



Universidade do Estado do Rio de Janeiro

Centro de Tecnologia e Ciências

Faculdade de Engenharia

Rafael Corrêa Gama de Oliveira

**Previsão de ozônio troposférico na Região Metropolitana do Rio de Janeiro
com base em técnicas de imputação de dados faltantes e calibração
multivariada**

Rio de Janeiro

2022

Rafael Corrêa Gama de Oliveira

Previsão de ozônio troposférico na Região Metropolitana do Rio de Janeiro com base em técnicas de imputação de dados faltantes e calibração multivariada

Tese apresentada, como requisito final para a obtenção do título de Doutor, ao Programa de Pós-Graduação em Engenharia Ambiental, da Universidade do Estado do Rio de Janeiro. Área de Concentração: Diagnóstico, Monitoramento e Modelagem Ambiental.

Orientador: Prof. Sergio Machado Corrêa

Coorientador: Prof. Alexandre Rodrigues Tôrres

Rio de Janeiro

2022

CATALOGAÇÃO NA FONTE
UERJ / REDE SIRIUS / BIBLIOTECA CTC/B

O48	Oliveira, Rafael Corrêa Gama de Previsão de ozônio troposférico na região metropolitana do Rio de Janeiro com base em técnicas de imputação de dados faltantes e calibração multivariada / Rafael Corrêa Gama de Oliveira. – 2022. 144f. Orientador: Sergio Machado Corrêa. Coorientador: Alexandre Rodrigues Tôrres. Dissertação (Mestrado) – Universidade do Estado do Rio de Janeiro, Faculdade de Engenharia. 1. Engenharia ambiental - Teses. 2. Ar - Controle de qualidade - Teses. 3. Ar - Poluição - Teses. 4. Ozônio atmosférico - Teses. 5. Aprendizado do computador - Teses. I. Corrêa, Sergio Machado. II. Tôrres, Alexandre Rodrigues. III. Universidade do Estado do Rio de Janeiro, Faculdade de Engenharia. IV. Título. CDU 628.512:004.89
-----	--

Bibliotecária: Júlia Vieira – CRB7/6022

Autorizo, apenas para fins acadêmicos e científicos, a reprodução total ou parcial desta tese, desde que citada a fonte.

Assinatura

Data

Rafael Corrêa Gama de Oliveira

Previsão de ozônio troposférico na Região Metropolitana do Rio de Janeiro com base em técnicas de imputação de dados faltantes e calibração multivariada

Tese apresentada, como requisito final para a obtenção do título de Doutor, ao Programa de Pós-Graduação em Engenharia Ambiental, da Universidade do Estado do Rio de Janeiro. Área de Concentração: Diagnóstico, Monitoramento e Modelagem Ambiental.

Aprovado em 27 de maio de 2022.

Banca Examinadora:

Prof. Dr. Sergio Machado Corrêa (Orientador)
Faculdade de Tecnologia - UERJ

Prof. Dr. Alexandre Rodrigues Torres (Coorientador)
Faculdade de Tecnologia - UERJ

Profa. Dra. Lilian Lefol Nani Guarieiro
Centro Universitário SENAI - CIMATEC

Profa. Dra. Debora Souza Alvim
Instituto Nacional de Pesquisas Espaciais - INPE

Prof. Dr. Leonardo Baptista
Faculdade de Tecnologia - UERJ

Prof. Dr. Eduardo Monteiro Martins
Faculdade de Tecnologia - UERJ

Rio de Janeiro
2022

AGRADECIMENTOS

Gostaria de agradecer primeiramente a Deus que me proporcionou saúde para terminar este trabalho. Em especial a minha esposa Camilla Lima Cunha, por total suporte e ensinamentos durante a dedicação ao estudo. A minha família como um todo, Selma, Gabriela, Romualdo, meus tios e tias, minha avó Cecília.

É preciso lembrar de agradecer aos Professores Fernando Martins, Sérgio Machado Corrêa e Alexandre Rodrigues Torres pelos ensinamentos e paciência.

Ao INEA pela disponibilização dos dados.

RESUMO

OLIVEIRA, Rafael Corrêa Gama de. *Previsão de ozônio troposférico na região metropolitana do Rio de Janeiro com base em técnicas de imputação de dados faltantes e calibração multivariada*. 145 f. Tese (Doutorado em Engenharia Ambiental) - Faculdade de Engenharia, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2022.

O ozônio troposférico tem um impacto negativo no meio ambiente e, consequentemente na saúde pública. Devido a este problema, este trabalho propõe estudar modelos estatísticos utilizando metodologia de aprendizado de máquina para a previsão de ozônio troposférico na Região Metropolitana do Rio de Janeiro. Técnicas de calibração multivariada baseada nos métodos Regressão de Mínimos Quadrados, Florestas Aleatórias, Máquinas de Vetor de Suporte e Rede Neural Artificial combinados com o algoritmo de imputação *MissForest*, foram aplicados para entender a interação entre ozônio e óxidos de nitrogênio, monóxido de carbono, velocidade do vento, radiação solar, temperatura, umidade relativa, entre outros poluentes, cujos dados foram coletados na Região Metropolitana do Rio de Janeiro em quatro estações de qualidade do ar de diferentes perfis, nas localidades de Adalgisa Nery, Porto das Caixas, Laboratório do INEA e Vila São Luiz, entre 2014 e 2018. Essas técnicas fornecem uma maneira fácil e viável de modelagem e análise de poluentes atmosféricos, a qual podem ser utilizadas e combinadas com outros métodos estatísticos para a validação do modelo. Os resultados mostraram que as técnicas quimiométricas Florestas Aleatórias, Máquinas de Vetor de Suporte e Rede Neural Artificial podem ser usadas na modelagem e previsão de concentrações de ozônio troposférico, com coeficiente de determinação da previsão de até 0,92. O erro quadrático médio de previsão varia entre 4,17 e 22,45 $\mu\text{g m}^{-3}$, dependendo das estações de monitoramento da qualidade do ar e estação do ano.

Palavras-chave: Qualidade do ar. Aprendizado de máquina. Dados faltantes. Linguagem R e Python. Ozônio. Troposfera.

ABSTRACT

OLIVEIRA, Rafael Corrêa Gama de. Forecasts of tropospheric ozone in the metropolitan area of Rio de Janeiro based on missing data imputation and multivariate calibration techniques. 145 f. Tese (Doutorado em Engenharia Ambiental) - Faculdade de Engenharia, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2022.

Tropospheric ozone has a negative impact on the environment and, consequently, on public health. Due to this problem, this work proposes to study statistical models using machine learning methodology to predict tropospheric ozone at the Metropolitan Region of Rio de Janeiro. Multivariate calibration techniques based on Least Squares Regression, Random Forests, Support Vector Machines and Artificial Neural Network combined with the MissForest imputation algorithm were applied to understand the interaction between ozone with nitrogen oxides, carbon monoxide, windy velocity, solar radiation, temperature, relative humidity, among other pollutants, whose data were collected at the Metropolitan Region of Rio de Janeiro at four air quality stations of different profiles, in the localities of Adalgisa Nery, Porto das Caixas, INEA Laboratory and Vila São Luiz, between 2014 and 2018. These techniques provide an easy and feasible way of modeling and analyzing atmospheric pollutants, which can be used and combined with other statistical methods for model validation. The results showed that the chemometric techniques Random Forests, Support Vector Machines and Artificial Neural Network can be used to the modeling and prediction of tropospheric ozone concentrations, with a calculated determination coefficient up to 0.92. The root mean squared error of prediction, ranges between 4.17 and 22.45 $\mu\text{g m}^{-3}$, depending on the air quality monitoring stations and season.

Keywords: Air quality. Machine learning. Missing data. R and Python languages. Ozone. Troposphere.

LISTA DE FIGURAS

Figura 1 - Esquema de decomposição das matrizes X e Y pela aplicação do PLS.....	38
Figura 2 - Esquema típico de uma rede neural artificial	44
Figura 3 - Fluxograma do estudo de previsão do Ozônio	52
Figura 4 - Organograma das ferramentas e métodos quimiométricos utilizados.....	53
Figura 5 - Fluxograma das etapas do estudo.....	54
Figura 6 - Divisão da bacia aérea da Região Metropolitana do Rio de Janeiro	55
Figura 7 - Localização da Estação meteorológica de Adalgisa Nery.....	56
Figura 8 - Localização da Estação meteorológica de INEA	57
Figura 9 - Localização da Estação meteorológica de VSL	57
Figura 10 - Localização da Estação meteorológica de PDC	58
Figura 11 - Boxplot normalizado da EAQA de ADN: dados brutos com todas as observações, antes da retirada de pontos anômalos (A) e dados refinados após a retirada de pontos anômalos (B).....	62
Figura 12 - <i>Boxplot</i> normalizado da EAQA de PDC: dados brutos com todas as observações, antes da retirada de pontos anômalos (A) e dados refinados após a retirada de pontos anômalos (B).....	63
Figura 13 - <i>Boxplot</i> normalizado da EAQA de VSL: dados brutos com todas as observações, antes da retirada de pontos anômalos (A) e dados refinados após a retirada de pontos anômalos (B).....	63
Figura 14 - <i>Boxplot</i> normalizado da EAQA de INE: dados brutos com todas as observações, antes da retirada de pontos anômalos (A) e dados refinados após a retirada de pontos anômalos (B).....	64
Figura 15 - RBOPCA para a EAQA ADN de 2014-2018 para a estação do ano inverno: antes da imputação dos dados (A) e após a imputação (B).....	65
Figura 16 - RBOPCA para a EAQA ADN de 2014-2018 para a estação do ano outono: antes da imputação dos dados (A) e após a imputação (B).....	65
Figura 17 - RBOPCA para a EAQA ADN de 2014-2018 para a estação do ano primavera: antes da imputação dos dados (A) e após a imputação (B).....	66

Figura 18- RBOPCA para a EAQA ADN de 2014-2018 para a estação do ano verão: antes da imputação dos dados (A) e após a imputação (B)	66
Figura 19- RBOPCA para a EAQA PDC de 2014-2018 para a estação do ano inverno: antes da imputação dos dados (A) e após a imputação (B).....	67
Figura 20- RBOPCA para a EAQA PDC de 2014-2018 para a estação do ano outono: antes da imputação dos dados (A) e após a imputação (B)	68
Figura 21- RBOPCA para a EAQA PDC de 2014-2018 para a estação do ano primavera: antes da imputação dos dados (A) e após a imputação (B).....	68
Figura 22- RBOPCA para a EAQA PDC de 2014-2018 para a estação do ano verão: antes da imputação dos dados (A) e após a imputação (B)	69
Figura 23- RBOPCA para a EAQA VSL de 2014-2018 para a estação do ano inverno: antes da imputação dos dados (A) e após a imputação (B).....	70
Figura 24- RBOPCA para a EAQA VSL de 2014-2018 para a estação do ano outono: antes da imputação dos dados (A) e após a imputação (B)	70
Figura 25- RBOPCA para a EAQA VSL de 2014-2018 para a estação do ano primavera: antes da imputação dos dados (A) e após a imputação (B).....	71
Figura 26- RBOPCA para a EAQA VSL de 2014-2018 para a estação do ano verão: antes da imputação dos dados (A) e após a imputação (B)	71
Figura 27- RBOPCA para a EAQA INE de 2014-2018 para a estação do ano inverno: antes da imputação dos dados (A) e após a imputação (B)	72
Figura 28- RBOPCA para a EAQA INE de 2014-2018 para a estação do ano outono: antes da imputação dos dados (A) e após a imputação (B)	73
Figura 29- RBOPCA para a EAQA INE de 2014-2018 para a estação do ano primavera: antes da imputação dos dados (A) e após a imputação (B).....	73
Figura 30- RBOPCA para a EAQA INE de 2014-2018 para a estação do ano verão: antes da imputação dos dados (A) e após a imputação (B)	74
Figura 31 - Comparação entre os valores médios medidos e os valores médios imputados para ADN_out (A), ADN_pri (B), ADN_ver (C), ADN_inv (D)	79
Figura 32- Comparação entre os valores médios medidos e os valores médios imputados para PDC_out (A), PDC_pri (B), PDC_ver (C), PDC_inv (D).....	81
Figura 33- Comparação entre os valores médios medidos e os valores médios imputados para VSL_out (A), VSL_pri (B), VSL_ver (C), VSL_inv (D).....	83

Figura 34- Comparação entre os valores médios medidos e os valores médios imputados para INE_out (A), INE_pri (B), INE_ver (C), INE_inv (D)	85
Figura 35- Distribuição por violino para ADN, PDC, VSL e INE para diferentes estações do ano.....	87
Figura 36- Diagrama de violino ADN, PDC, VSL, INE compiladas para as diferentes estações do ano para o O ₃	88
Figura 37- Diagrama de violino para ADN, PDC, VSL e INE compiladas para as diferentes horas do dia para o O ₃	89
Figura 38- Diagrama violino para os diferentes dias da semana para o O ₃ (µg m ⁻³)	90
Figura 39- Rosa dos ventos para a EAQA ADN: inverno (A), primavera (B), outono (C), verão (D), concentrações de O ₃ em µg m ⁻³	91
Figura 40- Rosa dos ventos para a EAQA PDC: inverno (A), primavera (B), outono (C) verão (D), concentrações de O ₃ em µg m ⁻³	92
Figura 41- Rosa dos ventos para a EAQA VSL: inverno (A), primavera (B), outono (C), verão (D), concentrações de O ₃ em µg m ⁻³	93
Figura 42- Rosa dos ventos para a EAQA INE: inverno (A), primavera (B), outono (C), verão (D), concentrações de O ₃ em µg m ⁻³	94
Figura 43- Modelos mais bem ajustados - Gráfico previsto versus medido para o ozônio para ADN EAQA.....	98
Figura 44- Soma da probabilidade do teste de Wilcoxon (A) e o viés (B) para todos os modelos de previsão do O ₃	99
Figura 45- RMSEP (µg m ⁻³) <i>versus</i> EAQA para todos os modelos de previsão do O ₃ troposférico	99
Figura 46- RMSEP (µg m ⁻³) <i>versus</i> EAQA e estações do ano	100
Figura 47- RMSEP (µg m ⁻³) <i>versus</i> estações do ano	100
Figura 48- Modelos mais bem ajustados - Gráfico previsto versus medido para o ozônio para PDC EAQA	102
Figura 49- Modelos mais bem ajustados - Gráficos previstos versus medidos para o ozônio para VSL EAQA.....	103
Figura 50- Modelos mais bem ajustados - Gráfico previsto versus medido para o ozônio para INE EAQA.....	104
Figura 51- Classificação Boruta para a ADN EAQA	108

Figura 52- Classificação Boruta para a VSL EAQA	108
Figura 53- Classificação Boruta para a PDC EAQA	108
Figura 54- Classificação Boruta para a INE EAQA	109

LISTA DE TABELAS

Tabela 1- Autores que aplicaram calibração multivariada em estudos de previsão de variáveis de qualidade do ar	23
Tabela 2- Esquema do cálculo da figura de mérito SWTP	49
Tabela 3 - Descrição da EAQA escolhida para o estudo	51
Tabela 4- Lista das ferramentas, comandos, pacotes e respectivas métricas, originários do R, utilizados neste trabalho	59
Tabela 5- Comparação de autovalores e variância explicada para cada modelo RBOPCA ...	67
Tabela 6- Comparação de autovalores e variância explicada para cada modelo RBOPCA a partir dos bancos de dados da EAQA PDC com NA e imputado.....	69
Tabela 7- Comparação de autovalores e variância explicada para cada modelo RBOPCA a partir dos bancos de dados da EAQA VSL com NA e imputado.....	72
Tabela 8- Comparação de autovalores e variância explicada para cada modelo RBOPCA ...	74
Tabela 9- Resumo das correlações entre as variáveis de entrada e concentração de ozônio para todas as EAQA, variáveis antes (preto) e após (azul-vermelho) imputação.	75
Tabela 10- Porcentagem de dados faltantes para cada EAQA analisada	77
Tabela 11- Número de observações, variáveis e <i>normalize root mean square error</i> (NRMSE) da imputação para cada EAQA escolhida para o estudo	78
Tabela 12- Resumo dos resultados obtidos pelos modelos PLS, RF, SVM e ANN para ADN EAQA	97
Tabela 13- Resumo dos resultados obtidos pelos modelos PLS, RF, SVM e ANN para PDC EAQA	101
Tabela 14- Resumo dos resultados obtidos pelos modelos PLS, RF, SVM e ANN para VSL EAQA	102
Tabela 15- Resumo dos resultados obtidos pelos modelos PLS, RF, SVM e ANN para o INE EAQA	103
Tabela 16- Resumo dos resultados das figuras de mérito com os valores imputados (input) e com valores faltantes (NA) para todas as EAQA	105
Tabela 17- Comparação dos modelos com 30 dias e 90 dias para a previsão de O ₃ para cada EAQA	110

Tabela 18- Comparação dos modelos já validados com os dados de Covid-19 de 2020 para a previsão de O ₃	112
---	-----

LISTA DE QUADROS

Quadro 1- Exemplo de conjunto de dados completo e conjuntos de dados com valores faltantes do tipo MCAR, MAR e NMAR	30
--	----

LISTA DE ABREVIATURAS E SIGLAS

ANN- Artificial Neural Network
ANP - Agência Nacional do Petróleo
ASTM - American Society for Testing and Materials
CART- Classificação e árvore de decisão
CETESB- Companhia de Tecnologia de Saneamento Ambiental
CFD- Computational Fluid Dynamics
CIT- California Institute of Technology
CMAQ- Community Multi-Scale Air Quality
CN- Cobertura de nuvem
CONAMA- Conselho Nacional do Meio Ambiente
COV- Compostos Orgânicos Voláteis
COVNM- Compostos Orgânicos Voláteis Não Metânicos
CO₂- Dióxido de carbono
COVNM- Composto Orgânico Volátil não metânico
CSA – TKCSA- Companhia Siderúrgica do Atlântico
CSN- Companhia Siderúrgica Nacional
CTDMPLUS- Complex Terrain Dispersion Model
DETRAN- Departamento de Trânsito
DV- Direção do Vento
EAQA- Estação Atmosférica de qualidade do ar
EPA- Environmental Protection Agency
GEE- Gases do Efeito Estufa
HCT- Hidrocarbonetos Totais
INEA- Instituto Estadual do Meio Ambiente
ISC- *Industrial Source Complex*
ISCST- *Industrial Source Complex Short Term*
KNN- Método dos vizinhos mais próximos
MLP-LM- Regressão Linear Múltipla – Levenberg–Marquardt
MLP- BP- Regressão Linear Múltipla – *Back Propagation*

MMA- Ministério do Meio Ambiente
MP- Material Particulado
HCNM- Non Methanic Hydrocarbons
OZIPR- Ozone Isopleth Package for Research
PCA- Análise de componentes principais
PI- Partículas inaláveis
P- Pressão atmosférica
PP- Precipitação Pluviométrica
ppmC- Partes por Milhão em Base de Carbono
PROCONVE- Programa de Controle da Poluição do Ar por Veículos Automotores
PTS- Partículas Totais em Suspensão
RMRJ- Região Metropolitana do Rio de Janeiro
ANN- Rede Neurais Artificiais
RS- Radiação Solar Global
RSU- Resíduos Sólidos Urbanos
SINDA- Sistema integrado de dados ambientais
SMAC- Secretaria de Meio Ambiente
SVM- *Support Vector Machine* (Máquinas de Vetores Suporte)
TEI- *Thermo Environmental Instruments*
T- Temperatura
UAM- *Urban Airshed Model*
UERJ- Universidade do Estado do Rio de Janeiro
U.S.EPA- United States Environmental Protection Agency
UR- Umidade Relativa
VV -Velocidade Escalar do Vento

SUMÁRIO

INTRODUÇÃO	18
Relevância Do Tema	22
Justificativa	23
Objetivo Geral	24
Objetivos Específicos	25
1. REVISÃO BIBLIOGRÁFICA.....	26
1.1 Dados Faltantes (<i>Missing Data</i>).....	29
1.1.1 k vizinhos mais próximos (KNN)	32
1.1.2 MissForest	33
1.2 Pré-Processamentos e/ou Transformações dos Dados	34
1.2.1 Análise de Componentes Principais (PCA)	34
1.2.2 Seleção de variáveis	35
1.2.3 Algoritmo de seleção de amostras.....	35
1.2.4. Normalização	36
1.3 Modelos de Regressão Multivariada	37
1.3.1 Mínimos Quadrados Parciais (PLS).....	37
1.3.2 Máquina de Vetores Suporte (SVM).....	39
1.3.3 Florestas Aleatórias (RF)	41
1.3.4 Rede Neural Artificial (ANN).....	42
1.4 Figuras de Mérito	46
2. METODOLOGIA	50
2.1 Descrição da Área Geográfica do Estudo	54
2.2 Linguagens Computacionais e Pacotes	58
3. RESULTADOS E DISCUSSÃO	61
3.1 Estatística Descritiva	61
3.1.1 Boxplot.....	61
3.1.2 PCA Robusto (ROBPCA)	64
3.1.3 Estudo de correlação de Pearson.....	75
3.1.4 Dados faltantes (data missing)	76
3.2 Imputação dos Dados (<i>Missforest</i>).....	77

3.3 Comparação dos dados imputados e medidos.....	78
3.4 Análise Exploratória dos Dados.....	86
3.4.1 Rosa dos ventos.....	90
3.4.2 Gráfico de Calendário	95
3.5. Previsão da Concentração de Ozônio Troposférico	95
3.6. Determinação da importância de cada variável na formação do Ozônio com auxílio da técnica de classificação Boruta.....	106
3.7. Exemplo Prático das Ferramentas Desenvolvidas	109
3.8. Previsão de O ₃ com os meses do Covid-19.....	111
4. CONCLUSÕES.....	113
REFERÊNCIAS	114
Apêndice A: Tabela A- Lista das EAQA estudadas na Tese	123
Apêndice B: Tabela B- Gráfico de Calendário para todas as EAQA.....	127

INTRODUÇÃO

A atmosfera de uma cidade é o resultado de diversos fatores, tais como as emissões das fontes móveis (veículos) e fixas (fábricas, residências, aterros, fontes naturais, entre outras), da meteorologia, da topografia, das transformações físicas e químicas que os poluentes emitidos pelas diversas fontes (poluentes primários) sofrem e são convertidos em poluentes secundários.

Segundo dados do Instituto Estadual do Ambiente (INEA) e da Companhia Ambiental de São Paulo (CETESB), mais de $\frac{3}{4}$ das emissões de poluentes das regiões metropolitanas de Rio de Janeiro e São Paulo, respectivamente, são oriundas de fontes móveis, pela queima dos variados combustíveis usados pela frota veicular, tais como gasolina, etanol, diesel, biodiesel e gás natural (INEA, 2015).

De acordo com o INEA (2015), as taxas de emissão por tipo de fonte na RMRJ, para os poluentes MP_{10} , SO_2 , NO_x , CO e HC apontam emissões totais (fixas e móveis) na ordem de 18,4; 63,3; 90,5; 321; 79,3 ($\times 1000 \text{ t}\cdot\text{ano}^{-1}$), respectivamente. As informações obtidas por meio do inventário apontam que, no universo das fontes consideradas, as fontes móveis são responsáveis por 77 % do total de poluentes emitidos para a atmosfera e as fontes fixas 23 %. A contribuição das fontes móveis para CO, NO_x e HC, são responsáveis por 98 %, 66,5 % e 67,3 % respectivamente.

Considerando a Região Metropolitana de São Paulo (RMSP), os veículos leves e pesados emitiram 97 % de monóxido de carbono (CO), 70 % de óxidos de nitrogênio (NO_x), 80 % de hidrocarbonetos (HC), 40 % de material particulado (MP) e 16 % de SO_x , segundo relatório da Agência Ambiental do Estado de São Paulo (CETESB, 2015), que correspondem a 129,17 mil toneladas de CO, 37,57 mil toneladas de HC, 72,35 mil toneladas de NO_x , 4,81 mil toneladas de MP e 6,71 mil toneladas de SO_x .

Avaliar a qualidade do ar é condição básica para o estabelecimento de políticas públicas de controle e melhoria da mesma e da qualidade de vida da população. Conhecendo-se os dados de monitoramento do ar é possível determinar o grau de controle e os recursos necessários para mitigar os impactos da poluição do ar no meio ambiente e na saúde humana (INEA, 2015).

O ozônio troposférico (O_3) tem um impacto negativo no clima (FISHMAN, 1991; FIORE et al., 2002), na vegetação (EMBERSON et al., 2018; FUHRER; SKÄRBY; ASHMORE, 1997; ZHANG et al., 2020) e no meio ambiente. É um dos responsáveis pelas

mudanças climáticas pois é um dos gases de efeito estufa, na troposfera (WANG et al., 2003; ORLANDO et al., 2010) e, portanto, afeta a saúde pública (LIPPMANN, 1991; WANG et al., 2003). Diante disso, a previsão dos níveis de ozônio é essencial para fornecer informações antecipadas ao público sobre a qualidade do ar. Pessoas sensíveis poderiam ser avisadas com antecedência sobre episódios de alta concentração de ozônio e, assim, reduzir a exposição à poluentes e danos à saúde. Além disso, a previsão de ozônio fornece informações sobre como políticas de redução de precursores estão sendo implementadas.

O ozônio é formado por um conjunto não linear complexo de reações de NO_x e COV na presença de luz solar. Começa com a dissociação fotoquímica de NO₂ com a formação de ozônio na ausência de COV, como mostrado nas Reações 1 a 3 (ATKINSON, 2000), onde M é uma molécula inerte.



No entanto, quando COV estão disponíveis, as reações de fotooxidação levam à formação de radicais alquil ou alquil peroxi ($R\dot{\text{O}}_2$), radicais alcoxi ou alcoxi substituído (RO) e radicais HO₂, que reagem com NO para converter NO em NO₂. O NO₂ fotolisa para formar O(³P) e em seguida, resulta na formação líquida de ozônio, conforme descrito nas Reações 4 a 8 (GERALDINO et al., 2020b).



Os radicais ($R\dot{\text{O}}_2$ e HO₂) reagem com o NO (reações 6 e 8) e convertem o NO em NO₂ sem consumo de ozônio. O ozônio formado é consumido pela Reação 3, mas devido a um

aumento de NO com uma conversão em NO₂, a formação de ozônio será aumentada quando a taxa de fotólise de NO₂ atingir um valor máximo (ATKINSON, 2000; SEINFELD; PANDIS, 2016) e será dependente da relação COV / NO_x (ORLANDO et al., 2010; SERGUEL; MORALES; LEIVA, 2012).

De acordo com a literatura, existem diversos modelos atmosféricos tradicionais, os quais utilizam equações matemáticas para descrever a dispersão da atmosfera e os processos físicos e químicos dentro da pluma, com o objetivo principal de calcular a concentração dos poluentes em vários lugares (HOLMES; MORAWSKA, 2006). Os modelos de poluição do ar podem classificar-se atendendo às suas aplicações, ou aos seguintes critérios, tais como: pela estrutura básica do modelo (determinísticos e não determinísticos (estocástico), estacionário ou dependente do tempo); pelo marco de referência (Euleriano ou Lagrangiano); pela dimensionalidade do domínio computacional (unidimensional, bidimensional, tridimensional de múltiplos níveis) e pelo método de resolução das equações (numérico ou analítico) (LORA, 2002).

Dentre os modelos tradicionais mais utilizados estão o *Ozone Isopleth Package for Research (OZIPR)* (GERY; CROUSE, 1990). De acordo com Corrêa (2003), o OZIPR permite a simulação de processos químicos e físicos que ocorrem na troposfera urbana, através de um modelo de trajetórias. Entende-se este modelo como uma coluna de ar que, na base, engloba toda a região de estudo e que se projeta para cima até a camada de mistura da atmosfera, como se fosse uma caixa com uma tampa móvel, a qual é função da altura da camada de mistura ao longo do dia. Toda a caixa é considerada perfeitamente homogênea e move-se de acordo com a trajetória do vento de modo a não se expandir horizontalmente. O *Complex Terrain Dispersion Model (CTDMPLUS)* é o modelo aplicado para fontes fixas e em estado estacionário, sendo um modelo da qualidade do ar que é aplicado para fontes de poluentes localizadas dentro de uma área ou perto de complexos topográficos. O modelo descreve três dimensões da pluma gaussiana e requer detalhes topográficos e descrição meteorológica detalhada (U.S. EPA, 2013). O *Community Multi-scale Air Quality (CMAQ)* é um modelo euleriano tridimensional de código aberto utilizado pelo governo dos EUA para implementar ações preventivas no controle dos padrões da qualidade do ar, como gerenciar cenários para cidades, estados e países. Além disso, o CMAQ combina o conhecimento da atmosfera terrestre e o modelo da qualidade do ar com técnicas de computação com multiprocessos. Geralmente

para estimar concentrações de ozônio, particulados, toxicidade e deposição ácida (U.S. EPA, 2013). O *Urban Airshed Model (UAM)* é o mais complexo modelo tridimensional desenvolvido pela U.S.EPA e auxilia a regulamentação de medidas preventivas no controle das emissões de poluentes nos EUA. Este modelo é utilizado para estudo fotoquímico da qualidade do ar, principalmente na formação do ozônio (U.S. EPA, 2013). O *California Institute of Technology (CIT)* é um modelo fotoquímico, baseado na solução numérica de equação de difusão atmosférica, que descreve a formação e transporte de poluentes quimicamente reativos, na camada limite planetária, incluindo a formação de ozônio. As características do modelo CIT para São Paulo são descritas nos estudos de Ulke e Andrade (2001), Sanchez-Ccoyllo *et al.*, (2006), Sánchez-Ccoyllo, *et al.*, (2007).

Entretanto, estes modelos descritos são muito trabalhosos e dispendiosos já que dependem de muitas variáveis de entrada, tais como: o monitoramento por múltiplas estações de qualidade do ar, sondagens verticais, especiação de compostos orgânicos voláteis (COV), inventários de emissões, concentrações iniciais de poluentes, dados meteorológicos horários de temperatura, pressão, umidade, direção e velocidade do vento, altura da camada de mistura, emissões primárias de óxidos de nitrogênio (NO_x), monóxido de carbono (CO) e COV, topografia, localização geográfica e data (para se calcular o fluxo solar actínico), cobertura de nuvens, coeficientes de deposição seca e úmida, modelo químico e a necessidade de uma boa capacidade de processamento numérico.

Os modelos de aprendizagem de máquinas como Redes Neurais Artificiais (ANN), Máquinas de Vetor de Suporte (SVM), Florestas Aleatórias (RF), Regressão de Mínimos Quadrados (PLS), entre outros, vêm conquistando uma grande parte dos pesquisadores mundiais, pois necessitam de menos variáveis de entrada, menos esforço computacional e as respostas destes modelos são satisfatórias em comparação com os modelos tradicionais. Neste contexto, o presente trabalho tem como proposta apresentar estes modelos de previsão de ozônio troposférico para a Região Metropolitana do Rio de Janeiro compreendidos entre os anos de 2014 e 2018.

Relevância Do Tema

O ozônio na troposfera (camada mais próxima da superfície terrestre) é um poluente secundário, formado pelas reações fotoquímicas do NO_x e COV, que são poluentes provenientes de fontes fixas e móveis. Este poluente é prejudicial à saúde humana atacando as vias respiratórias, tecidos pulmonares e os vegetais pela alteração natural na fotossíntese. O_3 é bem reativo e desempenha um papel na capacidade de oxidação da atmosfera. Ele e seu derivado fotoquímico o OH são os oxidantes mais importantes na redução da maioria dos gases (BRASSEUR *et al.*, 1999).

Atualmente o ozônio é o principal problema de poluição da Região Metropolitana do Rio de Janeiro (RMRJ) chegando a ultrapassar em alguns dias em 192 % o padrão de qualidade do ar no ano de 2017 (MARTINS *et al.*, 2017). Para a diminuição das concentrações de ozônio na atmosfera da cidade do Rio de Janeiro é necessário reduzir e controlar os precursores de ozônio que são os compostos orgânicos voláteis (COV) e os NO_x formadores desse poluente por processos fotoquímicos (CETESB, 2015; DA SILVA *et al.*, 2017, 2018; GIODA *et al.*, 2017; GERALDINO *et al.*, 2020a).

Estudos de Martins *et al.* (2017) na RMRJ nos períodos entre janeiro de 2012 e dezembro de 2013, os bairros que apresentaram o maior número de ultrapassagens de concentrações de O_3 foram Bangu, Irajá e Campo Grande com, respectivamente, 215, 189 e 77 ultrapassagens. Já os bairros de São Cristóvão, Centro e Copacabana apresentaram 19, 4 e 1 violações respectivamente.

Ozônio foi o único poluente que apresentou ultrapassagens em todas as estações de qualidade do ar, quando a comparação foi feita com o nova Resolução CONAMA nº491/2018 (CONAMA, 2018) que é de $140 \mu\text{g m}^{-3}$, máxima média móvel diária. Ao fazer a comparação com estes padrões é possível observar violações nas estações de Irajá, Bangu e Campo Grande com respectivamente 17, 12 e 7 médias horárias (MARTINS *et al.*, 2017). Porém, cabe ressaltar que na época destes monitoramentos a Resolução 03/1990 era a que estava em vigor, com um limite para ozônio de $160 \mu\text{g m}^{-3}$.

Justificativa

Trabalhos mais recentes, devido a evolução computacional e bastante citados na literatura acadêmica, aplicaram os modelos de calibração multivariada (Tabela 1) para avaliar a tendência/variabilidade, ou previsão, ou imputação dos dados, ou para entender a formação O_3 . Contudo, estudos com previsão e imputação de dados faltantes com a inclusão de 4 EAQA em diferentes bacias aéreas na RMRJ ainda não foram realizados e, portanto, existem lacunas na literatura para uma melhor compreensão da concentração de O_3 para as cidades brasileiras que possuem suas peculiaridades com o uso de biocombustíveis (etanol e biodiesel), GNV e veículos *flex fuel*.

Assim, o desenvolvimento de modelos de aprendizado de máquinas que contemple as quatro bacias aéreas na Região Metropolitana do Rio de Janeiro (91 estações meteorológicas, durante 5 anos), a qual possui diferentes relevos, meteorologias, microclimas, utilizando o banco de dados recentes, com inclusão de estudos de dados faltantes (*data missing*) pode contribuir para a melhor compreensão do perfil de ozônio troposférico.

Para se avaliar a qualidade do ar de uma cidade é preciso um investimento elevado em estações automáticas de qualidade do ar (EAQA) que meçam os poluentes listados na Resolução CONAMA 491/2018 (CONAMA, 2018), assim como os dados meteorológicos. Outro fator que encarece os custos é medir a especiação dos COV para conhecer a reatividade da mistura que tem grande influência na formação do ozônio. Na tese de doutorado de Silva (2016) foi estimado um conjunto mínimo de COV para se conhecer a reatividade da mistura com base em três escalas de reatividade (*Maximum Incremental Reactivity* - MIR, *Equal Benefit Incremental Reactivity* - EBIR e *Maximum Ozone Incremental Reactivity* - MOIR), desta forma reduzindo os custos das dispendiosas etapas das preconizadas na metodologia da U.S.EPA TO-15A (U.S.EPA, 2019).

Tabela 1- Autores que aplicaram calibração multivariada em estudos de previsão de variáveis de qualidade do ar

Autor	Variáveis estudadas	Figuras de mérito	Período estudado	Técnicas de calibração multivariada	Objetivo	Localização
Schuch <i>et al.</i> , 2019	CO, NO, NO ₂ , UR, RS, T, O ₃	R ²	20 anos	LM, QRM	Tendência e variabilidade do O ₃ na RMSP	São Paulo

Autor	Variáveis estudadas	Figuras de mérito	Período estudado	Técnicas de calibração multivariada	Objetivo	Localização
Garcia et al., 2018	SO ₂ , NO, NO ₂ , MP ₁₀ , O ₃ , C ₆ H ₆ , CO, C ₇ H ₈ , xilenos	RMSE	8 anos	SVM, ANN, VARMA, ARIMA	Previsão de MP ₁₀ em Oviedo (Espanha)	Espanha
Arroyo et al., 2018	O ₃ , NO, NO ₂ , CO, MP ₁₀ , SO ₂ , NO _x e variáveis meteorológicas	MSE	9 anos	MissForest, MLR, ANN	Imputação de dados de ozônio faltantes	Espanha
Gioda et al., 2017	NO _x , NO, NO ₂ , O ₃ , CO, HCT, BTEX e variáveis meteorológicas	-	2 anos	PCA	Entender a formação do O ₃ em duas Ilhas no RJ	Rio de Janeiro
Luna et al., 2014	NO _x , VV, T, CO, RS, UR, NO, NO ₂	R _{2cal} , R _{2ev} , RMSEC, RMSECV, RMSEP, R _{2pred}	5 meses	PCA, SVM, ANN	Previsão de O ₃ na RMRJ	Rio de Janeiro

Fonte: O autor, 2022.

O número de EAQA podem ser elevados, dependendo da extensão territorial da cidade, do seu relevo, das características de suas fontes fixas e móveis e sua distribuição espacial. O custo de uma EAQA é estimado em cerca de R\$ 500 mil mais aproximadamente 10 % deste valor anual para sua operação. Muitas cidades não dispõem de verbas para a implantação de diversas EAQA, assim como para manter sua operação, tratamento de dados e operações subsequentes. Diante disso esta tese foi pensada para ser uma ferramenta alternativa para as cidades poderem avaliar a qualidade do ar. Ao invés de adquirir e operar diversas EAQA, foi vislumbrada a possibilidade de usar uma EAQA móvel para coleta de dados em diferentes locais da cidade e a partir destes dados usar as ferramentas aqui propostas para fazer previsões futuras, embasar a aquisição de futuras EAQA, localização, restrições a serem impostas nas fontes emissoras, entre tantas possibilidades.

Objetivo Geral

O objetivo geral é analisar os dados das estações automáticas da qualidade do ar (EAQA), providas pelo Instituto Nacional do Meio Ambiente (INEA) localizadas na Região

Metropolitana do Rio de Janeiro (RMRJ), para os anos de 2014 a 2018 e investigar o comportamento das variáveis químicas e meteorológicas de forma descritiva e multivariada, propondo o preenchimento de dados faltantes para as previsões de concentrações de O_3 na troposfera, utilizando técnicas de aprendizado de máquinas.

Objetivos Específicos

Como desdobramentos do objetivo principal, podem ser considerados como objetivos específicos:

- Avaliar o comportamento dos dados meteorológicos e de poluentes das EAQA da RMRJ através de um pré-tratamento (estatística descritiva);
- Escolher uma EAQA de cada Bacia Aérea da RMRJ com perfis distintos para avaliar o potencial da técnica;
- Separar os dados de treinamento, validação e verificação;
- Desenvolver um modelo de preenchimento de dados faltantes utilizando técnicas de aprendizado de máquina;
- Avaliar o modelo de preenchimento de dados faltantes prevendo cenários da poluição de O_3 .

1. REVISÃO BIBLIOGRÁFICA

A literatura contém uma grande quantidade de trabalhos de modelagem sobre a previsão de ozônio (O_3). A maioria dos estudos utilizaram técnicas de inteligência artificial e estatística, como regressão multilinear (MLR), redes neurais artificiais (ANN), classificação e árvores de regressão (CART) e máquinas de vetores suporte (SVM).

Para o levantamento dos artigos na literatura acadêmica, foram utilizadas as seguintes bases de pesquisas para buscas de periódicos na área de estudo: *SCOPUS*, *Science Direct* e Periódicos Capes, com algumas das palavras - chave: *Machine learning*, *Ozone Forecast*, *Urban*. Foram encontrados mais de 100 artigos na literatura. Os artigos foram selecionados de acordo com sua relevância no meio acadêmico, através de número de citações e por ano de publicação, este último, se espera que os artigos mais recentes possuam melhores técnicas numéricas e de caracterização, além de acompanhar a evolução computacional.

A literatura dispõe de estudos, como os de Schuch *et al.* (2019), Alimissis *et al.* (2018), Sekar *et al.* (2016), Azid *et al.* (2014), Luna *et al.* (2014) e Tamas *et al.* (2014) que utilizaram ferramentas quimiométricas em seus trabalhos para prever o ozônio troposférico nas cidades.

Recentemente, um grupo de pesquisadores (SCHUCH *et al.* 2019) analisaram as tendências dos níveis de ozônio na troposfera de São Paulo de 1996 a 2017. Os autores estudaram a variabilidade e a tendência do ozônio usando variáveis independentes, como CO, NO, NO₂, velocidade do vento, radiação solar e temperatura, combinados com modelo de regressão interquartil (QRM). O modelo resultante apresentou coeficiente de determinação R^2 igual a 0,76 e permitiu concluir que os efeitos da radiação e da temperatura são os mais críticos na determinação dos maiores quartis de ozônio.

Alimissis *et al.* (2018) estudaram duas metodologias de interpolação uma por Redes Neurais Artificiais (ANN) e a outra por Regressão Linear Múltipla (MLR), utilizando dados de 13 redes de monitoramento da qualidade do ar localizada na área metropolitana de Atenas, na Grécia durante os anos de 2001 até 2013. Os resultados para os cinco poluentes atmosféricos (NO₂, NO, O₃, CO e SO₂) são comparados através do uso de um conjunto de medidas estatísticas de correlação e distribuição de resíduos. A estação de Peristeri apresentou um menor *root mean square error* (RMSE) de 11,39 $\mu g\ m^{-3}$ e um coeficiente de determinação (R^2) de 0,89 para a previsão do ozônio, utilizando um de ANN. As redes neurais artificiais se apresentaram,

na maioria dos casos, significativamente superior ao MLR, devido à sua capacidade de modelar de forma mais eficiente a variabilidade espacial complexa da poluição do ar.

Gioda *et al.* (2017) avaliaram a influência das emissões veiculares em duas ilhas localizadas na RMRJ, uma delas com frota considerável (Ilha do Governador) e outra sem a presença de veículos (Ilha de Paquetá). Os autores usaram médias horárias dos anos de 2012 e 2013 para o O₃, NO_x, CO, hidrocarbonetos aromáticos (BTEX), hidrocarbonetos totais (HCT) e dados meteorológicos. A análise de componentes principais (PCA) e o algoritmo Boruta foram utilizados para avaliar o comportamento dos dados e caracterizar o impacto da frota de veículos na qualidade do ar. Os resultados mostraram que os níveis de CO e NO_x foram 2 a 6 vezes maiores na Ilha do Governador do que na Ilha do Paquetá. Como esperado, os níveis de O₃ foram até 1,5 vezes maiores na Ilha de Paquetá do que na Ilha de Governador, o que pode ser explicado por estar relacionado ao processo de formação de HCT e NO_x na presença de luz solar.

Sekar *et al.* (2016) desenvolveram modelos de previsão horária de O₃ e NO_x com base em algoritmos de Árvore de Decisão, utilizando o *software* Matlab®, utilizaram árvore de regressão (REPTree), árvore M5 P e um *perceptron* multicamada usando Levenberg-Marquardt (MLP-LM) em Deli, Índia. Os dados dos poluentes foram provenientes de Deli e a estação de Safdarjung proveu os dados meteorológicos correspondentes aos anos de 2008-2010, a qual foram escolhidos para este estudo. Umidade relativa (UR), pressão atmosférica (P), temperatura (T), velocidade do vento (VV), direção do vento (DV), cobertura da nuvem (CN), radiação solar (RS), precipitação (PP), classe de estabilidade atmosférica, altura da camada de mistura e variáveis temporais, como o dia da semana e a hora do dia, foram utilizados como variáveis de entrada do melhor modelo encontrado. Este modelo de árvore M5 P apresentou melhor desempenho com um RMSE de 18,31 µg m⁻³ e R² de 0,82.

Azid *et al.* (2014), utilizando o *software* XLSTAT 2014, identificaram as fontes potenciais da variação da qualidade do ar e implementaram um modelo para previsão do índice de poluição do ar na Malásia. Entre as ferramentas estatísticas utilizadas neste trabalho foram o PCA (*Principal Component Analysis*), rede neural artificial (ANN- *Artificial Neural Network*), KNN (*Nearest Neighbor Method* – para dados faltantes) e concluíram que as variáveis CH₄, HCNM (hidrocarbonetos não metânicos), HCT (hidrocarbonetos totais), O₃ e

MP₁₀ foram classificadas como os maiores contribuintes na poluição atmosférica local na Malásia.

Luna *et al.*, (2014), utilizaram o *software* Matlab[®] R2008b version 7.7.0, com a finalidade de desenvolver uma análise a partir de modelos de aprendizagens de máquina para a previsão para o ozônio troposférico, a qual demonstrou ser uma ferramenta importante com o objetivo de prover políticas na prevenção de eventos onde haja alta concentração de poluentes. Demonstraram a potencialidade de se utilizar rede neurais artificiais (ANN) e as máquinas de vetor suporte (SVM) como métricas de ferramentas quimiométricas aplicando estatísticas na previsão de O₃ na cidade do Rio de Janeiro, Brasil. Uma estação de monitoramento móvel foi usada para coletar dados por hora em duas localidades, nomeadamente a área da Pontifícia Universidade Católica de julho a outubro de 2011 e outra na área da Universidade do Estado do Rio de Janeiro, de novembro de 2012 a março de 2013. NO₂, NO, CO, O₃, T, VV, RS e UR no ar foram utilizados como variáveis de entrada. O uso da análise de componentes principais (PCA) na redução de dimensão foi explorado. Os autores obtiveram os melhores resultados com os modelos de predição utilizando ANN com o *root mean square error of prediction* (RMSEP) igual a 16,51 µg m⁻³ e 8,10 µg m⁻³ e o R² de 0,653 e 0,886 para PUC e UERJ, respectivamente.

Tamas *et al.* (2014) utilizaram uma rede neural do tipo *Multilayer Perceptron* com o algoritmo Levenberg-Marquardt para treinar a rede (MLP-LM), com a finalidade de prever as próximas 24 horas de O₃ usando dados de 2008 a 2014 de estações urbanas e suburbanas (Canetto, Sposata) em Ajaccio, e (Giraud, Montesoro) em Bastia da ilha francesa da Córsega na França. O₃, NO₂, DV, RS, T, PP e variáveis temporais como hora do dia e os dias da semana foram utilizadas como variáveis de entrada. O menor RMSE encontrado foi de 15,90 µg m⁻³ na estação de Montesoro.

Com base no levantamento de informações e descrição do panorama da literatura, foi observado que em vários estudos as técnicas quimiométricas mais utilizadas nos trabalhos disponíveis na literatura envolvendo previsão de ozônio têm sido: ANN, SVM, RF e árvores de decisão. Além disso, independente da natureza do problema de previsão, esta área de estudo representa incontáveis possibilidades de combinações de muitas ferramentas estatísticas, trazendo benefícios à modelagem. No entanto, vale ressaltar a importância de se entender os dados iniciais e a integridade dos mesmos, visto que é comum encontrar dados faltantes em

uma base de dados, devido à vários fatores. Além disso é importante esgotar as técnicas convencionais mais simples (PLS, por exemplo) antes de se iniciar estudos com as ferramentas mais robustas, como técnicas não-lineares (RF, SVM e ANN).

1.1 Dados Faltantes (*Missing Data*)

É muito comum que os dados das estações meteorológicas e da qualidade do ar tenham dados faltantes em maior ou menor grau, o que pode dificultar a realização da análise dos dados e a construção de modelos a partir deles. Entre os principais problemas geralmente associados a valores faltantes, Farhangfar *et al.* (2007) relataram a perda de eficiência, as dificuldades no tratamento e análise dos dados e o viés resultante das diferenças entre os dados faltantes e completos. O estudo com dados faltantes segue duas abordagens, que podem ser ignoradas (removidas) ou imputadas (preenchidas) com novos valores. De acordo com Farhangfar *et al.* (2007), a primeira abordagem é aplicável apenas quando uma pequena quantidade de dados está faltando. Como em muitos casos os bancos de dados contêm uma quantidade relativamente grande de dados ausentes, é mais construtivo e praticamente viável considerar a imputação. Para Junninen *et al.* (2004), a situação pode ser resultado de amostragem insuficiente, erros nas medições ou falhas na aquisição e calibração dos dados. Quaisquer que sejam as razões, as descontinuidades representam um obstáculo significativo para a aplicação da técnica de previsão, que geralmente requer dados contínuos como condição para seu uso. Por este motivo, a imputação de dados é uma ferramenta importante para o preenchimento dos dados faltantes em um banco de dados.

Por muitas razões, os dados meteorológicos e os dados de poluentes nem sempre são completos ou válidos. Isso pode ser devido ao mau funcionamento do equipamento que opera em vários locais de monitoramento ou pode ser devido a quedas de energia. Este problema de dados faltantes pode ser tratado usando uma interpolação linear, por exemplo. Se os dados faltantes estiverem perdidos durante mais tempo (aproximadamente mais de um mês), os dados da estação próxima podem ser utilizados (caso a estação próxima seja inferior a 8 km); quando as estações próximas estiverem a uma distância superior a 8 km, então os valores médios das estações circundantes podem ser utilizados para preencher os dados faltantes (STEKHOVEN; BUEHLMANN, 2012).

É necessário entender por que os dados estão faltando. De acordo com Little e Rubin (2002) existem três mecanismos principais de dados faltantes (Quadro 1): *Missing Completely at Random* (MCAR), *Missing at Random* (MAR) e *Not Missing at Random* (NMAR).

O MCAR acontece quando o valor faltante não depende dos dados observados e nem dos não observados; constituindo-se como um evento randômico. Esse tipo de dado faltante pode ocorrer, por exemplo, se alguém resolve lançar uma moeda para o alto para decidir se uma disputa de cabo de guerra deve ou não deve ser iniciada (MARTINS, 2017).

Quadro 1- Exemplo de conjunto de dados completo e conjuntos de dados com valores faltantes do tipo MCAR, MAR e NMAR

Conjunto de dados completo						
x1	x2	x3	x4	x5	x6	classe
27	3258	123	111	210	188	a
29	4561	228	168	242	250	a
24	4587	13	275	97	133	a
36	2964	218	56	186	97	a
40	2809	16	143	48	78	a
25	2081	237	73	225	14	a
45	1947	38	140	184	76	a
49	810	50	260	118	274	a
42	4805	16	144	273	144	a
26	2819	298	12	107	292	a
42	2788	43	278	111	164	b
29	1519	269	19	192	28	b
29	1562	82	208	198	129	b
31	4315	213	35	297	133	b
29	4118	158	118	3	195	b
24	1297	190	300	79	19	b
49	4305	176	259	182	212	b
21	4318	100	177	7	176	b
28	4062	111	165	29	106	b
35	4471	61	84	102	291	b

Missing Completely at Random (MCAR)						
x1	x2	x3	x4	x5	x6	classe
27	3258	123	NA	210	188	a
29	4561	228	168	242	250	a
24	4587	13	275	97	133	a
36	2964	NA	56	NA	97	a
40	2809	16	143	48	78	a
25	NA	237	73	225	14	a
45	1947	NA	140	184	76	a
49	810	50	260	118	274	a
42	4805	16	144	273	144	a
26	2819	298	12	107	NA	a
42	2788	43	NA	111	164	b
29	NA	269	19	192	28	b
29	1562	82	208	198	129	b
31	4315	213	35	297	NA	b
29	4118	158	118	3	195	b
24	1297	190	300	79	19	b
49	4305	176	259	182	212	b
NA	4318	100	177	7	176	b
28	4062	111	165	29	NA	b
35	4471	61	84	NA	291	b

Missing at Random (MAR)						
x1	x2	x3	x4	x5	x6	classe
27	3258	123	111	210	188	a
29	NA	228	168	242	250	a

Not Missing at Random (NMAR)						
x1	x2	x3	x4	x5	x6	classe
27	3258	123	111	210	188	a
29	4561	228	168	242	250	a

<i>Missing at Random (MAR)</i>							<i>Not Missing at Random (NMAR)</i>						
24	NA	13	275	NA	133	a	24	4587	13	275	97	133	a
36	2964	218	56	186	97	a	36	2964	218	56	186	97	a
NA	2809	16	143	48	78	a	40	2809	16	143	48	78	a
25	2081	237	NA	225	14	a	25	2081	237	73	225	14	a
45	1947	38	140	184	76	a	45	1947	38	140	184	76	a
49	810	50	NA	NA	274	a	49	810	50	260	118	274	a
42	4805	16	144	273	144	a	42	4805	16	144	273	144	a

<i>Missing at Random (MAR)</i>							<i>Not Missing at Random (NMAR)</i>						
x1	x2	x3	x4	x5	x6	classe	x1	x2	x3	x4	x5	x6	classe
26	2819	298	12	107	NA	a	26	2819	298	12	107	292	a
42	2788	43	278	111	164	b	42	2788	43	278	111	164	b
NA	1519	269	19	192	28	b	29	1519	269	19	192	28	b
29	1562	82	208	198	129	b	29	1562	82	208	198	129	b
31	4315	NA	35	297	133	b	31	4315	213	35	297	133	b
NA	NA	158	118	3	195	b	29	4118	158	118	3	195	b
24	1297	190	300	79	19	b	24	1297	190	300	79	19	b
49	4305	176	259	182	212	b	49	4305	176	259	182	212	b
21	4318	100	177	7	176	b	21	4318	100	177	7	176	b
28	4062	111	165	29	106	b	28	4062	111	165	29	106	b
35	4471	61	84	102	291	b	35	4471	61	84	102	291	b

Fonte: adaptado de Misztal, 2013

O MAR sucede quando o dado faltante depende dos valores observados, em outras palavras, uma variável que contém os dados faltantes depende de uma variável com dados observados. Portanto, a falta se refere a uma variável particular. Por exemplo, suponha uma pesquisa na qual as mulheres são menos suscetíveis em relação a fornecer sua renda mensal pessoal. Se de antemão o sexo de todos os sujeitos e a renda para algumas mulheres forem conhecidos, então, a variável renda mensal, será do tipo MAR, pois depende da variável sexo (MARTINS, 2017).

O MNAR surge em ocasiões em que existe uma razão específica para o dado faltante, em outras palavras, está relacionado aos valores não observados. É muito comum quando as pessoas não querem revelar algo muito pessoal. Por exemplo, pessoas com depressão talvez rejeitem preencher uma pesquisa sobre depressão (MARTINS, 2017).

A eficácia da imputação dos dados depende não apenas da quantidade de dados faltantes, mas também das características dos padrões de dados faltantes. Além disso, o mecanismo de dados faltantes em dados de qualidade do ar é geralmente aleatório *missing at random* (MAR), no sentido de que a probabilidade de um valor estar ausente não depende do valor faltante (RUBIN, 1976).

Estudos de Stekhoven e Bühlmann (2012) compararam o método de imputação do *MissForest* à técnica de imputação por *k-nearest neighbors algorithm* (KNN) (TROYANSKAYA *et al.* 2001), *MissPALasso* (um método baseado no algoritmo *E- and M-Step* “EM”, proposto por Städler e Bühlmann (2010) e *Multiple Imputation by Chained Equations* (MICE) Van Buuren e Oudshoorn (1999). Eles mostraram que *MissForest* poderia superar outros métodos de imputação. Vale observar, entretanto, que em experimentos de simulação apenas o *Missing Completely at Random* (MCAR) foram analisados. É razoável, portanto, realizar experimentos adicionais para avaliar a utilidade do método de imputação *MissForest*.

1.1.1 *k* vizinhos mais próximos (KNN)

O método dos *k* vizinhos mais próximos (*k nearest neighbors*- KNN) é um dos algoritmos de mineração de dados mais simples e pertence à classe dos chamados aprendizados preguiçosos (COVER e HART 1967). O KNN realmente não obtém um modelo de dados de treinamento, mas simplesmente armazena os conjuntos de dados. Seu principal trabalho acontece no tempo de previsão. Dado um novo caso de teste, sua previsão é obtida pela busca de casos semelhantes nos dados de treinamento que foram armazenados. Os casos de treinamento mais semelhantes (ou seja, vizinhos) são usados para obter a previsão para o caso de teste dado. Quando se fala sobre vizinhos, sugere-se que há uma distância ou medida de proximidade que pode-se calcular entre amostras com base nas variáveis independentes.

Quando o número total de dados faltantes no banco de dados for muito pequeno, isto é, na ordem de 3 a 5 %, faixa comum de ser encontrada nas Estações Automáticas da Qualidade do Ar (EAQA), a implementação do KNN também pode ser aplicada com a finalidade de completar os dados faltantes e assim facilitar a análise de dados. Este método examina a distância entre cada ponto e o ponto mais próximo. O método do KNN é o esquema mais simples, onde os pontos finais das lacunas são usadas como estimativas para todos os valores

faltantes (JUNNINEN *et al.* 2004; DOMINICK *et al.* 2012). Uma relação que pode ser aplicada neste método é mostrada na Equação 1.

$$y = y_1 \text{ se } x \leq x_1 + [(x_2 - x_1)/2]$$

$$y = y_2 \text{ se } x > x_1 + [(x_2 - x_1)/2] \quad (\text{Equação 1})$$

Onde y é o interpolante, x é a variável obtida referente ao interpolante, y_1 e x_1 são as coordenadas do ponto de partida do intervalo e y_2 e x_2 são os pontos finais do intervalo.

1.1.2 MissForest

O algoritmo *MissForest* é baseado na técnica de floresta aleatória (RF) e permite a imputação de valores faltantes em basicamente qualquer tipo de dados. Ele pode manipular dados multivariados combinando variáveis contínuas e categóricas ao mesmo tempo, também não precisa de parâmetros de ajuste nem exige suposições sobre aspectos de distribuição dos dados. Stekhoven *et al.* (2012) mostraram em vários conjuntos de dados reais provenientes de dados biológicos e médicos distintos que a técnica do *MissForest* supera os métodos de imputação já estabelecidos na literatura, como a imputação de k vizinhos mais próximos (KNN) ou imputação multivariada. Entre as características mais importantes do *MissForest* está a vantagem de não realizar custosas validações cruzadas, pode ser implementado em dados com interações complexas e com relações não lineares, e pode ser aplicado em um conjunto de dados de alta dimensão onde o número de variáveis pode ser muito maior que o número de observações e ainda apresentar bons resultados na imputação (MISZTAL 2013; STEKHOVEN, DANIEL; BÜHLMANN, 2012).

Apesar de numerosos estudos na área de química da atmosfera, poucos autores relatam a aplicação de ferramentas de preenchimento de dados faltantes. Por exemplo, Arroyo *et al.* (2018) estudaram modelos de redes neurais artificiais para imputação de dados faltantes de ozônio na qualidade do ar em seis lugares diferentes na Espanha em 2018. As variáveis independentes usadas neste estudo foram CO, NO, NO₂, MP₁₀, SO₂, VV e estações do ano (SEA), com aproximadamente 10 % de dados faltantes. Técnicas de regressão linear múltipla (MLR), regressão não linear múltipla (MN-LR) e rede neural artificial (ANN) foram aplicadas. Os autores também estudaram separadamente cada estação (primavera, verão, outono e

inverno) e alcançaram os melhores resultados usando ANN para todos os cenários, com erro quadrático médio (MSE) variando entre $6,0 \cdot 10^{-6}$ até $8,0 \cdot 10^{-6}$.

1.2 Pré-Processamentos e/ou Transformações dos Dados

Segundo Ferreira (2015), o pré-processamento ou transformação dos dados deve ser realizado para diminuir variações indesejadas adquiridas durante a obtenção dos dados. Desta forma, o pré-tratamento se faz necessário para provocar melhorias na interpretação e simplificar o modelo, fazendo-o muitas vezes mais robusto e confiável contra essas variações indesejáveis.

1.2.1 Análise de Componentes Principais (PCA)

Após a imputação dos dados por *MissForest*, é aconselhável que os conjuntos de dados passem por uma etapa de análise exploratória, a fim de determinar o comportamento dos dados e extrair informações relevantes deles (PACK, 1991). A análise de componentes principais (PCA) foi o método estatístico selecionado para a análise exploratória.

PCA é um método não supervisionado aplicado a um conjunto de dados multivariado, com a finalidade de retratá-los em uma dimensão menor do que o conjunto de dados original, sem alterar a relação entre as amostras. A representação dessa redução na dimensionalidade é dada pelos componentes principais (PC), que podem ser definidos como uma combinação das variáveis que expressam a tendência dos dados. A ortogonalidade é o principal aspecto desse novo conjunto de dados, porém pode ser reconstruída a partir da combinação linear das variáveis originais (ABDI; WILLIAMS, 2010; CORDELLA, 2012; FERRER-RIQUELME, 2010). No PCA, uma matriz contendo variáveis independentes (X) é dividida em um produto de duas matrizes, *loads* (P) e *scores* (T), mais uma matriz de erro (E), conforme mostrado na Equação 2.

$$X = T \cdot P^t + E \quad (\text{Equação 2})$$

onde X é a matriz ($m \times n$), T é a matriz dos *scores* ($m \times \text{PC}$), P^t é a matriz de *loadings* transposta ($\text{PC} \times n$) e E é a matriz de erros ($m \times n$).

A escolha do número de PC utilizado no modelo de PCA pode ser definida pelo percentual da variação explicada, de forma que seja capturada a maior proporção da variação presente no conjunto de dados original (ABDI; WILLIAMS, 2010; CORDELLA, 2012).

O PCA é uma ferramenta importante nos estudos estatísticos, mas sabe-se que o PCA clássico é sensível a presença de *outliers*. De acordo com estudos de Ma e Aybat (2018), o PCA robusto (ROBPCA) pode ser utilizado para remover o efeito de erros brutos esparsos. Assim, para uma matriz de dados $\mathbf{X}$ ($m \times n$), o ROBPCA decompõe essa matriz em duas matrizes, conforme mostrado pela Equação 3.

$$\mathbf{X} = \mathbf{L} + \mathbf{S} \quad (\text{Equação 3})$$

onde $\mathbf{L}$ é uma matriz de baixo nível e $\mathbf{S}$ é uma matriz esparsa. O PCA robusto preza que $\mathbf{X}$ é uma superposição de $\mathbf{L}$ e $\mathbf{S}$. Logo, os erros mais grosseiros serão retidos pela matriz esparsa $\mathbf{S}$, possibilitando que a matriz de baixo nível $\mathbf{L}$ ainda possa se aproximar de $\mathbf{X}$. Em resumo, o ROBPCA fornece uma aproximação de baixa dimensão que é robusta para valores extremos.

Nesta tese o PCA e ROBPCA foram utilizados principalmente como um método exploratório para auxiliar na investigação do comportamento das observações coletadas, sendo capaz de separar as informações relevantes das redundantes e aleatórias, o que colabora na identificação de amostras atípicas.

1.2.2 Seleção de variáveis

O uso de técnicas de seleção de variáveis permite a construção de um modelo robusto e de fácil interpretação, uma vez que a escolha de regiões específicas pode minimizar os erros do modelo multivariado.

Os métodos de seleção de variáveis podem ser baseados em informações especializadas, no caso informações de características químicas, ou em algoritmos. As variáveis utilizadas nesta tese foram selecionadas, levando-se em consideração: (i) os coeficientes de correlação, (ii) a produção fotoquímica e (iii) o transporte de O_3 , as quais serão consideradas para a construção dos modelos.

1.2.3 Algoritmo de seleção de amostras

O algoritmo de Kennard-Stone (KS) é um dos mais conhecidos algoritmos para seleção de amostras. O algoritmo KS inicia os cálculos selecionando as duas amostras com a maior distância Euclidiana entre si no espaço $\mathbf{X}$. Para cada uma das amostras que permaneceram, calcula-se a distância mínima com relação às amostras já selecionadas. A partir de então, a

amostra com a maior distância mínima é retida, e o procedimento é repetido inúmeras vezes até que um determinado número de amostras seja selecionado (KENNARD; STONE, 1969).

O algoritmo de Kennard-Stone (KS) é utilizado na seleção das amostras do conjunto de validação (teste) e de calibração (treino) para o banco de dados. A partir do conjunto original dos dados, o algoritmo divide o total de amostras em dois conjuntos: treinamento que é responsável pela construção do modelo de regressão e o de teste que tem como objetivo avaliar a capacidade preditiva do modelo.

1.2.4. Normalização

O propósito da normalização dos dados originais é igualar a magnitude de cada observação (amostra), removendo assim as informações de distância de cada amostra com relação à origem, mas ao mesmo tempo preservando a direção (FERREIRA, 2015).

Na normalização, os valores de cada uma das variáveis de uma dada amostra i são divididos por um fator de normalização, podendo ser pela norma $\|X_i\|$ dessa amostra, por exemplo. Desta forma, todas as amostras passam a apresentar uma escala fixa. Na Equação 4, a expressão da normalização é apresentado para cada elemento de uma linha da matriz de dados (FERREIRA, 2015), com $j = 1, 2, \dots, J$.

$$x_{ij(norm)} = \frac{x_{ij}}{\|X_i\|} \quad (\text{Equação 4})$$

Segundo Ferreira, 2015, as normas mais utilizadas são: a norma *sup* ou (l_∞) , norma um (l_1) e a norma Euclideana (l_2) , apresentadas nas equações 5, 6 e 7, respectivamente.

$$\|X_i\|_\infty = \max_{1 \leq j \leq J} |x_{ij}| \quad (\text{Equação 5})$$

$$\|X_i\|_1 = \sum_{j=1}^J |x_{ij}| \quad (\text{Equação 6})$$

$$\|X_i\|_2 = \sqrt{\sum_{j=1}^J x_{ij}^2} \quad (\text{Equação 7})$$

1.3 Modelos de Regressão Multivariada

Nesta tese, as ferramentas aplicadas para obter modelos de regressão podem ser divididas em dois grupos: métodos lineares (PLS) e não lineares (SVM, RF e ANN). Além disso, foram utilizadas ferramentas de seleção de variáveis na tentativa de melhorar os modelos de previsão.

1.3.1 Mínimos Quadrados Parciais (PLS)

O PLS é um algoritmo de regressão que descreve modelos relacionados aos blocos de variáveis X e Y. Isso significa que as informações sobre as medidas das variáveis independentes (X) e as concentrações de ozônio (Y) são aplicadas simultaneamente na fase de calibração. O PLS é um dos mais utilizados e foi recomendado pela primeira vez por Herman Wold (WOLD, 1982). A informação das variáveis é compactada, o que a torna mais robusta e, como resultado, os modelos são menos onerosos para interpretar (MARTENS; NAES, 1989). De forma simplificada, o modelo PLS consiste na regressão entre os escores das matrizes X e Y. Ao construir o modelo, utiliza a matriz X ($m \times n$) no eixo X e o vetor Y ($n \times 1$) no eixo y. De forma simplificada, o modelo PLS consiste na regressão entre os escores das matrizes X e Y. Utiliza a matriz X ($m \times n$) no eixo X e o vetor Y ($n \times 1$) no eixo y na construção do modelo. Este modelo é considerado especialmente como uma relação externa entre as matrizes X e Y individualmente e, posteriormente, como uma relação interna que relaciona as duas matrizes (X e Y). Em termos práticos, o PLS assume que existem erros em ambas as matrizes, que são de igual importância. A relação externa para X pode ser expressa como a soma das novas matrizes, originadas da decomposição de X, conforme mostrado na Equação 8.

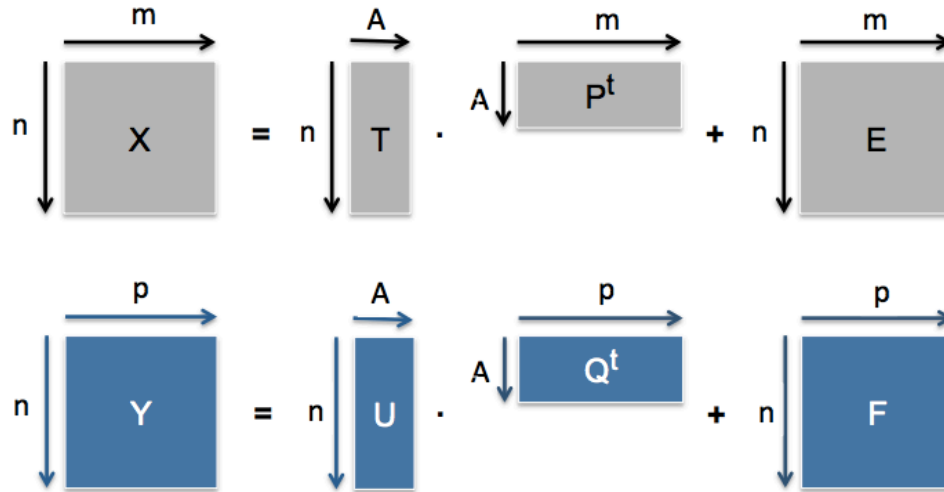
$$X = T \cdot P^t + E_x \quad (\text{Equação 8})$$

A relação externa Y segue o mesmo caminho, conforme apresentado na Equação 9:

$$Y = U \cdot Q^t + E_y \quad (\text{Equação 9})$$

onde X e Y são as matrizes decompostas, T e U são as matrizes dos escores ($m \times VL$), P^t e Q^t têm matrizes transpostas dos pesos ($VL \times n$) e E_x e E_y são as matrizes de erros das matrizes X e Y, respectivamente (BRERETON, 2003). A decomposição das matrizes X e Y pode ser observada no esquema da Figura 1:

Figura 1 - Esquema de decomposição das matrizes X e Y pela aplicação do PLS



Fonte: Adaptado de Wold *et al.* (2001) e Luna *et al.* (2017)

Neste estudo, o critério R de Wold foi aplicado para definir o número apropriado de variáveis latentes (VL) a serem incluídas na construção do modelo PLS. De acordo com os estudos de Stone e Wold, o critério R de Wold é baseado na validação cruzada, onde os dados (X e Y) são divididos em k blocos, e um modelo de variável latente é construído a partir de (k-1) blocos de dados (STONE, 1974; WOLD, 1978). Desta forma, o bloco excluído é usado para teste e um único *Predicted Residual Error Sum of Squares* (PRESS) é calculado. Este processo é repetido excluindo cada bloco de cada vez e, em seguida, o PRESS total é calculado para uma variável latente adicionando os valores de PRESS individuais. Para cada uma das variáveis latentes, o PRESS (Equação 10) total é calculado e, em seguida, uma série de valores de PRESS são obtidos (STONE, 1974; WOLD, 1978).

$$PRESS = \sum_{i=1}^n (y_i - \hat{y}_{(i)})^2 \quad (\text{Equação 10})$$

onde $\hat{y}_{(i)}$ é o estimador e y_i é um vetor resposta

Além disso, de acordo com Krzanowski, o critério R de Wold pode ser aplicado na escolha do número apropriado de variáveis latentes a serem incluídas em cada modelo de previsão por PLS (KRZANOWSKI, 1987). A diferença é que o critério R de Wold ajustado usa 0,95 e 0,90 como limiares, ao invés de adotar a unidade como um limiar como no critério R de Wold, dada a variabilidade da amostragem. O critério R de Wold sugere que uma variável

latente adicional não será incluída no modelo PLS, a menos que ofereça previsões significativamente melhores (KRZANOWSKI, 1987).

1.3.2 Máquina de Vetores Suporte (SVM)

Máquina de vetores suporte (SVM) é uma técnica de aprendizado de máquina desenvolvida por Vapnik, que foi originalmente desenvolvida para problemas de reconhecimento de padrões (VAPNIK e CHERVONENKIS, 1971). O modelo consiste em vários vetores suporte (amostras selecionadas do conjunto de calibração) e coeficientes de modelo não linear que definem o mapeamento não linear de variáveis no bloco de entrada x . O modelo permite a previsão da variável contínua do bloco y . Um grupo de cientistas usou a técnica SVM para estimar relações fonte-receptor altamente não lineares entre as emissões de precursores e as concentrações de poluentes. Tanaskuli *et al.* (2020) utilizaram a técnica de SVM para prever a concentração do ozônio troposférico na Malásia e utilizaram três estações de monitoramento para realização do estudo, obtendo um R^2 de 0,92412.

Feng *et al.* (2011) realizaram um estudo sobre previsão de ozônio troposférico em Beijing, China em 2011. Os autores utilizaram técnicas de SVM e algoritmo genético com a otimização por *Back Propagation neural network* (BPNN). O modelo utilizou as variáveis T, UR, VV e RS entres os períodos de março de 2009 a julho 2009. O coeficiente de correlação (R) encontrado para o melhor modelo foi de 0,87 e o RMSE foi de 36,56.

Na visão de Wang *et al.* (2008a), a aproximação por regressão SVM aponta a um problema de estimação de uma função com base em dados de um determinado conjunto, definido pela função: $G = \{(x_i, y_i)\}_{i=1}^l$ (onde $x_i \in R^n$ representa os vetores de entrada, $y_i \in R$ os valores desejados ou propriedades), que é produzido de Φ . A regressão por SVM aproxima a expressão na forma da Equação 11.

$$f(x) = \sum_{i=1}^l w_i \Phi_i(x) + b \quad (\text{Equação 11})$$

Em que, $\{\Phi_i(x)\}_{i=1}^l = 1$ representa as variáveis de entrada, $\{w_i\}_{i=1}^l = 1$ e b são os coeficientes. Estes coeficientes podem ser estimados pela minimização da função risco de acordo com a Equação 12, onde $L_\varepsilon(y, f(x))$ é retratada pela Equação 13.

$$R(C) = C \frac{1}{l} \sum_{i=1}^l L_{\varepsilon}(y_i, f(x_i)) + \frac{1}{2} \|w\|^2 \quad (\text{Equação 12})$$

$$L_{\varepsilon}(y_i, f(x_i)) = \begin{cases} |y - f(x)| - \varepsilon & |f(x) - y| \geq \varepsilon \\ 0 & \text{outros} \end{cases} \quad (\text{Equação 13})$$

Considerando que ε é um parâmetro esperado e $L_{\varepsilon}(y, f(x))$ é uma função perda (ou custo) ε -indiferente, que não penaliza erros menores que ε . O termo $\frac{1}{2} \|w\|^2$ é aplicado como uma medida do nivelamento da função e C é a constante de regularização que define a troca entre o erro de calibração e o nivelamento do modelo. Ao incluir no modelo as variáveis de folga (ζ, ζ^*) , têm-se as Equações 14 e 15 a serem minimizadas, atendendo as restrições indicadas pelas Equações 16 e 17 Wang *et al.* (2008a).

$$R(w, \zeta^*) = \frac{1}{2} \|w\|^2 + C^* \sum_{i=1}^n (\zeta_i + \zeta_i^*) \quad (\text{Equação 14})$$

$$w\Phi(x_i) + b - y_i \leq \varepsilon + \zeta_i \quad (\text{Equação 15})$$

$$y_i - w\Phi(x_i) - b - y_i \leq \varepsilon + \zeta_i \quad (\text{Equação 16})$$

$$\zeta, \zeta^* \geq 0 \quad (\text{Equação 17})$$

Conforme Wang *et al.* (2008a), caso a Equação 11 seja reescrita de maneira explícita, será obtida a Equação 18.

$$f(x, \alpha_i, \alpha_i^*) = \sum_{i=1}^l (\alpha_i - \alpha_i^*) K(x, x_i) + b \quad (\text{Equação 18})$$

Na Equação 19 são observados os multiplicadores Lagrangianos, α_i e α_i^* , que satisfazem a igualdade: $\alpha_i \cdot \alpha_i^* = 0$, $\alpha_i \geq 0$, $\alpha_i^* \geq 0$; e l é o número de vetores suporte (WANG *et al.*, 2008a; NAGUIB e DARWISH, 2012). Realizando $i = 1, \dots, l$ pode ser alcançado pela maximização da forma dual da Equação 19, considerando as restrições demonstradas nas Equações 20 e 21 (PLATT, SCHÜLKOPF,; BURGESS; SMOLA, 1998; VAPNIK, 1998; YEGANEH; MOTLAGH,; RASHIDI; KAMALAN, 2012).

$$\begin{aligned} \Phi(\alpha, \alpha^*) = & \sum_{i=1}^l y_i(\alpha_i - \alpha_i^*) - \varepsilon \sum_{i=1}^l (\alpha_i + \alpha_i^*) \\ & - \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(\alpha_i, \alpha_j) \end{aligned} \quad (\text{Equação 19})$$

$$\sum_{i=1}^l (\alpha_i - \alpha_i^*) = 0 \quad (\text{Equação 20})$$

$$\begin{aligned} 0 & \leq \alpha_i \leq C \\ 0 & \leq \alpha_i^* \leq C \end{aligned} \quad (\text{Equação 21})$$

Sabendo que $K(x_i, x_j) = \Phi(x_i)^t \cdot \Phi(x_j)$, é conhecida como função *Kernel*. Segundo Vapnik (1998) e Schölkopf e Smola (2002), as funções *Kernels* mais aplicadas são as lineares, quadráticas e *Radial Basis Function* (RBF), sendo que para os dados não lineares o *Kernel* RBF é o mais indicado. Para o modelo de regressão, dois tipos de vetores suporte são extensamente aplicados, o Nu-SVR e *Epsilon*-SVR. O algoritmo SVM utiliza a validação cruzada para selecionar as melhores faixas de valores para estes parâmetros e, então o modelo é concebido.

O algoritmo SVM possui parâmetros importantes que devem ser considerados, tais como os valores de *cost*, *gamma* e *epsilon*. O parâmetro *cost* retrata diretamente a penalidade associada a erros maiores que *epsilon*, uma vez que o aumento do valor do custo (*cost*) faz com que o ajuste do modelo seja mais próximo. O parâmetro *gamma* está relacionado ao aumento do número de vetores suporte, devido a sua principal característica que é controlar a forma de separação do hiperplano. Para a função de regressão SVM não há nenhuma penalidade associada aos pontos que estão previstos dentro do valor real da distância de *epsilon*, sendo que a diminuição do valor de *epsilon* faz com que o ajuste seja mais apropriado aos dados de calibração (BARRY M. WISE, NEAL B. GALLAGHER, RASMUS BRO, JEREMY M. SHAVER, WILLEM WINDIG, 2006).

1.3.3 Florestas Aleatórias (RF)

O algoritmo de floresta aleatória (RF), desenvolvido por Breiman, é uma das abordagens de aprendizagem de máquina (estatística) bem-sucedidas para implementações práticas, como ciência ambiental (ZHANG *et al.*, 2017), alimentos (LIU *et al.*, 2018), ecologia (CUTLER *et*

al., 2007) estudos de associação genética (CHEN e ISHWARAN, 2012) e engenharia (SHAHBAZI *et al.*, 2017). É feita uma distinção entre modelos estatísticos (distribuições de probabilidade para descrever dados) e modelos algorítmicos (modelos de caixa preta para fins de previsão e estimativa). Cox e Efron (2017) estudaram a diferença entre modelos estatísticos e algorítmicos, com ênfase na predição usando dados ruidosos, ao invés de tentar interpretar os dados. Em um futuro próximo, o papel das florestas aleatórias como uma estrutura genérica para modelagem preditiva parece ser a direção predominante na pesquisa relacionada a RF (TYRALIS; PAPACHARALAMPOUS; LANGOUSIS, 2019). Em suma, RF pode ser definido como um algoritmo de aprendizagem supervisionado que combina várias árvores de decisão de forma a construir uma floresta, melhorando a precisão, obtendo uma previsão estável e interpretação direta, este método adiciona aleatoriedade ao modelo e busca a característica mais crítica durante a divisão um nó, e geralmente resulta em um modelo melhor (BREIMAN, 2001; AMJAD *et al.*, 2018; WANG *et al.*, 2018).

1.3.4 Rede Neural Artificial (ANN)

As Redes Neurais Artificiais se iniciaram em 1943, quando o neurofisiologista Warren McCulloch e o matemático Walter Pitts, da Universidade de Illinois e Chicago, respectivamente e pertencentes a faculdade de medicina, fizeram uma analogia entre as células neuronais e o processo eletrônico em um artigo publicado no *Bulletin of Mathematical Biophysics* com o título denominado “*A Logical Calculus of the Ideas Immanent in Nervous Activity*” (MCCULLOCH e PITTS, 1943).

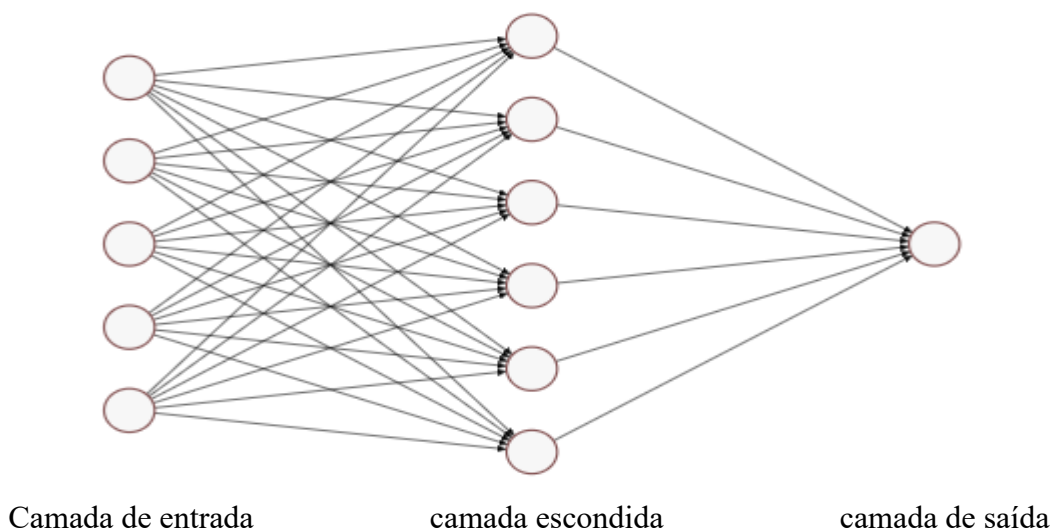
Após a implementação da técnica, diversos trabalhos utilizaram a ideia de Mcculloch e Pitts (1943) para a aplicação em várias áreas do conhecimento. A rede neuronal sustenta soluções em problemas complexos de classificação e regressão, para os quais a literatura reconhece os benefícios no qual esta técnica oferece em comparação às técnicas de modelagem estatística mais tradicionais. ANN são muito utilizadas para a modelagem de dados atmosféricos, uma vez que tais dados são conhecidos por serem complexos e por possuírem relações entre as variáveis dependentes e independentes não lineares. Entretanto computadores eram muito rudimentares naquela época, somente a partir da década de 90 aplicações mais sofisticadas se desenvolveram com mais intensidade com a evolução computacional. Redes Neurais Artificiais são técnicas computacionais que demonstram um algoritmo matemático

inspirado na estrutura neuronal de organismos inteligentes a qual adquirem conhecimento através da experiência (tentativa e erro).

Sousa *et al.* (2007), aplicaram *multiple linear regression* (MLR) e *feedforward artificial neural network* (FANN) para a previsão de O₃ na cidade do Porto, situado ao Nordeste de Portugal. Os autores utilizaram como variáveis NO, NO₂, O₃, T, UR e VV e aplicaram estudos de PCA para a seleção de variáveis de entrada dos modelos e o melhor resultado que obtiveram foram para o FANN com RMSE de 21,78 µg m⁻³ e com coeficiente de correlação (R) de 0,73.

Nos dias atuais, as ferramentas de modelagem são amplamente utilizadas em muitos campos científicos, especialmente nas ciências ambientais, como a poluição do ar (NAJAFPOOR *et al.*, 2014). Nesses casos, vários métodos estatísticos e computacionais podem ser usados para prever as concentrações de poluentes atmosféricos usando uma série de dados coletados antecipadamente. Nos últimos anos, o modelo de rede neural artificial (ANN) tem sido considerado um método de baixo custo operacional para obter valores de previsão confiáveis de poluentes do ar (AZID *et al.*, 2014). É um modelo computacional que replica a função simples de uma rede biológica e é usado para resolver funções complexas não-lineares. Os neurônios estão localizados nas camadas da rede do modelo. As camadas são definidas como a entrada, a saída e as camadas ocultas. Existem muitos tipos de diferentes redes neurais. A estrutura mais comum da rede neural é o "*feed forward*", onde o fluxo de dados das unidades de entrada para saída é estritamente avançado. O esquema típico desta estrutura é mostrado na Figura 2. As ANN são capazes de encontrar e identificar padrões complexos em conjuntos de dados que podem não ser bem descritos por uma fórmula matemática simples ou um conjunto de processos conhecidos (ABYANEH, 2014).

Figura 2- Esquema típico de uma rede neural artificial



Fonte: Autor, 2022.

Redes neurais artificiais (*artificial neural networks* - ANN) de forma análoga ao sistema nervoso humano, possuem nós em uma ou mais camadas, as quais estão ligadas por conexões, chamadas sinapses. Estudos de Fiorin *et al.* (2011) sugerem que esta técnica é capaz de armazenar conhecimento ao longo da sua evolução, e seu uso abrange problemas de ajuste de funções, reconhecimento de padrões, modelagem preditiva e possuem outras aplicações em diversas áreas. Este método tem uma capacidade de auto-organização e processamento temporal que permite resolver problemas distintos de alta complexidade. Neste contexto, *perceptrons* de múltiplas camadas (*perceptron multilayer* - MLP) é um tipo de rede neural artificial e tem se mostrado uma valiosa ferramenta para a predição, a aproximação da função e a classificação. Os benefícios da abordagem do MLP foram particularmente evidentes nas aplicações onde um modelo teórico completo não pode ser construído, e especialmente quando se trata de sistemas não lineares (GARDNER e DORLING, 1998). De acordo com Abdul-Wahab; Bakheit; Al-Alawi (2005), modelos com base em ANN tem o potencial de descrever relações não-lineares como aqueles que controlam a produção de O_3 . Além disso, esta técnica foi utilizada com sucesso por esses autores para prever a concentração de O_3 usando dados meteorológicos e de qualidade do ar.

Para construir uma rede neural artificial é necessário estabelecer uma arquitetura para a rede e, em seguida, usar um algoritmo para encontrar os pesos das conexões entre os nós. A rede pode conter muitas camadas intermediárias entre suas camadas de entrada e saída. Essas

camadas intermediárias são chamadas camadas ocultas e os nós incorporados nessas camadas são chamados de nós ocultos (ZHAO e HASAN, 2013). A rede pode usar tipos de funções de ativação, como as funções tangentes lineares, sigmóides (logísticas) e hiperbólicas, entre outras. No pacote da Linguagem R "nnet", a função sigmoid é o padrão utilizado para o modelo de classificação. Sua expressão é mostrada na Equação 22.

$$f(x) = \frac{e^x}{1 + e^x} \quad (\text{Equação 22})$$

O algoritmo *backpropagation* (retro-propagação) é o mais comumente empregado no treinamento supervisionado de redes MLP. Em uma primeira fase ocorre a propagação do sinal funcional (*feed-forward*) mantendo-se os pesos fixos de modo a gerar um valor de saída a partir das entradas fornecidas à ANN. Na segunda fase, as saídas são comparadas com os valores desejados, gerando um sinal de erro, que se propaga da saída para a entrada, ajustando-se os pesos de forma a minimizar o erro (retro-propagação do erro) (FIORIN *et al.*, 2011).

O algoritmo de *backpropagation* (ou seja BP) é usado em propagação em camadas ANN (HAGAN *et al.*, 1996). O algoritmo BP usa a aprendizagem supervisionada, o que significa que uma vez fornecido ao algoritmo com exemplos das entradas e saídas a qual uma rede é computada e, em seguida, o erro é calculado. Existem duas fases em cada iteração do algoritmo BP: i) A fase de avanço (*forward phase*): durante esta fase de avanço, as saídas dos neurônios no nível k são calculadas antes de computar as saídas no nível $k + 1$; ii) A fase de atraso, (*backward phase*) durante esta fase de atraso, os pesos no nível $k + 1$ são atualizados antes que os pesos no nível k sejam atualizados. Este algoritmo de BP permite usar o erro para os neurônios na camada $k + 1$ para estimar os erros para os neurônios na camada k .

Então, $D = \{(x_i, y_i) \mid i = 1, 2, \dots, N\}$ seja o conjunto de exemplos de treinamento. O objetivo do algoritmo de aprendizagem ANN é determinar um conjunto de pesos (*weights* - w) que minimizem um determinado estimador escolhido, como, por exemplo, a soma total de erros quadrados (Equações 23 e 24).

$$E(w) = \frac{1}{2} \times \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (\text{Equação 23})$$

onde y_i representa o valor de saída esperado do nó i , $\hat{y}$ é a saída gerada pelo nó i executando um somatório dos resíduos de sua entrada e a expressão de atualização de peso usada pelo método de descida gradiente.

$$\Delta w_{ij} = -\lambda \times \frac{\partial E(w)}{\partial w_{ij}} \quad (\text{Equação 24})$$

onde λ é a taxa de aprendizado.

1.4 Figuras de Mérito

Algumas figuras de mérito foram empregadas neste trabalho, o que é fundamental para a avaliação e comparação de modelos como: *root mean square error of calibration* (RMSEC), coeficiente de correlação da calibração (R_{cal}), *root mean square error of cross validation* (RMSECV), coeficiente de correlação da validação cruzada (R_{cv}), *root mean square error of prediction* (RMSEP), coeficiente de correlação de previsão (R_{prev}), coeficiente de determinação (R^2), viés (*bias*), *root square error of prediction* (RSEP) e *sum of Wilcoxon test probabilities* (SWTP). As figuras de mérito apresentadas foram utilizadas para avaliar qual modelo apresentou capacidade preditiva mais significativa.

O RMSEC é um parâmetro que fornece uma visão geral da capacidade do modelo, uma vez que é uma medida da diferença média entre o valor previsto e o valor real. Os valores RMSEC foram estimados de acordo com a Equação 25, já o coeficiente de correlação da calibração (R_{cal}), Equação 26. Ambos são obtidos a partir dos valores estimados pela validação cruzada e os valores conhecidos.

$$RMSEC = \sqrt{\frac{\sum_{i=1}^{I_c} (y_i - \bar{y})^2}{I_c - A}} \quad (\text{Equação 25})$$

$$R_{cal} = \frac{\sum_{i=1}^{I_c} (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{I_c} (\hat{y}_i - \bar{\hat{y}})^2 (y_i - \bar{y})^2}} \quad (\text{Equação 26})$$

onde $\hat{y}_i$ é o valor estimado da amostra i (não está incluída no modelo construído); y_i é i -th os valores de y e I_c é o número de amostras de calibração e A é o número de variáveis latentes.

A *root mean square error of cross validation* (RMSECV) esta demonstrada pela Equação 27 e o coeficiente de correlação da validação cruzada (R_{cv}) esta demonstrada pela Equação 28. Ambos estes parâmetros são obtidos a partir dos valores estimados pela validação cruzada (para esta Tese, os blocos contíguos) e os valores conhecidos. Conforme descrito por Wise *et al.* (2006); Ballabio e Consonni (2013), tem-se:

$$RMSECV = \sqrt{\frac{\sum_{i=1}^S (y_i - \bar{y})^2}{S}} \quad (\text{Equação 27})$$

$$R_{cv} = \frac{\sum_{i=1}^S (\hat{y}_i - \bar{y})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^S (\hat{y}_i - \bar{y})^2 (y_i - \bar{y})^2}} \quad (\text{Equação 28})$$

onde $\hat{y}_i$ é o valor estimado para a amostra i , não incluída na construção do modelo, S é o número de segmentos ou divisões no conjunto de dados e $\bar{y}$ é a média dos valores em y .

Outro parâmetro utilizado neste trabalho foi o RMSEP dado pela Equação 29. Conforme relatado por Mevik e Cederkvist (2004), para a etapa de predição, um conjunto de variáveis independentes foi fornecido ao modelo com componentes h para avaliar seu poder preditivo. E o coeficiente de correlação de previsão (R_{prev}) apresentado na Equação 30.

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{I_p} (\hat{y}_i - y_{i,measured})^2}{I_p}} \quad (\text{Equação 29})$$

$$R_{prev} = \frac{\sum_{i=1}^{I_p} (\hat{y}_i - \bar{y})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{I_p} (\hat{y}_i - \bar{y})^2 (y_i - \bar{y})^2}} \quad (\text{Equação 30})$$

onde $\hat{y}_i$ é o valor estimado do modelo previsto para a amostra i , I_p é o número da amostra no conjunto de dados de predição e $y_{i,measured}$ é obtido pelo valor medido e $\bar{y}$ é a média dos valores em y . Após a construção e validação do modelo, ele foi usado para prever novas amostras.

O coeficiente de determinação (R^2) é o parâmetro que explica o grau de ajuste do modelo e pode ser calculado pela razão entre a soma quadrada da regressão e a soma quadrada total, conforme mostrado na Equação 31 (CORNELL, 1987).

$$R^2 = \frac{SS_{regression}}{SS_{total}} = \frac{SS_{total} - SS_{residuals}}{SS_{total}} = 1 - \frac{SS_{residuals}}{SS_{total}} \quad (\text{Equação 31})$$

O viés (*bias*) é a tendência ou erro sistemático e pode ser descrito como a diferença média entre o valor predito e o valor medido, através da Equação 32 (FERREIRA, 2015).

$$bias = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n} \quad (\text{Equação 32})$$

onde n é o número de amostras do conjunto de teste, y_i é valor estimado pelo modelo, $\hat{y}_i$ é o valor real ou medido da propriedade.

O RSEP é outro parâmetro utilizado para avaliar o desempenho do modelo construído, conforme a Equação 33 (CORNELL, 1987).

$$RSEP(\%) = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i)^2}} \cdot 100\% \quad (\text{Equação 33})$$

O SWTP foi utilizado para comparar e escolher o modelo mais adequado para cada estações do ano (primavera, verão, outono e inverno) de cada uma das estações meteorológicas estudadas. De acordo com Cunha *et al.* (2020), o SWTP (Tabela 2) é baseado no teste de Wilcoxon, que é um teste não paramétrico e compara dois modelos a cada vez, calculando a diferença entre cada conjunto de pares e analisando essas diferenças. O vetor de resíduos de cada modelo foi utilizado para calcular a probabilidade do teste de Wilcoxon (WTP). Para que o modelo i seja considerado melhor do que o modelo j , $WTP_{i,j}$ deve ter um valor pequeno. Ao comparar um conjunto de N modelos, uma matriz $N \times N$ é construída aplicando o teste de Wilcoxon, onde $i = 1, \dots, N$ e $j = 1, \dots, N$. Segundo Cunha *et al.* (2020), as linhas dessa matriz contêm os resultados da comparação do modelo i com todos os modelos do conjunto. E a soma de todas as probabilidades da i -ésima linha da matriz, mostrada na Equação 34, foi descrito como um valor que resume o comportamento do modelo. O valor mínimo deste vetor indicará o melhor modelo.

$$SWTP_i = \sum_{j=1}^N WTP_{i,j} - WTP_{i,i} \quad (\text{Equação 34})$$

Tabela 2- Esquema do cálculo da figura de mérito SWTP

Modelos	PLS	SVM	RF	ANN	...	M _n	SWTP
PLS	0,5 +	WTP _{1,2} +	WTP _{1,3} +	WTP _{1,4} +	... +	WTP _{1,n} =	
SVM	WTP _{2,1}	0,5	WTP _{2,3}	WTP _{2,4}	...	WTP _{2,n} =	
RF	WTP _{3,1}	WTP _{3,2}	0,5	WTP _{3,4}	...	WTP _{3,n} =	
ANN	WTP _{4,1}	WTP _{4,2}	WTP _{4,3}	0,5	...	WTP _{4,n} =	
...	...	...	...	...	...	...	
M _n	WTP _{n,1}	WTP _{n,2}	WTP _{n,3}	WTP _{n,4} =	...	0,5	

Menor SWTP
=
Melhor Modelo

Fonte: adaptado de Cunha *et al.*, 2020.

2. METODOLOGIA

A Região Metropolitana do Rio de Janeiro (RMRJ), no sudeste do Brasil, é a segunda região mais populosa do país, com mais de 12 milhões de pessoas, formada pela cidade do Rio de Janeiro e outras 20 cidades. A cidade do Rio de Janeiro está localizada na costa do Atlântico Sul, próxima a latitude $-22^{\circ}54'$ e longitude $-43^{\circ}12'$, com uma população de aproximadamente 6,5 milhões de habitantes, 3 milhões de veículos e com o PIB superior a 500 bilhões (IBGE, 2019). De acordo com (IBGE, 2019) o município possui 40 %, 7 % e 20 % da frota veicular movidos a gasolina, etanol e gás natural veicular (GNV), respectivamente, enquanto 28 % são de tecnologia *flexfuel* (mistura de gasolina e etanol) e 4 % utilizam o diesel como combustível. Os veículos do tipo *flexfuel* podem usar entre etanol (100 %) ou gasolina (gasolina misturada com 27 % de etanol). Aproximadamente 87 % são automóveis de passageiros. Os veículos movidos a diesel são principalmente caminhões, ônibus, microônibus e utilitários. Existem aproximadamente 55.000 táxis e 150.000 serviços de transporte baseados em aplicativos.

Os dados de poluição atmosférica utilizados neste projeto são provenientes de 4 estações meteorológicas e de qualidade do ar (EAQA) da Região Metropolitana do Rio de Janeiro (RMRJ). O banco de dados foi construído a partir da média horária diária de dados (24 horas por dia) fornecidos pelo INEA (<http://200.20.53.7/qualiar/home/index>), os dados foram tratados e estão disponíveis nos repositórios de dados do *Mendeley data* (DOI: 10.17632 / hpc938kng2.1) de 1º de janeiro de 2014 a 31 de dezembro de 2018. Essas estações fornecem dados horários para diversas variáveis, tais como NO₂, NO, CO, O₃, RS, T, UR, VV, DV, entre outras, que serão as variáveis estudadas nesta tese. O maior erro de medição envolvendo instrumentos é de $0,98 \mu\text{g m}^{-3}$, para todos os poluentes.

Para visualizar os dados de forma mais clara, foi criada uma Tabela A (vide Apêndice A com 91 estações) com todas as EAQA estudadas para a Região Metropolitana do Rio de Janeiro. Nesta primeira etapa foi realizado a separação e classificação das EAQA dentro de cada bacia aérea. Entende-se por Bacia Aérea áreas formadas pela orientação das vertentes e da topografia, que influenciam a direção dos ventos de superfície e no transporte de poluentes.

Na Figura 3 é apresentado um fluxograma do estudo de previsão do ozônio. Na próxima etapa, para determinar qual EAQA seria escolhida, foi necessário investigar as variáveis disponíveis, a quantidade de dados faltantes em cada EAQA e as principais tipos de fontes de

emissão também foram classificadas, tais como: industrial (I), residencial (R) e veicular (V). Após esta etapa, foram selecionadas quatro EAQA (Tabela 3): Bacia 1 – Adalgisa Nery (ADN), Bacia 2- INEA (INE- 3 anos de dados somente), Bacia 3- Vila São Luiz (VSL), Bacia 4- Porto das Caixas (PDC). Por conseguinte, foi necessária a conversão e integração de unidades para o sistema internacional (ex: $\mu\text{g m}^{-3}$) e posteriormente um estudo de estatística descritiva foi aplicado, para identificar a quantidade de dados faltantes para cada EAQA envolvida no estudo e excluir dados anômalos (ex: *outliers* extremos) com o objetivo de mitigar os erros e as tendências nos dados.

Tabela 3 - Descrição da EAQA escolhida para o estudo

EAQA	Latitude (S)	Longitude (O)	Classe	Bacia
ADN	- 22,88875	- 43,715970	R/I	I
INE	- 22,98928	- 43,414962	V	II
VSL	- 22,78455	- 43,286388	I/V	III
PDC	- 22,70164	- 42,874543	I	IV

I – Industrial; R – Residencial; V - Veicular

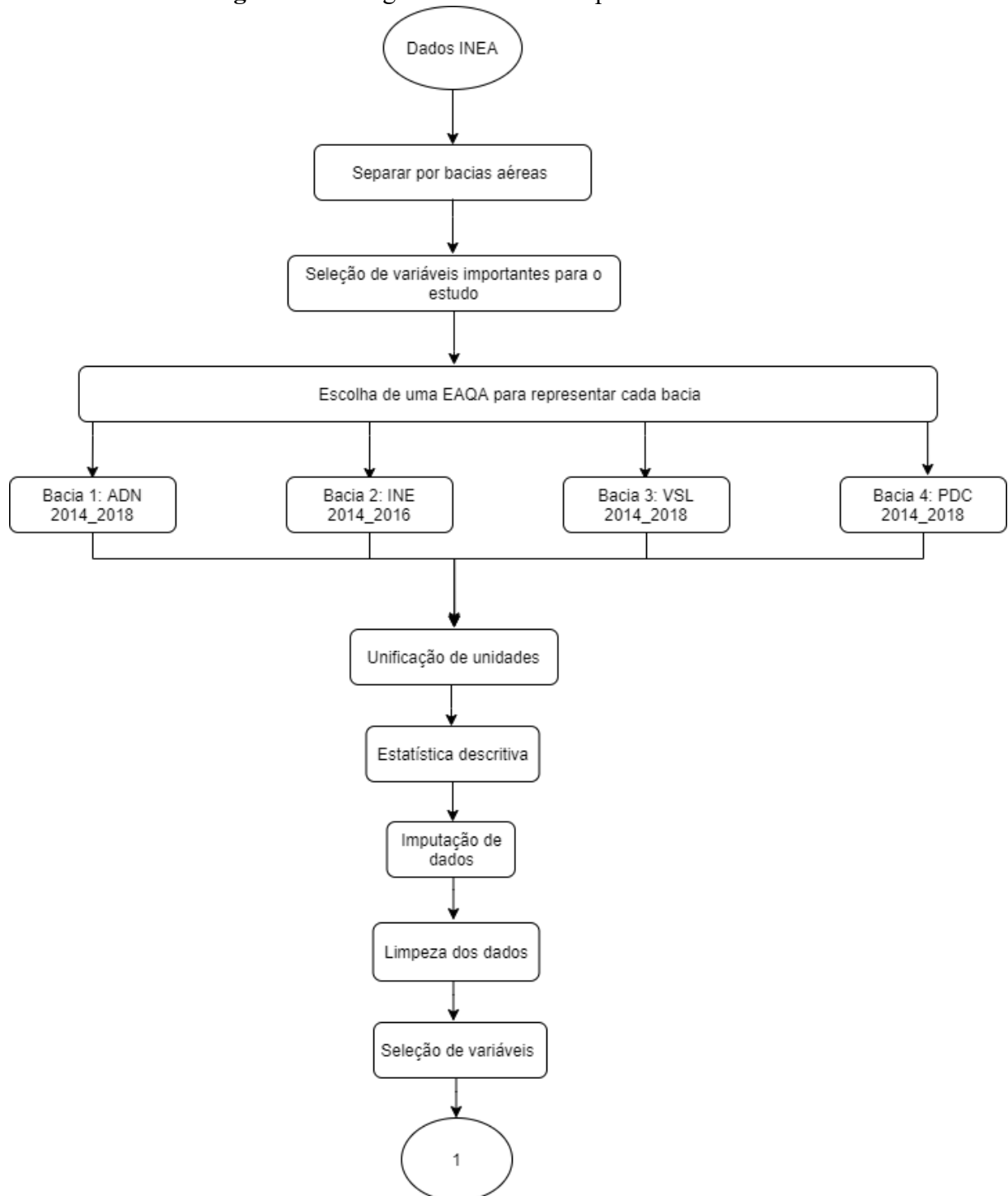
Fonte: O autor, 2022.

Logo após este estudo, foi aplicada a imputação dos dados faltantes com a utilização do algoritmo *MissForest* (com uma média de 10 rodadas) para se obter um banco de dados mas coeso. Todas as variáveis, incluindo aquelas que não são precursoras de ozônio, foram consideradas para imputação para evitar a perda de informações ou das características dos dados brutos.

Na etapa da limpeza dos dados, os dias chuvosos superiores a 4,0 mm foram retirados das bases de dados, uma vez que a chuva remove vários poluentes. Além disso, foram retirados os dados obtidos aos domingos e feriados, pois as emissões veiculares são drasticamente reduzidas nesses períodos e, portanto, as concentrações de poluentes são notavelmente diferentes dos dias da semana. Os dados das 19:30 às 5:30 h, obtidos à noite, também foram removidos porque a camada de mistura atmosférica é extremamente baixa, o que aumenta a concentração de alguns poluentes primários, e os processos fotoquímicos também são praticamente inexistentes. Um estudo de seleção de variáveis (correlação de Pearson) foi realizado, para conhecer a correlação entre as variáveis e o ozônio, com a finalidade de decidir quais variáveis deveriam ser escolhidas e/ou priorizadas para o estudo. Ventura *et al.* (2019)

utilizaram uma correlação de Pearson acima de 0,30 para priorizar quais variáveis seriam escolhidas para o modelo. Esta etapa, juntamente com o estudo de PCA, é muito importante para reduzir o esforço computacional e o erro acumulado.

Figura 3- Fluxograma do estudo de previsão do Ozônio



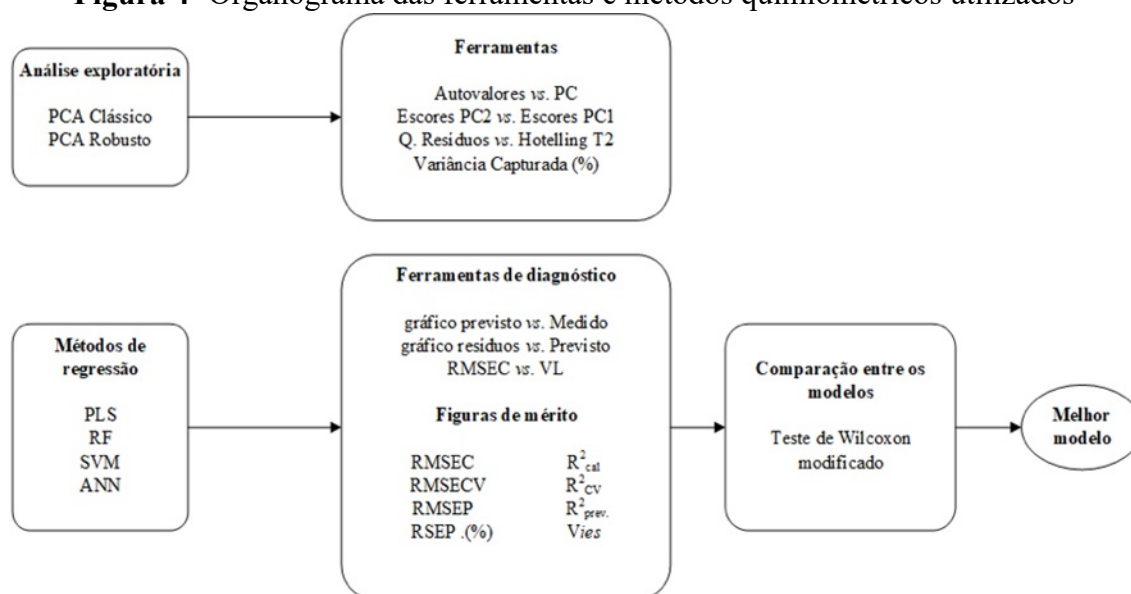
Fonte: O autor, 2022.

Na Figura 4 é apresentado um organograma dos métodos e das ferramentas aplicados a este estudo. A análise de componentes principais clássica (PCA) e robusta (ROBPCA) foram utilizadas como a principal ferramenta de análise exploratória dos dados e detecção de possíveis amostras atípicas.

O método de regressão linear aplicado foi a *Partial Least Squares Regression* (PLS) e os não-lineares foram o *Support Vector Machine* (SVM), *Random Forest* (RF) e *Artificial Neural Network* (ANN). Na aplicação do algoritmo PLS foram considerados alguns métodos de seleção de variáveis na tentativa de otimizar o modelo antes de testar os métodos não-lineares (SVM, RF e ANN).

Vale ressaltar que todos os modelos foram validados por uma etapa de validação interna. Como método de validação interna foi aplicado a validação cruzada k-fold=10.

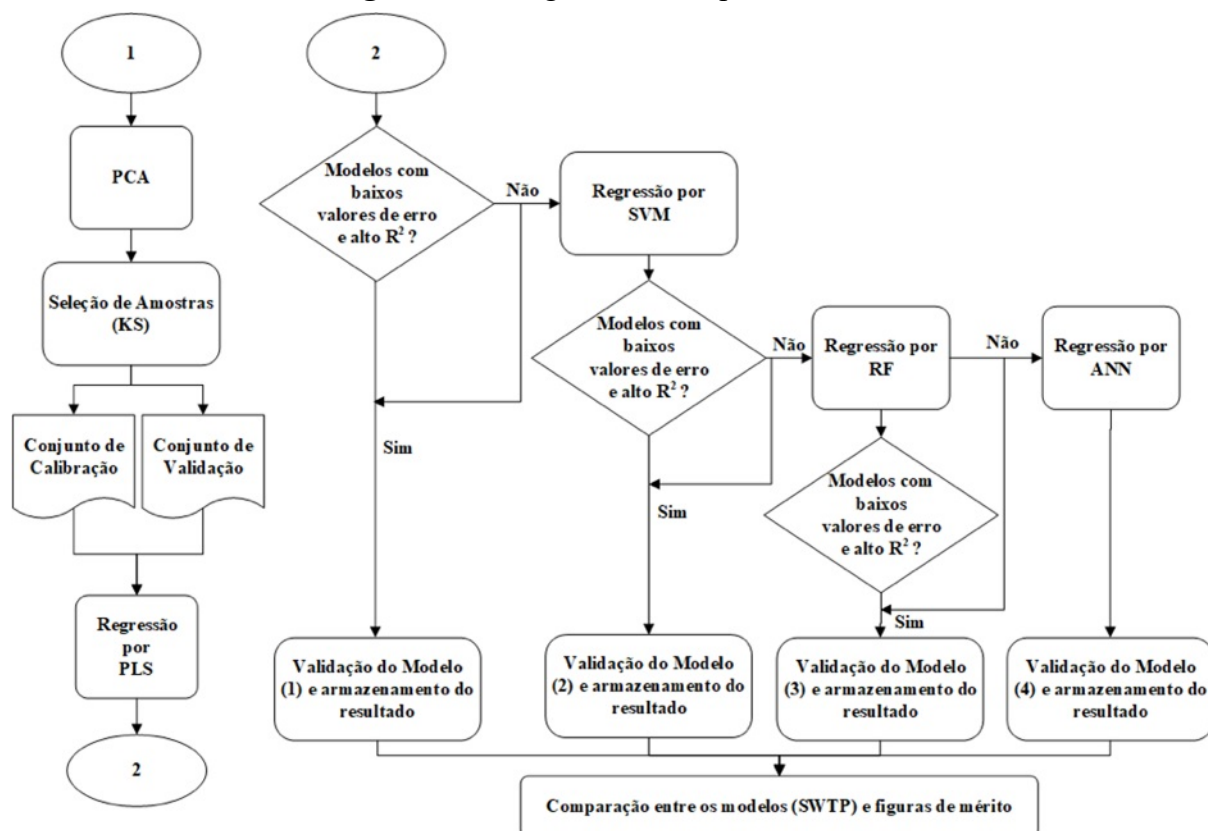
Figura 4- Organograma das ferramentas e métodos quimiométricos utilizados



Fonte: O autor, 2022.

No fluxograma da Figura 5 são apresentadas todas as fases do estudo desta tese, no qual foi contemplado desde a construção do banco de dados, o pré-tratamento deste banco, análise exploratória utilizando PCA clássico e robusto, utilização de algoritmo de seleção de amostras (algoritmo KS), seguido pela construção de modelos preditivos para o ozônio, comparação entre os modelos utilizando as figuras de mérito e o teste Wilcoxon pareado.

Figura 5- Fluxograma das etapas do estudo



Fonte: O autor, 2022.

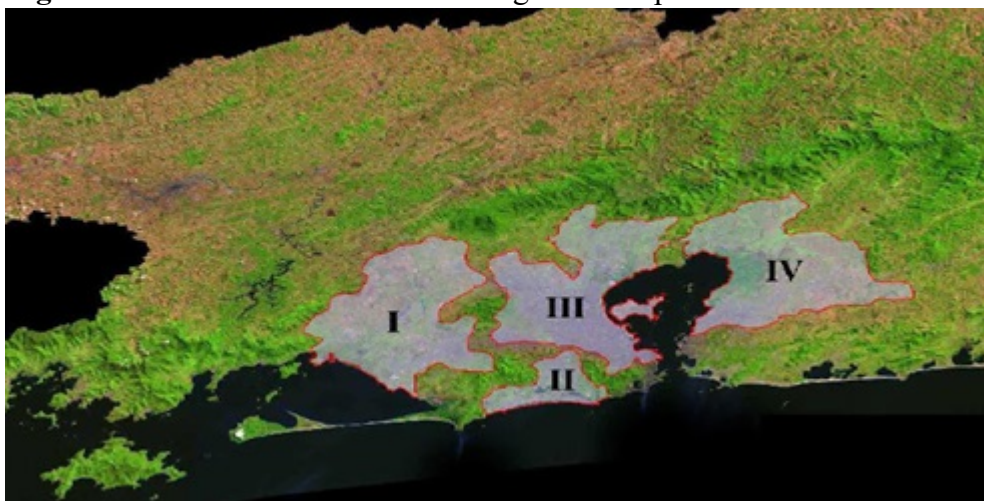
2.1 Descrição da Área Geográfica do Estudo

A RMRJ abrange uma área total de 4.690 km² (18 municípios - Rio de Janeiro, Mesquita, Nilópolis, São João de Meriti, Belford Roxo, Duque de Caxias, Nova Iguaçu, Japeri, Magé, Itaboraí, Tanguá, Queimados, Seropédica, Itaguaí, São Gonçalo, Maricá, Guapimirim e Niterói) e representa 11 % de todo o Estado do Rio de Janeiro (Figura 6). A RMRJ é uma das áreas mais industrializadas do Brasil e possui 4,7 milhões de veículos (IBGE, 2019). Em 2017, o número total de cidadãos é de cerca de 11 milhões e os altos níveis de poluição são um problema crítico de saúde na região (OLIVEIRA et al., 2017).

Nesta região são observadas atividades associadas aos setores: petroquímico, metalúrgico, geração de energia, polímeros, tintas, vernizes e produtos de química fina. Os municípios que se destacam no contexto regional pela produção industrial são: Duque de Caxias (pólo petroquímico de Campos Elíseos), Belford Roxo (indústria química), Niterói (indústria

naval, material de transporte, química, gráfica, e produtos alimentares), Nova Iguaçu (setor industrial moveleiro, produtos de perfumaria, bebidas e alimentos), São Gonçalo (minerais não metálicos, produtos alimentares, indústria farmacêutica e química), Seropédica (usinas termoeletricas) e o Distrito de Santa Cruz, no Rio de Janeiro (siderurgia com destaque à TKCSA e usinas termoeletricas, entre outros) (INEA, 2015).

Figura 6- Divisão da bacia aérea da Região Metropolitana do Rio de Janeiro



Fonte: INEA, 2015

Bacia aérea I: localizada na zona oeste da RMRJ, com cerca de 730 km²;

Bacia aérea II: localizada no município do Rio de Janeiro, envolve as regiões administrativas de Jacarepaguá e Barra da Tijuca, com cerca de 140 km²;

Bacia aérea III: compreende a zona norte do município do Rio de Janeiro e grande parte dos municípios da baixada fluminense, com cerca de 700 km²;

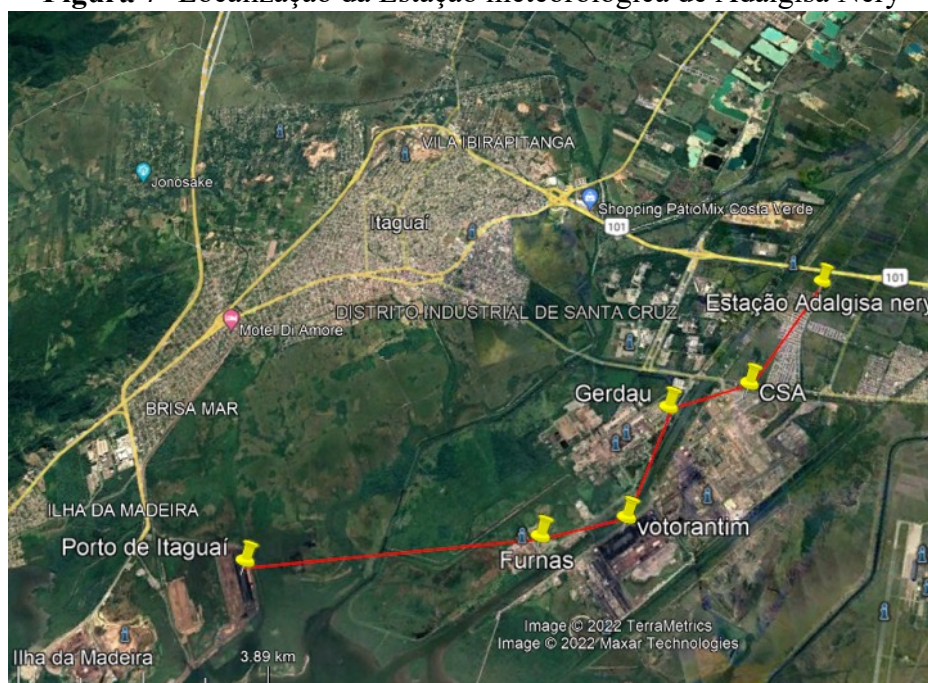
Bacia aérea IV: localizada a leste da baía de Guanabara possui 830 km² (INEA, 2015).

A qualidade do ar é um processo complexo e de difícil entendimento, devido às mudanças frequentes nas condições de dispersão atmosférica pela variação da meteorologia entre as diferentes estações do ano (MALBY *et al.*, 2013). Para o entendimento das características regionais para cada EAQA, um estudo da rosa dos ventos e variação no tempo, são úteis para capturar as características predominantes dos dados. Estes dados foram analisados de acordo com as técnicas mencionadas (estatística descritiva), com a finalidade de entender o comportamento das variáveis do estudo.

A EAQA da Estação de Adalgisa Nery (Figura 7), está localizada próxima a escola Municipal Adalgisa Nery – Rua Eduardo d’Aguiar Filho com coordenadas de Latitude: -

22,88875° e Longitude: -43,715970° . As distâncias das fontes fixas em km são: Furnas 5,89; Votorantim 4,84; CSA 1,97; Porto de Itaguaí 9,85; Gerdau-3,17. Mapeadas de acordo com a ferramenta do *Google Earth* para medir as distâncias entre as estações meteorológicas e as principais fontes emissoras.

Figura 7- Localização da Estação meteorológica de Adalgisa Nery

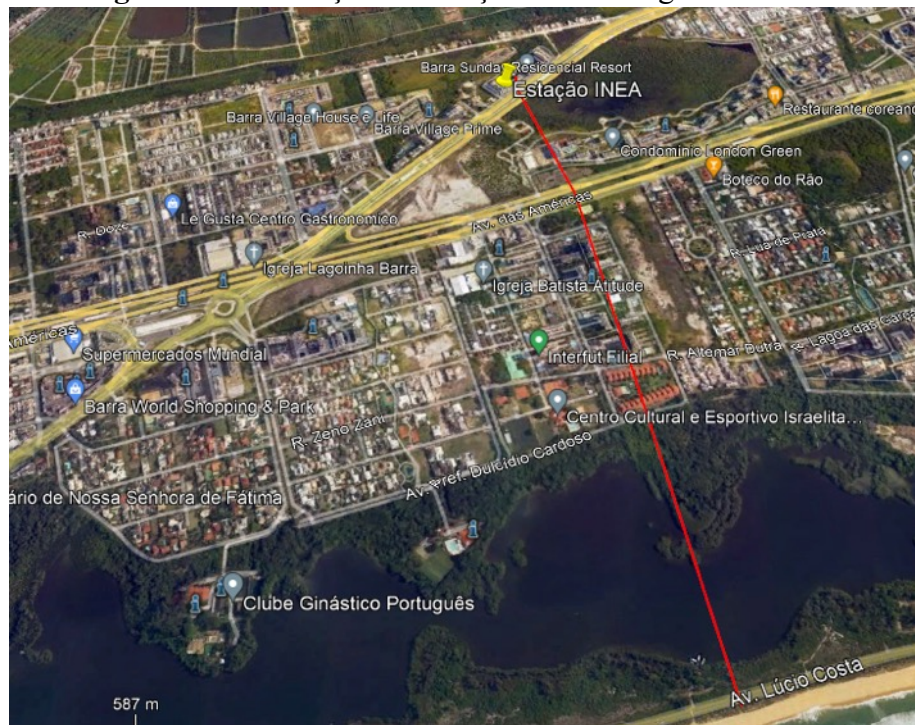


Fonte: Adaptado *Google Earth*, 2022.

A EAQA do INEA (Figura 8), está localizada próxima a Rua Salvador Allende nº 6300 com coordenadas de Latitude: -22,989281° e Longitude: -43,414962°. As distâncias das principais fontes móveis em km são: Avenida das Américas 0,53; Avenida Lúcio Costa 2,06. Mapeadas de acordo com a ferramenta do *Google Earth* para medir as distâncias entre as estações meteorológicas e as principais fontes emissoras.

A EAQA da Estação da Vila São Luiz (Figura 9), está localizada próxima ao Caxias Shopping na Rodovia Washington Luiz com coordenadas de Latitude: -22,78455° e Longitude: -43,286388°. As distâncias das principais fontes fixas em km são: Braskem 8,9; REDUC 7,21; Transpetro 8,84 e White Martins 8,85. Mapeadas de acordo com a ferramenta do *Google Earth* para medir as distâncias entre as estações meteorológicas e as principais fontes emissoras.

Figura 8- Localização da Estação meteorológica de INEA



Fonte: Adaptado *Google Earth*, 2022.

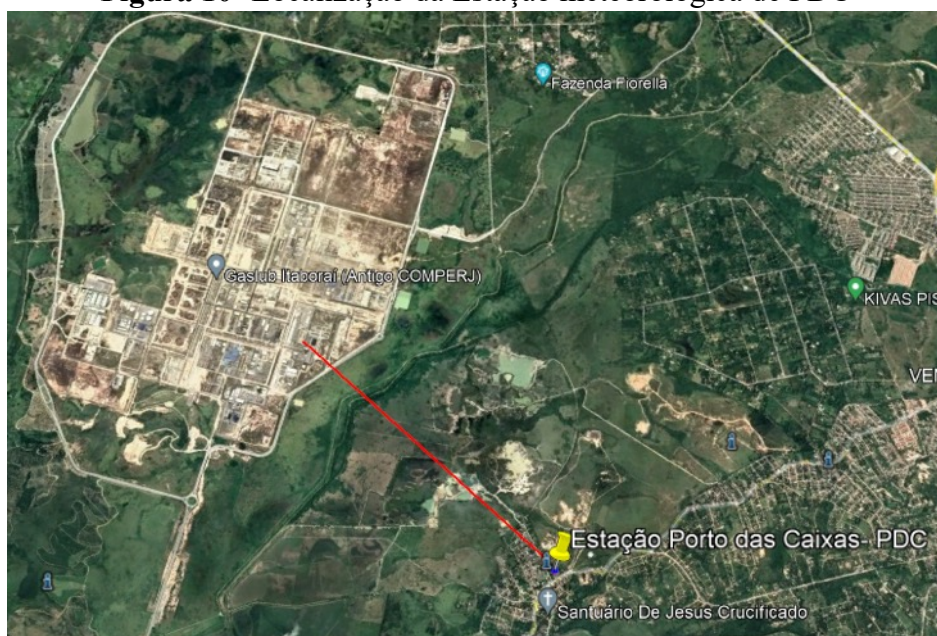
Figura 9- Localização da Estação meteorológica de VSL



Fonte: Adaptado *Google Earth*, 2022.

A EAQA da Estação de Porto das caixas (Figura 10), está localizada na Escola Estadual Professora Maria Inocência Ferreira, Rua da Conceição, nº 118 com coordenadas de Latitude: -22,70164° e Longitude: -42,8745430°. A distância da principal fonte fixa em km são: COMPERJ 3,71. Mapeadas de acordo com a ferramenta do *Google Earth* para medir as distâncias entre as estações meteorológicas e as principais fontes emissoras.

Figura 10- Localização da Estação meteorológica de PDC



Fonte: Adaptado *Google Earth*, 2022.

2.2 Linguagens Computacionais e Pacotes

O algoritmo *MissForest* foi utilizado para imputação dos dados faltantes, as linguagens R e Python foram aplicadas para realizar a previsão de O₃.

O *MissForest*, PCA, PLS, RF e SVM foram desenvolvidos utilizando a linguagem R (R CORE TEAM, 2022), e os pacotes *MissForest* (STEKHOVEN; BUEHLMANN, 2012; STEKHOVEN, DANIEL, 2013) *mdatools* (KUCHERYAVSKIY, 2015), *rrcov* (TODOROV; FILZMOSER, 2009), *pls* (MEVIK; WEHRENS; LILAND, 2015), *randomForest* (LIAW; WIENER, 2002), *e1071* (MEYER et al., 2018), *prospectr* (STEVENS; RAMIREZ-LOPEZ, 2013) e *openair* (CARSLAW; ROPKINS, 2012). Para a construção da ANN foi utilizado a linguagem Python com o pacote *MLPRegressor*. As outras ferramentas utilizadas na Tese estão apresentadas estão na Tabela 4.

Tabela 4- Lista das ferramentas, comandos, pacotes e respectivas métricas, originários do R, utilizados neste trabalho

Ferramentas	Tipo de Ferramenta	Comando	Pacote	Métrica	Referência
PCA	Análise exploratória	pca	{mdatools}	scale = FALSE; alpha=0.05; method="svd"; cv=10	Kucheryavskiy, 2015
PLS	Regressão de 1ª ordem	mvr	{pls}	method="simpls"; ncomp=10	Mevik et al., 2015
PCA Robusto	Análise exploratória	PcaHubert	{rrcov}	-	Todorov e Filzmoser, 2009
Kennard-Stone	Seleção de amostras	kenStone	{prospectr}	metric="mahal; k=100	Stevens e Ramirez-Lopez, 2013
Validação cruzada	Validação cruzada	crossval	{pls}	segments = 10; segment.type = "interleaved"	Mevik et al., 2015
RMSEC/ RMSECV/ RMSEP	Parâmetros	RMSEP	{pls}	-	Mevik et al., 2015
R2cal/ R2cv / R2val	Parâmetros	R2	{pls}	-	Mevik et al., 2015
Corplot	Seleção de variáveis	corPlot	{openair}	method=pearson	Sakar et al., 2007
SVM	Regressão	tune	{e1071}	kernel=c('linear', 'radial', 'polynomial', 'sigmoid')	Meyer et al., 2018
Florestas Aleatórias	Regressão	randomForest	{randomForest}	-	Liaw e Wiener, 2002

A rede neural foi construída no ambiente Python com o número de variáveis na camada de entrada variando de acordo com cada EAQA, para a camada escondida foi utilizada somente uma camada e o número de nós utilizados foram entre 40 e 100 respeitando o menor valor para o RMSEP encontrado, a função de ativação utilizada foi a *sigmoid* e a camada de saída é a variável dependente (O_3).

R e Python são uma linguagem estatística e um ambiente para computação de código aberto, que está disponível como *software* livre nos termos da Licença Pública Geral GNU do *Free Software Foundation* em código fonte. Ambas as linguagens funcionam em uma grande variedade de plataformas UNIX e sistemas similares (incluindo FreeBSD e Linux), Windows e MacOS, fornecendo uma grande variedade estatística (modelagem linear e não-linear, testes estatísticos clássicos, análise de séries temporais, classificação, agrupamento etc.), técnicas gráficas e é altamente extensível. Um dos pontos fortes destas linguagens é a facilidade de

interação com a comunidade existente na internet e a disponibilidade de pacotes (R- CORE TEAM, 2022 e PYTHON FOUNDATION, 2022).

3. RESULTADOS E DISCUSSÃO

3.1 Estatística Descritiva

3.1.1 Boxplot

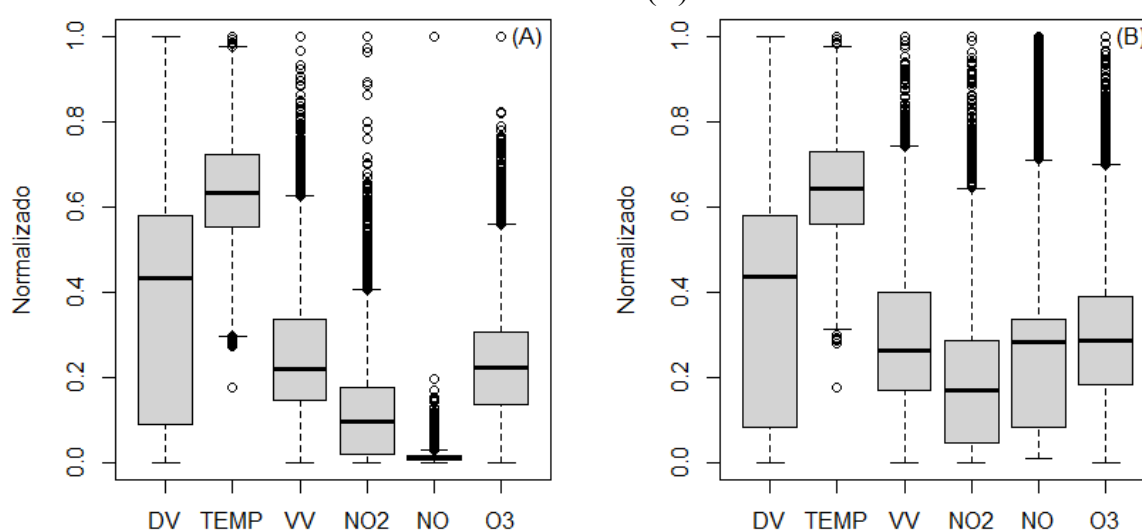
O *boxplot* foi utilizado para identificação e retirada de amostras anômalas do banco de dados de todas as EAQA envolvidas no estudo. Esta etapa é importante pois demonstram as propriedades estatísticas como: variabilidade (amplitude dos dados), média e *outliers*, este último estudo é de grande importância, pois existe a probabilidade destas amostras de não apresentar significado estatístico relevante para o modelo e provocar aumento do erro, já que estas amostras possivelmente não possuem informação adequada para estes bancos de dados. A retirada sucessiva de *outliers*, deve atender as resoluções da norma *American Society for Testing and Materials* (ASTM E1655-05, 2012), na qual a construção de um novo modelo, a partir da retirada de um conjunto de *outliers* de um modelo anterior, pode apresentar novos *outliers* não detectados no modelo anterior e isto pode ocorrer sucessivamente a cada novo modelo construído, sendo este comportamento classificado como efeito bola de neve (*snowballing effect*). Neste caso, a norma aconselha detectar e remover *outliers* somente até o segundo modelo. Para este trabalho cada EAQA foi cuidadosamente observada a quantidade de dados como possíveis *outliers*, pois dados meteorológicos e da qualidade do ar extraídos em ambientes externos, como é o caso das EAQA sofrem a interferências das intempéries naturais do ambiente.

Portanto, a retirada destes possíveis dados anômalos foi realizada uma única vez, através de estudos de *outliers* extremos, ou seja, foi identificado observações a qual possuem 3 vezes o intervalo interquartil (IQR- *interquartile range*). Mesmo identificando as amostras com 3 vezes o IQR, foi necessário observar minuciosamente se aquelas observações selecionadas, naquele específico período de tempo, foram realmente um caso atípico de ocorrer, ou não, pois a quantidade de observações para cada EAQA foi aproximadamente 16 mil, em outras palavras, se realmente aquelas amostras pudessem interferir nas características originais dos dados e representar alguma tendência para os modelos de previsão.

Todos os valores foram normalizados para melhor entendimento e comparação dos resultados obtidos pelo estudo dos *boxplots*. Somente as variáveis mais importantes foram selecionadas e estão apresentadas no *boxplot* para cada EAQA.

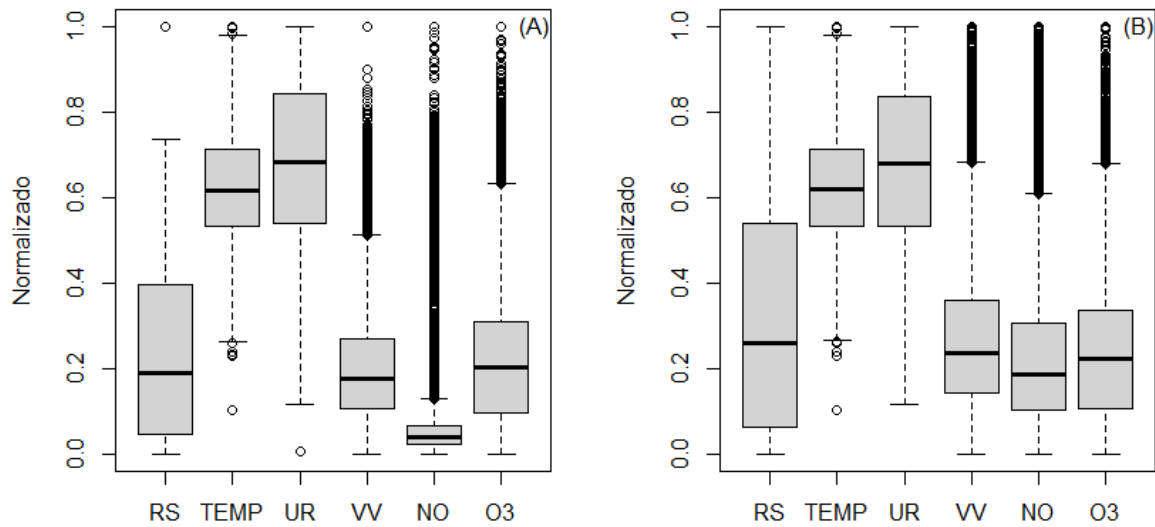
Para todas as EAQA (Figuras 11 a 14) é possível observar a variabilidade dos dados durante os 5 anos e para cada estação do ano (verão, outono, primavera e inverno), é comum as observações apresentarem dados com alta amplitude durante as estações do ano, fatores meteorológicos e ambientais (influência de proximidade de fontes fixas ou móveis).

Figura 11- Boxplot normalizado da EAQA de ADN: dados brutos com todas as observações, antes da retirada de pontos anômalos (A) e dados refinados após a retirada de pontos anômalos (B).



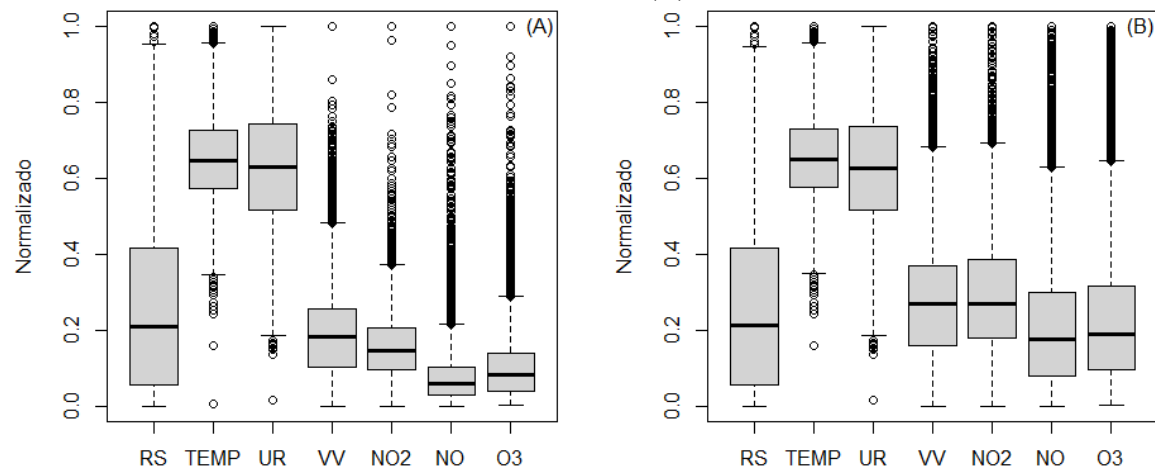
Fonte: O autor, 2022.

Figura 12- *Boxplot* normalizado da EAQA de PDC: dados brutos com todas as observações, antes da retirada de pontos anômalos (A) e dados refinados após a retirada de pontos anômalos (B).



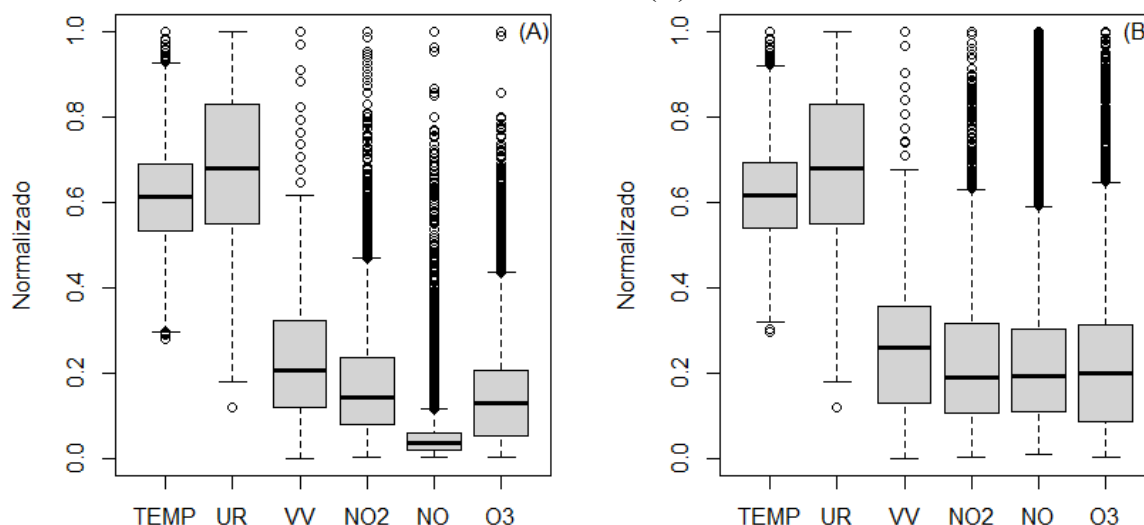
Fonte: O autor, 2022.

Figura 13- *Boxplot* normalizado da EAQA de VSL: dados brutos com todas as observações, antes da retirada de pontos anômalos (A) e dados refinados após a retirada de pontos anômalos (B).



Fonte: O autor, 2022.

Figura 14- *Boxplot* normalizado da EAQA de INE: dados brutos com todas as observações, antes da retirada de pontos anômalos (A) e dados refinados após a retirada de pontos anômalos (B).



Fonte: O autor, 2022.

Como pode ser observado nas Figuras 11 a 14, para a maioria das variáveis, uma única etapa de retirada de outliers não influenciou o perfil dos boxplots, com exceção para NO, que apresentava um boxplot com os quartis muito achatados, com um significativo número de outliers, que após a retirada, apresentou um boxplot mais compatível com as demais variáveis.

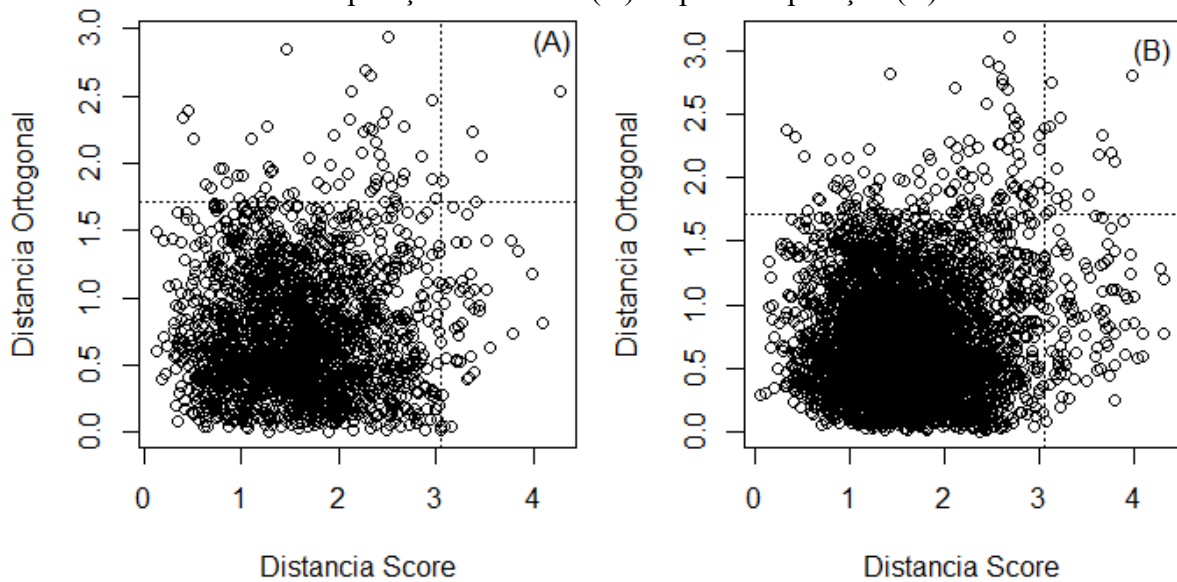
3.1.2 PCA Robusto (ROBPCA)

Um estudo de PCA robusto foi realizado para cada EAQA antes e após a imputação dos dados para verificar se a importância e as características estatísticas de cada variável foram mantidas. Em outras palavras, se não houveram mudanças significativas no número de componentes principais, autovalores e variância acumulativa explicada. O mesmo estudo foi realizado com o teste de correlação de Pearson e mostrou que as mesmas variáveis com as maiores correlações foram observadas antes e depois da imputação.

O ROBPCA foi utilizado como uma ferramenta para garantir e comprovar que o resultado obtido antes e após a imputação dos dados pelo algoritmo *MissForest* não foi alterado, ou seja, não alterou o comportamento e distribuição dos dados. Nas Figuras 15 a 30, pode ser observado que poucas observações apresentaram uma combinação de altos valores de distância ortogonal e distância entre os escores, ou seja, a minoria das observações está no quadrante

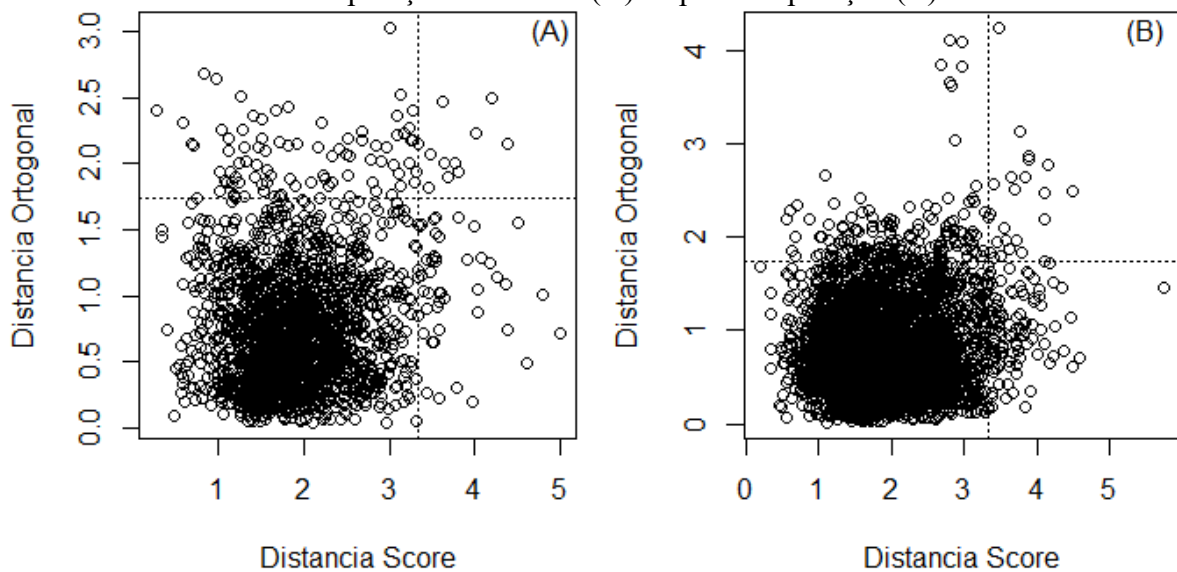
superior direito, contudo se considerar que os dados ambientais obtidos são do tipo aberto, sem nenhum controle, como experimentos realizados em laboratórios fechados, logo foi possível dizer que os resultados apresentados foram satisfatórios para todas as EAQA.

Figura 15- RBOPCA para a EAQA ADN de 2014-2018 para a estação do ano inverno: antes da imputação dos dados (A) e após a imputação (B)



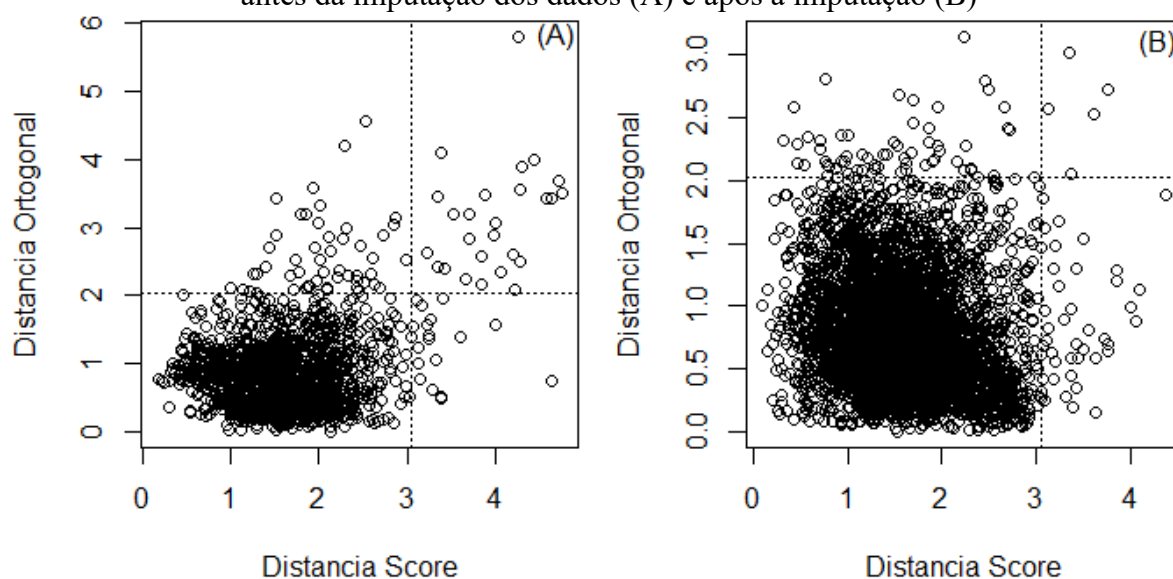
Fonte: O autor, 2022.

Figura 16- RBOPCA para a EAQA ADN de 2014-2018 para a estação do ano outono: antes da imputação dos dados (A) e após a imputação (B)



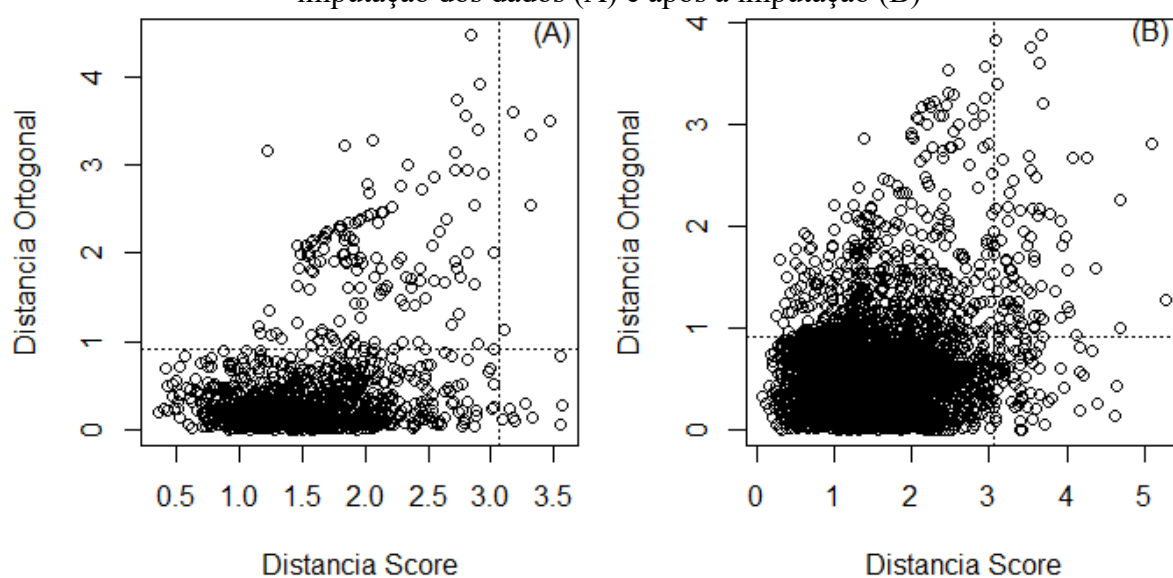
Fonte: O autor, 2022.

Figura 17- RBOPCA para a EAQA ADN de 2014-2018 para a estação do ano primavera: antes da imputação dos dados (A) e após a imputação (B)



Fonte: O autor, 2022.

Figura 18- RBOPCA para a EAQA ADN de 2014-2018 para a estação do ano verão: antes da imputação dos dados (A) e após a imputação (B)



Fonte: O autor, 2022.

A Tabela 5 mostra a variância explicada e os autovalores para cada modelo de RBOPCA aplicado para a EAQA de ADN para todos os anos compreendidos entre 2014 e 2018 e todas as estações do ano. Percebe-se com 3 a 4 números de componentes principais (PCs) foi possível capturar mais de 80 % de variância (foi utilizado o fator R de Wold para determinar o número de PCs). A variância capturada e os autovalores antes (com *not a number*-NA) e após a

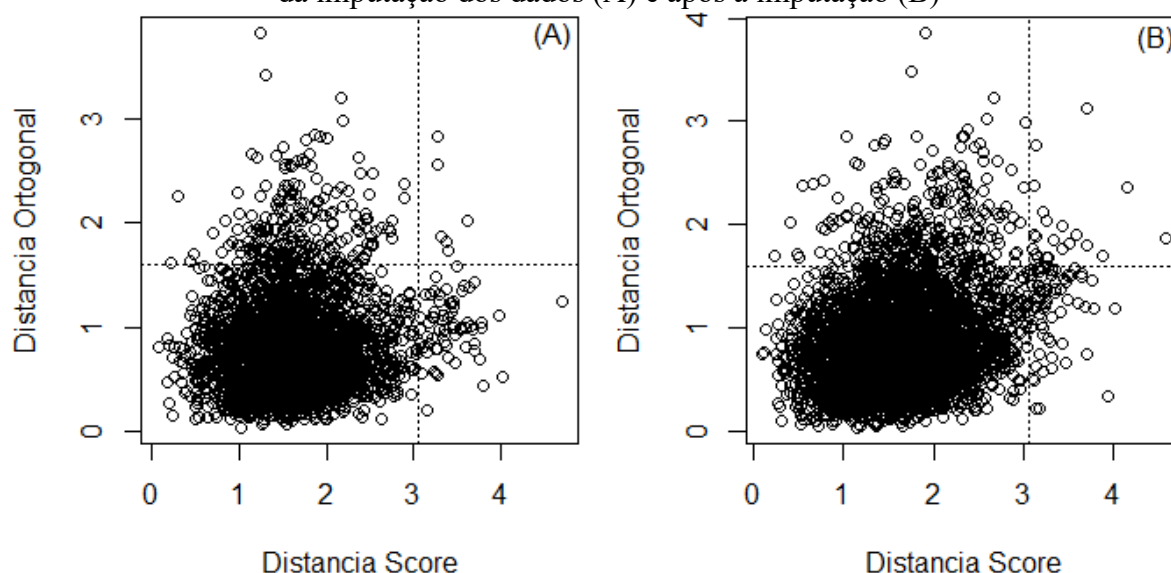
imputação (imputada) demonstra que não houve variação significativa dos valores encontrados em ambos os resultados, isto corrobora que a imputação dos dados não interferiu nas características originais dos dados.

Tabela 5- Comparação de autovalores e variância explicada para cada modelo RBOPCA a partir dos bancos de dados da EAQA ADN com NA e imputado

EAQA	Número de PCs	Variância capturada (%)	Autovalores
ADN_inv_NA	3	84,96	0,944
ADN_inv_imputada	3	86,29	0,923
ADN_ver_NA	3	88,10	0,501
ADN_ver_imputada	3	88,30	0,529
ADN_out_NA	4	86,90	0,636
ADN_out_imputada	4	85,90	0,713
ADN_pri_NA	3	80,30	0,705
ADN_pri_imputada	3	83,01	0,856

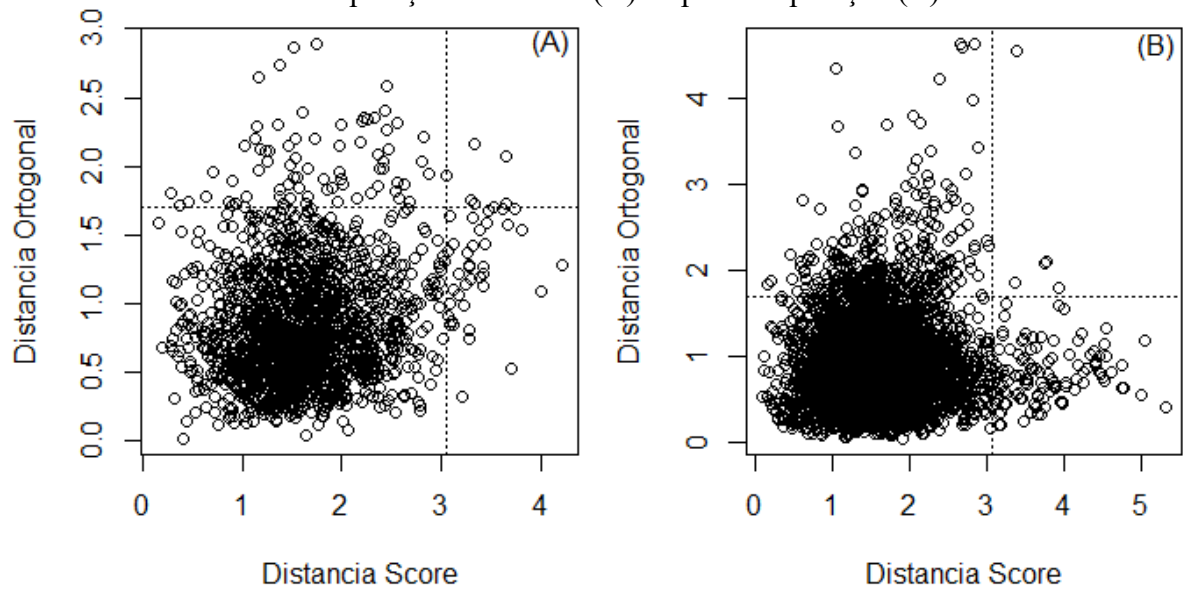
Fonte: O autor, 2022.

Figura 19- RBOPCA para a EAQA PDC de 2014-2018 para a estação do ano inverno: antes da imputação dos dados (A) e após a imputação (B)



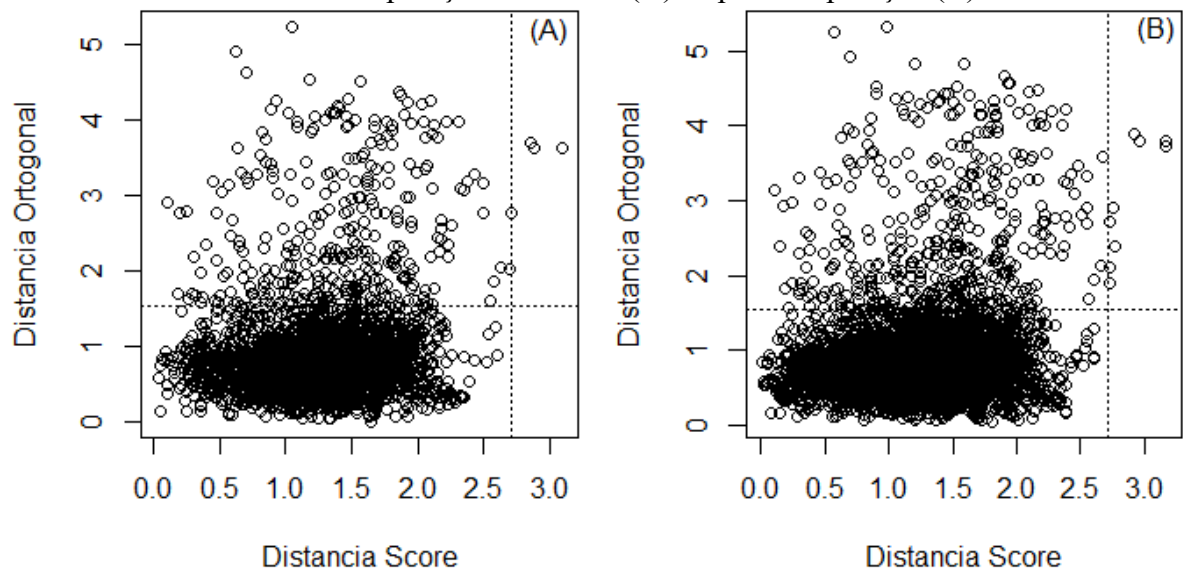
Fonte: O autor, 2022.

Figura 20- RBOPCA para a EAQA PDC de 2014-2018 para a estação do ano outono: antes da imputação dos dados (A) e após a imputação (B)



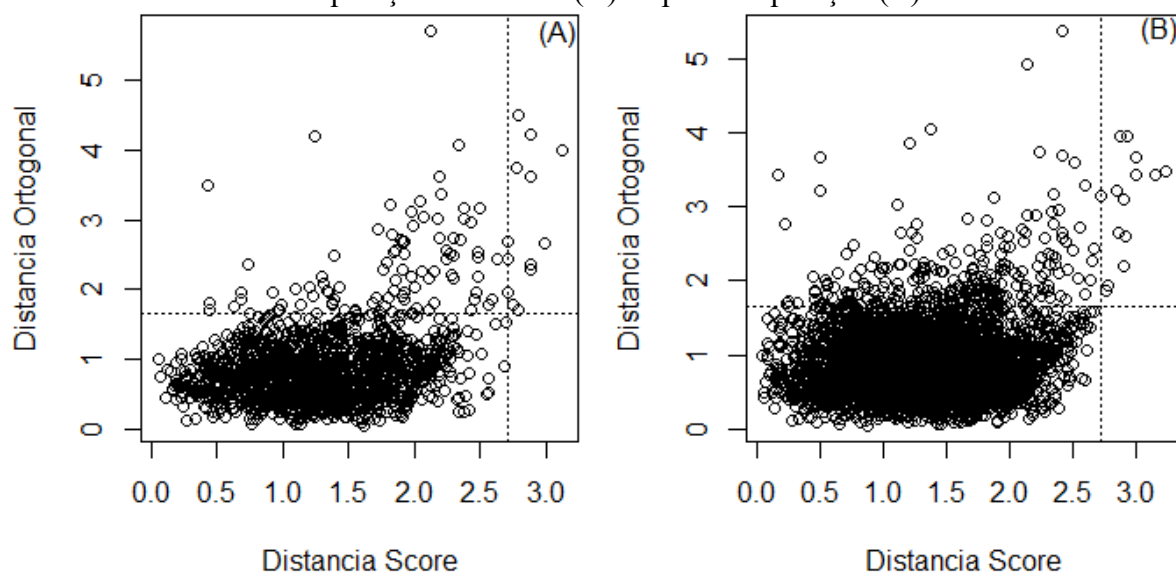
Fonte: O autor, 2022.

Figura 21- RBOPCA para a EAQA PDC de 2014-2018 para a estação do ano primavera: antes da imputação dos dados (A) e após a imputação (B)



Fonte: O autor, 2022.

Figura 22- RBOPCA para a EAQA PDC de 2014-2018 para a estação do ano verão: antes da imputação dos dados (A) e após a imputação (B)



Fonte: O autor, 2022.

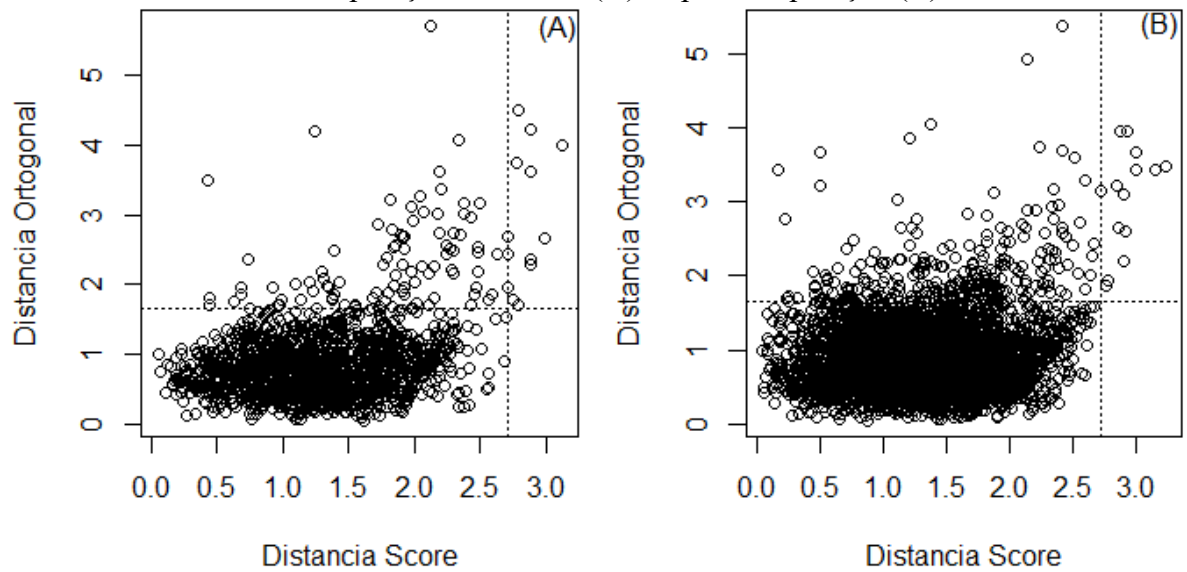
A Tabela 6 mostra a variância explicada e os autovalores para cada modelo de RBOPCA aplicado para a EAQA de PDC para todos os anos compreendidos entre 2014 e 2018 e todas as estações do ano. Percebe-se com 3 números de componentes principais (PCs) foi possível capturar mais de 83 % de variância. A variância capturada e os autovalores antes (com *not a number*-NA) e após a imputação (imputada) demonstra que não houve variação significativa dos valores encontrados em ambos os resultados, isto corrobora que a imputação dos dados não interferiu nas características originais dos dados.

Tabela 6- Comparação de autovalores e variância explicada para cada modelo RBOPCA a partir dos bancos de dados da EAQA PDC com NA e imputado

EAQA	Número de PCs	Variância capturada (%)	Autovalores
PDC_inv_NA	3	85,06	0,691
PDC_inv_imputada	3	84,83	0,655
PDC_ver_NA	3	94,66	0,700
PDC_ver_imputada	3	94,80	0,748
PDC_pri_NA	3	91,96	0,819
PDC_pri_imputada	3	92,11	0,797
PDC_out_NA	3	85,08	0,783
PDC_out_imputada	3	83,29	0,721

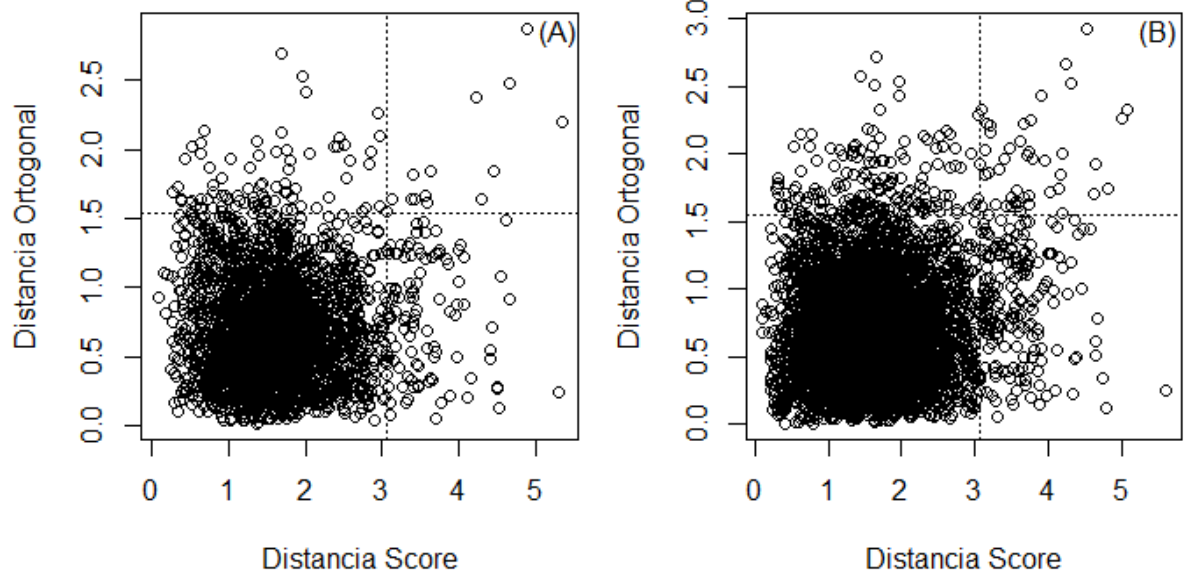
Fonte: O autor, 2022.

Figura 23- RBOPCA para a EAQA VSL de 2014-2018 para a estação do ano inverno: antes da imputação dos dados (A) e após a imputação (B)



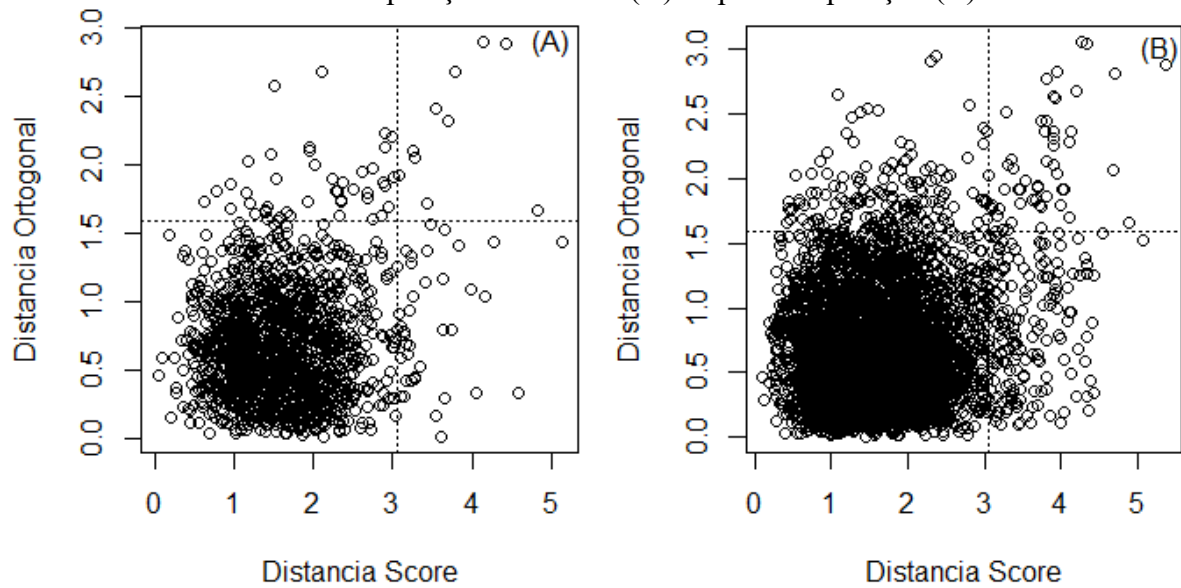
Fonte: O autor, 2022.

Figura 24- RBOPCA para a EAQA VSL de 2014-2018 para a estação do ano outono: antes da imputação dos dados (A) e após a imputação (B)



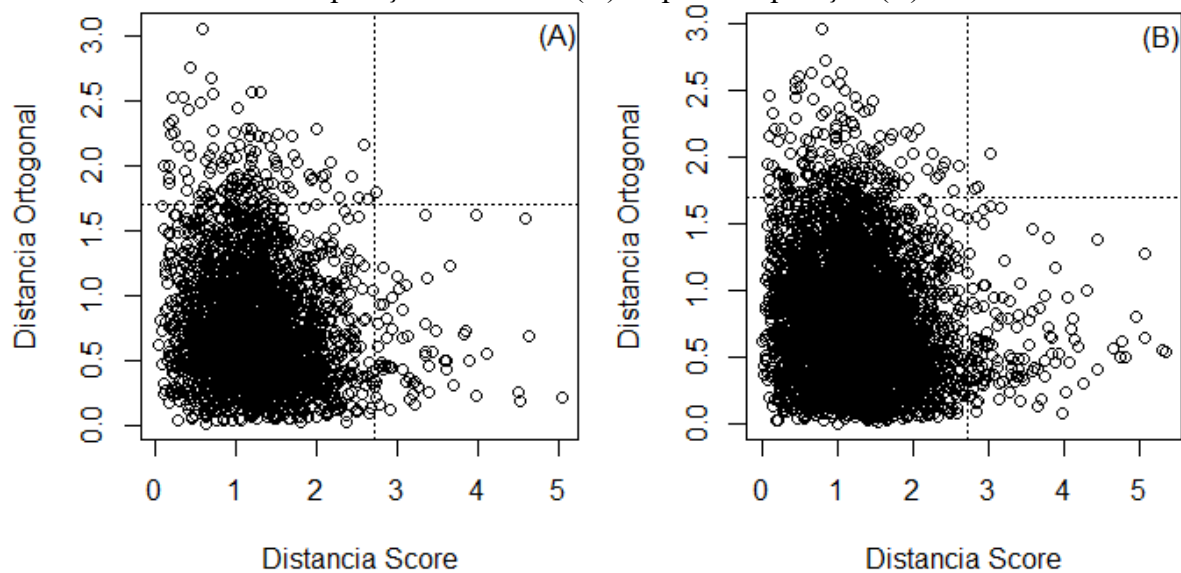
Fonte: O autor, 2022.

Figura 25- RBOPCA para a EAQA VSL de 2014-2018 para a estação do ano primavera: antes da imputação dos dados (A) e após a imputação (B)



Fonte: O autor, 2022.

Figura 26- RBOPCA para a EAQA VSL de 2014-2018 para a estação do ano verão: antes da imputação dos dados (A) e após a imputação (B)



Fonte: O autor, 2022.

A Tabela 7 mostra a variância explicada e os autovalores para cada modelo de RBOPCA aplicado para a EAQA de VSL para todos os anos compreendidos entre 2014 e 2018 e todas as estações do ano. Percebe-se com 3 números de componentes principais (PCs) foi possível capturar mais de 86 % de variância. A variância capturada e os autovalores antes (com *not a number*-NA) e após a imputação (imputada) demonstra que não houve variação significativa

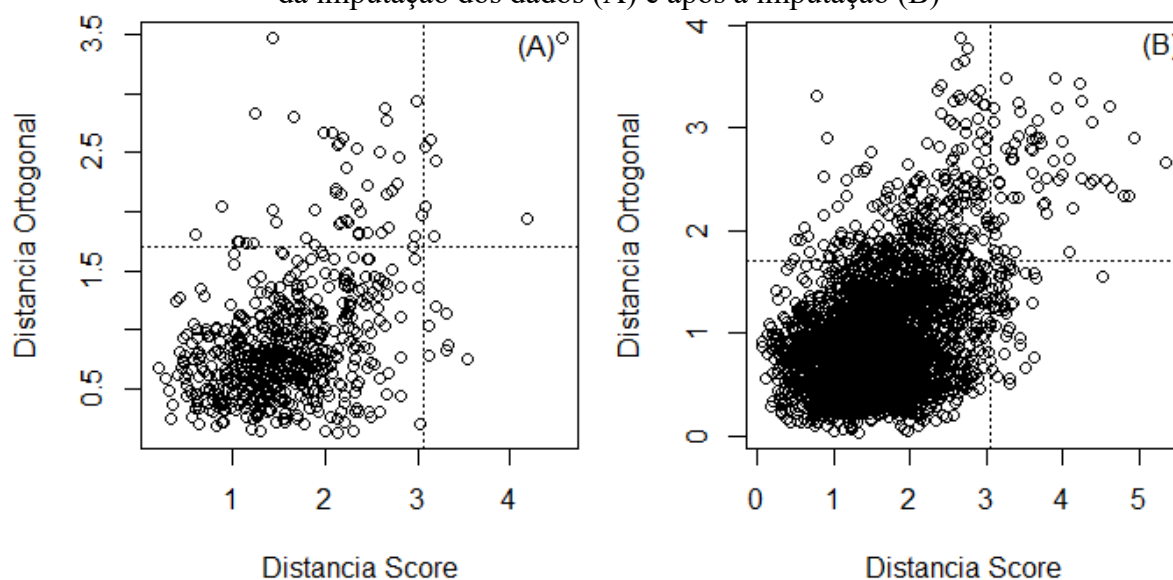
dos valores encontrados em ambos os resultados, isto corrobora que a imputação dos dados não interferiu nas características originais dos dados.

Tabela 7- Comparação de autovalores e variância explicada para cada modelo RBOPCA a partir dos bancos de dados da EAQA VSL com NA e imputado

EAQA	Número de PCs	Variância capturada (%)	Autovalores
VSL_inv_NA	3	87,21	0,800
VSL_inv_imputada	3	86,90	0,768
VSL_ver_NA	3	90,29	0,639
VSL_ver_imputada	3	91,79	0,612
VSL_pri_NA	3	94,24	0,578
VSL_pri_imputada	3	94,41	0,556
VSL_out_NA	3	87,72	0,810
VSL_out_imputada	3	86,77	0,836

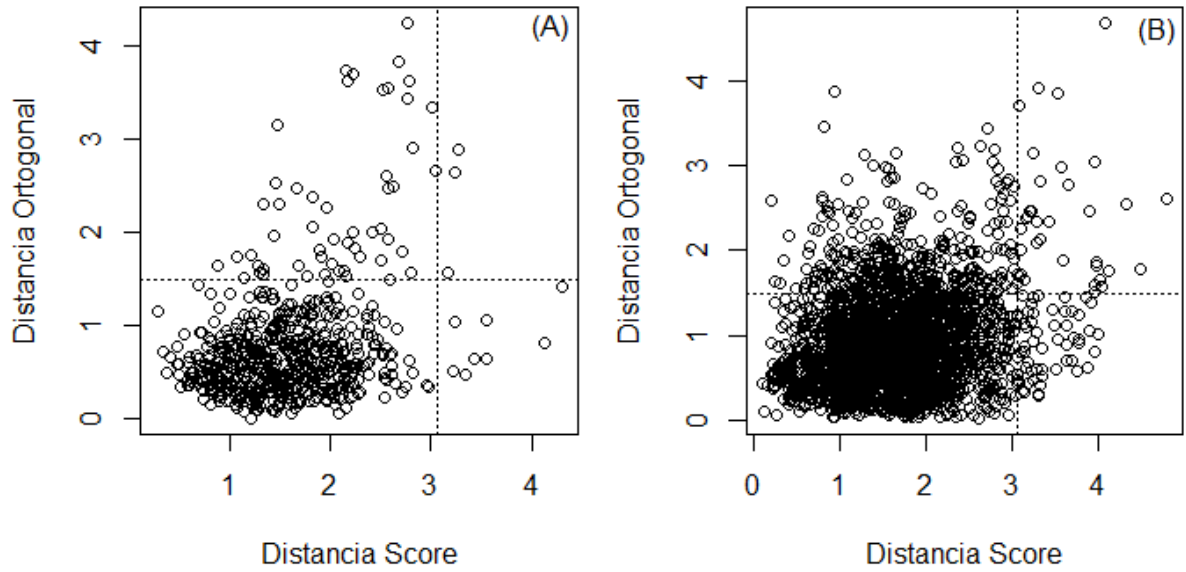
Fonte: O autor, 2022.

Figura 27- RBOPCA para a EAQA INE de 2014-2018 para a estação do ano inverno: antes da imputação dos dados (A) e após a imputação (B)



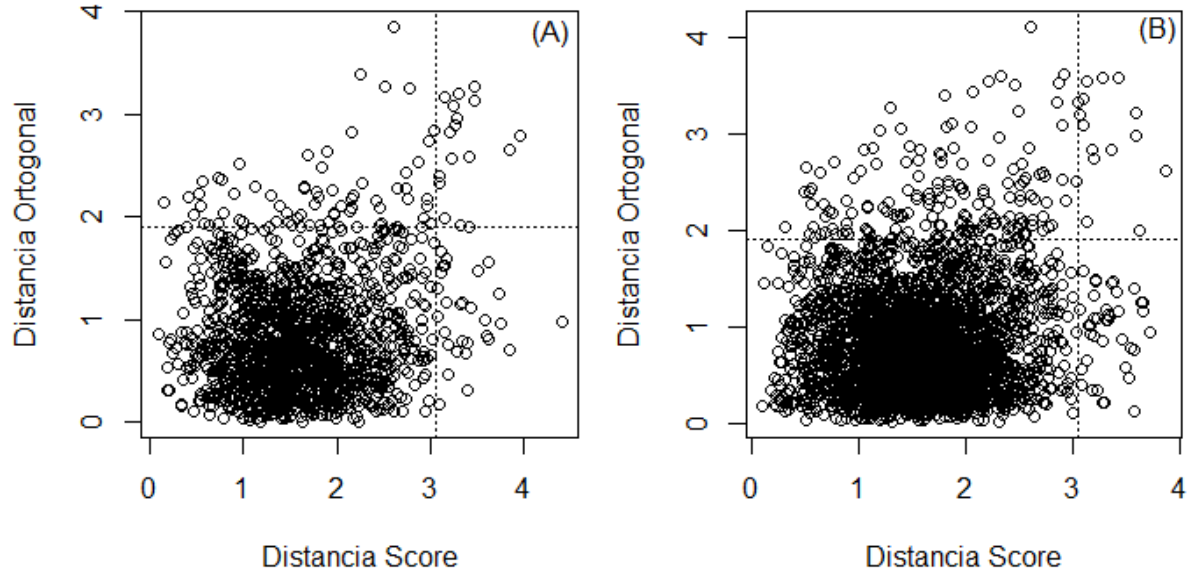
Fonte: O autor, 2022.

Figura 28- RBOPCA para a EAQA INE de 2014-2018 para a estação do ano outono: antes da imputação dos dados (A) e após a imputação (B)



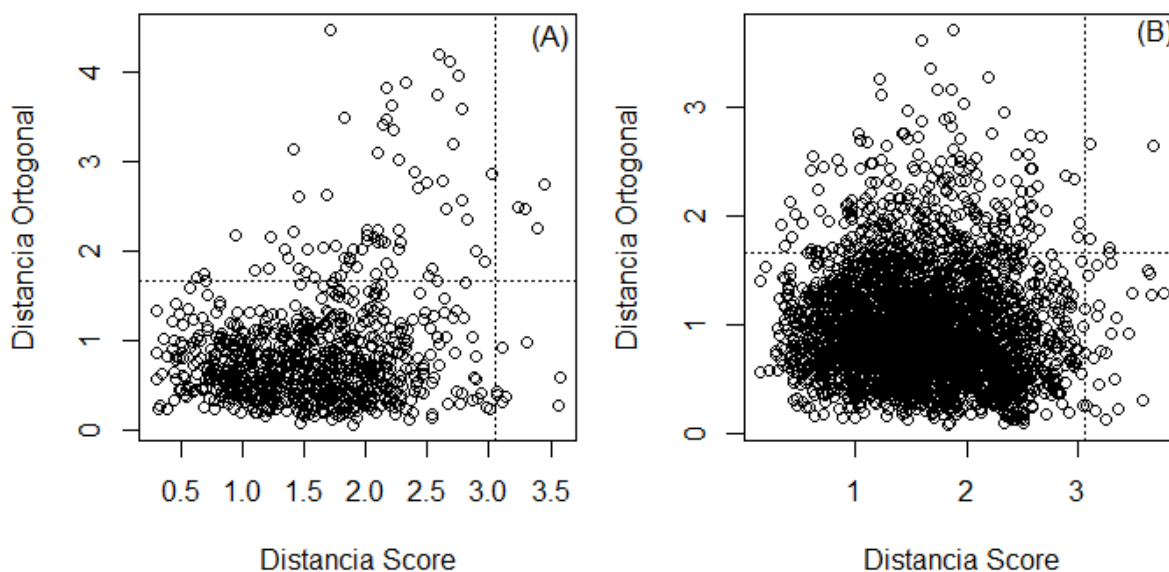
Fonte: O autor, 2022.

Figura 29- RBOPCA para a EAQA INE de 2014-2018 para a estação do ano primavera: antes da imputação dos dados (A) e após a imputação (B)



Fonte: O autor, 2022.

Figura 30- RBOPCA para a EAQA INE de 2014-2018 para a estação do ano verão: antes da imputação dos dados (A) e após a imputação (B)



Fonte: O autor, 2022.

A Tabela 8 mostra a variância explicada e os autovalores para cada modelo de RBOPCA aplicado para a EAQA de INE para todos os anos compreendidos entre 2014 e 2016 e todas as estações do ano. Percebe-se com 3 números de componentes principais (PCs) foi possível capturar mais de 80 % de variância. A variância capturada e os autovalores antes (com *not a number* - NA) e após a imputação (imputada) demonstra que não houve variação significativa dos valores encontrados em ambos os resultados, isto corrobora que a imputação dos dados não interferiu nas características originais dos dados.

Tabela 8- Comparação de autovalores e variância explicada para cada modelo RBOPCA a partir dos bancos de dados da EAQA INE com NA e imputado

EAQA	Número de PCs	Variância capturada (%)	Autovalores
INE_inv_NA	3	82,38	0,930
INE_inv_imputada	3	80,20	0,873
INE_ver_NA	3	83,27	0,655
INE_ver_imputada	3	80,24	0,687
INE_pri_NA	3	80,64	0,808
INE_pri_imputada	3	80,85	0,786
INE_out_NA	3	83,37	0,808
INE_out_imputada	3	82,56	0,739

Fonte: O autor, 2022.

3.1.3 Estudo de correlação de Pearson

A Tabela 9 mostra apenas as variáveis elegíveis para o estudo com correlação de Pearson com ozônio, acima de 0,20, valor escolhido para evitar carregar todas as variáveis que não tinham importância para o modelo e assim reduzir o tempo computacional. A matriz de correlação de Pearson revelou que as variáveis VV, RS, T e HORA em todas as EAQA apresentam níveis pequenos a moderados de correlação positiva com o ozônio, enquanto NO, NO₂, UR têm uma correlação negativa com o ozônio. Uma explicação para as diferentes correlações entre RS e T com o ozônio, mostradas pelo conjunto de dados, é que durante as estações de inverno, a intensidade da radiação solar é reduzida, pois a luz solar atinge a superfície da Terra com uma inclinação menor do que nas outras estações do ano. As correlações negativas com a UR podem estar associados à instabilidade atmosférica e grande cobertura de nuvens, o que pode diminuir os processos fotoquímicos na troposfera, com o O₃ sendo este consumido pela deposição úmida (NISHANTH *et al.*, 2012). Song *et al.* (2011) relataram resultados semelhantes, que uma alta umidade relativa na atmosfera pode diminuir a produção de O₃ até certo ponto. A variável VV (desde que os ventos não sejam muito fortes) tem uma correlação positiva com o O₃, isto pode estar ligado ao transporte de poluentes precursores de O₃ como NO e NO₂. O aumento da dispersão e a mistura desses poluentes atmosféricos emitidos em fontes mais próximas (ex: rodovias e fontes estacionárias), pode otimizar a formação de O₃ a partir destes precursores. Estes resultados corroboram com os trabalhos de Jones *et al.* (2010) e Guerra *et al.* (2011).

As concentrações de ozônio foram correlacionadas negativamente com NO e NO₂ como esperado, uma vez que esses poluentes são conhecidos precursores do ozônio, o que sugere que um aumento na concentração de ozônio segue uma queda nos níveis dessas variáveis. Outro sinal disso é a correlação negativa com o NO que é esperada pela reação entre O₃ e NO, ao formar NO₂ e O₂ (LUNA *et al.*, 2014).

Tabela 9- Resumo das correlações entre as variáveis de entrada e concentração de ozônio para todas as EAQA, variáveis antes (preto) e após (azul-vermelho) imputação.

EAQA	T		HORA		NO		NO ₂		VV		UR		RS		DV	
ADN_inv	0,65	0,65	0,56	0,55	-0,39	-0,40	-0,33	-0,34	0,25	0,25	-	-	-	-	-	-
ADN_pri	0,56	0,56	0,48	0,48	-0,44	-0,36	-0,31	-0,21	0,27	0,26	-	-	-	-	-	-

EAQA	T		HORA		NO		NO ₂		VV		UR		RS		DV	
ADN_ver	0,62	0,62	0,57	0,57	-0,30	-0,39	-0,23	-0,30	-	-	-	-	-	-	-	-
ADN_out	0,56	0,55	0,58	0,58	-0,33	-0,31	-0,20	-0,20	0,25	0,23	-	-	-	-	0,23	0,22
PDC_inv	0,63	0,63	0,40	0,40	-0,36	-0,34	-	-	0,21	0,21	-0,58	-0,58	0,36	0,36	-	-
PDC_pri	0,56	0,56	0,32	0,31	-0,20	-0,19	-	-	-	-	-0,52	-0,52	0,33	0,34	-	-
PDC_ver	0,45	0,47	0,36	0,35	-0,41	-0,43	-	-	-	-	-0,43	-0,44	0,32	0,32	-	-
PDC_out	0,51	0,51	0,42	0,42	-0,56	-0,53	-	-	0,22	0,22	-0,53	-0,52	0,33	0,33	-	-
VSL_inv	0,54	0,53	-	-	-0,52	-0,52	-	-	0,27	0,26	-0,52	-0,49	0,49	0,48	-	-
VSL_pri	0,46	0,43	-	-	-0,41	-0,42	-	-	-	-	-0,44	-0,39	0,38	0,38	-	-
VSL_ver	0,52	0,52	-	-	-0,50	-0,50	0,24	0,24	-	-	-0,47	-0,44	0,56	0,56	-	-
VSL_out	0,58	0,57	-	-	-0,49	-0,48	-	-	0,27	0,32	-0,53	-0,50	0,48	0,46	-	-
INE_inv	0,44	0,39	0,44	0,42	-0,53	-0,47	-0,35	-0,34	0,42	0,39	-0,55	-0,58	-	-	-	-
INE_pri	0,55	0,57	0,36	0,36	-0,65	-0,60	-0,65	-0,61	0,35	0,35	-	-	-	-	-	-
INE_ver	0,50	0,56	0,46	0,45	-0,23	-0,32	0,21	0,20	0,31	0,32	-0,36	-0,38	-	-	-	-
INE_out	0,41	0,42	0,46	0,45	-0,41	-0,34	-	-	0,30	0,24	-0,25	-0,27	-	-	-	-

Fonte: O autor, 2022.

As variáveis utilizadas nesta tese foram selecionadas levando-se em consideração: (i) os coeficientes de correlação de Pearson (conforme apresentados na Tabela 7); (ii) a produção fotoquímica, através das variáveis: T, HORA, NO, NO₂, RS, UR e (iii) o transporte de O₃, através das variáveis: VV e DV. Os resultados mostraram que não foram obtidas melhorias nos coeficientes quando as concentrações das variáveis CO e MP₁₀ foram incluídas no modelo (SOUSA et al., 2006). Um estudo de correlação de Pearson com e sem (Tabela 9) imputação foi realizado para estimar se a imputação alterou as características originais dos dados e ajudou a melhorar os modelos de previsão. A maioria dos casos mostrou uma correlação um pouco maior (cor azul) após a imputação de dados, já na minoria dos casos houve uma correlação negativa (cor vermelha), em ambos os casos não houve uma alteração significativa nas correlações antes e após a imputação dos dados, ou seja, o algoritmo *MissForest* foi capaz de imputar valores sem alterar as correlações entre a variável dependente e as variáveis independentes. O objetivo deste estudo foi compreender se existe a influência dos novos dados imputados, neste estudo que abrangeu todas as EAQA e estações do ano.

3.1.4 Dados faltantes (data missing)

A Tabela 10 mostra a quantidade percentual de dados faltantes para cada EAQA. Percebe-se que no caso, todas as variáveis dentro do conjunto de dados, o número de valores faltantes não é significativo, em particular para o ozônio. A EAQA INE é a que apresentou o

maior número de valores faltantes, pois esta estação registrou somente dados entre os anos de 2014 e 2016. Em relação a quantidade de dados faltantes para cada EAQA, isso pode ser explicado por alguns fatores como: a) quando os critérios de validação não foram alcançados, b) falta de manutenção e ou c) outro problema desconhecido.

Tabela 10- Porcentagem de dados faltantes para cada EAQA analisada

Variáveis	O ₃	T	NO	DV	NO ₂	VV	UR	RS
ADN_inv	4,70	3,77	4,39	-	4,72	5,36	-	-
ADN_pri	6,72	9,82	23,19	-	25,50	1,69	-	-
ADN_ver	6,71	3,38	22,37	-	23,59	-	-	-
ADN_out	6,62	1,76	8,56	3,76	9,01	3,83	-	-
PDC_inv	6,61	4,39	16,47	-	-	4,83	4,46	4,37
PDC_pri	14,54	10,39	14,00	-	-	-	10,51	10,41
PDC_ver	12,06	4,44	24,33	-	-	-	3,80	5,03
PDC_out	15,51	4,78	28,56	-	-	6,19	4,74	5,72
VSL_inv	7,23	8,19	6,46	-	-	4,48	24,78	6,64
VSL_pri	4,68	13,72	4,47	-	-	-	29,86	3,97
VSL_ver	17,33	13,61	23,09	-	22,76	-	21,85	13,54
VSL_out	16,56	14,55	15,38	-	-	21,72	31,55	19,71
INE_inv	13,37	12,36	45,26	-	45,47	6,32	48,78	-
INE_pri	21,11	10,64	44,08	-	44,08	4,78	-	-
INE_ver	16,24	11,31	59,03	-	59,24	6,96	32,33	-
INE_out	36,84	34,36	52,09	-	-	43,75	52,74	-

Fonte: O autor, 2022.

3.2 Imputação dos Dados (*Missforest*)

O algoritmo *MissForest* foi capaz de imputar valores em um conjunto de dados, conforme Tabela 11. Esse tempo de imputação foi de 3 segundos para cada iteração, em um total de 10 e foi utilizada a média destas 10 rodadas para a realização da imputação (usando um processador i7-6500U, 2 núcleos, 2592 MHz e 8 GB DDR4 2400 MHz). É esperado que quanto maior a quantidade de dados faltantes, maior o erro associado à imputação, visto que o algoritmo *MissForest* tenta capturar todas as características associadas entre as variáveis.

Todas as EAQA foram agrupadas em estações do ano para os anos de 2014 a 2018 para as EAQA – ADN, VSL, PDC e 2014 a 2016 para a EAQA INE. Com isso as características entre as variáveis dentro de cada uma das estações do ano foram preservadas.

Tabela 11- Número de observações, variáveis e *normalize root mean square error* (NRMSE) da imputação para cada EAQA escolhida para o estudo

EAQA	Número de observações	Variáveis	Média_NRMSE da imputação
ADN_inv	4531	T, VV, NO ₂ , NO, HORA	0,12
ADN_pri	4451	T, VV, NO ₂ , NO, HORA	0,10
ADN_ver	5274	T, NO ₂ , NO, HORA	0,13
ADN_out	4650	DV, T, VV, NO ₂ , NO, HORA	0,10
PDC_inv	5131	RS, T, UR, VV, NO, HORA	0,16
PDC_pri	4843	RS, T, UR, NO, HORA	0,16
PDC_ver	4736	RS, T, UR, NO, HORA	0,14
PDC_out	5111	RS, T, UR, VV, NO, HORA	0,16
VSL_inv	5229	RS, T, VV, NO, UR	0,12
VSL_pri	5211	RS, T, UR, NO	0,11
VSL_ver	5171	RS, T, UR, NO ₂ , NO	0,12
VSL_out	5170	RS, T, VV, NO, UR	0,12
INE_inv	3292	T, UR, VV, NO ₂ , NO, HORA	0,31
INE_pri	3140	T, VV, NO ₂ , NO, HORA	0,27
INE_ver	3267	T, UR, VV, NO ₂ , NO, HORA	0,26
INE_out	3271	T, UR, VV, NO, HORA	0,29

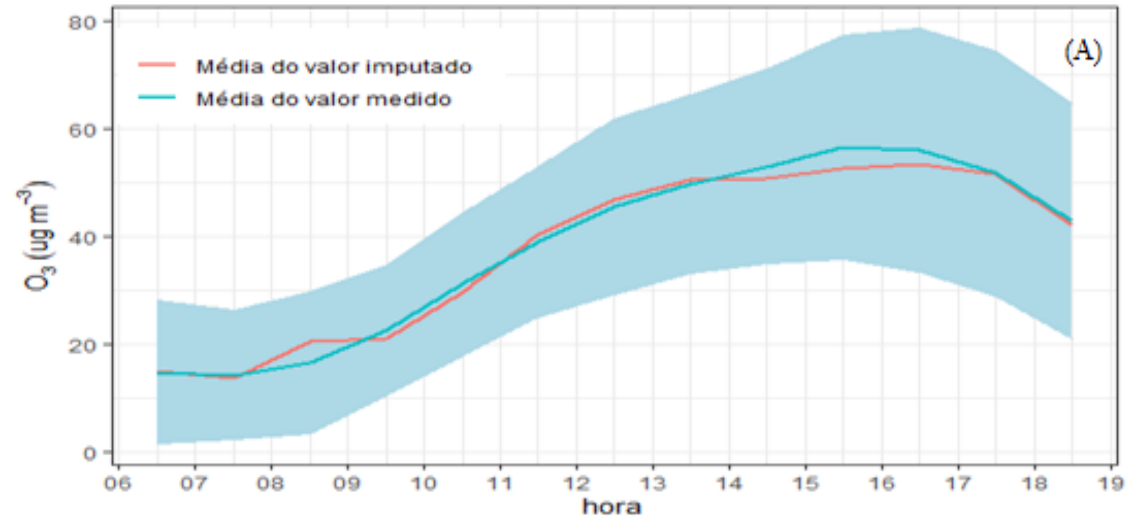
Fonte: O autor, 2022.

3.3 Comparação dos dados imputados e medidos

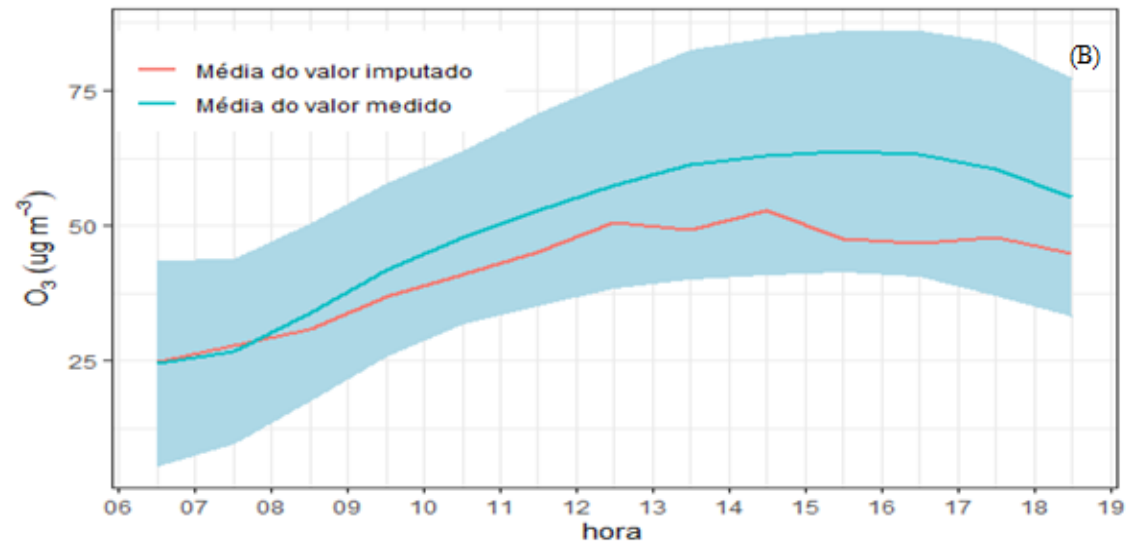
Uma visualização mais objetiva entre os dados imputados e os valores medidos podem ser observados nas Figuras 31, 32, 33 e 34. A faixa azul clara representa o desvio padrão dos valores medidos para cada horário, compreendidos entre 6:30 h até 18:30 h para cada EAQA e cada estação do ano (verão, outono, primavera e inverno). A linha azul mais escura representa a média do valor medido e a linha vermelha indica a média do valor imputado.

Todas as EAQA apresentadas mostram que os valores imputados estão dentro do desvio médio dos valores medidos, com exceção da EAQA do INE_ver (Figura 34 c) foi a única a apresentar os valores médios imputados em determinadas horas do dia maior que o desvio padrão. Este evento pode estar relacionado com a maior quantidade de dados faltantes principalmente da variável NO₂ (59,24% de acordo com a Tabela 10) e como pode ser observado na Figura 54, a classificação Boruta mostra que esta variável é mais importante em relação as demais, fazendo com que a imputação seja comprometida.

Figura 31 - Comparação entre os valores médios medidos e os valores médios imputados para ADN_out (A), ADN_pri (B), ADN_ver (C), ADN_inv (D)



Fonte: O autor, 2022.



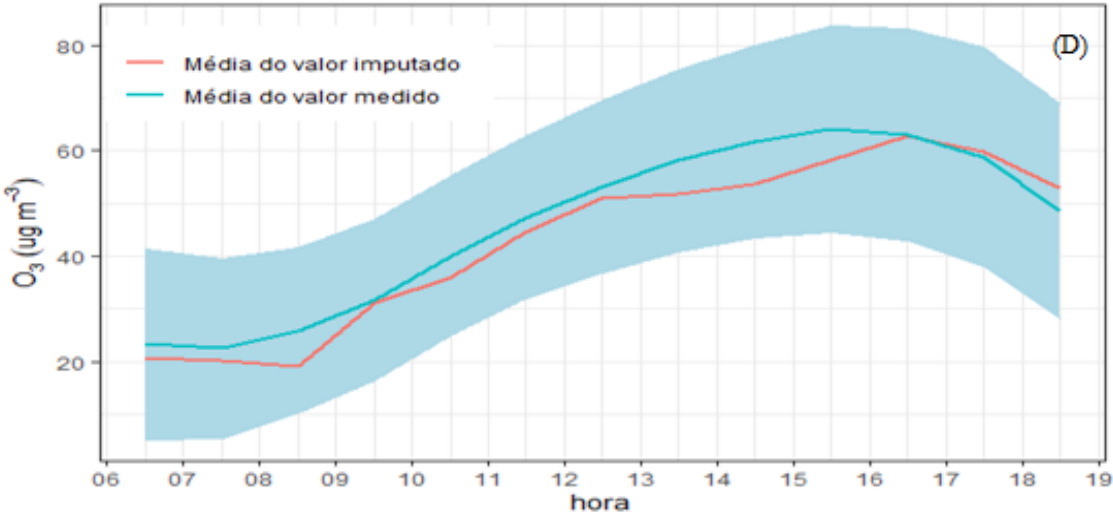
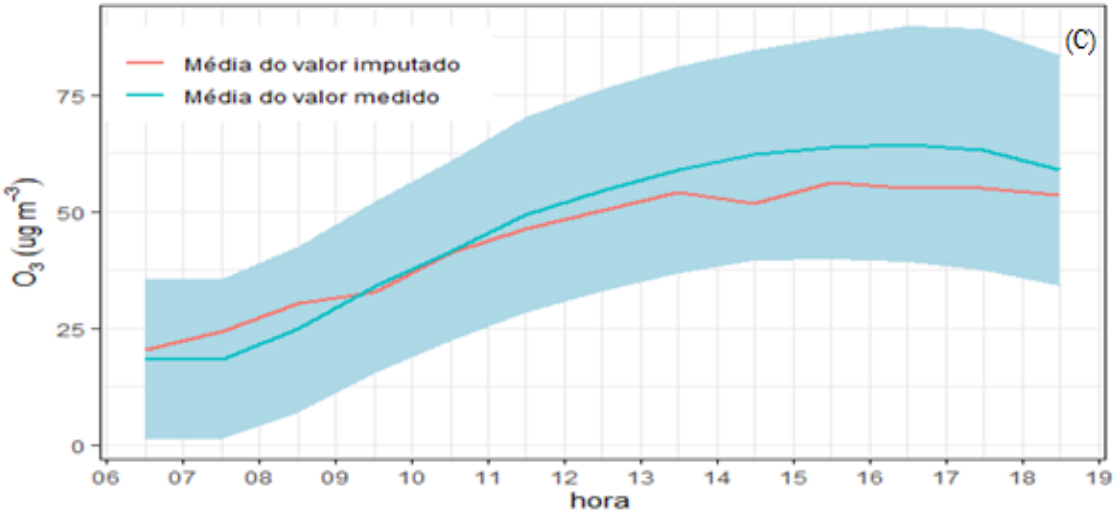
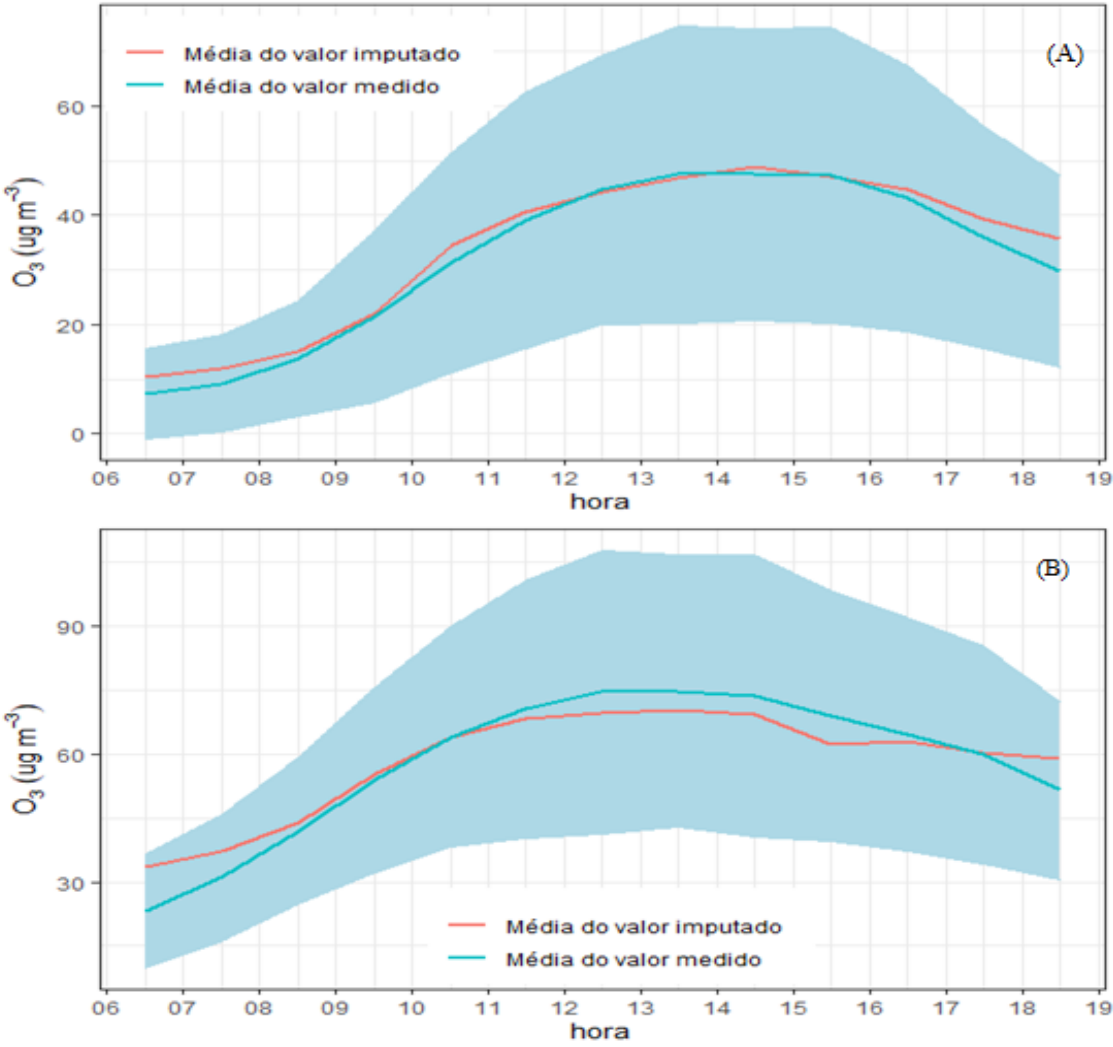


Figura 32- Comparação entre os valores médios medidos e os valores médios imputados para PDC_out (A), PDC_pri (B), PDC_ver (C), PDC_inv (D)



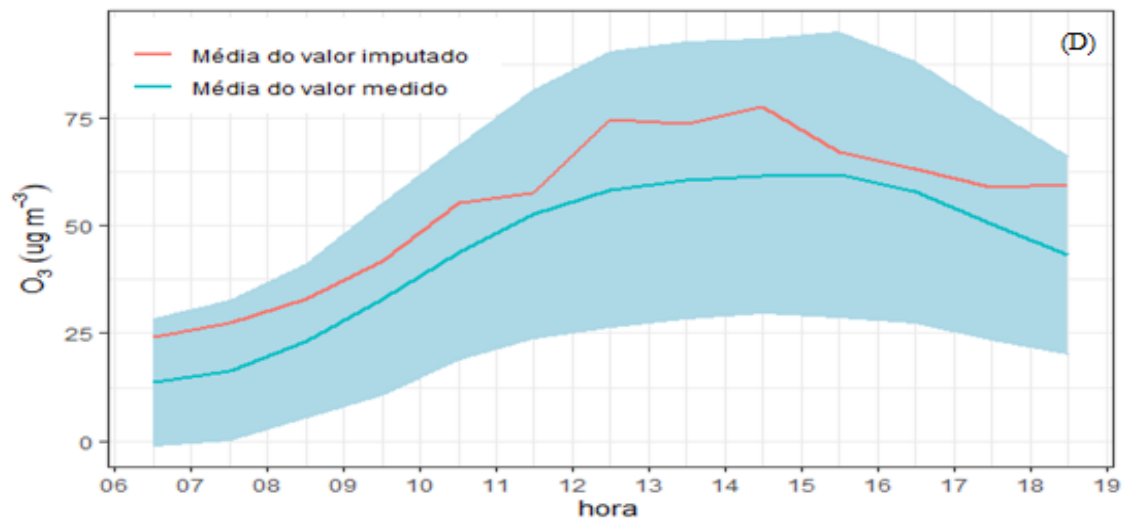
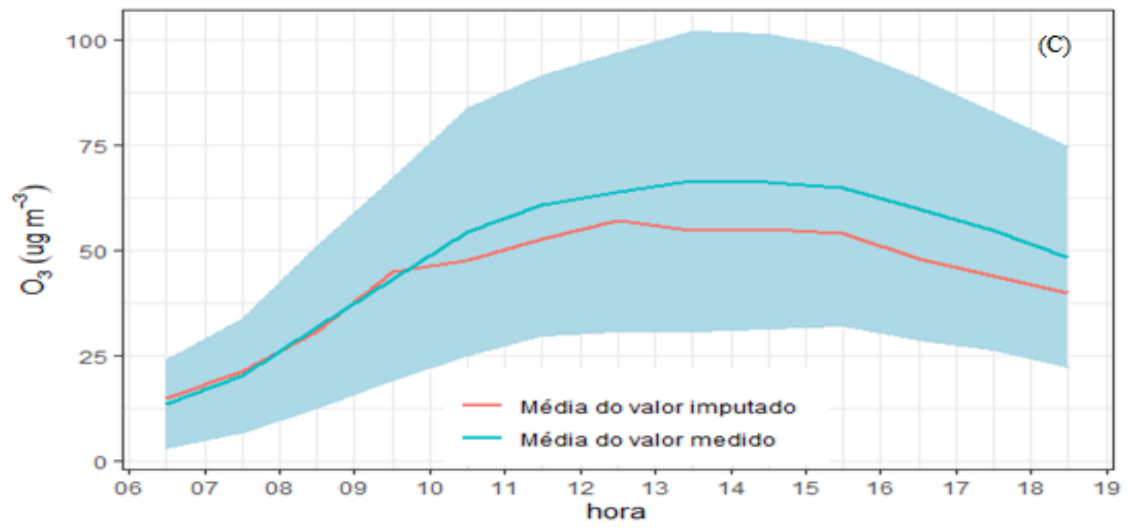
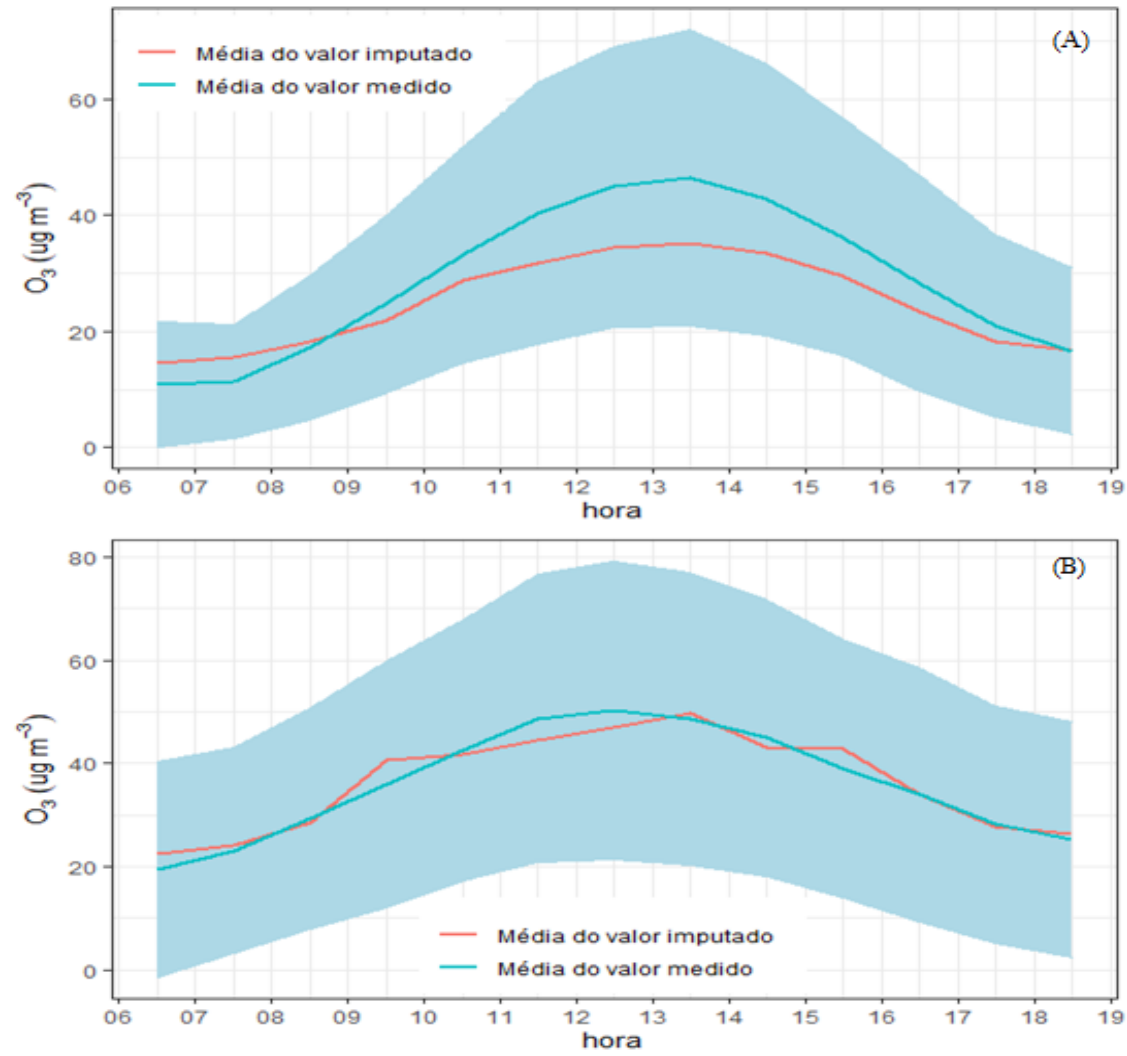


Figura 33- Comparação entre os valores médios medidos e os valores médios imputados para VSL_out (A), VSL_pri (B), VSL_ver (C), VSL_inv (D)



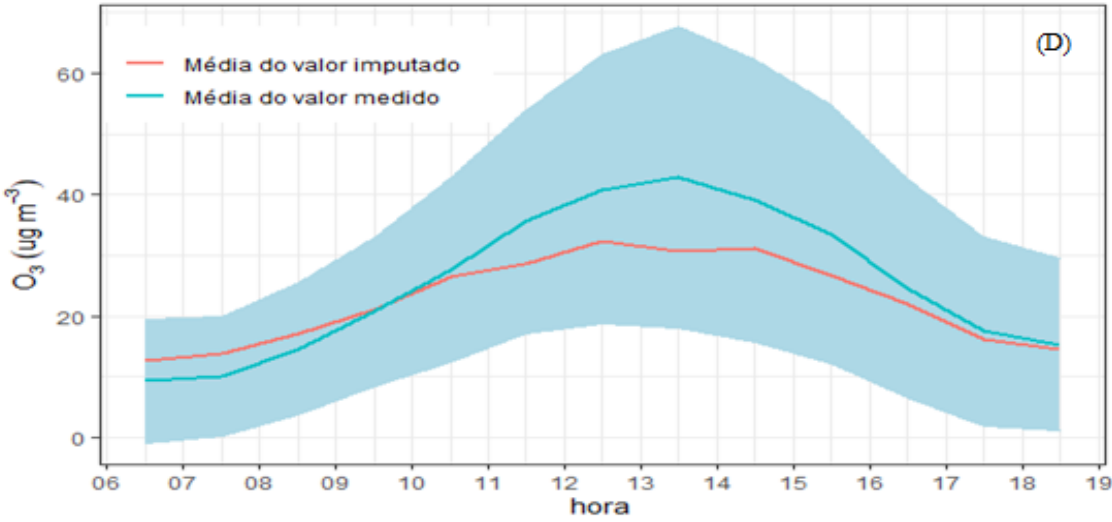
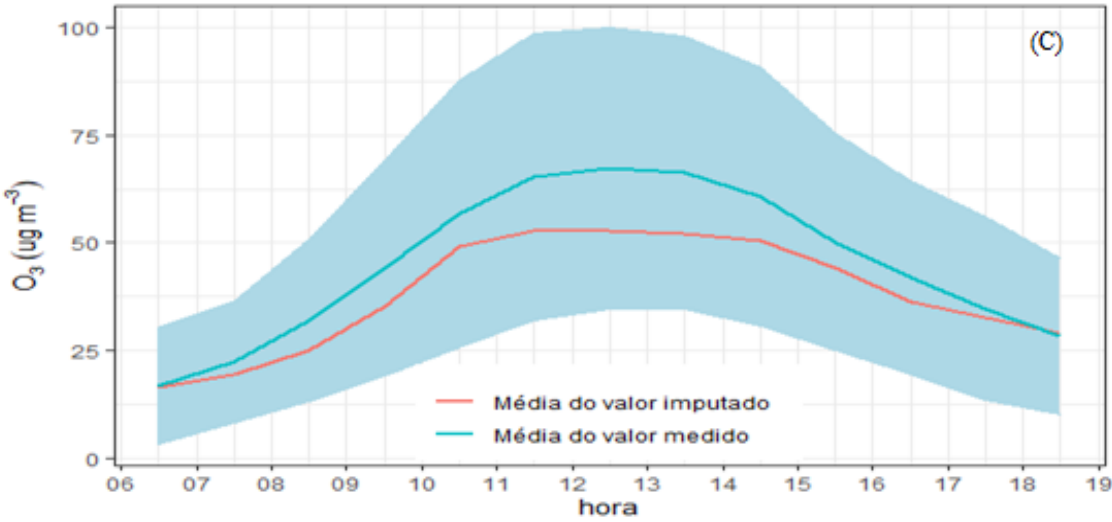
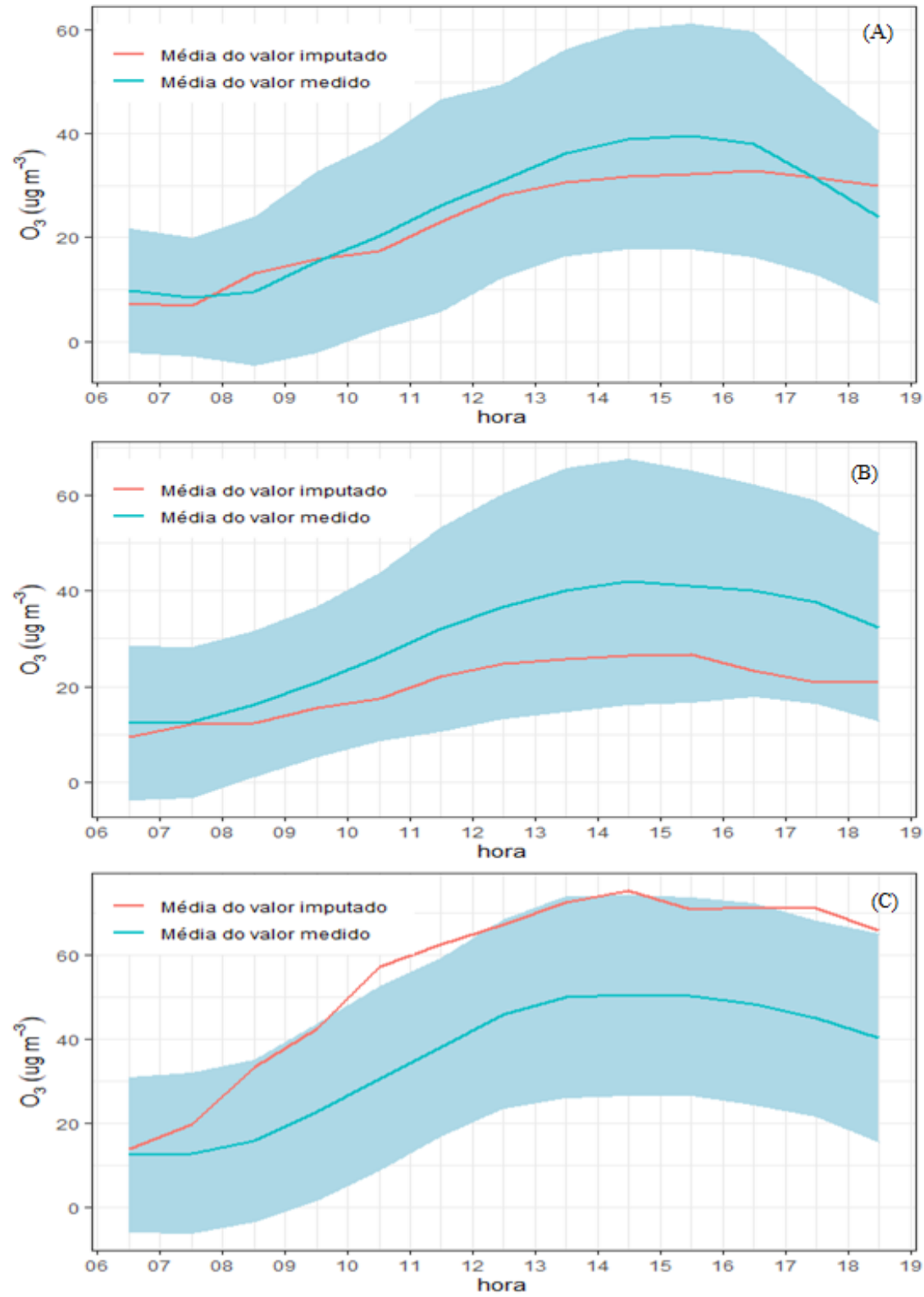
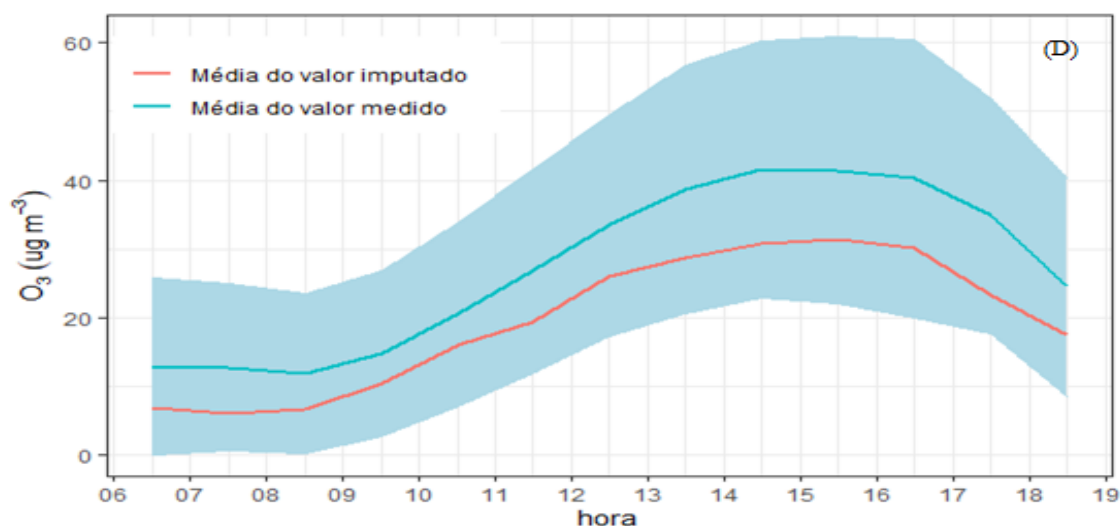


Figura 34- Comparação entre os valores médios medidos e os valores médios imputados para INE_out (A), INE_pri (B), INE_ver (C), INE_inv (D)





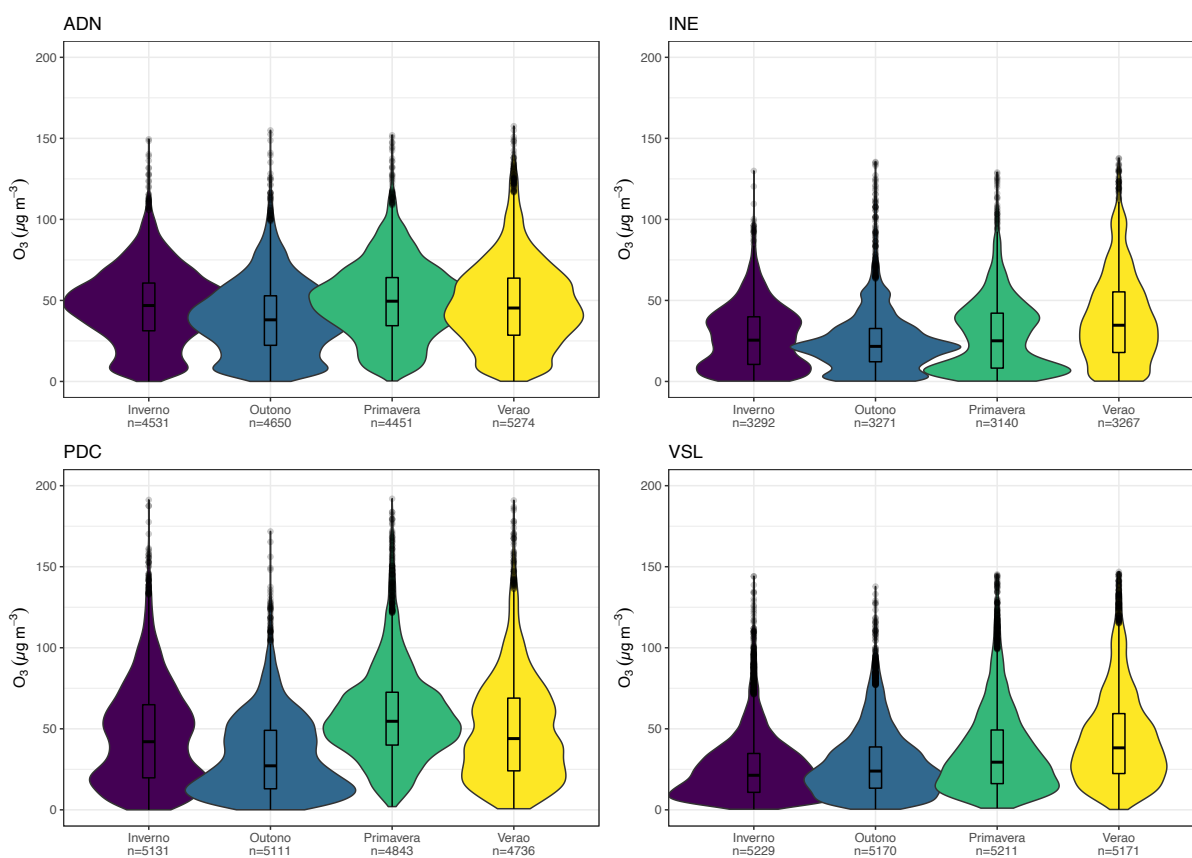
3.4 Análise Exploratória dos Dados

Uma análise exploratória foi realizada com o auxílio do gráfico de violino (Figura 35). A primeira observação a ser notada é que todas as EAQA apresentam uma quantidade de observações por estação do ano de forma homogênea, isto significa que não houve uma concentração desbalanceada entre as estações do ano, mais uma indicação que a imputação dos dados foi realizada de maneira mais abrangente possível. Também é possível perceber que os meses de primavera e verão apresentaram uma faixa maior de concentração de O_3 , devido apresentarem maiores temperatura em média e consequentemente radiação solar mais elevada. A explicação é que nos dias quentes os dias passam a ser mais longos, logo aumenta o processo fotoquímico na troposfera. Estes resultados também foram comprovados por Singh *et al.* (1997) e Kley *et al.* (1999).

A segunda observação é que a EAQA ADN concentra a maioria dos dados na faixa de $50 \mu g m^{-3}$ para todas as estações do ano e não apresenta muitas alterações ao longo destes 5 anos de estudo. A EAQA PDC demonstra uma dispersão maior entre as concentrações de O_3 em relação a EAQA de ADN, as estações de inverno, primavera e verão apresentaram maiores do que o outono. Este último mês apresentou a maioria das concentrações de O_3 menor de $25 \mu g m^{-3}$. Já para a EAQA de VSL é possível perceber nitidamente que os valores de concentração de O_3 são diferentes entre as estações do ano, nos meses de inverno, outono e primavera as concentrações de O_3 não ultrapassam a marca de $100 \mu g m^{-3}$, mas esta EAQA fica próxima de

uma estrada muito movimentada, a Rod. Washington Luiz, a qual pode estar interferindo no aumento da concentração de O_3 para o verão, devido uma parte desta EAQA está coberta pelo período de férias, diminuindo assim o fluxo veicular na via e diminuindo a concentração de NO na atmosfera provenientes destes veículos e de acordo com a Reação 3 o NO destrói o O_3 para formar NO_2 e O_2 (GERALDINO et al., 2020b). Para a EAQA INE é possível notar esta mesma característica que a VSL apresenta, visto que ambas são bem próximas de estradas bastantes movimentadas. A diferença é que para o inverno ela apresenta uma amplitude maior da concentração de O_3 e para o outono esta apresenta uma concentração maior na faixa dos $25 \mu g m^{-3}$ e para a primavera apresenta uma maior quantidade de concentração de O_3 abaixo de $25 \mu g m^{-3}$.

Figura 35- Distribuição por violino para ADN, PDC, VSL e INE para diferentes estações do ano

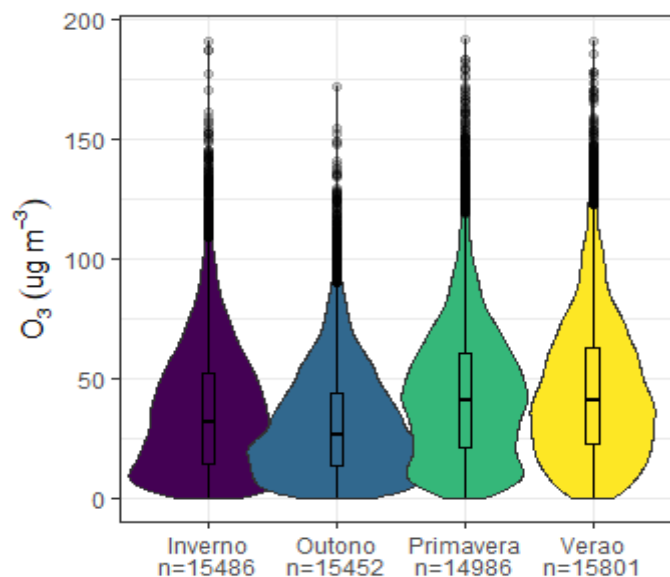


Fonte: O autor, 2022.

Uma maneira de observar o que pode estar acontecendo ao longo do ano, é estratificar em diferentes estações do ano. De acordo com a Figura 36 compiladas para todas as EAQA e todos os anos de estudo, que o mês de outono apresentou os menores valores de concentração

de O_3 , seguido do inverno. Já os meses primavera e verão apresentaram uma maior faixa de concentração de O_3 , devido a maior temperatura e radiação solar.

Figura 36- Diagrama de violino ADN, PDC, VSL, INE compiladas para as diferentes estações do ano para o O_3



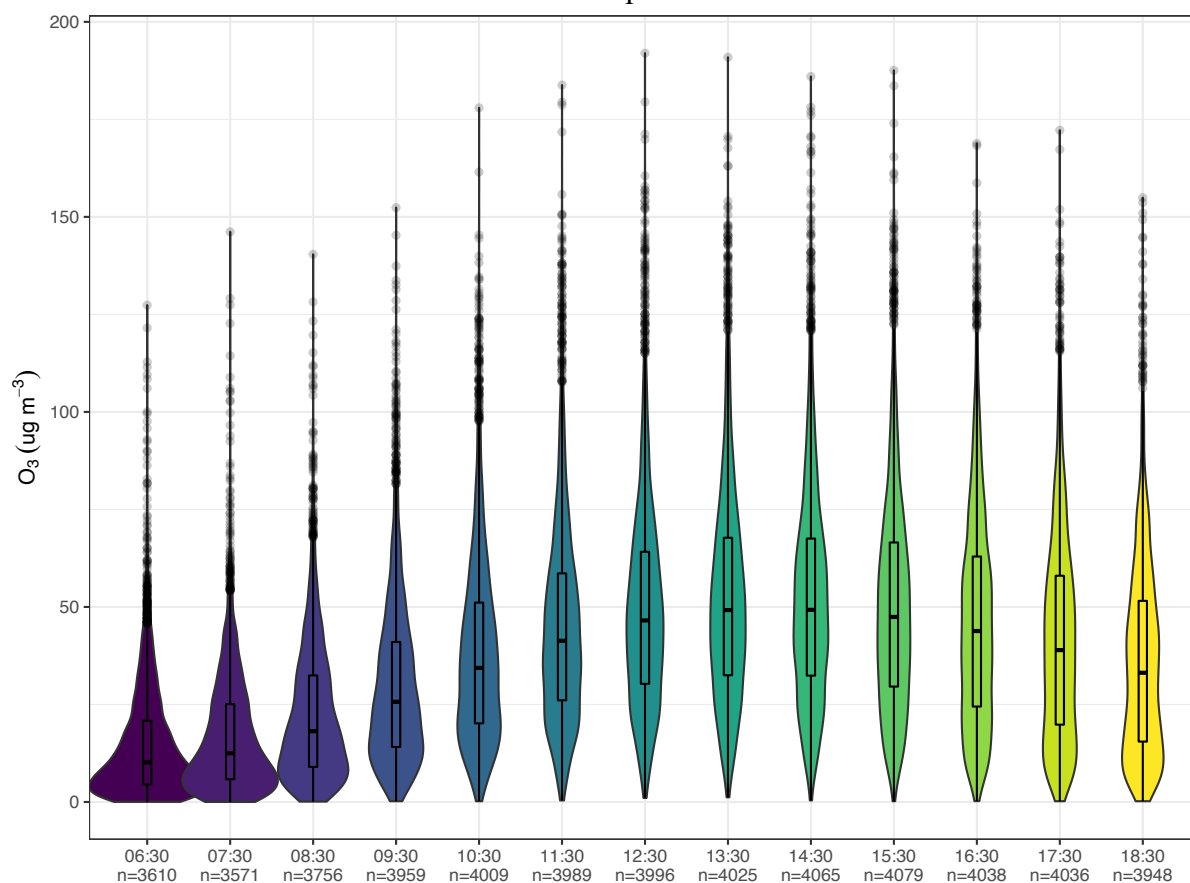
Fonte: O autor, 2022.

As concentrações de O_3 das 6:30 h até as 18:30 h são apresentadas na Figura 37, para todas as EAQA e todos os anos do estudo. Os números de observações estão na mesma proporção entre os diferentes dias da semana, logo, não houve uma maior concentração dos dados em um dia específico da semana. O O_3 é um poluente secundário que aumenta sua concentração ao longo do dia, devido a elevada temperatura e radiação solar. As 6:30 h pode representar uma parte desta concentração do O_3 devido ao resíduo acumulado do dia anterior, pois neste horário existe uma menor incidência direta de raios solares e a temperatura não está elevada.

Para as concentrações de O_3 para os dias da semana (Figura 38), não há muita variação na máxima alcançada, a qual ultrapassa uma concentração de $100 \mu g m^{-3}$ e a maioria dos valores de concentração de O_3 está na faixa abaixo de $50 \mu g m^{-3}$. Segundo Teixeira *et al.* (2009), como não há uma alteração significativa nas emissões das fontes primárias durante segunda a sexta-feira, quase nenhuma perturbação é encontrada nos níveis de concentração de O_3 nos dias da semana. Para sábado percebe-se um aumento da concentração de O_3 , isso pode ser explicado pelo fato de que a principal causa de níveis mais altos da concentração de ozônio no sábado é

a redução de emissões de NO_x ($\text{NO} + \text{NO}_2$) que podem ser atribuídas à redução do tráfego de veículos movidos à diesel, favorecendo assim uma formação mais rápida de ozônio devido à acúmulo de poluentes durante a semana (Reação 3) (GERALDINO et al., 2020b).

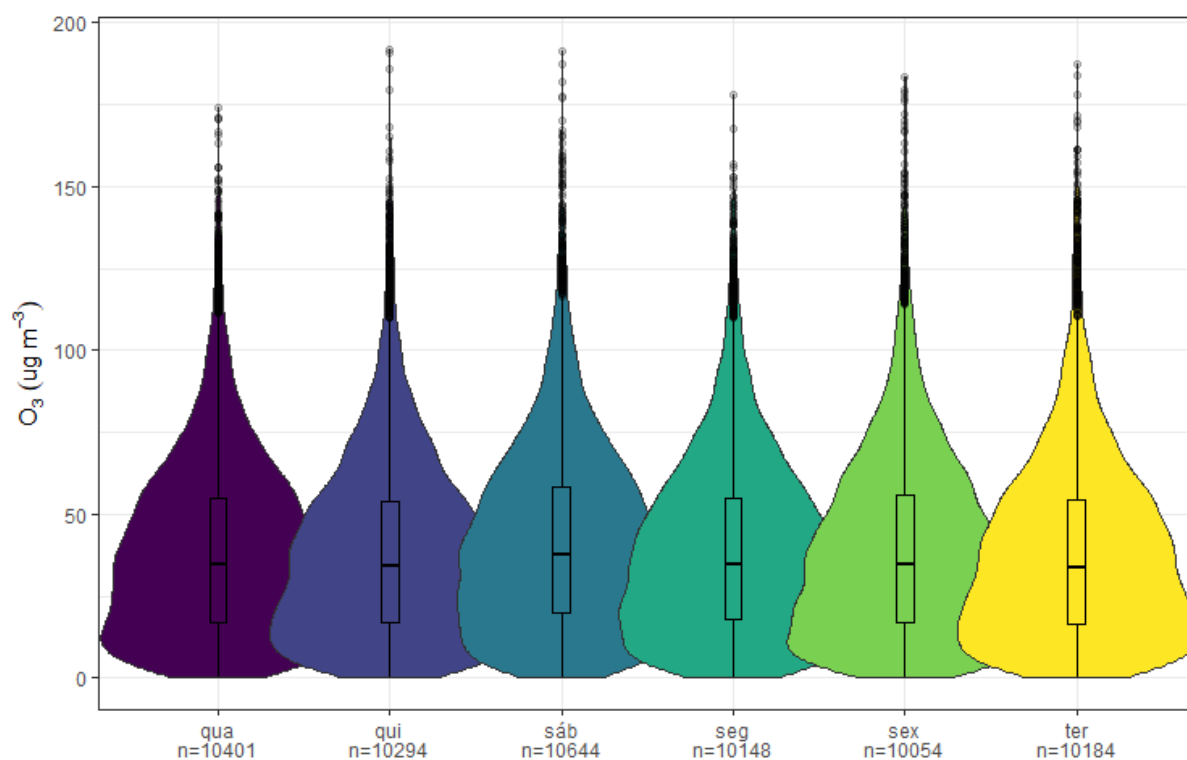
Figura 37- Diagrama de violino para ADN, PDC, VSL e INE compiladas para as diferentes horas do dia para o O_3



Fonte: O autor, 2022.

O aumento de O_3 durante sábado também pode ser explicado pela Reação 2, que mostra que a conversão de NO em NO_2 sem consumir uma molécula de ozônio, faz com que o ozônio acumule na troposfera. Esta reação ocorre na presença de COV, em particular, radicais peroxi (RO_2 , onde R é um grupo alquil) produzido durante a oxidação de moléculas de hidrocarbonetos, reagindo com NO para formar NO_2 , permitindo assim uma maior produção de ozônio (TEIXEIRA et al., 2009).

Figura 38- Diagrama violino para os diferentes dias da semana para o O_3 ($\mu g\ m^{-3}$)



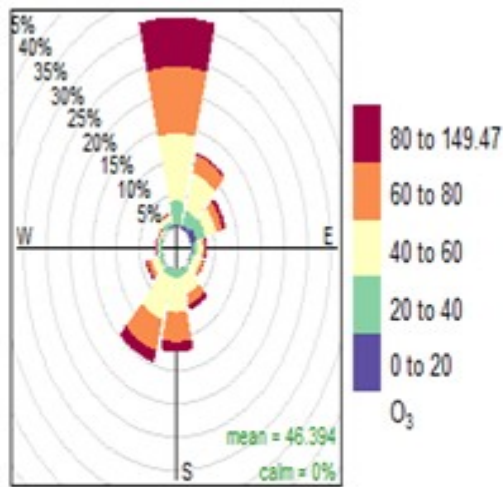
Fonte: O autor, 2022

3.4.1 Rosa dos ventos

A concentração de O_3 numa localidade específica é normalmente determinada pela direção do vento e velocidade do vento como um transporte de plumas a partir da fonte emissora. A Figura 39 apresenta a rosa dos ventos para a EAQA ADN para as diferentes estações do ano durante 2014 a 2018, apresentando um padrão para direção do vento diferente para cada estação do ano. A direção média do vento para o inverno e primavera contém uma certa similaridade pois os ventos se concentram-se majoritariamente no setor Norte, sugerindo uma direção do vento dominante para ADN nas estações de inverno e primavera vem do Norte e outra à Sudoeste, contribuindo mais de 50 % destes eventos. A direção do vento em Sudoeste, demonstra que o O_3 vem predominantemente da CSA, Gerdau, Votorantim que fazem parte do distrito industrial de Santa Cruz.

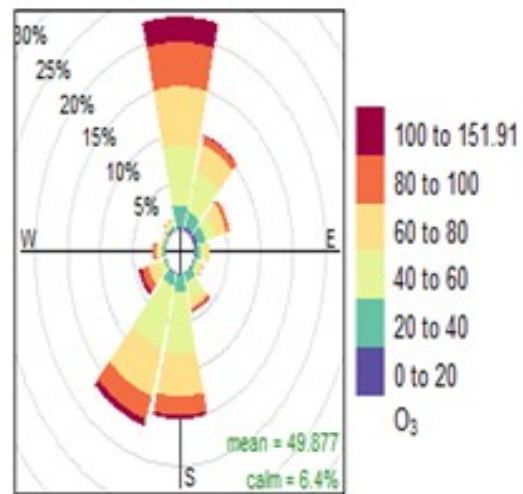
Figura 39- Rosa dos ventos para a EAQA ADN: inverno (A), primavera (B), outono (C), verão (D), concentrações de O₃ em µg m⁻³

(A)



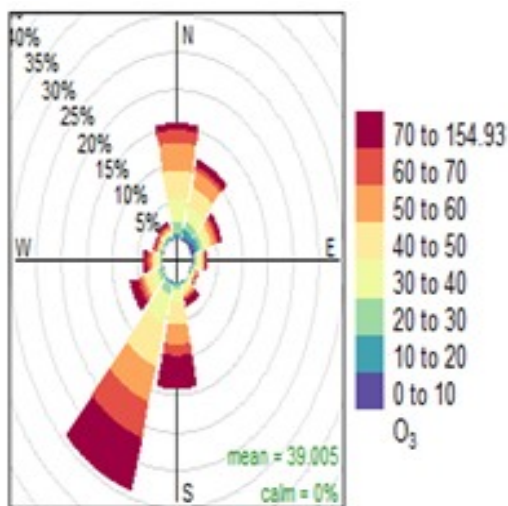
Proporção média da contribuição (%)

(B)



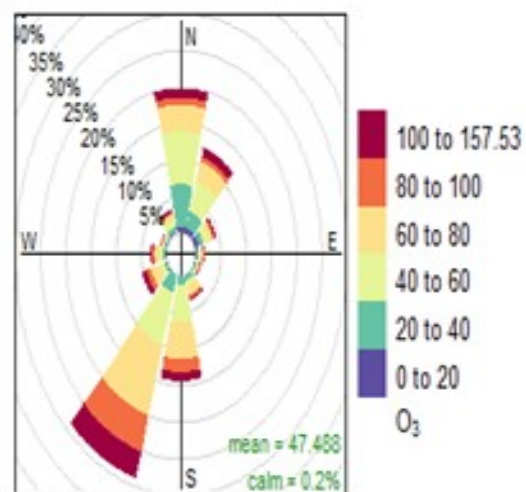
Proporção média da contribuição (%)

(C)



Proporção média da contribuição (%)

(D)



Proporção média da contribuição (%)

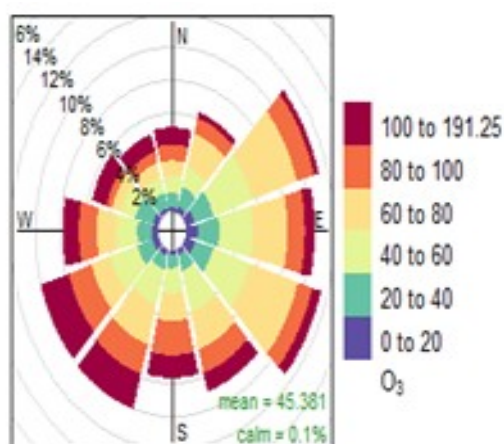
Fonte: O autor, 2022.

A Figura 40 apresenta a rosa dos ventos para a EAQA PDC para as diferentes estações do ano durante 2014 a 2018, apresentando um padrão para direção do vento diferente para cada estação do ano. A direção média do vento para todas as estações do ano contém uma certa direção pois os ventos se concentram-se na direção Sul, Leste e Oeste. Para estas direções de

vento, o O_3 provavelmente está vindo da cidade de Itaguaí e seus arredores, ou mesmo seus precursores sendo transportados, incrementando sua produção local.

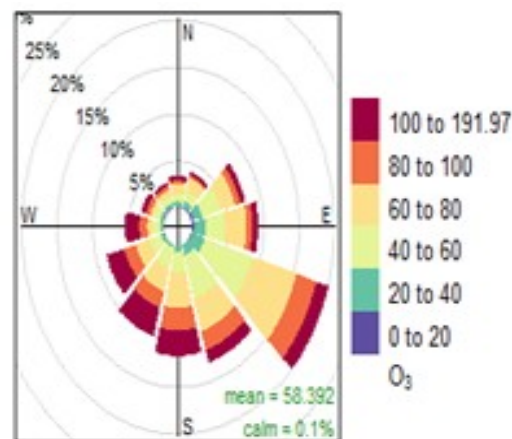
Figura 40- Rosa dos ventos para a EAQA PDC: inverno (A), primavera (B), outono (C) verão (D), concentrações de O_3 em $\mu g\ m^{-3}$

(A)



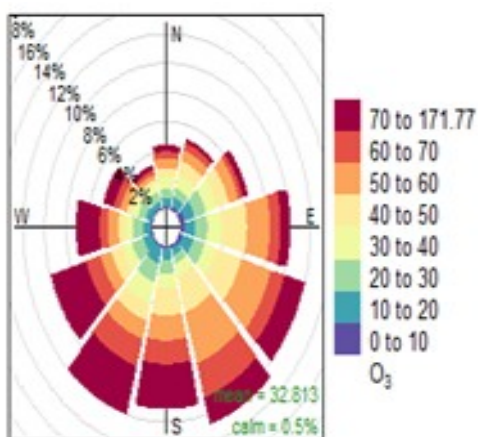
Proporção média da contribuição (%)

(B)



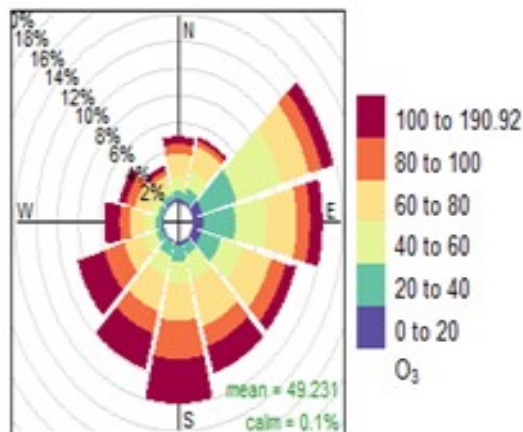
Proporção média da contribuição (%)

(C)



Proporção média da contribuição (%)

(D)



Proporção média da contribuição (%)

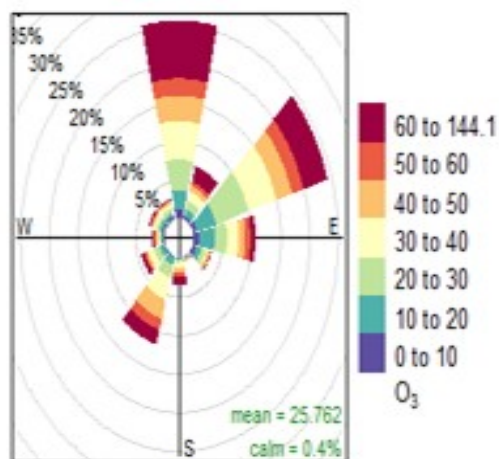
Fonte: O autor, 2022.

A Figura 41 apresenta a rosa dos ventos para a EAQA VSL para as diferentes estações do ano durante 2014 a 2018, apresentando um padrão para direção do vento diferente para cada estação do ano. A direção média do vento para o inverno e primavera contém uma certa similaridade pois os ventos se concentram-se majoritariamente no setor de 90°, sugerindo que o vento dominante para a VSL nas estações de inverno e primavera é de Norte e outra menor

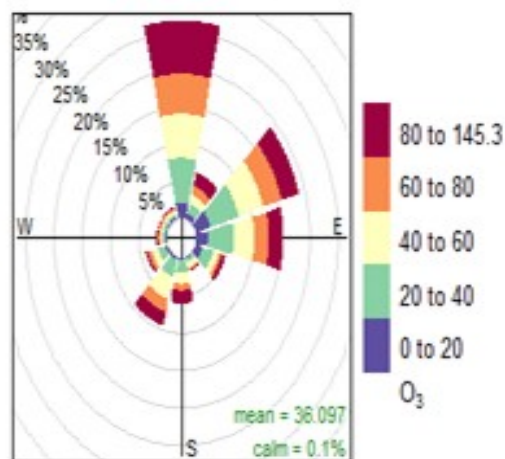
para à Sudoeste, contribuindo mais de 50 % destes eventos. Esta direção do vento indica que o O_3 está vindo proeminente da área da refinaria de Duque de Caxias (REDUC) e da interferência da Rod. Washigton Luiz.

Figura 41- Rosa dos ventos para a EAQA VSL: inverno (A), primavera (B), outono (C), verão (D), concentrações de O_3 em $\mu g m^{-3}$

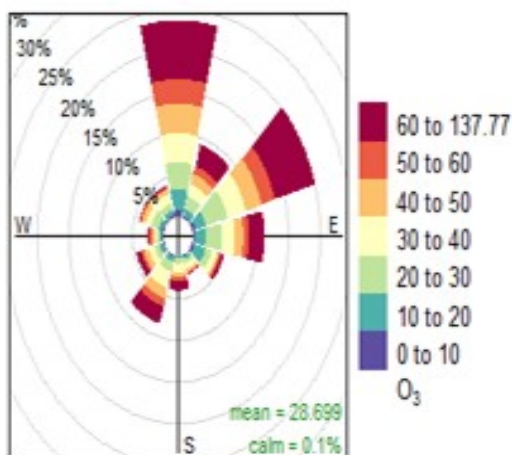
(A)



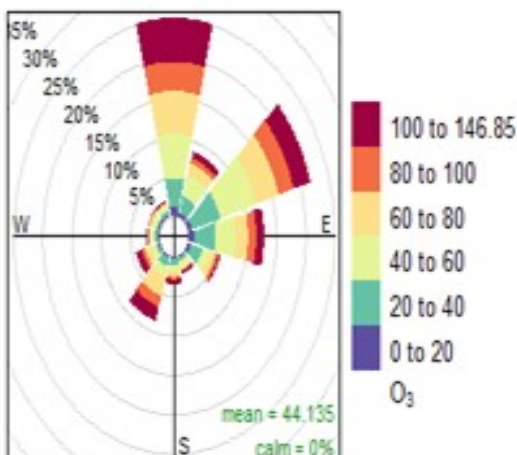
(B)



(C)



(D)



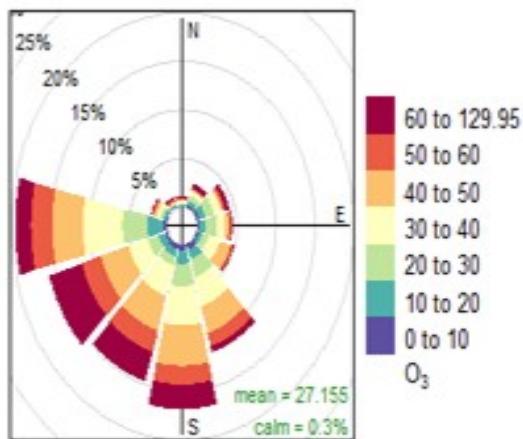
Fonte: O autor, 2022.

A Figura 42 apresenta a rosa dos ventos para a EAQA INE para as diferentes estações do ano durante 2014 a 2016, apresentando um padrão para direção do vento diferente para cada estação do ano. A direção média do vento para o inverno e primavera contém uma certa

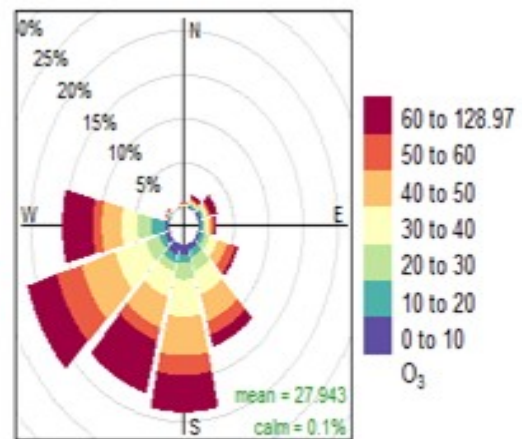
similaridade pois os ventos se concentram-se majoritariamente nos setores de 180° a 270°, sugerindo a vento dominante INE nas estações de inverno e primavera é Sudoeste, contribuindo mais de 50 % destes eventos. O vento em direção Sudoeste vindo do bairro do Recreio dos Bandeirantes.

Figura 42- Rosa dos ventos para a EAQA INE: inverno (A), primavera (B), outono (C), verão (D), concentrações de O_3 em $\mu g\ m^{-3}$

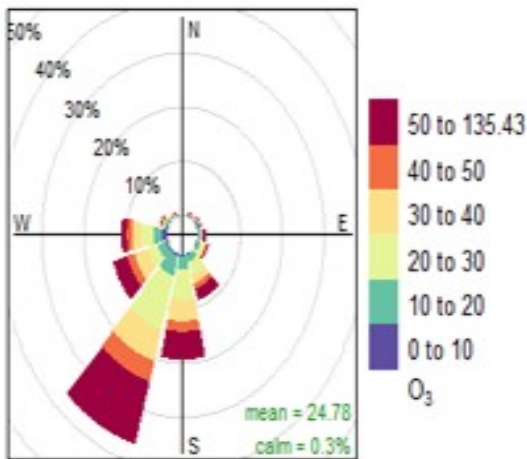
(A)



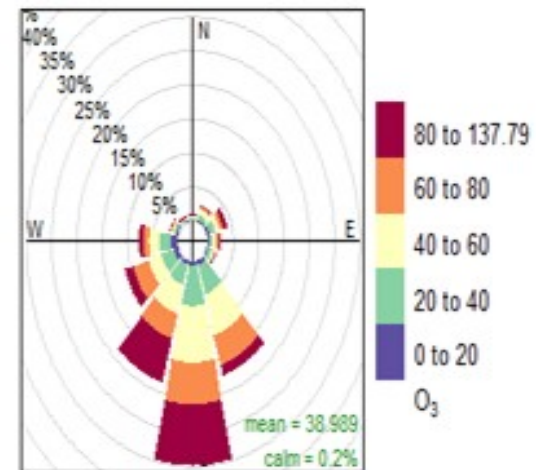
(B)



(C)



(D)



Fonte: O autor, 2022.

3.4.2 Gráfico de Calendário

O gráfico de calendário é uma ferramenta que mostra as médias diárias das concentrações de O_3 em 12 h. Todas as EAQA, ADN, PDC, VSL e INE foram disponibilizadas de 2014 a 2018. Para o padrão nacional de qualidade do ar (CONAMA, 2018) o limite de máxima média móvel obtida no dia para O_3 é de $140 \mu g m^{-3}$. Os gráficos de calendário estão disponíveis como Tabelas (B) e estão no apêndice B.

Para as estações no INE e ADN não houve ultrapassagens, porém para as Estações VSL e PDC, ambas excederam o limite. A EAQA VSL teve duas ultrapassagens em outubro de 2017, que foram nos dias da semana segunda e sábado. A EAQA PDC teve duas ultrapassagens, uma em outubro de 2015 numa segunda-feira e outra no mês próximo, em novembro de 2015 no sábado. Em 2015 teve o fenômeno do El Niño que é o superaquecimento das águas do Oceano Pacífico, e neste ano (2015) especificamente na estação do ano de primavera houve um aumento significativo da temperatura, o que pode ter influenciado no aumento da concentração de O_3 em PDC (SILVERSTONE, 2021). Como PDC fica rodeada por altas montanhas, a dispersão dos poluentes é prejudicada fazendo que o ar fique estagnado e os poluentes primários permanecem neste lugar, potencializando ainda mais a formação de O_3 . Conforme Teixeira *et al.* (2009) a concentração de O_3 pode ser maior nos fins de semana, pois existe menos quantidade de NO_x na atmosfera proveniente principalmente por fontes móveis, a qual participa na destruição de O_3 na troposfera, logo os mais elevados teores na segunda-feira são explicados pelo resíduo de poluente acumulado no dia anterior (domingo).

A EAQA VSL também apresenta problemas na dispersão dos poluentes, devido ao relevo e apresenta os maiores valores para a concentração de O_3 nos meses mais quentes principalmente na primavera.

3.5. Previsão da Concentração de Ozônio Troposférico

Os modelos PLS, RF, SVM e ANN foram construídos utilizando as variáveis independentes que apresentaram a maior correlação com o ozônio. Diferentes variáveis foram selecionadas para cada EAQA. A priori, os modelos que apresentaram o menor valor RMSEP para EAQA estudada, poderiam ser considerados adequados para a previsão de ozônio. Entre os modelos com menor RMSEP, é necessário comparar os modelos para encontrar aqueles com

os menores valores residuais entre os modelos estudados. O SWTP, proposto por Cunha *et al.* (2020), foi utilizado para comparar os valores residuais entre os modelos estudados. Um vetor contendo a soma da probabilidade do teste de Wilcoxon para todas as linhas foi obtido e, o valor mínimo deste vetor indicou o melhor modelo de predição de ozônio para cada EAQA. Os valores de SWTP para cada EAQA e estações do ano são mostrados nas Tabelas 12, 13, 14 e 15 para ADN, PDC, VSL e INE, respectivamente. O valor mínimo indica o melhor modelo dentre os modelos estudados.

Inicialmente, adotou-se o modelo PLS, mas este apresentou resultados piores do que os modelos SVM, RF e ANN para predição do ozônio troposférico, uma vez que este modelo é linear e os dados de poluentes atmosféricos são altamente não lineares. O critério de Wold (WOLD, 1978) foi aplicado para escolher um número adequado de variáveis latentes (VL) que pudessem ser consideradas para a construção do modelo PLS.

SVM, RF e ANN forneceram modelos mais aceitáveis para todos as EAQA, o que parece descrever a complexa relação entre a concentração de ozônio e as demais variáveis. Com relação ao tempo necessário do consumo da máquina para “rodar” o modelo, também houve uma grande diferença entre SVM, RF e ANN. O SVM gastou cerca de 9 horas e o RF juntamente com o ANN precisou de apenas 2 a 5 minutos. Talvez uma das explicações para que isso ocorra seja o número de parâmetros necessários para a função de ajuste do SVM, visto que para este último algoritmo ele testa de forma automática várias funções.

O fato das EAQAs estarem localizadas em locais distintos, com fontes distintas (industriais, veiculares, residenciais) pode ter contribuído para isso. Observou-se também que há uma alta variabilidade dos dados, pois há mudanças bruscas em algumas concentrações de poluentes atmosféricos, o que dificulta a previsão de O_3 em função das demais variáveis.

Segundo Cunha *et al.* (2020), o comportamento de aprendizagem do modelo SVM pode ser confirmado garantindo que o número de vetores de suporte (VS) seja menor que o número de observações utilizadas no conjunto de calibração. Caso contrário, um padrão comportamental de memorização seria identificado. Para a EAQA “ADN_ver” o menor SWTP foi obtido pelo modelo SVM, (conforme apresentado na Tabela 12). O modelo de calibração para “ADN_ver”, foi construído com 2623 vetores suportes e um conjunto de dados na calibração contendo 4219 observações. Neste caso, pode-se provar que o modelo SVM foi validado. No caso de “ADN_inv”, “ADN_pri” e “ADN_out” o RF teve um SWTP menor, então

estes foram escolhidos como os melhores EAQA. Os modelos mais bem ajustados são mostrados na Figura 43 para EAQA ADN.

A Figura 44 mostra o resultado do SWTP (A) e do viés da previsão (viés_prev) (B) compilado para todos os modelos utilizados. O PLS como é um método não linear, este método não conseguiu compreender o comportamento e características entre as variáveis e as observações, portanto já era esperado que este modelo iria apresentar alto valor para o SWTP e alta variação do viés. Para o ANN e SVM houve bastante variação entre os valores de SWTP e pouca variação do viés. O RF apresentou o menor valor para o SWTP e menor variação do viés em relação a todos os outros modelos, validando este último como o modelo mais adequado para a previsão de O₃ troposférico.

Tabela 12- Resumo dos resultados obtidos pelos modelos PLS, RF, SVM e ANN para ADN
EAQA

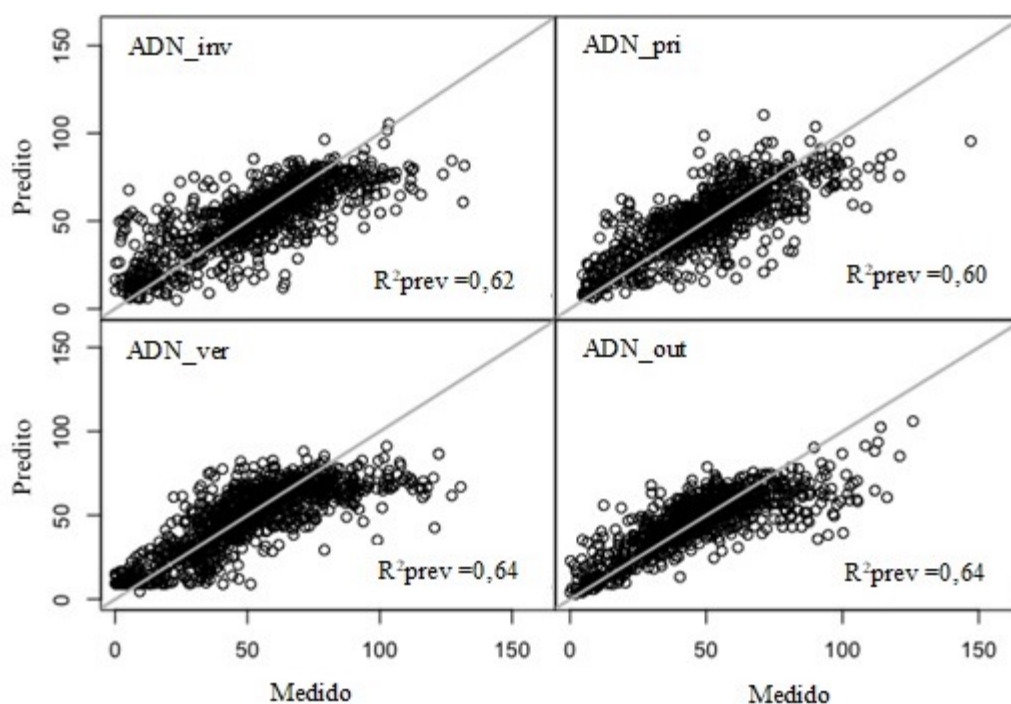
EAQA	Método	RMSEC	R ² cal	RMSEP	R ² prev	VS	VL	Viesprev	RMSECV	R ² cv	SWTP
ADN_inv	PLS	13,82	0,61	18,15	0,51	-	2	1,19	13,83	0,60	3,00
	RF	4,66	0,96	16,07	0,62	-	-	0,77	10,72	0,76	0,13
	SVM	9,90	0,80	15,81	0,63	1568	-	1,38	-	-	0,87
	ANN	10,84	0,76	15,63	0,64	-	-	1,62	-	-	2,00
ADN_out	PLS	15,76	0,50	17,74	0,38	-	3	3,49	15,78	0,50	3,00
	RF	4,92	0,95	13,52	0,64	-	-	0,76	11,71	0,73	0,04
	SVM	10,73	0,77	14,50	0,58	2826	-	1,18	-	-	1,00
	ANN	11,43	0,74	14,75	0,57	-	-	1,06	-	-	1,96
ADN_pri	PLS	16,38	0,53	15,23	0,53	-	2	-0,75	16,41	0,53	2,99
	RF	6,07	0,94	13,92	0,60	-	-	-0,22	13,88	0,66	0,45
	SVM	13,49	0,68	13,89	0,61	1959	-	0,62	-	-	1,10
	ANN	13,88	0,66	13,80	0,61	-	-	-0,64	-	-	1,45
ADN_ver	PLS	19,36	0,48	17,75	0,50	-	2	-5,62	19,39	0,48	3,00
	RF	7,33	0,93	15,49	0,62	-	-	-0,17	16,74	0,61	1,55
	SVM	16,92	0,60	15,11	0,64	2623	-	2,47	-	-	0,33
	ANN	17,47	0,58	14,95	0,64	-	-	0,17	-	-	1,12

Fonte: O autor, 2022.

O vies_prev é uma métrica que mostrou que alguns modelos obtiveram valores positivos e outros negativos, o que significa que as concentrações foram subestimadas ou superestimadas. No entanto, nenhum valor acima ou abaixo de zero foi muito significativo.

A Figura 45 mostra a figura de mérito RMSEP ($\mu\text{g m}^{-3}$) a qual indica que para todas as EAQA o PLS foi o método que apresentou maior erro, juntamente com o maior valor para o SWTP. Na Figura 46 indica os valores de RMSEP para as EAQA por estações do ano. As estações do INE apresentaram menor RMSEP ($\mu\text{g m}^{-3}$), isso pode ter ocorrido devido ao INE apresentar a maior quantidade de variável elegíveis, em contrapartida a estação PDC apresentou maiores valores para o RMSEP, isso ocorre porque estação possui dados mais dispersos que as outras EAQA.

Figura 43- Modelos mais bem ajustados - Gráfico previsto versus medido para o ozônio para ADN EAQA



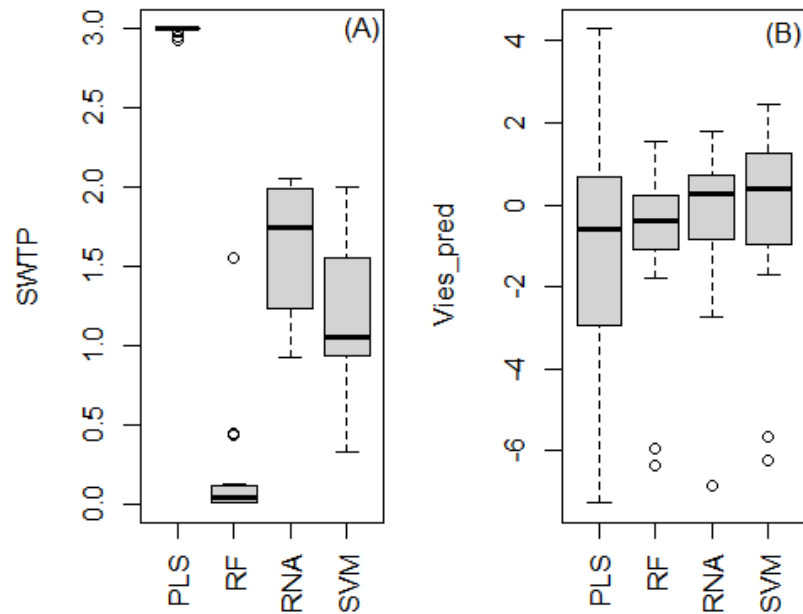
Fonte: O autor, 2022.

Para as estações do ano (Figura 47) *versus* RMSEP ($\mu\text{g m}^{-3}$) é evidente que nas estações do ano verão e inverno obtiveram um erro maior que outubro e primavera. Isto demonstra que para as estações antagônicas (inverno e verão), os modelos não conseguiram se ajustar igual ao outono e primavera.

O PDC EAQA está localizado em uma área que compreende basicamente fontes industriais, o que significa que a variabilidade dos dados é menor que os demais EAQA. Esta é também uma região com morros e depressões, o que piora a dispersão dos poluentes. Portanto, o PDC EAQA teve um valor de R^2_{prev} mais baixo (valor mais alto mostrado na Tabela 13, 0,64

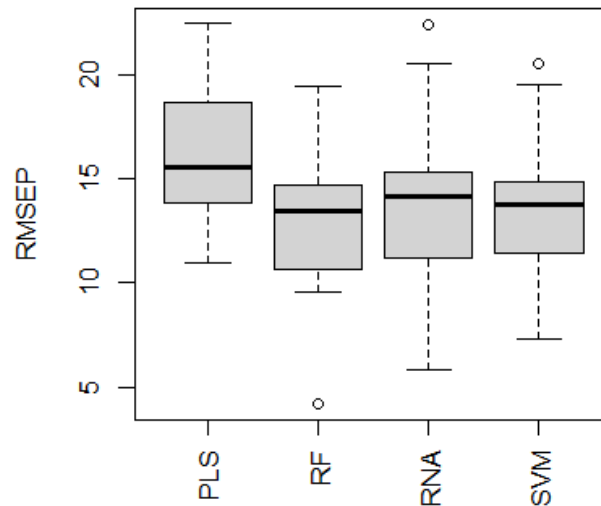
no outono) do que as outras EAQA estudadas. Os modelos mais bem ajustados são mostrados na Figura 48 para PDC EAQA.

Figura 44- Soma da probabilidade do teste de Wilcoxon (A) e o viés (B) para todos os modelos de previsão do O₃

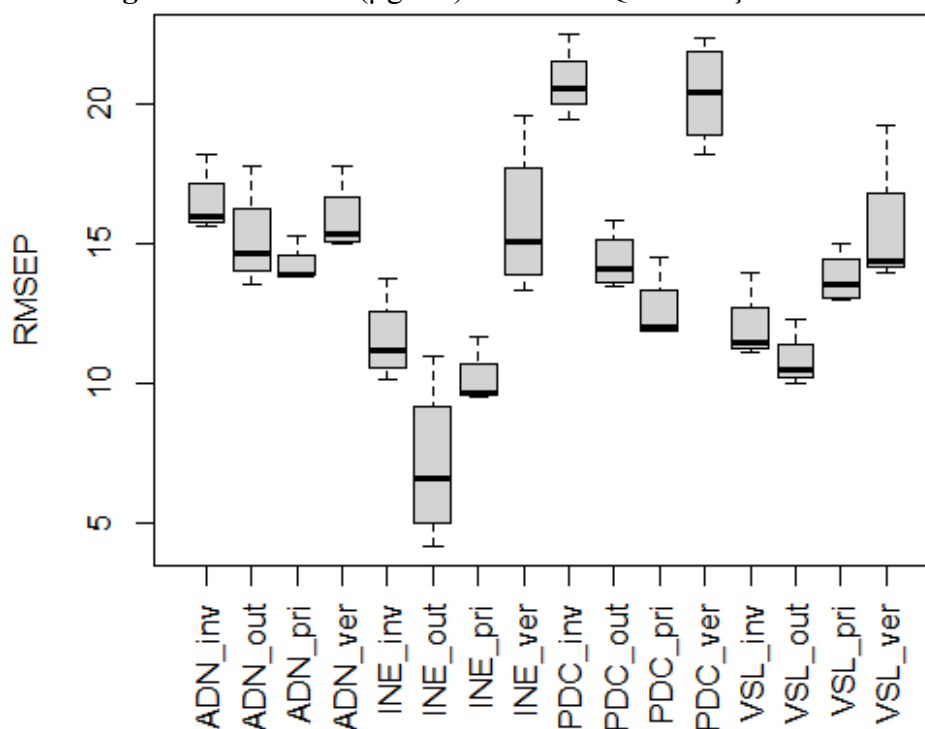


Fonte: O autor, 2022.

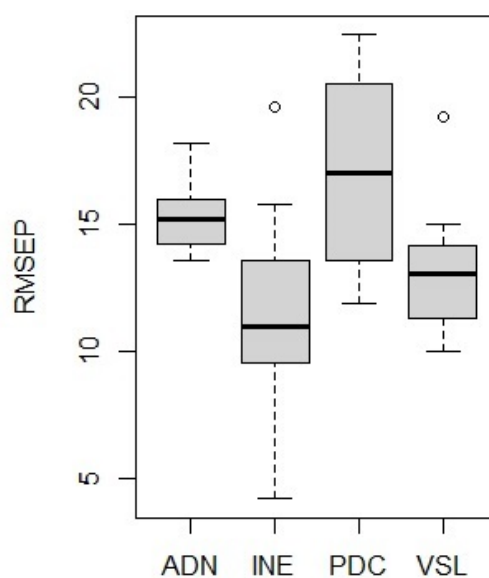
Figura 45- RMSEP ($\mu\text{g m}^{-3}$) versus EAQA para todos os modelos de previsão do O₃ troposférico



Fonte: O autor, 2022.

Figura 46- RMSEP ($\mu\text{g m}^{-3}$) *versus* EAQA e estações do ano

Fonte: O autor, 2022.

Figura 47- RMSEP ($\mu\text{g m}^{-3}$) *versus* estações do ano

Fonte: O autor, 2022.

O EAQA VSL está localizada em uma área que possui dois tipos de fontes poluentes: veicular e industrial, o que significa que a variabilidade e amplitude dos dados é maior que os

demais EAQA. Outro fator importante é a topografia, por se tratar de uma região com morros e depressões, o que leva a alterações na dispersão dos poluentes. Por esta razão, o VSL EAQA teve um valor de R^2_{prev} mais baixo (valor mais alto mostrado na Tabela 14, 0,74 no verão) do que o INE EAQA. Os modelos mais bem ajustados são mostrados na Figura 49 para VSL EAQA.

Tabela 13- Resumo dos resultados obtidos pelos modelos PLS, RF, SVM e ANN para PDC EAQA

EAQA	Método	RMSEC	R^2_{cal}	RMSEP	R^2_{prev}	VS	VL	Viesprev	RMSECV	R^2_{cv}	SWTP
PDC_inv	PLS	23,83	0,41	22,45	0,49	-	3	0,92	23,85	0,40	2,93
	RF	8,48	0,92	19,45	0,61	-	-	1,55	20,24	0,57	<0,01
	SVM	18,71	0,63	20,48	0,57	1997	-	1,32	-	-	1,71
	ANN	20,35	0,57	20,52	0,57	-	-	0,92	-	-	1,35
PDC_out	PLS	17,87	0,47	15,81	0,51	-	3	1,42	17,91	0,47	3,00
	RF	6,27	0,94	13,49	0,64	-	-	-0,31	14,90	0,63	0,09
	SVM	14,20	0,67	13,65	0,63	2405	-	0,61	-	-	0,91
	ANN	14,61	0,64	14,45	0,65	-	-	0,40	-	-	2,00
PDC_pri	PLS	24,62	0,36	14,48	0,37	-	3	-1,77	24,66	0,36	3,00
	RF	9,32	0,91	11,84	0,58	-	-	-1,77	21,60	0,51	0,06
	SVM	20,30	0,56	11,87	0,57	2201	-	-0,81	-	-	1,10
	ANN	21,16	0,53	12,11	0,55	-	-	-2,73	-	-	1,84
PDC_ver	PLS	27,15	0,32	21,26	0,37	-	3	-7,28	27,19	0,32	2,98
	RF	10,65	0,90	18,14	0,54	-	-	-6,37	24,85	0,43	<0,01
	SVM	24,55	0,44	19,52	0,47	1918	-	-6,25	-	-	0,99
	ANN	24,71	0,4	22,36	0,47	-	-	-2,34	-	-	2,00

Figura 48- Modelos mais bem ajustados - Gráfico previsto versus medido para o ozônio para PDC EAQA

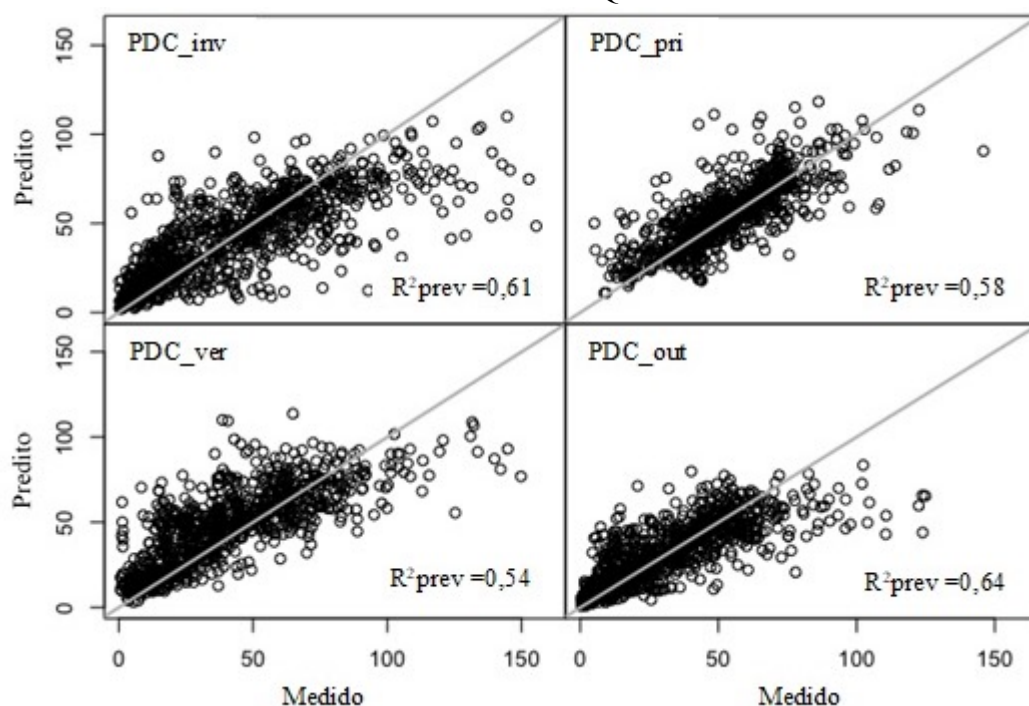
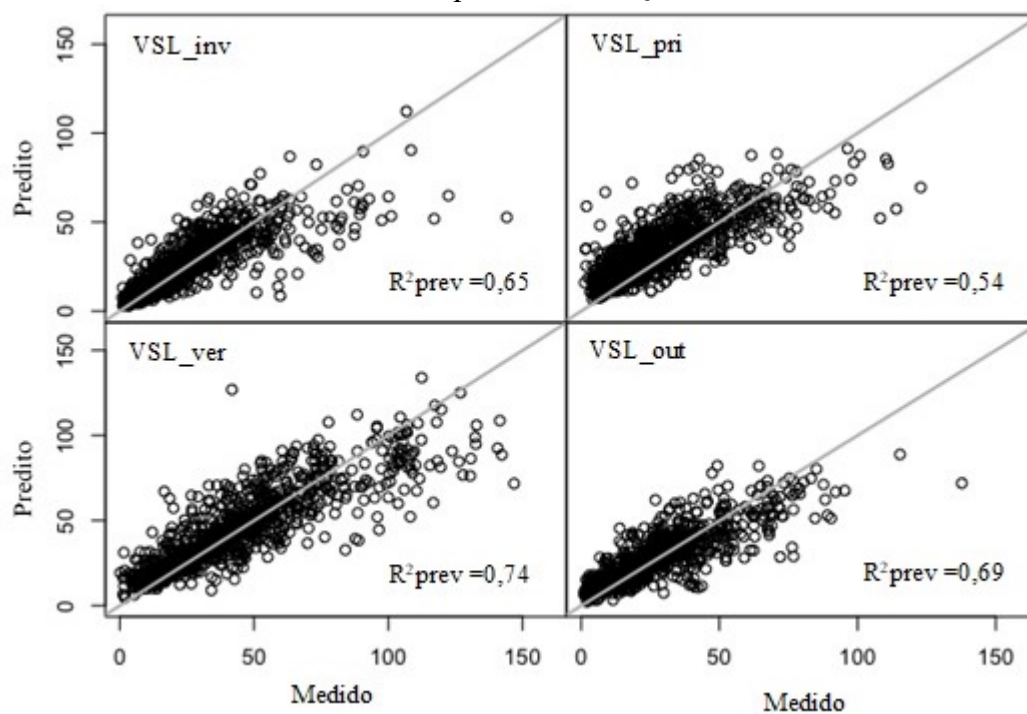


Tabela 14- Resumo dos resultados obtidos pelos modelos PLS, RF, SVM e ANN para VSL EAQA

EAQA	Método	RMSEC	R ² cal	RMSEP	R ² prev	VS	VL	Viesprev	RMSECV	R ² cv	SWTP
VSL_inv	PLS	14,37	0,51	13,91	0,46	-	4	0,40	14,39	0,51	3,00
	RF	5,00	0,94	11,11	0,65	-	-	-0,75	11,49	0,69	0,44
	SVM	11,04	0,71	11,39	0,64	1664	-	0,15	-	-	0,57
	ANN	10,56	0,74	11,45	0,63	-	-	0,54	-	-	1,99
VSL_out	PLS	15,39	0,49	12,25	0,54	-	4	-1,40	15,40	0,49	3,00
	RF	5,63	0,93	9,97	0,69	-	-	-0,45	13,22	0,63	<0,01
	SVM	12,81	0,65	10,42	0,67	1315	-	-0,77	-	-	1,41
	ANN	13,29	0,62	10,48	0,66	-	-	-1,06	-	-	1,58
VSL_pri	PLS	23,25	0,28	14,96	0,39	-	3	-6,66	23,27	0,28	3,00
	RF	9,47	0,88	12,95	0,54	-	-	-5,96	21,54	0,38	<0,01
	SVM	21,22	0,40	13,15	0,53	1613	-	-5,67	-	-	1,01
	ANN	21,86	0,36	13,92	0,47	-	-	-6,86	-	-	1,99
VSL_ver	PLS	20,34	0,51	19,20	0,51	-	2	-4,41	20,37	0,51	3,00
	RF	6,61	0,95	13,92	0,74	-	-	-1,23	15,72	0,71	0,12
	SVM	14,49	0,75	14,34	0,73	1477	-	-1,70	-	-	1,94
	ANN	15,49	0,72	14,34	0,73	-	-	0,44	-	-	0,93

Figura 49- Modelos mais bem ajustados - Gráficos previstos versus medidos para o ozônio para VSL EAQA



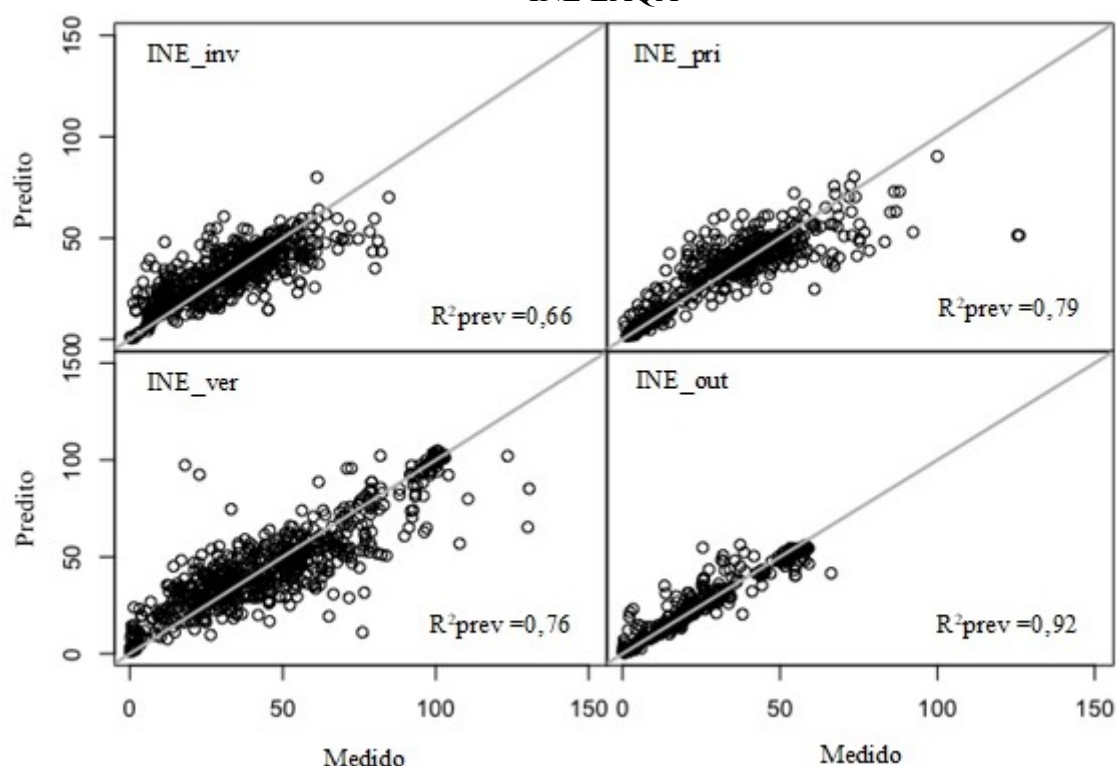
O EAQA INE mostrou que o R^2_{prev} foi melhor do que os outros EAQA, o que pode dever-se a 2 fatores: (i) a proximidade ao mar e naturalmente ter maior estabilidade na umidade relativa e consequentemente na temperatura, principalmente no outono conforme observado na Tabela 15 com R^2_{prev} igual a 0,92 e apenas fontes veiculares. Isso levou a uma menor variabilidade dos dados e o modelo foi capaz de entender esse comportamento e foi mais adequado a esses dados. (ii) O EAQA do INE possui uma grande quantidade de variáveis, com isso o algoritmo *MissForest* foi capaz de preencher as lacunas de forma mais assertiva, pois quanto maior o número de variáveis disponíveis maior a captura das características e comportamento entre as variáveis, obtendo assim um melhor desempenho do que as outras EAQA. Os modelos mais adequados são mostrados na Figura 50 para o INE EAQA.

Tabela 15- Resumo dos resultados obtidos pelos modelos PLS, RF, SVM e ANN para o INE EAQA

EAQA	Método	RMSEC	R ² cal	RMSEP	R ² prev	VS	VL	Viesprev	RMSECV	R ² cv	SWTP
INE_inv	PLS	13,40	0,54	13,76	0,37	-	4	0,41	13,43	0,54	3,00
	RF	4,67	0,94	10,14	0,66	-	-	-0,97	10,92	0,69	0,02
	SVM	10,26	0,73	11,39	0,57	1294	-	-0,42	-	-	1,87
	ANN	10,13	0,74	10,96	0,60	-	-	-0,03	-	-	1,11

EAQA	Método	RMSEC	R ² cal	RMSEP	R ² prev	VS	VL	Viesprev	RMSECV	R ² cv	SWTP
INE_out	PLS	16,52	0,33	10,94	0,42	-	3	-0,34	16,54	0,32	3,00
	RF	5,93	0,91	4,17	0,92	-	-	-0,44	13,38	0,56	<0,01
	SVM	13,81	0,53	7,32	0,74	623	-	-1,15	-	-	2,00
	ANN	13,16	0,57	5,86	0,83	-	-	-0,46	-	-	1,00
INE_pri	PLS	13,62	0,65	11,64	0,69	-	3	-0,41	13,65	0,64	3,00
	RF	4,86	0,95	9,52	0,79	-	-	0,10	10,87	0,77	0,05
	SVM	10,10	0,80	9,68	0,78	1205	-	0,71	-	-	1,30
	ANN	9,86	0,81	9,56	0,79	-	-	0,34	-	-	1,65
INE_ver	PLS	20,27	0,45	19,56	0,48	-	4	4,31	20,32	0,45	2,95
	RF	6,63	0,94	13,32	0,76	-	-	0,36	15,66	0,67	0,04
	SVM	13,55	0,76	14,36	0,72	1329	-	1,86	-	-	0,96
	ANN	15,19	0,69	15,78	0,68	-	-	1,78	-	-	2,05

Figura 50- Modelos mais bem ajustados - Gráfico previsto versus medido para o ozônio para INE EAQA



Esses resultados podem ser comparados com os de outros estudos da literatura, como Luna *et al.* (2014) e Sousa *et al.* (2006). Em seus trabalhos, os autores estudaram a previsão de ozônio em cidades e, em ambas, os melhores resultados foram alcançados com o algoritmo ANN. Na primeira, o melhor coeficiente de determinação para predição (R^2_{Prev}) foi igual a

0,89 e o menor RMSEP foi igual a 8,10 $\mu\text{g m}^{-3}$. No segundo, a melhor medida de variabilidade dos dados reproduzidos no modelo (R) foi igual a 0,73 e o erro quadrático médio (RMSE) foi igual a 21,78 $\mu\text{g m}^{-3}$.

Todos os resultados obtidos foram imputados, contudo um estudo foi realizado retirando todos os dados faltantes e aplicando o modelo que apresentaram o melhor resultado (menor SWTP) para cada EAQA, com a finalidade de comparar o RMSEP antes e após a imputação dos dados. A Tabela 16 apresenta as figuras de mérito obtidas sem a imputação.

Tabela 16- Resumo dos resultados das figuras de mérito com os valores imputados (input) e com valores faltantes (NA) para todas as EAQA

EAQA	Método	RMSEC	R ² cal	RMSEP	R ² prev	VS	Vies_prev	RMSECV	R ² cv
ADN_inv_input	RF	4,66	0,96	16,07	0,62	-	0,77	10,72	0,76
ADN_inv_NA	RF	5,48	0,94	11,32	0,74	-	7,00	12,39	0,72
ADN_out_input	RF	4,92	0,95	13,52	0,64	-	0,76	11,71	0,73
ADN_out_NA	RF	12,76	0,94	11,85	0,73	-	0,60	12,76	0,70
ADN_pri_input	RF	6,07	0,94	13,92	0,60	-	-0,22	13,88	0,66
ADN_pri_NA	RF	6,22	0,93	14,99	0,58	-	3,28	14,06	0,65
ADN_ver_input	SVM	16,92	0,60	15,11	0,64	2623	2,47	-	-
ADN_ver_NA	SVM	17,55	0,59	18,11	0,55	1497	6,00	-	-
PDC_inv_input	RF	8,48	0,92	19,45	0,61	-	1,55	20,24	0,57
PDC_inv_NA	RF	8,62	0,92	22,08	0,51	-	-1,84	20,35	0,57
PDC_out_input	RF	6,27	0,94	13,49	0,64	-	-0,31	14,90	0,63
PDC_out_NA	RF	6,17	0,93	15,24	0,62	-	0,27	14,64	0,62
PDC_pri_input	RF	9,32	0,91	11,84	0,58	-	-1,77	21,60	0,51
PDC_pri_NA	RF	8,88	0,91	22,41	0,53	-	-2,98	20,52	0,53
PDC_ver_input	RF	10,65	0,90	18,14	0,54	-	-6,37	24,85	0,43
PDC_ver_NA	RF	9,70	0,91	21,90	0,51	-	0,94	22,71	0,51
VSL_inv_input	RF	5,00	0,94	11,11	0,65	-	-0,75	11,49	0,69
VSL_inv_NA	RF	5,00	0,94	13,66	0,64	-	-1,20	11,75	0,70
VSL_out_input	RF	5,63	0,93	9,97	0,69	-	-0,45	13,22	0,63
VSL_out_NA	RF	5,76	0,93	15,01	0,53	-	1,42	13,38	0,62
VSL_pri_input	RF	9,47	0,88	12,95	0,54	-	-5,96	21,54	0,38
VSL_pri_NA	RF	9,43	0,89	23,68	0,36	-	1,84	21,36	0,43
VSL_ver_input	RF	6,61	0,95	13,92	0,74	-	-1,23	15,72	0,71
VSL_ver_NA	RF	6,79	0,95	15,73	0,73	-	0,04	16,23	0,71
INE_inv_input	RF	4,67	0,94	10,14	0,66	-	-0,97	10,92	0,69
INE_inv_NA	RF	4,32	0,95	12,67	0,64	-	-0,37	10,23	0,73
INE_out_input	RF	5,93	0,91	4,17	0,92	-	-0,44	13,38	0,56

EAQA	Método	RMSEC	R ² cal	RMSEP	R ² prev	VS	Vies_prev	RMSECV	R ² cv
INE_out_NA	RF	3,89	0,95	11,01	0,66	-	-1,68	8,93	0,74
INE_pri_input	RF	4,86	0,95	9,52	0,79	-	0,10	10,87	0,77
INE_pri_NA	RF	5,03	0,95	12,17	0,67	-	0,56	11,38	0,73
INE_ver_input	RF	6,63	0,94	13,32	0,76	-	0,36	15,66	0,67
INE_ver_NA	RF	5,23	0,95	13,20	0,70	-	1,27	15,60	0,71

De acordo com a Tabela 16 somente na EAQA de ADN inverno e outono os resultados com a retirada dos NAs apresentaram melhores resultados que os imputados. Isto se deve ao fato especificamente que esta estação apresentou pequenas quantidades de dados faltantes, em torno de 11 % no geral. Para o ano de 2018 a EAQA de ADN não possui dados entre maio e setembro, este evento pode ter influenciado no resultado desta EAQA, pois nestes meses estão compreendidos o outono e o inverno. A retirada dos NAs é um problema pois quando o banco de dados possui características de *missing data* do tipo MAR, toda a linha da observação é retirada do estudo, perdendo assim bastante informações relevantes sobre o banco.

3.6. Determinação da importância de cada variável na formação do Ozônio com auxílio da técnica de classificação Boruta

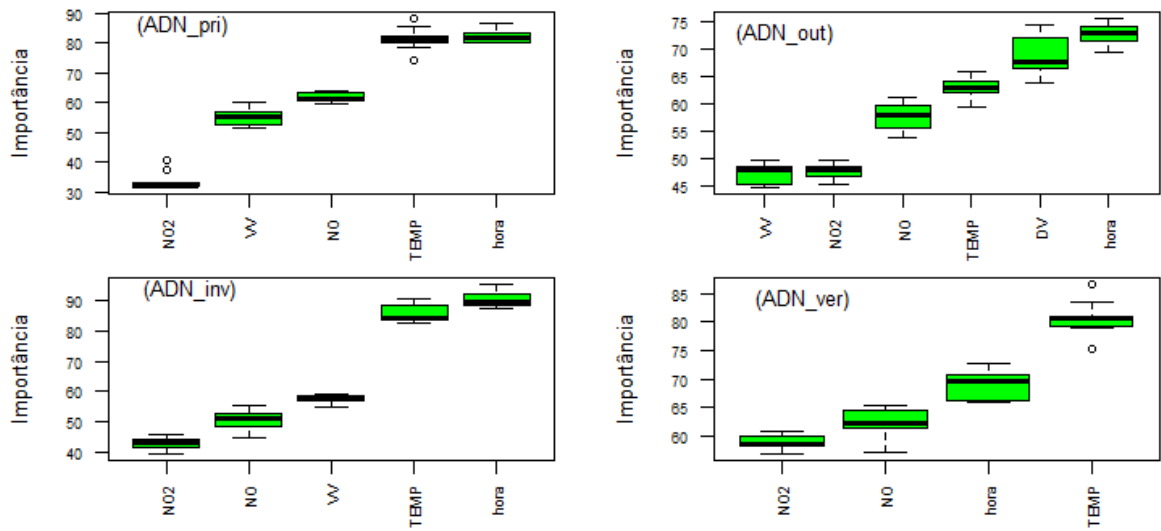
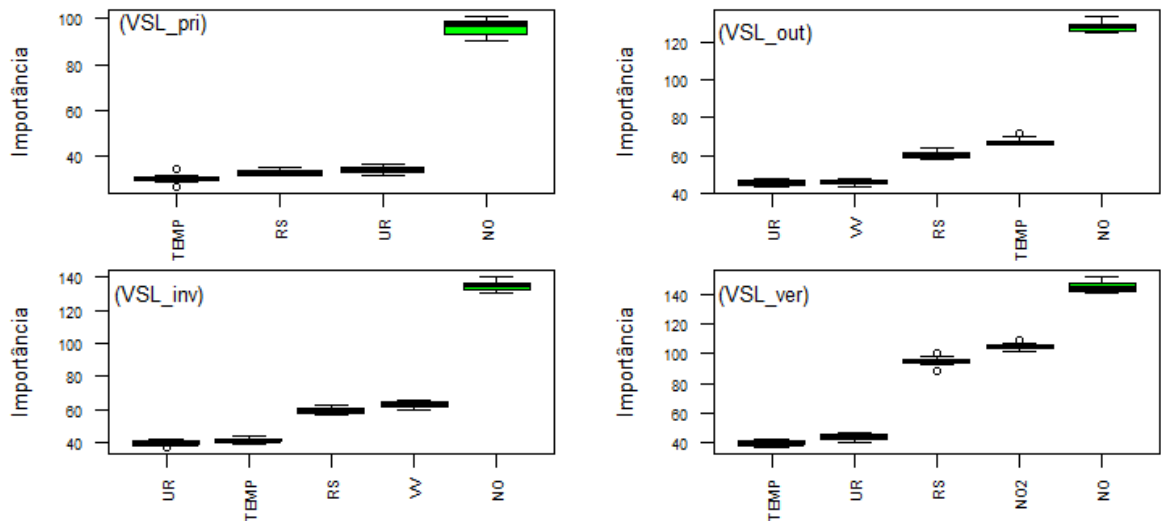
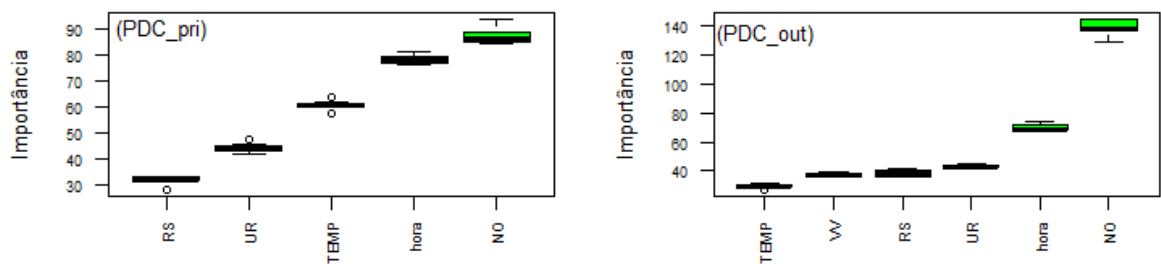
Para a EAQA de ADN (Figura 51) para todas as estações do ano, as variáveis meteorológicas (temperatura e hora) são as que tem maior contribuição na formação do O₃, este fato se deve que a principal via para esta formação provavelmente se deve a fotólise do NO₂ presente na atmosfera, de acordo com a Reação 1.

Para a EAQA de PDC (Figura 52) para todas as estações do ano, as variáveis NO e hora são as que mais contribuem para a formação de O₃, isto ocorre, proveniente de fonte móvel, já que existem pontos de ônibus próximas a estação meteorológica, principalmente por circulação de veículos pesados, como ônibus e caminhões.

Para a EAQA de VSL (Figura 53), para todas as estações do ano, a variável NO é a que possui maior contribuição na formação do O₃, devido à proximidade com a Rod. Washington Luiz que contém muito fluxo de veículos pesados diariamente.

Para a EAQA do INE (Figura 54), para as estações do ano (primavera e inverno), a variável NO₂ possui maior contribuição na formação do O₃, onde a estação meteorológica fica próxima a uma estação do BRT. Enquanto para as estações do ano (primavera e outono), as

variáveis que mais contribuem na formação do O_3 , são temperatura e hora, nestas sazonalidades a fotólise do NO_2 parece ter uma contribuição maior.

Figura 51- Classificação Boruta para a ADN EAQA**Figura 52- Classificação Boruta para a VSL EAQA****Figura 53- Classificação Boruta para a PDC EAQA**

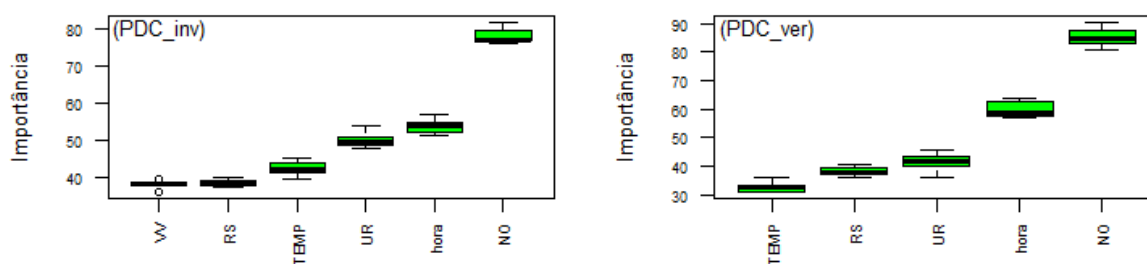
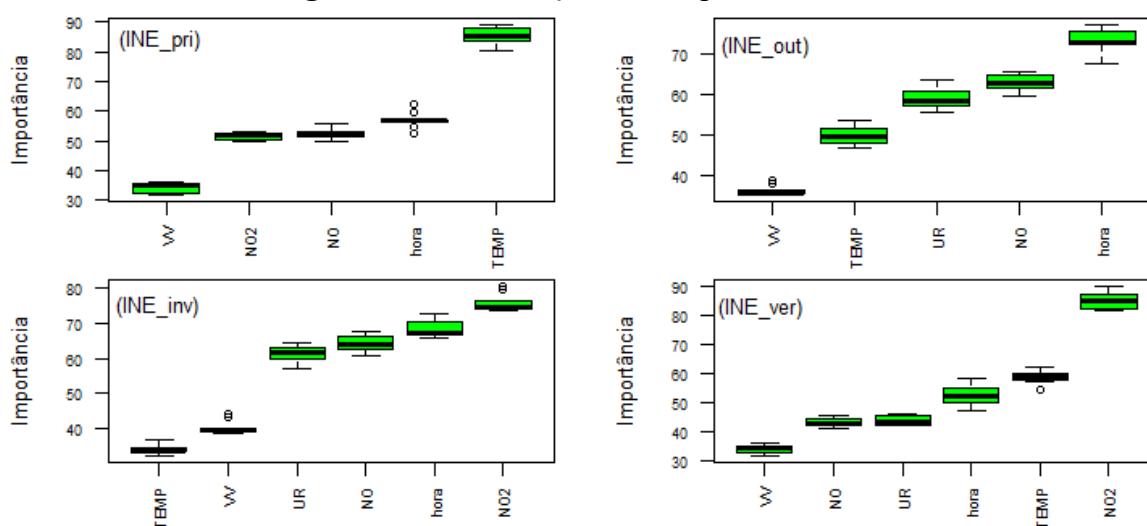


Figura 54- Classificação Boruta para a INE EAQA



3.7. Exemplo Prático das Ferramentas Desenvolvidas

Um estudo foi realizado para verificar a quantidade mínima de dias, que as EAQA podiam ficar coletando os dados, para que estes conjuntos de dados fossem validados e pudessem ser extraídas informações suficientes para as previsões numéricas e classificações. As estações do ano (primavera, verão, outono, inverno) possuem aproximadamente 90 dias cada e foram selecionados para este estudo, alguns períodos. Por exemplo: Os 30 ou 15 primeiros dias, os 30 ou 15 dias no meio da estação e os 30 ou 15 últimos dias de cada estação.

As figuras de mérito escolhidas para este estudo foram o RMSEP e O R^2_{prev} , pois estas duas métricas mostram a informação do menor erro e maior coeficiente de determinação da previsão. Os melhores resultados foram exatamente no meio da estação com 30 dias corridos, isto já era de se esperar pois entre uma estação do ano e outra existe um período de transição entre as duas estações do ano e consequentemente maior incerteza neste período. Para períodos de 15 dias não foram apresentados resultados satisfatórios.

Para o outono foram considerados os períodos compreendidos entre 20 de abril até 20 de maio, para o inverno de 20 de julho até 20 de agosto, para a primavera de 22 de outubro até 22 de novembro e para o verão de 20 de janeiro até 20 de fevereiro.

Foi utilizado o melhor modelo para cada EAQA, que na maioria delas foram o RF. O ADN_ver o melhor método apresentando era o SVM, porém, neste caso, ele não pode ser validado já que o número de vetores suporte para a construção do modelo foi igual ao número de observações dos dados de calibração, o que sugere comportamento de memorização e não de aprendizado do modelo. Logo foi realizado o RF para o ADN_ver.

De acordo com a Tabela 17, As EAQA INE_pri, PDC_pri, PDC_ver e VSL_pri, apresentaram maiores RMSEP em comparação com 90 dias. As estações mais quentes do ano verão e primavera, geralmente, possuem maior dispersão dos dados (conforme apresentado no RBOPCA das Figuras, 29, 21, 22, 25 respectivamente), isso faz com que na maioria dos casos é necessário que a estação meteorológica fique uma maior quantidade de tempo possível no lugar para capturar e entender melhor estas maiores distribuições das concentrações das variáveis no tempo.

Logo poderia se pensar em colocar uma EAQA apenas em um mês específico dentro de cada estação do ano, para a coleta de dados de apenas 30 dias, porém em se tratando que os dados possuem bastantes dados faltantes dependendo da EAQA, isto deve se fazer com cautela e mais dados e mais anos seriam necessários para definir este planejamento melhor.

O uso de uma estação móvel para coletar dados em diversas localidades por um período curto de tempo, com o objetivo de calibrar o modelo de previsão, gera uma economia para os órgãos ambientais de prefeituras e estados.

Tabela 17- Comparação dos modelos com 30 dias e 90 dias para a previsão de O₃ para cada EAQA

EAQA	Métodos	RMSEP (30 dias)	RMSEP	R ² prev (30 dias)	R ² prev
INE_out	RF	3,74	4,17	0,86	0,92
INE_inv	RF	9,81	10,14	0,64	0,66
INE_pri	RF	11,92	9,52	0,56	0,79
INE_ver	RF	12,72	13,32	0,73	0,76
ADN_inv	RF	13,78	16,07	0,66	0,62
ADN_out	RF	10,76	13,52	0,71	0,64
ADN_pri	RF	11,10	13,92	0,66	0,60

EAQA	Métodos	RMSEP (30 dias)	RMSEP	R ² prev (30 dias)	R ² prev
ADN_ver	SVM	14,26	15,11	0,64	0,64
ADN_ver	RF	14,70	-	0,62	-
PDC_inv	RF	19,13	19,45	0,54	0,61
PDC_out	RF	11,26	13,49	0,67	0,64
PDC_pri	RF	17,30	11,84	0,42	0,58
PDC_ver	RF	22,36	18,14	0,55	0,54
VSL_inv	RF	8,45	11,11	0,73	0,65
VSL_out	RF	7,73	9,97	0,66	0,69
VSL_pri	RF	21,25	12,95	0,38	0,54
VSL_ver	RF	11,49	13,92	0,81	0,74

3.8. Previsão de O₃ com os meses do Covid-19

Foi realizado um estudo com dados do INEA compreendidos entre 16 de março até 12 de abril de 2020. De acordo com o decreto estadual nº 46.973 de 16 de março de 2020, ficou restrita as atividades no Estado do Rio de Janeiro, ou seja, a quantidade de fontes móveis diminuiu nas principais vias do Estado e consequentemente isto afetou na formação de O₃ na troposfera. Como os dados estavam compreendidos de 16 de março até 12 de abril de 2020, a maioria destes dados estão no período do outono, logo para manter as características de sazonalidade, os modelos validados utilizados foram no período do outono.

Os modelos já validados nos estudos anteriores para os 5 anos para cada EAQA, foram utilizados para prever com um novo conjunto de dados do ano 2020. Porém os modelos não foram aptos para prever cenários com a pandemia do Covid-19 (Tabela 18). Isto se deve ao fato que as correlações entre as variáveis mudaram. Estas “novas” correlações apresentaram, na maioria dos casos, serem mais fracas em relação ao modelo validado, o que significa que o modelo não demonstrou ser eficiente para este evento Covid-19 (entenda-se eficiente em comparação com as figuras de mérito R²prev e RMSEP).

Na construção dos modelos umas das métricas escolhidas para as escolhas das variáveis foram as correlações acima de 0,20. Para a EAQA de ADN_out_2020 não entraria a variável DV, porém na classificação Boruta (Figura 51) para a EAQA ADN_out é a segunda variável mais importante em relação as demais. Para a EAQA de PDC_out_2020 a correlação de NO muda bastante, e de acordo com a Figura 53 a variável que mais importa é a variável NO e a

variável RS também não entraria no modelo pois esta estar abaixo de 0,20. A EAQA do INE não tinha dados disponíveis nestes períodos.

Tabela 18- Comparação dos modelos já validados com os dados de Covid-19 de 2020 para a previsão de O₃

EAQA	T	HORA	NO	NO ₂	VV	UR	DV	RS	R ² prev	RMSEP
ADN_out_2020	0,34	0,29	-0,25	-0,24	0,26	-	-0,05	-	<0,01	25,87
ADN_out	0,55	0,58	-0,31	-0,20	0,23	-	0,22	-	0,64	13,52
PDC_out_2020	0,42	0,36	-0,29	-	0,27	-0,35	-	0,16	0,31	21,37
PDC_out	0,51	0,42	-0,53	-	0,22	-0,52	-	0,33	0,64	13,49
VSL_out_2020	0,63	-	-0,49	-	0,49	-0,56	-	0,42	0,38	14,98
VSL_out	0,57	-	-0,48	-	0,32	-0,50	-	0,46	0,69	9,97

4. CONCLUSÕES

As técnicas PLS, RF, SVM e ANN foram usadas para estimar a complexa relação entre o ozônio e outras variáveis com base nos dados de cada EAQA. Como era esperado, as técnicas SVM, RF e ANN tiveram mais sucesso na previsão de O_3 para todas as EAQA estudadas, pois essas técnicas são baseadas na suposição de que as interações entre as variáveis independentes e o ozônio são não lineares. Além disso, o tempo de consumo da máquina também foi um parâmetro fundamental que foi destacado neste estudo.

O SWTP provou ser um algoritmo muito útil para comparação de modelos, principalmente porque as características entre variáveis independentes e ozônio não seguem uma distribuição normal.

Este estudo sugere que as técnicas quimiométricas de RF, SVM e ANN podem ser utilizadas na previsão de concentrações de ozônio na baixa troposfera, com R^2_{prev} até 0,92 e RMSEP entre 4,17 e 22,45 $\mu\text{g m}^{-3}$. Com poucas variáveis foi possível realizar uma boa previsão, sem a necessidade de ter uma estação meteorológica e da qualidade do ar com todas as instrumentações completas.

Em particular, este estudo sublinhou o potencial das ferramentas de imputação, como a aplicação do *MissForest* para capturar as interações não lineares entre o ozônio e outras variáveis. Além disso, os resultados forneceram suporte para o fato de que os padrões de variabilidade geralmente estão associados à interação de dados meteorológicos de nível regional.

A alternativa levantada por este estudo era a viabilidade de usar EAQA móvel para armazenar dados de diferentes locais e utilizar as ferramentas descritas para prever ozônio ou outros poluentes, ou mesmo eventos críticos, evitando assim a necessidade de instalação de EAQA dispendiosas em diferentes locais da cidade. Porém como existem quantidades de dados faltantes relevantes em determinadas EAQA, é necessário investigações aprofundadas para a mitigação destes dados ausentes, para posteriormente se pensar em criar um planejamento estratégico desta magnitude ocorra de forma eficaz.

REFERÊNCIAS

- ABDI, H. E.; WILLIAMS, L. J. Principal component analysis. *WIREs Comp Stat*, v. 2, n. 4, p. 433–459, 2010a.
- ABDI, H.; WILLIAMS, L. J. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, v. 2, n. 4, p. 433–459, 2010b.
- ABDUL-WAHAB, S. A.; BAKHEIT, C. S.; AL-ALAWI, S. M. Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. *Environmental Modelling & Software*, v. 20, n. 10, p. 1263–1271, out. 2005. Disponível em: <<https://linkinghub.elsevier.com/retrieve/pii/S1364815204002129>>.
- ABYANEH, Z. H. Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters. *J Environ Health Sci Eng*, v. 12, p. 40, 2014.
- ALIMISSIS, A. et al. Spatial estimation of urban air pollution with the use of artificial neural network models. *Atmospheric Environment*, v. 191, p. 205–213, 2018.
- AMJAD, A. et al. Raman spectroscopy based analysis of milk using random forest classification. *Vibrational Spectroscopy*, v. 99, p. 124–129, 2018.
- ARROYO, Á. et al. Neural models for imputation of missing ozone data in air-quality datasets. *Complexity*, v. 2018, 2018.
- ASTM E1655-05. Standard Practices for Infrared Multivariate Quantitative Analysis. *ASTM International*, v. 05, n. Reapproved 2012, p. 29, 2012.
- ATKINSON, R. Atmospheric chemistry of VOCs and NOx. *Atmospheric Environment*, v. 34, p. 2063–2101, 2000.
- AZID A, JUAHIR H, TORIMAN M, KAMARUDIN M, SAUDI A, H. C. Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: a case study in Malaysia. *Water Air Soil Pollut*, v. 225, p. 1–14, 2014.
- BALLABIO, D.; CONSONNI, V. Classification tools in chemistry. Part 1: linear models. PLS-DA. *Analytical Methods*, v. 5, n. 16, p. 3790–3798, 2013.
- BARRY M. WISE, NEAL B. GALLAGHER, RASMUS BRO, JEREMY M. SHAVER, WILLEM WINDIG, R. S. K. *Chemometrics tutorial for PLS toolbox and solo*. Eigenvector Research, Inc., Wenatchee. [s.l.] Eigenvector Research, Inc., 2006.
- BRASSEUR, G. P.; ORLANDO, J. J.; TYNDAL, G. S. *Atmospheric chemistry and global Change*. [s.l.] Oxford: Oxford University Press, 1999.
- BREIMAN, L. Random forests. *Machine Learning*, v. 45, p. 5–32, 2001.
- BRERETON, R. G. Signal Processing. In: *Chemometrics*. Chichester, UK: John Wiley & Sons, Ltd, 2003. 8p. 119–181.
- CARSLAW, D. C.; ROPKINS, K. openair - an R package for air quality data analysis. *Environmental Modelling & Software*, v. 27–28, p. 52–61, 2012.

- CETESB. *Relatório de qualidade do ar do estado de São Paulo 2014*. [s.l: s.n.]
- CHEN, X.; ISHWARAN, H. Random forests for genomic data analysis. *Genomics*, v. 99, n. 6, p. 323–329, 2012.
- CORDELLA, C. B. Y. PCA: The Basic Building Block of Chemometrics. *Analytical chemistry*, n. December, p. 1–46, 2012.
- CORNELL, J. A. Factors that Influence the Value of the Coefficient of Determination in Simple Linear and Nonlinear Regression Models. *Phytopathology*, v. 77, n. 1, p. 63, 1987.
- CORRÊA, S. M. *Qualidade do ar da cidade do Rio de Janeiro: Sinergia entre simulação e monitoramento*. 2003. Instituto de Química da Universidade Federal do Rio de Janeiro, 2003.
- COVER, T. AND HART, P. *Nearest neighbor pattern classification*. [s.l.] IEEE Transactions on Information Theory, 1967.
- COX, D. R.; EFRON, B. Statistical thinking for 21st century scientists. *Science Advances*, v. 3, n. 6, p. 1–6, 2017.
- CUNHA, C. L.; TORRES, A. R.; LUNA, A. S. Multivariate regression models obtained from near-infrared spectroscopy data for prediction of the physical properties of biodiesel and its blends. *Fuel*, v. 261, n. June, 2020.
- CUTLER, D. R. et al. Random Forests for Classification in Ecology. *Ecology*, v. 88, n. 11, p. 2783–2792, 2007.
- DA SILVA, C. M. et al. Speciation analysis of ozone precursor volatile organic compounds in the air basins of the Rio de Janeiro metropolitan area. *Revista Virtual de Química*, v. 9, n. 5, p. 1887–1909, 2017.
- DA SILVA, C. M. et al. A minimum set of ozone precursor volatile organic compounds in an urban environment. *Atmospheric Pollution Research*, v. 9, n. 2, p. 369–378, 2018.
- DOMINICK, D., JUAHIR, H., LATIF, M. T., ZAIN, S. M., & ARIS, A. Z. Spatial assessment of air quality patterns in Malaysia using multivariate analysis. *Atmospheric Environment*, v. 60, p. 172–181, 2012.
- EMBERSON, L. D. et al. Ozone effects on crops and consideration in crop models. *European Journal of Agronomy*, v. 100, n. June, p. 19–34, 2018.
- FARHANGFAR, A.; KURGAN, L. A.; PEDRYCZ, W. A novel framework for imputation of missing values in databases. *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans*, v. 37, n. 5, p. 692–709, 2007.
- FENG, Y. et al. Ozone concentration forecast method based on genetic algorithm optimized back propagation neural networks and support vector machine data classification. *Atmospheric Environment*, v. 45, n. 11, p. 1979–1985, 2011.
- FERREIRA, M. M. C. *Quimiometria: Conceitos, Métodos e Aplicações*. 1 ed. ed. Campinas, SP, Brasil: Editora da Unicamp, 2015.
- FERRER-RIQUELME, A. J. Statistical Control of Measures and Processes. In: *Comprehensive Chemometrics*. [s.l.] Elsevier, 2009. 1p. 97–126.

- FIGLIORE, A. M. et al. Linking ozone pollution and climate change: The case for controlling methane. *Geophysical Research Letters*, v. 29, n. 19, p. 2–5, 2002.
- FIGLIORE, D.V.; MARTINS, F.R.; SCHUCH, N.J.; PEREIRA, E. . Aplicações de redes neurais e previsões de disponibilidade de recursos energéticos solares. *Rev. Bras. Ens. Fis*, v. 33, p. 1309–1320, 2011.
- FISHMAN, J. THE GLOBAL CONSEQUENCES OF INCREASING TROPOSPHERIC OZONE CONCENTRATIONS. *Chemosphere*, v. 22, n. 7, p. 685–695, 1991.
- FUHRER, J.; SKÄRBY, L.; ASHMORE, M. R. Critical levels for ozone effects on vegetation in Europe. *Environmental Pollution*, v. 97, n. 1–2, p. 91–106, 1997.
- GARDNER, M. .; DORLING, S. . Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment*, v. 32, n. 14–15, p. 2627–2636, ago. 1998.
- GERALDINO, C. G. P. et al. Understanding high tropospheric ozone episodes in Bangu, Rio de Janeiro, Brazil. *Environmental Monitoring and Assessment*, v. 192, n. 3, 2020a.
- GERALDINO, C. G. P. et al. Understanding high tropospheric ozone episodes in Bangu, Rio de Janeiro, Brazil. *Environmental Monitoring and Assessment*, v. 192, n. 3, p. 156, mar. 2020b.
- GERY, M.; CROUSE, R. R. User's Guide for executing OZIPR. In: *U.S. Environmental Protection Agency*. [s.l: s.n.]
- GIODA, A. et al. Understanding ozone formation at two islands of Rio de Janeiro, Brazil. *Atmospheric Pollution Research*, v. 9, n. 2, p. 278–288, 2017.
- GUERRA, F. P.; MIRANDA, R. M. Influence of meteorology in the concentration of atmospheric pollutant PM_{2.5} in RJRM and MRSP. In: Proceedings of II Congress Brazilian Environmental Management, Parana, Brazil, *Anais...*2011.
- HAGAN, M.T.; DEMUTH, H. B.; BEALE, M. . *Neural Network Design*. [s.l.] Pws Pub, Boston, USA, 1996.
- HOLMES, N. S.; MORAWSKA, L. A review of dispersion modeling and its application to the dispersion of particles: an overview of different dispersion models available. *Atmospheric Environment*, v. 40, p. 5902–5928, 2006.
- IBGE. *Panorama das Cidades*. [s.l: s.n.]. Disponível em: <<https://agenciadenoticias.ibge.gov.br/agencia-sala-de-imprensa/2013-agencia-de-noticias/releases/25278-ibge-divulga-as-estimativas-da-populacao-dos-municipios-para-2019>>.
- INEA. *Relatório da Qualidade do Ar do Estado do Rio de Janeiro - Anos Base 2014*. [s.l.] Instituto Estadual do Ambiente., 2015. .
- JONES, A. M.; HARRISON, R. M.; BAKER, J. The wind speed dependence of the concentrations of airborne particulate matter and NO_x. *Atmospheric Environment*, v. 44, n. 13, p. 1682–1690, 2010.
- JUNNINEN, H. et al. Methods for imputation of missing values in air quality data sets.

Atmospheric Environment, 2004.

- KENNARD, R. W.; STONE, L. A. Computer Aided Design of Experiments. *Technometrics*, v. 11, n. 1, p. 137–148, 1969.
- KLEY, D., KLEINMANN, M., SANDERMAN, H., KRUPA, S. Photochemical oxidants: state of the science. *Environmental Pollution*, v. 100, p. 19–42, 1999.
- KRZANOWSKI, W. J. Croos-validation in principal component analysis. *Biometrics*, v. 43, p. 575–584, 1987.
- KUCHERYAVSKIY, S. *Mdatools: Multivariate Data Analysis for Chemometrics, R package version 0.7.0.*, 2015. .
- LIAW, A.; WIENER, M. Classification and Regression by randomForest. *R News*, v. 2, n. 3, p. 18–22, 2002.
- LIPPMANN, M. Health Effects of tropospheric:OZONE. *Environmental Science and Technology*, v. 25, n. 12, p. 1954–1962, 1991.
- LITTLE, R. J. A.; RUBIN, D. B. *Statistical analysis with missing data*. 2nd editio ed. New York: Wiley, 2014.
- LIU, W.; LIU, C.; YU, J.; ZHANG, Y.; LI, J.; CHEN, Y.; ZHENG, L. Discrimination of geographical origin of extra virgin olive oils using terahertz spectroscopy combined with chemometrics. *Food Chemistry*, v. 251, p. 86–92, 2018.
- LORA, E. E. S. *Prevenção e Controle da Poluição nos Setores Energéticos, Industrial e de Transporte*. [s.l.] Interciências, 2002.
- LUNA, A. S. et al. Prediction of ozone concentration in tropospheric levels using artificial neural networks and support vector machine at Rio de Janeiro, Brazil. *Atmospheric Environment*, v. 98, p. 98–104, 2014.
- MA S, A. N. Efficient optimization algorithms for robust principal component analysis and its variants. *Proceedings of the IEEE*, v. 106, p. 1411–1426, 2018.
- MALBY, A. R.; WHYATT, J. D.; TIMMIS, R. J. Conditional extraction of air–pollutant source signals from air quality monitoring. *Atmospheric Environment*, v. 74, p. 112–122, 2013.
- MARTENS, H.; NAES, T. Multivariate calibration. *Spectrochimica acta Part A: Molecular and biomolecular spectroscopy*, v. 44, p. 287–321, 1989.
- MARTINS, E. M.; MEIRELES, A. R.; MAGALHAES, F. R.; CARVALHO, J. B. B.; RIBEIRO, M. M. Concentrações de poluentes atmosféricos no Rio de Janeiro em relação a normas nacionais e internacionais. *Revista Internacional de Ciências*, v. 7, n. 1, p. 32–48, 2017.
- MARTINS, P. S. *Imputação de dados faltantes*. 2017. Universidade Federal Fluminense, 2017. Disponível em: <<https://app.uff.br/riuff/handle/1/14853>>.
- MEVIK, B. H.; CEDERKVIST, H. R. Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR). *Journal of Chemometrics*, v. 18, n. 9, p. 422–429, 2004.

- MEVIK, B. H.; WEHRENS, R.; LILAND, K. H. *pls: Partial Least Squares and Principal Component Regression. R package version 2.5-0*, 2015. .
- MEYER, D. et al. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. *R Package Version 17-0*, 2018.
- MISZTAL, M. Some Remarks on the Data Imputation Using “missForest” Method. *Acta Universitatis Lodziensis. Folia Oeconomica*, v. 285 Multiv, p. 169–179, 2013.
- NAGUIB, I. A.; DARWISH, H. W. Support vector regression and artificial neural network models for stability indicating analysis of mebeverine hydrochloride and sulpiride mixtures in pharmaceutical preparation: a comparative study. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, v. 86, p. 515–526, 2012.
- NAJAFPOOR AA, HOSSEINZADEH A, ALLAHYARI S, JAVID AB, E. H. Modeling of CO and NOx produced by vehicles in Mashhad. *Environ Health Eng Manag*, v. 1, p. 45–49, 2014.
- NISHANTH, T.; SATHEESH KUMAR, M. K.; VALSARAJ, K. T. Variations in surface ozone and NOx at Kannur: A tropical, coastal site in India. *Journal of Atmospheric Chemistry*, v. 69, n. 2, p. 101–126, 2012.
- OLIVEIRA, R. C. G. et al. A simulation study about the impact of biodiesel use on the atmosphere of Rio de Janeiro city. *Brazilian Journal of Chemical Engineering*, v. 34, n. 3, p. 727–738, 2017.
- ORLANDO, J. P. et al. Ozone precursors for the São Paulo Metropolitan Area. *Science of The Total Environment*, v. 408, n. 7, p. 1612–1620, mar. 2010. Disponível em: <<https://linkinghub.elsevier.com/retrieve/pii/S0048969709011759>>.
- PACK, S. Factor analysis in chemistry, (2nd edition), E. R. Malinowski, Wiley-Interscience, 1991. ISBN 0-471-53009-3. Price £43.70. *Journal of Chemometrics*, v. 5, n. 6, p. 545–545, 1991.
- PLATT, L. SCHÜLKOPF, B.; BURGESS, B.; SMOLA, A. Fast training of SVM using sequential optimization. *Advances in Kernel Methods-support Vector Learning. Cambridge*, p. 185–208, 1998.
- PYTHON SOFTWARE FOUNDATION. Python Software. n. August, p. 5–6, 2016. Disponível em: <<http://www.python.org>>.
- RUBIN, D. B. Biometrika Trust Inference and Missing Data Author (s): Donald B . Rubin Published by : Oxford University Press on behalf of Biometrika Trust Stable URL : <http://www.jstor.org/stable/2335739> Accessed : 12-06-2016 21 : 34 UTC. *Biometrika*, v. 63, n. 3, p. 581–592, 1976.
- SANCHEZ-CCOYLLO, O. R.; MARTINS, L. D.; YNOUE, R. Y.; ANDRADE, M. F. Impacts of ozone precursor limitation and meteorological variables on ozone concentration in São Paulo. *Atmospheric Environment*, v. 40, p. 552–562, 2006.
- SÁNCHEZ-CCOYLLO, O. R.; MARTINS, L. D.; YNOUE, R. Y.; ANDRADE, M. F. The impact on tropospheric ozone formation on the implementation of a program for mobile emissions control: a case study in São Paulo, Brazil. *Environmental Fluid Mechanics*,

- v. 7, p. 95–119, 2007.
- SCHÖLKOPF, B.; SMOLA, A. J. *Learning with Kernels*. [s.l.] Cambridge: MIT Press, MA, 2002.
- SCHUCH, D. et al. A two decades study on ozone variability and trend over the main urban areas of the São Paulo state, Brazil. *Environmental Science and Pollution Research*, v. 26, n. 31, p. 31699–31716, 2019.
- SEINFELD, J. H.; PANDIS, S. N. *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*. 3rd. ed. [s.l.] Wiley, 2016.
- SEKAR, C. et al. Modeling and prediction of hourly ambient ozone (O₃) and oxides of nitrogen (NO_x) concentrations using artificial neural network and decision tree algorithms for an urban intersection in India. *Journal of Hazardous, Toxic, and Radioactive Waste*, v. 20, n. 4, 2016.
- SERGUEL, R. J.; MORALES, R. G. E.; LEIVA, M. . Ozone weekend effect in Santiago, Chile. *Environ. Pollut.*, v. 162, p. 72–79, 2012.
- SHAHBAZI, B.; CHEHREH CHELGANI, S.; MATIN, S. S. Prediction of froth flotation responses based on various conditioning parameters by Random Forest method. *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, v. 529, p. 936–941, 2017.
- SILVA, C. M. *Estudo dos compostos orgânicos voláteis precursores de ozônio para a região metropolitana do Rio de Janeiro*. 2016. Universidade do Estado do Rio de Janeiro, 2016.
- SILVERSTONE, P. National oceanic and atmospheric administration (noaa). In: *The Navy of the Nuclear Age, 1947-2007*. [s.l.: s.n.]p. 311–316.
- SINGH, S., SARIN, M., SANMUGAM, P., SHARMA, N. Ozone distributions in the urban environment of Delhi during winter months. *Atmospheric Environment*, v. 31, n. 20, p. 3421–3427, 1997.
- SONG, F. et al. Relationships among the springtime ground-level NO_x, O₃ and NO₃ in the vicinity of highways in the US East Coast. *Atmospheric Pollution Research*, v. 2, n. 3, p. 374–383, 2011.
- SOUSA, S. I. V. et al. Prediction of ozone concentrations in Oporto city with statistical approaches. *Chemosphere*, v. 64, n. 7, p. 1141–1149, 2006.
- SOUSA, S. I. V. et al. Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. *Environmental Modelling and Software*, v. 22, n. 1, p. 97–103, 2007.
- STÄDLER, N.; STEKHOVEN, D. J.; BÜHLMANN, P. Pattern alternating maximization algorithm for missing data in high-dimensional problems. *Journal of Machine Learning Research*, v. 15, p. 1903–1928, 2014.
- STEKHOVEN, D. J. missForest: Nonparametric Missing Value Imputation using Random Forest. *R package version 1.4*, 2013.
- STEKHOVEN, D. J.; BÜHLMANN, P. Missforest-Non-parametric missing value imputation

- for mixed-type data. *Bioinformatics*, 2012.
- STEVENS, A.; RAMIREZ-LOPEZ, L. *An introduction to the prospectr package* R package *Vignette* R package version 0.1.3, , 2013. .
- STONE M. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, v. 36, n. 2, p. 111–147, 1974.
- TAMAS, W.; NOTTON, G.; PAOLI, C.; VOYANT, C.; NIVET, M.; BALU, A. Urban ozone concentration forecasting with artificial neural network in Corsica. *Mathematical modeling in Civil Engineering*, v. 10, p. 1–9, 2014.
- TANASKULI, M. et al. Ozone prediction based on support vector machine. *Indonesian Journal of Electrical Engineering and Computer Science*, v. 17, n. 3, p. 1461–1466, 2019.
- TEAM, R. C. *A language and environment for statistical computing*. [s.l.] . R Foundation for Statistical Computing, Vienna, Austria, 2022.
- TEIXEIRA, E. B. et al. Measurement of surface ozone and its precursors in an urban area in South Brazil. *Atmospheric Environment*, v. 43, n. 13, p. 2213–2220, 2009.
- TODOROV, V.; FILZMOSER, P. An object-oriented framework for robust multivariate analysis. *J Stat Softw*, v. 32, p. 1–47, 2009.
- TROYANSKAYA, O., CANTOR, M., SHERLOCK, G., BROWN, P., HASTIE, T., TIBSHIRANI, R., BOTSTEIN, D., AND ALTMAN, R. Missing value estimation methods for DNA microarrays. *Bioinformatics*, v. 17, p. 520–525, 2001.
- TYRALIS, H.; PAPACHARALAMPOUS, G.; LANGOUSIS, A. A brief review of random forests for water scientists and practitioners and their recent history in water resources. *Water (Switzerland)*, v. 11, n. 5, 2019.
- U.S. EPA. *Guideline for Developing an Ozone Forecasting Program*. *Environmental Protection Agency*. [s.l.] EPA-454/R-99-009, 2013. .
- U.S.EPA. *Method TO-15A - Determination of Volatile Organic Compounds (VOCs) in Air Collected in Specially Prepared Canisters and Analyzed by Gas Chromatography–Mass Spectrometry (GC-MS)*. [s.l.: s.n.]. Disponível em: <https://www.epa.gov/sites/default/files/2019-12/documents/to-15a_vocs.pdf>.
- ULKE, A. G.; ANDRADE, M. F. Modeling urban air pollution in São Paulo sensitivity of model predicted concentrations to different turbulence parameterizations. *Atmospheric Environment*, v. 35, p. 1747–1763, 2001.
- VAN BUUREN, S.; OUDSHOORN, K. *Flexible multivariate imputation by MICE* TNO *Prevention and Health* Leiden, The Netherlands. [s.l.] TNO Prevention Center, 1999. .
- VAPNIK, V. N. *Statistical Learning Theory*. [s.l.] Huddersfield: John Wiley and Sons, 1998.
- VAPNIK, V. N.; CHERVONENKIS, A. Y. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. In: *Measures of Complexity*. Cham: Springer International Publishing, 2015. 16p. 11–30.
- VENTURA L., M. B. R.; SANTOS, J. O.; GIODA, and A. Monitoring of air quality before the Olympic Games Rio 2016. *Annals of the Brazilian Academy of Sciences*, v. 91, 2019.

- WANG, W.; MEN, C.; LU, W. Online prediction model based on support vector machine. *Neurocomputing*, v. 71, p. 550–558, 2008.
- WANG, Y., HUANG, H. Y., ZUO, Z. T., & WANG, Y. Z. Comprehensive quality assessment of *Dendrobium officinale* using ATR-FTIR spectroscopy combined with random forest and support vector machine regression. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, v. 205, p. 637–648, 2018.
- WANG, X. et al. A study of ozone variation trend within area of affecting human health in Hong Kong. *Chemosphere*, v. 52, n. 9, p. 1405–1410, 2003.
- WARREN S. MCCULLOCH & WALTER PITTS. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, v. 5, p. 115–133, 1943.
- WOLD, H. Soft modelling: the basic design and some extensions. *Systems under indirect observation: Causality-structure-prediction*, p. 1–54, 1982.
- WOLD, S. Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models. *Technometrics*, v. 20, n. 4, p. 397–405, 1978.
- YEGANEH, B.; MOTLAGH, M. S. P.; RASHIDI, Y.; KAMALAN, H. Prediction of CO concentrations based on a hybrid Partial Least Square and Support Vector Machine model. *Atmospheric Environment*, v. 55, p. 357–365, 2012.
- ZHANG, H. et al. Prediction of soil organic carbon in an intensively managed reclamation zone of eastern China: A comparison of multiple linear regressions and the random forest model. *Science of the Total Environment*, v. 592, p. 704–713, 2017.
- ZHANG, X. et al. Metabolite profiling for model cultivars of wheat and rice under ozone pollution. *Environmental and Experimental Botany*, v. 179, n. May, p. 104214, 2020.
- ZHAO, Y.; HASAN, Y. A. Comparison of three classification algorithms for predicting pm2.5 in Hong Kong rural area. *Journal of Asian Scientific Research*, v. 3, p. 715–728, 2013.

Parte desta tese foi publicada na Revista *Environmental Monitoring and Assessment* em julho de 2021

Environ Monit Assess (2021) 193:531
<https://doi.org/10.1007/s10661-021-09333-2>



Forecasts of tropospheric ozone in the Metropolitan Area of Rio de Janeiro based on missing data imputation and multivariate calibration techniques

Rafael C. G. de Oliveira · Camilla L. Cunha · Alexandre R. Tôres · Sergio M. Corrêa

Received: 29 March 2021 / Accepted: 22 July 2021

© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2021

Abstract Multivariate calibration based on partial least squares, random forest, and support vector machine methods, combined with the MissForest imputation algorithm, was used to understand the interaction between ozone and nitrogen oxides, carbon monoxide, wind speed, solar radiation, temperature, relative humidity, and others, the data of which were collected by air quality monitoring stations in the metropolitan area of Rio de Janeiro in four distinct sites between, 2014 and, 2018. These techniques provide an easy and feasible way of modeling and analyzing air pollutants and can be used when coupled with other methods. The results showed that random forest and support vector machine chemometric techniques can be used in modeling and predicting tropospheric ozone concentrations, with a coefficient of determination for making predictions up to 0.92, a root-mean square error of calibration between 4.66

and 27.15 $\mu\text{g m}^{-3}$, and a root-mean square error of prediction between 4.17 and 22.45 $\mu\text{g m}^{-3}$, depending on the air quality monitoring stations and season.

Keywords Ozone · Missing data · MissForest · Support vector machine · Random forest · Wilcoxon test

Introduction

The atmosphere of a city is the outcome of several factors, such as emissions from mobile and stationary sources, meteorology, and topography, as well as the physical and chemical transformations of primary into secondary pollutants (Orlando et al., 2010).

Knowledge about air quality is a basic condition for ensuring good public policies for its control and improvement and, hence, for providing the public with a better quality of life. Through a knowledge of the interaction of the air pollutants, it is possible to determine the degree of control and the resources needed to mitigate the impacts of air pollution on the environment and human health (INEA, 2015).

Tropospheric ozone (O_3) has a negative impact on the climate (Fiore et al., 2002; Fishman, 1991), vegetation (Emberson et al., 2018; Fuhrer et al., 1997; Zhang et al., 2020), and the environment. It is responsible for climate change because it is a “greenhouse” gas (Sousa et al., 2006; Wang et al., 2003a) and hence affects public health (Lippmann, 1991; Wang

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1007/s10661-021-09333-2>.

R. C. G. de Oliveira · S. M. Corrêa (✉)
 Faculty of Engineering, Rio de Janeiro State University,
 Rua São Francisco Xavier, 524 Maracanã, Rio de Janeiro,
 RJ 20551-013, Brazil
 e-mail: sergiomc@uerj.br

C. L. Cunha · A. R. Tôres · S. M. Corrêa
 Faculty of Technology, Rio de Janeiro State University,
 Rodovia Presidente Dutra km 298, Resende,
 RJ 27537-000, Brazil

Published online: 28 July 2021

Springer

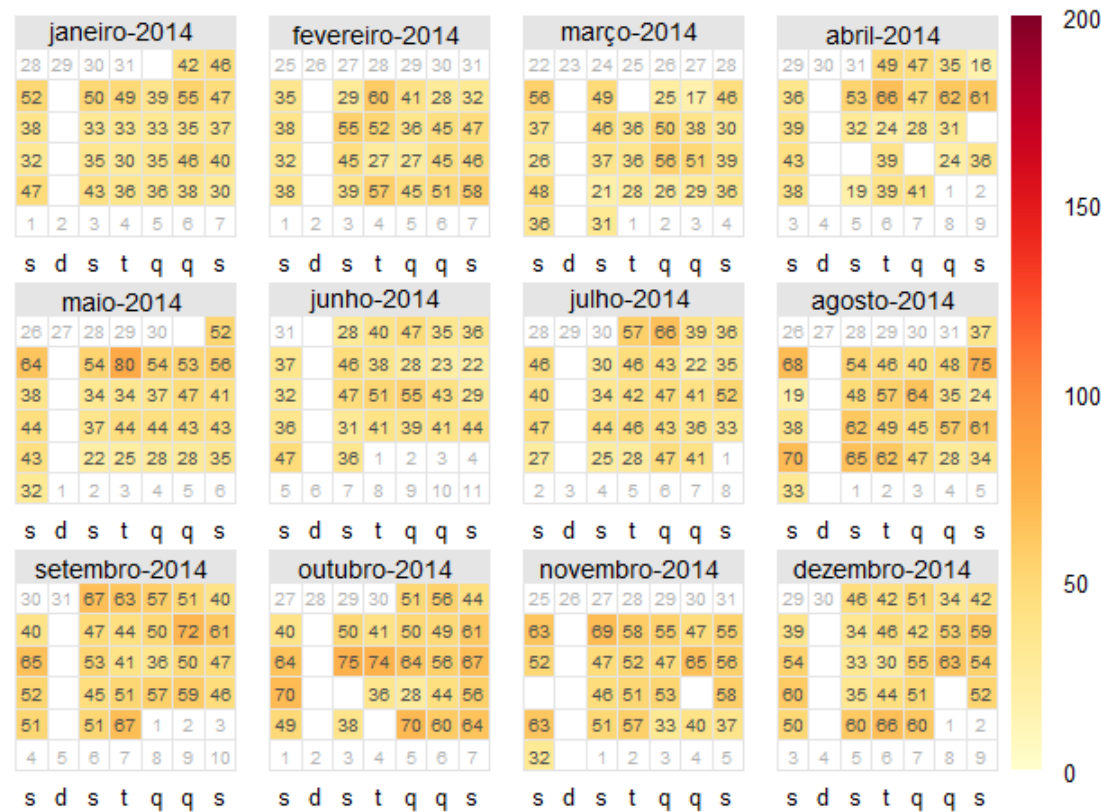
APÊNDICE A: TABELA A- LISTA DAS EAQA ESTUDADAS NA TESE

Variáveis	DV	MP ₁₀	PP	RS	TEMP	UR	VV	SO ₂	NO ₂	NO	CO	O ₃
legenda - Bac. I	(°)	(µg m ³)	(mm)	(W/m ²)	(°C)	(%)	(m/s)	(µg m ³)	(µg m ³)	(µg m ³)	(µg m ³)	(µg m ³)
legenda- Bac. II												
legenda- Bac. III												
legenda- Bac. IV												
Est. \ anos	4 5 6 7 8	4 5 6 7 8	4 5 6 7 8	4 5 6 7 8	4 5 6 7 8	4 5 6 7 8	4 5 6 7 8	4 5 6 7 8	4 5 6 7 8	4 5 6 7 8	4 5 6 7 8	4 5 6 7 8
Sambaetiba	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x
Coroa grande	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x		x x x	x x x		x x x
Eng. Pedreira	x x x x x	x x x x x			x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x
jardim Guandú	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x
Itacuruça		x x x x x							x x x	x x x		x x x
Adalgisa Nery	x x x x x	x x x x x			x x x x x		x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x
Ilha de Paquetá	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x
Ilha do Governador	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x
Piranema	x x x x x	x x x x x			x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x
Largo do Bodegão	x x x x x	x x x x x			x x x x x		x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x
Guapimirim	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x					
Brisamar		x x x x x										
Sítio Terezinha		x x x										
Vila Aparecida		x x x										
Vila Califórnia		x x x										
Auto do Jacú		x x x x x										
Fazenda Severina	x x x x x		x x x x x	x x x x x	x x x x x	x x x x x	x x x x x		x x x x x	x x x x x	x x x x x	x x x x x
Fazenda Aires	x x x x x		x x x x x	x	x x x x x	x x x x x	x x x x x		x x x x x	x x x x x	x x x x x	x x x x x
Cabiúnas	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x
Pesagro	x x x x x		x x x x x	x x x x x	x x x x x	x x x x x	x x x x x		x x x x x	x x x x x	x x x x x	x x x x x
Val Palmas	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x	x x x x x		x	x		x
Macuco		x x x x x										
Euclidelândia		x x x x x			x x x x x	x						

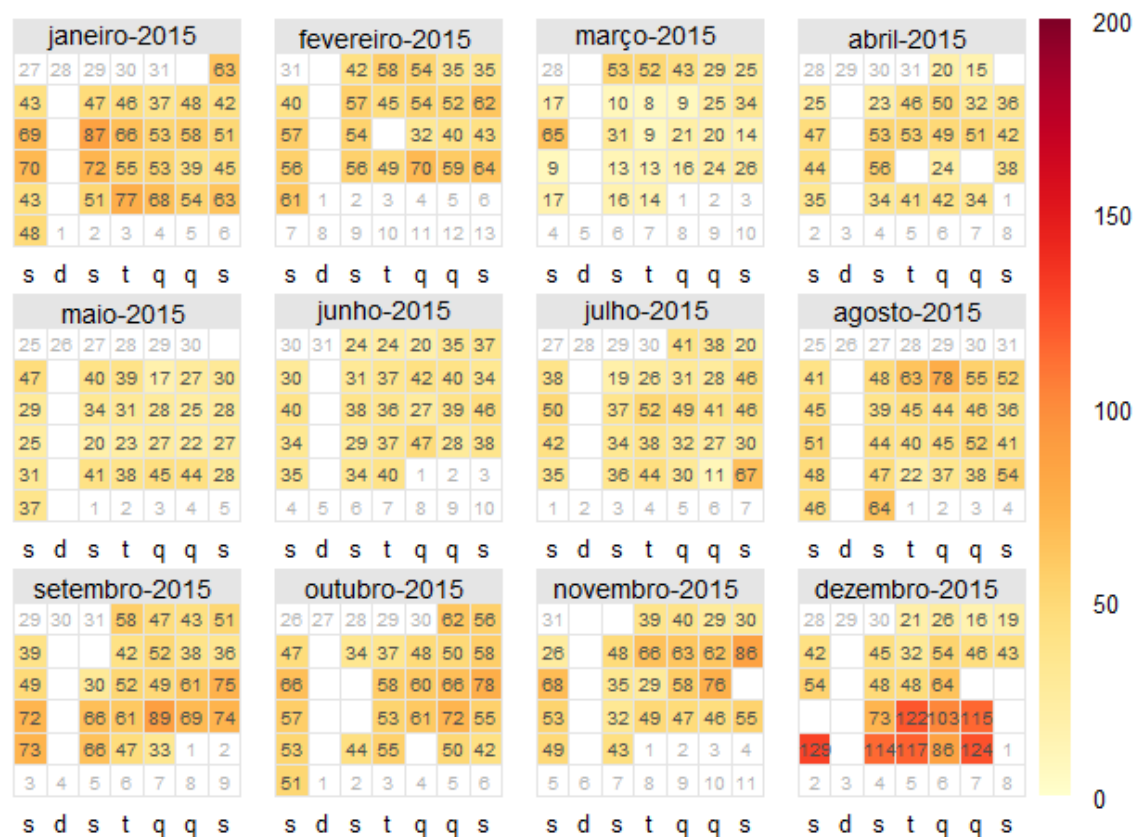
[illegible]

APÊNDICE B: TABELA B- Gráfico de Calendário para todas as EAQA

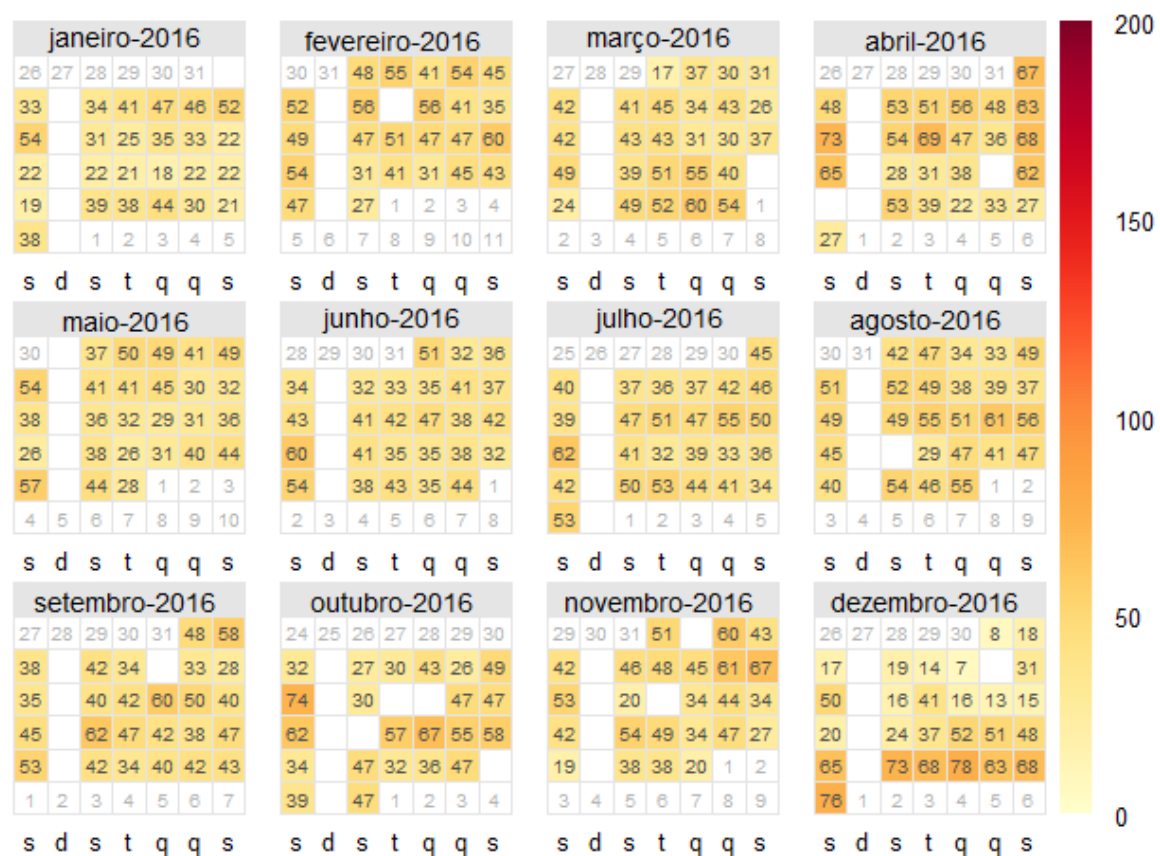
ANEXO B: Tabela B- Gráfico de Calendário para ADN em 2014

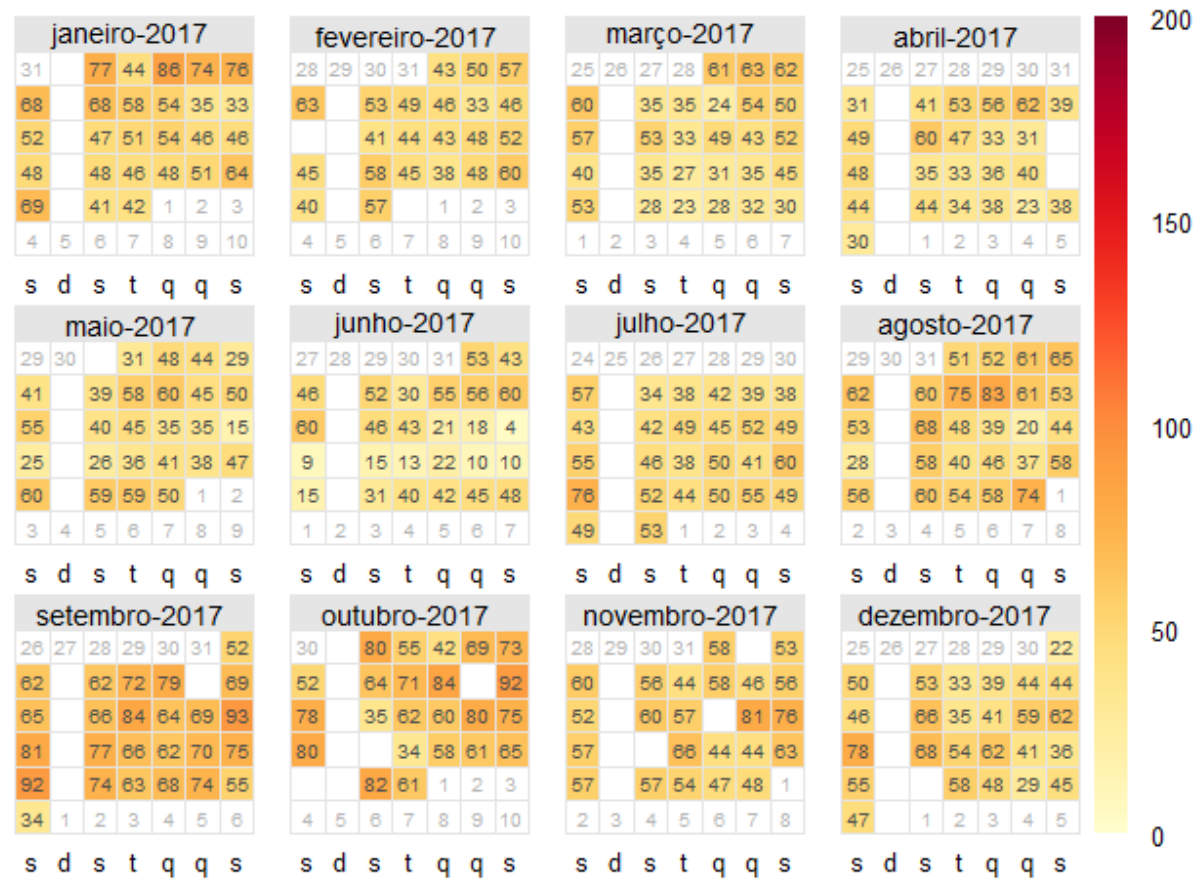


ANEXO B: Tabela B- Gráfico de Calendário para ADN em 2015 – continuação

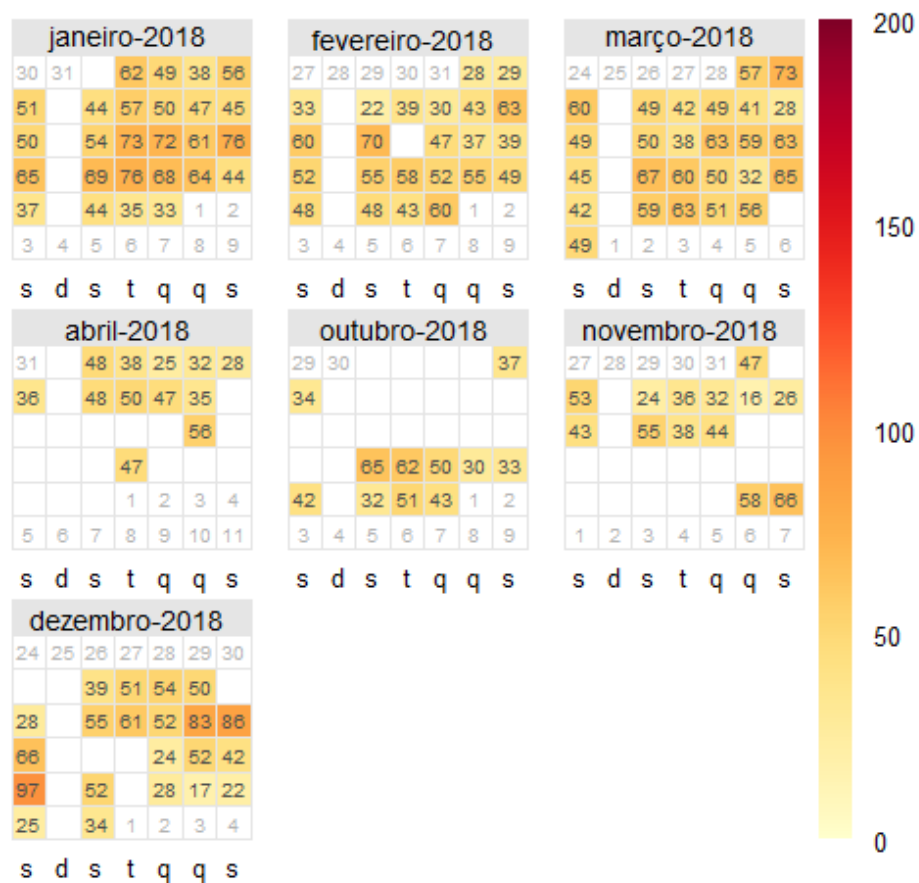


ANEXO B: Tabela B- Gráfico de Calendário para ADN em 2016 – continuação

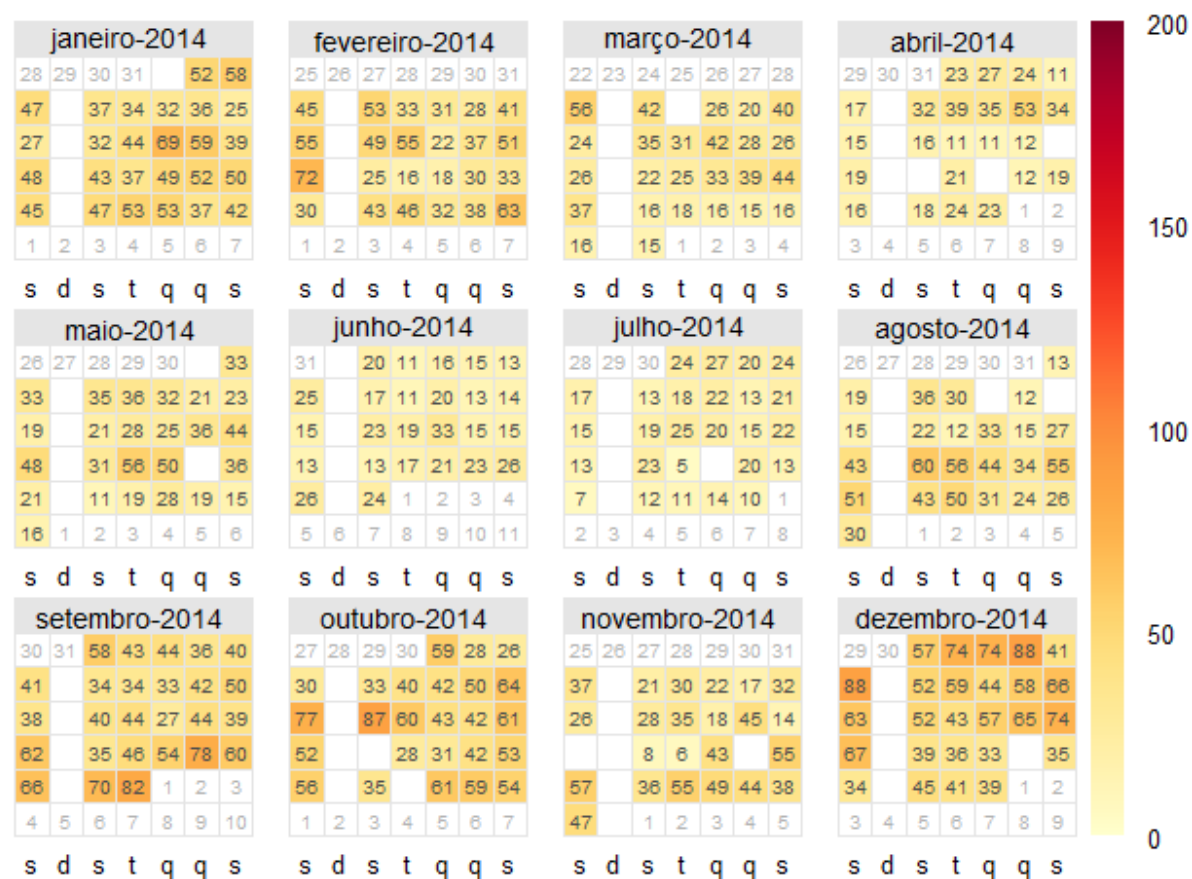


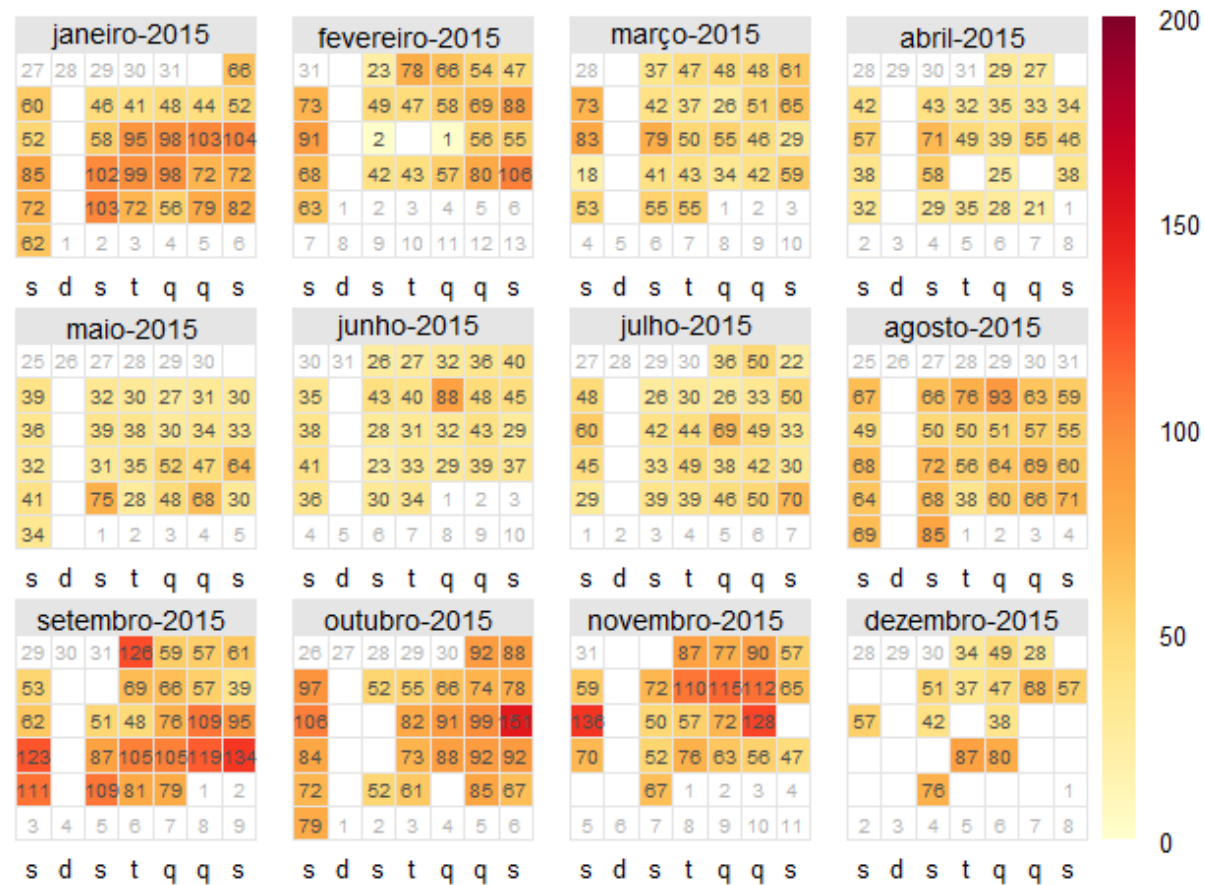


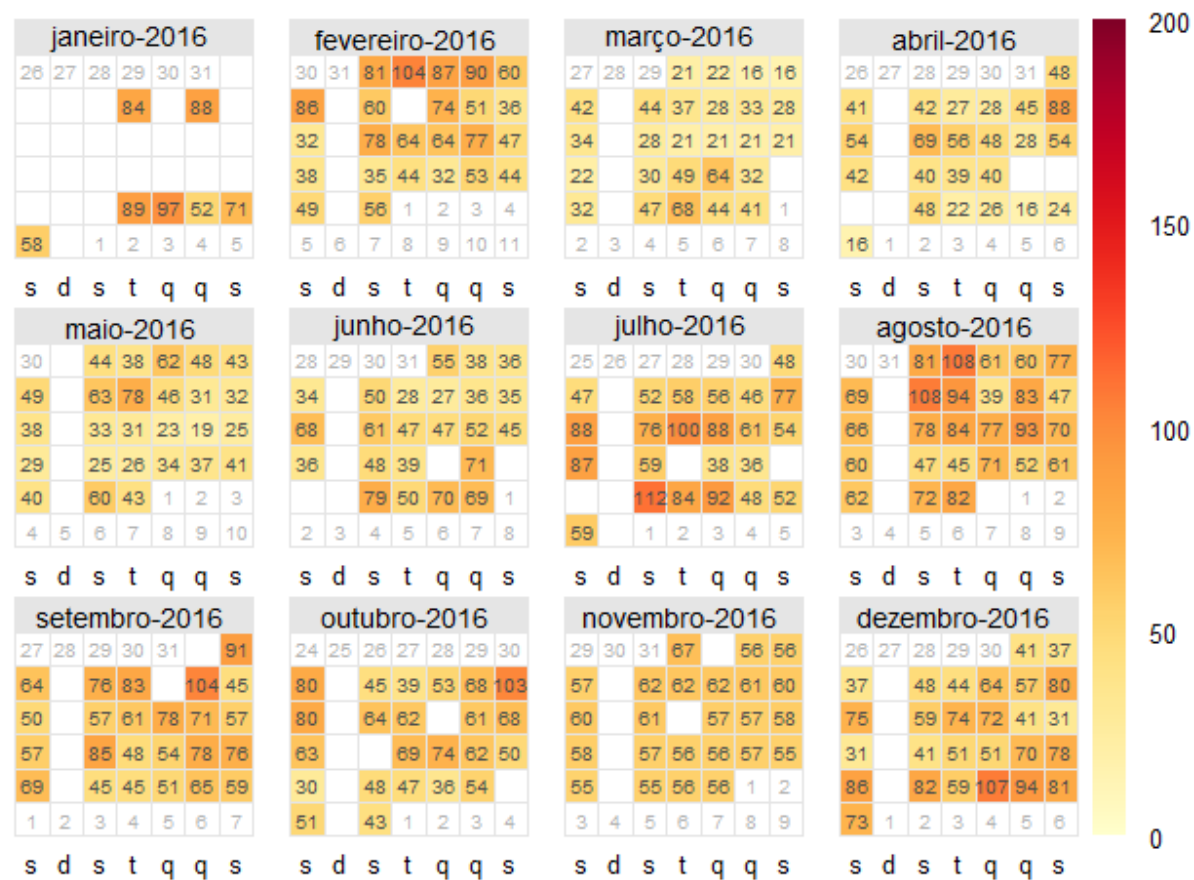
ANEXO B: Tabela B- Gráfico de Calendário para ADN em 2018 – continuação



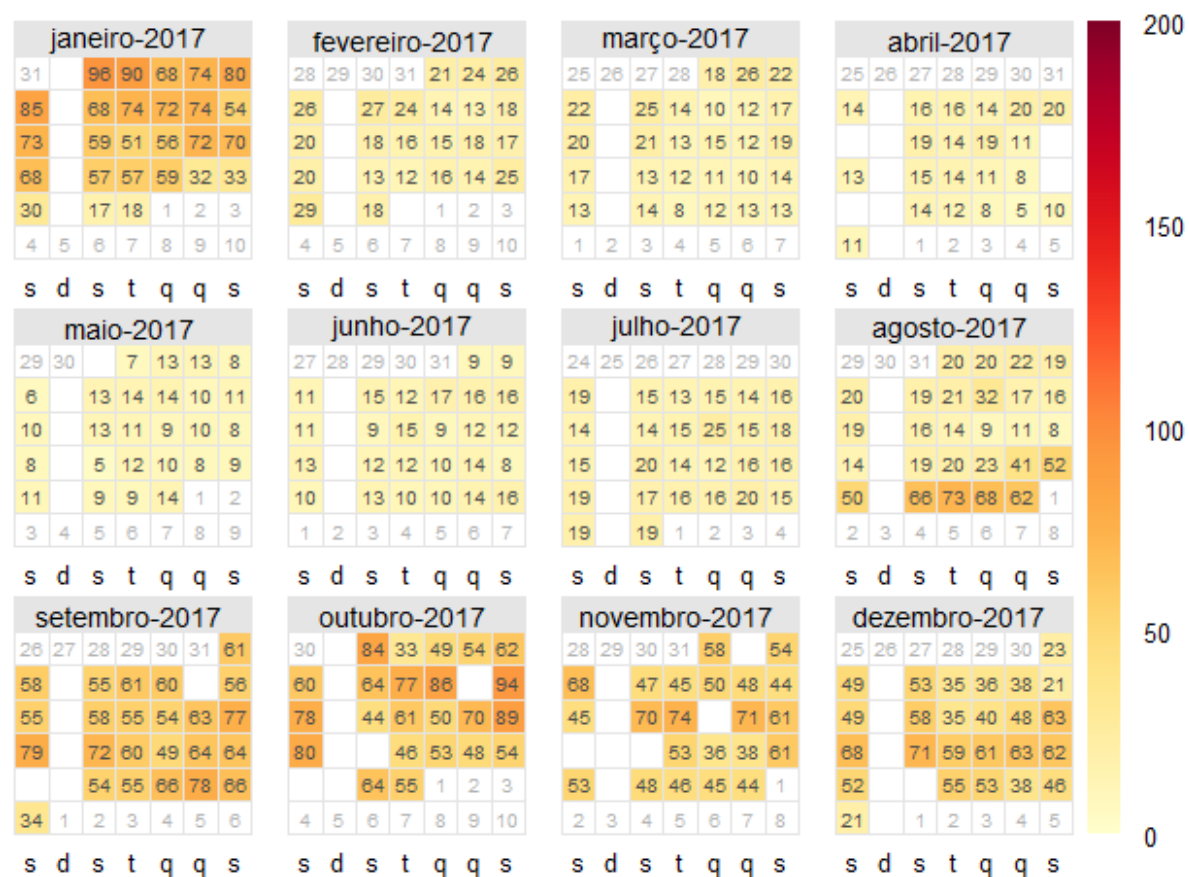
ANEXO B: Tabela B- Gráfico de Calendário para PDC em 2014 – continuação



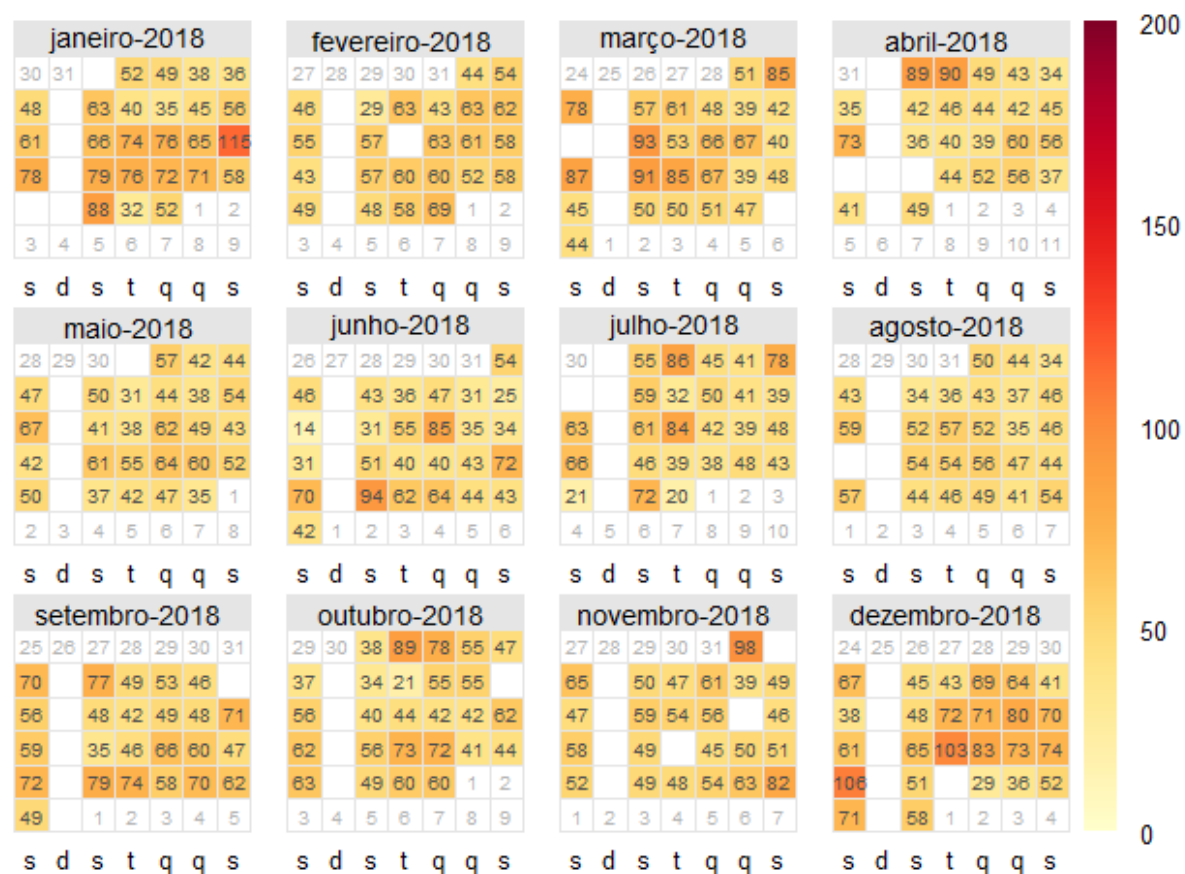




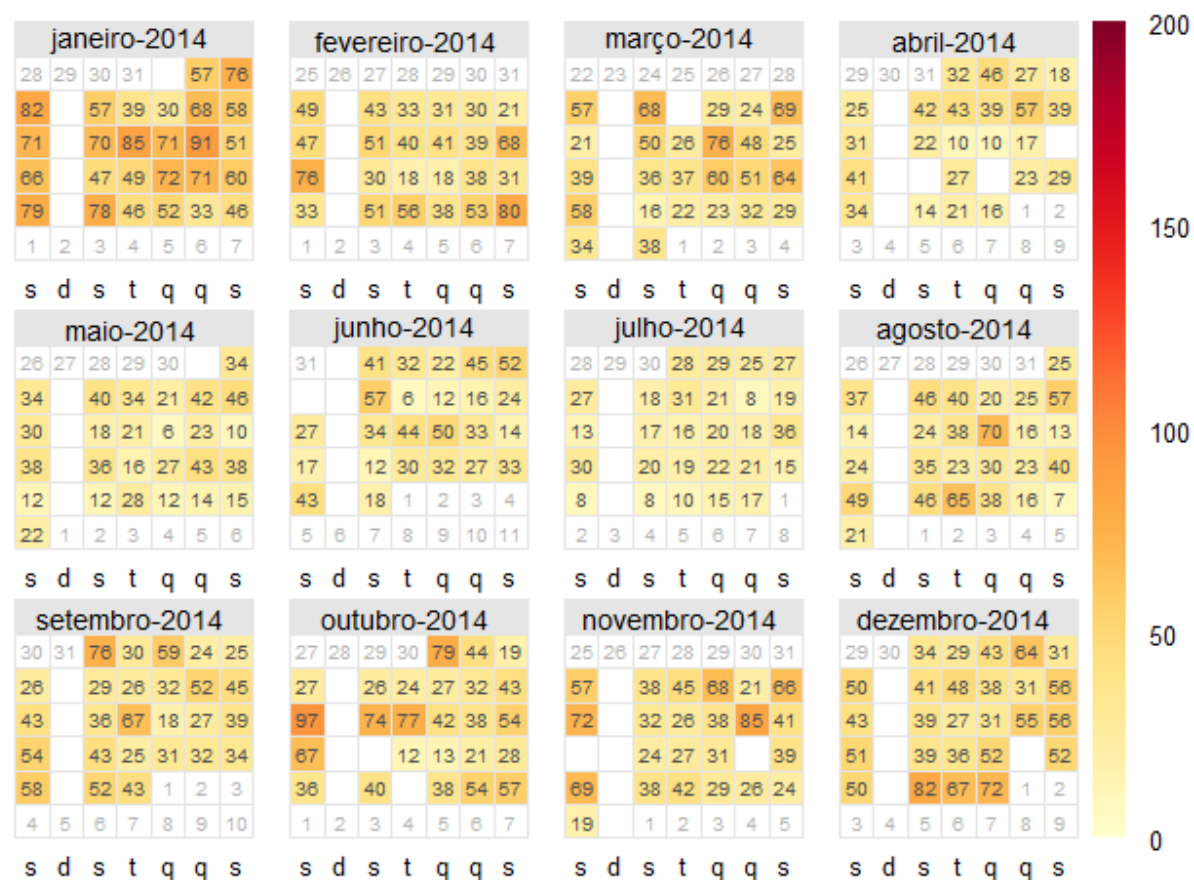
ANEXO B: Tabela B- Gráfico de Calendário para PDC em 2017 – continuação



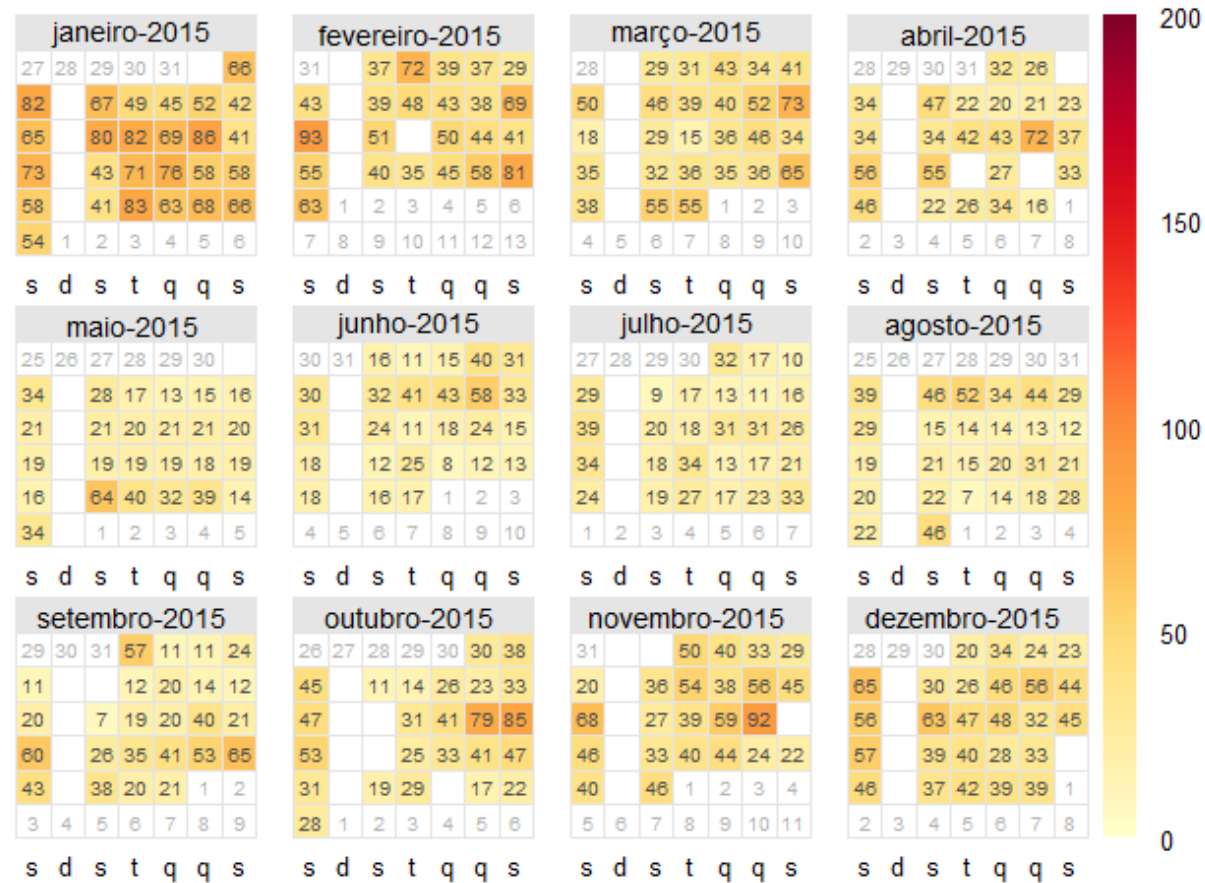
ANEXO B: Tabela B- Gráfico de Calendário para PDC em 2018 – continuação



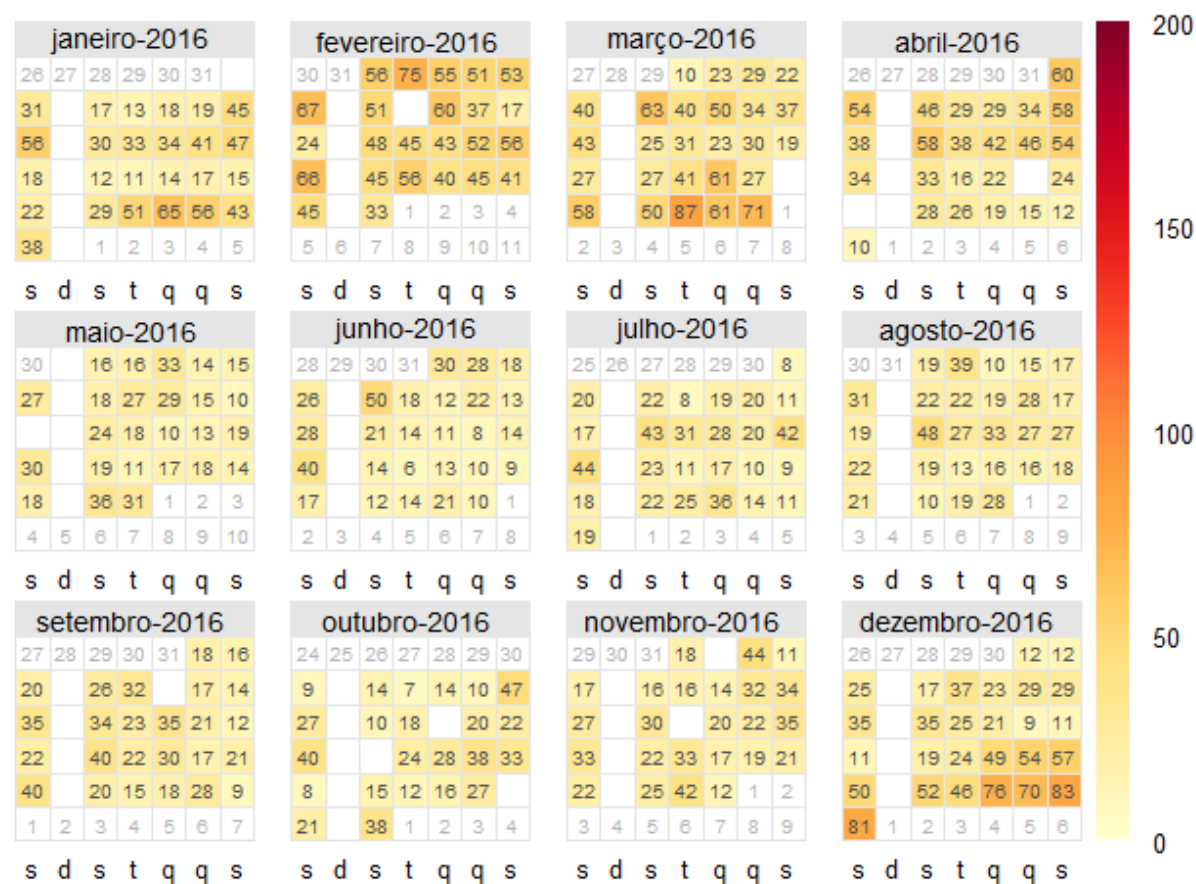
ANEXO B: Tabela B- Gráfico de Calendário para VSL em 2014 – continuação



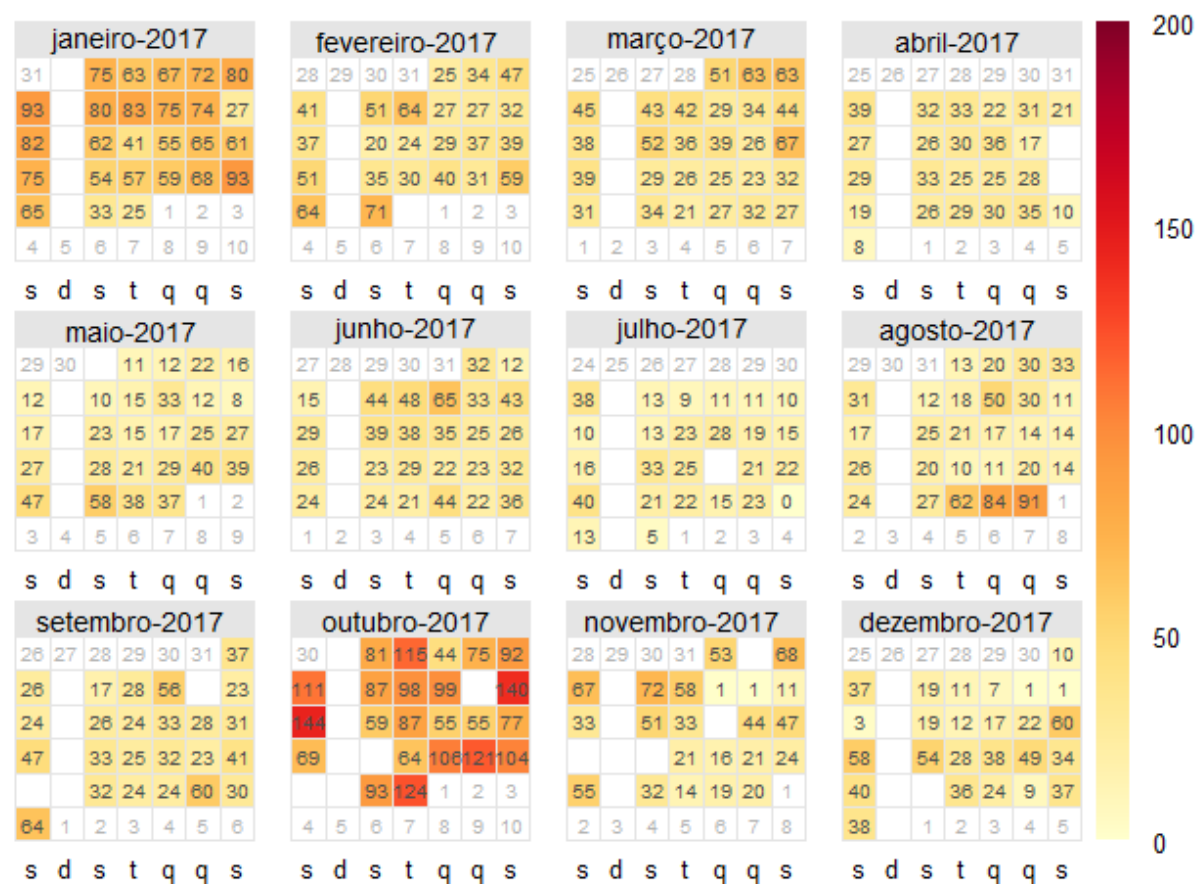
ANEXO B: Tabela B- Gráfico de Calendário para VSL em 2015 – continuação



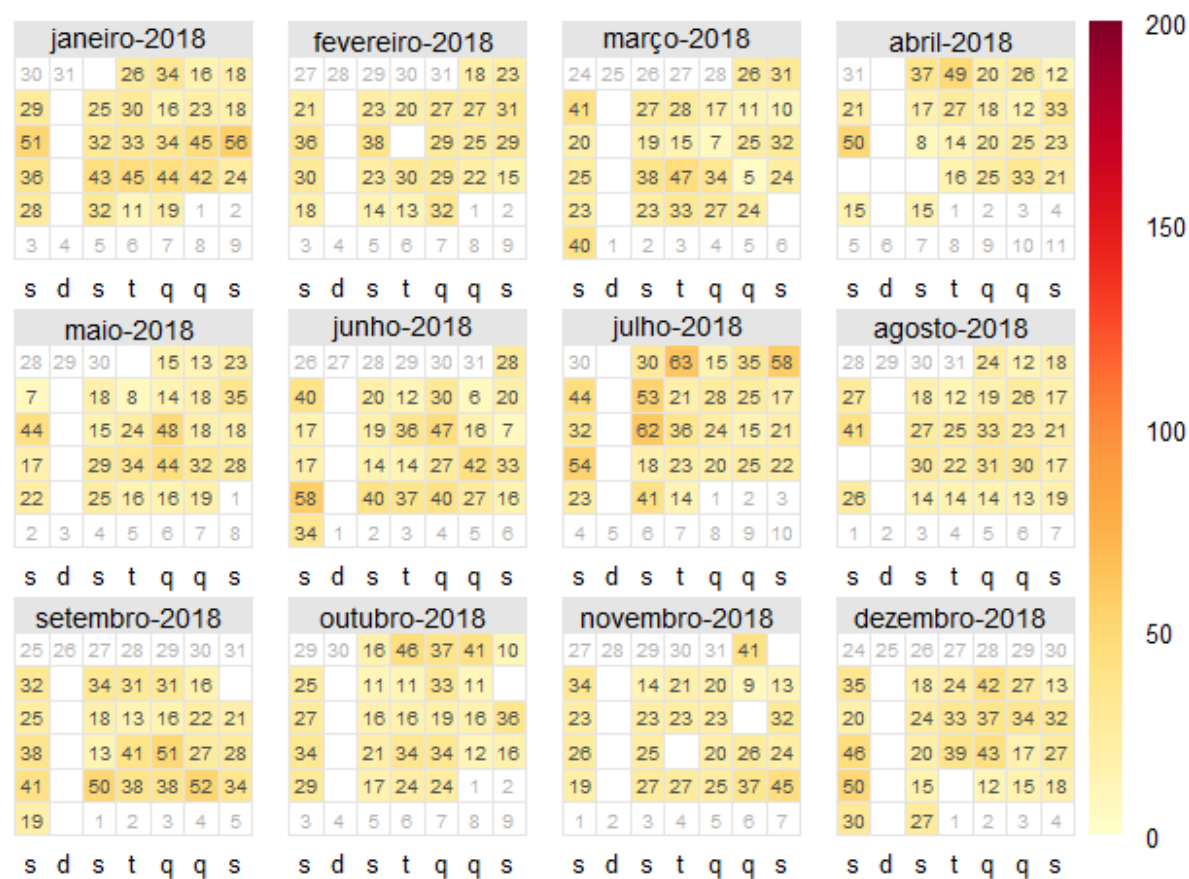
ANEXO B: Tabela B- Gráfico de Calendário para VSL em 2016 – continuação



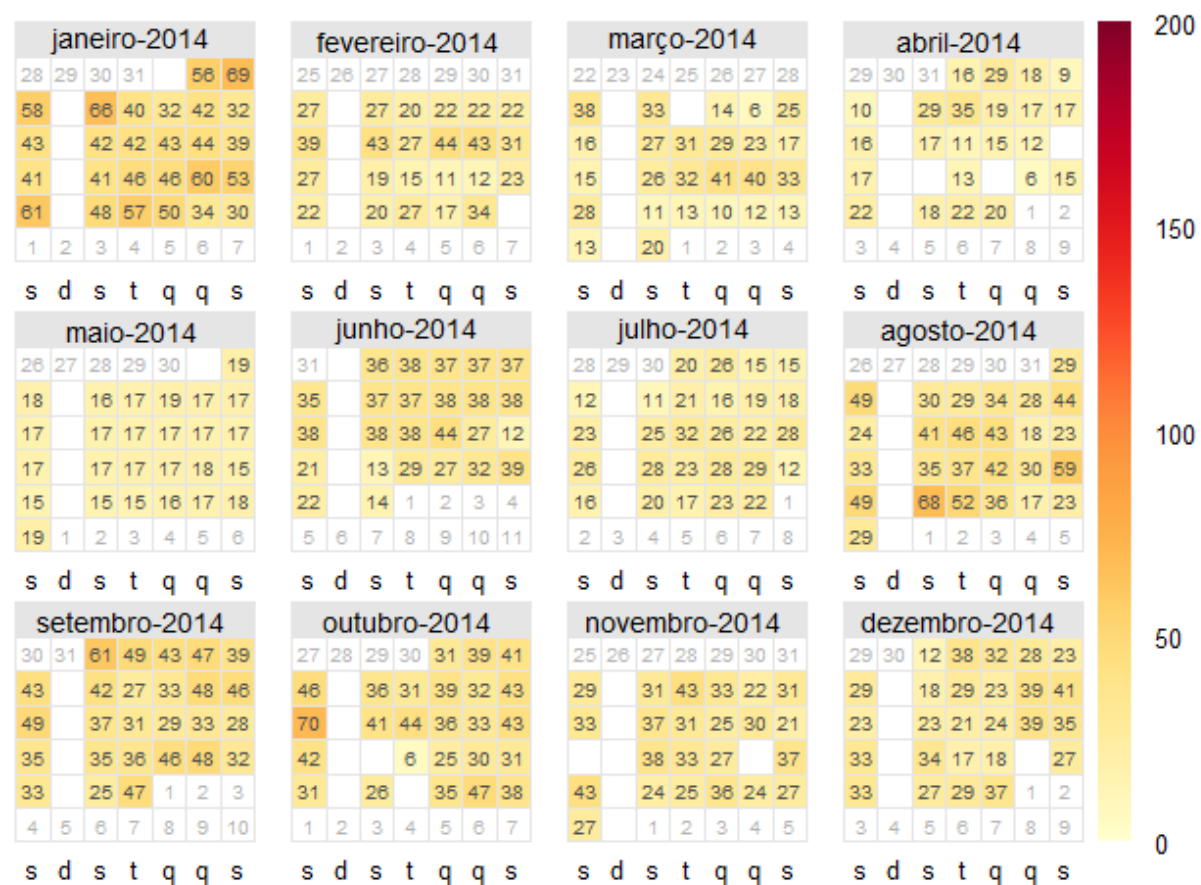
ANEXO B: Tabela B- Gráfico de Calendário para VSL em 2017 – continuação



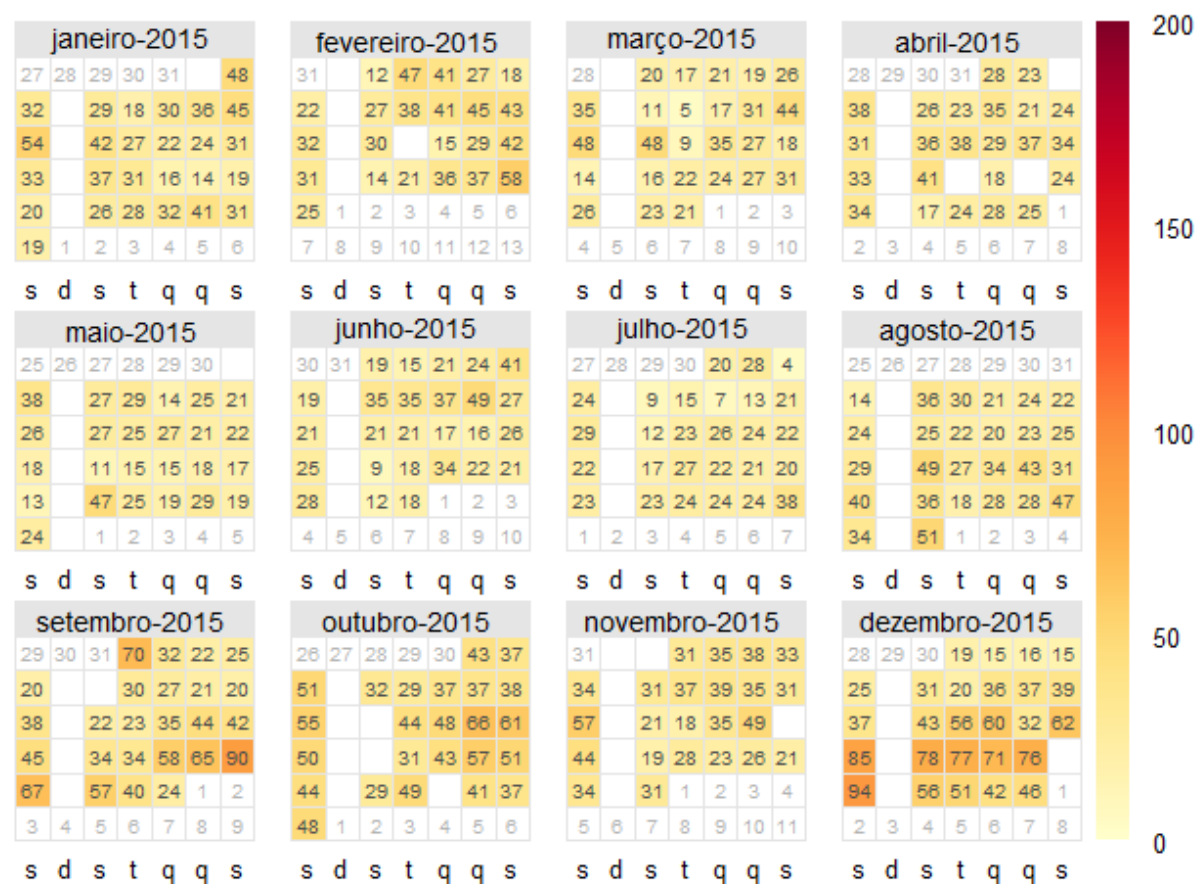
ANEXO B: Tabela B- Gráfico de Calendário para VSL em 2018 – continuação



ANEXO B: Tabela B- Gráfico de Calendário para INE em 2014 – continuação



ANEXO B: Tabela B- Gráfico de Calendário para INE em 2015 – continuação



ANEXO B: Tabela B- Gráfico de Calendário para INE em 2016 – continuação

