ás estações "IR" (Irajá) e "SP" (Tijuca), o que implica que estas são as estações cujo comportamento mediano do poluente é semelhante (veja Figura 6).

Tabela 12 - Resultados do teste de post-hoc de Dunn que aceitam Ho para o poluente O3 por estação.

Poluente	Grupo 1	Grupo 2	Estatística de teste	P-VALOR
O3	BG	PG	06742	1,00
O3	CG	IR	2,3020	$5{,}973e^{-l}$

Fonte: Autores (2023).

A Tabela destaca apenas os grupos comparados entre sí que ***não rejeitaram*** a hipótese nula, ou seja, obtiveram um p-valor maior do que o nível de significância de 5%. Sendo assim, dentre um total de 28 grupos, somente os 2 grupos listados destacam-se por possuírem os mesmos postos médios, o que implica que para "BG" (Bangu) e "PG" (Pedra de Guaratiba), e "CG" (Campo Grande) e "IR" (Irajá) o poluente O3 tem comportamento mediano semelhante (veja Figura 9).

Os resultados aqui expostos indicam que há, de uma maneira geral, uma diferença significativa na concentração dos poluentes entre as regiões de medição da mesma. Esta divergência pode ser explicada pela proximidade de algumas destas estações com vias de alto fluxo e com indústrias, vide estação de Bangu cuja concentração de O3 é a maior entre as estações. Outro fator que poderia implicar em uma maior concentração de poluentes é a própria localização e formação geográfica. A Tijuca, por exemplo, é um bairro cercado por montanhas, o que pode indicar uma dificuldade na fuga desses poluentes, que pode ser representado pela maior concentração de NO2 entre todas as estações.

4.3 Análise Temporal

Com a finalidade de compreender a evolução temporal dos poluentes foi efetuada una análise temporal considerando a abordagem ANOVA na qual os anos correspondem aos níves de cada fator sob análise, isto é, o níveis para cada poluente. Novamente, cada um dos poluentes foi submetido a testes que ajudaram na compreensão do comportamento desses dados analisados. Vale ressaltar que para essa etapa a variável fator em questão é a variável "ano", ou seja, as etapas da análise temporal seguem os mesmos passos da análise regional porém, considerando agora os anos das medidas dos poluentes. Sendo assim, o poluente PM2.5, como foi medido para todos os anos considerados, será considerado nessa parte da análise.

Para os poluentes PM10, PM2.5, NO2 e O3, seus respectivos testes de normalidade, a partir da aplicação do teste de Shapiro, e os respectivos testes de homogeneidade das variâncias, a partir do teste de Levene, geraram os seguintes resultados expressos nas Tabelas 13 – 16.

Tabela 13 - Resultados dos testes de Shapiro-Wilk e Levene para o poluente PM10 por ano.

Ano	Valor W Teste de Shapiro-Wilk	P-Valor Teste de Shapiro-Wilk	Aceita Ho para Shapiro-Wilk?	Graus de Liberdade Teste de Levene	Valor F Teste de Levene	P-Valor Teste de Levene	Aceita Ho para Levene?
2012	0,9430	0,5561	Sim	9	3,7908	$1{,}058e^{-4}$	Não
2013	0,9458	0,3076	Sim				

2014	0,9313	0,0178	Não				
2015	0,9776	0,0373	Não				
2016	0,9630	0,0388	Não				
2017	0,9562	$1{,}169e^{-4}$	Não				
2018	0,9302	$3{,}293e^{-7}$	Não				
2019	0,9919	0,9179	Sim				
2020	0,9664	0,0319	Não				
2021	0,9698	0,0123	Não				

Fonte: Autores (2023).

Tabela 14 - Resultados dos testes de Shapiro-Wilk e Levene para o poluente PM2.5 por ano.

Ano	Valor W Teste de Shapiro-Wilk	P-Valor Teste de Shapiro-Wilk	Aceita Ho para Shapiro-Wilk?	Graus de Liberdade Teste de Levene	Valor F Teste de Levene	P-Valor Teste de Levene	Aceita Ho para Levene?
2012	0,9222	$1{,}537e^{-6}$	Não	9	3,5243	$2{,}407e^{-4}$	Não
2013	0,9443	$5{,}590e^{-5}$	Não				
2014	0,9547	$3{,}324e^{-6}$	Não				
2015	0,9708	$5{,}507e^{-5}$	Não				
2016	0,9506	$2{,}998e^{-6}$	Não				
2017	0,9237	$4{,}462e^{-9}$	Não				
2018	0,9330	$1{,}899e^{-9}$	Não				
2019	0,9436	$2{,}073e^{-7}$	Não				
2020	0,9729	$6{,}610e^{-4}$	Não				
2021	0,9630	$5{,}122e^{-5}$	Não				

Fonte: Autores (2023).

Tabela 15 - Resultados dos testes de Shapiro-Wilk e Levene para o poluente NO2 por ano.

Ano	Valor W Teste de Shapiro-Wilk	P-Valor Teste de Shapiro-Wilk	Aceita Ho para Shapiro-Wilk?	Graus de Liberdade Teste de Levene	Valor F Teste de Levene	P-Valor Teste de Levene	Aceita Ho para Levene?
2012	0,9697	0,453	Sim	8	8,7835	$2{,}596e^{-11}$	Não
2013	0,9595	$6{,}358e^{-3}$	Não				
2014	0,9707	$2{,}287e^{-3}$	Não				

2015	0,9668	0,566	Sim				
2016	0,9618	$4{,}534e^{-2}$	Não				
2017	0,9852	0,871	Sim				
2018	0,9743	0,141	Sim				
2019	0,9403	0,057	Sim				
2020	0,9492	0,260	Sim				

Fonte: Autores (2023).

Tabela 16 - Resultados dos testes de Shapiro-Wilk e Levene para poluente O3 por ano.

Ano	**Valor W Teste de Shapiro-Wilk**	**P-Valor Teste de Shapiro-Wilk**	**Aceita Ho para Shapiro-Wilk?**	**Graus de Liberdade Teste de Levene**	**Valor F Teste de Levene**	**P-Valor Teste de Levene**	**Aceita Ho para Levene?**
2012	0,9843	0,686	Sim	9	4,2371	$2{,}081e^{-5}$	Não
2013	0,9840	0,192	Sim				
2014	0,9689	$3{,}199e^{-3}$	Não				
2015	0,9626	$1{,}492e^{-4}$	Não				
2016	0,9947	0,901	Sim				
2017	0,9724	$4{,}022e^{-2}$	Não				
2018	0,9757	$8{,}958e^{-2}$	Não				
2019	0,9610	$7{,}656e^{-2}$	Não				
2020	0,9923	0,898	Sim				
2021	0,9887	0,571	Sim				

Fonte: Autores (2023).

Os resultados das Tabelas 13 – 16 indicaram que, para todos os poluentes, deveríamos rejeitar a hipótese nula na maioria das estações, ou seja, a normalidade não foi aceita, uma vez que o p-valor de pelo menos uma das estações dos poluente resultou em valores menores do que 5%. De forma semelhante, devemos rejeitar também a hipótese nula do teste de Levene uma vez que o p-valor resultou em um valor menor do que 5% para todos os poluentes, ou seja, a homogeneidade das variâncias não foi aceita.

Sendo assim, deve-se então fazer uso da ANOVA não paramétrica. Ao aplicar o teste de Kruskal-Wallis para todos os poluentes, obten-se os resultados que estão apresentados na Tabela 17:

Tabela 17 - Resultados do teste de Kruskal-Wallis para os poluentes PM10, PM2.5, NO2 e O3 por ano.

Poluente	**Qui-Quadrado**	**Graus de Liberdade**	**P-VALOR**
PM10	37,576	9	$2{,}077e^{-5}$
PM2.5	65,376	9	$1{,}219e^{-10}$
NO2	127,36	8	$2{,}2e^{-16}$
O3	156,47	9	$2{,}2e^{-16}$

Fonte: Autores (2023).

Esse teste foi o responsável por realizar uma comparação entre grupos "ANOVA de 1 fator" para dados não normais e heterocedásticos. Sendo assim, após a aplicação do mesmo para todos os poluentes, observou-se um p-valor menor que $\alpha = 5\%$. Isso significou então que existem evidências de que pelo menos um dos grupos de cada poluente é diferente dos demais, isto é, pelo menos um ano é diferente dos outros. Porém, não se tem a informação de qual ou quais são diferentes.

O próximo passo foi a realização do teste de post-hoc de Dunn para todos os poluentes, com o objetivo de encontrar quais os grupos distintos. O resultado da aplicação dos testes post-hoc de Dunn para todos os anos de todos os poluentes estão demonstrados nas Tabela 18 – 21.

Tabela 18 - Resultados do teste de post-hoc de Dunn que rejeitam Ho para o poluente PM10 por ano.

Poluente	**Grupo 1**	**Grupo 2**	**Estatística**	**P-VALOR**
PM10	2015	2020	-4,6778	$1{,}305e^{-4}$
PM10	2016	2020	-4,1359	$1{,}590e^{-3}$
PM10	2017	2020	-4,4888	$3{,}222e^{-4}$

Fonte: Autores (2023).

A Tabela 18 destaca apenas os grupos comparados entre si para o PM10 que ***rejeitaram*** a hipótese nula. Todos os demais grupos apresentaram um p-valor maior do que o nível de significância, aceitando a hipótese nula. Sendo assim, esses grupos em destaque são os que não anualmente, para o poluente PM10, não são iguais sob H_o. Nesse caso, pode se observar que

ano de "2020" apresenta diferenças na mediana do poluente PM10 em relação aos anos "2015", "2016" e "2017". Para os anos restantes e suas combinações, não há evidências estatísticas que indiquem que há diferença entre os anos.

Tabela 19 - Resultados do teste de post-hoc de Dunn que rejeitam Ho para o poluente PM2.5.

Poluente	**Grupo 1**	**Grupo 2**	**Estatística**	**P-VALOR**
PM2.5	2012	2013	3,3994	$3{,}038e^{-2}$
PM2.5	2012	2016	4,0706	$2{,}109e^{-3}$
PM2.5	2013	2015	-3,4459	$2{,}561e^{-2}$
PM2.5	2013	2019	-4,7883	$7{,}568e^{-5}$
PM2.5	2013	2020	-3,5458	$1{,}761e^{-2}$
PM2.5	2013	2021	-3,3634	$3{,}464e^{-2}$
PM2.5	2014	2015	-3,2964	$4{,}405e^{-2}$
PM2.5	2014	2019	-4,8375	$5{,}916e^{-5}$
PM2.5	2014	2020	-3,3948	$3{,}090e^{-2}$
PM2.5	2015	2016	4,3168	$7{,}123e^{-4}$
PM2.5	2016	2019	-5,7953	$3{,}067e^{-7}$
PM2.5	2016	2020	-4,3659	$5{,}695e^{-4}$
PM2.5	2016	2021	-4,1541	$1{,}469e^{-3}$
PM2.5	2017	2019	-3,6039	$1{,}410e^{-2}$
PM2.5	2018	2019	-3,4314	$2{,}702e^{-2}$

Fonte: Autores (2023).

A Tabela 19 destaca apenas os grupos comparados entre sí para o PM2.5 que rejeitaram a hipótese nula, ou seja, obtiveram um p-valor menor do que o nível de significância de 5%. Dentre os 45 grupos totais, os 15 grupos listados destacam-se devido a possuírem uma diferença significativa de comportamento mediano do poluente. Aqui pode ser visto que os anos que apresentem maiores diferenças do poluente correspondem a: 2015, 2019, 2020 e 2021.

Levando em conta o mesmo poluente PM2.5, os gráficos gerados estão representados nas Figuras 11 e 12 representado, respectivamente, a visualização do gráfico de linhas de sua única estação e o Boxplot por ano, medidos em [$\mu g/m3$].

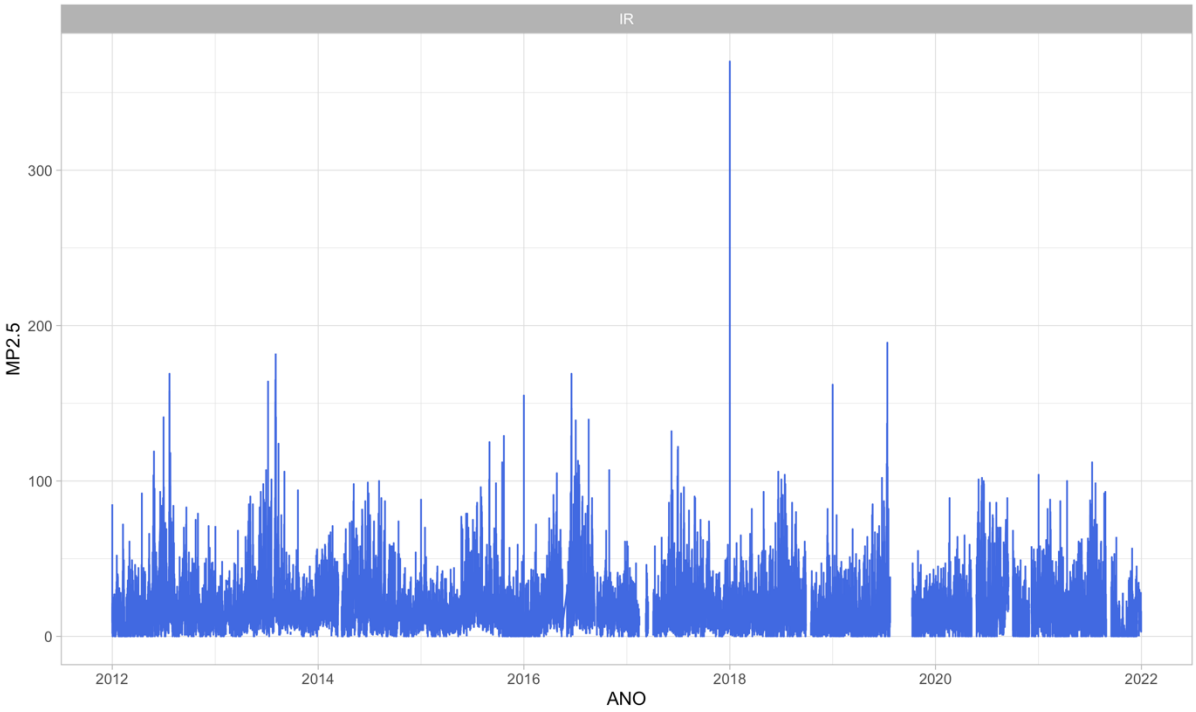

Figura 12- Série temporal das medições do poluente PM2.5 na estação "IR" (Irajá). Fonte: Autores (2023).

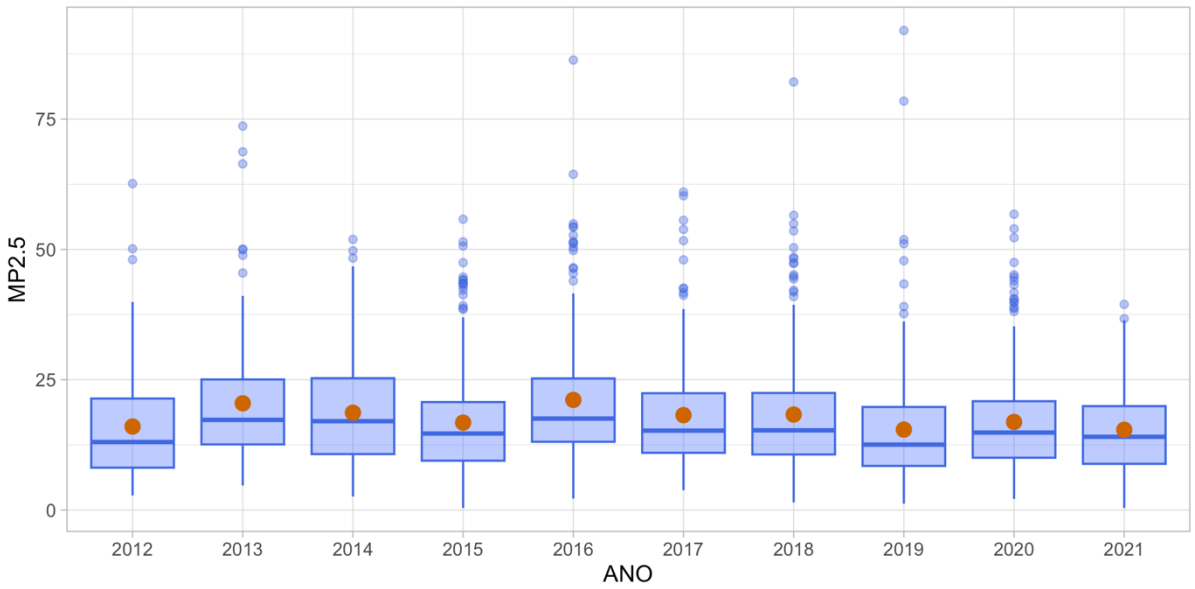

Figura 13 - Gráfico de Boxplot do poluente PM2.5 por ano. Fonte: Autores (2023).

Para o PM2.5 é possível notar que o poluente é medido apenas na estação “IR” (Irajá). Pela série temporal percebe-se que, similar aos outros poluentes há ausência de dados na estação, principalmente considerando os anos de 2019 e 2020. Detecta-se novamente a presença de valores extremos em algumas faixas de tempo.

Para o Boxplot de todos os anos do poluente PM2.5 é possível identificar a presença de valores extremos na maioria dos anos considerados, bem como observa-se uma distribuição das medidas do poluentes concentradas dentro da faixa da de 12,5 [$\mu g/m3$] até 25 [$\mu g/m3$] para todos os anos. Cabe destacar que no ano de 2016 o comportamento mediano do poluente atingiu seu valor mais elevado e no ano de 2021, evidenciou-se o menor valor médio.

Tabela 20 - Resultados do teste de post-hoc de Dunn que rejeitam Ho para o poluente NO2 por ano.

Poluente	**Grupo 1**	**Grupo 2**	**Estatística**	**P-VALOR**
NO2	2012	2014	4,0777	$1{,}637e^{-3}$
NO2	2012	2016	3,6638	$8{,}945e^{-3}$
NO2	2012	2017	3,8768	$3{,}809e^{-3}$
NO2	2013	2014	8,5707	$3{,}702e^{-16}$
NO2	2013	2015	5,2336	$5{,}983e^{-6}$
NO2	2013	2016	6,9762	$1{,}091e^{-10}$
NO2	2013	2017	6,6675	$9{,}362e^{-10}$
NO2	2013	2018	6,3220	$9{,}294e^{-9}$
NO2	2014	2020	-5,8168	$5{,}997e^{-7}$
NO2	2015	2020	-4,6368	$1{,}273e^{-4}$
NO2	2016	2020	-5,3534	$3{,}106e^{-6}$
NO2	2017	2020	-5,4538	$1{,}774e^{-6}$
NO2	2018	2020	-4,8305	$1{,}469e^{-5}$

Fonte: Autores (2023).

Estão reportados na Tabela 20 apenas os grupos comparados entre si para o poluente NO2 que rejeitaram a hipótese nula. Pode identificado que o ano em que o poluente tem um comportamento diferente dos outros é "2013" e "2020"se comparado com os anos de 2014 a 2018. Alguns anos de 2012 também possuem diferença; contudo, os pares de grupos (anos)

restantes não rejeitam H_o, o que significa que os poluentes medianos nesses grupos são iguais. Observa-se que 2012, 2013, 2020 e 2021 são os anos que ao comparar suas medianas (Figura 7) com os anos restantes, visualmente possuem uma diferença mais notável, o que é ratificado estatísticamente nos resultados da Tabela 20.

Tabela 21 - Resultados do teste de post-hoc de Dunn que rejeitam Ho para o poluente O3 por ano.

Poluente	**Grupo 1**	**Grupo 2**	**Estatística**	**P-VALOR**
O3	2012	2019	6,22	$2{,}21e^{-8}$
O3	2012	2020	4,64	$1{,}52e^{-4}$
O3	2012	2021	3,76	$7{,}63e^{-3}$
O3	2013	2019	8,12	$1{,}95e^{-14}$
O3	2013	2020	6,67	$1{,}13e^{-9}$
O3	2013	2021	5,70	$5{,}33e^{-7}$
O3	2014	2019	8,58	$4{,}16e^{-16}$
O3	2014	2020	7,19	$2{,}88e^{-11}$
O3	2014	2021	6,21	$2{,}33e^{-8}$
O3	2015	2019	7,61	$1{,}15e^{-12}$
O3	2015	2020	6,04	$6{,}61e^{-8}$
O3	2015	2021	4,98	$2{,}77e^{-5}$
O3	2016	2019	7,81	$2{,}47e^{-13}$
O3	2016	2020	6,29	$1{,}38e^{-8}$
O3	2016	2021	5,27	$5{,}83e^{-6}$
O3	2017	2019	5,57	$1{,}13e^{-6}$
O3	2017	2020	3,75	$7{,}67e^{-3}$
O3	2018	2019	5,26	$6{,}27e^{-6}$
O3	2018	2020	3,42	$2{,}75e^{-2}$
O3	2019	2021	-3,30	$4{,}23e^{-2}$

Fonte: Autores (2023).

Por fim, para o poluente O3 e de forma semelhante às análises dos poluentes anteriores, na Tabela 21, destaca-se apenas os grupos comparados entre si que rejeitaram a hipótese nula,

ou seja, obtiveram um p-valor menor do que o nível de significância de 5%. Aqui pode ser evidenciado que os anos "2019", "2020" e "2021" foram os que se comparados com outros anos, tiveram diferenças estatística no comportamento mediano do poluente O3, pois como pode ser visto na Figura 10, esse anos apresentam valores mais elevados do poluente.

Aqui, os resultados elucidados indicam que, de uma maneira geral, não houve diferença significativa na concentração dos poluentes entre os anos de medição, porém há como identificar uma alta na concentração de O3 nos últimos anos (principalmente em 2019) e uma diminuição na concentração de NO2 nos anos de 2020 e 2021 (além de 2012 e 2013 que também possuíram baixa). Esta diminuição na concentração de NO2 pode estar relacionada com a pandemia da COVID-19, que resultou no “lockdown” da população na cidade do Rio de Janeiro nos anos de 2020 e 2021, justamente os anos em que a concentação foi a menor. Com o confinamento, houve uma drástica redução no tráfego de automóveis e, com isso, na queima de combustíveis, que, por sua vez, é a principal atividade emissora de NO2.

A pouca divergência entre a concentração dos poluentes ano a ano pode ser explicada pelo período curto em que os dados aqui analisados estão compreendidos, podendo ser foco de trabalhos futuros uma análise que faça abranger décadas e décadas de dados, com o intuito de explicitar um possível aumento na concentração dos poluentes ao longo dos anos.

5. CONCLUSÕES

Este trabalho teve como objetivo compreender o comportamento dos poluentes PM10, PM2.5, NO2 e O3 que foram medidos em estações espalhadas pela cidade do rio de janeiro ao longo de uma faixa de 9 anos (2012 – 2021) e examinar as possíveis relações entre o comportamento das medidas desses poluentes com as estações de medição bem como com os anos.

As análises descritivas, forneceram um embasamento para que pudéssemos estabelecer uma prévia das possibilidades dos comportamentos dos dados, bem como, ajudaram na identificação tanto de valores ausentes, quanto de valores extremos.

Através da aplicação dos testes de Shapiro-Wilk e Levene para cada um dos poluentes foi possível comprovar que nenhum deles seguia uma distribuição normal nem possuía uma homogeneidade com relação às variâncias. Sendo assim, necessitou-se implementar ANOVA não paramétrica. Para a parte de análise regional, a maioria dos grupos comparados apresentou um comportamento distinto entre si para todos os poluentes considerados, já na parte de análise temporal, alguns conjuntos de grupos de anos apresentaram comportamentos semelhantes entre si, mas se observa mudanças mais marcantes do poluentes nos últimos anos, como o NO2 e O3, e menos marcantes, mas presentes, no PM10.

Através da aplicação desses testes para cada um dos poluentes foi possível comprovar as suposições, uma vez que obtivemos a confirmação, a partir dos resultados estatísticos, que o que havia sido apenas observado de maneira visual pelos gráficos e tratado apenas como suposição até aquela etapa, realmente ocorria na prática.

O que ocorreu na prática com relação aos dados era que em quase todas as regiões consideradas, as medições dos poluentes apresentavam um comportamento bastante distinto para cada uma delas. Esse ponto pode ser explicado por diversos fatores como por exemplo as condições territoriais distintas de cada uma das estações, às proximidades de determinadas estações com vias de alta intensidade de tráfego de automóveis, proximidade com fábricas e indústrias.

Já com relação aos anos nos quais os poluentes foram medidos, os comportamentos dessas medições variaram muito pouco de ano para ano. Sendo assim, admitimos que para os anos, houve uma interferência leve de fatores externos capaz de promover algumas alterações de medidas desses poluentes.

Por fim, quanto ao tema de trabalhos futuros pode-se destacar e mencionar que ha uma grande possibilidade de vertentes para o aprofundamento do tema que foi tratado nesse trabalho. Uma vertente possivel para futuras analises poderia ser a aplicação de uma ANOVA de duas vias ao inves da ANOVA de uma via que foi implementada. Sendo assim, deveria-se considerar como os dois fatores conjuntos as estações de medição dos poçuentes, bem como os anos. Com isso, poderia ser possível a realização de uma observação das medições dos poluentes ao longo dos anos dentro de cada estação, sendo possivel obter uma visão va variação das medições dos poluentes ao longo dos anos dentre de todas as estações consideradas.

6. REFERÊNCIAS BIBLIOGRÁFICAS

Compreender os Poluentes do Ar: Um Guia para a Matéria Particulada (PM). Disponível em: <https://www.crowcon.com/pt/blog/understanding-air-pollutants-a-guide-to-particulate-matter-pm/>. Acesso em: 20 set. 2023.

ECYCLE, E. Conheça os perigos do material particulado. Disponível em: <https://www.ecycle.com.br/material-particulado/>. Acesso em: 20 set. 2023.

FOX, John; WEISBERG, Sanford. An R Companion to Applied Regression. 3. ed. Thousand Oaks Ca: Sage, 2019.

HYNDMAN, Rob J.; FAN, Yanan. Sample Quantiles in Statistical Packages. The American Statistician, [S.L.], v. 50, n. 4, p. 361, nov. 1996. JSTOR. http://dx.doi.org/10.2307/2684934.
INEA. RELATÓRIO DA QUALIDADE DO AR DO ESTADO DO RIO DE JANEIRO: ano base 2018. Rio de Janeiro: Secretaria de Estado do Ambiente e Sustentabilidade, 2018. 163 p. Disponível em: https://www.inea.rj.gov.br/wp-content/uploads/2020/11/relatorio-qualidade-ar-2018.pdf. Acesso em: 10 nov. 2023.

LAZZARI, Angela Radünz. Comparação de técnicas estatísticas para analisar a relação entre doenças respiratórias e concentrações de poluentes atmosféricos. Ciência e Natura, Santa Maria, v. 35, n. 1, p. 098-105, jul. 2013.

MANRESA, A. P. Machine Learning to Predict High-Cost Hospitalizations. Dissertação (Mestrado). Departamento de Engenharia Industrial. Pontifícia Universidade Católica do Rio de Janeiro. Disponível em https://www.maxwell.vrac.puc-rio.br/49137/49137.PDF&ved=2ahUKEwjC1c7iyPb6AhWdA7kGHamfBM4QFnoECAcessado em: 01 de out. de 2023

MARTINS, D. Partículas Inaláveis - PM10 e a Qualidade do Ar Ambiente. Disponível em: <https://www.apopartner.pt/particulas-inalaveis-pm10-e-a-qualidade-do-ar-ambiente/>. Acesso em: 20 set. 2023.

MICRONICS. Entendendo a PM 2.5 e o Controle da Poluição Atmosférica. Disponível em: <https://www.micronicsinc.com/pt-br/filtration-news/particulate-matter/>. Acesso em: 20 set. 2023.

MONTORO, Edson Rui. Teste de Kruskal–Wallis. 2020. Disponível em: https://www.ermontoro.com/post/teste-de-kruskal-wallis. Acesso em: 18 out. 2023.

REIS, Edna Afonso; REIS, Ilka Afonso. Análise Descritiva de Dados. Belo Horizonte: Departamento de Estatística da Ufmg, 2002.

SCIENTIFIC, Industrial. Industrial Scientific: detectores de gás para dióxido de nitrogênio (no2). Detectores de gás para dióxido de nitrogênio (NO2). Disponível em: https://www.indsci.com/pt/industrial-scientific-tipos-de-g%C3%A1s-di%C3%B3xido-de-nitrog%C3%AAnio-no2#:~:text=O%20di%C3%B3xido%20de%20nitrog%C3%AAnio%20(NO2,diesel)%20e%20as%20centrais%20t%C3%A9rmicas. Acesso em: 10 out. 2023.

SHARMA, Shikha. O QUE É MATÉRIA PARTICULAR PM2.5? FONTES | IMPACTOS. 2021. Disponível em: https://www.pranaair.com/pt-pt/blog/particulate-matter-pm-2-5-sources-impacts-measures/. Acesso em: 08 out. 2023.

SOUZA, Amaury de; IKEFUT, Priscilla; GARCIA, Ana Paula; SANTOS, Debora; OLIVEIRA, Soetânia de. ANÁLISE DA RELAÇÃO ENTRE O3, NO E NO2 USANDO TÉCNICAS DE REGRESSÃO LINEAR MÚLTIPLA. Rio de Janeiro: Geographia, 2018.

R CORE TEAM. R: A language and environment for statistical computing. R Foundation for Statistical Computing. 2023. Disponível em: https://www.R-project.org/. Acesso em: 25 out. 2023.

RStudio TEAM: Integrated Development for R. RStudio. 2020. Disponível em: http://www.rstudio.com/. Acesso em: 24 out. 2023.

VIEIRA, Sonia. Bioestatística: tópicos avançados. 4. ed. Rio de Janeiro: Elsevier, 2018.

VIEIRA, Sonia. Introdução à Bioestatística. 6. ed. Rio de Janeiro: Gen Guanabara Koogan, 2021.

WICKHAM, Hadley; AVERICK, Mara; BRYAN, Jennifer; CHANG, Winston; MCGOWAN, Lucy D'agostino; FRANÇOIS, Romain; GROLEMUND, Garrett; HAYES, Alex; HENRY, Lionel; HESTER, Jim. Welcome to the tidyverse. Journal Of Open Source Software. Berkeley, nov. 2019.

WICKHAM, Hadley. Ggplot2: Elegant Graphics for Data Analysis. 2016. Disponível em: https://ggplot2.tidyverse.org. Acesso em: 25 out. 2023.

WICKHAM, Hadley; HESTER, Jim; BRYAN, Jennifer. Readr: Read Rectangular Text Data. 2023. Disponível em: https://readr.tidyverse.org. Acesso em: 25 out. 2023.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H. 2019. "Welcome to the tidyverse." Journal of Open Source Software, 4(43), 1686. doi:10.21105/joss.01686.

ZAR, Jerrold. Biostatistical Analysis. 5. ed. London: Pearson, 2014