



UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE TECNOLÓGICO
CURSO DE POSGRADUAÇÃO EM ENGENHARIA AMBIENTAL

Fernando Campo García

**Desenvolvimento de sistema de baixo custo para o monitoramento da
qualidade do ar**

Florianópolis
2024

Fernando Campo García

**Desenvolvimento de sistema de baixo custo para o monitoramento da
qualidade do ar**

Tese de Doutorado do Curso de Posgraduação
em Engenharia Ambiental do Centro de
Tecnológico da Universidade Federal de Santa
Catarina para a obtenção do título de Doutor em
Engenharia Ambiental.

Orientador: Prof. Leonardo Hoinaski, Dr.

Coorientador: Prof. Davide Franco, Dr.; Prof.
Alejandro Rafael García Ramírez, Dr.

Florianópolis
2024

Ficha catalográfica gerada por meio de sistema automatizado gerenciado pela BU/UFSC.
Dados inseridos pelo próprio autor.

García, Fernando Campo

Desenvolvimento de sistema de baixo custo para o monitoramento da qualidade do ar / Fernando Campo García ; orientador, Leonardo Hoinaski, coorientador, Davide Franco, coorientador, Alejandro Rafael García Ramírez, 2024.
422 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia Ambiental, Florianópolis, 2024.

Inclui referências.

1. Engenharia Ambiental. 2. Monitoramento da qualidade do ar. 3. Sensores de gases de baixo custo. 4. Modelos de regressão. 5. Sistema embarcado. I. Hoinaski, Leonardo. II. Franco, Davide. III. Ramírez, Alejandro Rafael García IV. Universidade Federal de Santa Catarina. Programa de Pós Graduação em Engenharia Ambiental. V. Título.

Fernando Campo García

**Desenvolvimento de sistema de baixo custo para o monitoramento da
qualidade do ar**

Esta Tese de Doutorado foi julgada adequado para obtenção do Título de “Doutor em Engenharia Ambiental” e aprovada em sua forma final pelo Curso de Posgraduação em Engenharia Ambiental.

Florianópolis, 18 de Março de 2024.

Profa. Alexandra Rodrigues Finotti, Dra.
Coordenadora do Curso

Banca Examinadora:

Prof. Leonardo Hoinaski, Dr.
Orientador

Prof. Pedro Luiz Borges Chaffe, Dr.
Avaliador
Universidade Federal de Santa Catarina

Profa. Taciana Toledo de Almeida
Albuquerque, Dra.
Avaliadora
Universidade Federal de Minas Gerais

Prof. Rizzieri Pedruzzi, Dr.
Avaliador
Universidade do Estado do Rio de Janeiro

Prof. Alejandro Durán Carrillo de Albornoz,
Dr.
Avaliador
Universidade de Havana, Cuba

À minha vó Olga, quem me ensinou a escrever.

AGRADECIMENTOS

Primeiramente agradeço a Deus por me conceder a inspiração, energia e o sustento para concluir este trabalho. Agradeço especialmente aos professores Henrique de Melo Lisboa, Alejandro Durán Carrillo de Albornoz e Alejandro García Ramírez que depositaram sua confiança em 2014 me concedendo a oportunidade que me conduziu até este estágio da minha carreira profissional, exatamente 10 anos depois. Minha profunda gratidão ao professor Leonardo Hoinaski que me ofereceu o desafio de assumir a empreitada deste projeto, juntamente com seu total apoio em todas as áreas, lutando lado a lado, suportando nos momentos difíceis e corrigindo nos momentos de vacilo. Meus agradecimentos vão também para o professor Davide Franco por me acolher e conduzir nos primeiros anos de doutorado.

Agradeço a minha esposa por suportar pacientemente os sacrifícios advindos de uma tese de doutorado, por perdoar minhas ausências e por encorajar nos momentos de desânimo. Aos meus sogros que me acolheram como um filho. À minha mãe por ser minha principal fã, por seus conselhos e seu sólido suporte. Ao meu caro irmão Andy Blanco por preparar o caminho para minha entrada no LCQAr e por seu apoio incondicional nos primeiros anos. Aos amigos que ganhei e que anseiam minha volta à vida social. Aos meus colegas de laboratório Camilo, Robson, Thiago e Otávio pela parceria e a amizade na ciência brasileira, a Gabriel Ratão e a Jean, que foram verdadeiros companheiros de luta e fizeram possível que este trabalho acontecesse. Ao PPGEA da UFSC que me abriu suas portas ao programa num momento de grandes incertezas e instabilidades.

Agradeço à empresa Dynamox e seu CEO Guillaume Barrault por me oferecer emprego quando a bolsa de doutorado não foi suficiente para o sustento, e por me possibilitar concluir a pesquisa concedendo todo o suporte e flexibilidade possíveis. Ao coordenador João Pedro dos Reis que tem se tornado um amigo e uma importante fonte de encorajamento. Por todo o conhecimento adquirido através dos meus colegas que possibilitou refinar muitas etapas da pesquisa, em especial ao próprio João e ao Daniel Schröder. Ao Marcos Barp por oferecer seu tempo e conhecimento para resolver "pepinos" de séries temporais.

À Mãe gentil brasileira, que em berço esplêndido me acolheu, enquanto Saturno devorava os seus filhos.

“Ó Senhor, que variedade há nas tuas obras! Fizeste todas com sabedoria; a terra está cheia das tuas riquezas. Também o vasto mar, onde se movem seres inumeráveis, animais pequenos e grandes. Ali passam os navios, e o Leviatã que formaste para nele se recrear. Todos esperam de ti que lhes dês o sustento a seu tempo.”
(Salmo de Davi)

RESUMO

O monitoramento da qualidade do ar tem experimentado uma mudança de paradigma com a incorporação de sensores de baixo custo. Estes equipamentos têm potencial para aumentar a resolução espaço-temporal dos dados de poluentes, assim como diversificar e simplificar as aplicações de monitoramento. Todavia, o volume e a diversidade de aplicações com este tipo de sensores encontra-se restrito pela baixa portabilidade de uma aplicação para outra e pela concentração de iniciativas maioritariamente em países desenvolvidos. Neste trabalho foi desenvolvido um sistema para o monitoramento de baixo custo da qualidade do ar: a iniciativa CLEAN (Collaborative Low-cost Environmental Air quality Network). O sistema foi desenvolvido com uma arquitetura modular possibilitando o reaproveitamento de funcionalidades comuns entre aplicações, direcionando assim os esforços de desenvolvimento à implementação dos requisitos específicos de cada aplicação particular. CLEAN engloba um conjunto de subsistemas que incluem: uma *API* para registro e acesso a dados de monitores de baixo custo, bibliotecas de *firmware* e elementos de *hardware* para o desenvolvimento de monitores de baixo custo. No trabalho também foram desenvolvidos monitores de baixo custo para comporem a rede CLEAN. Um deles foi instalado junto a uma estação de monitoramento de referência da qualidade do ar, para validação e correção das suas leituras por um período de 5 meses. No trabalho propõe-se uma metodologia de correção das leituras do equipamento, a partir das medições coletadas pela estação de referência. Para isso foram ajustados modelos de regressão baseados em redes neurais Perceptron Multicamadas, Florestas Aleatórias, K Vizinhos mais Próximos e regressão Linear Multivariada. Os modelos, diferentemente de outros trabalhos que realizam correções das leituras de sensores de baixo custo, consideraram, como variáveis de entrada, as leituras de diferentes modelos de sensores para obter as medidas de um determinado gás. No caso em questão, foram utilizados sensores de baixo custo do fabricante Alphasense, sensíveis aos gases Monóxido de Carbono (*CO*), Dióxido de Nitrogênio (*NO₂*), Ozônio (*O₃*) e Material Particulado (*MP10*). A metodologia de correção desenvolvida, considera o conjunto das leituras de todos os sensores e da temperatura para obter as medições de um único gás. Com essa metodologia obtiveram-se modelos de correção com valores de *R²* de até 0.6 para a medição de *O₃* utilizando uma rede neural Perceptron Multicamadas. Os modelos obtidos para correção das medidas de *MP10*, *NO₂* e *CO*, se comparados com as correções que consideravam um único modelo de sensor, apresentaram um incremento no valor de *R²* de $R^2 < 0$ até 0.18, 0.14 e 0.07 respectivamente. A metodologia proposta nesta pesquisa mostrou-se robusta para trabalhar com dados de baixa qualidade e acrescentar valor a leituras ruidosas e com falhas nos sensores. O trabalho traz como principais contribuições um marco de trabalho colaborativo para o desenvolvimento e interconexão de sensores de baixo custo, assim como uma metodologia para calibração dos sensores com dados muito ruidosos, em uma aplicação real e em ambiente não controlado.

Palavras-chave: Monitoramento da qualidade do ar, sensores de gases de baixo custo, sistemas embarcados, API, modelos de regressão

ABSTRACT

Air quality monitoring has experienced a paradigm shift with the incorporation of low-cost sensors. This equipment has the potential to increase the spatio-temporal resolution of pollutant data, as well as diversify and simplify monitoring applications. However, the volume and diversity of applications with this type of sensor are restricted by the low portability of one application to another and the concentration of initiatives mainly in developed countries. In this work, a system was developed for low-cost monitoring of air quality: the CLEAN (Collaborative Low-cost Environmental Air Quality Network) initiative. The system was developed with a modular architecture enabling the reuse of common functionalities between applications, thus directing development efforts to implement the specific requirements of each particular application. CLEAN encompasses a set of subsystems that include: an API for recording and accessing data from low-cost monitors, firmware libraries, and hardware elements for developing low-cost monitors. In the work, low-cost monitors were also developed to form part of the CLEAN network. One of them was installed next to an air quality reference monitoring station, to validate and correct its readings for a period of 5 months. The work proposes a methodology for correcting equipment readings, based on measurements collected by the reference station. For this, regression models based on Multilayer Perceptron neural networks, Random Forests, K Nearest Neighbors, and Multivariate Linear regression were adjusted. The models, unlike other works that perform corrections to readings from low-cost sensors, considered, as input variables, the readings from different sensor models to obtain measurements of a given gas. In the case in question, low-cost sensors from the manufacturer Alphasense were used, sensitive to the gases Carbon Monoxide (CO), Nitrogen Dioxide (NO₂), Ozone (O₃), and Particulate Matter (PM10). The correction methodology developed considers the set of readings from all sensors and temperature to obtain measurements of a single gas. With this methodology, correction models with R² values of up to 0.6 were obtained for measuring O₃ using a Multilayer Perceptron neural network. The models obtained to correct the measurements of PM10, NO₂, and CO, when compared with the corrections that considered a single sensor model, showed an increase in the R² value from R² < 0 to 0.18, 0.14, and 0.07 respectively. The methodology proposed in this research proved to be robust in working with low-quality data and adding value to noisy readings and sensor failures. The main contributions of the work are a collaborative work framework for the development and interconnection of low-cost sensors, as well as a methodology for calibrating sensors with very noisy data, in a real application and an uncontrolled environment.

Keywords: Air quality monitoring, low-cost gas sensors, embedded systems, API, regression models

LISTA DE FIGURAS

Figura 1 – Estado do monitoramento de referência da qualidade do ar no Brasil	29
Figura 2 – A iniciativa <i>Habitat Map</i> e o sensor <i>AirBeam</i>	32
Figura 3 – O portal <i>Air Quality Egg</i> e o sensor <i>Egg</i>	33
Figura 4 – Sistema de Monitoramento SELVA utiliza sensores <i>PurpleAir</i>	34
Figura 5 – Os <i>kits</i> de <i>Smart Citizen</i> e <i>Sensor.Community</i>	34
Figura 6 – Perceptron Multicamadas	42
Figura 7 – Componentes da iniciativa CLEAN	44
Figura 8 – Estrutura de desenvolvimento de aplicação de monitoramento de baixo custo	46
Figura 9 – Conjunto de bibliotecas utilizadas para <i>firmware</i> dos dispositivos CLEAN	48
Figura 10 – Diagramas de classes do pacote <i>Hardware Interfaces</i>	49
Figura 11 – Diagrama de classes do Módulo <i>Drivers</i>	50
Figura 12 – Diagramas de classes do módulo Sensors	52
Figura 13 – Diagrama de classes do pacote Data	55
Figura 14 – Processo de leitura de uma variável	56
Figura 15 – Aplicação web Renovar	57
Figura 16 – Estrutura da aplicação Web Renovar	58
Figura 17 – Entidades do banco de dados Renovar	59
Figura 18 – Camadas da aplicação <i>back-end</i> Renovar	60
Figura 19 – Classes de acesso ao dados da aplicação <i>back-end</i> Renovar	61
Figura 20 – Aplicação <i>front-end</i> da plataforma web Renovar	62
Figura 21 – Painéis da aplicação <i>front-end</i> de Renovar	62
Figura 22 – Painel de análise de dados da aplicação <i>back-end</i> Renovar	62
Figura 23 – Estrutura principal dos dispositivos. a) Medidor de gases fixo, e b) medidor móvel	63
Figura 24 – Ilustrações das versões (a) fixa e (b) móvel dos dispositivos de monitoramento	64
Figura 25 – A placa CLEAN Arduino Mega: (a) projeto PCB, (b) vista superior da placa, (c) vista inferior da placa.	65
Figura 26 – Mapa das estações de monitoramento em Tubarão e Capivari de Baixo	67
Figura 27 – Fluxograma das etapas de pré-processamento e as etiquetas de saída em cada uma delas	68
Figura 28 – Série temporal do sensor CO-B4	71
Figura 29 – Histograma das leituras do sensor CO-B4	72
Figura 30 – Série temporal do sensor pré-processada ($T = 15$ mins) e seu comportamento diário	72

Figura 31 – Série temporal das leituras de concentração de referência ($T = 1$ H) e seu comportamento diário	73
Figura 32 – Série temporal com $T = 1$ hr	73
Figura 33 – Relação dos dados de concentração de CO com a temperatura	74
Figura 34 – Séries temporais e gráficos de dispersão das medições de CO	75
Figura 35 – Série temporal dos sensores de O_3 modelo OX-B431	76
Figura 36 – Séries temporais dos sensores OX-B431 pré-processadas	77
Figura 37 – Histogramas e séries temporais horárias das leituras dos sensores OX-B431	79
Figura 38 – Relação entre as leituras dos sensores de O_3 e a temperatura	80
Figura 39 – Séries temporais e gráficos de dispersão das medições de O_3	81
Figura 40 – Série temporal do sensor NO2-B43F	82
Figura 41 – Histograma das leituras do sensor NO2-B43F	82
Figura 43 – Relação dos dados de concentração de NO_2 com a temperatura	84
Figura 44 – Séries temporais e gráficos de dispersão das medições de NO_2	85
Figura 45 – Série temporal das leituras do sensor OPC-N3	86
Figura 46 – Série temporal com $T = 1$ hr	87
Figura 47 – Relação dos dados de concentração de MP_{10} com a temperatura	88
Figura 48 – Séries temporais e gráficos de dispersão das medições de MP_{10}	88
Figura 49 – Fluxogramas do processo de correção das leituras do equipamento de baixo custo	90
Figura 50 – Processo de geração de combinações de variáveis, modelos e seus parâmetros	91
Figura 51 – Resultados dos modelos de regressão aplicados as leituras do sensor CO-B4	93
Figura 52 – Gráfico de dispersão das leituras do sensor CO-B4 e a estação de referência após aplicar modelos de regressão considerando a temperatura	94
Figura 53 – Desempenho dos modelos de regressão aplicados para inferir as leituras de concentração de CO medidas pela estação de referência	95
Figura 54 – Gráfico de dispersão das leituras do múltiplos sensores e a estação de referência para medição de CO	95
Figura 55 – Resultados dos modelos de regressão aplicados às leituras dos sensores OX-B431	97
Figura 56 – Gráfico de dispersão das leituras dos sensores de O_3 OX-B431 e a estação de referência após aplicar modelos de regressão considerando a temperatura	97
Figura 57 – Desempenho dos modelos de regressão aplicados para inferir as leituras de concentração de O_3 medidas pela estação de referência	98
Figura 58 – Gráfico de dispersão das leituras do múltiplos sensores e a estação de referência para medição de O_3	99

Figura 59 – Resultados dos modelos de calibração aplicados as leituras de NO_2 do sensor NO2-B43F	100
Figura 60 – Gráfico de dispersão das leituras do sensor de NO_2 NO2-B43F e a estação de referência após aplicar modelos de regressão considerando a temperatura	100
Figura 61 – Desempenho dos modelos de regressão aplicados para inferir as leituras de concentração de NO_2 medidas pela estação de referência	101
Figura 62 – Gráfico de dispersão das leituras de múltiplos sensores e a estação de referência para medição de NO_2	102
Figura 63 – Resultados dos modelos de calibração aplicados as leituras de MP_{10} do sensor OPC-N3	103
Figura 64 – Gráfico de dispersão das leituras do sensor de MP_{10} do OPC-N3 e a estação de referência após aplicar modelos de regressão considerando a temperatura	103
Figura 65 – Desempenho dos modelos de regressão aplicados para inferir as leituras de concentração de MP_{10} medidas pela estação de referência	104
Figura 66 – Gráfico de dispersão das leituras do múltiplos sensores e a estação de referência para medição de MP_{10}	105
Figura 67 – Representação de uma célula eletroquímica de dois eletrodos	120
Figura 68 – Potenciômetro para condicionamento de sensores eletroquímicos.	122
Figura 69 – Diagrama de blocos dos sistemas fixo (a) e móvel (b)	123
Figura 70 – Sensores dos fabricantes a) SPEC e b) Alphasense	125
Figura 71 – Interface entre os sensores e o microcontrolador Arduino. a) Alphasense, b) SPEC	129
Figura 72 – Interface entre o módulo cartão micro SD e o microcontrolador	130
Figura 73 – Interface entre o microcontrolador e os módulos a) RTC e b) GPS	130
Figura 74 – Interface entre o microcontrolador e o módulo de comunicação Wi-Fi .	131
Figura 75 – Instalação em campo do protótipo fixo	132
Figura 76 – Vista interior do protótipo fixo	133
Figura 77 – Diagrama de conexões do conjunto de sensores Alphasense	133
Figura 78 – Diagrama de conexões do conjunto de sensores Alphasense	135
Figura 79 – Sensor de Monôxido de Nitrogênio Alphasense da série B4	136
Figura 80 – Fluxograma do firmware programado para o microcontrolador Arduino MEGA	141
Figura 81 – Módulos e interfaces usados para controle e interface do RTC	150
Figura 82 – Fluxograma do firmware programado para o microcontrolador ESP8266	156
Figura 83 – Processo de atendimento a uma solicitação do mestre	160
Figura 84 – Fluxograma do processo após uma solicitação de DATA do mestre.	161
Figura 85 – Fluxograma do processo após uma solicitação TIME do mestre	163

Figura 86 – Série temporal dos sensores de SO_2 modelo SO2-B4	164
Figura 87 – Histogramas das leituras dos sensores SO2-B4	165

LISTA DE TABELAS

Tabela 1 – Metodologias de monitoramento de referência	27
Tabela 2 – Total de estações de monitoramento no Brasil	28
Tabela 3 – Requerimentos de desempenho dos instrumentos de monitoramento da qualidade do ar segundo área de aplicação	39
Tabela 4 – Relação de equipamentos presentes na estação de monitoramento de referência no município de Tubarão - SC	67
Tabela 5 – Contabilização das leituras do sensor CO-B4 por etiquetas	74
Tabela 6 – Contabilização dos dados por etiquetas das leituras do sensor 1 OX-B431	78
Tabela 7 – Contabilização dos dados por etiquetas das leituras do sensor 2 OX-B431	78
Tabela 8 – Contabilização das leituras do sensor NO ₂ -B43F por etiquetas	84
Tabela 9 – Contabilização dos dados por etiquetas das leituras de MP_{10} do sensor OPC-N3	87
Tabela 10 – Parâmetros ajustados para cada modelo nas buscas em <i>grid</i>	93
Tabela 11 – Resultados da calibração do sensor CO-B4	93
Tabela 12 – Resultados da calibração dos sensores OX-B431	96
Tabela 13 – Resultados da calibração das leituras de NO_2 do sensor NO ₂ -B43F . .	99
Tabela 14 – Resultados da calibração das leituras de MP_{10} do sensor OPC-N3 . .	102
Tabela 15 – Resultados dos melhores modelos considerando leituras de sensores por poluentes	105
Tabela 16 – Resultados dos melhores modelos por poluente considerando as leituras de todos os sensores nos modelos	106
Tabela 17 – Especificações técnicas dos ventiladores utilizados no equipamento fixo e móvel	124
Tabela 18 – Especificações técnicas dos sensores SPEC	126
Tabela 19 – Especificações técnicas dos sensores Alphasense	127
Tabela 20 – Lista de sensores utilizados no protótipo fixo	134
Tabela 21 – Principais componentes utilizados nos dispositivos CLEAN	137
Tabela 22 – Constantes e variáveis utilizadas para controlar a execução de cada funcionalidade no firmware	148
Tabela 23 – Tipos de solicitações representadas no tipo CommandTypes	160

LISTA DE SIGLAS

API Application Programming Interface

CO Monóxido de Carbono

O₃ Ozônio

NO₂ Dióxido de Nitrogênio

MP₁₀ Partículas inaláveis grossas com diâmetro 10 µm

SO₂ Dióxido de Enxofre

MLR Regressão linear multivariada (siglas em inglês)

EC Eletroquímico (siglas em inglês)

MP_{2.5-10} Partículas inaláveis grossas

CONAMA Conselho Nacional do Meio Ambiente

kNN k vizinhos mais próximos (siglas em inglês)

RF Florestas aleatórias (siglas em inglês)

ANN Redes neurais artificiais (siglas em inglês)

IQAr Índice da Qualidade do Ar

NDIR Infravermelho não dispersivo (siglas em inglês)

CO₂ Dióxido de Carbono

CH₄ Metano

PID Detectores de foto-ionização(siglas em inglês)

MOS Semicondutores de óxido metálico (siglas em inglês)

NO_X Óxidos de Nitrogênio

VOC Compostos orgânicos voláteis

H₂S Sulfato de Hidrogênio

GPS Global Positioning System

NH₃ Amônia

MP_{2.5} Partículas inaláveis grossas com diâmetro 2.5 µm

ppb Partes por bilhão

ppm Partes por milhão

NO Monóxido de Nitrogênio

DQO Objetivo da Qualidade dos Dados (siglas em inglês)

US EPA Agência de Proteção Ambiental Norte-americana (siglas em inglês)

ML Aprendizado de Máquinas (siglas em inglês)

RTC Relógio de Tempo Real (siglas em inglês)

UART Universal Asynchronous Receiver / Transmitter

NTP Network Time Protocol

CSV Valores Separados por Vírgula (siglas em inglês)

ISB Individual Sensor Board

IoT Internet das Coisas (siglas em inglês)

MVP Produto Viável Mínimo (siglas em inglês)

LCQAr Laboratório de Controle da Qualidade do Ar

REST Transferência de Estado Representacional (siglas em inglês)

PELT Pruned Exact Linear Time

AIC Akaike Information Criterion

BIC Bayesian Information Criterion

MSE Erro médio quadrático

JSON JavaScript Object Notation

IDE Ambiente de desenvolvimento integrado (siglas em inglês)

SSID Identificador de conjunto de serviços

URL Uniform Resource Locator

SUMÁRIO

	INTRODUÇÃO	21
1	ESTADO DA ARTE NO MONITORAMENTO DA QUALIDADE DO AR	25
1.1	MONITORAMENTO DE REFERÊNCIA DA QUALIDADE DO AR	26
1.2	MONITORAMENTO DE BAIXO CUSTO DA QUALIDADE DO AR	29
1.2.1	Princípios de funcionamento dos sensores de gases de baixo custo	30
1.2.2	Iniciativas de monitoramento de baixo custo da qualidade do ar	31
1.2.3	Limitações do monitoramento de baixo custo da qualidade do ar	35
1.3	CALIBRAÇÃO DOS SISTEMAS DE BAIXO CUSTO PARA MONITORAMENTO DA QUALIDADE DO AR	38
1.3.1	Redes Neurais Artificiais	42
1.3.2	Florestas Aleatórias	43
1.3.3	K Vizinhos Mais Próximos	43
2	CLEAN - COLLABORATIVE LOW-COST ENVIRONMENTAL AND AIR-QUALITY NETWORK	44
2.1	BIBLIOTECAS DE FIRMWARE	45
2.1.1	O módulo de interfaces de hardware	47
2.1.2	O módulo drivers	48
2.1.2.1	TimeDriver	49
2.1.2.2	GPSDriver	50
2.1.2.3	RTCDriver	51
2.1.2.4	DataContainer	51
2.1.2.5	HardStorage	51
2.1.3	O módulo Sensores	51
2.1.3.1	AlphaSenseISB	52
2.1.3.2	AlphaSenseCompensator	53
2.1.3.3	AlphaOXCompensator	53
2.1.3.4	Interface com sensores seriais	54
2.1.4	O módulo Data	54
2.2	A API RENOVAR	57
2.2.1	Banco de dados	58
2.2.2	A aplicação Back-end	59
2.2.3	A aplicação Front-end	61
2.3	DISPOSITIVOS DE HARDWARE DESENVOLVIDOS	61
2.3.1	Protótipos de monitores de qualidade do ar de baixo custo	62
2.3.2	A placa CLEAN Arduino MEGA	64

3	IDENTIFICAÇÃO DE INTERFERÊNCIAS E FALHAS NO MONITORAMENTO DE BAIXO CUSTO	66
3.1	CORREÇÃO DAS LEITURAS POR CO-LOCALIZAÇÃO	66
3.2	PRÉ-PROCESSAMENTO DOS DADOS	68
3.3	ANÁLISE DOS DADOS DE MONÓXIDO DE CARBONO	71
3.3.1	Dependência com a temperatura	73
3.3.2	Comparação das leituras do sensor CO-B4 com as medições de referência	75
3.4	ANÁLISE DOS DADOS DE OZÔNIO	75
3.4.1	Dependência com a temperatura	78
3.4.2	Comparação das leituras dos sensores OX-B431 com as medições de referência	80
3.5	ANÁLISE DOS DADOS DE DIÓXIDO DE NITROGÊNIO	80
3.5.1	Dependência com a temperatura	82
3.5.2	Comparação das leituras de NO_2 do sensor NO2-B43F com as medições de referência	83
3.6	ANÁLISE DOS DADOS DE MATERIAL PARTICULADO	84
3.6.0.1	Dependência com a temperatura	85
3.6.1	Comparação das leituras de MP_{10} do sensor OPC-N3 com as medições de referência	87
3.7	DISCUSSÃO	88
4	METODOLOGIA DE CORREÇÃO DE LEITURAS PROVENIENTES DE SENsoRES DE BAIXO CUSTO DA QUALIDADE DO AR	90
4.1	CORREÇÃO DAS LEITURAS DO SENSOR CO-B4 COM AS MEDIÇÕES DE REFERÊNCIA	92
4.2	CÁLCULO DA CONCENTRAÇÃO DE MONÓXIDO DE CARBONO A PARTIR DAS LEITURAS DO ARRANJO DE SENsoRES DE GASES	94
4.3	CORREÇÃO DAS LEITURAS DOS SENsoRES OX-B431 COM AS MEDIÇÕES DE REFERÊNCIA	96
4.4	CÁLCULO DA CONCENTRAÇÃO DE OZÔNIO A PARTIR DAS LEITURAS DO ARRANJO DE SENsoRES DE GASES	98
4.5	CORREÇÃO DAS LEITURAS DO SENSOR NO2-B43F COM AS MEDIÇÕES DE REFERÊNCIA	99
4.6	CÁLCULO DA CONCENTRAÇÃO DE DIÓXIDO DE NITROGÊNIO A PARTIR DAS LEITURAS DO ARRANJO DE SENsoRES DE GASES	101
4.7	CORREÇÃO DAS LEITURAS DE MP10 DO SENSOR OPC-N3 COM AS MEDIÇÕES DE REFERÊNCIA	101
4.8	CÁLCULO DA CONCENTRAÇÃO DE MP10 A PARTIR DAS LEITURAS DO ARRANJO DE SENsoRES DE GASES	104

4.9	DISCUSSÃO	104
4.9.1	Comparação dos resultados obtidos com o equipamento desenvolvido em relação a trabalhos similares	106
	CONCLUSÃO	108
	REFERÊNCIAS	111
	APÊNDICE A – SENsoRES DE GASES ELETROQUÍMICOS	120
	APÊNDICE B – PROTÓTIPOS DE MONITORES DA QUALIDADE DO AR DESENVOLVIDOS	123
B.1	TRANSPORTE DE GASES	125
B.2	SENSORIAMENTO	125
B.2.1	Sensores SPEC	126
B.2.2	Sensores Alphasense	126
B.3	CONDICIONAMENTO	127
B.3.1	Interface de condicionamento dos sensores Alphasense: A Placa de Sensoriamento Individual (ISB)	127
B.3.2	Interface de condicionamento dos sensores SPEC	128
B.4	MICROCONTROLADOR	129
B.4.1	Armazenamento dos dados	129
B.4.2	Controle de data e hora e geolocalização	129
B.4.3	Comunicação Wi-Fi	131
B.5	MONTAGEM DO PROTÓTIPO FIXO	132
B.6	MONTAGEM DO PROTÓTIPO MÓVEL	135
	APÊNDICE C – A PLACA <i>CLEAN ARDUINO MEGA</i>	136
C.1	MÓDULO DE SENSORIAMENTO	136
C.1.1	Sensores	136
C.1.1.1	Sensores eletroquímicos Alphasense.	136
C.1.1.2	Interface de condicionamento de sensores SPEC	137
C.1.1.3	Interface de condicionamento de sensores Alphasense	138
C.1.1.4	Contador ótico de partículas Alphasense para medição de material particulado	138
C.2	O MICROCONTROLADOR	138
C.2.1	Armazenamento dos dados	138
C.2.2	Relógio de tempo real	139
C.2.3	Comunicação Wi-Fi	139
	APÊNDICE D – O <i>FIRMWARE CLEAN</i>	140
D.1	CÓDIGO CLEAN ARDUINO MEGA	140
D.1.1	Identificação do dispositivo e seus sensores	142
D.1.2	Configuração: a função <i>setup()</i>	142
D.1.2.1	Serial, Serial1, Serial3	144
D.1.2.2	espIoT	144

D.1.2.3	SD	144
D.1.2.4	Rtc	144
D.1.3	Interrupção Serial3	145
D.1.4	Laço principal do programa: a função loop()	146
D.1.4.1	Leitura dos sensores	149
D.1.4.2	Armazenamento dos dados	152
D.1.4.3	Envio de dados via protocolo <i>HTTP</i>	153
D.1.4.4	Geolocalização	153
APÊNDICE E – O FIRMWARE DO MICROCONTROLADOR		
	ESP8266	155
E.1	CONFIGURAÇÃO E CONEXÃO WI-FI	155
E.1.1	NUM_WIFIS	156
E.1.2	WiFiCredentials wifiCreds[]	156
E.1.3	setup_wifi_connection<NUM_WIFIS>(wifiCreds)	157
E.1.4	espHTTP	157
E.1.5	espSerial	157
E.1.6	Serial	157
E.2	O LAÇO PRINCIPAL	158
E.2.1	O comando DATA: enviando um POST HTTP para a API Renovar	161
APÊNDICE F – LEITURAS DE SO2 ADQUIRIDAS PELO		
	SENSOR SO2-B4	164
ANEXO A – DOCUMENTAÇÃO DA API RENOVAR: ENDPOINTS		
	E REQUISIÇÕES	166
ANEXO B – REGISTRO DE PROGRAMA DE COMPUTADOR		
	CLEAN	220
ANEXO C – JUPYTER NOTEBOOKS COM METODOLOGIA		
	DE PRÉ-PROCESSAMENTO APLICADA ÀS	
	LEITURAS DOS SENsoRES CO-B4, OX-B431,	
	NO2-B43F E DE MP10 DO OPC-N3	222
ANEXO D – JUPYTER NOTEBOOKS COM METODOLOGIAS		
	DE CORREÇÃO APLICADAS ÀS LEITURAS DOS	
	SENsoRES OX-B431	364
D.1	METODOLOGIA CONSIDERANDO APENAS LEITURAS DOS SENsoRES OX-B431 E A TEMPERATURA	364
D.2	METODOLOGIA CONSIDERANDO AS LEITURAS DE TODOS OS SENsoRES E A TEMPERATURA	390

INTRODUÇÃO

A poluição do ar é de grande risco para a saúde, sendo responsável pela morte de cerca de sete milhões de pessoas em todo o mundo anualmente e de variadas doenças cardiorrespiratórias (WORLD HEALTH ORGANIZATION, 2021). Inclusive tem sido evidenciado que o risco de sofrer doenças, ou mesmo de morte, é maior nos grupos de população mais vulneráveis e marginalizadas, como as minorias raciais/étnicas e as pessoas de baixo nível socioeconômico (JBAILY *et al.*, 2022). Embora quase 99% da população mundial respire ar que excede os limites das diretrizes da OMS (WORLD HEALTH ORGANIZATION, 2021), a cobertura de monitoramento da qualidade do ar ainda é baixa e insuficiente (MUNIR *et al.*, 2019), especialmente em países de baixo e médio rendimento (VORMITTAG *et al.*, 2021; FERREIRA *et al.*, 2022). Um monitoramento representativo da poluição atmosférica é essencial para a gestão eficaz da qualidade do ar, já que a variabilidade espaço-temporal da concentração de poluentes é elevada, especialmente em ambientes urbanos (KUMAR *et al.*, 2015; MEAD *et al.*, 2013).

Os poluentes atmosféricos têm sido monitorados usando equipamentos complexos e caros em locais fixos (KANG *et al.*, 2022). Os altos custos desses instrumentos limitam sua implantação a apenas algumas estações por cidade, deixando grandes áreas geográficas descobertas (MUNIR *et al.*, 2019), restringindo a resolução espacial e a distribuição das redes de monitoramento (JIAO *et al.*, 2016; KUMAR *et al.*, 2015). No Brasil, por exemplo, as redes de monitoramento instaladas cobrem apenas aproximadamente 2% do total de municípios brasileiros, e seu desempenho a longo prazo é muitas vezes comprometido pela falta de manutenção e de pessoal qualificado (VORMITTAG *et al.*, 2021).

Buscando uma melhor compreensão do processo de poluição do ar e seus impactos, a ideia de sensoriamento em alta resolução espacial tem atraído a atenção da comunidade de qualidade do ar, e o uso de monitores de baixo custo tem ganhado popularidade (MOTLAGH *et al.*, 2020; KUMAR *et al.*, 2015). Devido à sua versatilidade - dada pelas suas pequenas dimensões e seu consumo energético reduzido (LEWIS; SCHNEIDEMESSER *et al.*, 2018) - e seus custos baixos (tanto de aquisição quanto de operação), esses dispositivos podem complementar a escassez espacial e temporal de redes certificadas de qualidade do ar e expandir o horizonte de novas aplicações de monitoramento (LEWIS; LEE *et al.*, 2016).

Alguns fabricantes desses dispositivos de baixo custo, como AQMesh, Vaisala, i-Blades, Libelium e Clarity fornecem, junto com suas plataformas de sensores, serviços de visualização e análise de dados, que são oferecidos como Software como Serviço (SaaS), Plataforma como Serviço (PaaS) ou Sensoriamento como Serviço. Outras iniciativas de dados abertos, lideradas por comunidades e instituições de investigação, fornecem recursos semelhantes. Alguns deles são o *Habitat Map* (HABITAT MAP, 2023), o *Air Quality Egg Portal* (AIR QUALITY EGG, 2023), *Sensor.Community* (SENSOR.COMMUNITY,

2023), o *PurpleAir* (PURPLEAIR, 2023), *Smart Citizen* (SMART CITIZEN, 2023), *uRADMonitor* (URAD MONITOR, 2023) e *IQAir*. O custo de aquisição das plataformas de sensores destas iniciativas varia entre 60,00 – 3.800,00 USD. Alguns deles disponibilizam APIs para registro e acesso aos dados dos monitores mas na maioria dos casos o acesso é pago e condicionado à aquisição de monitores. Em geral, essas plataformas disponibilizam os dados para visualização de forma gratuita mas o restante dos serviços tem custos adicionais. Dentre elas, apenas *Sensor.Community* e *Smart Citizen* fornecem guias para replicação dos dispositivos de monitoramento.

Da revisão realizada observou-se que as iniciativas de código aberto ainda são escassas. As que existem são direcionadas a aplicações específicas com arquiteturas que não buscam facilitar o seu reaproveitamento em outras aplicações de monitoramento. Especificamente no contexto brasileiro não foram identificados empreendimentos que fabricassem monitores de baixo custo próprios. Percebeu-se assim a necessidade de uma iniciativa que não apenas prossesse *software* e *hardware* para determinada aplicação, mas que, através da modularização e o reaproveitamento de código e *hardware*, facilitasse e acelerasse o desenvolvimento de novas aplicações.

Os monitores de baixo custo da qualidade do ar ainda devem alcançar níveis de confiabilidade apropriados para serem utilizados de forma estendida (PENZA, 2020). A literatura reporta diversos trabalhos que buscam reduzir o erro e a incerteza das medições dos sensores de baixo custo mediante a aplicação de modelos de calibração e de compensação (MAAG; ZHOU; THIELE, 2018; CONCAS; MINERAUD; LAGERSPETZ; VARJONEN; LIU *et al.*, 2021). Dentre os modelos de calibração aplicados, as regressões multivariadas paramétricas e não paramétricas têm produzido os melhores resultados (FENG *et al.*, 2019; CONCAS; MINERAUD; LAGERSPETZ; VARJONEN; LIU *et al.*, 2021). O motivo para o sucesso das regressões multivariadas está relacionado com a sua capacidade para considerar múltiplas variáveis. Isso têm sido vantajoso para os sensores de gases, cujas respostas sofrem pelas sensibilidades cruzadas com outros compostos (LEWIS; SCHNEIDEMESSER *et al.*, 2018), assim como pela influência das condições ambientais, como a temperatura (POPOOLA *et al.*, 2016) e a umidade relativa (PANG; SHAW; GILLOT *et al.*, 2018). Particularmente, os modelos paramétricos que utilizam regressões lineares multivariadas (*MLR*, siglas em inglês) (SPINELLE; GERBOLES *et al.*, 2015) costumam produzir os melhores resultados quando os sensores são expostos a concentrações elevadas de poluentes (KARAGULIAN *et al.*, 2019). Nesses cenários, a dinâmica linear entre a concentração e a saída dos sensores prevalece sobre as interferências de outras variáveis (HAGAN *et al.*, 2018). Já nos ambientes com baixos níveis de concentração, as respostas dos sensores são mais sensíveis às não-linearidades introduzidas pelas variações da temperatura e da umidade (HAGAN *et al.*, 2018). Nessas condições, os modelos não paramétricos superam o desempenho da *MLR* (KARAGULIAN *et al.*, 2019). Embora as regressões multivariadas, especialmente as não paramétricas, tenham alavancado o uso de

monitores de baixo custo, ainda é necessário buscar modelos e metodologias de calibração e correção dos dados que se adaptem às condições e características de operação de cada aplicação de sensoriamento (CONCAS; MINERAUD; LAGERSPETZ; VARJONEN; LIU *et al.*, 2021).

Diante do exposto e pelas lacunas observadas, o presente trabalho descreve o desenvolvimento de uma iniciativa para facilitar a colaboração no desenvolvimento de redes de monitoramento de baixo custo e a sua aplicação num cenário real de monitoramento. Sendo assim, define-se como **objetivo geral** da tese desenvolver uma rede colaborativa de medidores de baixo custo para monitoramento da qualidade do ar. Para isso foram definidos como objetivos específicos:

1. Desenvolver dispositivos de medição para serem adicionados à rede de monitoramento
2. Disponibilizar uma API para registro dos dados de monitoramento e acesso a eles em tempo real, assim como bibliotecas de *firmware* para o desenvolvimento dos dispositivos e sua comunicação com a API
3. Identificar erros e interferências nas leituras de um equipamento de baixo custo e comparar as leituras com os dados de uma estação de monitoramento de referência
4. Desenvolver uma metodologia de correção de leituras ruidosas ou em situação de falha de sensores

A iniciativa desenvolvida, sob o nome de CLEAN (Collaborative Low-cost Environmental Air-quality Network), inclui uma *api* aberta para registro e acesso de dados de monitores de baixo custo, bibliotecas de *software* bem documentadas para o desenvolvimento do *firmware* dos monitores e soluções de *hardware* (com documentação e tutoriais) para determinadas aplicações. A API permite que diversos dispositivos de monitoramento de baixo custo enviem seus dados geo-localizados para um servidor remoto para visualização e armazenamento em tempo real. Além disso, a API possibilita a integração com outras aplicações Web para visualização e análise de dados. Esses dados permanecem abertamente disponíveis para posterior processamento e análise. CLEAN também facilita a incorporação de novos periféricos e sensores de *hardware*, reutilizando o código das bibliotecas e garantindo a interação com a API independentemente do *hardware* desenvolvido para cobrir determinada aplicação de monitoramento. Dada a grande versatilidade dos sensores de baixo custo, muitas aplicações poderiam ser monitoradas a partir de diversos cenários contribuindo para uma maior disponibilidade de volumes de dados.

Os dispositivos de *hardware* desenvolvidos para a rede foram dois protótipos de monitores de baixo custo e três equipamentos com placas de circuito impresso. Estes últimos são denominados de monitores CLEAN Arduino Mega para fins de versionamento e identificação. Os cinco dispositivos utilizam sensores de gases eletroquímicos (*EC*) sensíveis

aos poluentes CO , NO_2 , SO_2 , O_3 e $MP_{2.5-10}$. Foram escolhidos esses gases por serem gases de referência para indicar a qualidade do ar segundo definido na Resolução no 491/2018 do CONAMA. Os resultados obtidos com os dois protótipos foram apresentados no congresso internacional CMAS 2020 (CAMPO *et al.*, 2020) e encontram-se fora do escopo da presente tese. Um dos monitores CLEAN Arduino Mega foi instalado no município de Tubarão, no estado de Santa Catarina, junto a uma estação de referência, para análise e correção dos seus dados. São aplicadas metodologias de pré-processamento e correção dos dados utilizando modelos multivariados baseados em *MLR*, *kNN*, *RF*, *ANN*. Os resultados são comparados em termos de R^2 , erro, e correlação com as medições de referência.

No Capítulo 1 apresenta-se uma revisão bibliográfica sobre o estado da arte no monitoramento da qualidade do ar, tanto o realizado para fins regulatórios como o de baixo custo, com especial foco no estado das redes de monitoramento no Brasil. No Capítulo 2 é apresentada a iniciativa CLEAN e seus componentes, i.e.: a API; as bibliotecas de *firmware* e os dispositivos de *hardware* desenvolvidos; e a aplicação *Web Renovar*. A iniciativa CLEAN foi também apresentada em artigo científico na revista Environmental Modelling & Software (CAMPO, 2021). O código principal dos dispositivos de monitoramento e as bibliotecas de *firmware* possuem registro de programa de computador no Instituto Nacional de Propriedade Industrial com número de registro BR512022001116-6. No Capítulo 3 são apresentadas as medições realizadas pelo equipamento instalado em campo, as metodologia aplicadas para pré-processar os dados, sua relação com a temperatura e com as leituras de referência. Por último no Capítulo 4 apresentam-se as metodologias aplicadas para corrigir as leituras do equipamento de baixo utilizando modelos de regressão multivariados.

1 ESTADO DA ARTE NO MONITORAMENTO DA QUALIDADE DO AR

A qualidade do ar e a poluição atmosférica estão estreitamente relacionadas. A poluição do atmosférica resulta da mistura de múltiplos e diversos compostos gasosos que podem ser emitidos por fontes antropogênicas – como processos industriais e de geração de energia, veículos automotores e queimadas – assim como por eventos naturais, como incêndios florestais e erupções vulcânicas (MMA, 2020). O nível de poluição de uma região, além da magnitude das emissões, depende de múltiplos fatores relacionados à topografia e condições meteorológicas, podendo afetar a área onde as emissões foram originadas, locais próximos e inclusive muito distantes (MMA, 2020; WILLIAMS *et al.*, 2014). A qualidade do ar, por outro lado, é uma medida qualitativa usada para indicar o nível de poluição presente no ar (WILLIAMS *et al.*, 2014). O termo abstrai a complexidade dos processos de poluição atmosférica por meio de um Índice de Qualidade do Ar *IQAr* que outorga uma nota de Boa até Péssima à qualidade do ar, relacionando a informação quantitativa de concentração de determinados poluentes com seus efeitos na saúde humana (FRANÇA *et al.*, 2019).

A exposição à poluição atmosférica é prejudicial para a saúde humana e ambiental, acarretando também em perdas econômicas para a sociedade (WORLD HEALTH ORGANIZATION, 2021). Uma série de efeitos adversos têm sido associados à exposição a poluentes atmosféricos, principalmente nos sistemas respiratório e cardiovascular, reduzindo a funcionalidade pulmonar e podendo produzir inflamação de tecidos. Várias doenças respiratórias como asma, bronquites e enfisema podem ser agravadas pela exposição a poluentes gasosos, assim como enfermidades cardiovasculares (WILLIAMS *et al.*, 2014; WORLD HEALTH ORGANIZATION, 2021). Da mesma forma a poluição atmosférica tem efeitos ecológicos negativos, danificando a vida vegetal e animal, e aumentando a acidificação de solos e águas (MMA, 2020; WILLIAMS *et al.*, 2014).

Com o objetivo de reduzir a poluição atmosférica e seu impacto na saúde e bem estar socio-ambiental, ações têm sido tomadas em escala global para regulamentar a qualidade do ar por meio de padrões da qualidade do ar definidos pelas legislações ambientais de diversos países. Esses padrões definem valores e tempos de exposição limites de concentração de determinado poluente que se ultrapassados causariam dano a saúde da população e ao meio ambiente (CONAMA, 2018).

A Agência de Proteção Ambiental norte-americana define como padrões da qualidade do ar as concentrações de Partículas inaláveis grossas ($MP_{2.5-10}$), dos oxidantes fotoquímicos (que inclui ozônio a nível de solo), o Monóxido de Carbono (CO), os óxidos de enxofre e de nitrogênio, e o chumbo (US EPA, 2021). Já a legislação europeia inclui, além desses compostos, ao benzeno (EUROPEU, 2021). A legislação brasileira, na resolução CONAMA No. 491/2018, determina como poluentes padrões o Partículas inaláveis grossas ($MP_{2.5-10}$), as partículas totais em suspensão, o Dióxido de Enxofre

(SO_2), Dióxido de Nitrogênio (NO_2), o Ozônio (O_3) ao nível de solo, o Monóxido de Carbono (CO) e a fumaça (CONAMA, 2018).

Para fins de cumprimento da regulamentação e gestão efetiva da qualidade do ar, o monitoramento dos poluentes atmosféricos se faz indispensável. O monitoramento realizado para esses fins é executado por instituições governamentais através de redes de monitoramento automáticas visando atingir dois objetivos básicos (FRANÇA *et al.*, 2019):

- Verificar o grau de exposição da população aos poluentes atmosféricos, considerando critérios de saúde pública
- Acompanhar tendências de médio e longo prazo para verificar a eficácia dos programas de controle, avaliando a necessidade de aprimoramentos

O monitoramento efetuado para esses fins utiliza estações de monitoramento de referência que se caracterizam por uma alta precisão e confiabilidade nas suas medições. Contudo, as suas leituras estão restritas ao local de medição e não consideram a alta variabilidade espacial das concentrações de poluentes. Devido aos altos preços dessas estações, as redes de monitoramento costumam ter baixa resolução espacial e os equipamentos são instalados em localidades afastadas das fontes de emissão para evitar máximos pontoais. Dessa forma, os dados compilados proveem apenas informações globais relativas a níveis de concentrações de fundo, o que dificulta a compilação de informação confiável e representativa de uma área específica (KUMAR *et al.*, 2015).

A resolução espacial das redes de monitoramento é tão importante quanto a sua confiabilidade. Por exemplo, a medição em escalas de até 100 m (ou microescala) permite mensurar o impacto de fontes veiculares ou de fontes fixas difusas (não efetuadas por chaminés), que emitem poluentes em altas concentrações a escala local (FRANÇA *et al.*, 2019). Medições em microescala podem complementar, e até substituir, os modelos numéricos que predizem o fluxo das massas de ar em ambientes urbanos, assim como facilitar a elaboração de mapas de concentração de poluentes em tempo real e em alta resolução espacial, reconhecer tendências e identificar áreas segundo os níveis de poluição (KUMAR *et al.*, 2015). Contudo, até o momento, devido à onerosidade das metodologias de medição para fins regulatórios, as redes de monitoramento caracterizam-se por serem espacialmente pouco distribuídas.

1.1 MONITORAMENTO DE REFERÊNCIA DA QUALIDADE DO AR

As redes de referência são compostas por estações de monitoramento, certificadas por órgãos competentes, as quais registram ininterruptamente, e em tempo real, as concentrações dos poluentes na atmosfera (CETESB, 2020). As estações podem ser fixas ou móveis e a qualidade das suas leituras é garantida mediante procedimentos padrões de calibração dos instrumentos, de coleta de dados e de pós-processamento. As informações

coletadas pelas estações da rede são enviadas a um computador central, na forma de médias horárias, por meio de sistemas de telemetria. Ali os dados são processadas com base nos padrões legais estabelecidos e podem ser disponibilizados na forma de boletins diários ou relatórios anuais, como um resumo das condições da poluição atmosférica dentro de determinada área (CETESB, 2020). Os métodos de medição de referência utilizados para cada poluente nas estações são documentados no Guia Técnico para o Monitoramento e Avaliação da Qualidade do Ar (FRANÇA *et al.*, 2019), e resumem-se na Tabela 1.

Tabela 1 – Metodologias de monitoramento de referência

Poluentes	Métodos de Medição
Partículas totais em suspensão	Amostrador de Grande Volume (AGV)
Dióxido de enxofre	Fluorescência na região ultravioleta
Dióxido de nitrogênio	Quimiluminescência em fase gasosa
Monóxido de carbono	Fotometria infravermelha não dispersiva
Ozônio	Quimiluminescência
Fumaça	Refletância da luz

Fonte: (FRANÇA *et al.*, 2019)

As redes de monitoramento de referência caracterizam-se por uma elevada precisão e confiabilidade. Contudo, fatores como preço, consumo energético, complexidade de operação e de instalação dificultando seu uso de forma mais extensa. A literatura reporta preços dessas tecnologias em torno de \$ 15.000 – \$ 100.000 (CONCAS; MINERAUD; LAGERSPETZ; VARJONEN; LIU *et al.*, 2021) e consumo de potência aproximadamente entre 0.2 e 1 kW (PIEDRAHITA *et al.*, 2014), sem contar os custos advindos dos procedimentos de instalação, operação e manutenção (KUMAR *et al.*, 2015; FERREIRA *et al.*, 2022). Sendo assim, a onerosidade da solução como um todo limita o número de estações financeiramente viáveis e reduz a resolução e distribuição espacial das medições, inclusive em países e regiões com elevados índices de desenvolvimento econômico como os Estados Unidos e a Europa (KUMAR *et al.*, 2015; JIAO *et al.*, 2016).

No Brasil o monitoramento de referência é ainda bastante limitado. Algumas iniciativas governamentais e de órgãos de pesquisa, a nível nacional e estadual, têm sido implementadas para facilitar o acesso a dados sobre a qualidade do ar como, por exemplo: (CETESB, 2020), (IEMA, 2020) e (IEMA/ES, 2020). No entanto, ainda assim, as redes de monitoramento que têm sido instaladas cobrem escassos pontos das cidades brasileiras e seu desempenho a longo prazo muitas vezes se vê comprometido por falta de manutenção e de pessoal qualificado (OYAMA; ZAMBONI, 2017). Estudos realizados entre os anos 2021 e 2022 sobre o estado do monitoramento da qualidade do ar no Brasil apontam que, das 27 unidades federativas, menos da metade (entre 11 - 12) monitoram a qualidade do ar e atendem a regulamentação vigente (VORMITTAG *et al.*, 2021; FERREIRA *et al.*, 2022). A Tabela 2 resume as unidades federativas reportadas em cada trabalho que possuem rede de monitoramento de referência ativas e gerenciadas por algum órgão público ou entidades

Tabela 2 – Total de estações de monitoramento no Brasil

UF	Total de estações privadas ¹	Total de estações públicas (2021) ¹	Total de estações públicas (2022) ²
AC	2	29	-
BA	-	-	10
CE	-	-	1
DF	-	4	5
ES	6	9	17
GO	-	2	2
MG	32	-	53
MS	3	-	3
PE	3	1	4
PR	-	16	15
RJ	96	65	91
RS	11	2	6
SP	-	90	79

Fonte: (VORMITTAG *et al.*, 2021; FERREIRA *et al.*, 2022)

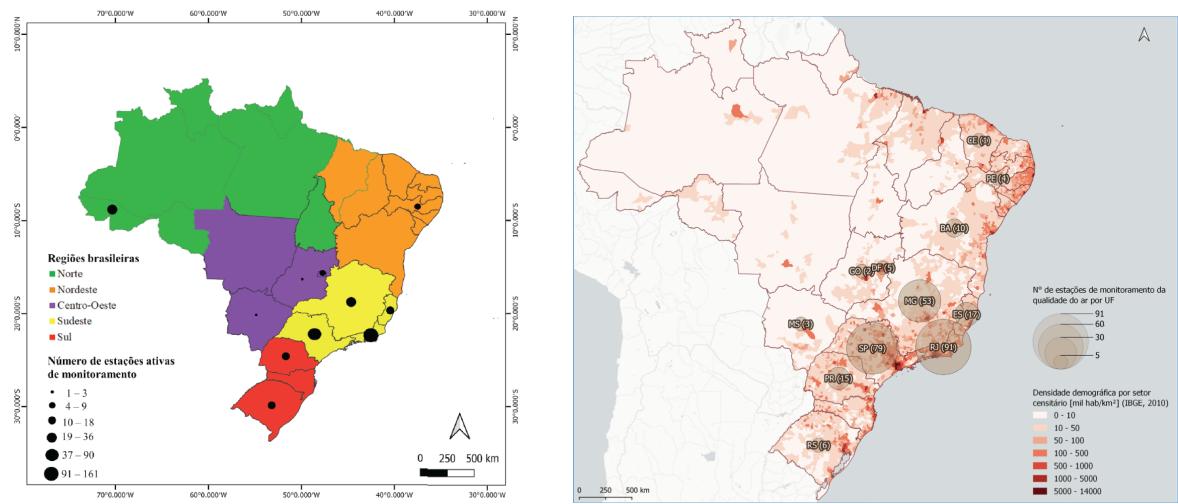
privadas. Existem algumas inconsistências nos dados mostrados em cada trabalho, já que o número de estações muda de forma abrupta em alguns estados, mas é possível extrair uma visão geral da abrangência das redes de monitoramento no país. As Figuras 1a e 1b mostram mapas com o total de estações em cada região do Brasil conforme foi reportado pelos dois trabalhos mencionados. Observa-se nos mapas que as regiões sul e sudeste concentram quase a totalidade das estações (VORMITTAG *et al.*, 2021; FERREIRA *et al.*, 2022) e que na região norte, onde se verificam emissões de poluentes em larga escala provindos de incêndios florestais, não há qualquer monitoramento (FERREIRA *et al.*, 2022). Vale ressaltar também que, embora não incluído nos dados levantados pelos dois trabalhos anteriores, o estado de Santa Catarina conta na atualidade com 3 estações de monitoramento de referência operadas pela Diamante Geração de Energia Ltda. localizadas nos municípios de Tubarão e Capivari de Baixo.

A pesar da sua elevada confiabilidade, a distribuição espacial limitada das redes de monitoramento de referência as habilita para proverem apenas informações globais, relativas a níveis de concentração de fundo (KUMAR *et al.*, 2015). Isto dificulta a compilação de informação confiável e representativa de áreas específicas. Esse fato constitui uma limitante para o estudo dos processos associados à poluição atmosférica. Dada a alta variabilidade espaço-temporal da concentração dos poluentes, especialmente em ambientes urbanos (MEAD *et al.*, 2013), a resolução espacial das redes de monitoramento é tão relevante quanto a confiabilidade das suas estações (JIAO *et al.*, 2016).

¹ Dados levantados por Vormittag e colaboradores (VORMITTAG *et al.*, 2021)

² Dados levantados por Ferreira e colaboradores (FERREIRA *et al.*, 2022)

Figura 1 – Estado do monitoramento de referência da qualidade do ar no Brasil



Fonte: (VORMITTAG *et al.*, 2021)

Fonte: (FERREIRA *et al.*, 2022)

Diante disso, se faz necessário a busca de novas soluções que possibilitem incrementar o número de monitores viáveis nas redes de monitoramento sem afetar a qualidade das medições. Todavia, especial cuidado deve ser tomado para não comprometer a qualidade dos dados obtidos, já que, como apontado por Emily Snyder e colaboradores, ter dados pouco confiáveis é mais prejudicial do que não ter dados, pois podem conduzir a decisões desacertadas (SNYDER *et al.*, 2013).

1.2 MONITORAMENTO DE BAIXO CUSTO DA QUALIDADE DO AR

Segundo Emily Snyder e colaboradores o monitoramento da qualidade do ar tem experimentado uma mudança de paradigma na forma como os dados são coletados (SNYDER *et al.*, 2013). Os avanços recentes em instrumentação eletrônica unido a necessidade de soluções alternativas que complementem as técnicas de monitoramento tradicionais, têm contribuído para um crescente interesse no desenvolvimento de sensores de qualidade do ar de baixo custo (KUMAR *et al.*, 2015; LEWIS; SCHNEIDEMESSER *et al.*, 2018). Esses novos sensores possuem características essenciais como tamanho e peso reduzidos, baixo consumo de potência, baixo custo e facilidade de uso, que os colocam em vantagem com relação aos instrumentos de referência, e que têm criado as condições para aprimorar uma série de aplicações de monitoramento e gerar novas (SNYDER *et al.*, 2013; LEWIS; SCHNEIDEMESSER *et al.*, 2018).

Este tipo de tecnologia possibilitaria a agências públicas, entidades reguladoras e de pesquisa utilizar um maior volume de sistemas de monitoramento, e assim diversificar e complementar as aplicações de monitoramento para fins de pesquisa e regulamentação e validar modelos atmosféricos (LEWIS; SCHNEIDEMESSER *et al.*, 2018). Estes sistemas

podem prover informação temporal qualitativa relevante sobre o nível de poluição atmosférica em uma determinada localidade por períodos de dias a meses (CASTELL; SCHNEIDER *et al.*, 2018), como por exemplo, os momentos do dia em que a poluição é maior ou menor (ZIMMERMAN *et al.*, 2018), ou observar a sua variação ao longo do tempo (CASTELL; DAUGE *et al.*, 2017). Igualmente, os sistemas de baixo custo têm sido utilizados para detectar de áreas com níveis de concentração elevados nas cidades (MEAD *et al.*, 2013) e elaborar mapas de poluição (HUANG *et al.*, 2019).

Para Emily Snyder e colaboradores o uso deste tipo de sensores pode levar a uma melhor proteção da saúde pública e do meio ambiente (SNYDER *et al.*, 2013). Sensores de baixo custo podem ser usados para prover informação representativa de exposição pessoal a poluentes, com elevada resolução temporal (MEAD *et al.*, 2013; JERRETT *et al.*, 2017). Igualmente, esta nova forma de medição, por ser de baixo custo e de fácil operação, pode ser instalada em comunidades (MAHAJAN *et al.*, 2020), escolas e áreas residenciais (CASTELL; SCHNEIDER *et al.*, 2018), provendo informação à população sobre a qualidade do ar que respiram, e colocando os processos de monitoramento e medição nas mãos de comunidades e indivíduos (LEWIS; SCHNEIDEMESSER *et al.*, 2018).

1.2.1 Princípios de funcionamento dos sensores de gases de baixo custo

Segundo seu princípio de funcionamento, os sensores de baixo custo utilizados para a medição de poluentes na atmosfera podem ser classificados em: sensores de material particulado e sensores de gases (MAAG; ZHOU; THIELE, 2018). Os sensores de material particulado funcionam baseados em princípios de detecção óticos que medem o espalhamento ou a adsorção da luz pelas partículas (RAI *et al.*, 2017). Já os sensores de gases podem ser subdivididos em duas classes: os óticos e os que dependem da interação entre o material transdutor e o composto gasoso (SNYDER *et al.*, 2013).

O princípio ótico de medição consiste em expor o composto gasoso a um feixe de luz, com determinado comprimento de onda, e medir o efeito dessa interação com um detector fotossensível (RAI *et al.*, 2017). Dentro desse grupo, os sensores mais comuns são os infravermelhos não-dispersivos (*NDIR*, por suas siglas em inglês), usados para medir CO_2 e CH_4 , e os detectores de foto-ionização (*PID*, por suas siglas em inglês) que utilizam luz ultravioleta e são usados para medir compostos orgânicos voláteis (SNYDER *et al.*, 2013).

Dentre os sensores de gases que operam a partir da interação entre o material transdutor e o composto gasoso, os mais populares são: os semicondutores de óxido metálico (*MOS*, por suas siglas em inglês) e os de princípio eletroquímico (*EC*, por suas siglas em inglês). Esses são os sensores mais comumente utilizados para medir gases tóxicos como CO , NO_x , O_3 e SO_2 (LEWIS; SCHNEIDEMESSER *et al.*, 2018).

Os sensores *EC*, em comparação com os *MOS*, costumam ter um menor consumo

de potência, maior seletividade, menores limites de detecção e uma relação linear com a concentração. Os sensores *MOS*, por sua parte, têm custos menores e seu condicionamento eletrônico costuma ser mais simples (RAI *et al.*, 2017). Neste trabalho serão abordados apenas os sensores eletroquímicos por serem os mais utilizados para o monitoramento de baixo custo da qualidade do ar e pela relação linear entre suas respostas e a concentração de gás.

É importante ressaltar que o atual estado da arte dos sensores de baixo custo impossibilita que eles substituam os métodos de medição de referência. Contudo, resultados promissores têm sido encontrados em várias áreas, principalmente naquelas em que as técnicas convencionais não poderiam ser implementadas devido a limitações como a pouca portabilidade, o elevado consumo energético e o custo.

1.2.2 Iniciativas de monitoramento de baixo custo da qualidade do ar

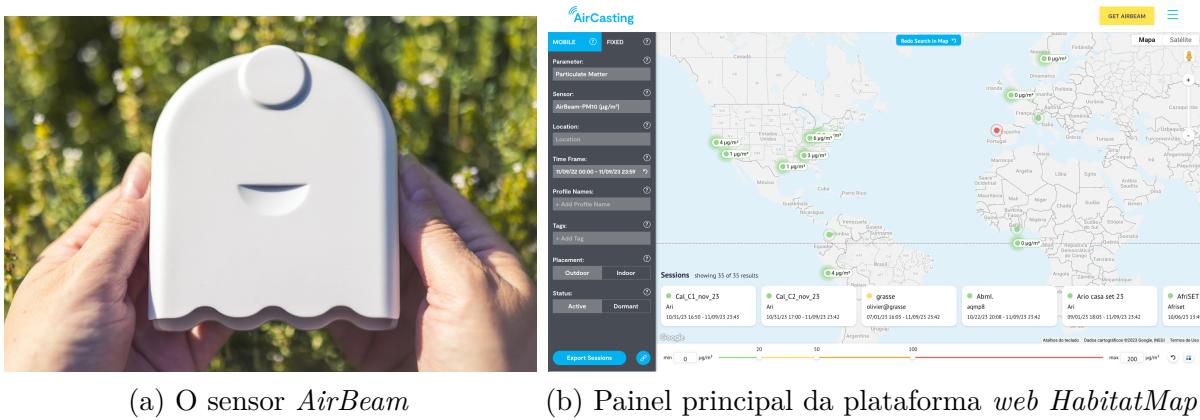
Diversas iniciativas de monitoramento têm implementado e disponibilizado recursos para a coleta e o acesso a dados de redes de monitoramento da qualidade do ar compostas por sensores de gases de baixo custo. Fabricantes como AQMesh, Vaisala, i-Blades, Libelium e Clarity fornecem, junto com os equipamentos de monitoramento, serviços de visualização e de análise de dados, que são oferecidos como Software como Serviço (SaaS), Plataforma como Serviço (PaaS) ou Sensoriamento como Serviço. Os custos destes equipamentos oscila entre 1.000,00 - 10.000,00 EUR, sem incluir custos de operação, e o acesso aos dados dos sensores está condicionado a subscrições nos serviços de nuvem providos pelos próprios fabricantes (KARAGULIAN *et al.*, 2019). Destes, apenas a Clarity fornece um mapa aberto para visualização geo-localizada e das séries históricas das leituras dos sensores instalados pelo globo. Com relação à propriedade intelectual, a exceção da Libelium, os equipamentos dos outros fabricantes mencionados encontram-se na categoria de "caixa preta".

Outras iniciativas, lideradas por comunidades e instituições de investigação, fornecem recursos para o acesso aberto aos dados e alguns deles incluem também código e hardware aberto. Alguns delas são o *Habitat Map* (HABITAT MAP, 2023), *IQAir* (IQAIR, 2023), *Air Quality Egg Portal* (AIR QUALITY EGG, 2023), o projeto *Sensor.Community* (SENSOR.COMMUNITY, 2023), *PurpleAir* (PURPLEAIR, 2023), *Smart Citizen* (SMART CITIZEN, 2023), e o mapa da rede *uRADMonitor* (URAD MONITOR, 2023). O custo de aquisição das plataformas de sensores destas iniciativas varia entre 60,00 – 3.750,00 USD. A maior parte das iniciativas comunitárias disponibilizam os dados e o acesso à plataforma web de monitoramento de forma gratuita, mas é necessário adquirir um sensor para se conectar à rede.

Habitat Map disponibiliza seu sensor *AirBeam* por 249,00 USD. *AirBeam* mede material particulado, temperatura e umidade relativa, possui pequenas dimensões, peso reduzido e é portátil (HABITAT MAP, 2023). O acesso aos dados dos monitores ativos

estão acessíveis de forma aberta e gratuita na plataforma *HabitatMap*, mas o registro de sensores nela está condicionado à compra de sensores *AirBeam*. As informações sobre a fabricação do sensor e da API de monitoramento não estão acessíveis publicamente. O *HabitatMap* disponibiliza também medições de sensores *PurpleAir* e de estações de referência. A Figura 2 ilustra o sensor e o painel principal da plataforma.

Figura 2 – A iniciativa *Habitat Map* e o sensor *AirBeam*



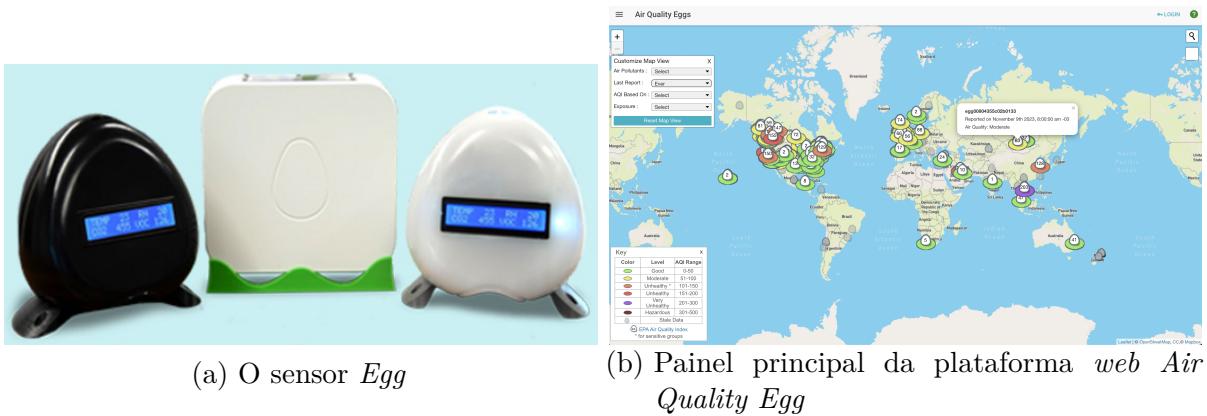
Fonte: (HABITAT MAP, 2023)

Air Quality Egg é uma plataforma de monitoramento de baixo custo com foco no ensino básico e fundamental para abordar temas sobre qualidade do ar, poluição atmosférica e ciência cidadã. Assim como *HabitatMap* os dados de monitoramento estão acessíveis mas para o registro de dados deve ser feito por meio de um dos seus sensores, chamados de *Egg*. É possível adquirir diferentes modelos de *Egg* que medem temperatura, umidade relativa, pressão atmosférica, CO_2 , SO_2 , NO_2 , $MP_{2.5-10}$, O_3 , CO , VOC e H_2S . Os modelos também podem incluir módulo GPS e bateria. Dependendo da configuração de sensor escolhida os valores podem oscilar desde 650.00 USD até 1485.00 USD. Além disso, para acessar outros recursos da plataforma é necessária uma subscrição, que é gratuita para indivíduos mas tem um custo adicional de 2000.00 USD para organizações.

uRADMonitor, *IQAir* e *PurpleAir* possuem modelos de negócio semelhantes aos mencionados acima, no sentido de ser necessário adquirir um sensor para registrar dados nas plataformas. O que os diferencia, contudo, é que disponibilizam mais de um modelo de sensor assim como acesso aos dados e serviços de software através de APIs. *uRADMonitor* vende 12 modelos de monitores da qualidade do ar com preços que vão desde 200.00 uSD até 3750.00 USD dependendo da quantidade de gases que medem, a aplicação e precisão. Os sensores utilizados por *uRADMonitor* são diversos e podem medir $MP_{2.5-10}$, CO , SO_2 , O_3 , NO_2 , formaldeído, VOC , NH_3 , radiação e outros gases tóxicos. Não existe um custo adicional pelo acesso à API além do custo dos dispositivos.

IQAir oferece dois modelos de sensores, um para medições em ambientes internos e outro para ambientes externos, ambos pelo preço de 300.00 USD. Além disso, *IQAir* disponibiliza acesso a sua API de qualidade do ar que inclui acesso a dados de estações de

Figura 3 – O portal Air Quality Egg e o sensor Egg

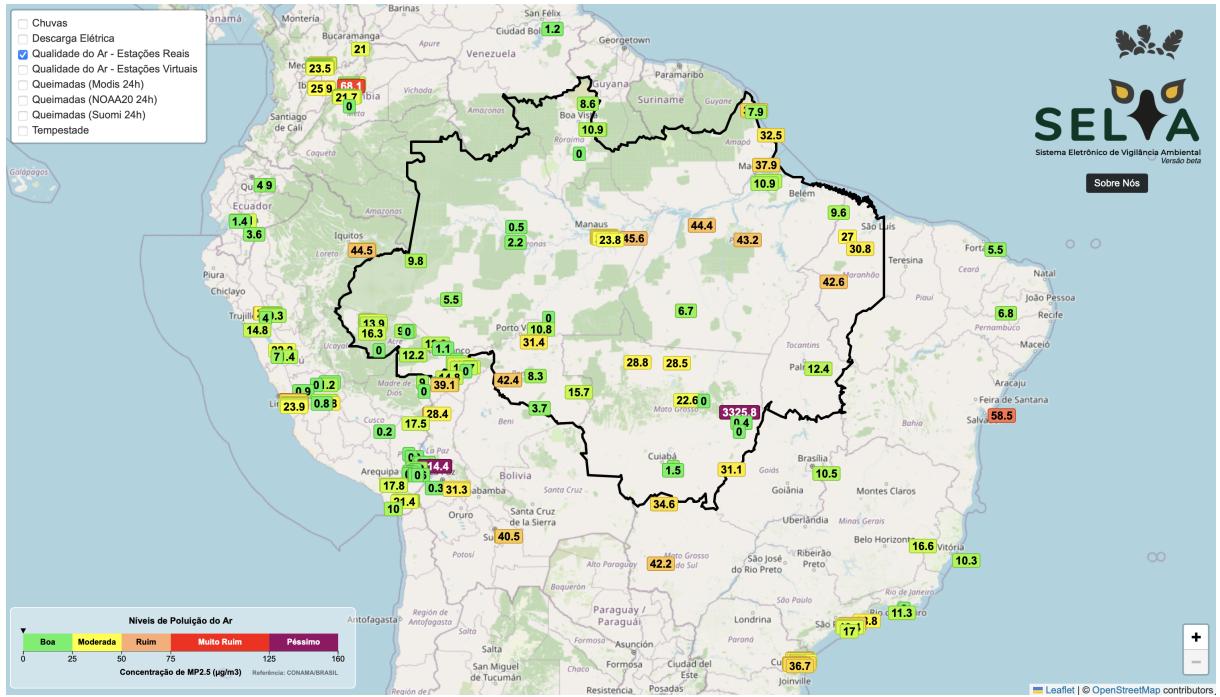


Fonte: (AIR QUALITY EGG, 2023)

monitoramento de referência, dados de sensores de qualidade do ar, modelos de previsão de qualidade do ar e dados meteorológicos. Existem três planos de subscrição para acessar a API. Um deles é gratuito mas oferece recursos limitados e apenas acesso a dados do índice de qualidade do ar e dados meteorológicos. Os outros dois planos possibilitam um maior número de chamadas à API e oferecem dados de concentração de poluentes ($MP_{2.5-10}$, CO , SO_2 , NO_2 e O_3), dados históricos e prognósticos meteorológicos e de qualidade do ar por 7 dias. O valor desses planos é de 400.00 a 1000.00 USD por mês ou 4000.00 a 10 000.00 ao ano.

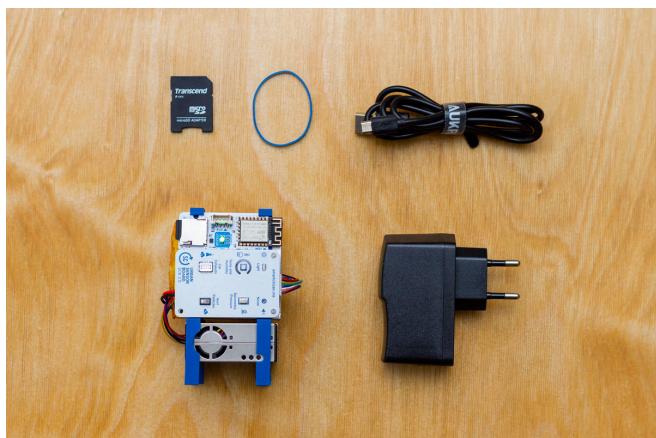
PurpleAir disponibiliza 4 modelos de sensores de qualidade do ar cujos preços oscilam entre 200.00 USD e 300.00 USD. Todos os modelos medem $MP_{2.5}$, o que os diferencia é se são para medição em ambiente externo ou interno e recursos ergonômicos adicionais. O acesso a API não é cobrado mas para a leitura dos dados dos sensores é solicitado pagamento. Vale destacar que sensores *PurpleAir* foram instalados para uso no Sistema Eletrônico de Vigilância Ambiental (SELVA) (RIBEIRO; SOUZA, 2021) desenvolvido numa parceria entre a Universidade do Estado do Amazonas, a *CUOMO Foundation*, a Fundação Universitas de Estudos do Amazônicos e o Ministério Público do Estado do Amazonas, que monitora a concentração de $MP_{2.5}$ em diversas localidades brasileiras. SELVA disponibiliza sensores de qualidade do ar a voluntários que queiram se incorporar ao sistema de monitoramento, mas não inclui nenhuma documentação sobre a fabricação desses sensores nem API própria para acesso aos dados. No Brasil, sensores *PurpleAir* tem sido utilizados para caracterizar o comportamento temporal da concentração de MP_{10} na Amazônia brasileira e mensurar a influência dos incêndios florestais que acontecem nesta região e no Cerrado (JANG; JUNG, 2023; CONNERTON et al., 2023).

Das iniciativas de acesso aberto revisadas, apenas a *Sensor Community* e a *Smart Citizen Map* são efetivamente de código aberto. Além de disponibilizarem seus monitores para venda, ambas fornecem guias para replicação dos seus dispositivos de monitoramento

Figura 4 – Sistema de Monitoramento SELVA utiliza sensores *PurpleAir*

Fonte: (RIBEIRO; SOUZA, 2021)

e para utilização das suas APIs para acesso e registro de dados de monitoramento. Contudo, a documentação é apenas para replicação dos dispositivos já existentes, de acordo com um conjunto de instruções de hardware e software, e não visam facilitar o seu reaproveitamento para criação de novas topologias para aplicações diferentes. Para isso *Smart Citizen* disponibiliza o *Smart Citizen Kit* e *Sensor.Community* disponibiliza o *Sensor Kit #1*, conforme ilustrado nas Figuras 5a e 5b. Os valores dos *kits* são de 119.00 USD e 50.00 EUR respectivamente.

Figura 5 – Os *kits* de *Smart Citizen* e *Sensor.Community*(a) *Smart Citizen Kit* de *Smart Citizen*

Fonte: (SMART CITIZEN, 2023)

(b) *Sensor Kit #1* de *Sensor.Community*

Fonte: (SENSOR.COMMUNITY, 2023)

Estes esforços tem expandido as redes de monitoramento de poluição atmosférica e incorporado as comunidades e os cidadãos no acesso a dados de qualidade do ar. Contudo, o acesso aos dispositivos ainda é limitada, especialmente nos países em desenvolvimento onde o custo do dólar costuma ser mais elevado. Como a maioria destes instrumentos são comerciais, enquadram-se na categoria de “caixa preta” e não facilitam a sua replicação. Por outro lado, as iniciativas de código e de *hardware* aberto desenvolvem com foco em aplicações específicas e seguindo topologias que não visam necessariamente a modularização para reaproveitamento em outras aplicações. Isso ainda representa uma limitante dada a heterogeneidade de aplicações de monitoramento, já que sempre que uma nova aplicação surgir, novos códigos e novos *hardwares* precisarão ser desenvolvidos. Dentro do contexto brasileiro, não foram encontradas iniciativas de produção de sensores de baixo custo em solo nacional.

1.2.3 Limitações do monitoramento de baixo custo da qualidade do ar

Os sensores de gases de baixo custo, como qualquer sistema de medição, possuem fontes de erro internas que são inerentes ao seu próprio funcionamento (MAAG; ZHOU; THIELE, 2018). Estes erros são geralmente conhecidos e fáceis de determinar. Além desses, existem fontes de interferência externas, relacionadas às suas condições de operação, que são mais difíceis de detectar e controlar. Inclusive, sensores de um mesmo fabricante e de um mesmo lote de fabricação, podem apresentar comportamentos diferentes perante a influência de fontes externas (ALPHASENSE, 2013b; CASTELL; DAUGE *et al.*, 2017).

Os fabricantes normalmente definem um intervalo de medição onde os sensores apresentam melhor desempenho. O limite inferior desse intervalo é conhecido como limite de detecção, e todos os valores inferiores a ele são considerados como ruído (MAAG; ZHOU; THIELE, 2018). Esse parâmetro é determinado em condições de laboratório, por isso em condições de operação não controladas, o seu valor pode sofrer alterações levando a erros nas medições. Por exemplo, em ambientes com muitas fontes de ruído electromagnético, ou um circuito de alimentação de energia pouco robusto a flutuações na tensão elétrica, podem aumentar a amplitude do ruído elétrico e, com ele, o limite de detecção dos sensores, afetando a sua resolução.

Outro tipo de erro sistemático que se manifesta internamente são os erros de *offset* e de sensibilidade. Como estes erros são não aleatórios são relativamente fáceis de remover mediante calibrações de laboratório (SPINELLE; ALEIXANDRE; GERBOLES, 2013). As derivas são outra fonte interna de erros produto de alterações na sensibilidade dos sensores devido principalmente ao seu envelhecimento, que dificultam seu uso para monitoramento a longo prazo (FENG *et al.*, 2019). As derivas podem ser eliminadas com re-calibrações periódicas. A frequência das re-calibrações depende do sensor e da concentração de poluentes a que é exposto, podendo chegar a ser quinzenal (CONCAS; MINERAUD; LAGERSPETZ; VARJONEN; LIU *et al.*, 2021).

Fatores externos ao funcionamento dos sensores também produzem interferências nas medições. A influência desses fatores, principalmente das condições ambientais, são identificados por grande parte dos autores como um dos principais desafios no tratamento das respostas dos sensores (MEAD *et al.*, 2013; POPOOLA *et al.*, 2016; RAI *et al.*, 2017; BARON; SAFFELL, 2017). Esses problemas são característicos de medições feitas em ambientes externos, em condições e ambientes reais, não controladas, ao contrário das medições tomadas em condições de laboratório sob as quais o desempenho dos sensores costuma ser muito melhor (CASTELL; DAUGE *et al.*, 2017).

Variações na temperatura e na umidade do ambiente afetam a sensibilidade e o valor de linha base dos sensores (POPOOLA *et al.*, 2016; PANG; SHAW; GILLOT *et al.*, 2018). Particularmente, tem sido observado que em ambientes externos, onde os níveis de concentração costumam encontrar-se na ordem dos ppb, as variações na temperatura e na umidade relativa alteram o valor de linha base em maior medida que a sensibilidade (POPOOLA *et al.*, 2016).

Os fabricantes de sensores muitas vezes disponibilizam informações sobre a relação entre as respostas dos sensores e as variáveis ambientais, junto a modelos lineares de compensação (ALPHASENSE, 2013b; SPEC SENSORS, 2016a). No entanto, essas informações são obtidas a partir de testes de laboratório que simulam condições reais (SPINELLE; ALEIXANDRE; GERBOLES, 2013), sendo válidas apenas em níveis de concentração na ordem dos ppm e em condições similares às dos testes (LEWIS; SCHNEIDEMESSER *et al.*, 2018). Dessa forma, as soluções para compensar os efeitos da temperatura e a umidade relativa providas pelos fabricantes resultam insuficientes para aplicações em campo (PANG; SHAW; GILLOT *et al.*, 2018). Essas informações são especialmente limitadas para aplicações de monitoramento móvel, onde os sensores são expostos a transientes de concentração bruscos e condições ambientais variadas (DELAINE; LEBENTAL; RIVANO, 2019).

Uma solução que o fabricante Alphasense tem aplicado nos seus sensores é a incorporação de um quarto eletrodo, chamado eletrodo auxiliar, cujo sinal de saída é utilizado para compensar os efeitos de variáveis ambientais no valor de linha base do sensor (ALPHASENSE, 2019a). Este eletrodo tem uma composição similar ao eletrodo de trabalho e provê um sinal de corrente de linha base que acompanha as variações do eletrodo de trabalho decorrentes das mudanças na temperatura, a umidade relativa e a pressão, podendo ser subtraída para obter a resposta do sensor ao gás (BARON; SAFFELL, 2017). Idealmente deveria funcionar assim, contudo, na prática tem sido demonstrado que o eletrodo auxiliar não é capaz de acompanhar as variações do eletrodo de trabalho em todo o intervalo de temperaturas de operação, e que portanto, uma simples subtração é insuficiente para gerar uma resposta confiável (WEI *et al.*, 2018). Cross e colaboradores também comprovaram que as correções recomendadas pelo fabricante Alphasense utilizando o eletrodo auxiliar não produzem os níveis de acurácia requeridos

nas medições em ambientes externos (CROSS *et al.*, 2017).

Tem sido reportado que variações bruscas na umidade relativa e na pressão ambiente produzem picos nas respostas dos sensores que invalidam as leituras por intervalos de tempo de até 40 minutos (ALPHASENSE, 2013b; LEWIS; SCHNEIDEMESSER *et al.*, 2018). Igualmente ambientes com valores de umidade muito extremos ou muito poluídos podem saturar os sensores ocasionando falhas e reduzindo sua sensibilidade (ALPHASENSE, 2013b).

Outros fatores que influenciam no desempenho dos sensores são as variações nos níveis de concentração do local de instalação. Os sensores eletroquímicos, por exemplo costumam ter melhor desempenho em locais onde os níveis de poluição são elevados (CASTELL; DAUGE *et al.*, 2017), já que nestas condições a dinâmica senso-gás predomina sobre o efeito das variáveis interferentes (HAGAN *et al.*, 2018). Por exemplo, Hagan e colaboradores comprovaram que o efeito da umidade relativa poderia ser ignorado ao calibrar um sensor *EC* sensível a SO_2 em todo o intervalo de medição do sensor, contudo, para medições abaixo de 25 ppb, o efeito da umidade relativa mostrou-se significativa (HAGAN *et al.*, 2018). Nuria Castell e colaboradores obtiveram coeficientes de correlação maiores com sensores instalados em locais com trânsito intenso do que nos locais com trânsito leve (CASTELL; DAUGE *et al.*, 2017). Eles também constataram que o coeficiente de correlação dos sensores testados em locais com trânsito intenso caiu de forma considerável durante o período de férias, quando o trânsito pelo local foi reduzido (CASTELL; DAUGE *et al.*, 2017). Por esse motivo, outro dos desafios dos monitores de baixo custo é que os instrumentos mantenham um bom desempenho independentemente dos níveis de concentração encontrados no local (CONCAS; MINERAUD; LAGERSPETZ; VARJONEN; PUOLAMÄKI *et al.*, 2019).

Outro problema comum dos sensores de baixo custo é a sensibilidade cruzada, que é a sensibilidade que os sensores têm a outros gases além do gás de interesse (MAAG; ZHOU; THIELE, 2018). Por exemplo, é bem conhecido que os sensores de O_3 são também sensíveis ao NO_2 (PANG; SHAW; LEWIS *et al.*, 2017; ALPHASENSE, 2019a). Outros estudos têm encontrado também sensibilidade a NO_2 em sensores de NO e SO_2 (LEWIS; LEE *et al.*, 2016), assim como a O_3 e CO_2 em sensores de NO_2 (LEWIS; SCHNEIDEMESSER *et al.*, 2018). Este efeito é especialmente desvantajoso em ambientes externos onde o ar é formado por uma mistura complexa de compostos gasosos. Por isso, se faz necessário assegurar que a leitura de um sensor corresponda ao gás para o qual foi projetado sem a interferência de outros compostos.

Uma forma de abordar o problema da sensibilidade cruzada é otimizando o material do eletrodo de trabalho durante a fabricação do sensor, para facilitar ou catalisar apenas as reações do gás de interesse (R. STETTER; LI, 2008). Igualmente, a seletividade pode ser melhorada no circuito de condicionamento, fixando o potencial de trabalho em um valor que favoreça as reações para o gás objeto de estudo (R. STETTER; LI,

2008; ALPHASENSE, 2013a). Em ambientes externos essas abordagens costumam ser insuficientes, sendo necessário utilizar arranjos de sensores nas medições e aplicar técnicas de calibração multivariadas que considerem a resposta global do arranjo (MAAG; ZHOU; THIELE, 2018).

Dada a multiplicidade e complexidade dos fatores que influenciam as medições dos sensores de baixo custo, se faz necessário o estudo cuidadoso e a aplicação de técnicas de calibração que garantam níveis de precisão e confiabilidade aceitáveis para cada aplicação de monitoramento. Isso é de importância especialmente nas aplicações de monitoramento móvel, em que os sensores são expostos a transientes bruscos nas condições de operação.

Com o intuito de regulamentar o uso de monitores de baixo custo segundo a confiabilidade das suas leituras, a Diretiva Europeia para a Qualidade do Ar definiu o Objetivo da Qualidade dos Dados (*DQO*) como requisito necessário para o uso destes instrumentos para fins de indicação da concentração de poluentes atmosféricos (EU, 2008). O *DQO* representa o nível de incerteza aceitável nas medições destes instrumentos sendo de 50% para $MP_{2.5-10}$, 30% para O_3 e 25% para CO , NO_X e SO_2 . A Agência de Proteção Ambiental Norte-americana (*US EPA*), por outro lado, sugere a avaliação dos instrumentos de monitoramento considerando cinco áreas de aplicação (Tabela 3). Cada área tem valores de precisão, viés e completude dos dados que devem ser cumpridos para que um monitor de baixo custo possa ser utilizado em aplicações dentro dessa área (WILLIAMS *et al.*, 2014). Numa direção um pouco diferente dos dois órgãos anteriores, em um estudo mais recente conduzido por Lidia Morawska, foi proposto que, dado o amplo leque de aplicações de monitoramento, os requisitos de desempenho dos monitores de baixo custo fossem definidos de acordo com cada aplicação prescindindo assim de métricas padrões (MORAWSKA *et al.*, 2018). Esta última abordagem, contudo, exige um conhecimento profundo da aplicação e dos requerimentos de desempenho associados (MORAWSKA *et al.*, 2018).

1.3 CALIBRAÇÃO DOS SISTEMAS DE BAIXO CUSTO PARA MONITORAMENTO DA QUALIDADE DO AR

A calibração implica na obtenção de um modelo matemático que converta o parâmetro medido (e.g. adsorção de luz, tensão, condutividade, corrente elétrica, etc) na variável de saída desejada (e.g. concentração do poluente). Na maioria dos casos, os sensores de gases já possuem uma calibração de fábrica. Inclusive, sensores digitais como os do fabricante SPEC Sensors, já são disponibilizados com um *software* embarcado que entrega as leituras de concentração calibradas e compensadas pelos efeitos da temperatura e a umidade relativa (SPEC SENSORS, 2017). Outros fabricantes, como Alphasense, disponibilizam constantes de calibração e modelos lineares de compensação que podem ser aplicados à saída dos sensores para obter as leituras de concentração (ALPHASENSE,

¹ Pontos de elevada concentração de determinado poluente

Tabela 3 – Requerimentos de desempenho dos instrumentos de monitoramento da qualidade do ar segundo área de aplicação

Área de aplicação	Poluentes	Erro	Completude dos dados
Educação e Informação	Todos	< 50%	$\geq 50\%$
Identificação e Caracterização de <i>Hotsspots</i> ¹	Todos	< 30%	$\geq 75\%$
Monitoramento Complementar	Todos os regulados pelo CONAMA e os VOC	< 20%	$\geq 80\%$
Exposição pessoal	Todos	< 30%	$\geq 80\%$
Monitoramento de referência	O_3 CO, SO_2 NO_2 $MP_{2.5-10}$	< 7% < 10% < 15% < 10%	$\geq 75\%$

Fonte: (WILLIAMS *et al.*, 2014)

2019a).

Os modelos e parâmetros disponibilizados nos manuais de usuário, folhas técnicas e notas de aplicação dos sensores de gases são obtidos através de testes de laboratório realizados pelo próprio fabricante em condições controladas. Por isso, quando testados pelo usuário da tecnologia em condições laboratoriais semelhantes, o desempenho dos dispositivos costuma ser favorável. Contudo, as informações providas por estes meios são insuficientes para o amplo e diverso espectro de condições de operação dos sensores (MORAWSKA *et al.*, 2018). Por este motivo, a literatura recomenda a execução de rotinas de testes e calibração que comprovem o comportamento dos sensores e determinem sua aptidão para a aplicação em que serão utilizados (WILLIAMS *et al.*, 2014; LEWIS; SCHNEIDEMESSER *et al.*, 2018).

As rotinas de calibração dos sensores em condições de laboratório possibilitam detectar e eliminar fontes de erro internas relacionadas com a fabricação e princípio de funcionamento dos dispositivos, como por exemplo: limites dinâmicos, erros sistemáticos de offset e sensibilidade e não linearidades (SPINELLE; ALEIXANDRE; GERBOLES, 2013). Nestas rotinas os sensores são expostos a diferentes níveis de concentração dentro de uma câmara de medição que tem acoplado um sistema de diluição e instrumentos de referência para medição de gases. Nesta câmara são gerados, dentro dos intervalos de interesse, cenários artificiais que representem as condições médias de temperatura, umidade relativa, pressão e níveis de concentração de poluentes onde se espera que o sensor será utilizado no longo prazo (SPINELLE; ALEIXANDRE; GERBOLES, 2013). Porém, é consenso na literatura que estas calibrações não são suficiente para validar o desempenho dos sensores em aplicações reais, já que são incapazes de abranger a totalidade de condições operacionais possíveis (MAAG; ZHOU; THIELE, 2018; LEWIS; SCHNEIDEMESSER *et al.*, 2018; MORAWSKA *et al.*, 2018). A presença e as variações não controladas de

múltiplas espécies gasosas e das condições atmosféricas fazem necessária uma segunda etapa de calibração, em campo, que produza resultados mais consistentes com a realidade da aplicação.

As rotinas de calibração em campo possibilitam avaliar o comportamento dos dispositivos de baixo custo em condições que melhor se aproximem do cenário de aplicação de longo prazo. Nelas, os sensores são colocados junto a estações de referência previamente calibrados e validados por agências reguladoras. Para cobrir as diversas condições atmosféricas, a sazonalidade, e a gama de valores de concentração dos gases de interesse e dos gases interferentes, a literatura recomenda uma duração mínima de três meses para a execução das rotinas de calibração e dos testes de validação (SPINELLE; ALEIXANDRE; GERBOLES, 2013). Todavia, tem sido comprovado que sensores calibrados e testados com bom desempenho em determinado local, não reproduzem os mesmos resultados em outras localidades (ZIMMERMAN *et al.*, 2018; HAGAN *et al.*, 2018; CONCAS; MINERAUD; LAGERSPETZ; VARJONEN; PUOLAMÄKI *et al.*, 2019).

Vários fatores ocasionam isso. Um deles está dado pela distribuição log-normal que caracteriza às leituras de concentração, o que dificulta a aquisição de amostras de concentração elevada, em quantidades representativas para treinar os modelos de calibração. Isso ocasiona que as medições dos sensores produzam erros não desprezíveis quando expostos a ambientes com concentrações de poluentes maiores do que as encontradas durante o treinamento do modelo (ZIMMERMAN *et al.*, 2018). Igualmente, a variação das condições atmosféricas, dos níveis de concentração de poluentes e das fontes de emissão de um local para outro, dificulta a aplicação de um mesmo modelo de calibração em locais diferentes, por exemplo: ambientes urbanos vs. ambientes rurais, áreas residenciais vs. áreas com movimento intenso de veículos, ou áreas com predominância de fontes industriais vs. fontes naturais (HAGAN *et al.*, 2018; ZIMMERMAN *et al.*, 2018).

A literatura reporta várias técnicas de calibração e de compensação que têm sido aplicadas nos sensores de baixo custo para o monitoramento da qualidade do ar com resultados promissores. De forma geral, os modelos de calibração utilizados podem ser agrupados em: modelos paramétricos univariados, modelos paramétricos multivariados e modelos multivariados não-paramétricos. Outros autores também têm aplicado algoritmos de compensação que removem os efeitos da temperatura e da umidade relativa das respostas dos sensores.

A regressão univariada é a técnica de calibração paramétrica mais básica reportada na literatura. Esses modelos de calibração ajustam o sinal de saída do sensor (corrente, tensão, resistência elétrica, etc.) para seguir uma curva com valores de concentração de referência, sem considerar a interferência de outras variáveis como a temperatura, a umidade relativa e as sensibilidades cruzadas (MAAG; ZHOU; THIELE, 2018). Estes modelos caracterizam-se por produzir os piores resultados em termos de erro e incerteza das medições em campo.

A regressão linear multivariada é uma regressão linear de múltiplas variáveis. Pelo fato de considerarem outras variáveis além das leituras de um só sensor, este modelo de regressão produz melhores resultados do que a regressão univariada (KARAGULIAN *et al.*, 2019). No entanto, tem sido verificado que em concentrações na ordem dos ppb estes modelos não são capazes de produzir resultados satisfatórios já que, nesta gama de valores, a temperatura e a umidade relativa produzem não linearidades nas respostas dos sensores que o modelo linear é incapaz de reproduzir (HAGAN *et al.*, 2018). Já em concentrações maiores seu desempenho costuma ser tão bom quanto os modelos de aprendizado de máquina, pois nestes intervalos prevalece a dinâmica linear entre concentração do gás e a saída do sensor (HAGAN *et al.*, 2018). Estes modelos, por serem paramétricos, conseguem extrapolar, com erro baixo, novos valores fora do intervalo de treinamento. Isto é vantajoso pois valores altos de concentração são pouco frequentes durante o treinamento em campo, devido à distribuição log-normal que caracteriza as variáveis de concentração.

A regressão não paramétrica, baseada em técnicas de aprendizado de máquina, possibilita elaborar modelos de calibração, a partir de dados coletados, sem a necessidade de complexos modelos paramétricos não-lineares (MAAG; ZHOU; THIELE, 2018). Estas técnicas têm se mostrado uma alternativa eficiente em termos de erro e incerteza, e a maioria deles conseguem resultados iguais ou melhores em comparação com o restante dos modelos de calibração (FENG *et al.*, 2019). Especificamente nos intervalos de baixas concentrações os modelos de aprendizado de máquina são superiores aos paramétricos já que conseguem lidar com as não linearidades que afetam as leituras dos sensores nestes intervalos (MALINGS *et al.*, 2019). Contudo, estes modelos são mais limitados em ambientes com concentração elevada de poluentes devido a dois fatores principais: (1) a limitação para extrapolar valores fora do intervalo de treinamento (HAGAN *et al.*, 2018) e (2) a distribuição log-normal que caracteriza às variáveis de concentração, limitando a quantidade de amostras de concentração elevada que podem ser utilizadas para treinamento (ZIMMERMAN *et al.*, 2018).

Uma alternativa que vem produzindo resultados favoráveis é a aplicação de modelos de calibração híbridos. Estes modelos combinam técnicas de aprendizado de máquina e regressões lineares multivariadas para calibrar os sensores em duas faixas de concentração: altas e baixas. Dessa forma, o modelo híbrido consegue aproveitar o bom desempenho das técnicas de aprendizado de máquina nas baixas concentrações e o dos modelos lineares multivariados nas altas. Dois trabalhos têm reportado a aplicação desta alternativa e em ambos os resultados foram satisfatórios, já que os autores conseguiram reduzir a influência da temperatura e a umidade nas medições assim como reutilizar o mesmo modelo de calibração em diferentes locais com uma boa correlação e baixos valores de erro e incerteza (HAGAN *et al.*, 2018; MALINGS *et al.*, 2019).

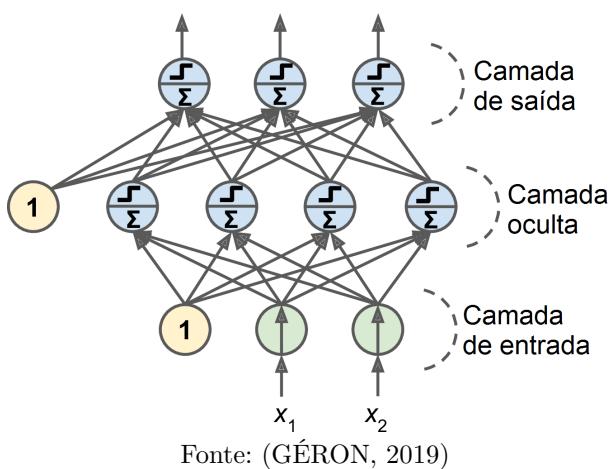
Dada a variedade de técnicas de aprendizado de máquina, é interessante comparar o desempenho daquelas que já tem sido utilizadas em modelos de calibração híbridos,

i.e.: regressão pelos k-vizinhos mais próximos (HAGAN *et al.*, 2018) e as florestas aleatórias (MALINGS *et al.*, 2019). Um algoritmo de aprendizado de máquina bem popular (FENG *et al.*, 2019) dentre os modelos de calibração de monitores de baixo custo é a regressão por redes neurais artificiais (CONCAS; MINERAUD; LAGERSPETZ; VARJONEN; PUOLAMÄKI *et al.*, 2019). Como os modelos híbridos ainda são bem incipientes é interessante comparar o desempenho dessas três técnicas de *ML* dentro de um modelo híbrido. Também, pela robustez que os modelos híbridos têm demonstrado sob condições ambientais variadas, resulta interessante testar seu desempenho em aplicações de monitoramento móvel. Até o momento, não se tem conhecimento de estudos que tenham sido realizados nessa direção.

1.3.1 Redes Neurais Artificiais

As redes neurais artificiais (*ANN*) são sistemas de processamento paralelo inspirados nos neurônios biológicos. Uma das topologias mais populares de redes neurais é a Perceptron multicamadas (Figura 6) (GÉRON, 2019). A rede contém uma camada de entrada, que são as entradas do modelo; uma camada de saída e uma ou várias camadas ocultas. As camadas de saída e as ocultas estão compostas por neurônios artificiais que aplicam uma função de transferência sigmoidal sobre suas entradas (FENG *et al.*, 2019). O objetivo da rede é predizer os valores de saída com o mínimo valor de erro. Para isso é aplicado um algoritmo de otimização que ajusta os pesos das funções de transferência dentro dos neurônios para minimizar o erro. O algoritmo de otimização mais conhecido é o de Retropropagação do Erro (GÉRON, 2019).

Figura 6 – Perceptron Multicamadas



As *ANN* têm sido muito usadas para problemas de predição e generalização já que conseguem modelar inúmeros sistemas. Especificamente nos sistemas de baixo custo para monitoramento de gases têm se mostrado uma solução eficaz (SPINELLE; GERBOLES *et al.*, 2015, 2017; DE VITO *et al.*, 2018). Principalmente em intervalos

baixos de concentração, esta alternativa supera aos modelos *MLR*, pois consegue capturar melhor as não linearidades produzidas pela interferência de outras variáveis. Contudo, em intervalos de concentração elevados, seu desempenho cai já que estes modelos de aprendizado de máquina não conseguem extrapolar valores fora do intervalo de treinamento.

1.3.2 Florestas Aleatórias

A Floresta Aleatória é uma técnica de aprendizado supervisionado de máquinas. Uma floresta aleatória está constituída por um conjunto de árvores de decisão construídos a partir do conjunto de dados de treinamento (GÉRON, 2019). O valor médio obtido do conjunto de árvores é depois usado para predizer novos dados de entrada. A idéia das florestas aleatórias é que a decisão de um conjunto de modelos medíocres é melhor que a de um único modelo bom. Isso também os torna menos propensos a sobre-ajustes.

Em comparação com as redes neurais, as florestas aleatórias precisam de um volume de dados menor e são menos intensas computacionalmente (MONTANTES, 2020). Vários trabalhos sobre monitoramento de gases de baixo custo têm aplicado modelos de calibração deste tipo com resultados satisfatórios, comparáveis aos obtidos com redes neurais (KARAGULIAN *et al.*, 2019). Elas também têm como vantagem que seus resultados são interpretáveis.

1.3.3 K Vizinhos Mais Próximos

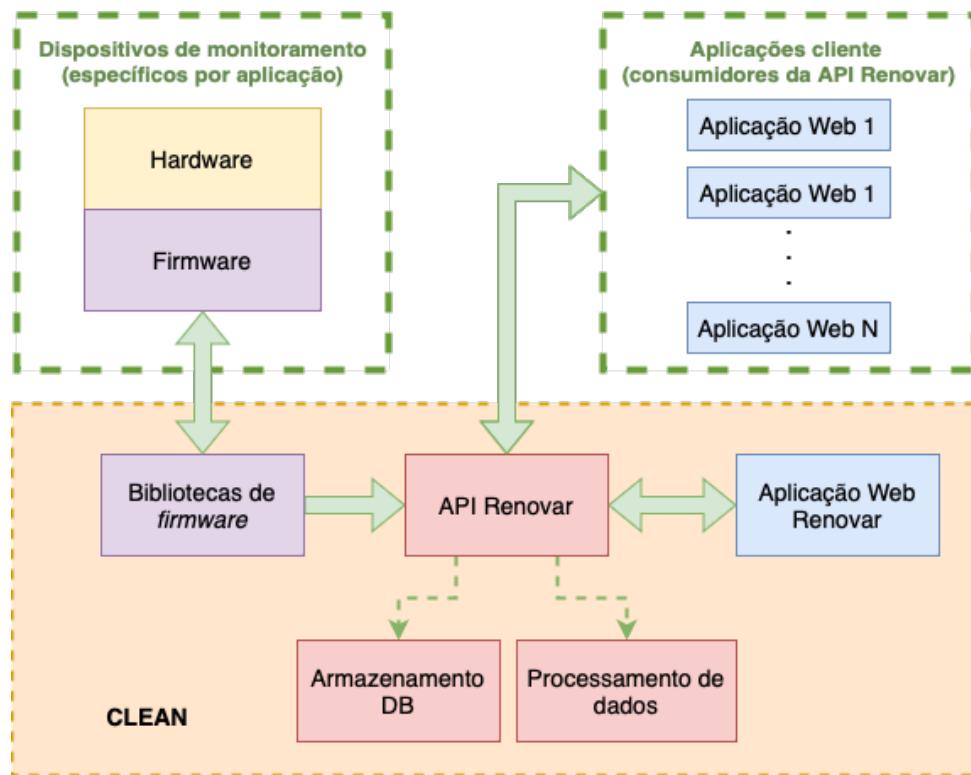
O método *kNN* baseia-se na suposição de que coisas semelhantes se encontram próximas. O algoritmo prediz o valor de determinada amostra de uma variável a partir dos valores das *k* amostras conhecidas mais próximas (ALTMAN, 1992). Para predizer o valor de uma amostra desconhecida, o algoritmo seleciona, dentro do conjunto de treinamento, as *k* amostras que se encontram mais próximas dela, considerando as variáveis de entrada do modelo. Logo, o algoritmo calcula o novo valor como a média dos valores de saída das amostras de treinamento (KRAMER, 2013). Como medida para a distância entre as amostras costuma-se utilizar a distância euclidiana ou a distância Manhattan. O número de vizinhos é determinado mediante uma grade de busca durante o treinamento do modelo para garantir baixo erro na predição de amostras desconhecidas (MILLER, 2019). Este algoritmo consegue modelar bem sistemas não-lineares e o tempo de treinamento é curto, porém o tempo de predição é longo e para conjuntos de dados muito extensos o tempo de processamento pode ser muito demorado. Outra desvantagem deste algoritmo é que seus resultados são difíceis de interpretar pois não é possível saber qual é o peso de cada variável dentro do modelo.

2 CLEAN - COLLABORATIVE LOW-COST ENVIRONMENTAL AND AIR-QUALITY NETWORK

A iniciativa CLEAN³ consiste numa plataforma colaborativa de código aberto para promover e facilitar o desenvolvimento de monitores de qualidade do ar de baixo custo e o acesso remoto a informações sobre a qualidade do ar em tempo real. Possui três componentes principais, conforme enumerado a seguir e ilustrado de forma gráfica na Figura 7.

1. API Renovar para acesso aos dados de dispositivos de monitoramento
2. Bibliotecas de *firmware* baseadas no framework Arduino para programação dos dispositivos de monitoramento e comunicação com a *api*
3. Aplicação *web* Renovar para visualização e análise dos dados na forma de mapas, séries históricas e gráficos de *box-plot*

Figura 7 – Componentes da iniciativa CLEAN



Fonte: Desenvolvido pelo autor (2023)

Além disso, a iniciativa disponibiliza a documentação do *hardware* e do *firmware* dos monitores desenvolvidos dentro desse contexto para facilitar a sua replicação e

³ O conteúdo deste capítulo foi publicado em artigo científico na revista Environmental Modelling & Software (CAMPO, 2021) e possui registro de programa de computador no Instituto Nacional de Propriedade Industrial com número BR512022001116-6

a colaboração entre os diferentes agentes. Todos os guias e documentação relativos ao desenvolvimento do hardware e firmware dos dispositivos até agora concebidos, as bibliotecas implementadas e as ferramentas de desenvolvimento estão abertas e disponíveis gratuitamente na página inicial de CLEAN (CAMPO, 2021).

A API Renovar disponibiliza os recursos necessários para a comunicação remota com dispositivos de monitoramento, o registro das leituras dos dispositivos e o acesso aos dados para visualização, processamento e análise. Dentro da iniciativa CLEAN pretende-se que a API seja uma forma de centralizar o registro (através de dispositivos de monitoramento) e o acesso (através de aplicações cliente) a dados da qualidade do ar de monitores de baixo custo desenvolvidos por parceiros, eximindo eles da necessidade de desenvolver soluções próprias desde zero. Assim, CLEAN pretende acelerar e facilitar a expansão de aplicações de monitoramento de baixo custo e de análise de dados.

As bibliotecas de *firmware* providenciam uma interface entre o *hardware* e *firmware* específico para cada aplicação de monitoramento e a API, facilitando também o processo de desenvolvimento, e permitindo aos parceiros focarem na aplicação em si, sem investir muito esforço na integração com a API. As bibliotecas de *firmware* também providenciam abstrações para os componentes de *hardware*, como sensores, controladores e periféricos, contribuindo ao reaproveitamento de código. A medida que os parceiros forem incorporando novos componentes nas suas aplicações, novas abstrações podem ser adicionadas às bibliotecas, expandindo o leque de interfaces a componentes de *hardware*.

A aplicação *web* Renovar desenvolvida dentro do contexto de CLEAN é uma aplicação *front-end* consumidora da API. Ela é uma alternativa para iniciar qualquer aplicação de monitoramento de baixo custo já que disponibiliza a visualização de mapas dos dispositivos, séries temporais dos dados dos sensores e gráficos de *box-plots* dos dados. Pode funcionar também como modelo para outras aplicações consumidoras que venham a ser desenvolvidas.

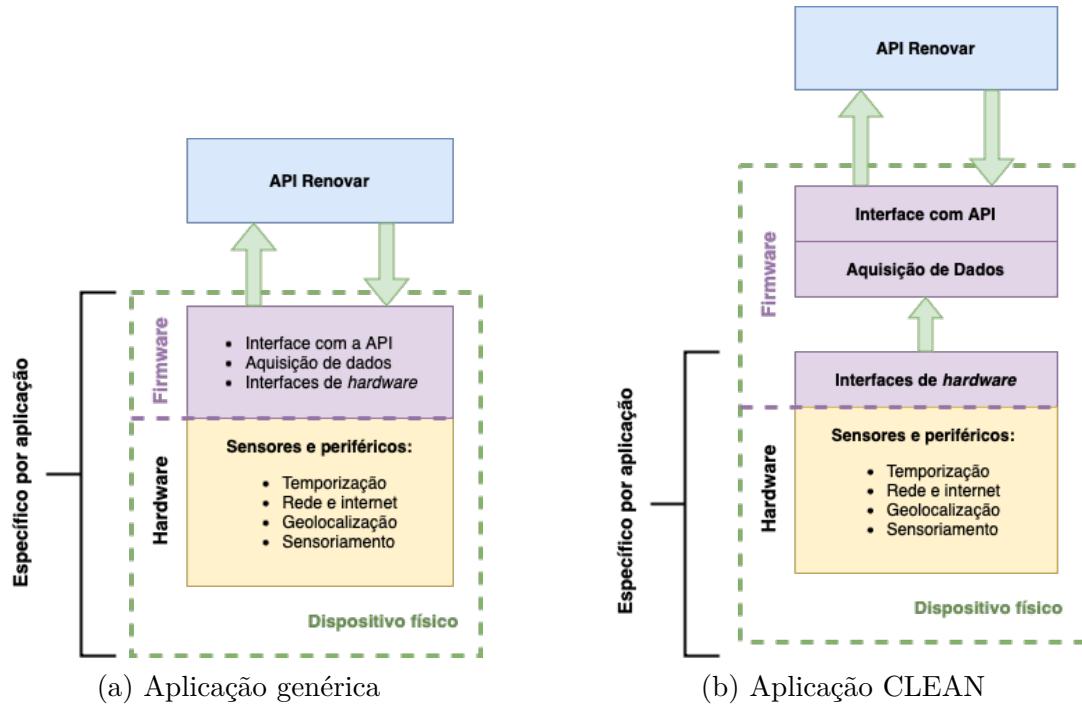
A continuação são descritas as estruturas das bibliotecas de *firmware*, da *API* e os dispositivos de *desenvolvidos*. Nos Apêndices D e E descrevem-se também os códigos gravados nos microcontroladores Arduino e ESP8266 dos equipamentos CLEAN.

2.1 BIBLIOTECAS DE FIRMWARE

O *firmware* dos dispositivos foi desenvolvido no *Framework Arduino*, que é uma abstração de códigos-fonte e bibliotecas comuns a diversas plataformas de *hardware*. O *framework* fornece bibliotecas de código escritas em C/C++ para programação de microcontroladores e interação com dispositivos periféricos. Esta estrutura torna possível escrever programas para controlar uma ampla gama de placas microcontroladoras de Arduino e de outros fabricantes.

Para programar todas as funcionalidades do *firmware* CLEAN, o código foi estruturado em um conjunto de classes em C++ separados em módulos ou bibliotecas.

Figura 8 – Estrutura de desenvolvimento de aplicação de monitoramento de baixo custo



Fonte: Desenvolvido pelo autor (2023)

Esta estrutura foi concebida visando sua reutilização em outras plataformas de microcontroladores e outros componentes de *hardware* suportados no Framework Arduino (como ESP8266 da Espressif) e também para facilitar a revisão e manutenção do código. Para melhor compreender a utilidade desta estrutura comparem-se os diagramas da Figura 8.

A Figura 8a representa a estrutura de desenvolvimento de uma aplicação genérica de monitoramento. Para a criação de um monitor de qualidade do ar de baixo custo que envie dados para a API requer-se de um dispositivo físico composto por uma camada de *hardware*, com sensores e periféricos, e uma camada de *firmware* que realiza o interfaceamento do elementos de *hardware*, a aquisição dos dados e sua transmissão para a API. Seguindo essa topologia, cada nova aplicação que for abordada precisaria de novas iterações pelas mesmas etapas de criação e desenvolvimento com mínimo reaproveitamento de iterações anteriores.

A Figura 8b ilustra a solução proposta pela iniciativa CLEAN. O intuito é desacoplar o que é específico de cada aplicação e o que é comum para todas as aplicações que visem utilizar a API Renovar para registro dos dados de monitoramento. Dessa forma, os componentes de *hardware* e o *firmware* para interfacear estes componentes ficariam sob responsabilidade dos desenvolvedores das aplicações, enquanto as funcionalidades relacionadas à aquisição de dados e comunicação com a *api* estariam cobertas dentro do conjunto de classes e funcionalidades disponibilizadas pelas bibliotecas de CLEAN. Dessa forma maximiza-se o reaproveitamento de código minimizando o esforço para implantar

novas soluções de monitoramento.

As classes desenvolvidas para o projeto estão distribuídas em quatro módulos principais, conforme mostrado na Figura 9: o módulo *Hardware Interfaces*, o módulo *System Drivers*, o módulo *Sensors* e o módulo *Data*.

O módulo *Hardware Interfaces* agrupa todas as classes e estruturas utilizadas para se comunicar com o *hardware* periférico, como sensores, módulos de temporização, módulos de geolocalização e módulos de armazenamento. Os *Drivers* implementam funcionalidades que podem ser utilizadas pelo programa principal independentemente do *hardware* utilizado em cada dispositivo. O pacote *Sensors* está no mesmo nível dos *Drivers* e pode ser interpretado como um conjunto de *drivers* especiais para os sensores, mas com a particularidade de ser específico para cada fabricante. Por fim, o pacote Dados engloba todas as funcionalidades relacionadas à preparação de dados de sensores para armazenamento e transmissão. Este pacote abstrai as informações de concentração adquiridas pelos sensores de gás a partir de detalhes específicos sobre o funcionamento e operação de seu *hardware*.

2.1.1 O módulo de interfaces de *hardware*

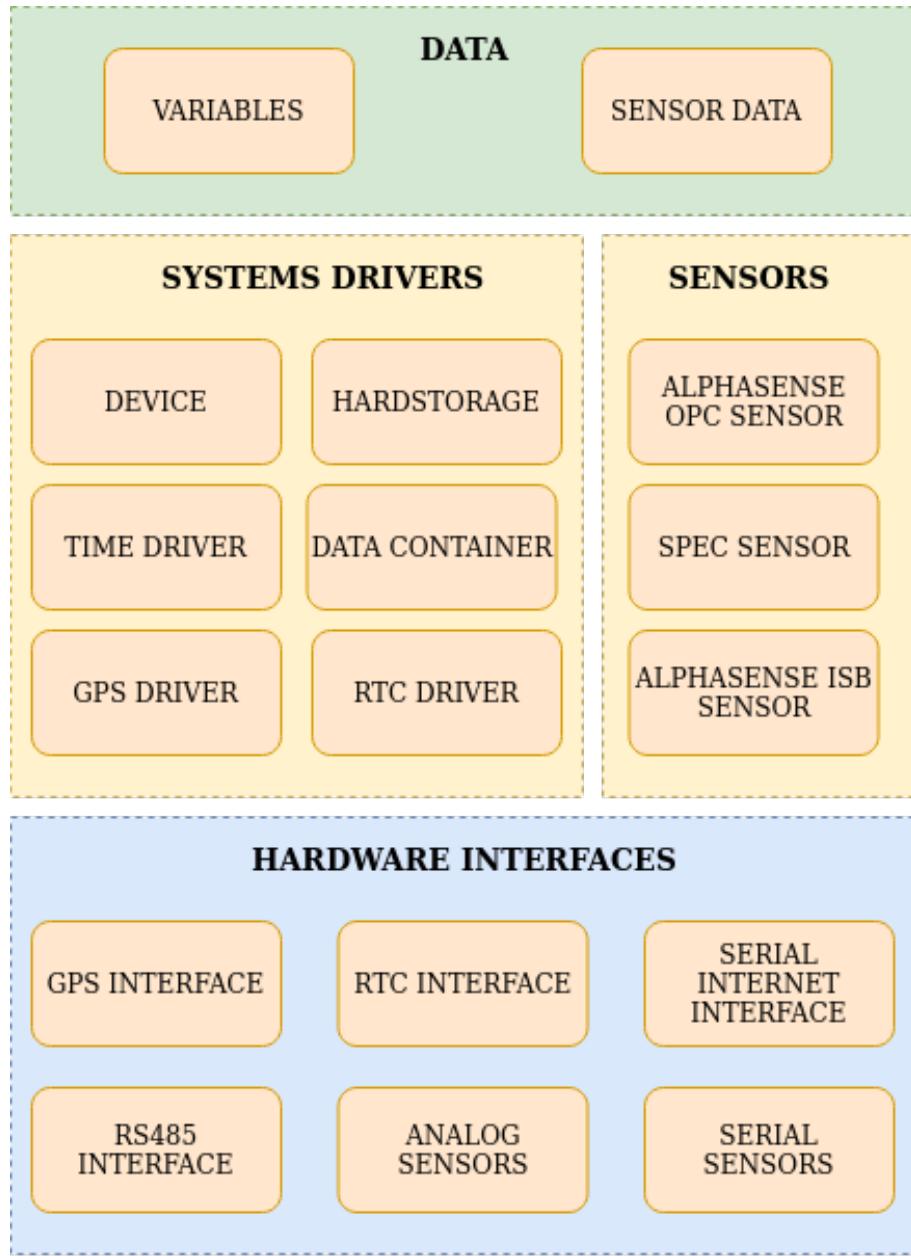
As Interfaces de *Hardware* abrangem as funcionalidades relacionadas à comunicação e interface de sensores de gás, módulos de geoposicionamento (GPS) e relógio de tempo real (RTC) que foram utilizados nos equipamentos desenvolvidos. O modo de operação e a saída dos sensores e de cada dispositivo de *hardware* determinarão o seu esquema de conexão ao microcontrolador e a forma como sua leitura é implementada no *firmware*.

A Figura 10 mostra um diagrama das classes que foram implementadas para a versão atual do *firmware*. As classes `SerialSensorInterface` e `AnalogSensorInterface` implementam interfaces para sensores digitais e analógicos, respectivamente. `SerialSensorInterface`, em particular, implementa uma interface para um sensor digital conectado através de um barramento UART ou RS-485, por meio das classes filhas `UARTSensorInterface` e `RS485SensorInterface`.

Cada classe implementa seu próprio método `sense()` que recebe como parâmetros um ponteiro para um objeto `Stream` (geralmente uma porta serial do microcontrolador), e um ponteiro para um `SerialParser`, que analisa as cadeias de caracteres com comandos ou dados enviados pelo sensor digital. O `SerialParser` é implementado numa camada superior pelas classes do módulo `Sensor`.

A interface com um dispositivo serial para conexão à internet foi implementada através das classes `SerialInternetInterface` e `ESPSerialInterface`, esta última representando a conexão com o microcontrolador ESP8266. Foram criadas mais duas interfaces para módulos GPS e RTC. Na versão atual do *firmware*, foram utilizadas as bibliotecas `TinyGPSPlus` e `RtcDS3221` para cada módulo respectivamente, porém, qualquer outra biblioteca ou módulo também pode ser usado, desde que seja criado

Figura 9 – Conjunto de bibliotecas utilizadas para *firmware* dos dispositivos CLEAN

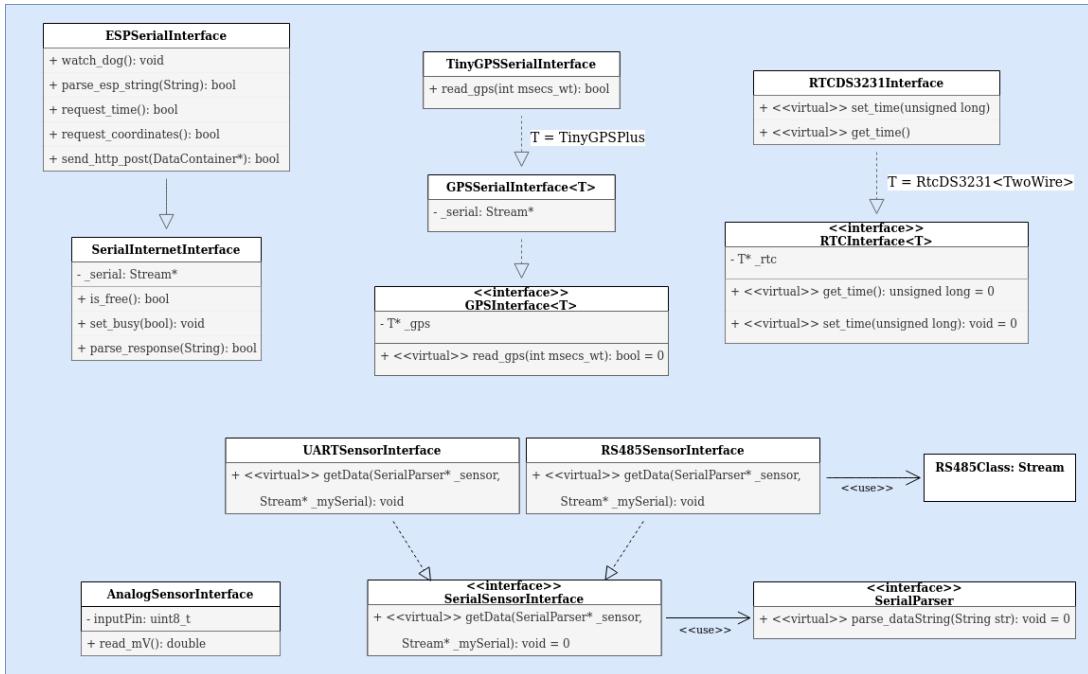


como uma classe filha de `GPSSerialInterface` e `RTCInterface`. Para isso, as classes filhas deverão implementar os métodos virtuais: `readGPS()`, `set_time()` e `get_time()` respectivamente.

2.1.2 O módulo *drivers*

Os *Drivers* atuam como uma camada intermediária entre as Interfaces de *Hardware* e o programa principal. Eles abstraem o *hardware* dos dispositivos do código principal, permitindo sua reutilização independentemente dos módulos e bibliotecas utilizadas em um nível inferior. Alguns *drivers* implementados para o *firmware* foram:

- O driver `HardStorage`, para armazenamento de dados em cartão *SD*;

Figura 10 – Diagramas de classes do pacote *Hardware Interfaces*

- O **RTCDriver** para a Interface RTC;
- O **GPSDriver** para a Interface GPS;
- O **TimeDriver** para gerenciamento das fontes de tempo no dispositivo, as quais podem provir de um módulo RTC, um módulo GPS ou de um servidor NTP

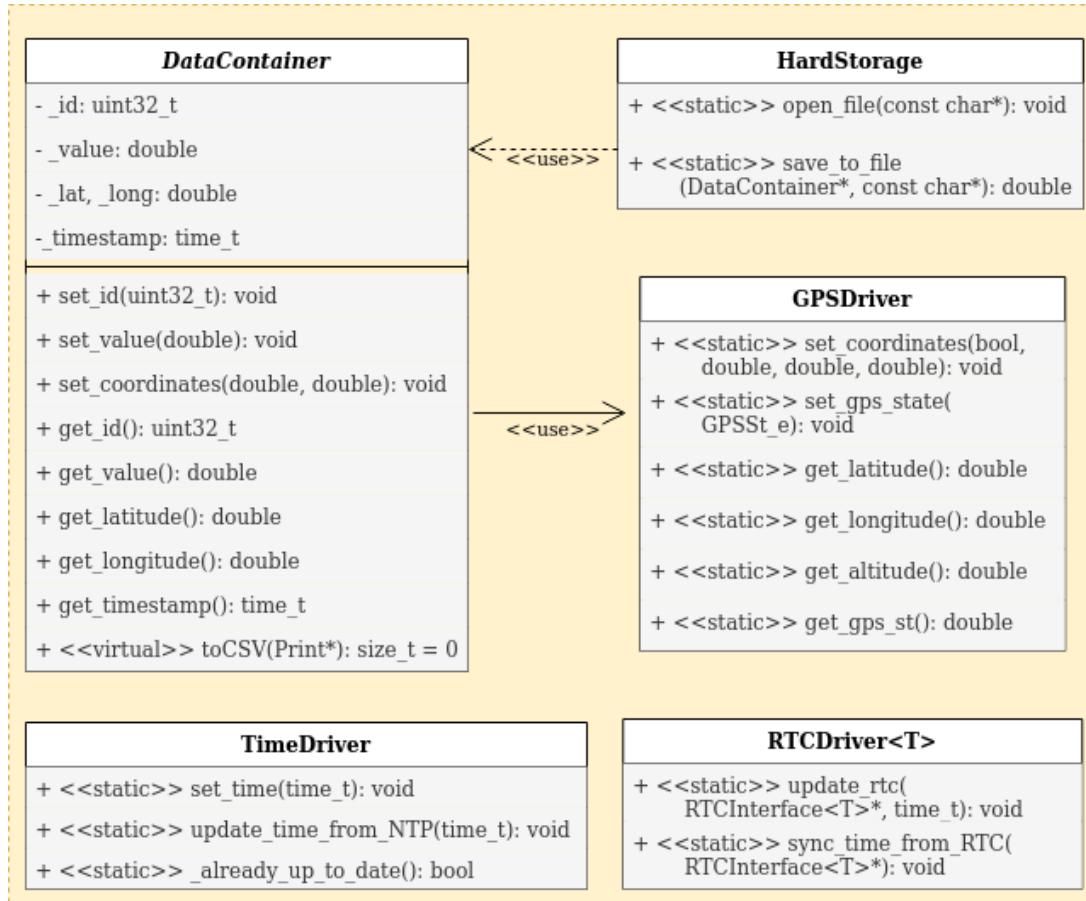
Esses quatro *drivers* usam métodos estáticos, o que significa que podem ser usados sem necessidade de ter um objeto implementado no código.

Os outros dois *drivers* que foram implementados estão relacionados ao tratamento dos dados. São eles o **DataContainer** e o **Smoother**. A Figura 11 mostra o diagrama de classes deste módulo. A continuação são resumidos alguns dos principais métodos e atributos de cada uma das classes pertencentes ao módulo **Drivers**.

2.1.2.1 TimeDriver

Esta classe registra a data e hora internas do sistema e fornece métodos para retornar informações de data e hora em diferentes formatos. O método **set_time(time_t)** define a data e hora do sistema. Internamente, ele invoca o método **setTime()** da biblioteca **Time.h** do framework Arduino. Recebe como parâmetro um número inteiro de 32 bits contendo a data e hora fornecidas por alguma fonte de relógio externa (um módulo GPS, um módulo **rtc** ou um servidor **ntp**).

Figura 11 – Diagrama de classes do Módulo Drivers



2.1.2.2 GPSDriver

Esta classe controla a interface com um módulo gps, armazena as informações das coordenadas geográficas do sistema e fornece métodos para acessá-las, sendo eles:

- `static get_latitude(): double`
- `static get_longitude(): double`
- `static get_altitude(): double`
- `static get_gps_st(): GPSSt_e`

Esses métodos fornecem as informações de geolocalização armazenadas no `GPSDriver`, bem como o estado dessas informações. As informações de geolocalização podem estar OK ou desatualizadas. Esses dois valores são retornados como uma enumeração do tipo `GPSSt_e`.

Outros dois métodos definem as coordenadas geográficas do sistema e o estado dessa informação. Esses métodos são chamados por uma instância de `GPSInterface`. Eles são:

- `static set_coordinates(): void`

- static set_gps_state(): void

2.1.2.3 RTCDriver

Esta classe controla a interface com um módulo RTC. O método `update_rtc(RTCInterface*, time_t)` é chamado sempre que o módulo RTC precisa ser atualizado. Recebe como parâmetro um ponteiro para a instância do `RTCInterface` que será atualizada e a data e hora. O método `sync_time_from_RTC(RTCInterface*)` retorna a data e hora a partir do ponteiro ao tipo `RTCInterface` passado como parâmetro

2.1.2.4 DataContainer

Esta é uma classe abstrata que contém informações sobre a leitura de uma variável. Essas informações são: o identificador da variável que está sendo medida e o valor dessa variável; as coordenadas; e a data e hora onde o valor foi medido. Objetos desta classe são usados para armazenar dados no cartão SD e para enviar postagens *HTTP*. O método `toCSV(Print*)` é um método virtual puro para formatar os dados de uma leitura de variável e armazená-los em um arquivo CSV. Por se tratar de um método virtual, ele deve ser implementado pelas classes filhas de `DataContainer`. Desta forma cada aplicação pode ter seu próprio formato de armazenamento das informações.

2.1.2.5 HardStorage

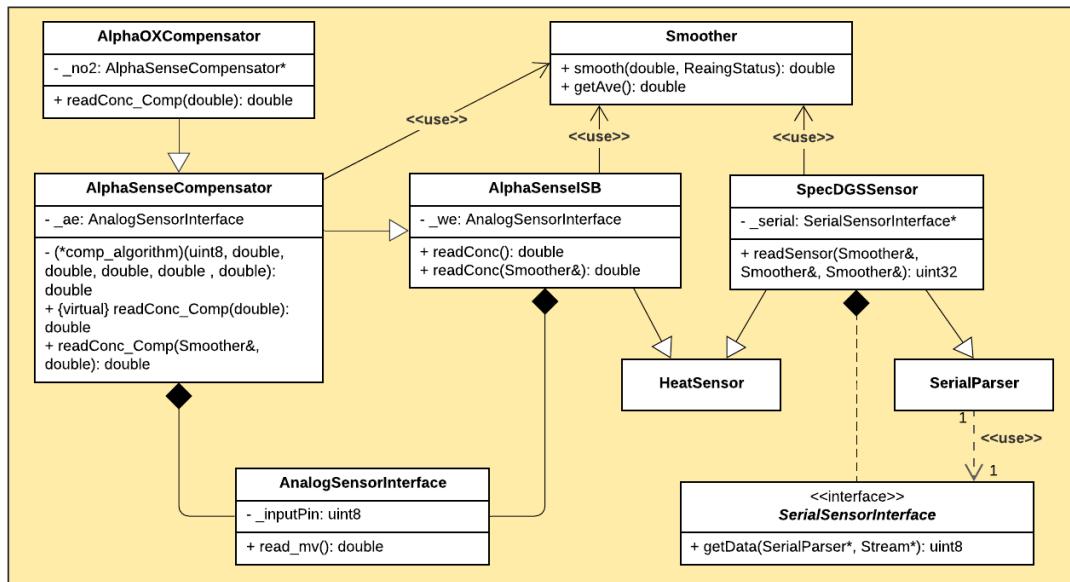
Esta classe contém os métodos para leitura e gravação de e para um cartão SD. Para leitura, o método `open_file(const char*)` abre o arquivo no qual as operações de leitura/gravação serão executadas. O método recebe o nome do arquivo como parâmetro. Já para a escrita no cartão, o método `save_to_file<T>(DataContainer*, const char*)` grava dados em um arquivo no cartão SD. O nome do arquivo é passado como parâmetro, juntamente com os dados a serem salvos. A função espera um ponteiro para um `DataContainer`, que na versão atual do *firmware* são objetos do tipo `SensorData`. O objeto `SensorData` implementa o método `toCSV(Print*)`, que recebe um ponteiro para o arquivo e armazena os dados nele.

2.1.3 O módulo Sensores

As classes deste pacote encapsulam a lógica de leitura de cada sensor, considerando as especificações de cada fabricante. Eles fazem uso das interfaces de sensores implementadas no pacote de Interfaces de *Hardware*. Dois fabricantes de sensores foram utilizados no *hardware* dos equipamentos desenvolvidos no contexto deste trabalho: Alphasense e SPEC Sensors. As interfaces dos sensores Alphasense e SPEC diferem na forma como foram implementadas. As saídas dos sensores Alphasense são dois sinais de tensão analógicos. Os sensores SPEC, por outro lado, fornecem os valores de concentração

de gás, temperatura e umidade em uma cadeia de caracteres que é enviada através de uma interface UART. A Figura 12 mostra um diagrama das classes implementadas para este módulo.

Figura 12 – Diagramas de classes do módulo Sensors



A base para o interfaceamento dos sensores Alphasense é a leitura de duas entradas analógicas do microcontrolador utilizando a função `analogRead()` do *framework* Arduino. Por esse motivo, a classe base para modelagem dos sensores Alphasense é a classe `AnalogSensorInterface`. Ela representa uma entrada analógica identificada pelo atributo `_inputPin`, e seu método `read_mv()` converte o valor digital adquirido pelo conversor analógico-digital do Arduino, em um valor de tensão entre 0 – 5 V. Este método pode receber como parâmetro uma referência a um objeto do tipo `Smoother`, que por sua vez deve estar associado a um objeto `Variable`. Assim, são vinculadas as variáveis físicas modeladas no *firmware* com a respectiva interface de *hardware*; neste caso uma entrada analógica.

A classe `HeatSensor` representa um sensor que precisa de um tempo de aquecimento para funcionar. A lógica que determina a validade das leituras dos sensores é implementada dentro desta classe, levando em consideração um período de aquecimento para cada sensor. Do `HeatSensor` derivam as classes que representam os sensores Alphasense e SPEC, uma vez que ambos são sensores eletroquímicos amperométricos que requerem um intervalo de aquecimento para garantir que as leituras sejam válidas. A continuação resumem-se as principais propriedades das classes relacionadas ao interfaceamento dos sensores de gases.

2.1.3.1 AlphaSenseISB

Esta classe representa um sensor Alphasense com um circuito de condicionamento do tipo ISB. O sufixo "ISB" indica que o circuito de condicionamento usado é a placa de

detectação individual do fabricante do sensor. Esta classe não incorpora nenhum algoritmo de compensação.

Atributos: `_we: AnalogSensorInterface`: Este é um atributo privado que representa a entrada analógica conectada ao eletrodo de trabalho (WE) do sensor

Métodos: `readConc(): double` `readConc(Smooth&): double` Estes são métodos públicos que convertem o valor de tensão lido pelo atributo `_we` em um valor de concentração, levando em consideração a sensibilidade do sensor informada pelo fabricante. A referência ao objeto `Smooth` associa o sensor à variável física correspondente e retorna um valor suave das leituras da variável.

2.1.3.2 AlphaSenseCompensator

Derivado do `AlphaSenseISB`, representa um sensor Alphasense com um algoritmo de compensação. Os sensores da série Alphasense B4 podem usar diferentes algoritmos de compensação dependendo do gás ao qual são sensíveis. Por isso, cada algoritmo é inerente a cada objeto e não à classe

Atributos: `_ae: AnalogSensorInterface`: Este é um atributo privado que representa a saída do eletrodo auxiliar (AE) do sensor eletroquímico. O valor de saída deste eletrodo é usado nos algoritmos de compensação.

Métodos: `(*comp_algorithm)(uint8, double, double, double, double): double`: Este é um ponteiro para a função que implementa o algoritmo de compensação. As funções recebem como parâmetros as variáveis necessárias para o cálculo do algoritmo, dentre eles a temperatura.

`virtual readConc_Comp(double): double` `readConc_Comp(Smooth&, double): double` Estes são métodos públicos que leem os valores de tensão armazenados nos atributos `_we` (herdados do `AlphaSenseISB`) e `_ae`. Eles aplicam o algoritmo de compensação correspondente e retornam um valor de concentração. Ambos os métodos recebem como parâmetros a temperatura ambiente e uma referência a um objeto `Smooth`, como no `AlphaSenseISB`.

2.1.3.3 AlphaOXCompensator

Este é um caso especial para sensores de ozônio que utilizam um algoritmo de compensação. Os sensores de ozônio medem, na verdade, a soma das concentrações de ozônio e dióxido de nitrogênio, portanto, o valor da concentração de dióxido de nitrogênio é exigido pelo algoritmo de compensação.

Atributos: `_no2: AlphaSenseCompensator*`: Para acessar o sensor de dióxido de nitrogênio, a classe `AlphaOXCompensator` usa um ponteiro para um objeto `AlphaSenseCompensator` que representa o sensor de dióxido de nitrogênio.

Métodos: `readConc_Comp(double): double`: Este método lê o valor da concentração do sensor de ozônio e aplica um algoritmo de compensação considerando

também a concentração de dióxido de nitrogênio. Para vincular essas leituras a um objeto do tipo **Variable**, a classe **AlphaOXCompensator** utiliza o mesmo método **readConc_Comp()** herdado da classe **AlphaSenseCompensator**, que recebe uma referência a um objeto do tipo **Smoother**.

2.1.3.4 Interface com sensores seriais

A interface com os sensores SPEC é realizada através da classe abstrata **SerialSensorInterface**. Esta classe fornece métodos para a leitura dos sensores através da porta serial do microcontrolador. A comunicação entre os sensores e o Arduino pode ser implementada através de uma interface UART ou através de um barramento RS-485. Ambas as interfaces de comunicação são modeladas nas classes **UARTSensorInterface** e **RS485SensorInterface**, que derivam de **SerialSensorInterface**.

A classe **specDGS_sensor** funciona como uma camada intermediária entre a interface de *hardware* e as classes do módulo **Data**. Por representar um sensor eletroquímico que necessita de um período de aquecimento, esta classe também herda da classe **HeatSensor**. As instâncias de **specDGS_sensor** têm a finalidade de ler e analisar as cadeias de caracteres enviadas pelos sensores SPEC, com as medições de temperatura, umidade e concentração de gás. Esta classe também é responsável por validar as medições levando em consideração o tempo de aquecimento dos sensores e possíveis erros na comunicação serial. O método **readSensor()** lê os valores de concentração, temperatura e umidade e os disponibiliza aos objetos de tipo **Variable** correspondentes por meio das referências **Smoother** que recebe como parâmetros.

O atributo **_serial** da classe **specDGS_sensor** é um ponteiro para um objeto do tipo **SerialSensorInterface**, que é atribuído durante a construção de cada instância **specDGS_sensor**. O ponteiro pode ser um objeto do tipo **UARTSensorInterface** ou **RS485SensorInterface**, dependendo apenas da interface de comunicação implementada no *hardware*. Os objetos **specDGS_sensor** representam os sensores SPEC, e as instâncias derivadas da classe abstrata **SerialSensorInterface** representam a interface com esses sensores, que no *hardware* é uma única porta serial.

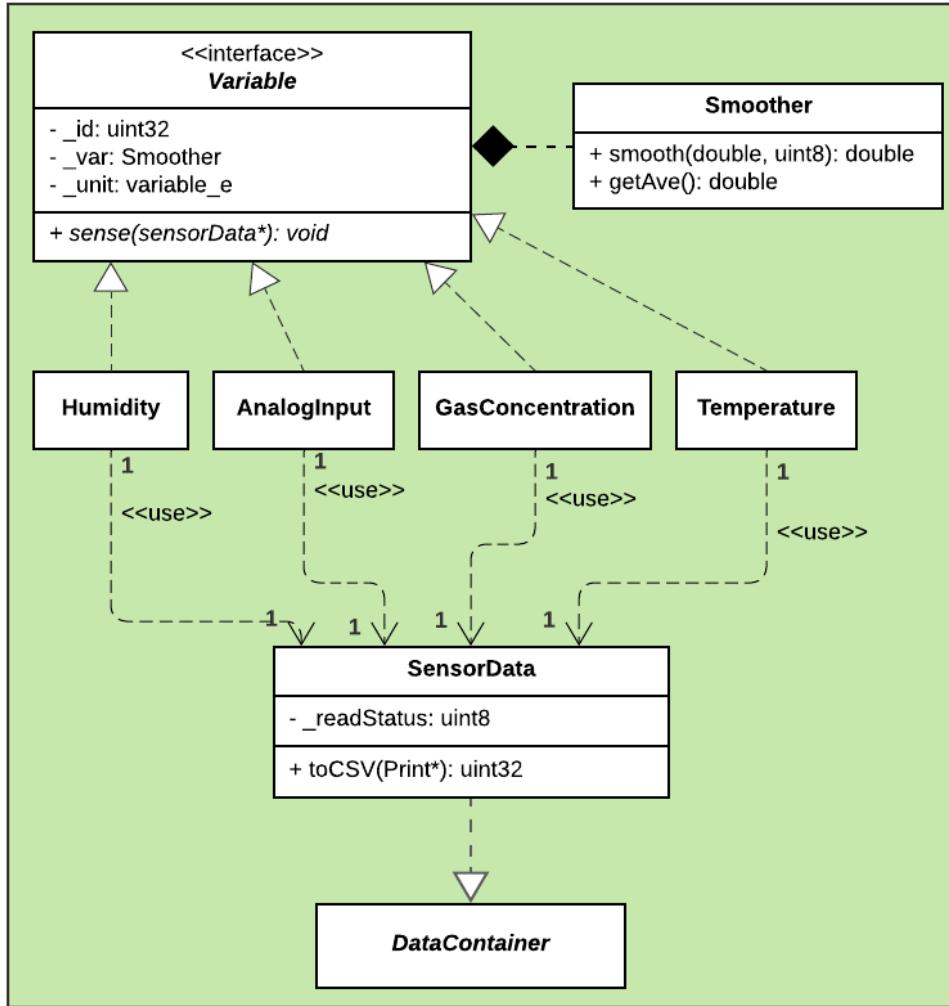
2.1.4 O módulo Data

A Figura 13 mostra o diagrama de classes do módulo **Data**. Como já mencionado, este módulo funciona como uma camada intermediária que prepara e formata as medições obtidas no *hardware* do sensor para seu armazenamento e transmissão. É formado por duas classes principais: **Variable** e **SensorData**.

A classe **SensorData** método prepara os dados para transmissão remota e armazenamento local. Cada objeto do **SensorData** está associado a um único objeto do tipo **Variable**, que representa uma variável física com um identificador único. Vale ressaltar que, embora no *firmware* cada variável física seja representada por um único

identificador, no *hardware* uma ou mais dessas variáveis podem estar vinculadas a um mesmo transdutor. O número de identificação que representa cada variável física é o que associa cada objeto da classe **Variable** ao objeto do tipo **SensorData** correspondente. Este número é armazenado em cada classe nos atributos `_sensorID` e `_id` como valores inteiros de 32 bits.

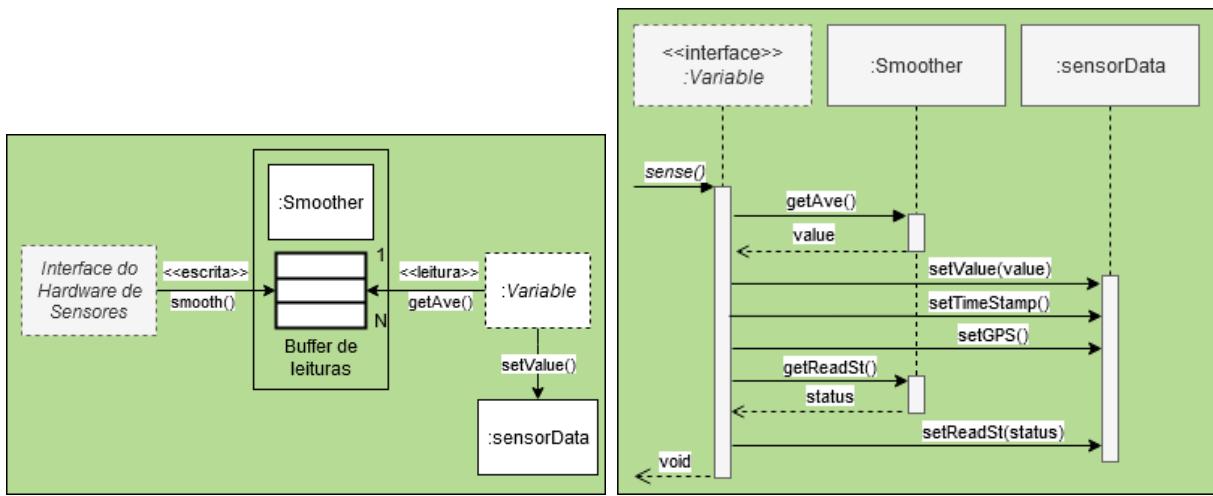
Figura 13 – Diagrama de classes do pacote Data



Objetos do tipo **SensorData** contêm o valor das variáveis físicas às quais estão associados, juntamente com informações sobre a data, hora e local onde as medições foram feitas. O valor de cada variável é armazenado no atributo `_value`, o qual pode ser um dado bruto medido em determinado instante de tempo ou uma média de valores adquiridos durante uma janela temporal. O método `toCSV()` consolida e prepara as informações do valor medido pelo sensor, sua geolocalização, e a data e hora em que a medição foi realizada, no formato CSV.

A classe **Variable** atua como uma camada intermediária entre a camada de *hardware* do sensor e a classe **SensorData**. Os objetos desta classe representam as variáveis físicas que estão sendo monitoradas, porém, não contêm suas quantidades, pois esses valores

Figura 14 – Processo de leitura de uma variável

(a) Processo de escrita e leitura no buffer da classe *Smoother*(b) Diagrama de sequências do método *sense()*

Fonte: Desenvolvido pelo autor (2023)

estão armazenados em objetos do tipo `SensorData`. Como já mencionado, o atributo `_id` contém o identificador da variável física que representa. O atributo `_unit` representa a unidade de medida da variável física que está sendo monitorada. O atributo `_var`, do tipo `Smoother`, funciona como um buffer de memória no qual os objetos que implementam a interface de *hardware* dos sensores podem colocar as amostras da variável medida. Desse mesmo buffer, o objeto `SensorData` associado pode extrair o valor médio das amostras. O número de amostras de cada buffer depende da capacidade que for programada. Os diagramas nas Figuras 14a e 14b ilustram este processo.

Os objetos que representam os sensores, gravam as leituras de cada variável física no buffer de amostragem através do método `smooth()` da classe `Smoother` (14a). Já a classe `Variable` acessa a média das amostras invocando o método `getAve()` do atributo `_var`. Esse valor médio é transferido para o objeto `SensorData` associado por meio do método `setValue()`. O processo de leitura e transferência do valor médio das amostras para o objeto `SensorData` acontece dentro do método `sense()` definido na classe `Variable`. A Figura 14b mostra o diagrama de sequência para este método.

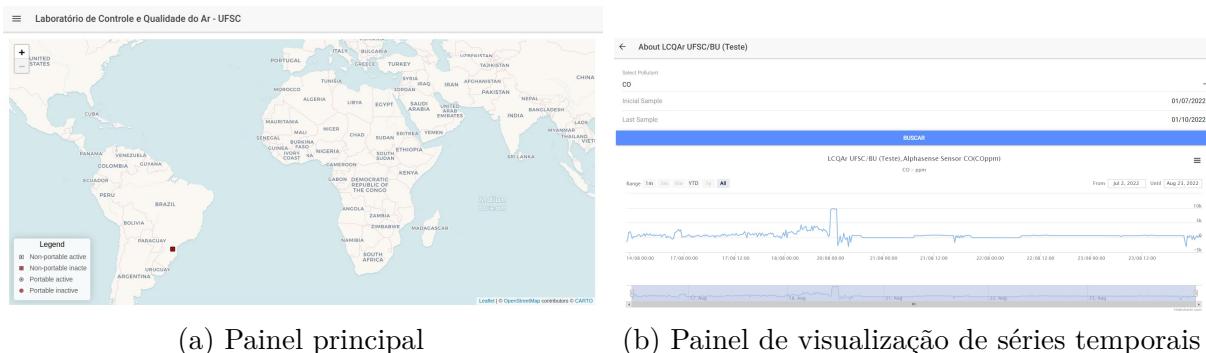
A função `sense()` é um método virtual puro, portanto, as instâncias que derivam da classe abstrata `Variable` devem implementá-la. A sequência principal de ações executadas pelo método é comum a todas as classes derivadas. Quando o método `sense()` é invocado, a classe filha de `Variable` acessa, através do método `getAve()` de `Smoother`, o valor médio das amostras. Este valor é então passado para o objeto do tipo `SensorData` associado através do método `setValue()`. O objeto do tipo `SensorData`, por sua parte, armazena a data, a hora e o local onde foi feita a medição, bem como o status da leitura (método `getReadSt()`).

Os tipos de variáveis físicas implementadas na versão atual do *firmware* foram: Temperatura, Umidade, Concentração de gás e Entrada Analógica. Esta última representa uma tensão analógica que pode ser lida como um sinal de tensão entre 0 – 5 V. Essas variáveis foram modeladas em classes filhas de **Variable** como **Temperature**, **Humidity**, **GasConcentration** e **AnalogInput**. Por serem classes filhas, todas possuem os mesmos atributos de **Variable**, mas cada uma implementa seu próprio método **sense()**.

2.2 A API RENOVAR

Renovar é uma plataforma Web que fornece dados de sensores de ar para visualização e uma API para integração de sensores de ar IoT. Um MVP da plataforma foi desenvolvido em parceria com o Departamento de Informática e Estatística da UFSC (TEIXEIRA, 2018) e foi continuado no LCQAr nos anos seguintes. A plataforma é composta por um banco de dados, um serviço de *back-end* – desenvolvido em linguagem Java utilizando Spring Boot –, e uma aplicação *front-end* criada em Angular, com Ionic e TypeScript. Seu acesso é gratuito e aberto para pesquisas e análises ambientais. A Figura 15a ilustra o painel principal da plataforma, o qual consiste em um mapa que mostra a localização dos dispositivos de monitoramento. A Figura 15b ilustra o painel de séries temporais, onde o usuário consegue visualizar o histórico de determinada variável num intervalo de tempo selecionável.

Figura 15 – Aplicação web Renovar

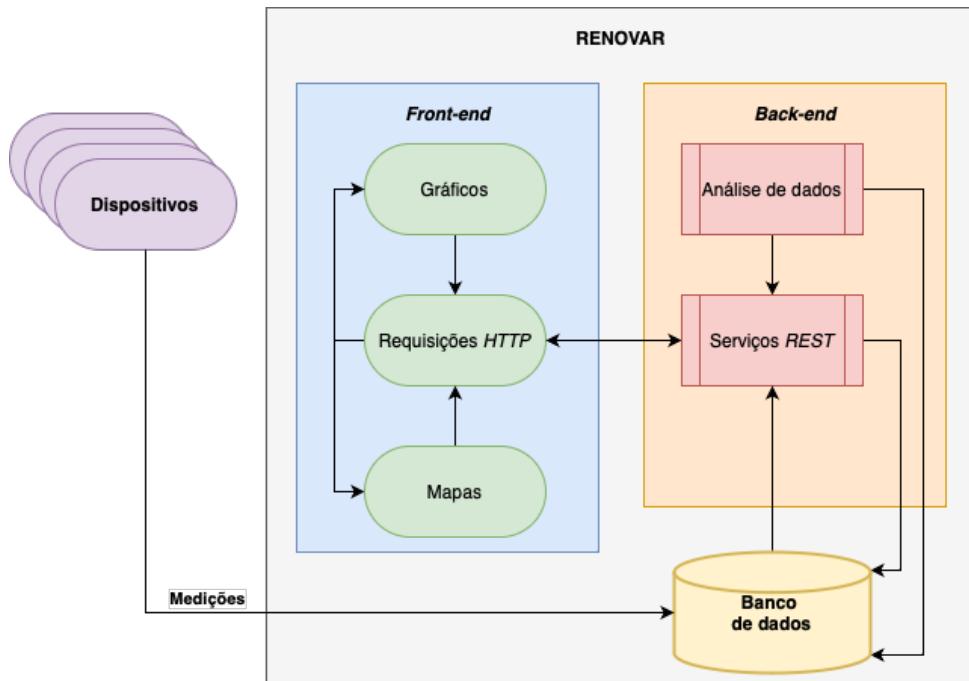


Fonte: Desenvolvido pelo autor (2023)

O sistema recebe dados de dispositivos IoT, como concentração de poluentes atmosféricos, temperatura e umidade relativa. Os dados são armazenados em um banco de dados como séries temporais que podem ser visualizadas online na plataforma *web*. O software consiste em (1) um banco de dados MySQL que armazena as leituras do dispositivo e demais dados necessários à plataforma, como usuários, dispositivos cadastrados, poluentes e unidades; (2) um back-end RESTful, desenvolvido em Java utilizando Spring Boot, que é responsável por coletar dados do banco de dados e prepará-los para o frontend; e (3) o front-end, desenvolvido para ser multiplataforma,

disponibilizando a interface com o usuário conforme ilustrado nas imagens da Figura 15. O banco de dados, backend e frontend estão hospedados em um servidor da Universidade Federal de Santa Catarina.

Figura 16 – Estrutura da aplicação Web Renovar



Fonte: Desenvolvido pelo autor (2023)

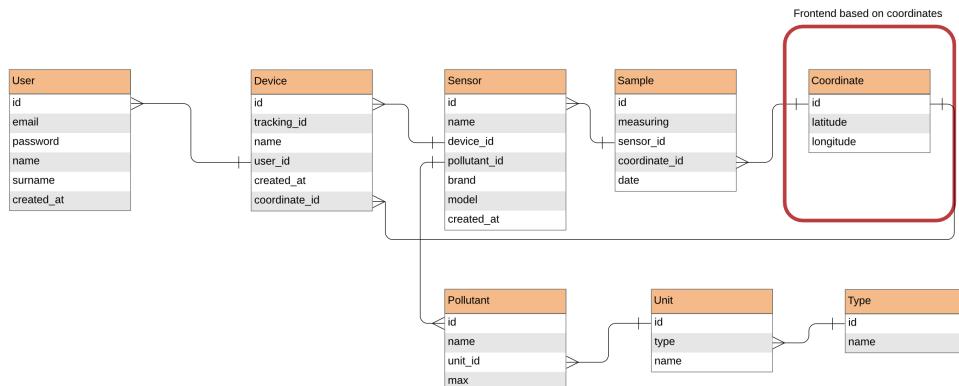
A Figura 16 ilustra o funcionamento do serviço em geral. Os dispositivos coletam dados ambientais e os enviam para o banco de dados pela internet. O backend recebe solicitações do frontend e coleta os dados necessários do banco de dados. Caso os dados necessitem de algum tratamento (ex.: cálculo de valores médios ou filtragem), o backend executa as operações necessárias e envia as informações processadas de volta ao frontend. O frontend, por outro lado, implementa a interface com o usuário e gera os resultados das operações solicitadas, como visualizar os dados como séries temporais e baixar os dados como arquivo CSV.

2.2.1 Banco de dados

O banco de dados foi construído utilizando *MySQL* como linguagem de consulta e *phpMyAdmin* para administração. A Figura 17 ilustra a estrutura do banco, que consiste em oito entidades, cada uma com a sua função, atributos e relacionamentos. Elas são as entidades *User*, *Device*, *Sensor*, *Sample*, *Coordinates*, *Pollutant*, *Unit* e *Type*.

A entidade *Device* representa na prática os dispositivos de coleta, mais especificamente os monitores de qualidade do ar. A entidade *Sensor* representa um sensor de determinada variável física. Relaciona-se com a entidade *Device* com uma cardinalidade 1:N, ou seja, um dispositivo pode ter N sensores enquanto um sensor pode

Figura 17 – Entidades do banco de dados Renovar



Fonte: Desenvolvido pelo autor (2023)

pertencer apenas a um dispositivo. *Pollutant* representa as variáveis que são monitoradas no meio ambiente pelos dispositivos IoT. A relação entre *Pollutant* e *Sensor* também está caracterizada por uma cardinalidade 1:N, já que um sensor possui um único poluente, mas um mesmo poluente pode estar associado a N sensores. Cada poluente tem associado também uma unidade de medida e um tipo de unidade. A entidade *Sample* representa uma leitura de determinada variável física; existem N amostras por sensor. Cada amostra tem associada uma coordenada, que é a localização geográfica onde o valor da amostra será mostrado no mapa. Os dispositivos para medição em locais fixos também tem associados uma Coordenada e consequentemente, todas as amostras dos sensores desse dispositivo devem possuir os mesmos valores de latitude e longitude. Por último, a entidade *User* é utilizada para controle de acesso à plataforma. Embora o acesso aos dados de Renovar não precise de uma etapa de login, a escrita ou envio de dados para a plataforma precisa de um cadastro prévio. Assim, cada dispositivo deve ter um usuário associado para conseguir armazenar as leituras no banco de dados de Renovar.

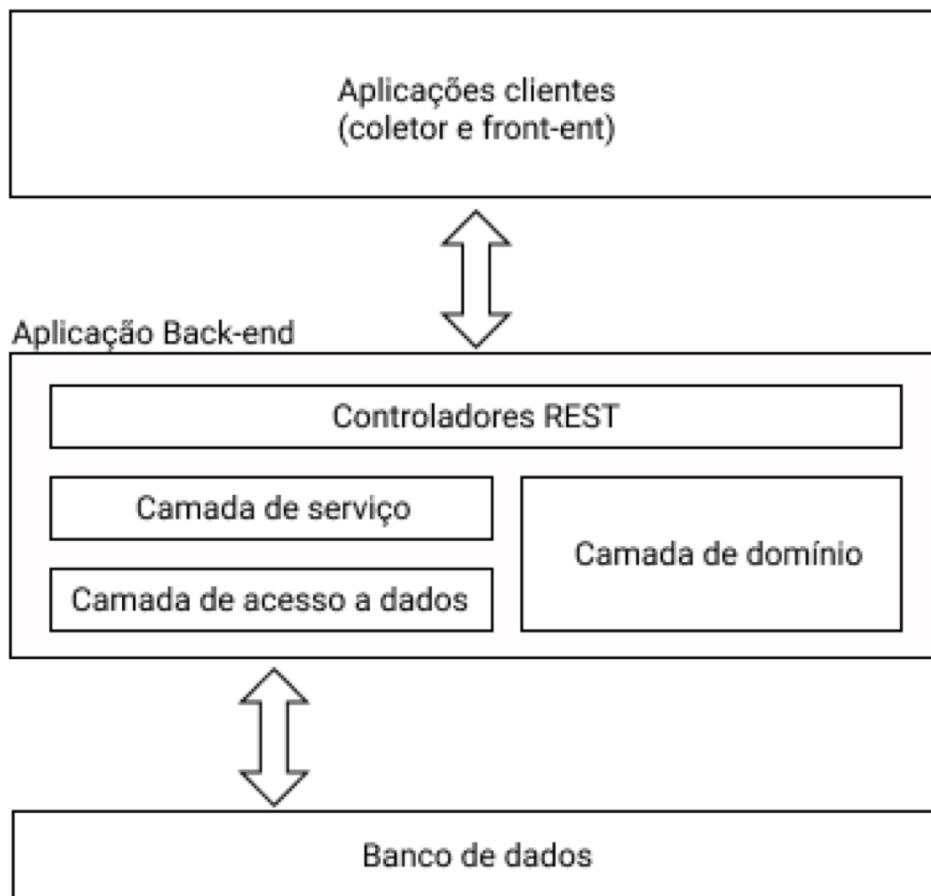
2.2.2 A aplicação *Back-end*

O backend é uma aplicação autônoma construída na linguagem de programação Java, com o framework *Spring Boot* e a ferramenta *Maven*. Esta aplicação é responsável por receber e organizar as leituras enviadas pelos dispositivos IoT, no banco de dados e atender as requisições *HTTP* do lado do cliente (*front-end*). Igualmente a aplicação realiza algumas operações básicas de análise de dados como filtragem e agrupação para gráficos tipo *box-plots*.

Quatro camadas compõem a aplicação: os controladores REST, a camada de serviço, a camada de acesso aos dados e a camada de domínio (Figura 18) (TEIXEIRA, 2018). São os controladores os que estabelecem os *endpoints*, mensagens, conteúdos e cabeçalhos de cada recurso do *back-end*. A Camada de Serviço está abaixo dos controladores REST e funciona como um mediador entre a Camada de Acesso aos Dados e a Camada de

Comunicação. É ela também quem define as regras de negócio da aplicação. A Camada de Domínio replica a modelagem do banco de dados em classes que são utilizadas pelos distintos componentes da aplicação; cada classe representa uma tabela. Por último a Camada de Acesso aos Dados tem a responsabilidade de acessar diretamente o banco de dados e disponibilizar a informação para as camadas superiores. A Figura 19 ilustra as classes utilizadas na Camada de Acesso aos Dados.

Figura 18 – Camadas da aplicação *back-end* Renovar



Fonte: (TEIXEIRA, 2018)

O processo de coleta das medições enviadas pelos dispositivos de monitoramento na aplicação *back-end* consiste, de forma simplificada, no seguinte fluxo de execução. Os controladores REST processam as requisições de tipo *POST HTTP* enviadas pelos dispositivos IoT, e transferem as informações até a camada de serviço. Ali os dados são validados e instâncias da entidade *Sample* são criadas e transferidas até a camada de acesso aos dados onde as amostras são armazenadas na tabela correspondente do banco de dados. As requisições e os *endpoints* disponibilizados na API Renovar pelos controladores REST são detalhados na documentação da API disponibilizada no Anexo A.

Figura 19 – Classes de acesso ao dados da aplicação *back-end* Renovar

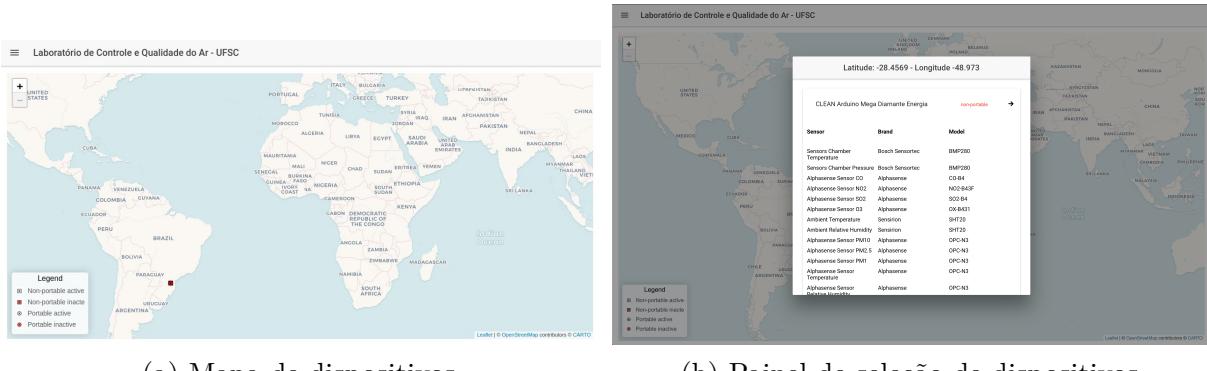
Fonte: Desenvolvido pelo autor (2023)

2.2.3 A aplicação *Front-end*

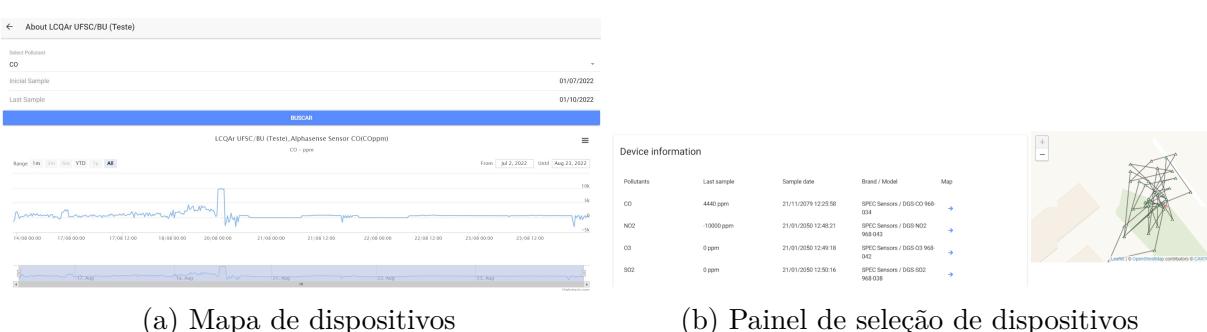
A aplicação *front-end* foi construída usando o *framework Angular*, junto com outras ferramentas como *HighCharts* e *HighStock* para os gráficos e *Leaflet* para mapas. A tela principal mostra um mapa com todos os dispositivos (Figura 20a). A cor dos dispositivos no mapa indica seu estado: verde para ativo (i.e.: o dispositivo está transmitindo novas medidas), vermelho para inativo (o dispositivo não tem enviado novas medidas em um determinado intervalo de tempo). Quando o usuário seleciona um dispositivo no mapa, um *pop-up* é aberto com informações sobre esse e todos os dispositivos que se encontram nas mesmas coordenadas geográficas, juntamente com informações sobre seus sensores (Figura 20b). A partir desse ponto, o usuário pode acessar a página do dispositivo e escolher um poluente para visualizar sua série histórica dentro de um intervalo de datas (Figura 21a). Se o aparelho for portátil, o usuário também pode ver outro mapa que mostra onde cada amostra foi coletada, conforme mostrado na Figura 21b. Uma seção de análise de dados possibilita visualizar os valores das leituras dos sensores em gráficos de *box-plots*, agrupando as medições por ano mês ou semana, conforme ilustra a Figura 22.

2.3 DISPOSITIVOS DE HARDWARE DESENVOLVIDOS

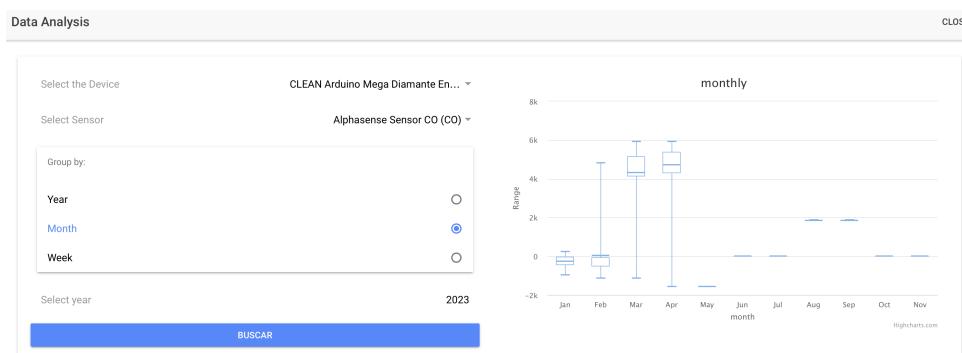
Dentro do contexto da iniciativa CLEAN foram desenvolvidos dispositivos para medição da qualidade do ar. Dois desses dispositivos foram protótipos para validação da ideia, concebidos para medições em locais fixos e medições móveis. Numa segunda etapa foram desenvolvidos dispositivos mais robustos com placas de circuito impresso e quadros elétricos para instalação em campo. A continuação serão descritos os equipamentos produzidos.

Figura 20 – Aplicação *front-end* da plataforma web Renovar

Fonte: Desenvolvido pelo autor (2023)

Figura 21 – Painéis da aplicação *front-end* de Renovar

Fonte: Desenvolvido pelo autor (2023)

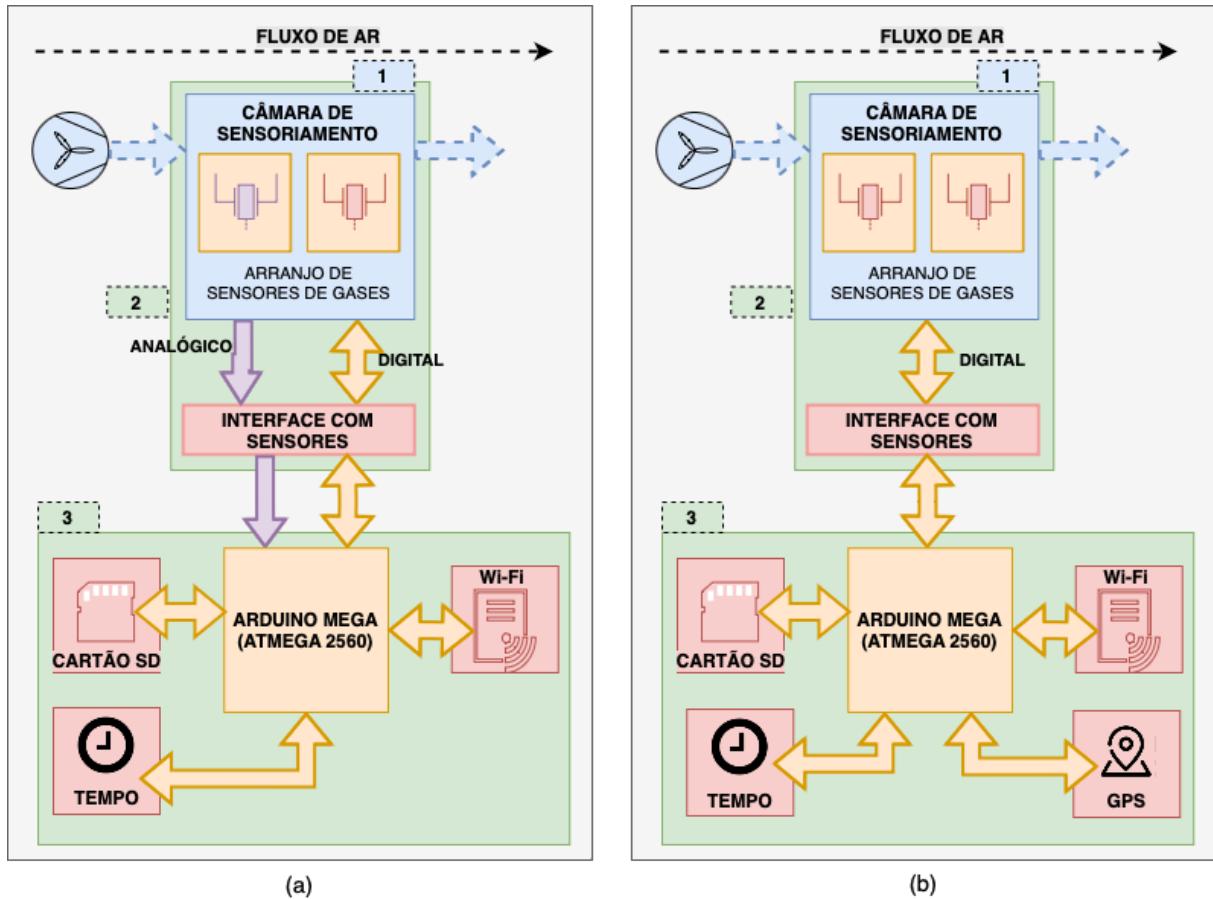
Figura 22 – Painel de análise de dados da aplicação *back-end* Renovar

Fonte: Desenvolvido pelo autor (2023)

2.3.1 Protótipos de monitores de qualidade do ar de baixo custo

Foram concebidos dois protótipos de baixo custo para medição de poluentes atmosféricos (CAMPO *et al.*, 2020), um para monitoramento fixo e outro para monitoramento móvel. O hardware de ambos os dispositivos, conforme mostrado na Figura 23, é composto por três blocos principais: 1) transporte de gás, 2) sensoriamento e 3) microcontrolador. O estágio de transporte de gás captura o ar ambiente nos sensores,

Figura 23 – Estrutura principal dos dispositivos. a) Medidor de gases fixo, e b) medidor móvel



Fonte: Desenvolvido pelo autor (2023)

que produzem um sinal analógico proporcional à concentração do gás. O microcontrolador, que é um Microchip ATmega2560 embarcado em uma plataforma Arduino Mega, captura as respostas dos sensores e as transforma em dados de concentração de gás. O hardware também obtém a hora e o local onde cada medição foi coletada. O microcontrolador armazena essas informações em um cartão micro SD e as transmite para um servidor web hospedado na Superintendência de Tecnologia da Informação e Comunicação da Universidade, rodando o aplicativo Renovar Web. Uma conexão Wi-Fi é estabelecida por um microcontrolador ESP8266 para transmissão de dados. Um relógio em tempo real e um módulo GPS fornecem informações de data, hora e geolocalização, respectivamente.

A versão fixa dos dispositivos de monitoramento (Figura 23a) utiliza seis sensores eletroquímicos do fabricante de sensores Alphasense e quatro sensores eletroquímicos do fabricante SPEC Sensors. Para alimentação de energia do dispositivo utiliza-se uma fonte de 12VCC. Este dispositivo não incorpora módulo *GPS* para geolocalização. A versão móvel (Figura 23b), por outro lado, utiliza apenas quatro sensores eletroquímicos do fabricante SPEC Sensors. O dispositivo é alimentado por um banco de energia de 5VCC

Figura 24 – Ilustrações das versões (a) fixa e (b) móvel dos dispositivos de monitoramento



Fonte: Desenvolvido pelo autor (2023)

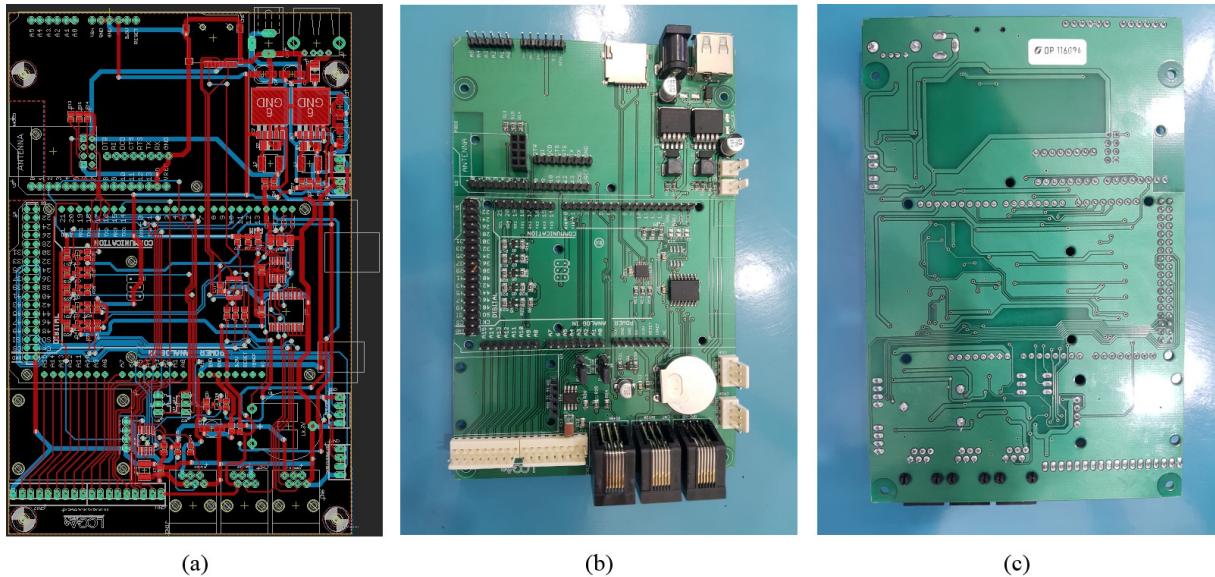
através de uma conexão USB. A Figura 3 ilustra ambos protótipos na versão fixa e móvel. Mais detalhes sobre os dispositivos podem ser encontrados no Apêndice B.

2.3.2 A placa CLEAN Arduino MEGA

Com base nos resultados obtidos pelos protótipos e nas experiências alcançadas, foi desenvolvida uma versão mais compacta e atualizada para monitoramento fixo. Esta versão foi chamada de *CLEAN Arduino Mega Board* por causa do microcontrolador Arduino Mega que ela usa como processador principal. A composição do hardware é muito semelhante à dos protótipos, mas os módulos foram montados em uma única *PCB*. A Figura 25 ilustra o projeto da *PCB* e uma das placas fabricadas. A *PCB* foi criada no *software Eagle*, e os arquivos do projeto estão disponíveis nos repositórios do LCQAr da UFSC.

A Tabela 21 do apêndice C mostra os principais componentes de hardware utilizados na placa, que requer uma tensão de alimentação de 12V, 2A através de um conector de alimentação P4. Possui entradas analógicas para 6 placas de sensores Alphasense da série ISB, barramento RS-485 para futuras expansões, três saídas digitais e conectores para alimentação de ventoinhas de 12V e 5V. A placa foi concebida para suportar conexões Wi-Fi e *GPRS* à Internet. Essas conexões não podem ser utilizadas simultaneamente, o que dependerá de cada aplicação. O usuário pode configurar a placa para usar um ou outro e terá que adaptar o firmware do microcontrolador Arduino correspondentemente.

Figura 25 – A placa CLEAN Arduino Mega: (a) projeto PCB, (b) vista superior da placa, (c) vista inferior da placa.



Fonte: Desenvolvido pelo autor (2023)

3 IDENTIFICAÇÃO DE INTERFERÊNCIAS E FALHAS NO MONITORAMENTO DE BAIXO CUSTO

Utilizando uma das placas CLEAN Arduino Mega descritas no Capítulo 2, foi desenvolvido um equipamento de baixo custo para monitoramento dos poluentes CO , NO_2 , SO_2 , O_3 e MP_{10} . Para medição desses poluentes o instrumento conta com a seguinte relação de sensores:

- Um sensor eletroquímico do fabricante Alphasense modelo CO-B4 para medição da concentração de CO
- Dois sensores eletroquímicos do fabricante Alphasense modelo SO2-B4 para medição da concentração de SO_2 ⁴
- Um sensor eletroquímico do fabricante Alphasense modelo NO2-B43F para medição da concentração de NO_2
- Dois sensores eletroquímicos do fabricante Alphasense modelo OX-B431 para medição da concentração de O_3
- Um contador ótico de partículas do fabricante Alphasense modelo OPC-N3 para medição da concentração de MP_{10}
- Um sensor de temperatura para medição da temperatura dentro da câmara de medição dos gases.

O instrumento foi instalado junto a uma estação de referência no município de Tubarão por um período de 5 meses, desde 21/11/2022 até 21/04/2023. O objetivo da instalação foi testar o dispositivo em campo, numa aplicação real e adquirir leituras de concentração dos sensores para posteriormente calibrar o instrumento com os dados da estação de referência.

3.1 CORREÇÃO DAS LEITURAS POR CO-LOCALIZAÇÃO

A co-localização consiste em instalar o equipamento de baixo custo junto a uma estação de monitoramento de referência para fins de validação e correção. Para cobrir as diversas condições atmosféricas, a sazonalidade, e a gama de valores de concentração dos gases de interesse e dos gases interferentes, a literatura recomenda uma duração de três a seis meses para a execução das rotinas de testes e validação (SPINELLE; ALEIXANDRE; GERBOLES, 2013).

⁴ As leituras desses sensores foram descartadas na primeira etapa de análise por apresentar muitos dados ruidosos e contínuas alterações no valor de linha base. Para visualizar as leituras dos sensores remeta-se ao Apêndice F.

Na aplicação apresentada neste trabalho, o instrumento de baixo custo desenvolvido foi instalado por cinco meses junto a uma estação de referência localizada no bairro Vila Moema, no município de Tubarão - SC. A estação é uma das três que fazem parte atualmente da rede de monitoramento do estado de Santa Catarina, operadas pela Diamante Geração de Energia Ltda. (IMA/SC, 2023). O município de Tubarão é vizinho de Capivari de Baixo, onde se encontra o complexo termelétrico Jorge Lacerda, operado também pela Diamante Geração de Energia Ltda. A Figura 26 ilustra a localização geográfica das estações de monitoramento na região de Capivari de Baixo, e Tubarão. Na Tabela relacionam-se os equipamentos presentes na estação Vila Moema onde foi instalado o instrumento de baixo custo desenvolvido.

Figura 26 – Mapa das estações de monitoramento em Tubarão e Capivari de Baixo



Fonte: (IMA/SC, 2023)

Tabela 4 – Relação de equipamentos presentes na estação de monitoramento de referência no município de Tubarão - SC

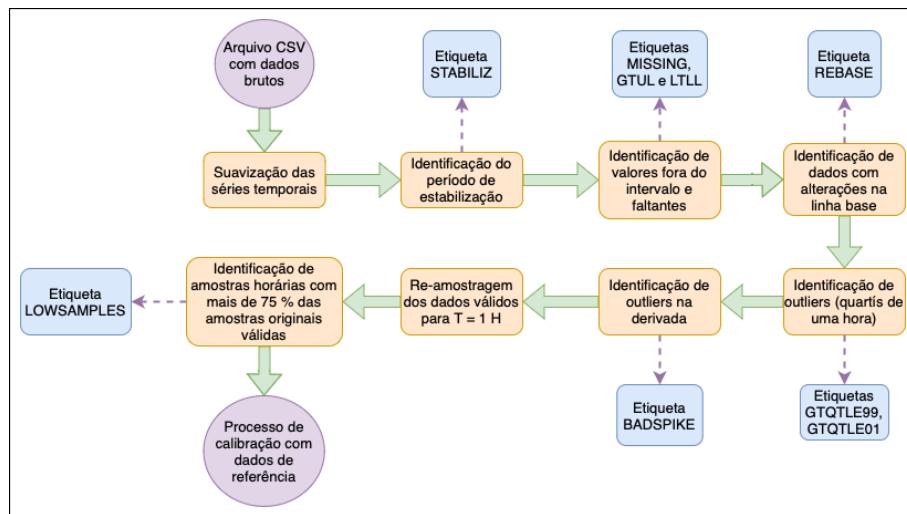
Equipamento	Poluentes	Princípio de operação	Fabricante
Monitor APNA-370	NO, NO ₂ e NO _x	Quimioluminescência	Horiba
Monitor APOA-370	O ₃	Adsorção ultravioleta	Horiba
Monitor APSA-370	SO ₂	Fluorescência ultravioleta	Horiba
Monitor APMA-370	CO	Modulação cruzada infravermelha sem dispersão	Horiba
Monitor BAM 1020	MP2.5, MP10	Atenuação de raios beta	Met One

Os dados registrados pelos equipamentos de referência são compilados em relatórios diários onde as leituras de concentração são registradas como médias horárias. Para a validação do equipamento de baixo custo, o Instituto do Meio Ambiente de Santa Catarina disponibilizou os relatórios de concentrações de poluentes no período em questão. O instrumento desenvolvido registra as leituras dos sensores com um período de amostragem de 15 minutos. Os dados adquiridos foram filtrados, pré-processados e re-amostrados em médias horárias antes de efetuar as rotinas de correção das leituras com base nas medições da estação de referência.

3.2 PRÉ-PROCESSAMENTO DOS DADOS

A primeira etapa no tratamento das leituras dos sensores de gases foi o pré-processamento dos dados para reduzir ruído, identificar falhas e valores atípicos. Para isso os dados foram etiquetados seguindo alguns pontos dos procedimentos descritos por *AQMesh*, Ottosen e Kumar (OTTOSEN; KUMAR, 2019) e o Guia para monitoramento da qualidade do ar o IMA (INSTITUTO DE ENERGIA E MEIO AMBIENTE, 2019). O pré-processamento foi realizado utilizando a linguagem de programação Python e *Jupyter Notebooks*. A continuação descrevem-se cada uma das fases de pré-processamento aplicadas aos dados dos sensores, summarizadas graficamente na Figura 27.

Figura 27 – Fluxograma das etapas de pré-processamento e as etiquetas de saída em cada uma delas



1. Suavização das curvas de dados com uma janela temporal de 1 hora: Primeiramente foi realizada a suavização das séries originais buscando reduzir o impacto de flutuações de curto prazo e realçar padrões de longo prazo, dado que os dados de referência encontram-se em períodos horários.
2. Remoção do Período de Estabilização: Os primeiros 7 dias da série temporal foram desconsiderados. Essa decisão fundamenta-se na consideração de que nesse período

inicial após a instalação o sensor encontra-se num estado de estabilização propenso a flutuações e ajustes. A remoção desses dados iniciais visa mitigar possíveis distorções na análise decorrentes desse período transitório.

3. Remoção de valores fora de intervalo de medição: Em seguida, foram removidos os valores abaixo da resolução do sensor e acima do valor máximo do sensor. Tal procedimento tem por objetivo eliminar possíveis ruídos, registros muito extremos e falhas no sistema de medição que poderiam comprometer a integridade da série temporal.
4. Remoção de valores com alterações na linha base: Para a detecção dos pontos em que a linha base das leituras sofreu alterações aplicou-se algoritmo PELT (KILLICK; FEARNHEAD; ECKLEY, 2012). Este método foi aplicado por Ottosen e Kumar (OTTOSEN; KUMAR, 2019) para detectar mudanças abruptas na média e/ou na variância das séries temporais de sensores de gases. As leituras que apresentaram alterações na linha base foram desconsiderados já que nesses intervalos a distribuição dos dados mudou, dificultando a análise.
5. Remoção de outliers por quartis: Consistiu na remoção dos quartis 1% e 99% dos dados agrupados por hora. Esse processo foi conduzido após a divisão dos dados em 24 grupos, representando cada hora do dia. Os quartis foram calculados individualmente dentro desses grupos, resultando em uma eliminação robusta de extremos estatísticos que poderiam introduzir viés na análise.
6. Identificação e remoção de picos e outliers na derivada dos dados: Este passo envolveu a identificação e eliminação de picos na primeira derivada dos dados que excediam o valor máximo encontrado na derivada da série de dados de referência. Essa estratégia visa mitigar efeitos de variações abruptas, frequentemente associadas a falhas no sensor ou interferências externas.
7. Re amostragem para período de 1 hora: Este passo consistiu na re amostragem dos dados para um período de 1 hora. Essa prática foi adotada para comparar e calibrar com os dados de referência que estão em períodos de uma hora.
8. Remoção de médias com amostras insuficientes: Por fim, foram excluídas as médias de cada hora que não atingiram mais de 75% de amostras válidas, equivalente a pelo menos 3 amostras por hora. Essa medida visa garantir a robustez estatística das médias horárias, excluindo intervalos com dados insuficientes para uma análise significativa.

Em cada estágio do pré-processamento, os dados eram etiquetados para diferenciar as leituras válidas das que apresentaram algum tipo de anomalia que inviabilizasse seu uso. A seguir são listadas as etiquetas utilizadas para identificar cada amostra; entre parênteses

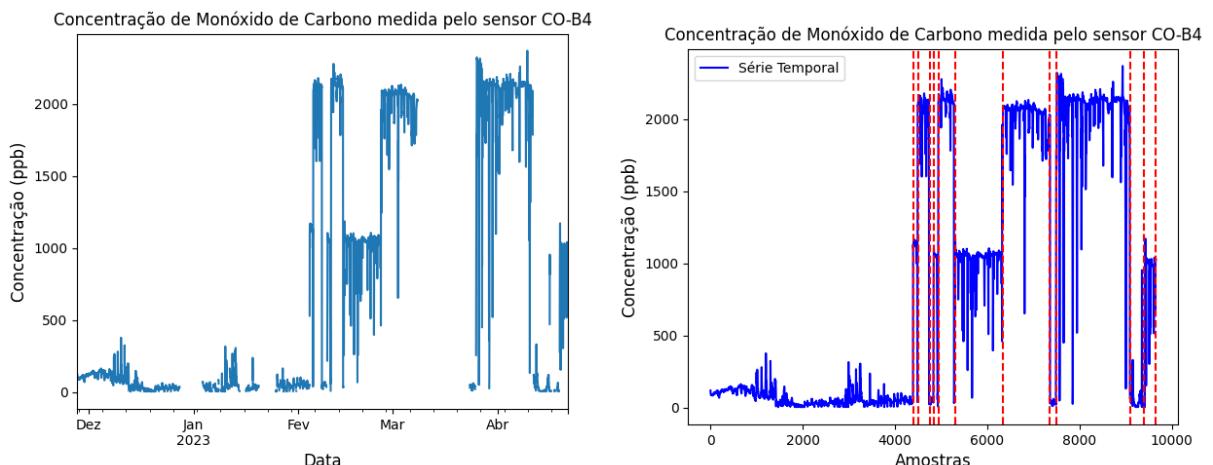
colocasse o nome da etiqueta utilizada dentro do código. O Anexo C contém o código desenvolvido em *Jupyter Notebooks* para executar esta etapa da análise dos dados.

- Dados faltantes (*MISSING*): Esta etiqueta representa amostras faltantes na amostragem dos dados.
- Estabilização (*STABILIZ*): A estabilização é um período em que um sensor fornece dados não confiáveis por não estar em estado de equilíbrio. Uma vez que o sensor se estabilizar em seu ambiente após ter sido movido recentemente ou após ter sido instalado pela primeira vez, ele fornecerá dados utilizáveis. Este processo leva 2 dias para ser concluído para sensores eletroquímicos, não sendo necessário para o sensor de material particulado modelo OPC de Alphasense (Manual de procedimento de operação padrão AQMesh). Neste trabalho foi utilizado um período de 7 dias para garantir uma maior estabilidade do sensor.
- Dados acima do Limite Superior (*GTUL*): Essa etiqueta está relacionada a valores que ultrapassam o limite superior do sensor. As especificações do fabricante ou valores conhecidos do poluente podem ser utilizados para definir esse limite. Por exemplo, o sensor Alphasense CO-B4 tem um limite de 1.000 ppm, equivalente a 1.000.000 ppb. Os valores que ultrapassaram esses limites para cada sensor foram removidos por serem indícios de alguma falha ou mal funcionamento do sistema de medição.
- Dados abaixo do Limite Inferior (*LTLL*): Esta etiqueta sugere valores que estão abaixo da resolução do sensor e pode estar relacionada a possíveis valores negativos. A definição desse limite depende das especificações do sensor utilizado. O sistema de medição desenvolvido identifica valores faltantes ou NaN com o valor -9999.99. Por esse motivo, valores abaixo de -1000 ppb nos dados representam estas amostras depois de passar pelo processo de suavização.
- Dados com alteração do valor de linha base (*REBASE*): Nas séries temporais dos sensores foram identificados alterações no valor de linha base das leituras. Observou-se também que nesses intervalos de variação da linha base a distribuição dos dados também estava alterada. Por esse motivo, os dados nesses intervalos foram marcados e removidos.
- Dados do quartil 99 % (*GTQTL99*): Essa etiqueta está relacionada aos valores que se encontram no quartil 99 % no histograma dos dados. Valores dentro desse intervalo foram etiquetados para remoção.
- Dados do quartil 1 % (*LTQTL01*): Essa etiqueta está relacionada aos valores que se encontram no quartil 1 % no histograma dos dados. Valores dentro desse intervalo foram etiquetados para remoção.

- Baixo número de amostras (*LOWSAMPLES*): De acordo com o Guia de Monitoramento da Qualidade do Ar (IMA), pelo menos 3/4 das medições de uma hora devem ser válidas para o cálculo da média horária. No sistema desenvolvido, como o período de amostragem foi de 15 minutos, para uma média horária ser considerada válida, ela deve ser calculada por 3 ou 4 pontos.
- Transições abruptas nos dados (*BADSPIKE*): Transições muito abruptas na série de dados também foram identificados. Para isso foram calculadas as derivadas das séries temporais e os valores comparados com o valor máximo das derivadas das séries temporais de referência. Valores de derivadas na série de dados acima desse valor máximo foram identificados como transições muito abruptas e etiquetados para remoção.
- Dados inválidos de temperatura *INVALID_ENV*: Aos dados de temperatura também foram aplicadas as etiquetas descritas anteriormente, com exceção de *STABILIZ* e *BADSPIKE*. Os valores de concentração de gás adquiridos no mesmo instante de algum dado inválido de temperatura foram marcados como *INVALID_ENV* para remoção.

3.3 ANÁLISE DOS DADOS DE MONÓXIDO DE CARBONO

Figura 28 – Série temporal do sensor CO-B4

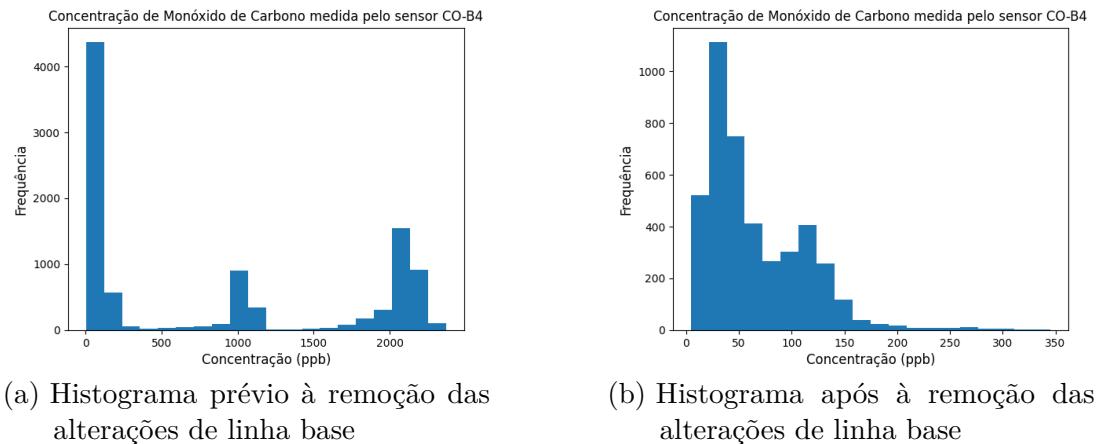


(a) Série temporal do sensor depois de remover valores fora de intervalo (b) Pontos de alteração da linha base detectados pelo algoritmo PELT

A metodologia de pre-processamento dos dados descrita acima foi aplicada as leituras obtidas pelo sensor de *CO*. A Figura 28a mostra a série temporal do sensor depois de removidos os valores fora de intervalo. Observa-se que a partir do mês de fevereiro ocorreram seguidas alterações de linha base, detectadas pelo algoritmo PELT, conforme se ilustra na Figura 28b. No histograma das leituras do sensor em questão (Figura 29a)

observa-se que as mudanças de linha base alteraram também a distribuição dos dados nesse período. Dado que as mudanças de linha base se mantiveram de forma continuada a partir desse mês, para o restante das análises apenas foram consideradas as leituras anteriores à primeira alteração da linha base do sensor. A Figura 29b mostra o histograma dos dados das leituras prévias às alterações de linha base.

Figura 29 – Histograma das leituras do sensor CO-B4



Depois de pré-processadas as leituras do sensor obtiveram-se os resultados ilustrados na Figura 30. A Figura mostra a série de dados pré-processados do sensor CO-B4, juntamente com um gráfico de caixas que representa o comportamento diário das leituras agrupadas por hora do dia. Neste último percebe-se um comportamento periódico nos dados, observando-se maiores valores de amplitude e dispersão durante o período diurno. Ao comparar esse comportamento com os valores de referência percebe-se que a componente diária é mais evidente no sensor CO-B4, indicando uma possível relação com a sazonalidade diária da temperatura.

Figura 30 – Série temporal do sensor pré-processada ($T = 15$ mins) e seu comportamento diário

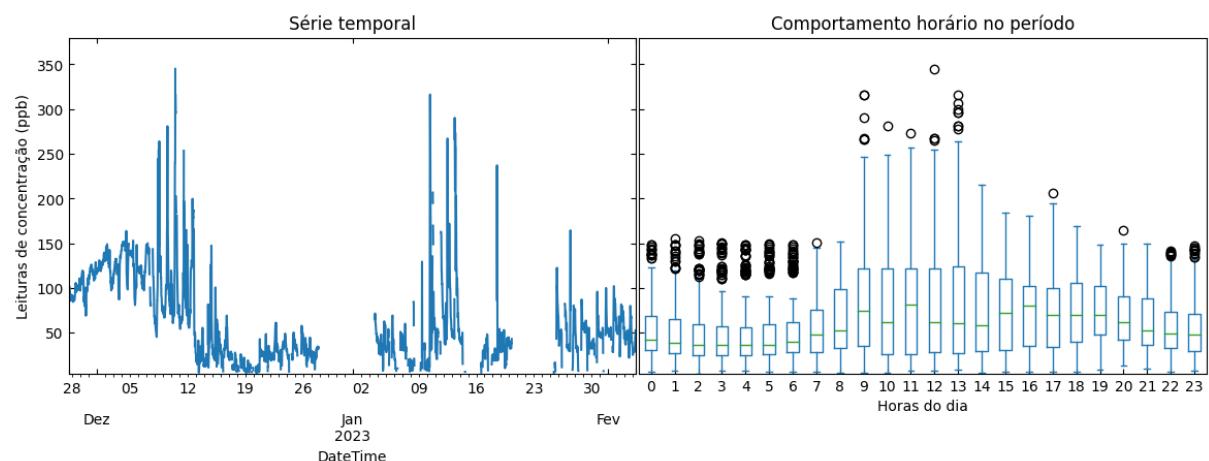
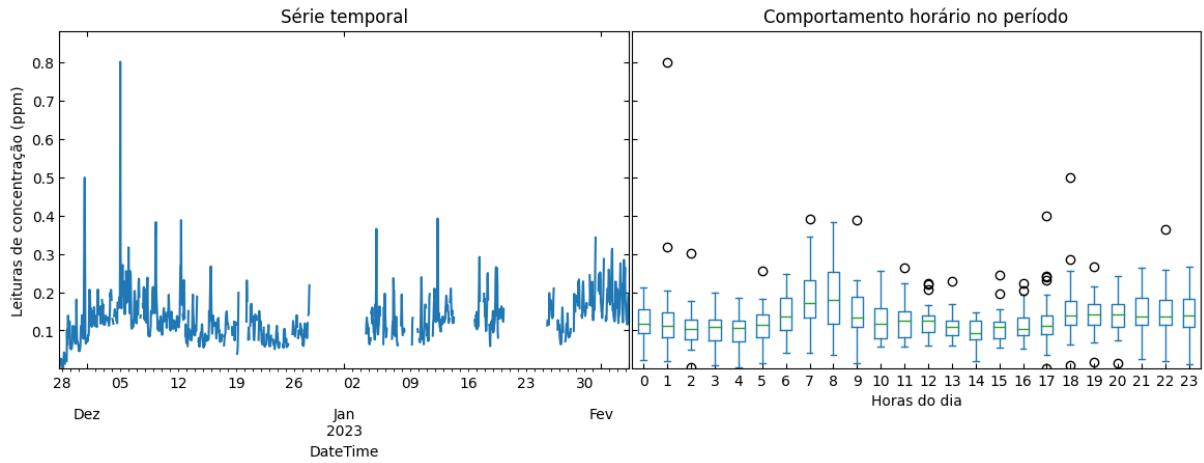
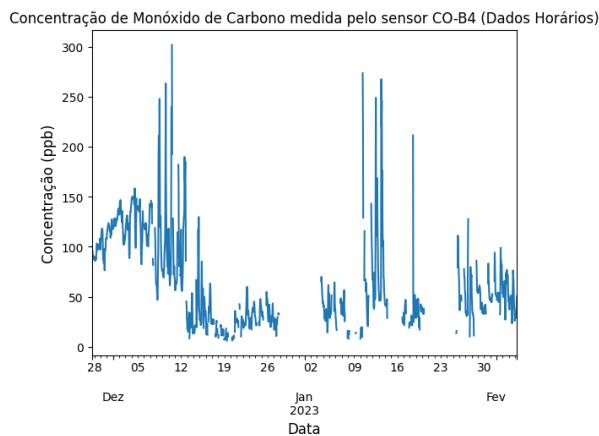


Figura 31 – Série temporal das leituras de concentração de referência ($T = 1 \text{ H}$) e seu comportamento diário



A Figura 41b mostra a série temporal das leituras do sensor consideradas como válidas, com período de amostragem horário.

Figura 32 – Série temporal com $T = 1 \text{ hr}$



A Tabela 5 mostra a contagem dos dados etiquetados para períodos de 15 minutos e de 1 hora. Observa-se que dos 17647 pontos de dados, que representavam as amostras adquiridas com um período de 15 minutos no intervalo de 20/11/2022 até 23/05/2023, 4270 foram aproveitados como dados válidos, o que representa um 24 % aproximadamente dos dados originais. Ao re-amostrar esses 4270 pontos em dados horários obtiveram-se 1048 amostras horárias de concentração válidas (aproximadamente 64 % dos dados) para realizar a calibração.

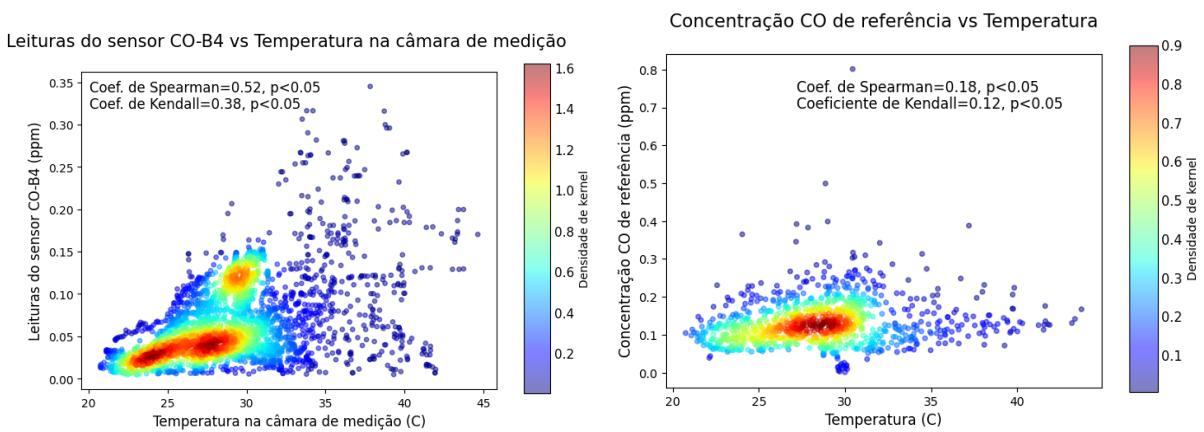
3.3.1 Dependência com a temperatura

Investigou-se a existência de correlação entre as leituras do sensor de CO e as variações de temperatura medida no interior da câmara de medição. Para tal, foram

Tabela 5 – Contabilização das leituras do sensor CO-B4 por etiquetas

Série temporal T = 15 mins			Série temporal T = 1 hr		
Etiquetas	No.	% amostras amostras	Etiquetas	No.	% amostras amostras
MISSING	5756	32.62 %	LOWSAMPLES	603	36.52 %
LTLL	1560	8.84 %	VALID	1048	63.48 %
GTUL	0.0	0.0 %			
STABILIZING	673	3.81 %			
BADSPIKE	3	0.02 %			
LTQTLE01	63	0.36 %			
GTQTLE99	63	0.36 %			
REBASE	5259	29.80 %			
VALID	4270	24.20 %			
TOTAL	17647		TOTAL	1651	

empregados os testes estatísticos de correlação de Spearman e Kendall, por serem métodos não paramétricos que exploram a relação monotônica entre variáveis. Os resultados desses testes revelaram coeficientes de correlação significativos. O coeficiente de Spearman calculado foi de 0.52, com um valor de p inferior a 0.05, indicando uma correlação estatisticamente significativa entre as leituras do sensor e a temperatura. De maneira semelhante, o coeficiente de Kendall foi de 0.38, também com $p < 0.05$, reforçando a presença de uma associação significativa. Ao avaliar a hipótese nula de ausência de correlação, os resultados forneceram evidências robustas para sua rejeição, sugerindo que há uma correlação entre as leituras do sensor de *CO* e as variações de temperatura. A Figura 33a mostra um gráfico de dispersão entre os dados do sensor e a temperatura, ilustrando os resultados de correlação obtidos.

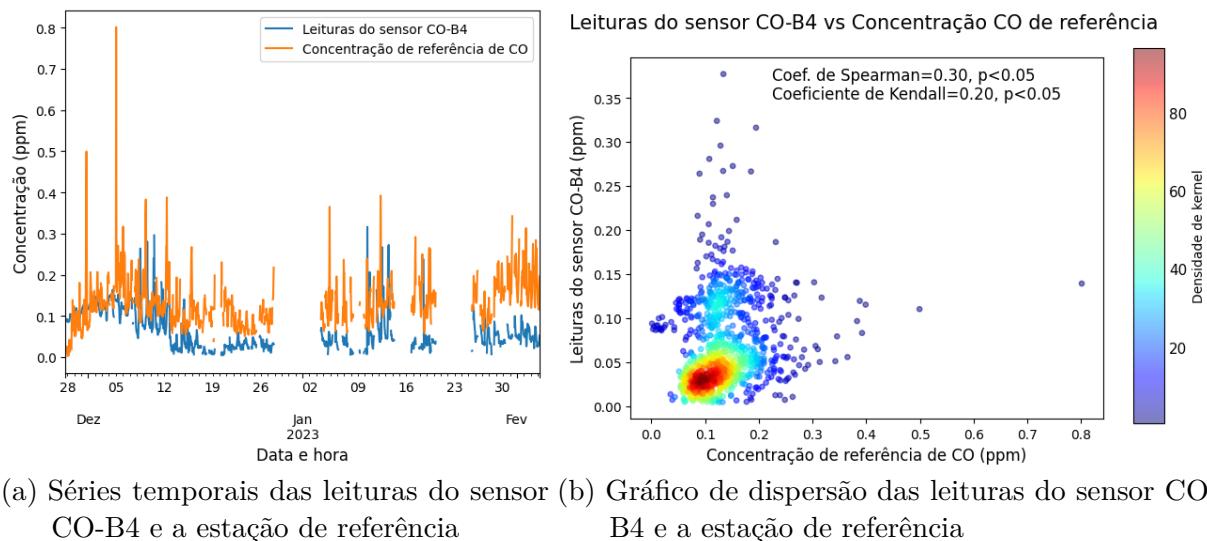
Figura 33 – Relação dos dados de concentração de *CO* com a temperatura

- (a) Relação entre as leituras do sensor CO-B4 (ppm) e a temperatura (°C)
- (b) Relação entre os valores de concentração de referência (ppm) e a temperatura (°C)

Os resultados obtidos nos testes estatísticos podem ser corroborados no gráfico

de dispersão entre as variáveis. Na Figura 33a observam-se dois núcleos principais de dados. O menor deles comporta os valores de concentração entre 0.10 e 0.15 ppm dentro de um intervalo de temperatura de 28 a 30°C. No segundo núcleo encontram-se valores de concentração entre 0 e 0.06 ppm aproximadamente e de 22 até 30°C. Neste último grupo de dados é possível apreciar uma clara relação linear entre as leituras do sensor e a temperatura. Ao analisar a relação entre as medições de concentração de referência e a temperatura, também se observa alguma correlação, embora em menor medida, com coeficientes de Spearman e Kendall de 0.18 e 0.12 respectivamente (Figura 33b).

Figura 34 – Séries temporais e gráficos de dispersão das medições de *CO*



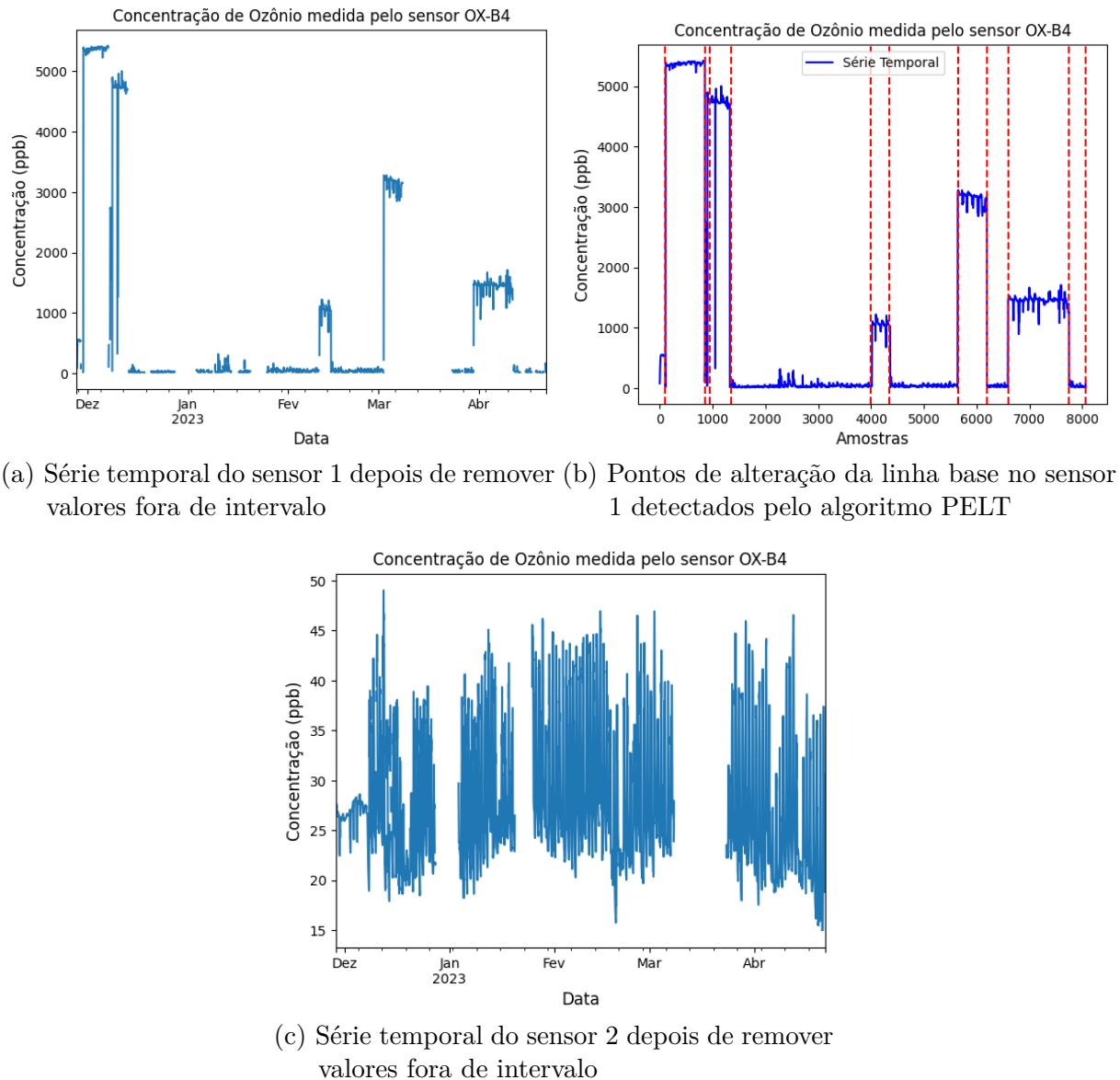
(a) Séries temporais das leituras do sensor CO-B4 e a estação de referência (b) Gráfico de dispersão das leituras do sensor CO-B4 e a estação de referência

3.3.2 Comparação das leituras do sensor CO-B4 com as medições de referência

Nas Figuras 34a e 34b apresentam-se as leituras de *CO* obtidas pelo sensor CO-B4 de Alphasense e a estação de referência. Observa-se que as leituras do sensor CO-B4 em geral subestimaram os valores de concentração de referência. Os testes de Spearman e Kendall revelaram a existência de correlação entre as medições com o sensor de baixo custo e a referência com coeficientes de 0.3 e 0.2 respectivamente.

3.4 ANÁLISE DOS DADOS DE OZÔNIO

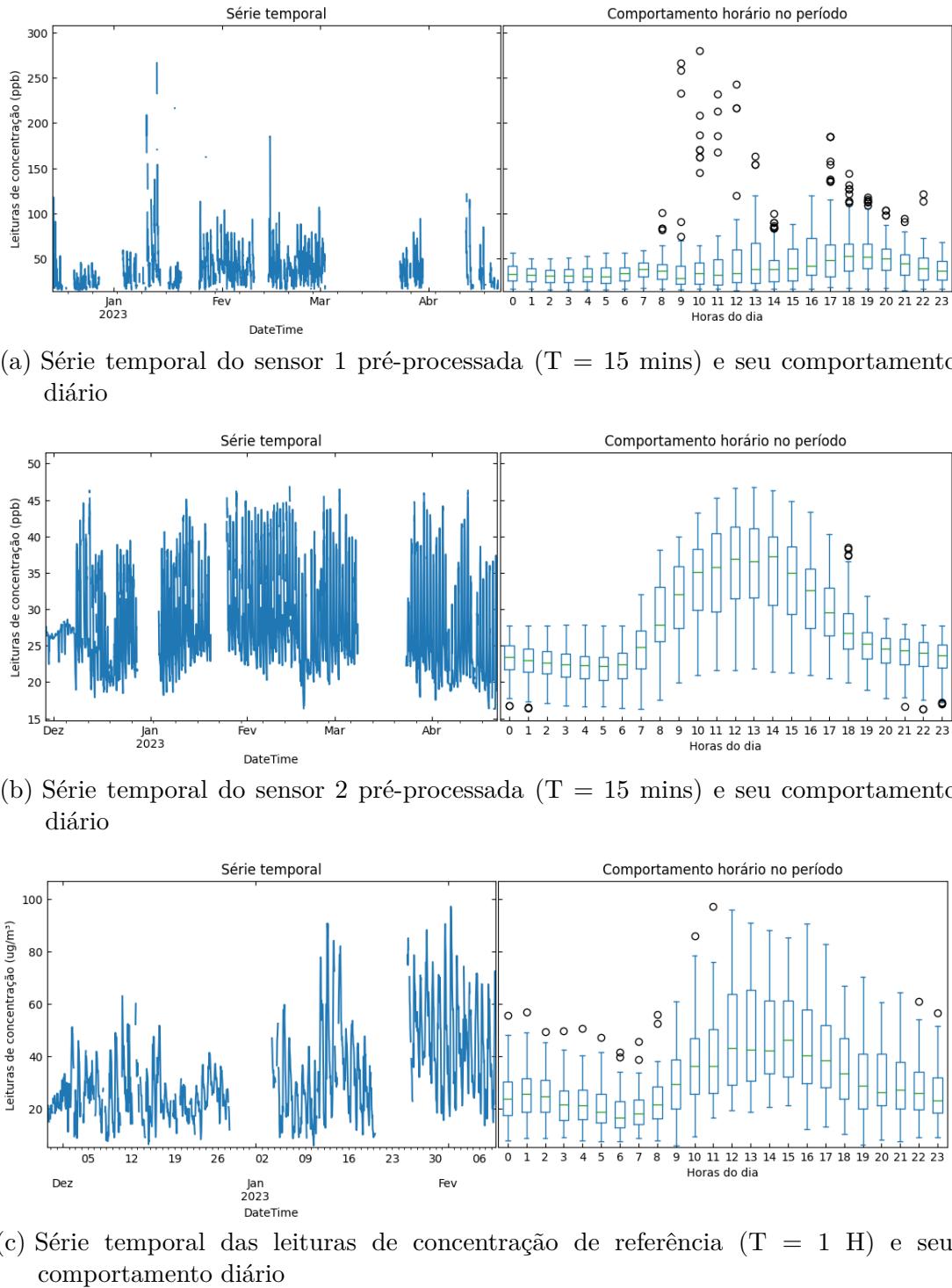
Para a medição de O_3 foram utilizados dois sensores do modelo OX-B431. As Figuras 35a e 35c mostram as séries temporais dos sensores depois de removidos os valores fora de intervalo. Observa-se que o sensor 1 sofreu alterações no valor de linha base nos meses de dezembro, fevereiro, março e abril, que foram detectadas pelo algoritmo PELT, conforme ilustrado na Figura 35b. As amostras dentro dos intervalos que apresentaram variações na linha base no sensor 1 foram etiquetados correspondentemente para sua remoção. As leituras do sensor 2 não apresentaram alterações na linha base.

Figura 35 – Série temporal dos sensores de O_3 modelo OX-B431

Depois de pré-processadas as leituras dos sensores de ozônio obtiveram-se os resultados ilustrados nas Figuras 36a e 36b. Os gráficos mostram as séries dos dados pré-processados dos dois sensores OX-B431 juntamente com o comportamento diário das medições ao longo do período, agrupados por hora do dia. Na Figura 36c observa-se uma clara componente diária nas leituras de concentração, que coincide com o comportamento apresentado pelas medições de referência (Figura 36c). Esse comportamento é esperado nas medições de O_3 já que a variável é influenciada pela luz solar.

Histogramas das leituras dos sensores OX-B431 são mostrados nas Figuras 37a e 37b. Os dados adquiridos pelo sensor 1 apresentaram um comportamento log-normal. Já o sensor 2 produziu leituras com uma componente log-normal e uma outra componente de menor frequência nos valores de concentração acima 30 ppb. Esta última representa a componente com sazonalidade diária mencionada acima. As séries re amostradas em

Figura 36 – Séries temporais dos sensores OX-B431 pré-processadas



períodos de 1 hora são mostradas nas Figuras 37c e 37d.

Nas Tabelas 6 e 7 contabilizam-se os dados dos sensores de O_3 para períodos de 15 minutos e de 1 hora. Observa-se que no sensor 1, dos 14625 pontos de dados, que representavam as amostras adquiridas com um período de 15 minutos no intervalo de 20/11/2022 até 21/04/2023, 4413 foram aproveitados como dados válidos, o que representa um 30 % aproximadamente dos dados originais. Ao re-amostrar esses 4413 pontos em dados

Tabela 6 – Contabilização dos dados por etiquetas das leituras do sensor 1 OX-B431

Série temporal T = 15 mins			Série temporal T = 1 hr		
Etiquetas	No.	% amostras amostras	Etiquetas	No.	% amostras amostras
<i>MISSING</i>	2750	18.80 %	<i>LOWSAMPLES</i>	2020	65.67 %
<i>LTL</i>	3134	21.43 %	<i>VALID</i>	1056	34.33 %
<i>GTUL</i>	0	0.0 %			
<i>STABILIZING</i>	514	3.51 %			
<i>BADSPIKE</i>	56	0.38 %			
<i>LTQTL01</i>	102	0.70 %			
<i>GTQTL99</i>	64	0.44 %			
<i>REBASE</i>	3592	24.56 %			
<i>VALID</i>	4413	30.17 %			
TOTAL	14625		TOTAL	3076	

horários obtiveram-se 1056 amostras horárias de concentração válidas (aproximadamente 34 % dos dados) para realizar a calibração. Já no sensor 2, dos 14542 pontos de dados, que representavam as amostras adquiridas no intervalo de 21/11/2022 até 21/04/2023, 10814 foram aproveitados como dados válidos, o que representa um 74 % aproximadamente dos dados originais. Ao re-amostrar esses 10814 pontos válidos em dados horários obtiveram-se 2685 amostras horárias de concentração válidas (aproximadamente 77 % dos dados) para realizar a calibração.

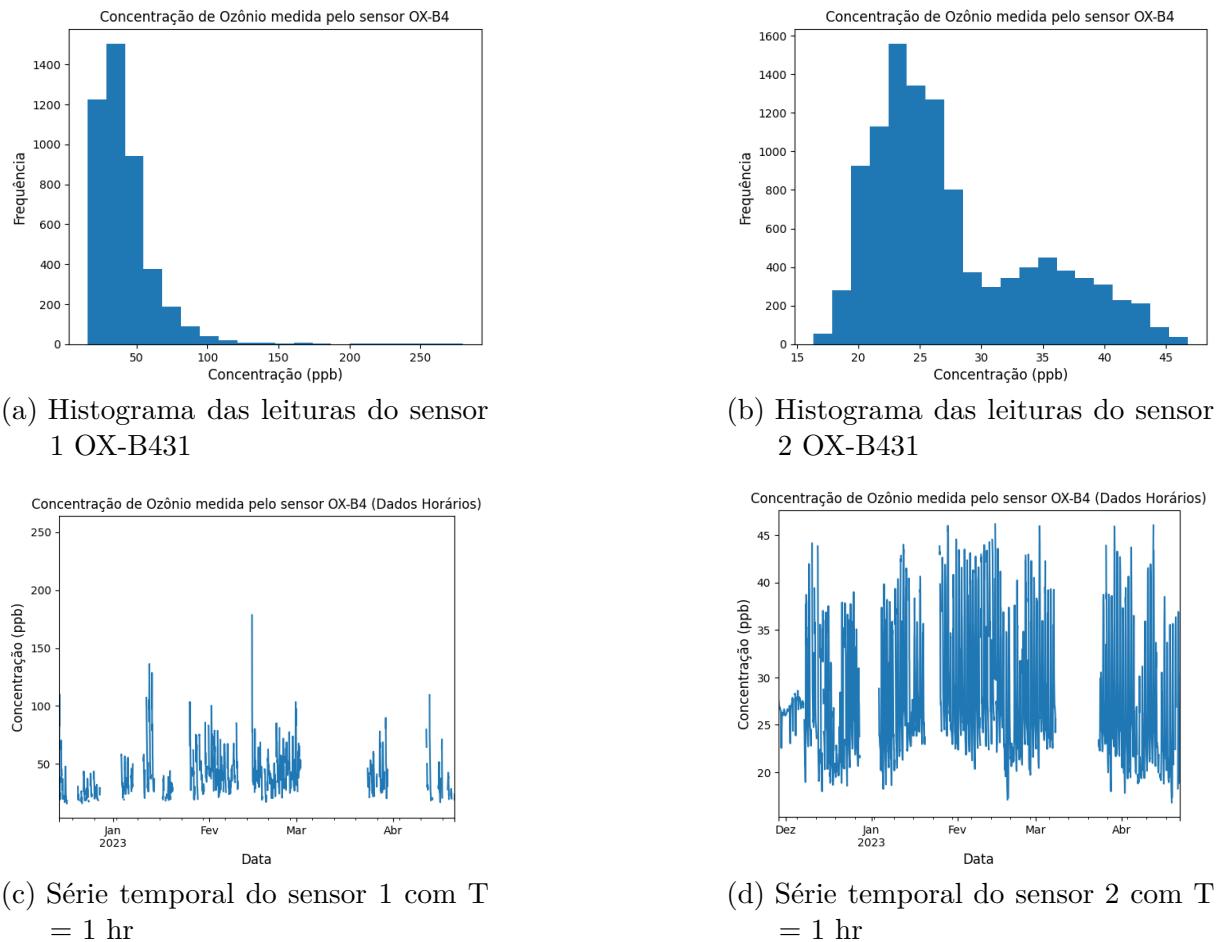
Tabela 7 – Contabilização dos dados por etiquetas das leituras do sensor 2 OX-B431

Série temporal T = 15 mins			Série temporal T = 1 hr		
Etiquetas	No.	% amostras amostras	Etiquetas	No.	% amostras amostras
<i>MISSING</i>	2734	18.80 %	<i>LOWSAMPLES</i>	783	22.58 %
<i>LTL</i>	49	0.34 %	<i>VALID</i>	2685	77.42 %
<i>GTUL</i>	0	0.0 %			
<i>STABILIZING</i>	673	4.63 %			
<i>BADSPIKE</i>	0	0.0 %			
<i>LTQTL01</i>	125	0.86 %			
<i>GTQTL99</i>	147	1.01 %			
<i>REBASE</i>	0	0.0 %			
<i>VALID</i>	10814	74.36 %			
TOTAL	14542		TOTAL	3468	

3.4.1 Dependência com a temperatura

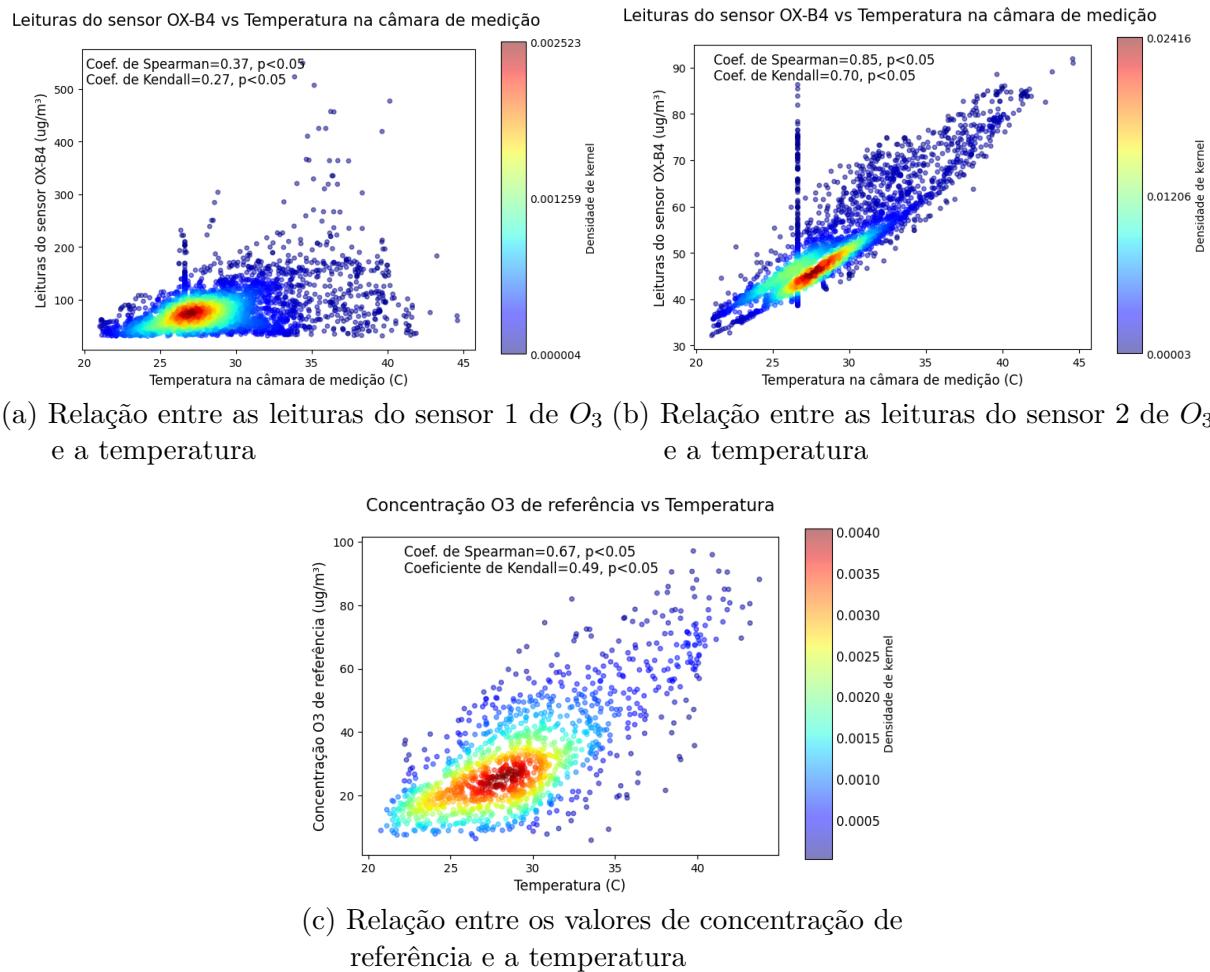
Investigou-se a existência de correlação entre as leituras dos sensores de O_3 e as variações de temperatura medida no interior da câmara de medição. Os resultados dos

Figura 37 – Histogramas e séries temporais horárias das leituras dos sensores OX-B431



testes estatísticos de Spearman e Kendall revelaram coeficientes de correlação significativos, conforme se ilustra nas Figuras 38a e 38b. Os coeficientes de Spearman calculados foram de 0.37 e 0.85 para os sensores 1 e 2 respectivamente, com valores de p inferiores a 0.05, indicando uma correlação estatisticamente significativa entre as leituras dos sensores e a temperatura. De maneira semelhante, os coeficientes de Kendall foram de 0.27 e 0.70 respectivamente, também com $p < 0.05$, reforçando a presença de uma associação significativa. Ao avaliar a hipótese nula de ausência de correlação, os resultados forneceram evidências para sua rejeição, sugerindo a existência de uma correlação entre as leituras dos sensores de O_3 e as variações de temperatura.

Os resultados obtidos nos testes estatísticos podem ser corroborados nos gráficos de dispersão entre as variáveis nas Figuras 38a e 38b. Delas comprova-se uma maior correlação com a temperatura no sensor 2, que coincide com o comportamento sazonal diário observado anteriormente. As leituras de referência também apresentaram uma relação linear com a temperatura, com coeficientes de Spearman e Kendall de 0.69 e 0.51 respectivamente, conforme se ilustra na Figura 38c.

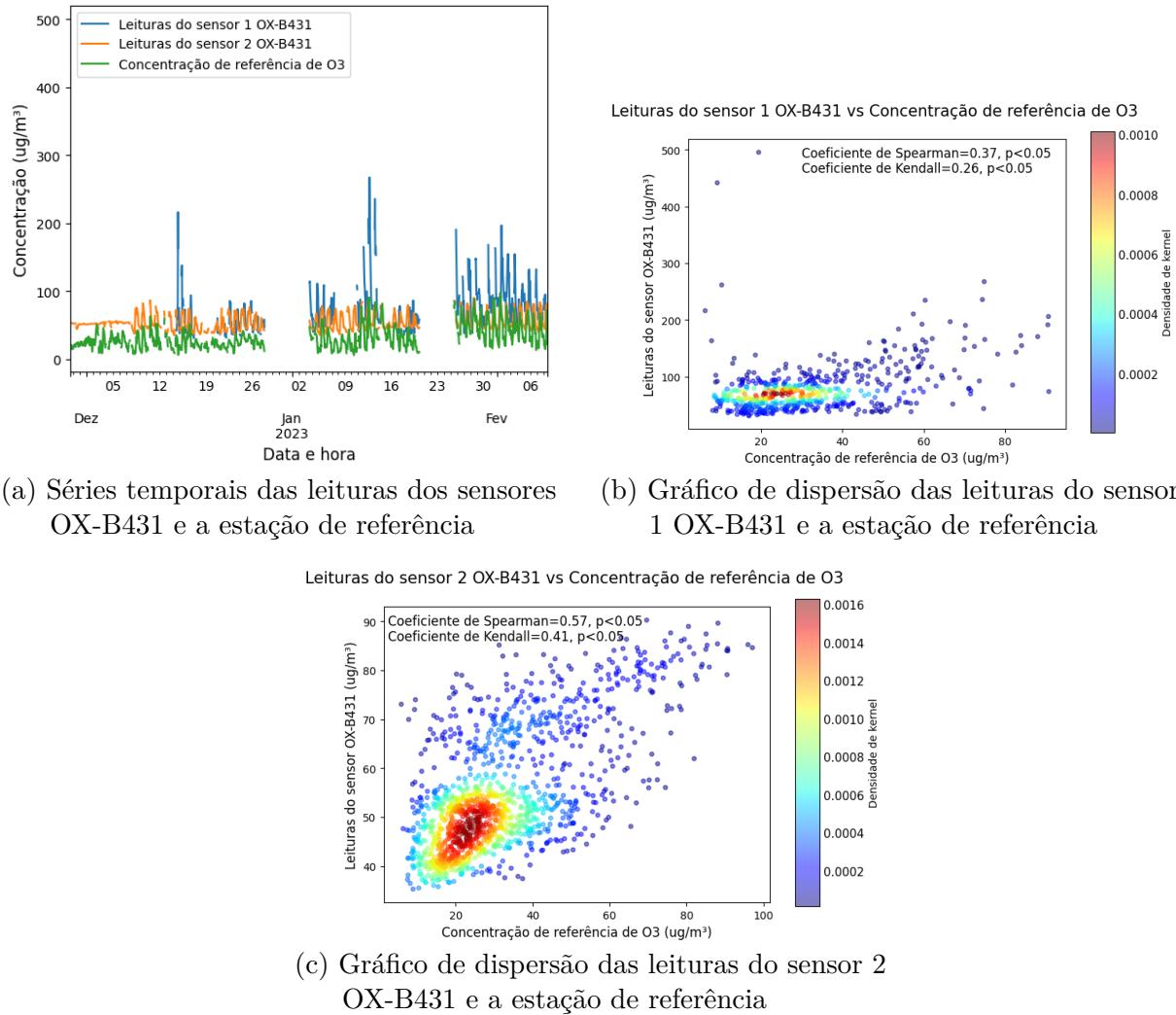
Figura 38 – Relação entre as leituras dos sensores de O_3 e a temperatura

3.4.2 Comparação das leituras dos sensores OX-B431 com as medições de referência

Nas Figuras 39a, 39b e 39c apresentam-se as leituras de O_3 obtidas pelos sensores OX-B431 de Alphasense e a estação de referência. Observa-se que as leituras do sensor 1 superestimaram os valores de concentração de referência. Os testes de Spearman e Kendall revelaram a existência de correlação entre as medições com o sensor 1 e a referência com coeficientes de 0.38 e 0.27 respectivamente, e de 0.59 e 0.42 respectivamente para o sensor 2.

3.5 ANÁLISE DOS DADOS DE DIÓXIDO DE NITROGÊNIO

A Figura 40a mostra a série temporal do sensor depois de removidos os valores fora de intervalo. Observa-se que a partir do final do mês de dezembro ocorreram seguidas alterações de linha base, detectadas pelo algoritmo PELT, conforme se ilustra na Figura 40b. Dado que as mudanças de linha base se mantiveram de forma continuada a partir desse mês, para o restante das análises apenas foram consideradas as leituras anteriores à

Figura 39 – Séries temporais e gráficos de dispersão das medições de O_3 

primeira alteração da linha base do sensor. A Figura 41a mostra o histograma dos dados das leituras prévias às alterações de linha base, os quais apresentaram uma distribuição log-normal. A Figura 41b mostra a série temporal das leituras do sensor NO2-B43F após aplicar a metodologia de pré-processamento e re-amostrar os dados para um período de 1 hora.

A Figura 42a mostra a série de dados pré-processados do sensor NO2-B43F, juntamente com um gráfico de caixas que representa o comportamento diário das leituras agrupadas por hora do dia. Neste último não é possível perceber um padrão de comportamento muito claro nos dados ao longo do dia, mas em geral observa-se que foram registradas valores de concentração mais altos entre as 14 - 19 hrs. Na Figura 42b apresenta-se a série temporal da concentração de referência e seu comportamento diário ao longo do mesmo período. Nas leituras de referência também não é possível observar um padrão diário nos dados.

A Tabela 8 mostra a contagem dos dados etiquetados para períodos de 15 minutos e de 1 hora. Observa-se que dos 17647 pontos de dados, que representavam as amostras

Figura 40 – Série temporal do sensor NO2-B43F

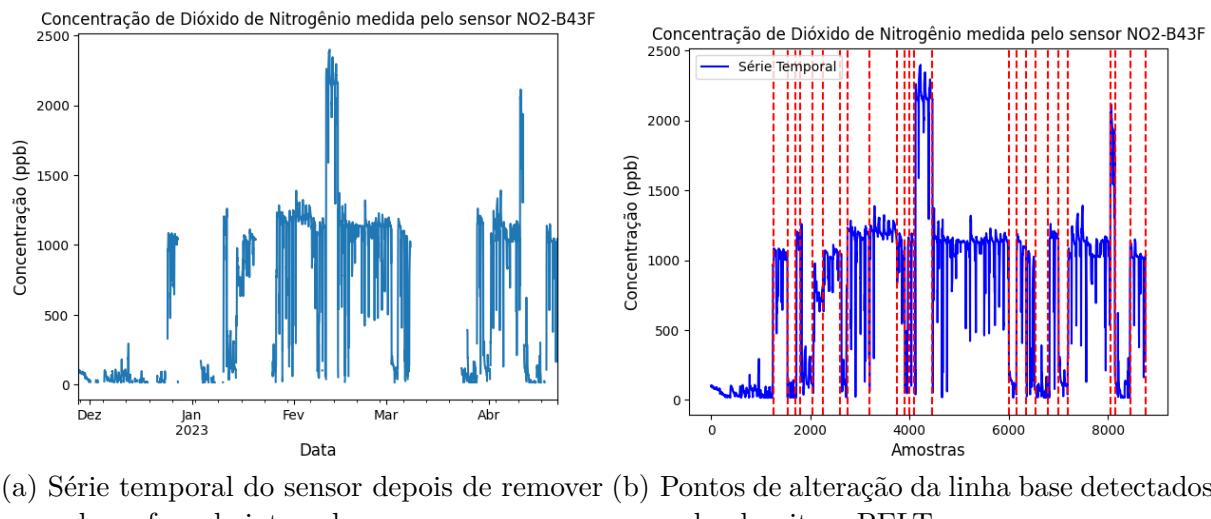
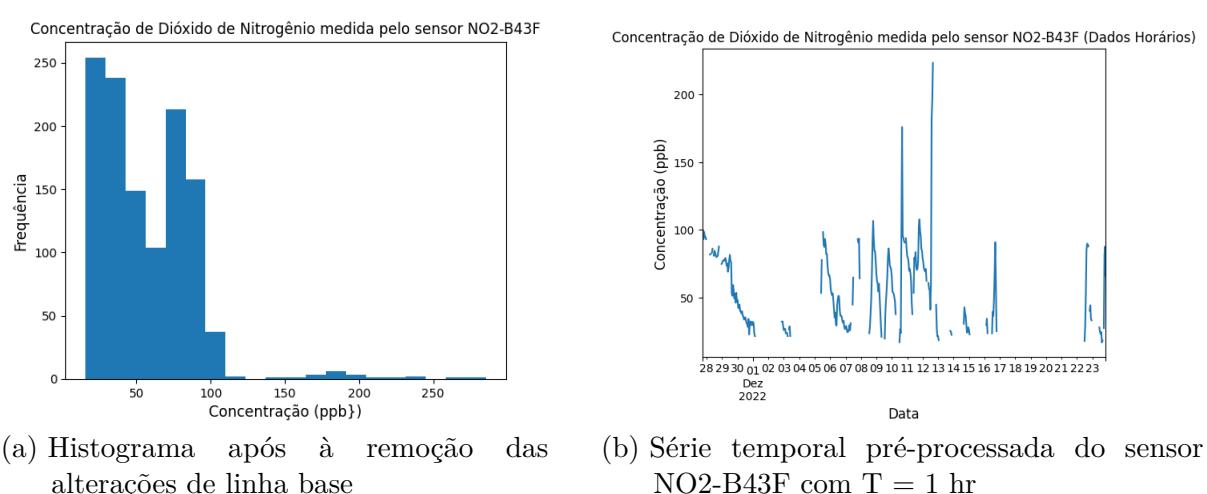


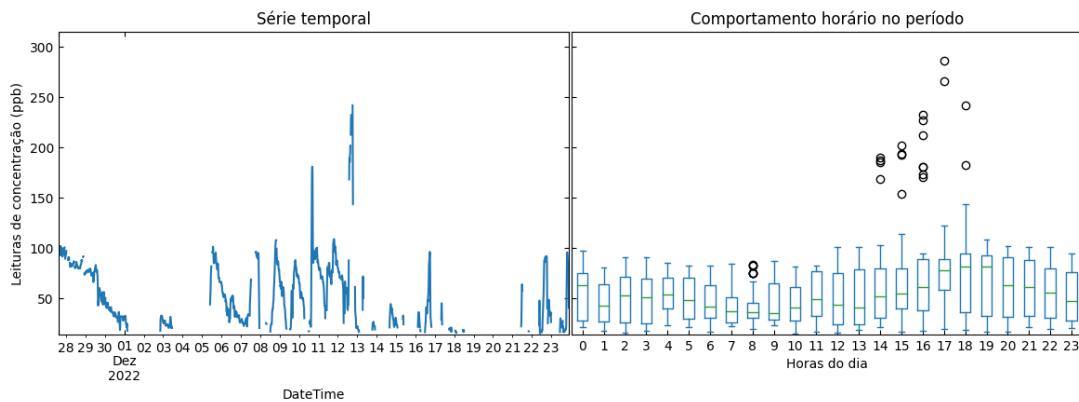
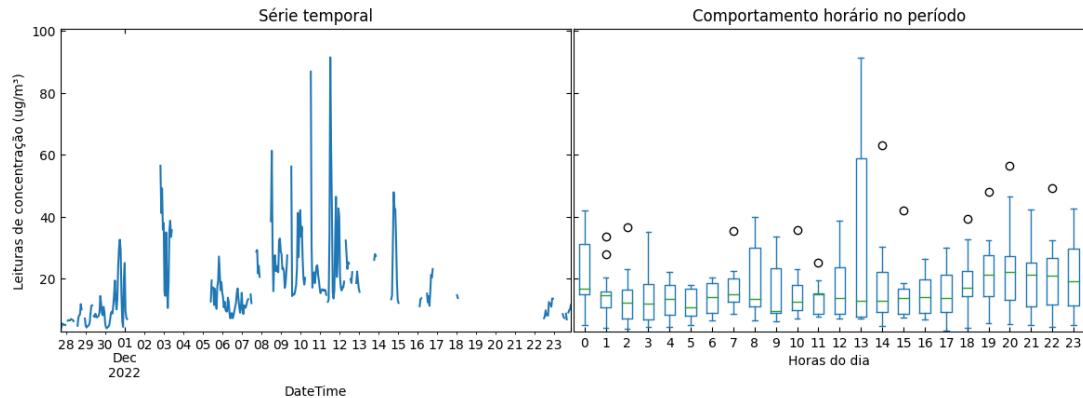
Figura 41 – Histograma das leituras do sensor NO2-B43F



adquiridas com um período de 15 minutos no intervalo de 20/11/2022 até 23/05/2023, 1175 foram aproveitados como dados válidos, o que representa um 7 % aproximadamente dos dados originais. Ao re-amostrar esses 1175 pontos em dados horários obtiveram-se 285 amostras horárias de concentração válidas (aproximadamente 45 % dos dados) para realizar a calibração.

3.5.1 Dependência com a temperatura

Investigou-se a existência de correlação entre as leituras do sensor de NO_2 e as variações de temperatura medida no interior da câmara de medição. Para tal, foram empregados os testes estatísticos de correlação de Spearman e Kendall. Os resultados desses testes revelaram coeficientes de correlação baixos. O coeficiente de Spearman calculado foi de 0.09, com um valor de p inferior a 0.5. De maneira semelhante, o coeficiente de

(a) Série temporal do sensor pré-processada ($T = 15$ mins) e seu comportamento diário(b) Série temporal das leituras de concentração de referência ($T = 1$ H) e seu comportamento diário

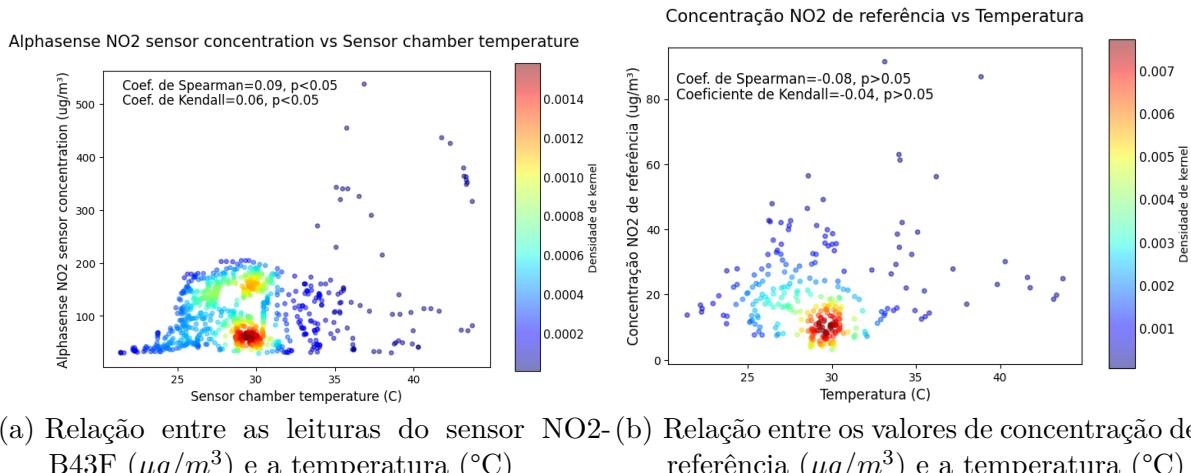
Kendall foi de 0.06, também com $p < 0.05$. Ao avaliar a hipótese nula de ausência de correlação, os resultados forneceram evidências robustas para sua rejeição, sugerindo que existe correlação, embora baixa, entre as leituras do sensor de NO_2 e as variações de temperatura. A Figura 43a mostra um gráfico de dispersão entre os dados do sensor e a temperatura, ilustrando os resultados de correlação obtidos. As leituras de referência mostraram um comportamento diferente, não apresentando correlação com a variável temperatura. Os testes estatísticos de Spearman e Kendall mostraram que não foi possível rejeitar a hipótese nula de não existência de correlação entre as variáveis (Figura 43b).

3.5.2 Comparação das leituras de NO_2 do sensor NO2-B43F com as medições de referência

Nas Figuras 44a e 44b apresentam-se as leituras de NO_2 obtidas pelo sensor NO2-B43F de Alphasense e a estação de referência. Observa-se que as leituras do sensor superestimaram os valores de concentração de referência com valores aproximadamente 5 vezes maiores. Os testes de Spearman e Kendall revelaram que não foi possível rejeitar a hipótese nula de que não existe correlação entre os dados do sensor e da estação de referência.

Tabela 8 – Contabilização das leituras do sensor NO2-B43F por etiquetas

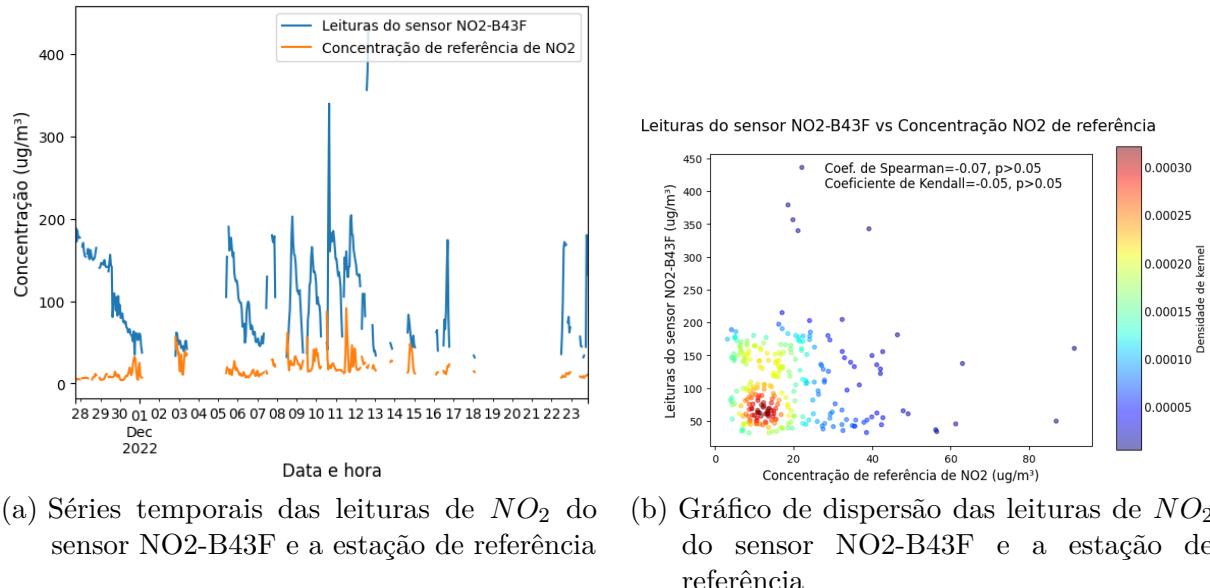
Série temporal T = 15 mins			Série temporal T = 1 hr		
Etiquetas	No.	% amostras	Etiquetas	No.	% amostras
	amostras		amostras		
MISSING	5767	32.68 %	LOWSAMPLES	347	54.91 %
LTL	2438	13.82 %	VALID	285	45.09 %
GTUL	0.0	0.0 %			
STABILIZING	673	3.81 %			
BADSPIKE	1	0.01 %			
LTQTL01	32	0.18 %			
GTQTL99	36	0.20 %			
REBASE	7525	42.64 %			
VALID	1175	6.66 %			
TOTAL	17647		TOTAL	632	

Figura 43 – Relação dos dados de concentração de NO_2 com a temperatura

3.6 ANÁLISE DOS DADOS DE MATERIAL PARTICULADO

A Figura 45a mostra a série temporal do sensor de MP_{10} depois de removidos os valores fora de intervalo. Os resultados do pré-processamento das leituras do sensor são ilustrados nas Figuras 45b e 45c que apresentam, respectivamente, o histograma dos dados e a série pré-processada do sensor OPC-N3 juntamente com o comportamento diário das medições ao longo do período agrupadas por hora do dia.

O sensor de material particulado é um sensor digital com algoritmos próprios de filtragem, correção e ajuste de linha base. Por esse motivo, as leituras do sensor não apresentaram alterações na linha base nem foram tão ruidosas como as dos sensores eletroquímicos. Da mesma forma não é possível identificar um padrão de comportamento diário tão claro quanto os outros sensores. Em comparação com o comportamento diário das leituras de referência (Figura 45d) observa-se que estas também não apresentaram

Figura 44 – Séries temporais e gráficos de dispersão das medições de NO_2 

um padrão diário muito evidente.

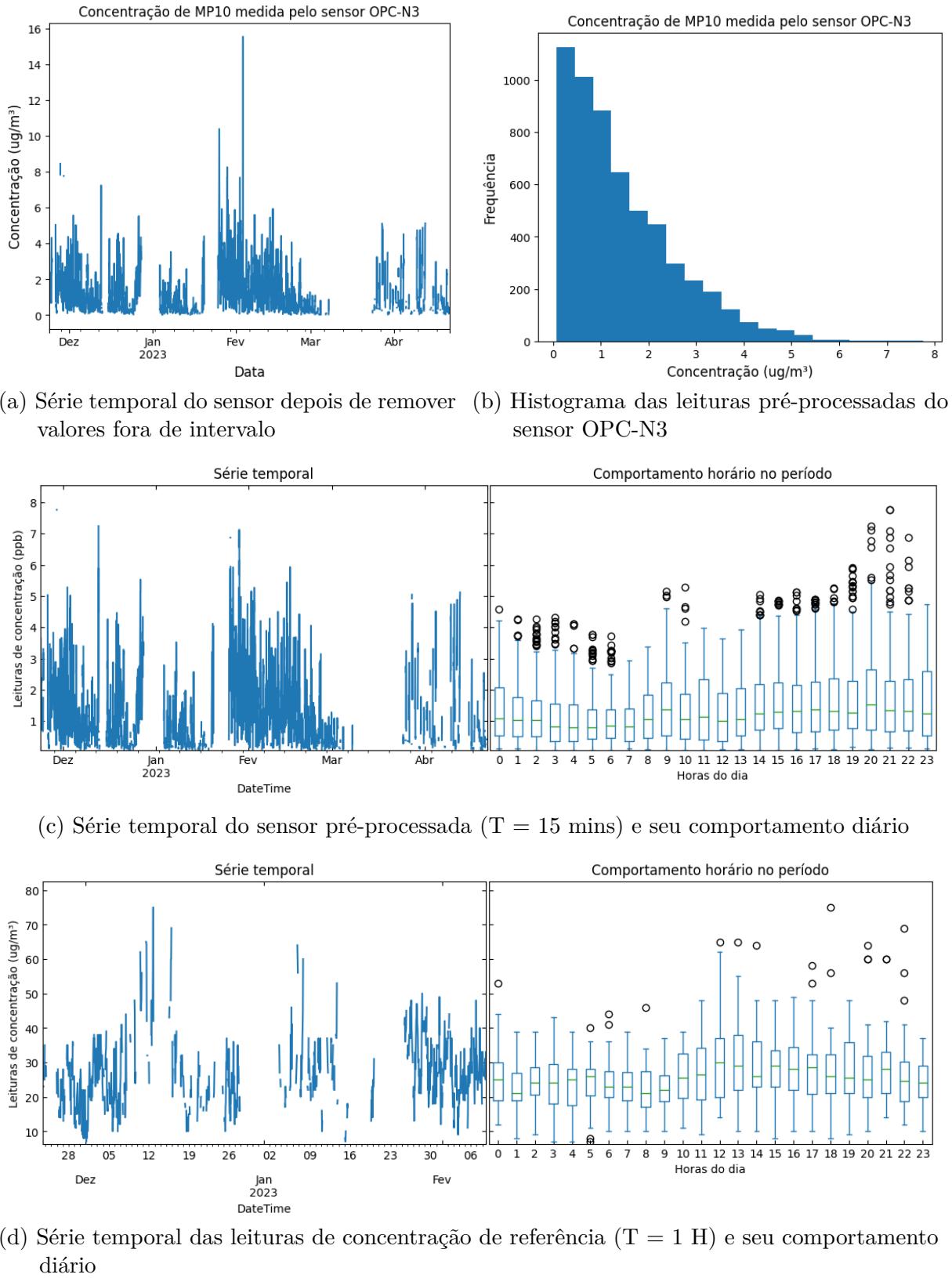
Os dados do sensor de MP_{10} do OPC-N3 foram re-amostrados para um período de 1 hora para serem calibrados com as leituras de referência; a Figura 46 mostra a série temporal resultante da re-amostragem.

Na Tabela 9 contabilizam-se os dados para períodos de 15 minutos e de 1 hora. Observa-se que dos 14541 pontos de dados, que representavam as amostras adquiridas com um período de 15 minutos no intervalo de 21/11/2022 até 21/04/2023, 5668 foram aproveitados como dados válidos, o que representa um 39 % aproximadamente dos dados originais. Ao re-amostrar esses 5668 pontos em dados horários obtiveram-se 1291 amostras horárias de concentração válidas (aproximadamente 36 % dos dados) para realizar a calibração. Vale salientar que nos dados de MP_{10} não foram encontradas alterações na linha base nem dados de estabilização já que esse sensor não precisa desse intervalo prévio as medições.

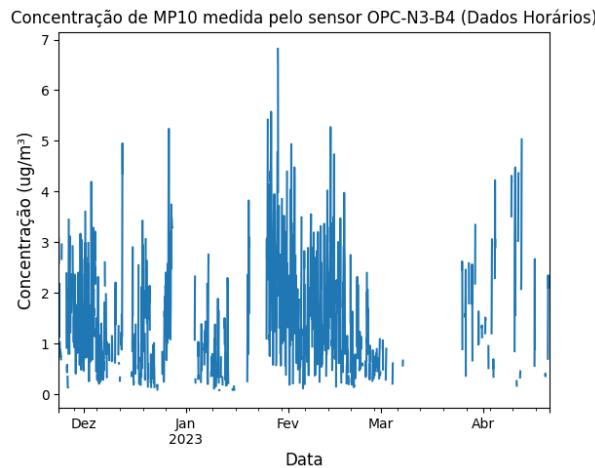
3.6.0.1 Dependência com a temperatura

Investigou-se a existência de correlação entre as leituras do sensor de $mp10$ e as variações de temperatura medida no interior da câmara de medição. Os resultados dos testes estatísticos de Spearman e Kendall revelaram coeficientes de correlação estatisticamente significativos, conforme se ilustra na Figura 47a. O coeficiente de Spearman resultou em 0.29 com um valor de p inferior a 0.05, indicando uma correlação estatisticamente significativa entre as leituras dos sensores e a temperatura. De maneira semelhante, o coeficiente de Kendall foi de 0.20, também com $p < 0.05$, reforçando a presença de uma associação significativa. Ao avaliar a hipótese nula de ausência de correlação, os resultados forneceram evidências para sua rejeição, sugerindo a existência

Figura 45 – Série temporal das leituras do sensor OPC-N3



de uma correlação entre as leituras dos sensores de MP_{10} e as variações de temperatura. Os resultados obtidos nos testes estatísticos podem ser corroborados no gráfico de

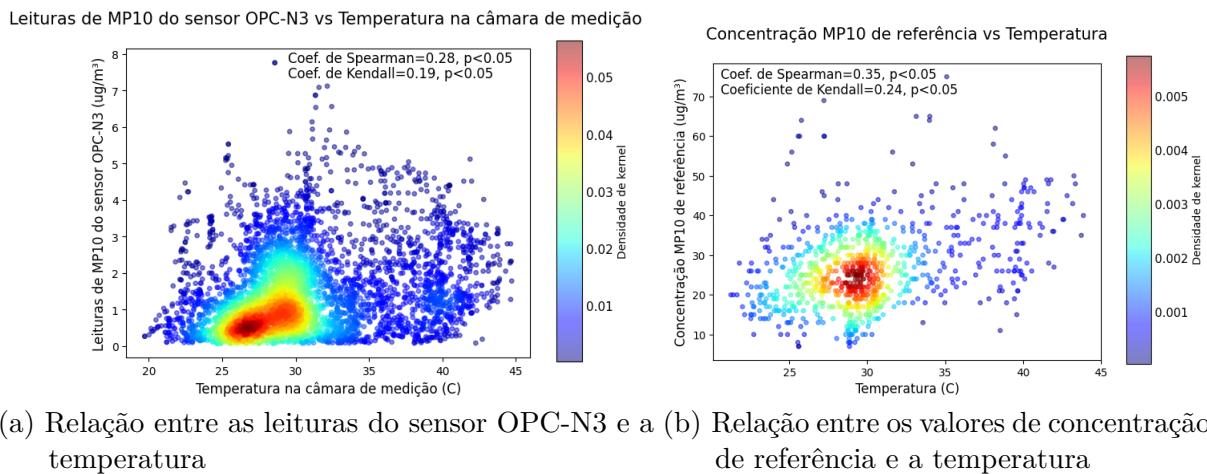
Figura 46 – Série temporal com $T = 1\text{ hr}$ Tabela 9 – Contabilização dos dados por etiquetas das leituras de MP_{10} do sensor OPC-N3

Série temporal $T = 15\text{ mins}$			Série temporal $T = 1\text{ hr}$		
Etiquetas	No. amostras	% amostras	Etiquetas	No. amostras	% amostras
<i>MISSING</i>	6481	44.57 %	<i>LOWSAMPLES</i>	2291	63.96 %
<i>LTL</i>	1759	12.10 %	<i>VALID</i>	1291	36.04 %
<i>GTUL</i>	0	0.0 %			
<i>STABILIZING</i>	0	0.0 %			
<i>BADSPIKE</i>	430	2.96 %			
<i>LTQTL01</i>	117	0.80 %			
<i>GTQTL99</i>	86	0.59 %			
<i>REBASE</i>	0	0.0 %			
<i>VALID</i>	5668	38.98 %			
TOTAL	14541		TOTAL	3582	

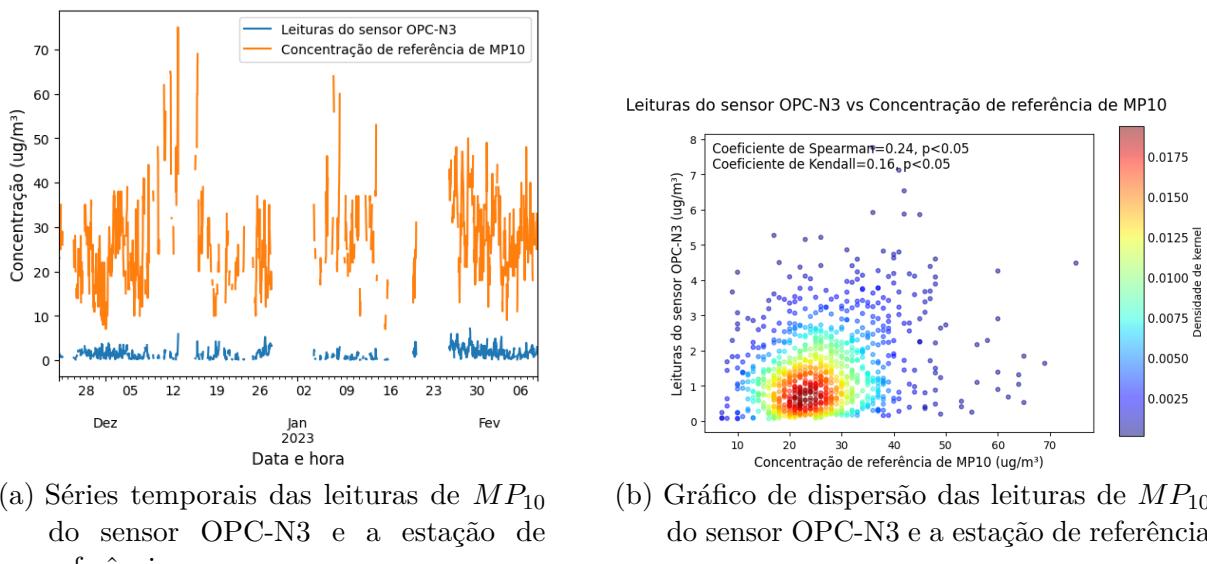
dispersão entre as variáveis. Na Figura 47a observa-se que o núcleo principal dos dados de concentração, entre 0 e $2\text{ }\mu\text{g}/\text{m}^3$, mostrou uma tendência crescente com as variações de temperatura entre 25 a 30°C . Da mesma forma, ao analisar a relação entre as medições de concentração de referência e a temperatura, observa-se correlação, com coeficientes de Spearman e Kendall de 0.35 e 0.25 respectivamente (Figura 47b).

3.6.1 Comparação das leituras de MP_{10} do sensor OPC-N3 com as medições de referência

Nas Figuras 48a e 48b apresentam-se as leituras de MP_{10} obtidas pelo sensor OPC-N3 de Alphasense e a estação de referência. Observa-se que as leituras do sensor OPC-N3 subestimaram os valores de concentração de referência com valores aproximadamente 10 vezes menores. Os testes de Spearman e Kendall revelaram a existência de correlação entre as medições com o sensor de baixo custo e a referência com coeficientes de 0.3 e 0.2

Figura 47 – Relação dos dados de concentração de MP_{10} com a temperatura

respectivamente.

Figura 48 – Séries temporais e gráficos de dispersão das medições de MP_{10} 

3.7 DISCUSSÃO

Dos resultados descritos anteriormente observa-se que de modo geral a porcentagem das leituras que foram aproveitadas como válidas foi baixa, sendo os principais motivos a perda de dados (dados com leituras inválidas ou NaN) e as alterações de linha base. As leituras de MP_{10} apresentaram a maior porcentagem de dados válidos com aproximadamente 39 % das leituras. Isso pode ter sido possível pelo fato do sensor OPC-N3 ter uma saída digital, produzindo maior estabilidade nas suas leituras. O sensor NO2-B43F teve a menor porcentagem com menos de 7 % de dados válidos. Quase 43 % das suas leituras sofreram alterações na linha base que impossibilitaram

seu aproveitamento, o restante foram dados perdidos ou no período de estabilização. Dentre os sensores eletroquímicos utilizados, os que melhor desempenharam em termos de quantidade de leituras aproveitáveis foram os dois sensores OX-B431, que tiveram 30 e 74 % de leituras válidas aproximadamente.

Os dados perdidos estão associados a falhas no equipamento como quedas de energia, falhas na leitura dos sensores que produziram valores NaN, erros de escrita no cartão de memória e erros na obtenção de carimbos de data e hora para os dados. Por outro lado, o motivo das alterações frequentes de linha base não foi possível descobrir a partir das informações com que se conta até o momento. O fato das alterações ocorrerem em momentos distintos para cada sensor faz pensar que não foram produzidas por falhas generalizadas na operação do equipamento, e sim com o funcionamento de sensores particulares.

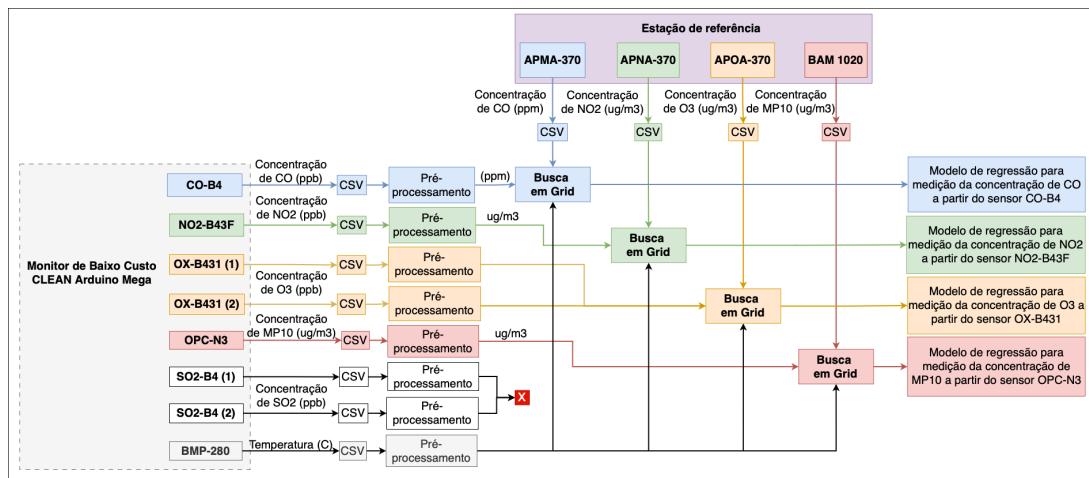
Uma vez identificados os dados válidos comprovou-se que todos apresentaram distribuições log-normais, conforme o esperado para dados de concentração. Os sensores CO-B4 e OX-B431 (2) mostraram uma pequena bimodalidade ocasionada por uma possível dependência com a temperatura, já que foi identificado um perfil sazonal ao longo do dia coincidente com o perfil de temperatura diário. Em geral, com a exceção do sensor NO2-B43F, observou-se certa correlação entre as leituras dos sensores e a temperatura, principalmente nos dados de O_3 . Esse comportamento era esperado dada a influência da radiação na concentração de O_3 no ambiente, e pôde ser corroborado na análise das concentrações de referência. O sensor CO-B4 também gerou leituras correlacionadas com a temperatura, que não coincidiu com o comportamento dos dados de referência. A correlação das medições de referência foi baixa ($\rho = 0.18$) em comparação com o sensor de baixo custo ($\rho = 0.52$). É possível que essa suposta relação com a temperatura, inclusive no sensor de referência, tenha estado associada com padrões diários de poluição, sendo que a maior atividade antropogênica acontece na luz do dia, onde as temperaturas são mais elevadas. Contudo, o coeficiente de correlação mais elevado no sensor de baixo custo pode ser indício de uma certa influência com a temperatura, conforme já tem sido reportado na literatura. Nas leituras de NO_2 e MP_{10} a relação com a temperatura foi pouco perceptível.

De modo geral, a exceção dos sensores de O_3 , o restante das leituras dos sensores de baixo custo apresentaram baixa correlação com as medições de referência, e diferenças elevadas na escala das medições. A maior correlação dos dados de ozônio pode explicar-se pelo próprio comportamento diário da variável que foi registrado pelos sensores, mas observou-se que o sensor OX-B431 (1) superestimou em quase 3 vezes as leituras originais, já o sensor 2 mostrou valores mais coerentes. A mesma diferença na ordem dos dados foi observada nas leituras do sensor NO2-B43F, com variações na escala de aproximadamente 4 vezes em relação com os dados originais.

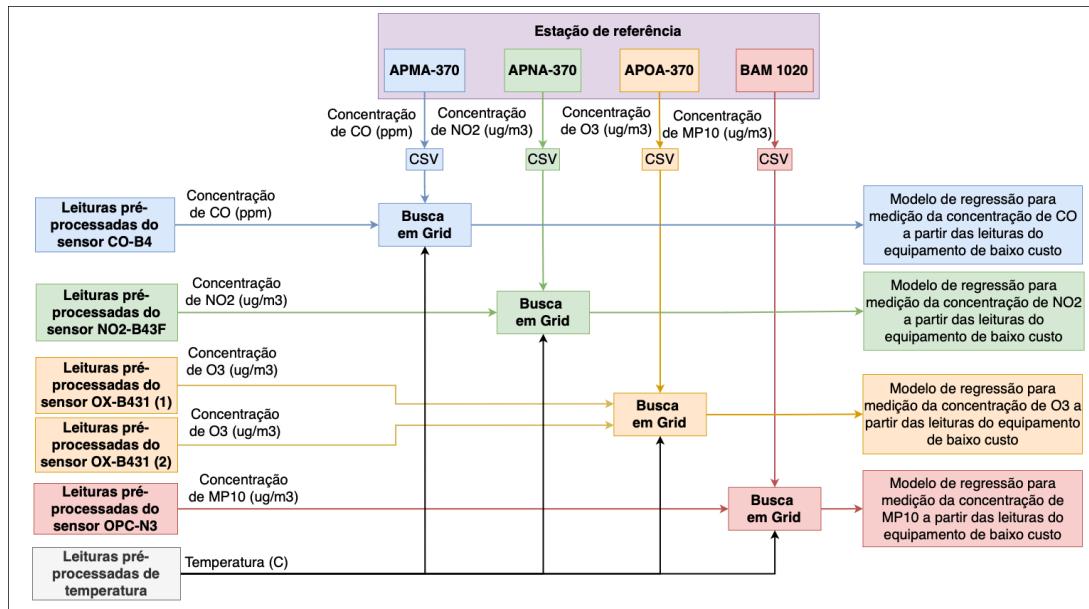
4 METODOLOGIA DE CORREÇÃO DE LEITURAS PROVENIENTES DE SENsoRES DE BAIXO CUSTO DA QUALIDADE DO AR

No Capítulo 3 foi apresentado o processo de co-localização do dispositivo de baixo custo com a estação de referência e foram comparadas as leituras de cada equipamento. A última etapa do processamento, que será abordada neste capítulo, consistiu em desenvolver uma metodologia para corrigir as medições adquiridas pelo equipamento de baixo custo a partir dos dados de referência da estação certificada.

Figura 49 – Fluxogramas do processo de correção das leituras do equipamento de baixo custo



(a) Etapa de correção das leituras de cada sensor

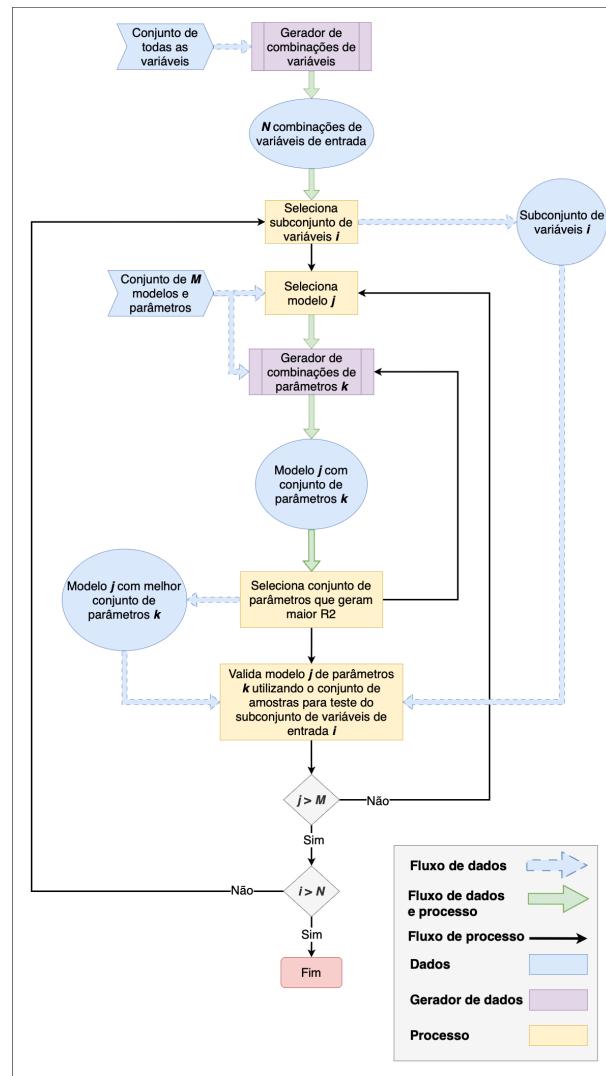


(b) Etapa de correção das leituras do equipamento considerando todos seus sensores

O processo de correção foi realizado em duas etapas, ilustradas nas Figuras 49a e 49b. O Anexo D apresenta o código utilizado para aplicar as correções às leituras dos sensores OX-B431. Em um primeiro momento (Figura 49a), aplicou-se uma busca em

grid para encontrar o modelo de regressão que melhor corrigisse as leituras dos sensores de cada poluente. Nesta primeira etapa, foram consideradas como entradas ao modelo apenas as leituras dos sensores específicos para cada composto juntamente com os valores de temperatura. No segundo momento (Figura 49b) aplicou-se também uma busca em *grid* para encontrar um modelo de correção por poluente, mas dessa vez considerando as leituras de todos os sensores de baixo custo no equipamento, além da temperatura. Os dados de entrada nas etapas de correção provieram da etapa de pré-processamento dos dados descrita no Capítulo 3.

Figura 50 – Processo de geração de combinações de variáveis, modelos e seus parâmetros



Para a correção das leituras adquiridas pelos sensores foram explorados diferentes modelos de regressão multivariados, i.e.: o Perceptron Multicamadas (MLP), Regressão Linear Multivariada (MLR), K Vizinhos mais Próximos (KNN) e as Florestas Aleatórias (RF). Para encontrar o modelo que melhor explicasse os dados foram realizadas buscas em *grid* que combinaram diferentes parâmetros e variáveis de entrada. Os modelos foram avaliados utilizando validações cruzadas com $k = 3$ considerando métricas essenciais, i.e.:

o coeficiente de determinação (R^2), o erro médio quadrático (RMSE) e o erro absoluto médio (MAE). Como a complexidade dos modelos aumenta com o incremento das variáveis independentes, também foram consideradas as métricas de AIC e BIC na avaliação dos modelos de regressão.

A Figura 50 apresenta um fluxograma da metodologia aplicada nas duas etapas de correção. O processo consiste em três laços. O laço mais externo itera ao longo de uma lista com as distintas combinações de variáveis de entrada. Esta lista, dependendo da etapa de correção, pode estar composta pelas leituras de um ou dois sensores de um mesmo poluente com a temperatura, ou por todos os sensores do equipamento e a temperatura. Para gerar as combinações de variáveis foi considerado que cada sensor poderia representar apenas uma entrada, e que cada combinação deveria ter pelo menos um sensor do poluente em questão. Exemplificando, a primeira etapa de correção das leituras de O_3 iteraria sobre a lista de combinatórias formada por:

```
[[Leituras OX-B431 (1)],  
 [Leituras OX-B431 (2)],  
 [Leituras OX-B431 (1), Leituras OX-B431 (2)],  
 [Leituras OX-B431 (1), Leituras de temperatura],  
 [Leituras OX-B431 (2), Leituras de temperatura],  
 [Leituras OX-B431 (1), Leituras OX-B431 (2), Leituras de temperatura]]
```

Já na segunda etapa de correção das leituras de O_3 não poderia ser considerada, por exemplo, a seguinte combinação de variáveis de entrada, já que esta não possui nenhuma leitura de sensor de O_3 .

```
[Leituras CO-B4, Leituras NO2-B43F, Leituras de OPC-N3, Leituras de Temperatura]
```

Uma vez gerado o conjunto de combinações de variáveis de entrada, o segundo e terceiro laço iteram por conjuntos de modelos de regressão e de parâmetros desses modelos. Em cada iteração do terceiro laço são selecionados os melhores parâmetros para cada modelo de regressão. Para isso utilizou-se a classe *GridSearchCV* do pacote *sklearn.model_selection* de *Python*. A Tabela 10 resume os parâmetros que ajustados para cada modelo. Uma vez selecionados os parâmetros, o segundo laço calcula os valores de R^2 , RMSE, MAE, AIC e BIC produzidos por cada modelo utilizando validações cruzadas com $k = 3$.

4.1 CORREÇÃO DAS LEITURAS DO SENSOR CO-B4 COM AS MEDIÇÕES DE REFERÊNCIA

A partir dos dados de referência e das leituras de concentração e temperatura adquiridas pelo monitor em questão, foi realizada uma busca em *grid* para encontrar as melhores combinações de parâmetros e variáveis de entrada aos modelos de regressão. As variáveis que foram testadas como entrada foram as leituras de concentração do sensor CO-B4 e a temperatura no interior da câmara de medição. Na Tabela 11 resumem-se os melhores modelos encontrados pela busca em *grid* para corrigir as leituras do sensor

Tabela 10 – Parâmetros ajustados para cada modelo nas buscas em *grid*

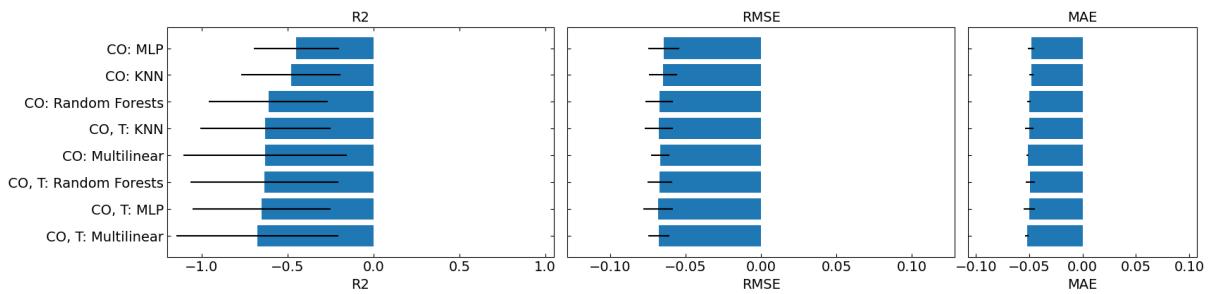
Modelo	Parâmetros
MLP	No. de camadas ocultas
	Coeficiente <i>alpha</i>
KNN	No. de vizinhos
RF	No. de estimadores
	Profundidade máxima
	No. de amostras para dividir um nó
	No. mínimo de amostras por folha
MLR	—

CO-B4. Os mesmos resultados são ilustrados graficamente na Figura 51 que apresenta o desempenho dos modelos e as variáveis de entrada considerando os valores de r2, RMSE e MAE.

Tabela 11 – Resultados da calibração do sensor CO-B4

Var.	Modelo	R2	RMSE	MAE	ρ
CO	MLP:	-0.45 ± 0.25	-0.06 ± 0.01	-0.05	0.47
	MLR:	-0.63 ± 0.48	-0.07 ± 0.01	-0.05	0.33
	KNN:	-0.48 ± 0.29	-0.06 ± 0.01	-0.05	0.46
	RF:	-0.61 ± 0.35	-0.07 ± 0.01	-0.05	0.39
CO, T	MLP:	-0.65 ± 0.40	-0.07 ± 0.01	-0.05 ± 0.01	0.50
	MLR:	-0.68 ± 0.47	-0.07 ± 0.01	-0.05	0.31
	KNN:	-0.63 ± 0.38	-0.07 ± 0.01	-0.05	0.53
	RF:	-0.63 ± 0.43	-0.07 ± 0.01	-0.05	0.47

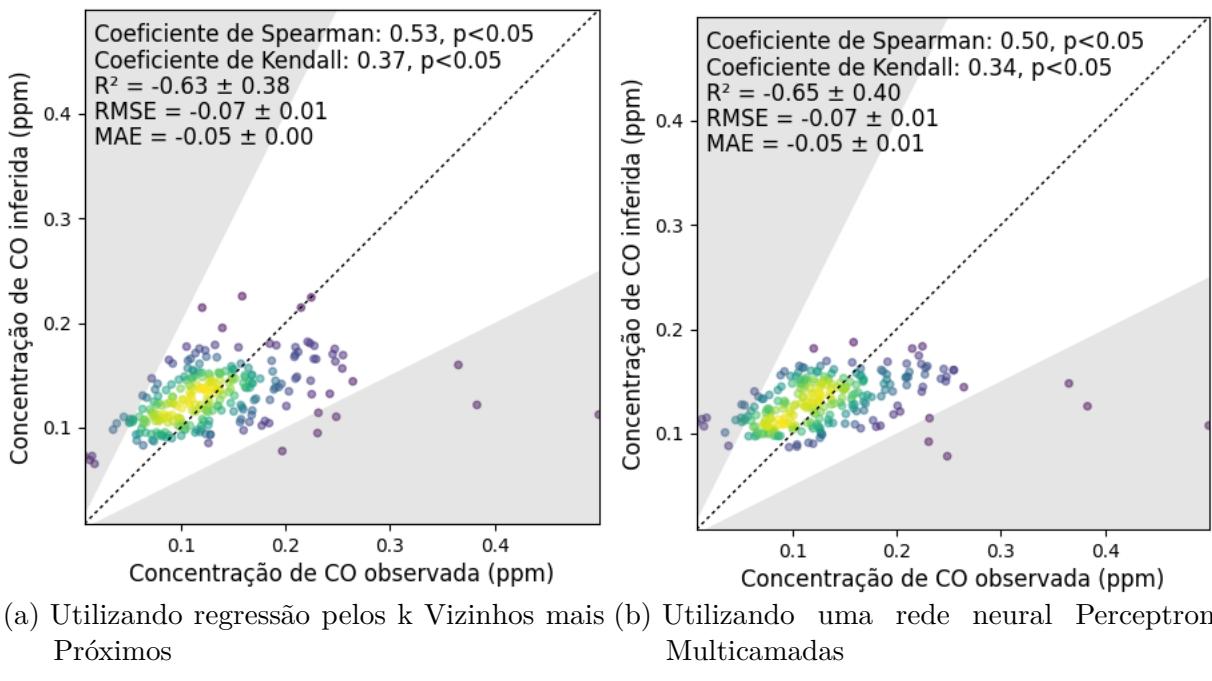
Figura 51 – Resultados dos modelos de regressão aplicados as leituras do sensor CO-B4



Como se observa, todas as variantes de modelos e variáveis de entrada apresentaram valores de R2 negativos, indicando que nenhum dos modelos foi capaz de explicar a variância na variável dependente, i.e. a concentração real. Contudo, os modelos não lineares produziram maiores coeficientes de correlação do que as regressões lineares, e

de igual modo, os modelos não lineares que incluíram a temperatura, apresentaram melhorias na correlação em comparação com os que não consideraram a temperatura como variável de entrada. Os maiores coeficientes de Spearman foram de 0.53 com uma regressão KNN considerando a temperatura, e de 0.50 utilizando uma rede neural MLP também considerando a temperatura. As Figuras 52a e 52b apresentam os resultados ao aplicar os modelos de k Vizinhos mais Próximos e Perceptron Multicamadas.

Figura 52 – Gráfico de dispersão das leituras do sensor CO-B4 e a estação de referência após aplicar modelos de regressão considerando a temperatura

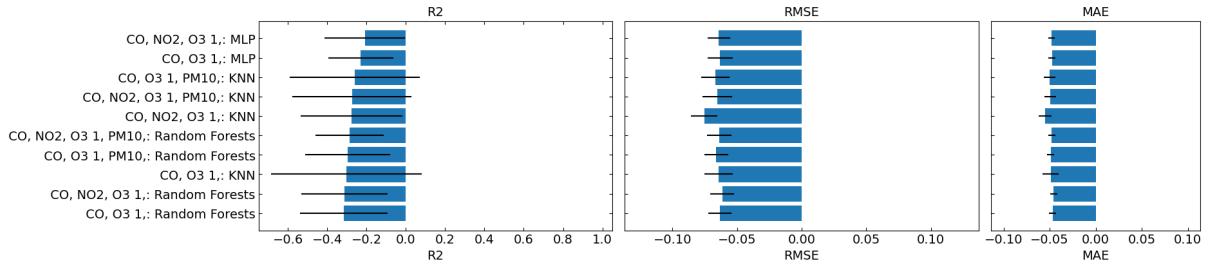


4.2 CÁLCULO DA CONCENTRAÇÃO DE MONÓXIDO DE CARBONO A PARTIR DAS LEITURAS DO ARRANJO DE SENSORES DE GASES

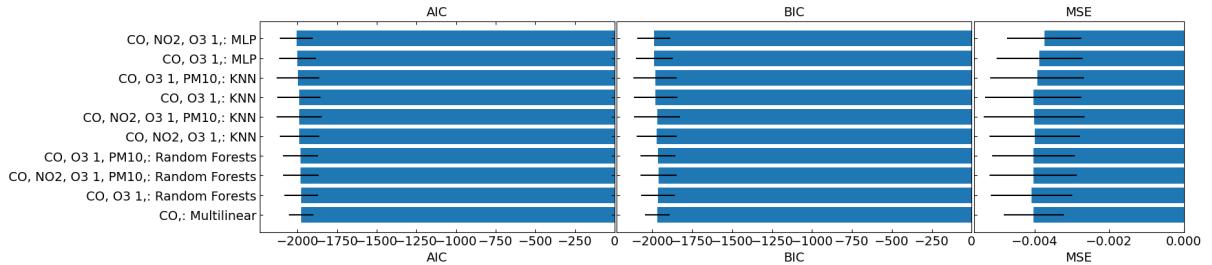
A Figura 53a apresenta os valores de R² dos 10 melhores modelos de calibração calculados para as leituras de CO. Observa-se que apesar do valor médio de R² obtido nas validações cruzadas continuar sendo negativo, obtiveram-se máximos de até aproximadamente 0.1 para alguns conjuntos de dados de teste nas validações cruzadas. Em geral os modelos que produziram os melhores resultados foram baseados em regressões pelos k vizinhos mais próximos e redes neurais Perceptron Multicamadas. Com relação as variáveis de entrada dos modelos, todos os que produziram melhores resultados consideraram as leituras do sensor 1 de O₃. Nenhum deles considerou a temperatura.

Ao comparar os modelos em termos de sua complexidade, observa-se uma sobreposição entre os que desempenharam melhor em termos de representação dos dados originais (maiores valores de R²) e os que desempenharam melhor em termos de complexidade (menores valores de AIC e BIC). A Figura 53b compara os valores de

Figura 53 – Desempenho dos modelos de regressão aplicados para inferir as leituras de concentração de *CO* medidas pela estação de referência

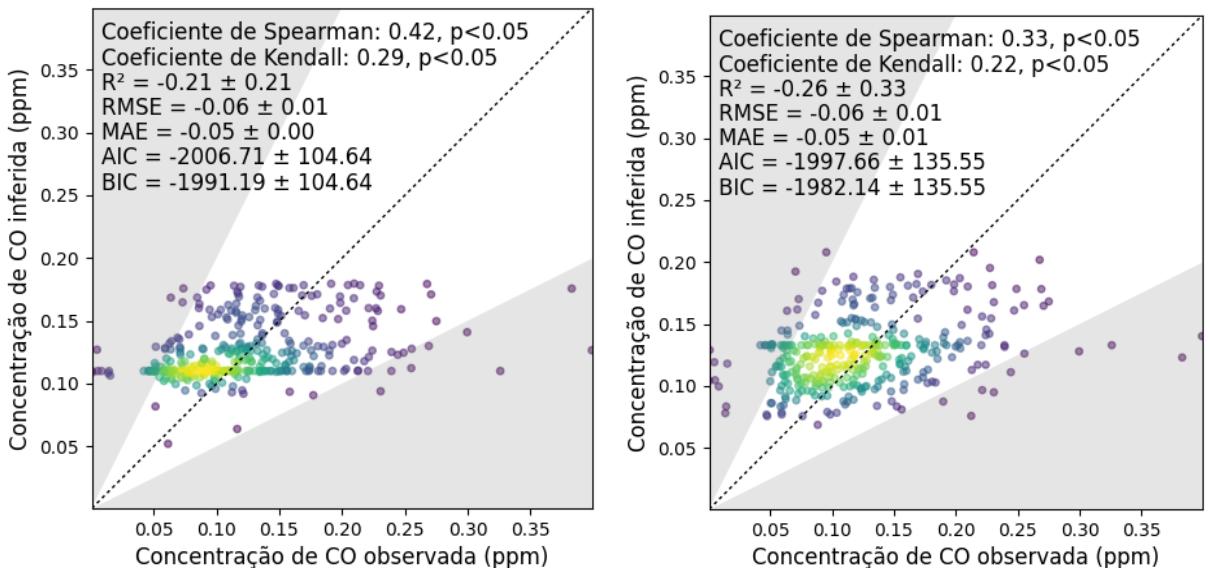


(a) Valores de R2, RMSE e MAE obtidos pelos 10 modelos com maiores valores de R2



(b) Modelos com menores valores de AIC e BIC

Figura 54 – Gráfico de dispersão das leituras do múltiplos sensores e a estação de referência para medição de *CO*



(a) Utilizando modelo de regressão com uma rede neural Perceptron Multicamadas; variáveis independentes: leituras de sensores CO-B4, NO2-B43F e OX-B431 (1)

(b) Utilizando modelo de regressão pelos k vizinhos mais próximos; variáveis independentes: leituras de sensores CO-B4, OX-B431 (1) e de MP_{10} medido pelo OPC-N3

AIC, BIC e MSE dos 10 modelos que obtiveram menores valores de AIC. Por último as Figuras 54a e 54b mostram gráficos de dispersão entre os valores de saída dos modelos de calibração e os dados de referência de *CO*. Os gráficos mostram os resultados dos

modelos com melhores valores de R2.

Tabela 12 – Resultados da calibração dos sensores OX-B431

Var.	Modelo	R2	RMSE	MAE	ρ
O_3 (1)	MLP:	-0.62 ± 0.46	-18.52 ± 2.88	-14.33 ± 2.13	0.15
	MLR	-0.41 ± 0.29	-17.64 ± 3.78	-13.81 ± 2.76	0.14
	KNN:	-0.48 ± 0.32	-18.36 ± 5.32	-13.65 ± 4.35	0.13
	RF:	-0.61 ± 0.63	-18.20 ± 3.30	-14.53 ± 2.29	0.12
O_3 (1), T	MLP:	0.40 ± 0.16	-11.29 ± 1.55	-8.71 ± 1.30	0.67
	MLR:	0.29 ± 0.21	-12.17 ± 1.28	-9.50 ± 1.17	0.64
	KNN:	0.34 ± 0.21	-11.68 ± 1.22	-8.92 ± 1.12	0.66
	RF:	0.33 ± 0.19	-11.87 ± 1.34	-9.10 ± 1.20	0.63
O_3 (2)	MLP:	0.16 ± 0.13	-13.72 ± 2.11	-10.87 ± 1.66	0.54
	MLR	0.09 ± 0.14	-14.34 ± 2.50	-11.35 ± 2.13	0.56
	KNN:	0.03 ± 0.35	-14.27 +/- 1.49	-11.27 ± 1.39	0.54
	RF:	-0.03 ± 0.37	-14.76 ± 1.47	-11.55 ± 1.34	0.52
O_3 (2), T	MLP:	0.23 ± 0.17	-13.16 ± 2.94	-10.14 ± 2.62	0.67
	MLR:	0.38 ± 0.22	-11.44 ± 0.72	-8.86 ± 0.86	0.69
	KNN:	0.19 ± 0.34	-12.91 ± 0.81	-9.89 ± 0.92	0.70
	RF:	0.28 ± 0.22	-12.44 ± 1.11	-9.54 ± 1.08	0.67
O_3 (1), O_3 (2)	MLP:	0.24 ± 0.15	-13.01 ± 1.82	-10.02 ± 1.55	0.60
	MLR	0.14 ± 0.16	-13.85 ± 2.26	-10.81 ± 2.02	0.58
	KNN:	0.12 ± 0.24	-13.90 ± 1.86	-10.86 ± 1.72	0.58
	RF:	0.13 ± 0.24	-13.70 ± 1.54	-10.64 ± 1.43	0.57
O_3 (1), O_3 (2), T	MLP:	0.29 ± 0.21	-12.37 ± 1.59	-8.89 ± 0.82	0.71
	MLR:	0.39 ± 0.21	-11.37 ± 0.82	-8.82 ± 0.88	0.69
	KNN:	0.19 ± 0.34	-12.97 ± 1.18	-9.92 ± 1.37	0.72
	RF:	0.30 ± 0.19	-12.34 ± 1.61	-9.57 ± 1.58	0.68

4.3 CORREÇÃO DAS LEITURAS DOS SENsoRES OX-B431 COM AS MEDIÇÕES DE REFERÊNCIA

A partir dos dados de referência e das leituras de concentração e temperatura adquiridas pelo monitor em questão, foi realizada uma busca em grid para encontrar as melhores combinações de parâmetros e variáveis de entrada a modelos de regressão. As variáveis que foram testadas como entrada foram as leituras de concentração de O_3 dos dois sensores OX-B431 e a temperatura no interior da câmara de medição. Na Tabela 12

resumem-se os melhores modelos encontrados pela busca em *grid* para calibrar as leituras dos sensores OX-B431. São mostradas as diferentes combinações de variáveis de entrada testadas em cada iteração da busca pelos melhores modelos de regressão. Os mesmos resultados são ilustrados graficamente na Figura 55 que apresenta o desempenho dos modelos e as variáveis de entrada considerando os valores de R², RMSE e MAE.

Figura 55 – Resultados dos modelos de regressão aplicados às leituras dos sensores OX-B431

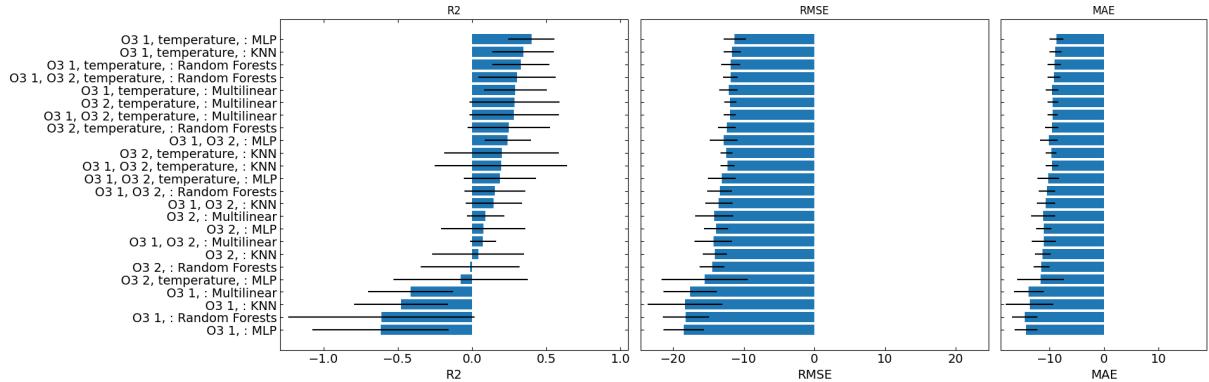
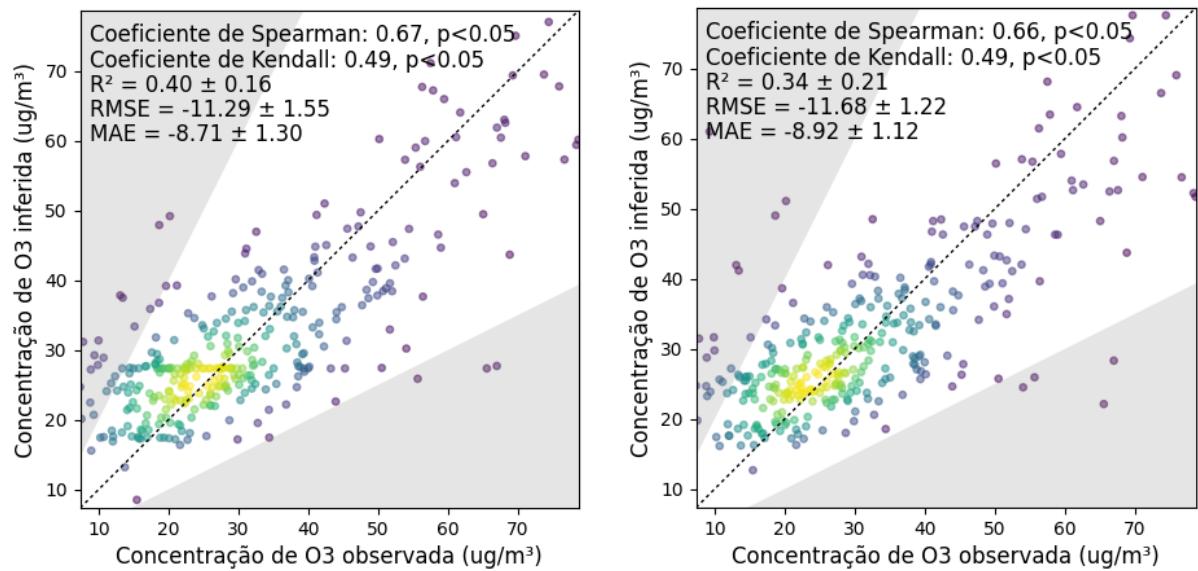


Figura 56 – Gráfico de dispersão das leituras dos sensores de O_3 OX-B431 e a estação de referência após aplicar modelos de regressão considerando a temperatura



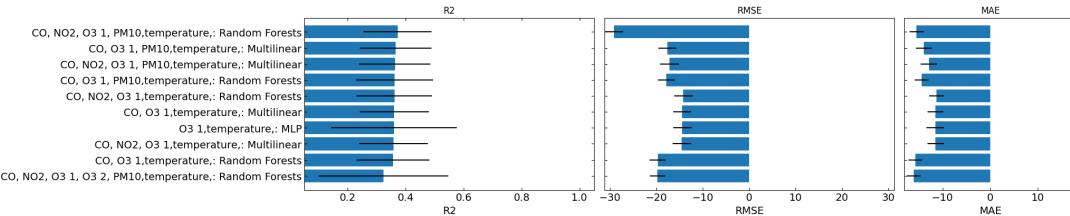
(a) Utilizando uma rede neural Perceptron Multicamadas obtiverem-se os melhores resultados de R², RMSE e MAE considerando as leituras do sensor 1 e a temperatura

(b) Utilizando uma regressão pelos k vizinhos mais próximos considerando as leituras do sensor 1 e a temperatura obtiveram-se resultados semelhantes

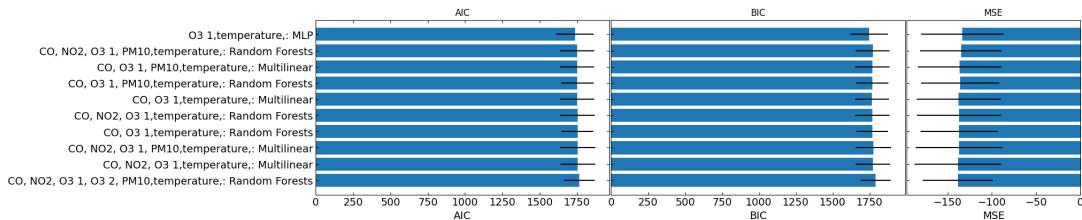
De modo geral observa-se que os modelos que consideraram a temperatura como variável de entrada produziram os melhores resultados de R², erro e correlação. A dupla

de variáveis sensor OX-B431 (1) e temperatura ocupou os primeiros lugares com diferentes modelos de regressão. Ao remover a temperatura, os modelos que melhor desempenharam consideraram os dois sensores de O_3 ; os sensores individualmente não produziram bons resultados. As Figuras 56a e 56b apresentam gráficos de dispersão com as leituras de referência e as inferidas pelos dois melhores modelos em termos de R2, RMSE e MAE.

Figura 57 – Desempenho dos modelos de regressão aplicados para inferir as leituras de concentração de O_3 medidas pela estação de referência



(a) Valores de R2, RMSE e MAE obtidos pelos 10 modelos com maiores valores de R2

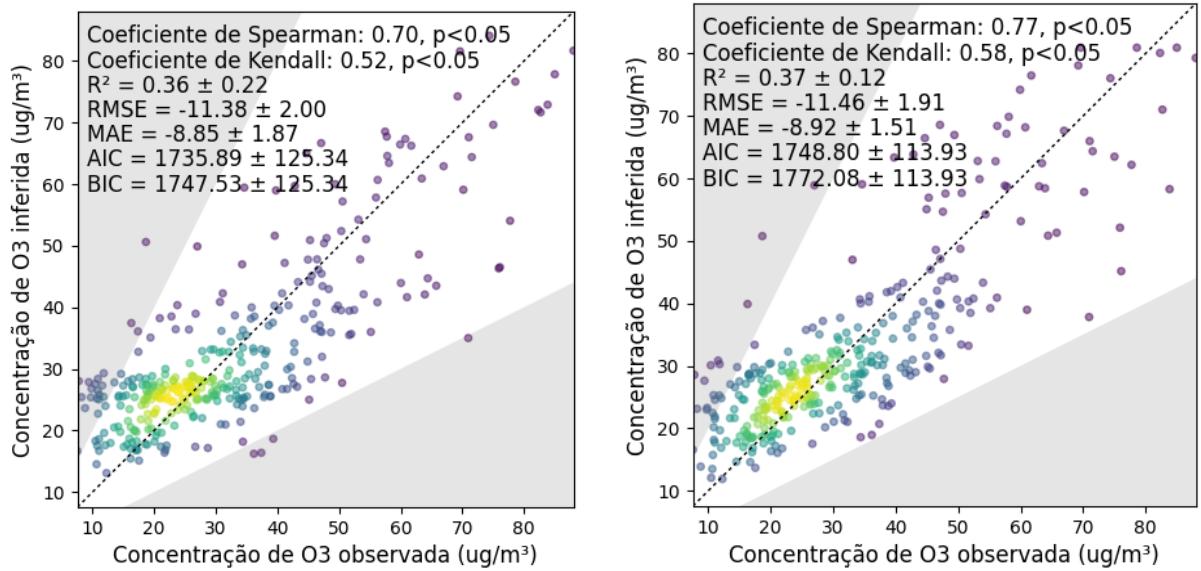


(b) Modelos com menores valores de AIC e BIC

4.4 CÁLCULO DA CONCENTRAÇÃO DE OZÔNIO A PARTIR DAS LEITURAS DO ARRANJO DE SENSORES DE GASES

As Figuras 57a e 57b apresenta os valores de R2, RMSE, MAE, AIC e BIC dos 10 melhores modelos de calibração calculados para as leituras de O_3 . Observa-se que os valores de R2 desses 10 modelos oscilaram entre 0.1 e 0.6, e que todos incluíram a temperatura. O Perceptron Multicamadas que considerou apenas as leituras do sensor OX-B431 (1) e a temperatura gerou o maior valor de $R^2 = 0.6$, mas também gerou um dos menores de aproximadamente 0.1. Este modelo também apresentou a menor complexidade em termos de coeficientes AIC e BIC. O modelo que teve o maior valor de R2 médio também apresentou o maior erro médio quadrático, e os que se encontraram nas posições 5 - 8 geraram os menores valores de RMSE. A Figura 58a mostra os resultados de aplicar uma rede neural Perceptron Multicamadas com variáveis de entrada OX-B431 (1) e temperatura, que gerou o maior R2, um dos menores RMSE e menor complexidade. A Figura 58b, por sua parte, mostra os resultados obtidos com Florestas Aleatórias e variáveis de entrada CO-B4, NO2-B43F, OX-B431 (1), MP_{10} do OPC-N3 e temperatura, que foi a combinação com maior R2 médio e a segunda com menor complexidade segundo os valores de AIC e BIC.

Figura 58 – Gráfico de dispersão das leituras do múltiplos sensores e a estação de referência para medição de O_3



- (a) Utilizando uma rede neural perceptron multicamadas com variáveis independentes: leituras de sensor OX-B431 (1) e temperatura
- (b) Utilizando modelo de regressão de Florestas Aleatórias com variáveis independentes: leituras de sensores CO-B4, NO2-B43F, OX-B431 (1), sensor de MP_{10} OPC-N3 e temperatura

Tabela 13 – Resultados da calibração das leituras de NO_2 do sensor NO2-B43F

Var.	Modelo	R2	RMSE	MAE	ρ
NO_2	MLP:	-3.95 ± 5.68	-12.87 ± 3.90	-9.94 ± 2.52	–
	MLR:	-3.80 ± 6.39	-12.61 ± 3.74	-9.62 ± 2.17	–
	KNN:	-3.73 ± 5.49	-12.84 ± 3.67	-9.92 ± 2.10	–
	RF:	-4.39 ± 6.07	-13.70 ± 3.50	-10.33 ± 2.08	–
NO_2, T	MLP:	-2.84 ± 4.36	-11.76 ± 3.40	-9.00 ± 1.81	0.59
	MLR:	-3.62 ± 5.88	-12.57 ± 3.40	-9.70 ± 1.87	–
	KNN:	-2.62 ± 4.52	-12.02 ± 3.33	-8.88 ± 1.54	0.61
	RF:	-2.66 ± 4.51	-11.93 ± 3.39	-9.18 ± 2.03	0.63

4.5 CORREÇÃO DAS LEITURAS DO SENSOR NO2-B43F COM AS MEDIÇÕES DE REFERÊNCIA

A partir dos dados de referência e das leituras de concentração e temperatura adquiridas pelo monitor em questão, foi realizada uma busca em grid para encontrar as melhores combinações de parâmetros e variáveis de entrada a modelos de regressão. As variáveis que foram testadas como entrada foram as leituras de concentração de NO_2 do sensor NO2-B43F e a temperatura no interior da câmara de medição. Na Tabela 13 resumem-se os melhores modelos encontrados pela busca em *grid* para calibrar as leituras

do sensor de baixo custo. Os mesmos resultados são ilustrados graficamente na Figura 59 que apresenta o desempenho dos modelos e as variáveis de entrada considerando os valores de R², RMSE e MAE.

Figura 59 – Resultados dos modelos de calibração aplicados as leituras de NO₂ do sensor NO2-B43F

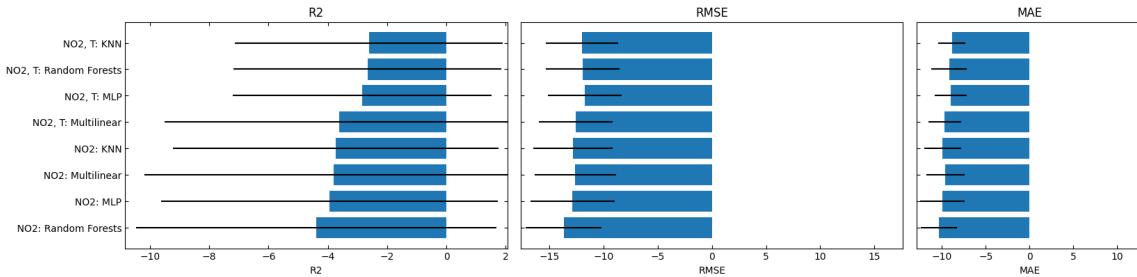
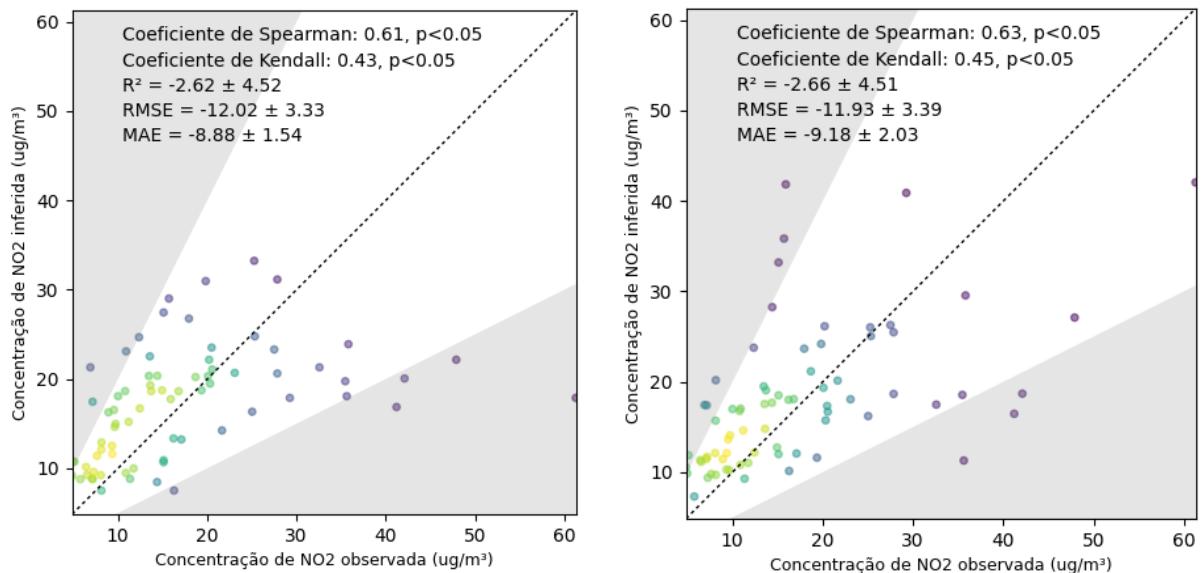


Figura 60 – Gráfico de dispersão das leituras do sensor de NO₂ NO2-B43F e a estação de referência após aplicar modelos de regressão considerando a temperatura

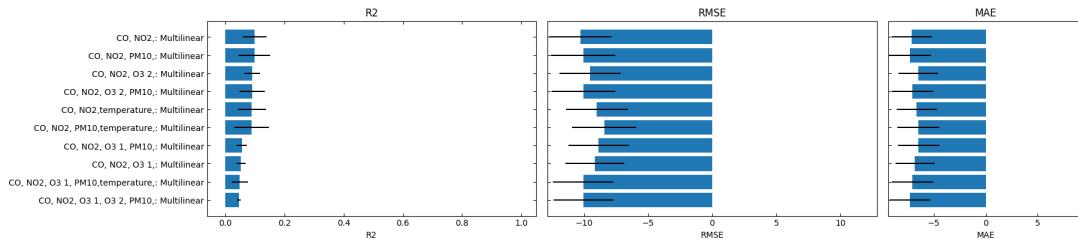


(a) Utilizando uma regressão pelos k vizinhos mais próximos considerando a temperatura obteve-se um ρ de 0.61

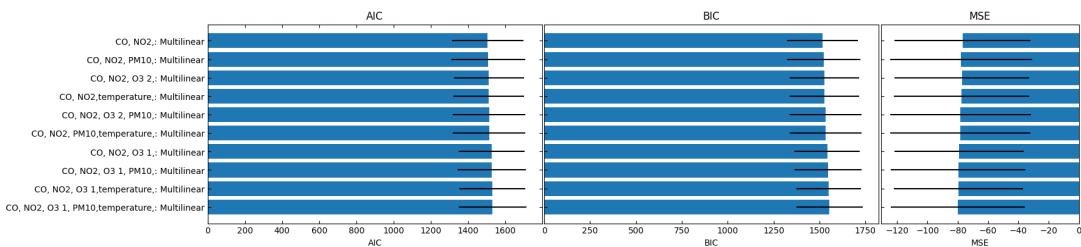
(b) Utilizando uma regressão por Florestas Aleatórias considerando a temperatura obteve-se um valor de ρ de 0.63

Como se observa todas as variantes de modelos e variáveis de entrada apresentaram valores de R² negativos, indicando que nenhum dos modelos foi capaz de explicar a variância na variável dependente, i.e. a concentração real. Observa-se também que nenhum dos modelos uni-variados conseguiu que os dados observados e os inferidos apresentassem alguma correlação. Por outro lado, os modelos multivariados, a exceção da regressão linear, que incluíram a temperatura como variável de entrada, produziram coeficientes de correlação entre a concentração real e a medida pelo sensor, entre 0.59 - 0.63. As Figuras 60a e 60b apresentam os resultados ao aplicar os modelos de k Vizinhos mais Próximos e Florestas Aleatórias.

Figura 61 – Desempenho dos modelos de regressão aplicados para inferir as leituras de concentração de NO_2 medidas pela estação de referência



(a) Valores de R2, RMSE e MAE obtidos pelos 10 modelos com maiores valores de R2



(b) Modelos com menores valores de AIC e BIC

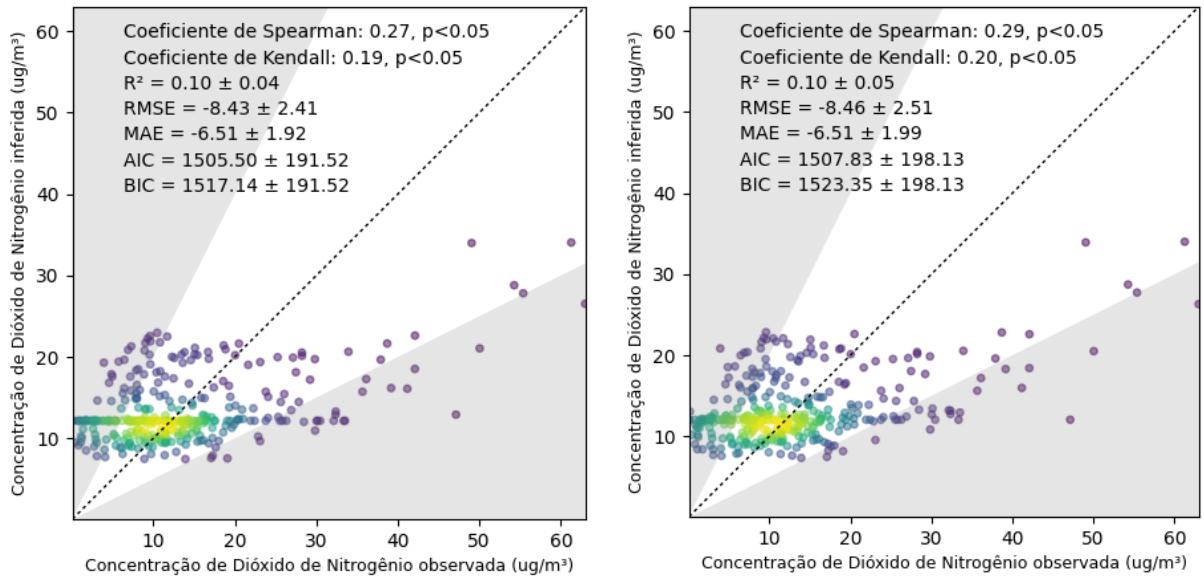
4.6 CÁLCULO DA CONCENTRAÇÃO DE DIÓXIDO DE NITROGÊNIO A PARTIR DAS LEITURAS DO ARRANJO DE SENSORES DE GASES

A Figura 61a apresenta os valores de R2 dos 10 melhores modelos de calibração calculados para as leituras de NO_2 . Observa-se que os valores de R2 desses 10 modelos apresentaram valores de R2 em média positivos, com valores máximos de até aproximadamente 0.2, todos obtidos a partir de regressões lineares. Com relação as variáveis de entrada observa-se que todos os 10 modelos consideraram o CO com variações nos restantes das variáveis para cada modelo. Com relação à complexidade dos modelos (Figura 61) observa-se que o ranqueamento por AIC coincidiu bastante com o ranqueamento por R2. As Figuras 62a e 62b mostram os resultados obtidos com os dois modelos com maior R2 médio, i.e.: regressões lineares com variáveis de entrada leituras de sensores CO-B4 e NO2B43F, e leituras de sensores CO-B4, NO2B43F e sensor de MP_{10} OPC-N3, respectivamente. As figuras mostram gráficos de dispersão entre os dados calibrados por esses modelos e as leituras de referência.

4.7 CORREÇÃO DAS LEITURAS DE MP10 DO SENSOR OPC-N3 COM AS MEDIÇÕES DE REFERÊNCIA

A partir dos dados de referência e das leituras de concentração e temperatura adquiridas pelo monitor em questão, foi realizada uma busca em grid para encontrar as melhores combinações de parâmetros e variáveis de entrada a modelos de regressão. As variáveis que foram testadas como entrada foram as leituras de concentração de MP_{10} do sensor OPC-N3 e a temperatura no interior da câmara de medição. Como modelos

Figura 62 – Gráfico de dispersão das leituras de múltiplos sensores e a estação de referência para medição de NO_2



- (a) Utilizando modelo de regressão linear multivariado com variáveis independentes: leituras de sensores CO-B4, e NO₂-B43F
(b) Utilizando modelo de regressão linear multivariado com variáveis independentes: leituras de sensores CO-B4, NO₂-B43F e sensor de MP_{10} OPC-N3

Tabela 14 – Resultados da calibração das leituras de MP_{10} do sensor OPC-N3

Var.	Modelo	R2	RMSE	MAE	ρ
MP_{10}	MLP:	-0.05 ± 0.036	-9.77 ± 0.75	-7.41 ± 0.49	0.18
	MLR:	-0.01 ± 0.03	-9.57 ± 0.98	-7.26 ± 0.62	0.17
	KNN:	-0.14 ± 0.07	-10.18 ± 0.62	-7.71 ± 0.45	0.13
	RF:	-0.19 ± 0.11	-10.33 ± 0.52	-7.80 ± 0.40	0.22
MP_{10}, T	MLP:	-0.29 ± 0.27	-10.68 ± 0.96	-8.33 ± 0.78	0.47
	MLR:	0.10 ± 0.08	-9.04 ± 1.12	-6.73 ± 0.69	0.37
	KNN:	-0.02 ± 0.09	-9.57 ± 0.46	-7.29 ± 0.27	0.45
	RF:	-0.17 ± 0.27	-10.12 ± 0.32	-7.89 ± 0.28	0.46

de regressão foram testados: o Perceptron Multicamadas (MLP), a Regressão Linear Multivariada (MLR), os K Vizinhos mais Próximos (KNN) e as Florestas Aleatórias (RF). Na Tabela 14 resumem-se os melhores modelos encontrados pela busca em *grid* para calibrar as leituras do sensor OPC-N3. Os mesmos resultados são ilustrados graficamente na Figura 63 que apresenta o desempenho dos modelos e as variáveis de entrada considerando os valores de r2, RMSE e MAE.

Como se observa, apenas o modelo de regressão linear com variáveis independentes temperatura e concentração medida pelo sensor OPC-N3 conseguiu explicar a variância na variável dependente, i.e. a concentração real. Os valores de R2 obtidos nas validações cruzadas realizadas para este modelo foram de 0.10 ± 0.08 , e os coeficientes de correlação

Figura 63 – Resultados dos modelos de calibração aplicados as leituras de MP_{10} do sensor OPC-N3

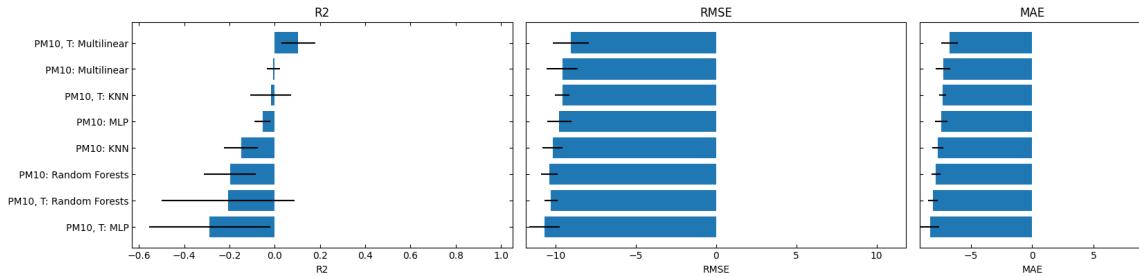
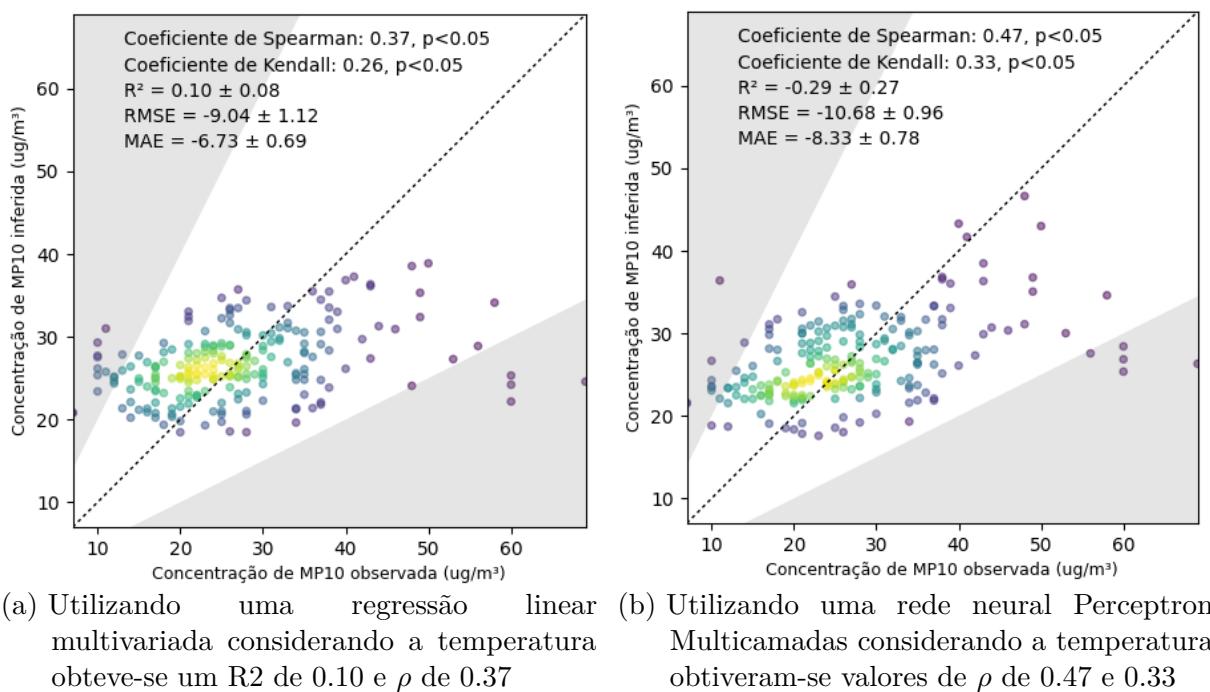
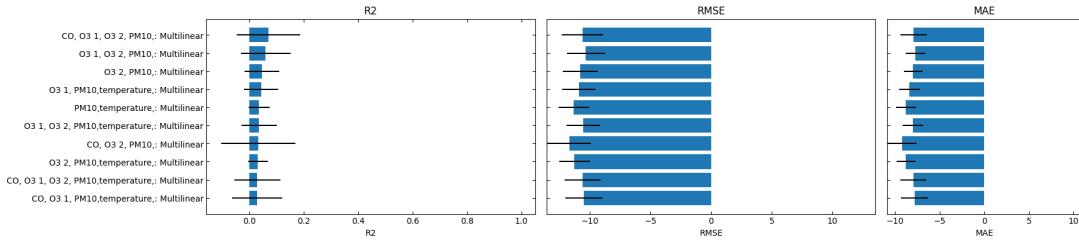


Figura 64 – Gráfico de dispersão das leituras do sensor de MP_{10} do OPC-N3 e a estação de referência após aplicar modelos de regressão considerando a temperatura

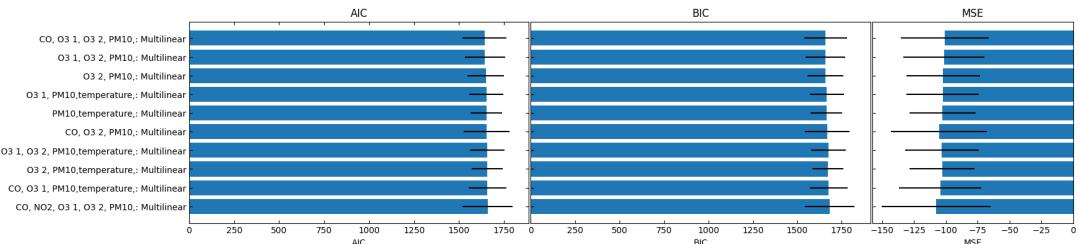


de Spearman e Kendall entre a concentração inferida na calibração e a concentração real foram de 0.37 e 0.26 respectivamente. Os erros RMSE e MAE estiveram próximos de $10 \mu\text{g}/\text{m}^3$, que é um valor relativamente baixo considerando a diferença elevada entre os valores de referência e as leituras do sensor (aproximadamente 10 vezes). De modo geral, os modelos não lineares que incluíram a temperatura como variável dependente, apresentaram melhorias na correlação entre a concentração real e a medida pelo sensor, com coeficientes de correlação de até 0.47. As Figuras 64a e 64b apresentam os resultados ao aplicar o modelo de regressão linear considerando leituras do sensor e temperatura e ao aplicar uma rede neural Perceptron Multicamadas considerando as mesmas variáveis de entrada.

Figura 65 – Desempenho dos modelos de regressão aplicados para inferir as leituras de concentração de MP_{10} medidas pela estação de referência



(a) Valores de R2, RMSE e MAE obtidos pelos 10 modelos com maiores valores de R2



(b) Modelos com menores valores de AIC e BIC

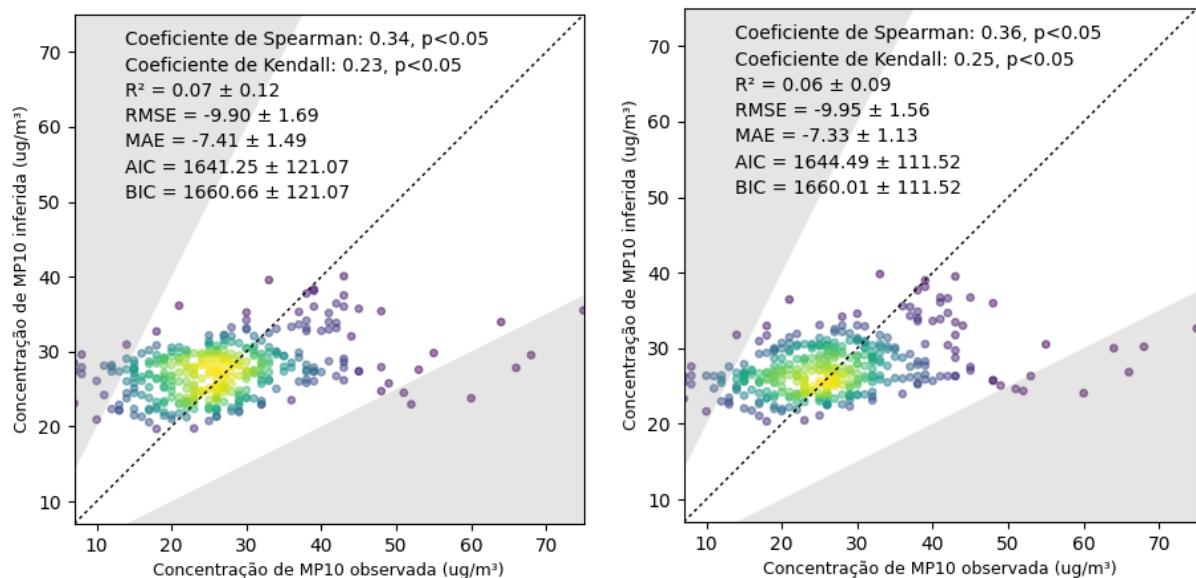
4.8 CÁLCULO DA CONCENTRAÇÃO DE MP10 A PARTIR DAS LEITURAS DO ARRANJO DE SENsoRES DE GASES

A Figura 65a apresenta os valores de R2 dos 10 melhores modelos de calibração calculados para as leituras de MP_{10} . Observa-se que os valores de R2 desses 10 modelos apresentaram valores de R2 em média positivos, com valores mínimos e máximos oscilando entre -0.1 e 0.2, todos obtidos a partir de regressões lineares. Com relação as variáveis de entrada observa-se maior variabilidade do que nos casos analisados anteriormente, mas em geral destaca-se a presença de sensores de O_3 em 9 dos modelos e em segundo lugar da temperatura, presente em 6. Nenhum dos modelos incluiu leituras do sensor NO2-B43F. Com relação à complexidade dos modelos (Figura 65) observa-se bastante coincidência com ranqueamento de R2. As Figuras 66a e 66b mostram os resultados obtidos com os dois modelos com maior R2 médio, i.e.: regressões lineares com variáveis de entrada leituras de sensores CO-B4, OX-B431 (1 e 2) e sensor de MP_{10} OPC-N3, e leituras de sensores OX-B431 (1 e 2) e sensor de MP_{10} OPC-N3, respectivamente. As figuras mostram gráficos de dispersão entre os dados calibrados por esses modelos e as leituras de referência.

4.9 DISCUSSÃO

Os modelos de regressão obtidos considerando apenas as leituras de sensores específicos para cada poluente tiveram um mal desempenho. Apenas as leituras dos sensores OX-B431 apresentaram coeficientes de determinação acima de 0.2, com máximos de até 0.6 e valores médios de 0.4 aproximadamente. O restante dos sensores produziu valores negativos de R2, indicando que a variância dos dados reais de concentração não

Figura 66 – Gráfico de dispersão das leituras do múltiplos sensores e a estação de referência para medição de MP_{10}



pôde ser explicada pelos modelos de regressão encontrados para os dados dos sensores utilizados. Contudo, a relação entre os dados de saída dos modelos e as leituras originais foi maior em comparação com os dados brutos, chegando a coeficientes de Spearman entre 0.47 e 0.61 entre todos os sensores. É importante ressaltar que os modelos que geraram melhores resultados, todos eles consideraram a temperatura e saída do sensor como variáveis independentes. A Tabela 15 resume os melhores resultados obtidos para cada conjunto de sensores.

Resulta interessante destacar a diferença de comportamento entre os dois sensores OX-B431. Mesmo sendo sensores do mesmo modelo e do mesmo lote, eles apresentaram comportamentos muito distintos com relação as leituras de referência e a dependência com a temperatura. Isso corrobora o reportado na literatura com relação ao comportamento dos sensores de baixo custo.

Tabela 15 – Resultados dos melhores modelos considerando leituras de sensores por poluentes

Var.	Modelo	R2	RMSE	MAE	ρ
$CO + T$	KNN	-0.63 ± 0.47	-0.07 ± 0.01	-0.05	0.53
$O_3 (1) + T$	MLP	0.40 ± 0.16	-11.29 ± 1.55	-8.71 ± 1.30	0.67
$NO_2 + T$	RF	-2.66 ± 4.51	-11.93 ± 3.39	-9.18 ± 2.03	0.63
$MP_{10} + T$	MLP	-0.29 ± 0.27	-10.68 ± 0.96	-8.33 ± 0.78	0.47

Na Tabela 16 resumem-se os resultados dos melhores modelos por poluente ao considerar as leituras de todos os sensores como variáveis de entrada. Observa-se que de modo geral os parâmetros utilizados para medir o desempenho dos modelos melhorou com relação à Tabela 15. Os valores de R₂, embora ainda pequenos, passaram a ser positivos com exceção do CO, que no entanto mostrou valores de R₂ máximo de aproximadamente 0.1. As leituras de O₃ mantiveram-se no mesmo intervalo de valores de R₂, assim como os valores de erro quadrático e erro absoluto para o restante dos poluentes. Os modelos escolhidos apresentaram menores valores de coeficientes de correlação de Spearman devido a que o parâmetro escolhido para a otimização dos modelos foi o coeficiente de determinação R₂.

Tabela 16 – Resultados dos melhores modelos por poluente considerando as leituras de todos os sensores nos modelos

Poluente	Modelo	R2	RMSE	MAE	ρ
CO	MLP	-0.21 ± 0.21	-0.06 ± 0.01	-0.05	0.42
O ₃	MLR	0.37 ± 0.12	-11.46 ± 1.91	-8.92 ± 1.51	0.77
NO ₂	MLR	0.10 ± 0.04	-8.43 ± 2.41	-6.51 ± 1.92	0.27
MP ₁₀	MLR	0.07 ± 0.12	-9.90 ± 1.69	-7.41 ± 1.49	0.34

4.9.1 Comparação dos resultados obtidos com o equipamento desenvolvido em relação a trabalhos similares

Na revisão de (KANG *et al.*, 2022), os autores compilam os resultados reportados para uma série de equipamentos e sensores de baixo custo em um total de 112 artigos científicos. Ao compararmos os resultados obtidos no presente trabalho com o conjunto recolhido por Kang e co-autores, observa-se que, em geral, o instrumento desenvolvido teve um desempenho inferior à maioria dos trabalhos com sensores de baixo custo. A continuação localizamos os resultados obtidos para cada poluente dentro do espectro de reportes compilados por esses autores.

Entre os resultados reportados para a medição de MP₁₀, nenhum trabalho apresentou valores de R₂ inferiores a 0.2. A mediana de R₂ entre todos os trabalhos que utilizaram sensores Alphasense para medições em ar ambiente foi de 0.68. Já ao considerar trabalhos que utilizaram sensores de outros fabricantes, a mediana teve um valor de 0.71. Diante disso conclui-se que o desempenho das medições de MP₁₀ do nosso instrumento, com o qual obteve-se um R₂ máximo de 0.19, foi inferior ao total de trabalhos compilados em (KANG *et al.*, 2022).

Com relação à medição de CO, Kang reporta uma mediana de R₂ de 0.78 quando considerados todos os trabalhos que mediram dito poluente no ar ambiente. Já a mediana dos trabalhos que utilizaram sensores eletroquímicos Alphasense foi de 0.82. O valor máximo de R₂ obtido no presente trabalho para CO foi de 0.1, que representa menos do 10 % do total de trabalhos reportados por Kang que realizam medições em ar ambiente.

O valor mínimo de R2 reportado para medições com sensores Alphasense em ar ambiente foi de 0.3

As medições de NO_2 produziram melhores resultados. As medianas dos trabalhos em ar ambiente e em ar ambiente com sensores Alphasense foram de 0.57 e 0.68 respectivamente. O valor máximo de 0.14 obtido neste trabalho encontra-se entre o 10 % dos trabalhos que utilizaram sensores Alphasense, os quais reportaram valores de R2 entre 0 - 0.2 aproximadamente. Já se considerados todos os trabalhos realizados em ar ambiente, o valor máximo obtido pelo nosso equipamento está entre 10 - 25 % dos trabalhos compilados por Kang e co-autores.

Os melhores resultados obtiveram-se para as medições de O_3 . A mediana de R2 para medições em ar ambiente reportada por Kang foi de 0.78, e de 0.70 para medições com sensores Alphasense. O valor de R2 médio de 0.37 obtido no presente trabalho foi representado entre 20 e 25 % do total de trabalhos em área externa e entre 20 e 25 % do total de trabalhos em área externa com sensores Alphasense. Já o valor máximo de 0.5 obtido pelo equipamento desenvolvido representa um pouco menos do 25 % do total de trabalho e um pouco mais de 25 % do total de trabalhos com sensores Alphasense.

Apesar de que os resultados em comparação com outros trabalhos foram inferiores, ao considerar a evolução das leituras desde os dados brutos até os modelos multivariados considerando todos os sensores percebe-se uma melhoria na qualidade das leituras do instrumento desenvolvido. Assim, a metodologia de calibração proposta apresenta-se como mais uma alternativa para calibração de outros monitores de baixo custo. Essa metodologia, contudo, tem a ressalva de aumentar a complexidade dos modelos de calibração.

Cabe destacar que a temperatura ganhou menor destaque nesses modelos de regressão, já que apenas os modelos de O_3 incluíram a temperatura como variável independente. Por outro lado, todos os poluentes a exceção do NO_2 consideraram as leituras de pelo menos um dos sensores OX-B431. Dada a forte relação desses sensores com a temperatura, é possível que a sua inclusão tenha substituído a relação com a temperatura. A ausência de temperatura no modelo de NO_2 faz sentido dada a relação mostrada entre as variáveis.

CONCLUSÃO

Neste trabalho foi desenvolvida uma rede colaborativa para monitorização da qualidade do ar de baixo custo, a iniciativa CLEAN. Este iniciativa promove a colaboração para o desenvolvimento de plataformas de monitoramento de baixo custo, com custos mais baixos e com maior flexibilidade do que as atuais iniciativas de acesso aberto. A rede colaborativa pode incorporar um grupo mais amplo e diversificado de especialistas e entusiastas do monitoramento ambiental, aumentando a quantidade de dados disponíveis publicamente, diversificando as aplicações de monitoramento e melhorando a cobertura espaço-temporal da monitorização da qualidade do ar, especialmente nos países em desenvolvimento. CLEAN proporciona recursos de hardware e bibliotecas de firmware para diversas aplicações e usuários, assim como oferece uma aplicação backend e uma API REST, que facilita a integração de novos dispositivos à rede. Essa estrutura completa, aliada a uma aplicação frontend para visualização intuitiva dos dados, estabelece um ambiente para a integração de novos usuários e monitores, aumentando e diversificando o número de aplicações de monitoramento. O diferencial desta iniciativa com outras similares levantadas na revisão bibliográfica é que a estrutura modular das bibliotecas de firmware e a API para envio e leitura de dados para e do banco de dados contribuem na incorporação de novos colaboradores provém flexibilidade se para adaptar às especificações de cada aplicação e ao mesmo tempo uma estrutura sólida como base do desenvolvimento.

A iniciativa CLEAN promove a ciência cidadã, facilitando o processo de desenvolvimento dos dispositivos a um público mais vasto e disponibilizando dados sobre a qualidade do ar para análise e visualização. A rede de colaboradores pode se expandir para desenvolvedores, pesquisadores, amadores e estudantes que possam utilizar as ferramentas disponíveis no CLEAN para educação, prototipagem, uso pessoal e pesquisa. Uma cobertura mais ampla dos dados sobre a qualidade do ar disponibilizará uma quantidade considerável de dados, facilitando o acesso à informação ambiental para tornar as cidades e as povoações mais inclusivas, seguras, resilientes e sustentáveis. O acesso a informações representativas e fiáveis ajuda os investigadores e os decisores a encontrar soluções que promovam o florescimento de cidades mais inteligentes e saudáveis. Nos países em desenvolvimento, onde o acesso a tecnologias caras é mais limitado, iniciativas como o CLEAN são alternativas interessantes às redes de monitorização altamente dispendiosas que podem contribuir para o crescimento econômico sustentado e para um aumento da qualidade de vida dos cidadãos. Esta iniciativa tem potencial para expandir a comunidade de monitoramento do ar, especialmente no território brasileiro, e facilitar o acesso a informações sobre poluição atmosférica em regiões onde os dados regulatórios são escassos ou mesmo inexistentes.

Como parte da pesquisa também foram desenvolvidos cinco dispositivos de medição de qualidade do ar, baseados no framework Arduino. Dois deles foram protótipos para

medições em local fixo e medições móveis. Os outros três dispositivos foram desenvolvidos a partir de uma placa de circuito impresso versionada como CLEAN Arduino Mega possuindo melhor acabamento e robustez. Um desses equipamentos foi instalado por um período de 5 meses junto a uma estação de referência no município de Tubarão para validação e correção das suas leituras. Os outros dois equipamentos foram instalados posteriormente no mesmo local e se encontram coletando dados até o momento. Das leituras coletadas pelo primeiro dispositivo CLEAN Arduino Mega instalado em campo observou-se que as principais causas de dados inválidos foram as alterações de linha base e os períodos sem dados coletados devido a quedas de energia no local, falhas na comunicação com os sensores e falhas do equipamento. Versões futuras podem melhorar as interfaces de comunicação com os sensores utilizando placas dedicadas, com microcontrolador próprio que enviem as leituras de concentração por um canal digital e não mais analógico. Dessa forma isola-se a etapa de aquisição do restante do equipamento, adicionando maior proteção a ruído e interferências elétricas. Prover aos equipamentos de uma infraestrutura própria de energia e acesso à internet são outras ações que podem ser tomadas para acrescentar robustez e autonomia aos dispositivos, deixando-os menos propensos a falhas originadas na infraestrutura do local de instalação.

Em comparação com outros trabalhos que instalam equipamentos de baixo custo junto a estações certificadas em campo, o equipamento desenvolvido teve um baixo desempenho na medição dos poluentes. Contudo, com a metodologia proposta de correção de dados foi possível agregar valor às leituras. Corroborou-se que, considerando as leituras do conjunto de sensores é possível corrigir, ainda que com certo grau de erro, um conjunto de medições muito ruidosas e com dados faltantes, como foi o caso das medições de NO_2 . Além da metodologia de correção proposta, foi desenvolvida uma metodologia para o pré-tratamento dos dados utilizando várias técnicas reportadas na literatura. Essa metodologia possibilita detectar falhas e remover dados corrompidos, e pode ser incorporada aos equipamentos de medição para realizar validações dos dados em linha e em tempo real. Por exemplo, se uma batelada de medições coletadas pelo equipamento não tiver o 75 % dos dados válidos ela poderia ser descartada ou marcada para posterior correção por um serviço de nuvem. Se um equipamento apresentar sucessivas bateladas com falha, um alerta poderia ser gerado indicando uma falha de funcionamento.

Como próximos passos, recomenda-se uma atenção especial para o aprimoramento do hardware de aquisição de sinal, com o objetivo de reduzir o ruído e estabilizar o sinal de linha base. Essas melhorias podem resultar em medições ainda mais precisas e confiáveis, consolidando a contribuição significativa deste trabalho na promoção do monitoramento de baixo custo da qualidade do ar em diversos contextos. Recomenda-se também adicionar uma rotina de validação dos dados de forma contínua com notificações do estado do equipamento para o serviço em nuvem e adicionar hardware aos equipamentos para funcionarem de forma independente da infraestrutura do local de instalação, utilizando

conjunto de bateria e painel solar e explorando tecnologias de conectividade GPRS ou LoRa.

REFERÊNCIAS

AIR QUALITY EGG. Air Quality Egg - Web Portal. [S.l.: s.n.], 2023.
Disponível em: <https://airqualityegg.com/portal/>.

ALPHASENSE. Alphasense Application Note AAN 104 How Electrochemical Gas Sensors Work. [S.l.], 2013. P. 1–4. Disponível em:
https://www.alphasense.com/WEB1213/wp-content/uploads/2013/07/AAN_104.pdf.

_____. Alphasense Application Note AAN 110 ENVIRONMENTAL CHANGES: TEMPERATURE, PRESSURE, HUMIDITY. [S.l.], 2013. P. 1–6.
Disponível em:
https://www.alphasense.com/WEB1213/wp-content/uploads/2013/07/AAN_110.pdf.

_____. Alphasense Application Note AAN 803-05 Correcting for Background Currents in Four Electrode Toxic Gas Sensors. [S.l.], mar. 2019. P. 1–6.

_____. NO-B4 Nitric Oxide Sensor. Great Notley, UK, jul. 2019. P. 1–2.
Disponível em: www.alphasense.com.

ALTMAN, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. **American Statistician**, v. 46, n. 3, p. 175–185, 1992. ISSN 15372731. DOI: 10.1080/00031305.1992.10475879.

BARON, Ronan; SAFFELL, John. Amperometric Gas Sensors as a Low Cost Emerging Technology Platform for Air Quality Monitoring Applications: A Review. **ACS Sensors**, v. 2, n. 11, p. 1553–1566, nov. 2017. ISSN 2379-3694. DOI: 10.1021/acssensors.7b00620. Disponível em:
<https://pubs.acs.org/doi/10.1021/acssensors.7b00620>.

CAMPO, Fernando. The CLEAN Initiative. [S.l.: s.n.], 2021. Disponível em:
<https://lcqar.ufsc.br/novo/index.php/documentacao-clean/>.

CAMPO, Fernando *et al.* DEPLOYMENT OF MOBILE AND FIXED AIR SENSOR PLATFORMS IN THE CITY OF FLORIANÓPOLIS, BRAZIL: PRELIMINARY RESULTS. In: 19TH Annual CMAS Conference, Chapel Hill, North Carolina. Chapel Hill, North Carolina: [s.n.], out. 2020. P. 1–6. Disponível em: <https://www.cmascenter.org/conference/2020/abstracts/Campo-et-al-2020.pdf>.

CASTELL, Nuria; DAUGE, Franck R. *et al.* Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? **Environment International**, Pergamon, v. 99, p. 293–302, fev. 2017. ISSN 0160-4120. DOI: 10.1016/J.ENVINT.2016.12.007. Disponível em:
<https://www.sciencedirect.com/science/article/pii/S0160412016309989>.

CASTELL, Nuria; SCHNEIDER, Philipp *et al.* Localized real-time information on outdoor air quality at kindergartens in Oslo, Norway using low-cost sensor nodes. **Environmental Research**, Academic Press Inc., v. 165, p. 410–419, ago. 2018. ISSN 10960953. DOI: 10.1016/j.envres.2017.10.019.

CETESB. **Redes de Monitoramento | Qualidade do Ar.** [S.l.: s.n.], 2020. Disponível em: <https://cetesb.sp.gov.br/ar/redes-de-monitoramento/>.

CONAMA. **Resolução n. 491, de 19 de novembro de 2018. Dispõe sobre padrões de qualidade do ar.** [S.l.: s.n.], 2018.

CONCAS, Francesco; MINERAUD, Julien; LAGERSPETZ, Eemil; VARJONEN, S.; PUOLAMÄKI, K. *et al.* A Gap Analysis of Low-Cost Outdoor Air Quality Sensor In-Field Calibration. **undefined**, 2019.

CONCAS, Francesco; MINERAUD, Julien; LAGERSPETZ, Eemil; VARJONEN, Samu; LIU, Xiaoli *et al.* **LOW-COST OUTDOOR AIR QUALITY MONITORING AND SENSOR CALIBRATION: A SURVEY AND CRITICAL ANALYSIS A PREPRINT.** [S.l.], 2021. Disponível em:
http://www.chinadaily.com.cn/china/2016-02/22/content_23595631.htm.

CONNERTON, Patrick *et al.* Use of Low-Cost Sensors for Environmental Health Surveillance: Wildfire-Related Particulate Matter Detection in Brasília, Brazil. **Atmosphere**, Multidisciplinary Digital Publishing Institute, v. 14, p. 1796, 12 dez. 2023. ISSN 2073-4433. DOI: 10.3390/atmos14121796. Disponível em:
<https://www.mdpi.com/2073-4433/14/12/1796>.

CROSS, Eben S. *et al.* Use of electrochemical sensors for measurement of air pollution: correcting interference response and validating measurements. **Atmospheric Measurement Techniques**, v. 10, n. 9, p. 3575–3588, set. 2017. ISSN 1867-8548. DOI: 10.5194/amt-10-3575-2017. Disponível em:
<https://www.atmos-meas-tech.net/10/3575/2017/>.

DE VITO, S. *et al.* Calibrating chemical multisensory devices for real world applications: An in-depth comparison of quantitative machine learning approaches. **Sensors and Actuators B: Chemical**, Elsevier, v. 255, p. 1191–1210, fev. 2018. ISSN 0925-4005. DOI: 10.1016/J.SNB.2017.07.155. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0925400517313692?via%3Dihub>.

DELAINE, Florentin; LEBENTAL, Bérénice; RIVANO, Hervé. **In Situ Calibration Algorithms for Environmental Sensor Networks: A Review.** v. 19. [S.l.]: Institute of Electrical e Electronics Engineers Inc., ago. 2019. P. 5968–5978. DOI: 10.1109/JSEN.2019.2910317.

EU. **Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe.** [S.l.]: Official

Journal of the European Union L 152/1 of 11.6.2008, 2008. P. 1–44. Disponível em: <https://eur-lex.europa.eu/>.

EUROPEU, Parlamento. Resolução do Parlamento Europeu, de 25 de março de 2021, sobre a aplicação das Diretivas relativas à qualidade do ar ambiente: Diretiva 2004/107/CE e Diretiva 2008/50/CE (2020/2091(INI)). [S.l.: s.n.], mar. 2021. Disponível em:
<https://www.eea.europa.eu/themes/air/health-impacts-of-air-pollution..>

FENG, Shaobin *et al.* Review on Smart Gas Sensing Technology. **Sensors**, Multidisciplinary Digital Publishing Institute, v. 19, n. 17, p. 3760, ago. 2019. ISSN 1424-8220. DOI: 10.3390/s19173760. Disponível em:
<https://www.mdpi.com/1424-8220/19/17/3760>.

FERREIRA, André L. *et al.* RECOMENDAÇÕES PARA A EXPANSÃO E A CONTINUIDADE DAS REDES DE MONITORAMENTO DA QUALIDADE DO AR NO BRASIL. [S.l.: s.n.], jul. 2022. P. 1–16. Disponível em: <https://energiaeambiente.org.br/produto/recomendacoes-para-a-expansao-e-a-continuidade-das-redes-de-monitoramento-da-qualidade-do-ar-no-brasil>.

FRANÇA, André Luiz Felisberto *et al.* GUIA TÉCNICO PARA O MONITORAMENTO E AVALIAÇÃO DA QUALIDADE DO AR. Brasília, DF, 2019. P. 1–135. Disponível em:
<https://www.mma.gov.br/agenda-ambiental-urbana/qualidade-do-ar.html>.

GÉRON, Aurélien. Mãoz à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow. [S.l.]: Alta Books, 2019. ISBN 9788550809021. Disponível em: <https://books.google.com.br/books?id=Z0mvDwAAQBAJ>.

HABITAT MAP. AirCasting. [S.l.: s.n.], 2023. Disponível em:
http://aircasting.habitatmap.org/mobile_map.

HAGAN, David H. *et al.* Calibration and assessment of electrochemical air quality sensors by co-location with regulatory-grade instruments. **Atmospheric Measurement Techniques**, Copernicus GmbH, v. 11, n. 1, p. 315–328, jan. 2018. ISSN 18678548. DOI: 10.5194/amt-11-315-2018.

HUANG, Keyong *et al.* Estimating daily PM2.5 concentrations in New York City at the neighborhood-scale: Implications for integrating non-regulatory measurements. **Science of the Total Environment**, Elsevier B.V., v. 697, p. 134094, dez. 2019. ISSN 18791026. DOI: 10.1016/j.scitotenv.2019.134094.

IEMA. Qualidade do Ar. [S.l.: s.n.], 2020. Disponível em:
<http://www.qualidadedoar.org.br/>.

- IEMA/ES. **IEMA - Qualidade do Ar.** [S.l.: s.n.], 2020. Disponível em: <https://iema.es.gov.br/qualidadedoar/redesdemonitoramento/automaticaramqar>.
- IMA/SC. **Rede de Monitoramento - SC.** [S.l.: s.n.], 2023. Disponível em: <https://www.ima.sc.gov.br/index.php/qualidade-ambiental/menu-qualidade-ar/rede-de-monitoramento#:~:text=Localizada%20na%20Rua%20Dr.,Vereadores%20da%20Cidade%20de%20Tubar%C3%A3o.&text=Localizada%20na%20Rua%20Ant%C3%B4nia%20Goularte,370%20da%20Cidade%20de%20Tubar%C3%A3o>.
- INSTITUTO DE ENERGIA E MEIO AMBIENTE. **Qualidade do Ar.** [S.l.: s.n.], 2019. Disponível em: <http://www.qualidadedoar.org.br/>.
- IQAIR. **IQAir.** [S.l.: s.n.], 2023. Disponível em: <https://www.iqair.com/us/air-quality-monitors>.
- JANG, Yu-Woon; JUNG, Gi-Woong. Temporal Characteristics and Sources of PM2.5 in Porto Velho of Amazon Region in Brazil from 2020 to 2022. **Sustainability**, Multidisciplinary Digital Publishing Institute, v. 15, p. 14012, 18 set. 2023. ISSN 2071-1050. DOI: 10.3390/su151814012. Disponível em: <https://www.mdpi.com/2071-1050/15/18/14012>.
- JBAILY, Abdulrahman *et al.* Air pollution exposure disparities across US population and income groups. **Nature**, Nature Research, v. 601, n. 7892, p. 228–233, jan. 2022. ISSN 14764687. DOI: 10.1038/s41586-021-04190-y. Disponível em: <https://doi.org/10.1038/s41586-021-04190-y>.
- JERRETT, Michael *et al.* Validating novel air pollution sensors to improve exposure estimates for epidemiological analyses and citizen science. **Environmental Research**, Academic Press, v. 158, p. 286–294, out. 2017. ISSN 0013-9351. DOI: 10.1016/J.ENVRES.2017.04.023. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0013935117307351?via%3Dihub>.
- JIAO, Wan *et al.* Community Air Sensor Network (CAIRSENSE) project: evaluation of low-cost sensor performance in a suburban environment in the southeastern United States. **Atmospheric Measurement Techniques**, v. 9, p. 5281–5292, 2016. DOI: 10.5194/amt-9-5281-2016. Disponível em: www.atmos-meas-tech.net/9/5281/2016/.
- KANG, Ye *et al.* **Performance evaluation of low-cost air quality sensors: A review.** v. 818. [S.l.]: Elsevier B.V., abr. 2022. P. 151769. DOI: 10.1016/j.scitotenv.2021.151769.
- KARAGULIAN, F *et al.* **Review of sensors for air quality monitoring.** Luxembourg, 2019. P. 1–54. ISBN 978-92-76-09255-1. DOI: 10.2760/568261. Disponível em: <https://op.europa.eu/en/publication-detail/-/publication/c5c8aadb-b33d-11e9-9d01-01aa75ed71a1/language-en>.

- KILLICK, R.; FEARNHEAD, P.; ECKLEY, I. A. Optimal detection of changepoints with a linear computational cost. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 107, n. 500, p. 1590–1598, dez. 2012. ISSN 01621459. DOI: 10.1080/01621459.2012.737745. Disponível em: <https://www.tandfonline.com/doi/full/10.1080/01621459.2012.737745>.
- KRAMER, Oliver. K-Nearest Neighbors. In: DIMENSIONALITY Reduction with Unsupervised Nearest Neighbors. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. P. 13–23. ISBN 978-3-642-38652-7. DOI: 10.1007/978-3-642-38652-7\}2. Disponível em: https://doi.org/10.1007/978-3-642-38652-7_2.
- KUMAR, Prashant *et al.* The rise of low-cost sensing for managing air pollution in cities. **Environment International**, Pergamon, v. 75, p. 199–205, fev. 2015. ISSN 0160-4120. DOI: 10.1016/J.ENVINT.2014.11.019. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0160412014003547?via%3Dihub>.
- LEWIS, Alastair C.; LEE, James D. *et al.* Evaluating the performance of low cost chemical sensors for air pollution research. **Faraday Discussions**, The Royal Society of Chemistry, v. 189, n. 0, p. 85–103, jul. 2016. ISSN 1359-6640. DOI: 10.1039/C5FD00201J. Disponível em: <http://xlink.rsc.org/?DOI=C5FD00201J>.
- LEWIS, Alastair C.; SCHNEIDEMESSER, Erika von *et al.* **Low-cost sensors for the measurement of atmospheric composition: overview of topic and future applications**. Geneva, mai. 2018. P. 1–60. ISBN 9789263112156. Disponível em: http://www.wmo.int/pages/prog/arep/gaw/documents/Draft_low_cost_sensors.pdf%20http://eprints.whiterose.ac.uk/135994/#.XmE_tzYtmFc.mendeley.
- MAAG, Balz; ZHOU, Zimu; THIELE, Lothar. A Survey on Sensor Calibration in Air Pollution Monitoring Deployments. **IEEE Internet of Things Journal**, Institute of Electrical e Electronics Engineers Inc., v. 5, n. 6, p. 4857–4870, dez. 2018. ISSN 23274662. DOI: 10.1109/JIOT.2018.2853660.
- MAHAJAN, Sachit *et al.* A citizen science approach for enhancing public understanding of air pollution. **Sustainable Cities and Society**, v. 52, p. 101800, 2020. ISSN 2210-6707. DOI: <https://doi.org/10.1016/j.scs.2019.101800>. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2210670719317020>.
- MALINGS, Carl *et al.* Development of a general calibration model and long-term performance evaluation of low-cost sensors for air pollutant gas monitoring. **Atmospheric Measurement Techniques**, v. 12, n. 2, p. 903–920, fev. 2019. DOI: 10.5194/amt-12-903-2019. Disponível em: <https://www.atmos-meas-tech.net/12/903/2019/>.
- MEAD, M.I. *et al.* The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks. **Atmospheric Environment**, Pergamon, v. 70, p. 186–203, mai. 2013. ISSN 1352-2310. DOI: 10.1016/J.ATMOSENV.2012.11.060.

Disponível em:

<https://www.sciencedirect.com/science/article/pii/S1352231012011284>.

MILLER, Max. **The Basics: KNN for classification and regression | by Max Miller | Towards Data Science.** [S.l.: s.n.], 2019. Disponível em:
<https://towardsdatascience.com/the-basics-knn-for-classification-and-regression-c1e8a6c955>.

MMA. **Qualidade do Ar.** [S.l.: s.n.], 2020. Disponível em:
<https://www.mma.gov.br/cidades-sustentaveis/qualidade-do-ar>.

MONTANTES, James. **3 Reasons to Use Random Forest Over a Neural Network—Comparing Machine Learning versus Deep Learning | by James Montantes | Towards Data Science.** [S.l.: s.n.], 2020. Disponível em:
<https://towardsdatascience.com/3-reasons-to-use-random-forest-over-a-neural-network-comparing-machine-learning-versus-deep-f9d65a154d89>.

MORAWSKA, Lidia *et al.* Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: How far have they gone? **Environment International**, Pergamon, v. 116, p. 286–299, jul. 2018. ISSN 0160-4120. DOI: 10.1016/J.ENVINT.2018.04.018. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0160412018302460?via%3Dihub>.

MOTLAGH, Naser Hossein *et al.* Toward Massive Scale Air Quality Monitoring. **IEEE Communications Magazine**, Institute of Electrical e Electronics Engineers (IEEE), v. 58, n. 2, p. 54–59, fev. 2020. ISSN 0163-6804. DOI: 10.1109/mcom.001.1900515.

MUNIR, Said *et al.* Analysing the performance of low-cost air quality sensors, their drivers, relative benefits and calibration in cities—a case study in Sheffield. **Environmental Monitoring and Assessment**, v. 191, n. 2, 2019. ISSN 15732959. DOI: 10.1007/s10661-019-7231-8.

OTTOSEN, Thor Bjørn; KUMAR, Prashant. Outlier detection and gap filling methodologies for low-cost air quality measurements. **Environmental Science: Processes and Impacts**, v. 21, n. 4, 2019. ISSN 20507895. DOI: 10.1039/c8em00593a.

OYAMA, Beatriz Sayuri; ZAMBONI, Ademilson. AIR QUALITY MONITORING AND DATA AVAILABILITY IN BRAZIL. In: _____. **Air Quality Conference Brazil, 3rd CMAS South America.** Vitória - ES, Brazil: Fundação Espírito Santo de Tecnologia, 2017. P. 170–172.

PANG, Xiaobing; SHAW, Marvin D.; GILLOT, Stefan *et al.* The impacts of water vapour and co-pollutants on the performance of electrochemical gas sensors used for air quality monitoring. **Sensors and Actuators B: Chemical**, v. 266, p. 674–684, 2018. ISSN 09254005. DOI: 10.1016/j.snb.2018.03.144.

- PANG, Xiaobing; SHAW, Marvin D.; LEWIS, Alastair C. *et al.* Electrochemical ozone sensors: A miniaturised alternative for ozone measurements in laboratory experiments and air-quality monitoring. **Sensors and Actuators, B: Chemical**, Elsevier B.V., v. 240, p. 829–837, mar. 2017. ISSN 09254005. DOI: 10.1016/j.snb.2016.09.020.
- PENZA, Michele. Low-cost sensors for outdoor air quality monitoring. In: ADVANCED Nanomaterials for Inexpensive Gas Microsensors. [S.l.]: Elsevier, 2020. P. 235–288. DOI: 10.1016/b978-0-12-814827-3.00012-8.
- PIEDRAHITA, R. *et al.* The next generation of low-cost personal air quality sensors for quantitative exposure monitoring. **Atmospheric Measurement Techniques**, v. 7, n. 10, p. 3325–3336, out. 2014. DOI: 10.5194/amt-7-3325-2014. Disponível em: <https://www.atmos-meas-tech.net/7/3325/2014/>.
- POPOOLA, Olalekan A.M. *et al.* Development of a baseline-temperature correction methodology for electrochemical sensors and its implications for long-term stability. **Atmospheric Environment**, Pergamon, v. 147, p. 330–343, dez. 2016. ISSN 1352-2310. DOI: 10.1016/J.ATMOSENV.2016.10.024. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1352231016308317?via%3Dihub>.
- PURPLEAIR. PurpleAir | Real-time Air Quality Monitoring. [S.l.: s.n.], 2023. Disponível em: <https://www2.purpleair.com/>.
- R. STETTER, Joseph; LI, Jing. Amperometric Gas SensorsA Review. **Chemical Reviews**, v. 108, n. 2, p. 352–366, jan. 2008. DOI: 10.1021/cr0681039.
- RAI, Aakash C. *et al.* End-user perspective of low-cost sensors for outdoor air pollution monitoring. [S.l.: s.n.], 2017. DOI: 10.1016/j.scitotenv.2017.06.266.
- RIBEIRO, Igor; SOUZA, Rodrigo. Sistema Eletrônico de Vigilância Ambiental. [S.l.: s.n.], 2021. Disponível em: <https://appselva.com.br>.
- SENSOR.COMMUNITY. Luft Daten Project Map. [S.l.: s.n.], 2023. Disponível em: <https://sensor.community/pt/>.
- SMART CITIZEN. Smart Citizen. [S.l.: s.n.], 2023. Disponível em: <https://smartcitizen.me/kits/>.
- SNYDER, Emily G. *et al.* The Changing Paradigm of Air Pollution Monitoring. **Environmental Science & Technology**, American Chemical Society, v. 47, n. 20, p. 11369–11377, out. 2013. ISSN 0013-936X. DOI: 10.1021/es4022602. Disponível em: <http://pubs.acs.org/doi/10.1021/es4022602>.
- SPEC SENSORS. Application Note AN-104 Short-Term Environment Effects on Performance. [S.l.], 2016. P. 1–6. Disponível em:

<http://www.spec-sensors.com/wp-content/uploads/2016/06/SPEC-AN-104-Environmental-Effects.pdf>.

SPEC SENSORS. **Digital Gas Sensor Developer Kit 968-045.** [S.l.], 2017. P. 1–10. Disponível em: https://www.spec-sensors.com/wp-content/uploads/2017/01/DG-SDK-968-045_9-6-17.pdf.

_____. **SPEC Sensor Operation Overview SPEC Sensor TM Operation and Performance Considerations.** [S.l.], 2016. P. 1–6. Disponível em: <http://www.spec-sensors.com/wp-content/uploads/2016/05/SPEC-Sensor-Operation-Overview.pdf>.

SPINELLE, Laurent; ALEIXANDRE, Manuel; GERBOLES, Michel. **Protocol of evaluation and calibration of low-cost gas sensors for the monitoring of air pollution.** [S.l.], 2013. P. 46. ISBN 9789279326912. DOI: 10.2788/9916. Disponível em: <http://publications.jrc.ec.europa.eu/repository/handle/JRC83791>.

SPINELLE, Laurent; GERBOLES, Michel *et al.* Field calibration of a cluster of low-cost available sensors for air quality monitoring. Part A: Ozone and nitrogen dioxide. **Sensors and Actuators B: Chemical**, Elsevier, v. 215, p. 249–257, ago. 2015. ISSN 0925-4005. DOI: 10.1016/J.SNB.2015.03.031. Disponível em: <https://www.sciencedirect.com/science/article/pii/S092540051500355X>.

_____. Field calibration of a cluster of low-cost commercially available sensors for air quality monitoring. Part B: NO, CO and CO₂. **Sensors and Actuators B: Chemical**, v. 238, p. 706–715, jan. 2017. ISSN 09254005. DOI: 10.1016/j.snb.2016.07.036. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S092540051631070X>.

TEIXEIRA, Francisco Sacco Flores Almeida. Renovar: Um MVP para monitorar a qualidade do ar, p. 1–157, 2018. Disponível em: <https://repositorio.ufsc.br/handle/123456789/192181>.

URAD MONITOR. **PM2.5 - uRADMonitor.** [S.l.: s.n.], 2023. Disponível em: <https://www.uradmonitor.com/>.

US EPA. **Criteria Air Pollutants | US EPA.** [S.l.: s.n.], 2021. Disponível em: <https://www.epa.gov/criteria-air-pollutants>.

VORMITTAG, Evangelina da Motta P. A. de Araújo *et al.* Análise do monitoramento da qualidade do ar no Brasil. **Estudos Avançados**, Instituto de Estudos Avançados da Universidade de São Paulo, v. 35, n. 102, p. 7–30, ago. 2021. ISSN 1806-9592. DOI: 10.1590/s0103-4014.2021.35102.002. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-40142021000200007&tlang=pt.

WEI, Peng *et al.* Impact Analysis of Temperature and Humidity Conditions on Electrochemical Sensor Response in Ambient Air Quality Monitoring. **Sensors**, Multidisciplinary Digital Publishing Institute, v. 18, n. 2, p. 59, jan. 2018. ISSN 1424-8220. DOI: 10.3390/s18020059. Disponível em: <http://www.mdpi.com/1424-8220/18/2/59>.

WESTBROEK, P. Fundamentals of electrochemistry. In: ANALYTICAL Electrochemistry in Textiles. [S.l.]: Elsevier Inc., jan. 2005. P. 3–36. ISBN 9781855739192. DOI: 10.1533/9781845690878.1.1.

WILLIAMS, Ron *et al.* **Air Sensor Guidebook**. Washington, DC: U.S. Environmental Protection Agency, EPA/600/R-14/159 (NTIS PB2015-100610), 2014. P. i–65. Disponível em: https://cfpub.epa.gov/si/si_public_record_report.cfm?Lab=NERL&dirEntryId=277996.

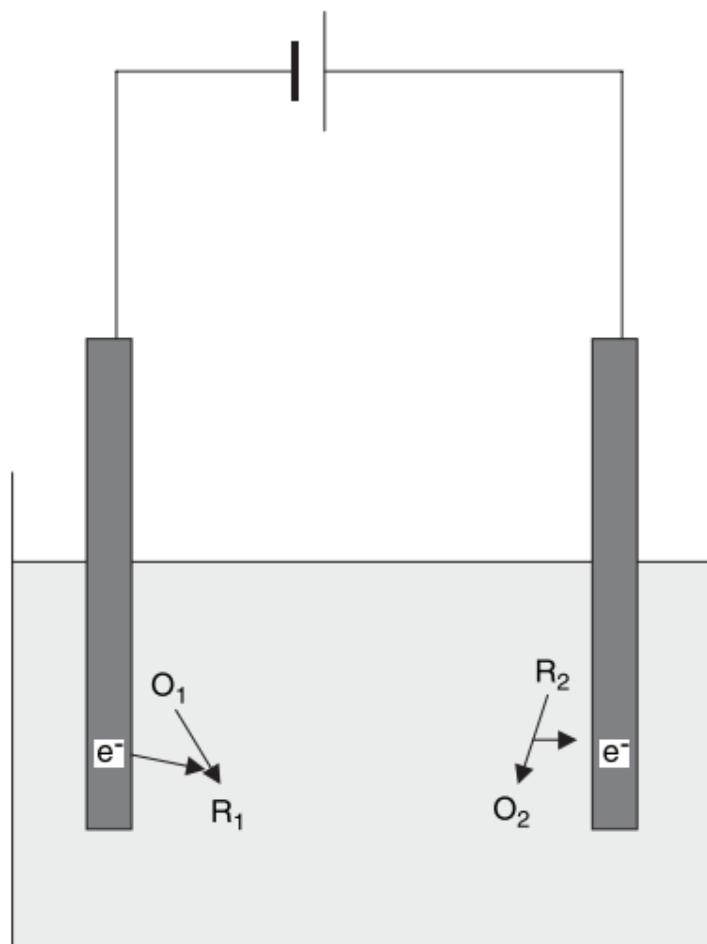
WORLD HEALTH ORGANIZATION. WHO global air quality guidelines: particulate matter (PM_{2.5} and PM₁₀), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. [S.l.: s.n.], 2021. ISBN 978-92-4-003422-8. Disponível em: <https://www.who.int/publications/i/item/9789240034228>.

ZIMMERMAN, Naomi *et al.* A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. **Atmospheric Measurement Techniques**, v. 11, n. 1, p. 291–313, jan. 2018. ISSN 1867-8548. DOI: 10.5194/amt-11-291-2018. Disponível em: <https://www.atmos-meas-tech.net/11/291/2018/>.

APÊNDICE A – SENSORES DE GASES ELETROQUÍMICOS

Os sensores eletroquímicos funcionam baseados no princípio de conversão de energia química em energia elétrica e vice-versa, e consistem, basicamente, em uma célula contendo uma solução eletrolítica na qual são submersos eletrodos metálicos interconectados por um circuito externo (WESTBROEK, 2005). Dentro da célula originam-se reações de oxidação-redução entre cada um dos eletrodos e a solução eletrolítica. As reações de redução removem elétrons do material do eletrodo, que passa a funcionar como catodo. Nas reações de oxidação, o eletrodo que faz o papel de anodo ganha elétrons e produz-se uma espécie oxidada. Esse intercâmbio de cargas entre os eletrodos e a solução analítica, equivalente a uma corrente elétrica fluindo do anodo para o catodo, é proporcional à velocidade das reações redox nas superfícies dos eletrodos (R. STETTER; LI, 2008). A Figura 67 mostra a estrutura básica de uma célula eletrolítica amperométrica de dois eletrodos.

Figura 67 – Representação de uma célula eletroquímica de dois eletrodos.



Fonte: (WESTBROEK, 2005).

Segundo seu princípio de operação, os sensores eletrolíticos são classificados em amperométricos, potenciométricos e condutimétricos (R. STETTER; LI, 2008). Neste trabalho apenas são abordados os sensores amperométricos por serem os mais comumente utilizados no monitoramento de baixo custo.

Os sensores amperométricos apresentam uma estrutura básica de três eletrodos: o eletrodo de trabalho, o eletrodo contador e o eletrodo de referência. O eletrodo de trabalho é a superfície onde acontece a reação de interesse entre o material do eletrodo, a solução eletrolítica e o gás sob estudo, que formam uma interface de três fases. Para aumentar a seletividade dos sensores, costuma-se aplicar algum catalisador na superfície do eletrodo para facilitar ou catalisar determinadas reações. Como este eletrodo está em contato direto com o ar ambiente, corre risco de envenenamento se exposto a certos gases que possam ser adsorvidos no catalisador ou que possam reagir com ele produzindo outros compostos químicos que inibam sua ação (ALPHASENSE, 2013a; WESTBROEK, 2005; R. STETTER; LI, 2008; BARON; SAFFELL, 2017).

A reação no eletrodo de trabalho pode ser de oxidação ou de redução, dependendo da natureza da substância gasosa de interesse. Na ausência de gás, a célula eletrolítica encontra-se em equilíbrio, mas, ao entrar em contato com a substância gasosa, as moléculas do gás produzem um desbalanço nas reações redox, e, como resultado, o eletrodo pode ganhar ou perder elétrons, obtendo assim uma carga elétrica (ALPHASENSE, 2013a; R. STETTER; LI, 2008).

A carga elétrica gerada no eletrodo de trabalho é balanceada com uma reação oposta no eletrodo contador. Se a reação no eletrodo de trabalho for de oxidação, então no contador se produzirá uma reação de redução complementar, e vice-versa. Como resultado, uma corrente elétrica é gerada, proporcional à velocidade das reações nas superfícies dos eletrodos, que por sua vez é proporcional à concentração do gás (ALPHASENSE, 2013a; R. STETTER; LI, 2008; WESTBROEK, 2005).

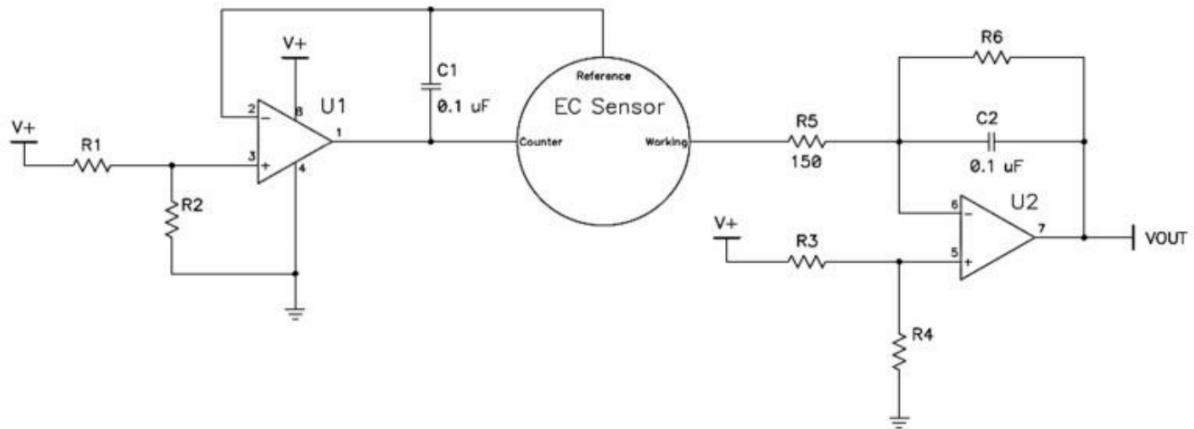
Para assegurar que o sensor se encontra operando dentro de uma região de trabalho desejada, é utilizado o eletrodo de referência. Ele é encarregado de fixar a tensão do eletrodo de trabalho em um valor constante sem que este valor (i.e.: a tensão de operação) seja alterado pela circulação da corrente elétrica, entre os eletrodos contador e de trabalho, que resulta dos processos de transdução. Já o eletrodo contador é deixado com uma carga flutuante que dependerá unicamente das reações eletroquímicas decorrentes da exposição ao gás e do fluxo de elétrons resultante. A corrente de saída do sensor é o resultado da diferença de potencial entre os eletrodos contador e de trabalho (ALPHASENSE, 2013a; BARON; SAFFELL, 2017).

A corrente de saída dos sensores eletroquímicos é muito pequena, geralmente na ordem dos nanoampères (R. STETTER; LI, 2008), fazendo necessária a utilização de uma etapa posterior de condicionamento. A função principal da etapa de condicionamento é transformar o valor da variável elétrica de saída do transdutor em um dado que possa ser

lido por um sistema de aquisição e que represente a quantidade física sendo medida, em um determinado instante de tempo. Comumente, a saída da etapa de condicionamento é um sinal de tensão, ajustado para representar, proporcionalmente, dentro de um determinado intervalo, a variável física de interesse. Esse sinal de tensão pode então ser lido por um conversor analógico-digital. Adicionalmente, podem ser contempladas subetapas intermediária de amplificação, filtragem ou alisamento, ajuste de zero e de ganho, e linearização.

No caso dos sensores eletroquímicos, a configuração mais utilizada no circuito de condicionamento é o potenciostato (ALPHASENSE, 2013a; SPEC SENSORS, 2016b). Este circuito controla o potencial aplicado aos eletrodos de trabalho e de referência e converte a corrente que circula entre os eletrodos de trabalho e contador em um valor de tensão proporcional. Um diagrama simplificado de um sensor eletroquímico e um potenciostato é ilustrado na Figura 68.

Figura 68 – Potenciostato para condicionamento de sensores eletroquímicos.



Fonte: (SPEC SENSORS, 2016b).

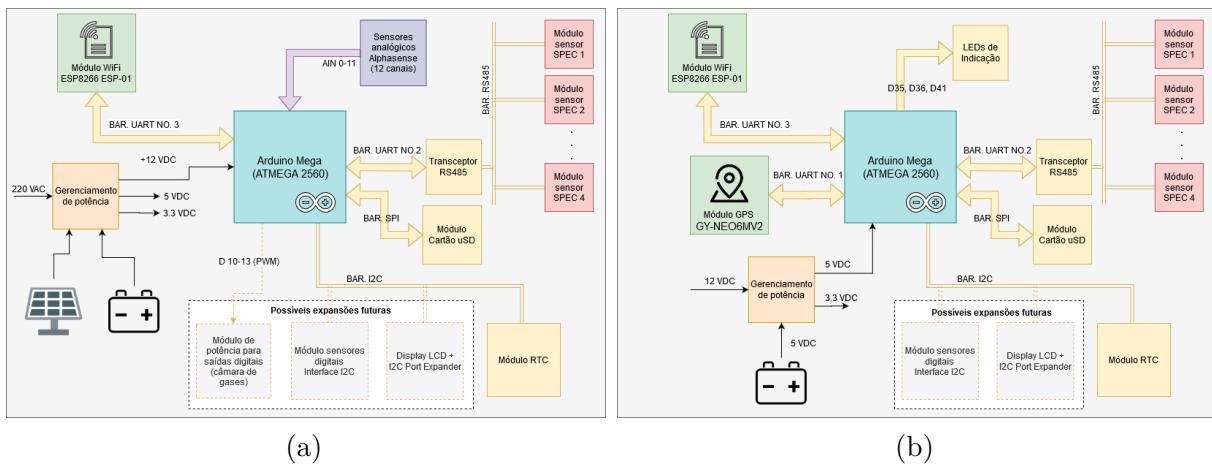
O potencial do eletrodo de referência é estabelecido pelo amplificador U1 da Figura 68, no pino 2. O amplificador U1 também garante a circulação da corrente elétrica pelo eletrodo contador, mantendo constante a tensão de referência. O amplificador U2 fixa o potencial do eletrodo de trabalho no pino 5, e converte em tensão (VOUT) a corrente que circula entre esse eletrodo e o contador. Dessa forma, é possível medir a tensão no pino VOUT como um valor proporcional à concentração do gás em contato com a superfície do sensor (SPEC SENSORS, 2016b).

APÊNDICE B – PROTÓTIPOS DE MONITORES DA QUALIDADE DO AR DESENVOLVIDOS

Foram desenvolvidos dois protótipos para o monitoramento da qualidade do ar ilustrados na Figura 24 do Capítulo 2. Eles foram baseados na plataforma Arduino Mega 2560, que utiliza o microcontrolador ATMega2560 da Microchip. Um deles foi projetado para medição fixa em um local, e o outro para monitoramento de forma móvel. Este último, além de prover a informação temporal associada a cada leitura de concentração de poluente, inclui a localização onde a medição foi tomada. Os dispositivos foram projetados para a medição de poluentes regulados na Resolução CONAMA No. 491/2018, sendo eles: CO , NO_2 , SO_2 e O_3 . Além desses gases, a versão fixa também inclui um sensor de sulfeto de hidrogênio (H_2S).

A Figura 69 mostra diagramas com os módulos de *hardware* que compõem os sistemas de medição fixo e móvel, sem incluir o processo de transporte dos gases. A versão fixa do sistema de monitoramento (Figura 69a) utiliza seis sensores da empresa Alphasense sensíveis aos gases CO , NO_2 , SO_2 , O_3 e H_2S . Além destes, também estão instalados quatro sensores da empresa SPEC, sensíveis aos mesmos poluentes com exceção do H_2S . A conexão entre a plataforma Arduino Mega e os sensores SPEC é realizada pela porta serial UART2 do microcontrolador através de um barramento RS-485. Já a leitura dos sensores da Alphasense é realizada pelas entradas analógicas AI0 – AI11 do microcontrolador.

Figura 69 – Diagrama de blocos dos sistemas fixo (a) e móvel (b)



Fonte: Desenvolvido pelo autor (2023)

A versão móvel (Figura 69b) utiliza apenas sensores da SPEC para a medição de gases. Os modelos SPEC utilizados nesta versão são os mesmos que na versão fixa, e utilizam a mesma configuração para se comunicar de forma serial com o Arduino Mega. Um diferencial desta versão com relação à fixa, além de não utilizar sensores Alphasense, é

Tabela 17 – Especificações técnicas dos ventiladores utilizados no equipamento fixo e móvel

Características	Versão fixa	Versão móvel
Descrição	Ventilador cooler 40mm 12VDC	Ventilador cooler 40mm 5VDC
Marca	GC	GDT
Tamanho	40 x A: 40 x C: 10mm	L: 40 x A: 40 x C: 10mm
Corrente nominal	80 ± 10% mA	140 ± 10% mA
Tensão nominal	12 V	Entre 5 e 7V
Ruído	16 ± 10% dBA	16 ± 10% dBA
Velocidade rotação	5000 ± 10% RPM	7000 ± 20% RPM
Fluxo de ar	4.2 CFM	6.12 CFM
Peso	12 g	14 g
Consumo potência	1.2 W	0.8 W
Vida útil	35000 hr	50000 hr

a inclusão de um módulo GPS para georreferenciar as medições dos poluentes. O módulo utilizado é o GY-NEO6MV2 que se comunica com o microcontrolador através da UART1.

Além dos dispositivos mencionados acima, cada unidade de monitoramento inclui um módulo ESP8266 ESP-01 conectado à porta serial UART3 do Arduino, para comunicação Wi-Fi. Ambas unidades utilizam também um módulo de cartão micro SD para o armazenamento dos dados e um módulo de Relógio de Tempo Real (RTC) para manter a informação de data e hora. Outros periféricos como sensor de pressão, monitor LCD ou atuadores para controlar o transporte dos gases, podem ser adicionados através do barramento I2C em versões futuras.

Os módulos que compõem ambos sistemas de medição funcionam com tensões de alimentação tanto de 3.3 V como 5 V. Para fornecer esses níveis de voltagem foi utilizado um módulo-fonte que utiliza dois reguladores AMS1117. Um dos reguladores fornece uma saída 3.3 V e o outro 5 V. Ambos reguladores conseguem fornecer até 1 A de corrente de saída. O módulo-fonte possui dois canais de entrada de tensão. Um canal possui um conector Jack P4 para tensões entre 9 – 15 V, enquanto o outro possui um conector USB fêmea para fornecer uma tensão de 5 V.

O sistema fixo é alimentado com uma tensão de 12 V, aplicada no conector Jack P2 do módulo-fonte. Os 12 V de tensão podem ser provenientes de uma fonte conectada à rede elétrica, ou de um controlador solar conectado a um painel solar e uma bateria de 12 V. Já o sistema móvel pode ser alimentado por qualquer carregador de bateria portátil com saída em formato USB de 5 V e mínimo 2 A de corrente.

As seções seguintes descrevem cada um dos blocos que compõem os protótipos desenvolvidos.

B.1 TRANSPORTE DE GASES

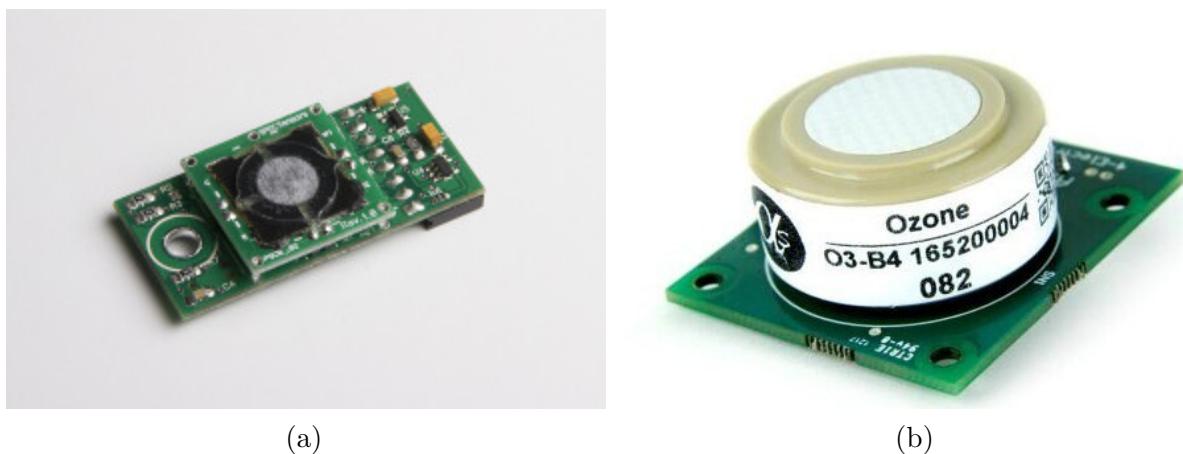
A etapa de transporte de gases é tida como a entrada do sistema. Sua função é capturar amostras do ar no ambiente e direcioná-las para o conjunto de sensores. Nos protótipos desenvolvidos, esta etapa é formada por dois ventiladores de corrente direta e uma câmara de gases. Os ventiladores conduzem o ar desde o ambiente de monitoramento até o interior da câmara. Esta última, por outro lado, consiste em um volume que retém o ar amostrado. No interior dela, os sensores são expostos às porções de ar coletadas para extrair informação de alguns poluentes que podem estar nelas contidos.

As configurações dos ventiladores variam de acordo com a versão do protótipo que os contêm. A versão fixa utiliza ventiladores com tensão nominal de 12 V, e a instalação mecânica deles foi realizada em série para conseguir maior pressão no fluxo do ar. Os ventiladores utilizados na versão móvel, por outro lado, têm uma tensão nominal de 5 V e foram colocados em paralelo. Suas características principais estão dispostas na Tabela 17.

B.2 SENSORIAMENTO

A etapa de sensoriamento consiste em um arranjo de sensores de gases eletroquímicos e os circuitos de condicionamento analógico ou interfaces digitais correspondentes. Os sensores utilizados nos protótipos variam da versão fixa para a móvel, mas de forma geral foram instalados sensores do tipo *screen-printed* fabricados pela SPEC Sensors LLC., e sensores da série B4 da Alphasense Ltd. Os dispositivos de transdução escolhidos são sensíveis aos poluentes regulados na Resolução CONAMA No. 491/2018: CO , NO_2 , SO_2 , O_3 . Além desses gases, foi monitorado também o sulfeto de hidrogênio (H_2S) com um sensor de Alphasense. A modo de ilustração, a Figura 70 mostra os modelos SPEC DGS-CO 968-034 e Alphasense O3-B4, utilizados na medição de CO e O_3 .

Figura 70 – Sensores dos fabricantes a) SPEC e b) Alphasense



A versão fixa dos instrumentos desenvolvidos contém arranjos de sensores das

empresas SPEC e Alphasense. Já o equipamento móvel dispõe apenas de um arranjo de sensores da SPEC Sensors. A seguir são descritas características dos sensores de cada fabricante.

B.2.1 Sensores SPEC

Os sensores da SPEC possuem a configuração característica de três eletrodos (de trabalho, contador e de referência). A sigla SPEC provém do inglês Screen-Printed Electro-Chemical, que é a tecnologia de manufatura utilizada por esse fabricante para produzir seus sensores. Essa tecnologia possibilita fabricar sensores de gases eletroquímicos de alta performance em um encapsulamento fino e de um custo menor que os encapsulamentos mais volumosos, utilizados tradicionalmente para fabricar sensores eletroquímicos (SPEC SENSORS, 2016b). A Figura 70a mostra o sensor SPEC DGS-CO 968-034, utilizado na medição de monóxido de carbono, em sua placa de condicionamento. A Tabela 18 resume as principais características dos sensores que foram utilizados dessa empresa.

B.2.2 Sensores Alphasense

Os sensores da série B4, da Alphasense, utilizam, além dos três eletrodos característicos do princípio de medição eletroquímico, um quarto eletrodo chamado de Eletrodo Auxiliar. Sua função é gerar uma corrente com um valor de intensidade muito próximo ao valor da corrente de fundo do zero (zero background current). Dessa forma é possível compensar a saída dos sensores do efeito desta corrente de zero ou de linha base. A Tabela 19 resume as principais características dos sensores que foram utilizados desse

Tabela 18 – Especificações técnicas dos sensores SPEC

<i>Características</i>	<i>CO</i>	<i>NO₂</i>	<i>SO₂</i>	<i>O₃</i>
Modelo	DGS-CO 968-034	DGS-NO2 968-043	DGS-SO2 968-038	DGS-O3 968-042
Intervalo de medição	0 - 1000 ppm	0 - 5 ppm	0 - 20 ppm	0 - 5 ppm
Resolução	100 ppb	20 ppb	50 ppb	20 ppb
Tensão nominal	3.3 V	3.3 V	3.3 V	3.3 V
Consumo de potência	12 mW	14 mW	12 mW	14 mW
Tempo de resposta*	< 30 s	< 30 s	< 30 s	< 30 s
Temperatura de operação	-20 – 40 °C	-20 – 40 °C	-20 – 40 °C	-20 – 40 °C
Umidade relativa de operação	15 – 95 %	15 – 95 %	15 – 95 %	15 – 95 %

Tabela 19 – Especificações técnicas dos sensores Alphasense

<i>Características</i>	<i>CO</i>	<i>NO₂</i>	<i>SO₂</i>	<i>O₃</i>	<i>H₂S</i>
Modelo	CO-B4	NO ₂ -B43F	SO ₂ -B4	OX-B431	H ₂ S-B4
Intervalo de medição (ppm)	0 - 1000	0 - 20	0 - 100	0 - 20	0 - 100
Resolução (ppb)	4	15	5	15	1
Tempo de resposta (s)*	< 30	< 80	< 60	< 80	< 60
Temperatura de operação (°C)	-30 - 50	-30 - 40	-30 - 50	-30 - 40	-30 - 50
Umidade relativa de operação (%)	15 - 90	15 - 85	15 - 90	15 - 85	15 - 90

fabricante.

B.3 CONDICIONAMENTO

A configuração mais utilizada nos circuitos de condicionamento dos sensores eletroquímicos é o potencióstato. Este circuito controla o potencial aplicado ao eletrodo de trabalho e converte a corrente desse eletrodo em um valor de tensão.

A empresa Alphasense disponibiliza para os sensores da série B4 uma placa de condicionamento chamada de *Individual Sensor Board* (ISB). Esta placa transforma o sinal de corrente de saída do sensor em um sinal de tensão proporcional ao valor de concentração do gás. A SPEC Sensors, por sua vez, disponibiliza uma placa com um microcontrolador dedicado, que condiciona a saída do transdutor e entrega o dado de concentração mediante uma interface digital serial.

B.3.1 Interface de condicionamento dos sensores Alphasense: A Placa de Sensoriamento Individual (ISB)

As placas ISB da Alphasense possuem circuitos de potencióstato compatíveis com a família de sensores B4, de quatro eletrodos. Nesses circuitos, tanto o eletrodo de trabalho quanto o auxiliar possuem etapas de amplificação equivalentes. As tensões de saída destes dois eletrodos são disponibilizados em um conector Molex de 6 vias junto com os canais para a alimentação da placa. A tensão de alimentação das placas ISB pode ser entre 3.5 e 6.4 VDC; nos protótipos desenvolvidos foi utilizada uma tensão de 5 VDC.

O diagrama ilustrado na Figura 71a apresenta as conexões realizadas entre a plataforma Arduino e os sensores da Alphasense através das placas ISB. Percebe-se que cada conjunto composto por um sensor e seu respectivo circuito de condicionamento, ocupa duas entradas analógicas do microcontrolador; uma entrada para o eletrodo auxiliar (AE) e outra para o eletrodo de trabalho (WE). No total foram utilizadas os canais analógicos A0 – A11.

B.3.2 Interface de condicionamento dos sensores SPEC

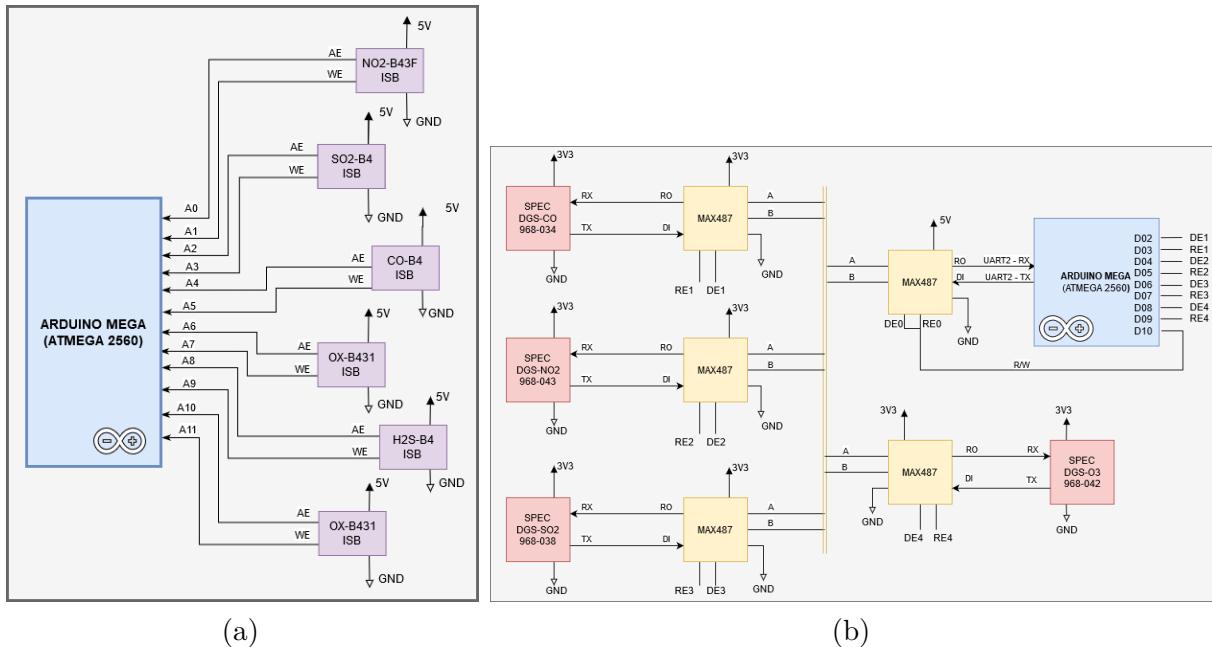
Assim como os sensores da Alphasense, os sensores da empresa SPEC também utilizam um circuito de condicionamento de potenciómetro. Na mesma placa de condicionamento, a SPEC tem incorporado um microcontrolador dedicado e sensores de temperatura e umidade. O microcontrolador converte o valor de tensão de saída do potenciómetro em um valor digital de concentração de gás, e realiza uma compensação, por software, dos efeitos da temperatura e a umidade na medição.

O kit de condicionamento SPEC funciona como uma camada de abstração no que diz respeito ao tratamento e condicionamento das informações, garantindo uma fácil integração com os sistemas de monitoramento. O dispositivo disponibiliza as informações de data e hora, o valor de concentração em ppm/ppb e as leituras de temperatura e umidade através de uma interface serial UART. De igual modo, podem ser realizadas operações como calibração, ajuste de zero e span, configuração dos sensores, e seleção de modo de operação de baixo consumo de energia, através de uma biblioteca com comandos pré definidos (SPEC SENSORS, 2017).

A Figura 71b apresenta um diagrama da conexão do arranjo de sensores da SPEC à plataforma Arduino Mega. Os sensores e o microcontrolador ATMega2560 são acoplados a um barramento RS-485 mediante o transceptor MAX487. Esse transceptor provê uma interface entre os meios de comunicação serial UART e RS-485. O barramento RS-485 consiste basicamente em dois fios A e B que fornecem o meio físico para a transmissão de níveis de tensão que representam os dados seriais enviados pelos diferentes dispositivos. O nível do sinal transmitido através do barramento é determinado pela tensão diferencial entre os conectores A e B, independentemente da voltagem de alimentação dos dispositivos conectados. Como mostra a figura, os transceptores dos sensores são alimentados com uma tensão de 3.3 VDC, enquanto o transceptor do Arduino é alimentado pelo mesmo sinal de 5 VDC que o microcontrolador.

É possível conectar múltiplos dispositivos a um mesmo barramento RS-485 (máximo até 128), sendo necessária a ação de um controlador que determine quem acessa o meio físico a cada instante, para evitar colisões. O microcontrolador ATMega2560 realiza essa função através das saídas digitais D02 – D10. Esses sinais digitais controlam o estado das entradas RE_i e DE_i de cada MAX487, a fim de habilitar/desabilitar cada transceptor para operações de escrita/leitura.

Figura 71 – Interface entre os sensores e o microcontrolador Arduino. a) Alphasense, b) SPEC



Fonte: Desenvolvido pelo autor (2023)

B.4 MICROCONTROLADOR

A etapa de processamento engloba todas as funcionalidades de controle, temporização, geolocalização, aquisição dos dados dos sensores, comunicação e armazenamento dos dados. Todas essas funções são gerenciadas pelo microcontrolador ATMega2560 da Microchip, embarcado na plataforma Arduino Mega 2560. Mais detalhes sobre o firmware desenvolvido para o controle da etapa de processamento são abordados no Apêndice D. A continuação descrevem-se cada um dos módulos que compõem esta etapa.

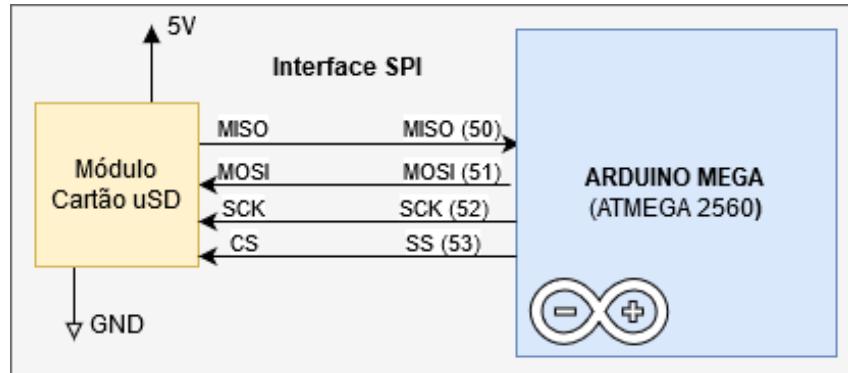
B.4.1 Armazenamento dos dados

Para o armazenamento dos dados foi utilizado um módulo para fazer leitura e escrita diretamente em um cartão micro SD. A comunicação é feita por meio de uma interface SPI, conforme se mostra na Figura 72. O nível de sinal é de 3.3V, mas o módulo possui divisores de tensão nos seus pinos que possibilitam uma ligação direta com placas que trabalham com 5 V, como o Arduino. O módulo é alimentado com uma tensão de 5 V, e suporta cartões Micro SD e Micro SDHC de alta velocidade.

B.4.2 Controle de data e hora e geolocalização

Para manter o controle da data e hora do sistema fixo, e assim acrescentar informação temporal às leituras de gases, foi utilizado o módulo de Relógio de Tempo

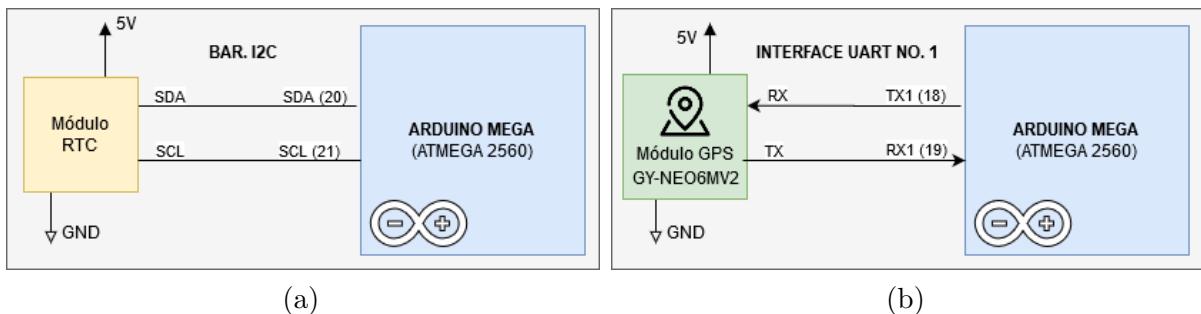
Figura 72 – Interface entre o módulo cartão micro SD e o microcontrolador



Fonte: Desenvolvido pelo autor (2023)

Real (RTC) DS1307. O DS1307 é um relógio/calendário de baixo consumo de potência que utiliza um barramento I^2C bidirecional para a transferência de dados desde (e para) o microcontrolador. O relógio/calendário provê informação de segundos, minutos, horas, dia, mês e ano, incluindo ajuste automático de ano bissexto e de meses com menos de 31 dias. O DS1307 é alimentado por uma tensão de 5 V, e também possui um circuito que detecta falhas de energia e automaticamente aciona a alimentação através de uma bateria. Quando isso sucede, o relógio/calendário mantém a contagem do tempo em um modo de baixo consumo (consumo de corrente menor que 500 nA), estendendo o tempo de vida útil da bateria. A Figura 73a mostra como é realizada sua conexão ao microcontrolador no sistema desenvolvido.

Figura 73 – Interface entre o microcontrolador e os módulos a) RTC e b) GPS



Fonte: Desenvolvido pelo autor (2023)

Na versão móvel, ambos os controles da data e hora e geolocalização são realizados por um mesmo dispositivo GPS, o módulo NEO6MV2. O NEO6M é um receptor GPS de baixo consumo de potência e pequenas dimensões que o tornam uma opção interessante para dispositivos móveis. O módulo possui uma antena integrada, com precisão de aproximadamente 5 metros, e tecnologia para supressão de congestionamentos na comunicação e interferências. A conexão entre o módulo e a plataforma Arduino é realizada através de um barramento serial UART a uma taxa de transferência padrão de 9600 bauds (Figura 73b). Ele pode ser alimentado com uma tensão de 3.3 ou 5 V e seu

consumo de corrente em pleno funcionamento chega a 45 mA. Seus pinos de entrada são compatíveis com níveis de tensão TTL e suportam tensões tanto de 5 como de 3.3 V, independentemente da tensão de alimentação.

B.4.3 Comunicação Wi-Fi

Para a comunicação Wi-Fi é utilizado o módulo ESP-01. Esse módulo incorpora o sistema integrado em um único chip (SoC, System on Chip) ESP8266EX, da empresa Espressif, e uma antena embarcada com ganho de potência de 3dBi, garantindo um alcance de até 90 metros em espaços abertos. O SoC ESP8266EX integra um processador de 32 bits, o Tensilica L106, que implementa os protocolos TCP/IP e o 802.11 b/g/n WLAN MAC. Ele possui como vantagens um baixo consumo de energia atrelado a uma velocidade de clock de 80 MHz. Sua memória RAM, disponível em aplicações em que o sistema está configurado como estação é de aproximadamente 36 kB. O módulo ESP-01 disponibiliza, para armazenar o programa de usuário, uma memória FLASH de 1MB externa que pode ser acessada por um barramento SPI. O módulo disponibiliza quatro portas digitais que são utilizadas principalmente para programar a FLASH de usuário e uma porta serial UART.

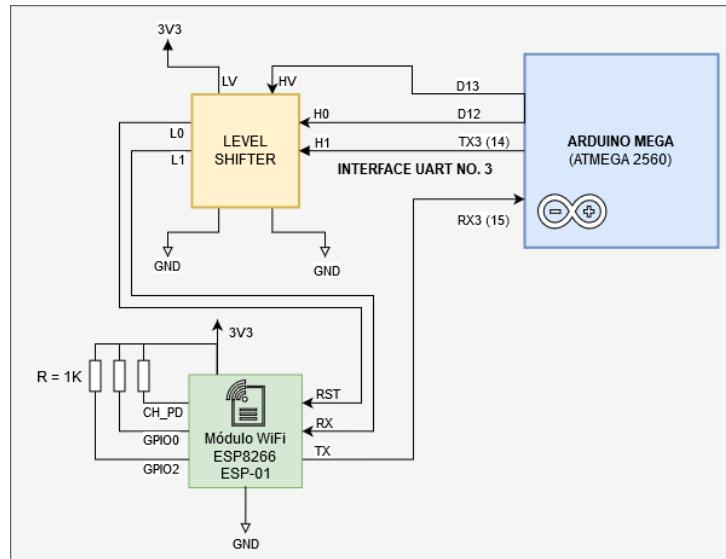


Figura 74 – Interface entre o microcontrolador e o módulo de comunicação Wi-Fi

Fonte: Desenvolvido pelo autor

A Figura 74 apresenta as conexões realizadas entre o ESP-01 e a plataforma Arduino. O módulo opera com uma tensão de 3.3 V, por esse motivo é utilizado um circuito intermediário, um elevador de nível (level shifter) para converter os níveis de tensão de 5 V para 3.3 V, e vice-versa. O pino CH_PD corresponde ao chip enable do ESP-01 e deve ser conectado a um resistor de *pull-up* de 1 kΩ, assim como as entradas GPIO0 e GPIO2. Essas entradas são utilizadas para configurar o ESP8266 em modo gravação (para gravar o programa de usuário) ou modo estação. A figura mostra a configuração do modo

estaçao, com ambas entradas conectadas à 3.3 V por meio de resistores de *pull-up* de 1 kΩ. O pino de entrada RST tem como função reiniciar o módulo. Como esse pino é “ativo baixo”, cada vez que uma tensão de 0 V for aplicada nessa porta o módulo será reiniciado. No circuito desenvolvido, o Arduino pode reiniciar o ESP8266 através da saída digital D12. Já a saída D13 do microcontrolador Arduino é encarregada de manter uma tensão de referência de 5 V no elevador de nível para possibilitar a conversão dos níveis de tensão.

B.5 MONTAGEM DO PROTÓTIPO FIXO

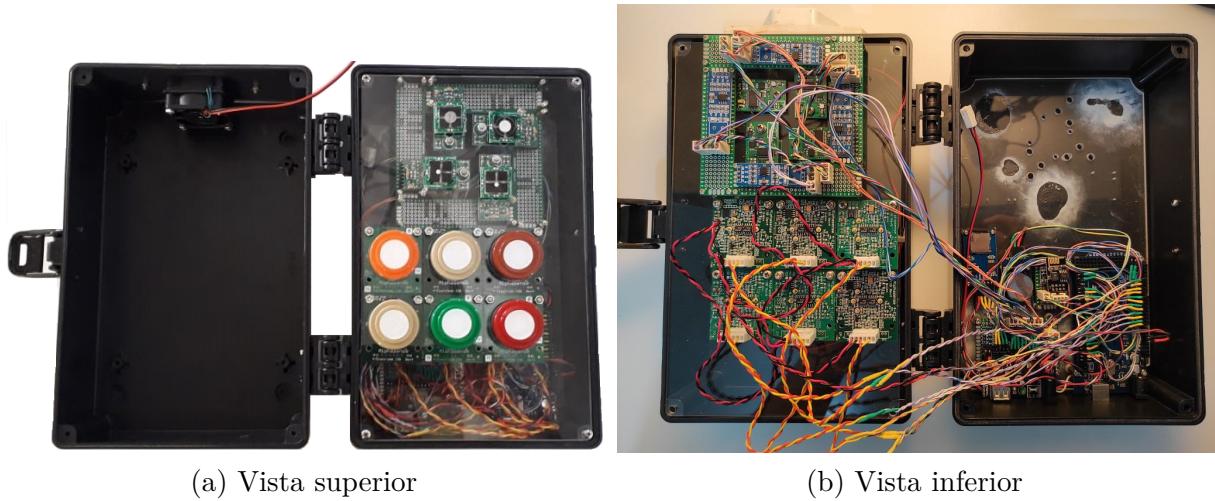
A seguir descreve-se brevemente a montagem e interligação dos elementos de hardware que compõem o protótipo de medição fixa. A Figura 75 mostra o protótipo de monitor fixo instalado em campo. O quadro externo é a caixa ambiental modelo Atlantic 352 00 da Cemar & Legrand com nível de proteção IP66.

Figura 75 – Instalação em campo do protótipo fixo



As Figuras 76a e 76b mostram o módulo de sensoriamento, que é a parte fundamental de todo o sistema. Nele são contidos todos os elementos que compõem o sistema e que foram descritos anteriormente.

Figura 76 – Vista interior do protótipo fixo



Um sistema de transporte de gases, composto por duas ventoinhas de 12VDC, coleta amostras do ar ambiente para dentro da câmara. A entrada é composta por uma flange de 50 mm de diâmetro (que serve para acoplar a câmara no restante do sistema de transporte de gases) e um filtro de tecido. As dimensões das ventoinhas são 40x40mm, e foram fixadas com quatro parafusos M2x30mm com porca e arruela. Dentro do volume da câmara, as superfícies dos sensores de gás interagem com os componentes gasosos e produzem um sinal de resposta proporcional à concentração do gás. A Tabela 20 resume os sensores e placas de condicionamento que foram utilizados nessa versão do equipamento.

Uma placa de acrílico foi utilizada para fixar os sensores de maneira correta e isolá-los do hardware do fluxo de ar. As conexões elétricas para levar os sinais de saída dos sensores até o Arduino foram feitas com fios de seção 0,2mm², soldando “headers” nas pontas e isolando-as corretamente com duto termoretrátil. As Figuras 77a e 77b ilustram respectivamente diagramas de conexão da alimentação elétrica e dos eletrodos dos sensores ao Arduino.

Figura 77 – Diagrama de conexões do conjunto de sensores Alphasense

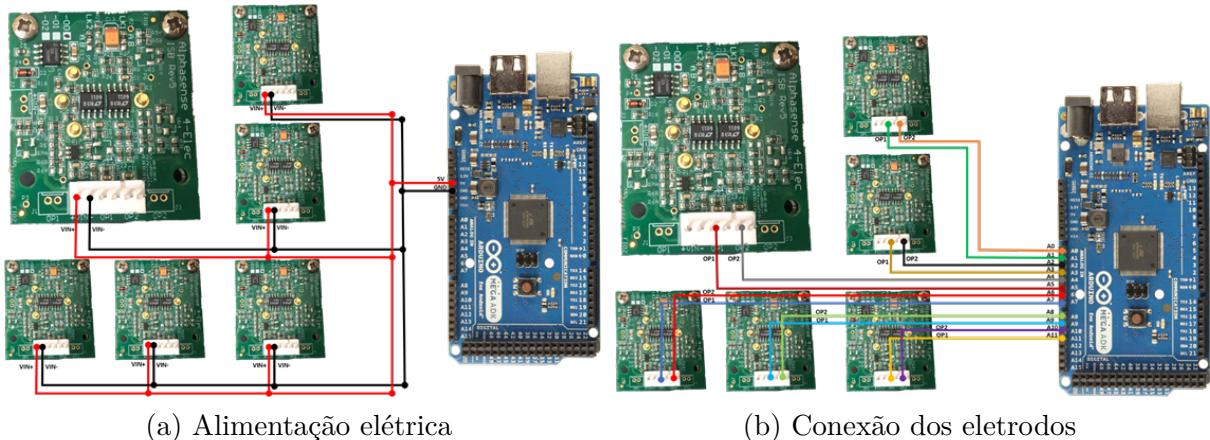


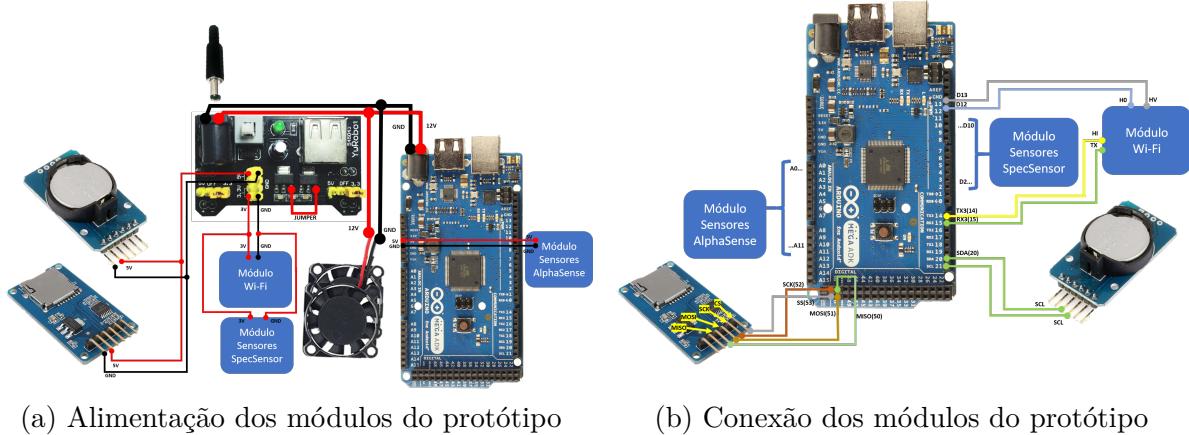
Tabela 20 – Lista de sensores utilizados no protótipo fixo

Qtd.	Ítem	Descrição	Fabricante
1	CO-B4	Sensor de <i>CO</i>	Alphasense
1	H2S-B4	Sensor de <i>H₂S</i>	Alphasense
1	SO2-B4	Sensor de <i>SO₂</i>	Alphasense
1	NO-B4	Sensor de <i>NO</i>	Alphasense
1	NO2-B43F	Sensor de <i>NO₂</i>	Alphasense
1	OX-B431	Sensor de <i>O₃</i>	Alphasense
1	NH3-B1	Sensor de <i>NH₃</i>	Alphasense
3	CO/H2S/SO2 4-electrodes ISB	Placa de condicionamento para sensores da série B4 que medem <i>CO</i> , <i>H₂S</i> e <i>SO₂</i>	Alphasense
1	NO 4-electrodes ISB	Placa de condicionamento para sensores da série B4 que medem <i>NO</i>	Alphasense
1	NO2/O3 4-electrodes ISB	Placa de condicionamento para sensores da série B4 que medem <i>NO₂</i> e <i>O₃</i>	Alphasense
1	NH3 4-electrodes ISB	Placa de condicionamento para sensores da série B4 que medem <i>NH₃</i>	Alphasense
1	DGS-O3-968- 042_9-6-17	Sensor de <i>O₃</i> para IoT	SPEC Sensors
1	DGS-SO2-968- 038	Sensor de <i>SO₂</i> para IoT	SPEC Sensors
1	DGS-NO2-968- 043-9-6-17	Sensor de <i>NO₂</i> para IoT	SPEC Sensors
1	DGS-CO-968- 034	Sensor de <i>CO</i> para IoT	SPEC Sensors

A fixação dos sensores *SPEC* na placa de acrílico mencionada anteriormente foi feita através de placas de prototipagem confeccionadas artesanalmente. Nas placas foram soldados “headers” fêmeas encima dos quais os sensores foram montados. Nas placas também foram instalados os transceptores MAX487 que criam o barramento RS-485 para a conexão serial com o microcontrolador da placa Arduino. As placas são conectadas ao barramento através de fios e conectores do tipo MOLEX. As placas de prototipagem com os sensores foram fixadas diretamente à placa de acrílico com espaçadores M2.

Após a montagem do conjunto de sensores e prefixação dos componentes eletrônicos da câmara de medição, foi realizada a ligação elétrica de alimentação e comunicação de todos os componentes envolvidos no sistema. O diagrama de alimentação é mostrado na Figura 78a. Vale salientar que a alimentação da placa Arduino foi feita diretamente com 12V com fios soldados no conector P2 de entrada. Já a conexão do restante dos módulos com a placa Arduino é ilustrada na Figura 78b.

Figura 78 – Diagrama de conexões do conjunto de sensores Alphasense



B.6 MONTAGEM DO PROTÓTIPO MÓVEL

O equipamento mede poluentes da legislação ambiental brasileira (**BRASIL.MINISTERIO**) que são: CO , NO_2 , SO_2 , O_3 e H_2S . Para isso utiliza um conjunto de quatro sensores do fabricante SPEC Sensor que contemplam a medição desse poluentes. O controle da estapa de monitoramento, armazenamento e envio de dados é baseado na plataforma Arduino Mega 2560, que utiliza o microcontrolador ATMega2560 da Microchip. Para operar com sucesso, o sistema inclui: módulo Wi-Fi, módulo de cartão SD, módulo GPS e indicadores LED operacionais.

APÊNDICE C – A PLACA CLEAN ARDUINO MEGA

A continuação são descritos os principais módulos que compõem a placa CLEAN Arduino Mega e os protótipos desenvolvidos. A Tabela 21 mostra os principais componentes de hardware utilizados sem considerar os sensores.

C.1 MÓDULO DE SENSORIAMENTO

C.1.1 Sensores

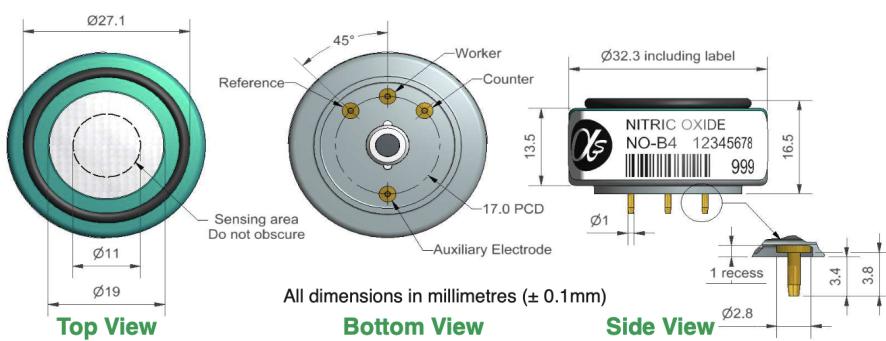
Nos sistemas desenvolvidos foram utilizados sensores do fabricante *Alphasense* de princípio eletroquímico para medição de gases e contadores ópticos de partículas para medição de material particulado.

C.1.1.1 Sensores eletroquímicos Alphasense.

Alphasense fabrica sensores eletroquímicos amperométricos. Especificamente, os sensores da série B4 foram selecionados para os monitores desenvolvidos, já que são indicados pelo fabricante para a medição de baixas concentrações de gases. Estes sensores incorporam um quarto eletrodo, denominado eletrodo auxiliar, que compensa os efeitos da temperatura e da umidade relativa nas leituras dos sensores (BARON; SAFFELL, 2017). A Figura 79 ilustra um sensor Alphasense da série B4, suas dimensões e disposição dos eletrodos. Para mais informações sobre os efeitos das variáveis ambientais nas respostas dos sensores, confira a nota de aplicação AAN 110 da *Alphasense*. Para informações adicionais sobre os sensores da série B4 da *Alphasense*, como especificações elétricas, dimensões e pinagem, consulte as fichas técnicas dos modelos de sensores listados na Tabela 1.

Sensores eletroquímicos amperométricos produzem uma corrente de saída que é proporcional à concentração do gás. Para ler este sinal elétrico com um sistema de aquisição de dados, a corrente de saída deve ser transformada em um sinal de tensão. Para isso, o

Figura 79 – Sensor de Monôxido de Nitrogênio Alphasense da série B4



Fonte: (ALPHASENSE, 2019b).

Tabela 21 – Principais componentes utilizados nos dispositivos CLEAN

Item	Descrição	Modelo e fabricante
Arduino Mega 2560	Placa microcontroladora baseada no microcontrolador Microchip ATmega2560	Arduino MEGA 2560 Rev3, de Arduino
DS3231 RTC	Relógio em tempo real (RTC) I2C com oscilador de cristal compensado por temperatura integrado	DS3231, by Maxim Integrated
Soquete micro SD	Soquete TF / micro SD tipo PUSH-PUSH	KLS1-TF-007, da KLS Electronic
Buffer CI 74XX125	Portas de buffer de barramento quádruplas com saídas de 3 estados para buffer de pinos de cartão micro SD	74HC125, da Texas Instruments
Cartão MicroSD	Cartão Micro SD Classe 10 de 16 GB	microSDHC SanDisk Ultra, da SanDisk
Módulo GPS	Módulo GPS NEO-6M c/ antena	GY-GPS6MV2, por u-blox
Módulo GPRS	Shield Arduino - GSM GPRS SIM900 com antena Quad Band	SIM900, da SIMCom
Módulo Wi-Fi	Módulo serial Wi-Fi ESP-01 ESP8266	ESP8266, da Expressif
Sensor BMP280	Sensor digital de temperatura e pressão BMP280 I2C	BMP280, da Bosch Sensortec
Sensor SHT20	Sensor de umidade e temperatura SHT20 I2C	SHT20, por Sensirion

circuito mais utilizado é o potenciómetro. *Alphasense* e *SPEC* fornecem placas de circuito potenciómetro para acoplar facilmente seus sensores a um sistema de monitoramento.

C.1.1.2 Interface de condicionamento de sensores SPEC

Os sensores digitais para IoT fornecidos pela *SPEC* são compostos por um transdutor eletroquímico montado em uma placa com um circuito potenciómetro, que converte a saída do sensor (corrente elétrica) a tensão. Os sensores incorporam também um microcontrolador e um sensor de temperatura e umidade relativa. O microcontrolador adquire o sinal de tensão do potenciómetro como valores de concentração de gás e realiza uma compensação em software para reduzir os efeitos da temperatura e a umidade relativa nas leituras do sensor. Os valores de concentração, temperatura e umidade relativa são transmitidos através de uma interface UART seguindo um protocolo serial

definido pelo fabricante. Esta placa de condicionamento atua como uma camada de abstração para condicionamento de sinal que permite a fácil integração dos sensores a qualquer sistema de monitoramento. Para obter mais informações sobre os sensores SPEC, como especificações elétricas, dimensões, pinagem e protocolo serial, verifique as fichas técnicas dos sensores (Tabela 1) e do kit de desenvolvimento do sensor de gás digital 968-045 (REF).

C.1.1.3 Interface de condicionamento de sensores Alphasense

A Alphasense fornece placas de sensores individuais para seus sensores de gás de 4 eletrodos da série B4 (REF). Essas placas incorporam circuitos potencióstatos equivalentes para os eletrodos de trabalho e auxiliar. As saídas de cada canal do potencióstato foram conectadas às entradas analógicas do microcontrolador Arduino MEGA, conforme mostrado na Figura A.3. Os sinais AE e WE representam os sinais correspondentes ao eletrodo auxiliar e de trabalho respectivamente. Seis sensores foram utilizados no protótipo, sendo utilizadas assim doze entradas analógicas do microcontrolador (A0 – A11). Cada módulo ISB foi alimentado com uma tensão de 5 V. Para obter mais detalhes sobre a montagem e conexão dos sensores Alphasense, consulte o Guia de montagem de sensores Alphasense.

C.1.1.4 Contador ótico de partículas Alphasense para medição de material particulado

O sensor da Alphasense modelo OPC-N3 é um contador ótico de partículas para medição de PM10, PM2.5, e PM1. Possui uma saída digital através de um barramento SPI, podendo ser acoplado a um microcontrolador para leitura das suas variáveis. O sensor pode enviar as informações de concentração na forma de histogramas ou em valores absolutos em $\mu\text{g}/\text{m}^3$. Além dos valores de concentração o sensor também envia leituras de temperatura e umidade relativa.

C.2 O MICROCONTROLADOR

O microcontrolador Arduino MEGA 2560 coordena as tarefas associadas à aquisição e armazenamento de dados, temporização, geolocalização e comunicação. O firmware para este protótipo está disponível no repositório de firmware. Para obter detalhes sobre a estrutura do firmware e bibliotecas de firmware, consulte a documentação do firmware.

C.2.1 Armazenamento dos dados

Para armazenamento dos dados foi utilizado um módulo micro SD conectado ao microcontrolador através de uma Interface Periférica Serial (SPI). O cartão micro SD funciona com 3.3 V, mas o módulo inclui buffers e um regulador de tensão que permite conexão direta ao Arduino SPI e fonte de alimentação de 5 V, conforme mostra a figura 6.

C.2.2 Relógio de tempo real

Para monitorar a data e a hora de forma contínua foi utilizado o módulo DS1307 Real-Time Clock (RTC). Este módulo é um relógio/calendário de baixo consumo de energia que fornece informações sobre segundos, minutos, horas, dia, data, mês e ano (REF). A data do final do mês é ajustada automaticamente para meses com menos de 31 dias, incluindo correções para anos bissextos. O DS1307 possui um circuito sensor de energia integrado que detecta falhas de energia e alterna automaticamente para a fonte de backup por meio de uma bateria. A operação de cronometragem continua enquanto a peça opera no modo de baixo consumo de energia da fonte de reserva. O módulo se conecta ao Arduino MEGA através da interface I2C e é alimentado com 5V, conforme mostra a figura 7.

C.2.3 Comunicação Wi-Fi

Para a comunicação Wi-Fi foi utilizado o módulo ESP-01 (Figura 8). Este módulo incorpora o microcontrolador ESP8266 junto com uma antena embarcada com ganho de potência de 3dBi e alcance de até 90 m. O ESP8266 é um System on Chip (SoC), fabricado pela Espressif Systems, que integra o microprocessador Tensilica L106 de 32 bits e implementa os protocolos TCP/IP e 802.11 b/g/n WLAN MAC (REF). O ESP-01 também incorpora uma memória flash de 512 kB para programação, que é acessível ao ESP8266 via SPI. Ele também possui oito pinos que são utilizados para alimentação, conexão à porta serial do ESP8266 e conexão aos quatro GPIOs do ESP8266, conforme mostrado na Figura 8. Para mais detalhes sobre a pinagem do ESP-01 e como programar e conectar este módulo para o Arduino MEGA, consulte o Guia de programação do módulo ESP-01. Uma descrição do firmware que desenvolvemos para o microcontrolador ESP8266 pode ser encontrada em The ESP8266 Firmware.

O módulo ESP-01 fornece a conexão a uma rede Wi-Fi para o Arduino MEGA. Conforme mostrado na Figura 9, um circuito de mudança de nível é necessário para fazer a interface com os pinos do Arduino como resultado das diferentes tensões de operação das placas. A comunicação entre os microcontroladores ATMega2560 e ESP8266 é implementada através de uma interface UART (UART3 na placa Arduino), seguindo um protocolo de comunicação que é descrito detalhadamente no Guia de Programação do Módulo ESP-01. O Arduino atua como mestre do ESP8266, cuja única iniciativa é estabelecer conexão com a Internet. Uma vez estabelecida a conexão, o Arduino pode enviar comandos para criar posts HTTP, obter o horário da internet ou obter as coordenadas de geolocalização do Google; para obter mais detalhes, consulte o Guia de programação do módulo ESP-01. O microcontrolador Arduino também pode redefinir o ESP8266 através do pino D12 GPIO.

APÊNDICE D – O FIRMWARE CLEAN

O *firmware* dos dispositivos foi desenvolvido para o microcontrolador Microchip ATMega2560 embarcado em uma plataforma Arduino Mega. O código foi implementado na linguagem de programação C/C++ utilizando o *framework* de Arduino, disponível na IDE PlatformIO para o editor de código Microsoft Visual Studio (VSCode).

Para a programação de todas as funcionalidades do *firmware*, o código foi estruturado em um conjunto de classes. Essa estrutura foi concebida visando seu reaproveitamento em outros microcontroladores suportados no Framework Arduino, como o ESP8266 da Espressif, e também para facilitar a revisão e manutenção do código. As classes desenvolvidas para o projeto estão distribuídas em três pacotes de bibliotecas: o pacote *IoT*, o pacote *Data* e o pacote *Hardware Interfaces*.

O pacote *IoT* encapsula os processos associados à conexão na rede Wi-Fi, comunicação com o servidor e envio de dados pelo protocolo *HTTP*. Já o pacote *Data* engloba todas as funcionalidades relacionadas à preparação dos dados dos sensores para seu armazenamento e transmissão. Este pacote permite abstrair as informações de concentração adquiridos pelos sensores de gases, de detalhes específicos sobre o funcionamento e operação do hardware destes sensores. Ele atua como uma camada intermediária entre as tarefas de aquisição, e as de armazenamento e transmissão dos dados. Por último, o pacote *Hardware Interfaces* agrupa as classes e estruturas utilizadas para interfacear todo o hardware periférico ao microcontrolador utilizado, como sensores, módulos de temporização, módulos de geolocalização e módulos de armazenamento.

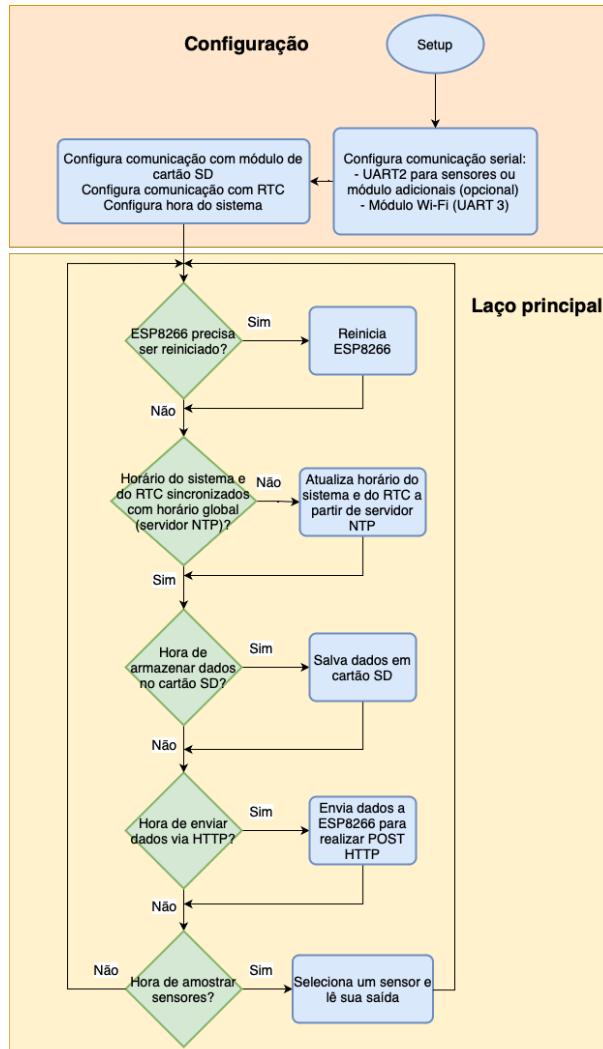
D.1 CÓDIGO CLEAN ARDUINO MEGA

O código consiste em duas partes: uma para configuração (*setup*) e outra para o laço (*loop*) de execução principal. A versão atual do código possui quatro funcionalidades principais que estão de acordo com o hardware do monitor, sendo elas:

1. Amostragem, ou leitura das saídas dos sensores
2. Armazenamento das informações de concentração de gás em um cartão SD
3. Envio dos dados dos sensores para o servidor Renovar através do microcontrolador ESP8266
4. Leitura das coordenadas geográficas do local onde ocorre cada leitura

A Figura 80 mostra um fluxograma do código programado para o ATMega2560. Como acontece com todo programa do Arduino Framework, o código é executado em duas funções principais: `setup()` e `loop()`. Na versão atual do firmware, o `setup` prepara a comunicação entre os módulos externos (i.e.: RTC, Wi-Fi, cartão de memória e sensores

Figura 80 – Fluxograma do firmware programado para o microcontrolador Arduino MEGA



Fonte: Desenvolvido pelo autor (2023)

seriais) e o microcontrolador. A função `loop()` verifica se o ESP8266 não responde há algum tempo, e, se for o caso, é enviado um sinal RESET para o ESP8266. O resto da função está dividido em quatro seções que são executadas periodicamente e controlam as funcionalidades mencionadas.

O firmware inclui mais uma funcionalidade que está separada do fluxo principal do programa: a função `serialEvent3()` que trata os eventos de interrupção da porta serial UART3. Ela é executada cada vez que novos dados são recebidos no buffer dessa interface de hardware que estabelece a comunicação entre o Arduino Mega e o microcontrolador ESP8266. A seguir, são descritas as diferentes seções do código.

D.1.1 Identificação do dispositivo e seus sensores

Uma parte crucial do firmware é a identificação do dispositivo e dos sensores conectados a ele. Essas identificações deverão corresponder às que tenham sido previamente configuradas no servidor da aplicação Web Renovar, já que serão utilizadas pela aplicação backend para atualizar o banco de dados. A definição do dispositivo é feita na seguinte linha:

```
unsigned long Device::id = <THE NUMBER OF YOUR DEVICE>;
```

Depois disso, devem ser definidos os IDs dos sensores, de forma que cada um deles represente uma variável na aplicação Renovar. O código da Lista D.1.1 exemplifica a definição dos identificadores de 6 sensores previamente registrados na aplicação Renovar.

Listing D.1 – Definição dos identificadores dos sensores de um dispositivo

```
enum iotId_e {
    // The ID of the CO gas concentration read from a sensor
    CO_ID      = 156,
    // The ID of the NO2 gas concentration read from a sensor
    NO2_ID     = 157,
    // The ID of the O3 gas concentration read from a sensor
    O3_ID      = 158,
    // The ID of the SO2 gas concentration read from a sensor
    SO2_ID     = 159,
    // The ID of the temperature variable
    TEMP_ID    = 160,
    // The ID of the relative humidity variable
    RHUM_ID    = 161
};
```

D.1.2 Configuração: a função setup()

A função `setup()` prepara a comunicação entre os módulos externos e o microcontrolador. O código desta função é mostrado na Lista D.1.2. Primeiramente o programa configura as portas seriais que serão utilizadas para comunicação com os sensores seriais utilizando uma interface RS-458 (RS485_2) e ESP8266 (Serial3). Cada porta serial é inicializada a uma taxa de transmissão previamente definida no código, conforme será descrito posteriormente. A porta serial UART0 (Serial) é usada para depuração. O RS485_2 implementa uma interface serial ao barramento RS485 que conecta os sensores ao microcontrolador. A função também reinicia o microcontrolador ESP8266 através do objeto `espIoT`, inicializa as interfaces para o cartão SD e módulos RTC e define a hora do sistema. A continuação são descritos os objetos, constantes e funções usados nesta parte do código.

Listing D.2 – Definição dos identificadores dos sensores de um dispositivo

```

void setup ()
{
    Serial . begin (9600);
#ifndef FIXED_DEVICE
    Serial1 . begin (GPSBaud);
#endif
    Serial3 . begin (9600UL);

    pinMode (13 , OUTPUT);
    digitalWrite (13 , HIGH);

    espIoT . restart ();

    SD . begin (CHIPSEL_PIN);

    Rtc . Begin ();
    if (!Rtc . GetIsRunning ())
    {
        Rtc . SetIsRunning (true);
    }
#ifndef DS3231
    Rtc . Enable32kHzPin (false);
    Rtc . SetSquareWavePin (DS3231SquareWavePin_ModeNone);
#endif
    time_t now = Rtc . GetDateTime () . Epoch32Time ();
    setSyncProvider (sync_time);
    setSyncInterval (5*SECS_PER_MIN);
    TimeDriver :: config (TIMEZONE_SEC);
    SHT20 . initSHT20 ();
    BMP280 . begin (BMP280_ADDRESS_ALT, BMP280_CHIPID);
    Alpha_OPC . begin ();
#ifndef FIXED_DEVICE
    GPSDriver :: set_coordinates (true,
                                DeviceFixedLocation :: LATITUDE,
                                DeviceFixedLocation :: LONGITUDE,
                                DeviceFixedLocation :: ALTITUDE);
#endif
    mLastTime = millis ();
}

```

```

mLastTimeGPS = mLastTime;
mLastTimeSD = mLastTime;
mLastTimeHTTP = mLastTime;
}

```

D.1.2.1 Serial, Serial1, Serial3

Esses objetos, declarados no *framework* Arduino, representam as portas UART do microcontrolador. Os objetos são inicializados pela função `begin()`, que recebe a taxa de transmissão da comunicação serial. O objeto `Serial` representa a porta serial `UART0` que é usada para depuração do programa. Já os objetos `Serial1` e `Serial3`, que representam as portas seriais `UART1` e `UART3` do Arduino Mega, são utilizadas para comunicação com o módulo GPS e o microcontrolador `ESP8266` respectivamente.

A variável `GPSBaud` é uma constante que define a velocidade em *bits* por segundo da comunicação serial entre o microcontrolador `ATMega` do Arduino e o módulo GPS. O valor predeterminado é 9600 baúdios, definido na biblioteca `serial-geo-interface`, mas pode ser redefinido segundo a aplicação.

D.1.2.2 espIoT

Este é um objeto da classe `ESPSerialInterface` definido na biblioteca `serial-internet-interface`. Este objeto controla a comunicação com o `ESP8266` conectado ao `UART3` do Arduino. A função `restart()` envia um sinal de *RESET* para o `ESP8266`. O objeto `espIoT` é definido no código da seguinte forma: `ESPSerialInterface espIoT(&Serial3);`

D.1.2.3 SD

Este é um objeto do `SDClass` declarado no núcleo do Arduino para interfacear módulos de cartão SD. Ele é inicializado com um método `begin()` que recebe o pino digital que se conecta ao pino CS do módulo. O pino digital utilizado para a versão atual do hardware e firmware é definido no arquivo `hardstorage.h` da seguinte forma: `#define CHIPSEL_PIN 53`

D.1.2.4 Rtc

Este é um objeto da classe `RtcDS3231` definida na biblioteca `Rtc` de Makuna. O objeto `Rtc` é declarado da seguinte forma:

```

#define I2C Wire
RtcDS3231<TwoWire> Rtc(I2C);

```

A instância `Rtc` é inicializada com a função `begin()` e posteriormente o código verifica se o módulo está funcionando por meio de uma chamada ao método `GetIsRunning()`. Caso não esteja rodando, o código chama o método `SetIsRunning(true)`. Caso o módulo RTC DS3231 tenha sido configurado incorretamente, o código também redefine seu status através do seguinte código:

```
#ifdef DS3231
    Rtc.Enable32kHzPin(false);
    Rtc.SetSquareWavePin(DS3231SquareWavePin_ModeNone);
#endif
```

A chamada ao método `Rtc.GetDateTime().Epoch32Time()` da classe `RtcDS3231` obtém a hora atual do módulo RTC no formato UNIX. Já os métodos `setSyncProvider(getExternalTime getTimeFunction)` e `setSyncInterval(time_t interval)` são funções da biblioteca `Time` que permitem a sincronização automática da hora do sistema com uma fonte de relógio determinada. Neste caso, a fonte utilizada para sincronização é o módulo RTC. A função `setSyncProvider()` recebe um ponteiro para uma função que retorna a data e hora atual como uma variável de tipo `time_t`. Neste caso, a função que é passada como ponteiro é `sync_time()`, declarada anteriormente no código conforme mostrado abaixo:

```
RtcDS3231<TwoWire> Rtc(I2C);
RTCDS3231Interface My_RTCInterface(&Rtc);
time_t sync_time() {
    return RTCDriver<RtcDS3231<TwoWire>>::
        sync_time_from_RTC(&My_RTCInterface);
}
```

A função `setSyncInterval()` recebe o período de sincronização da hora do sistema, que neste caso foi definido como 5 segundos.

D.1.3 Interrupção Serial3

O código da função que trata a interrupção do UART3 é mostrado na Lista D.1.3. Cada vez que os dados estiverem disponíveis no *buffer* de entrada da porta serial, o objeto `espIoT` irá analisar a cadeia de caracteres recebida.

Listing D.3 – Código para tratamento da interrupção da porta serial UART3

```
void serialEvent3() {
    if (Serial3.available()) {
        String buffer = Serial3.readStringUntil(';' );
        Serial3.flush();
        espIoT.parse_esp_string(buffer);
```

```

    }
}
```

D.1.4 Laço principal do programa: a função loop()

A sequência de instruções do programa da placa CLEAN Arduino Mega é executado dentro de um laço infinito definido na função `loop()` do *Framework* Arduino. Esta função é responsável por tratar quatro funcionalidades que foram mencionadas anteriormente, i.e.: amostragem, armazenamento, envio de dados e geolocalização. A Lista D.1.4 mostra o código da função `loop()`.

Listing D.4 – Código do laço de execução do programa

```

void loop()
{
    static bool sd_ok = false;

    espIoT.watch_dog();

    if (( ! TimeDriver :: _already_up_to_date ()) ) espIoT.request_time ();
    if (( ! My_RTCInterface .is_up_to_date () ))
        RTCDriver<RtcDS3231<TwoWire>>::update_rtc(&My_RTCInterface , now ());

    // /*
    if (( millis () - mLastTime) >= SAMPLE_ITERATION_PERIOD_MS)
    {
        mLastTime = millis ();
        static uint8_t index = 0;
        Vars [index]→smooth( sensors [index]→read ());
        index = (index >= numSensors - 1) ? 0 : index + 1;
    }
    // */

    // /*
    if (( millis () - mLasteuSD) >= uSD_TIME_MSEC)
    {
        mLasteuSD = millis ();
        static uint8_t data_index_uSD = 0;
        Vars [data_index_uSD]→sense(&data );
```

```

char* filename = (char*)malloc(strlen_P(
    filenames [data_index_uSD])+1);
strcpy_P(filename , filenames [data_index_uSD]);
if(open_file(filename))
    sd_ok = save_to_file(&data , filename );
else SD.begin(CHIPSEL_PIN);
free( filename );

data_index_uSD = (data_index_uSD >= numSensors-1) ? 0 :
    data_index_uSD + 1;
}

// */

if(( millis () - mLastTimeHTTP) >= HTTP_TIME_MSEC)
{
    mLastTimeHTTP = millis ();
    static uint8_t data_index_iot = 0;

    Vars [data_index_iot]->sense(&data );
    readings [0] = &data ;
    if (!espIoT.send_http_post(&data)) print_debug("Couldn't post!");
    data_index_iot = (data_index_iot >= numSensors-1) ? 0 :
        data_index_iot + 1;
}

// */

if(( millis () - mLastTimeGPS) >= MSECS_GPSOUTDATE)
{
    static uint8_t gps_tries = 0;
    print_debug(" [MAIN] GPS");
    mLastTimeGPS = millis ();
    if (!gps.read_gps(MSECS_GPSOUTDATE/2))
    {
        if (gps_tries++ > 7)
        {
            GPSDriver::set_coordinates(true,
                DeviceDefaultLocation::LATITUDE,

```

```

        DeviceDefaultLocation ::LONGITUDE,
        DeviceDefaultLocation ::ALTITUDE);
    gps_tries = 8;
}
}
else {
    gps_tries = 0;
}
//
}
}

```

As quatro funcionalidades principais que o código executa periodicamente são controladas pelas variáveis `mLastTime`, `mLastTimeGPS`, `mLastTimeuSD` e `mLastTimeHTTP`, que armazenam a marca o instante de tempo em que cada funcionalidade é executada. As constantes `uSD_TIME_MSEC`, `HTTP_TIME_SEC`, `SAMPLE_ITERATION_PERIOD_MS` e `MSECS_GPSOUTDATE` representam os períodos em que cada funcionalidade deve ser executada, conforme está resumido na Tabela 22. Em cada ciclo do laço, o objeto `espIoT` chama ao método `watch_dog()` para verificar se existe alguma solicitação enviada ao ESP8266 cujo tempo de espera tenha expirado. Caso isso aconteça, o ESP8266 será reiniciado. A função `loop()` também verifica se o microcontrolador atualizou seu horário a partir de um servidor de data e hora da Internet, caso contrário, uma solicitação é enviada ao ESP8266 para retornar o horário atual da Internet.

Tabela 22 – Constantes e variáveis utilizadas para controlar a execução de cada funcionalidade no firmware

Funcionalidade	Período	Constante definida no código	Variável de controle
Amostragem	6 segundos	<code>SAMPLE_ITERATION_PERIOD_MS</code>	<code>mLastTime</code>
Armazenamento	60 segundos	<code>uSD_TIME_MSEC</code>	<code>mLastTimeuSD</code>
Envio de dados	60 segundos	<code>HTTP_TIME_MSEC</code>	<code>mLastTimeHTTP</code>
Geolocalização	70 segundos	<code>MSECS_GPSOUTDATE</code>	<code>mLastTimeGPS</code>

Fonte: Desenvolvido pelo autor (2023)

O código também verifica se o módulo RTC foi atualizado com o horário da internet. Caso contrário, ele chama o método `update_rtc()` da classe `RTCDriver`. Esta classe é um *template* para controlar as funcionalidades relacionadas a um módulo RTC genérico, como atualizar seu horário, por exemplo. O método `update_rtc()` recebe um ponteiro para uma instância da classe `RTCInterface`, que cria uma interface para o hardware de

qualquer módulo RTC. `RTCInterface` é uma classe abstrata, portanto, para instanciar essa interface em um módulo RTC, uma classe concreta deve ser herdada dele. No presente caso, esta instância foi implementada no objeto `My_RTCInterface`, declarado previamente no código conforme mostrado abaixo.

```
class RTCDS3231Interface : public RTCInterface<RtcDS3231<TwoWire>>
{
public:
    RTCDS3231Interface( RtcDS3231<TwoWire>* rtc ) :
        RTCInterface<RtcDS3231<TwoWire>>( rtc ) {}

    virtual void set_time( time_t t ) {
        RtcDateTime dt;
        dt.InitWithEpoch32Time( t );
        _rtc->SetDateTime( dt );
    }

    virtual time_t get_time() {
        return _rtc->GetDateTime().Epoch32Time();
    }
};

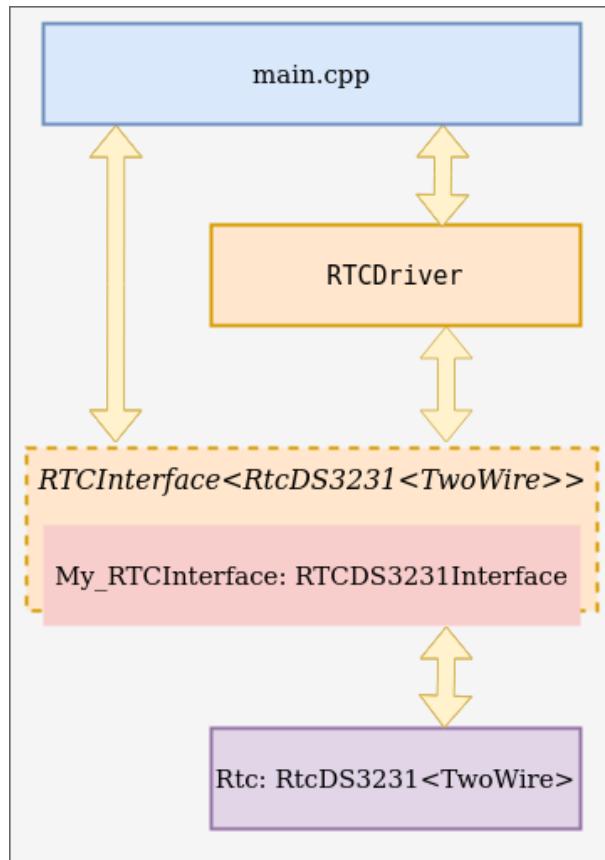
#define I2C Wire
RtcDS3231<TwoWire> Rtc( I2C );
RTCDS3231Interface My_RTCInterface(&Rtc);
```

Como pode ser observado, a classe `RTCDS3231Interface` herda de `RTCInterface`. Quando o objeto daquela classe é declarado, ele recebe em seu construtor uma referência a um objeto `rtc`, que neste caso representa o próprio módulo DS3231. Resumindo, o objeto `rtc` representa o módulo DS3231; o objeto `My_RTCInterface` implementa a interface entre o módulo RTC e o código principal; e a classe `RTCDriver` controla as funcionalidades do módulo dentro do código. Essa relação é representada no diagrama da Figura 81.

D.1.4.1 Leitura dos sensores

A seção do código que lê os sensores de gás é mostrada na Lista D.1.4.1. O código primeiro verifica se o tempo de leitura dos sensores já passou e atualiza a variável `mLastTime`. Esta seção do código basicamente itera entre as listas `Vars` e `sensors` para obter a leitura de cada sensor e atribuir a variável correspondente. A lista `sensors` contém todos os sensores conectados na placa CLEAN, já a lista `Vars` contém as variáveis que cada sensor representa; um representa a camada de *hardware* enquanto o outro representa

Figura 81 – Módulos e interfaces usados para controle e interface do RTC



Fonte: Desenvolvido pelo autor (2023)

uma camada de mais alto nível dos dados, as variáveis físicas. Dessa forma a informação de mais baixo nível contida no sinal de saída dos sensores é transportada para camadas de dados superiores que possibilita seu armazenamento e transmissão remota. Na Lista D.1.4.1 ilustra-se o código que implementa cada uma das listas.

Listing D.5 – Código para leitura dos sensores

```

if (( millis () - mLastTime) >= SAMPLE_ITERATION_PERIOD_MS)
{
    mLastTime = millis ();
    static uint8_t index = 0;
    Vars [index]→smooth( sensors [index]→read ());
    index = (index >= numSensors - 1) ? 0 : index + 1;
}
  
```

Listing D.6 – Declaração das listas de sensores e variáveis

```

Sensor* sensors [numSensors] =
{
    ( SensorInterface<Adafruit_BMP280>*&)IntTempSensor ,
  
```

```

( SensorInterface<Adafruit_BMP280>*)&IntPresSensor ,
( SensorInterface<AlphaSenseCompensator>)
    (new AlphaSenseCompensatorSensor(&Alpha_COComp)) ,
( SensorInterface<AlphaSenseCompensator>)
    (new AlphaSenseCompensatorSensor(&Alpha_NO2Comp)) ,
( SensorInterface<AlphaSenseCompensator>)
    (new AlphaSenseCompensatorSensor(&Alpha_SO2_1_Comp)) ,
( SensorInterface<AlphaSenseCompensator>)
    (new AlphaSenseCompensatorSensor(&Alpha_O3_1_Comp)) ,
( SensorInterface<AlphaSenseCompensator>)
    (new AlphaSenseCompensatorSensor(&Alpha_O3_2_Comp)) ,
( SensorInterface<AlphaSenseCompensator>)
    (new AlphaSenseCompensatorSensor(&Alpha_SO2_2_Comp)) ,
( SensorInterface<DFRobot_SHT20>*)&ExtTempSensor ,
( SensorInterface<DFRobot_SHT20>*)&ExtHumSensor ,
( SensorInterface<AlphasenseOPC>)*(new
    AlphaSenseOPCPM10Sensor(&Alpha_OPc)) ,
( SensorInterface<AlphasenseOPC>)*(new
    AlphaSenseOPCPM2_5Sensor(&Alpha_OPc)) ,
( SensorInterface<AlphasenseOPC>)*(new
    AlphaSenseOPCPM1Sensor(&Alpha_OPc)) ,
( SensorInterface<AlphasenseOPC>)*(new
    AlphaSenseOPCTempSensor(&Alpha_OPc)) ,
( SensorInterface<AlphasenseOPC>)*(new
    AlphaSenseOPCHumSensor(&Alpha_OPc)) )
};

Variable* Vars [ numSensors ] =
{
    new Temperature(TEMPERATURE_ID, SI_TEMP_Celsius, BUFFER_SIZE) ,
    new Pressure(PRESSURE_ID, SI_PRES_Pascal, BUFFER_SIZE) ,
    new GasConcentration(ALPHA_CO_ID, SI_CONC_ppb, BUFFER_SIZE) ,
    new GasConcentration(ALPHA_NO2_ID, SI_CONC_ppb, BUFFER_SIZE) ,
    new GasConcentration(ALPHA_SO2_1_ID, SI_CONC_ppb, BUFFER_SIZE) ,
    new GasConcentration(ALPHA_OX_1_ID, SI_CONC_ppb, BUFFER_SIZE) ,
    new GasConcentration(ALPHA_OX_2_ID, SI_CONC_ppb, BUFFER_SIZE) ,
    new GasConcentration(ALPHA_SO2_2_ID, SI_CONC_ppb, BUFFER_SIZE) ,
    new Temperature(EXT_TEMPERATURE_ID, SI_TEMP_Celsius, BUFFER_SIZE) ,
    new Humidity(EXT_HUMIDITY_ID, SI_HUMD_Relative, BUFFER_SIZE) ,

```

```

new GasConcentration(PM10_ID, SI_CONC_ug, BUFFER_SIZE),
new GasConcentration(PM25_ID, SI_CONC_ug, BUFFER_SIZE),
new GasConcentration(PM01_ID, SI_CONC_ug, BUFFER_SIZE),
new Temperature(OPC_TEMPERATURE_ID, SI_TEMP_Celsius, BUFFER_SIZE),
new Humidity(OPC_HUMIDITY_ID, SI_HUMD_Relative, BUFFER_SIZE)
};

```

D.1.4.2 Armazenamento dos dados

A seção do código que armazena os dados em um cartão SD é mostrada na Lista . O código primeiro verifica se o tempo de armazenamento dos dados já passou e atualiza a variável `mLastTimeuSD`. Esta seção do código transfere os dados de cada variável para um objeto `data`, que é do tipo `SensorData` e que é usado para armazenar as informações da variável. O método utilizado para transferir as informações de `Variable` para `SensorData` é a função `sense()`, que recebe um ponteiro para `SensorData`. Este objeto `data` é definido anteriormente no código conforme mostrado abaixo. Depois que os dados forem transferidos para a instância de `SensorData`, eles são armazenados em um arquivo no cartão SD.

Listing D.7 – Sequência de armazenamento de dados

```

SensorData data;

// ...

void loop()
{
    // ...
    if(( millis () - mLastTimeuSD ) >= uSD_TIME_MSEC)
    {
        mLastTimeuSD = millis ();
        static uint8_t data_index_uSD = 0;
        Vars [data_index_uSD]→sense (&data );

        char* filename = (char*)malloc (strlen_P (
            filenAMES [data_index_uSD])+1);
        strcpy_P (filename , filenAMES [data_index_uSD]);
        if(open_file (filename ))
            sd_ok = save_to_file (&data , filename );
        else SD.begin (CHIPSEL_PIN);
        free (filename );
    }
}

```

```

    data_index_uSD = (data_index_uSD >= numSensors-1) ? 0 :
                    data_index_uSD + 1;
}
// ...
}

```

D.1.4.3 Envio de dados via protocolo *HTTP*

A seção do código que envia os dados ao ESP8266 para postagem em um servidor HTTP é mostrada abaixo. Assim como as demais seções do código, esta seção primeiro verifica se o tempo de envio dos dados já passou e atualiza a variável `mLastTimeHTTP`. Depois disso, ele utiliza o último objeto `SensorData` armazenado para enviar as informações adquiridas para a variável correspondente. O objeto `espIoT` enviará uma cadeia de caracteres com um objeto JSON contendo as informações a serem postadas pelo ESP8266. O método `send_http_post()` recebe um ponteiro para um `DataContainer`. Como a classe `SensorData` herda de `DataContainer`, cada item nos dados pode ser convertido em um ponteiro desse tipo

```

if (( millis () - mLastTimeHTTP ) >= HTTP_TIME_MSEC)
{
    mLastTimeHTTP = millis ();
    static uint8_t data_index_iot = 0;

    Vars [ data_index_iot ]->sense (&data );
    readings [ 0 ] = &data ;
    if (!espIoT . send _http _post (&data )) print _debug ( " Couldn ' t post ! " );
    data_index_iot = ( data_index_iot >= numSensors -1) ? 0 :
                    data_index_iot + 1;
}

```

D.1.4.4 Geolocalização

Por fim, a continuação mostra a seção de código que atualiza as informações de geolocalização do módulo GPS. Assim como nas outras seções do código, esta seção primeiro verifica se passou o tempo para atualizar as informações do GPS e atualiza a variável `mLastTimeGPS`. Para ler o módulo GPS, o objeto `gps` invoca o método `readGPS()`. Este método toma como parâmetro o tempo máximo que o Arduino deve aguardar uma resposta do módulo GPS, neste caso `MSECS_GPSOUTDATE/2`. O objeto `gps` é uma instância da classe `TinyGPSSerialInterface` que está previamente definida no código conforme mostrado. O construtor deste objeto recebe uma referência à porta serial utilizada para comunicação com o módulo, neste caso `Serial1`. Em caso de falha na comunicação

com o módulo GPS, depois da sétima tentativa, é setado um valor padrão previamente configurado.

```
TinyGPSSerialInterface gps(&Serial1);  
  
// ...  
  
void loop()  
{  
    // ...  
  
    if ((millis() - mLastTimeGPS) >= MSECS_GPSOUTDATE)  
    {  
        static uint8_t gps_tries = 0;  
        print_debug("[MAIN] GPS");  
        mLastTimeGPS = millis();  
        if (!gps.read_gps(MSECS_GPSOUTDATE/2))  
        {  
            if (gps_tries++ > 7) {  
                GPSDriver::set_coordinates(  
                    true,  
                    DeviceDefaultLocation::LATITUDE,  
                    DeviceDefaultLocation::LONGITUDE,  
                    DeviceDefaultLocation::ALTITUDE);  
                gps_tries = 8;  
            }  
        }  
        else {  
            gps_tries = 0;  
        }  
    }  
}
```

APÊNDICE E – O FIRMWARE DO MICROCONTROLADOR ESP8266

O *firmware* do microcontrolador ESP8266 foi desenvolvido na linguagem de programação C/C++ utilizando o *Framework* Arduino. O código foi programado na IDE PlatformIO para o editor de código Microsoft Visual Studio (VSCode). A versão atual do firmware do ESP8266 possui três funcionalidades principais, que são:

1. Prover conexão à Internet a um microcontrolador principal, que atua como "mestre", via uma rede *Wi-Fi*
2. Envio dos dados coletados pelo microcontrolador "mestre" para o servidor web Renovar mediante *POST HTTP*
3. Obter a data e hora atuais de um servidor NTP

A Figura 82 mostra um fluxograma do código programado para o microcontrolador ESP8266. Em primeiro lugar, o programa estabelece uma conexão à Internet através de uma rede *Wi-Fi*. Uma vez estabelecida a conexão, o programa se mantém escutando a conexão serial com o microcontrolador "mestre". Para cada solicitação recebida do "mestre", o ESP8266 executa a operação associada à solicitação e envia seu resultado de volta para o "mestre". As mensagens trocadas entre o ESP8266 e o outro microcontrolador são cadeias de caracteres no formato JSON.

Na versão atual do *firmware*, o "mestre" pode realizar dois tipos de solicitações. Pode solicitar o envio de uma requisição para a API Renovar na forma de POST HTTP com leituras dos sensores ou pode solicitar a data e hora de um servidor NTP.

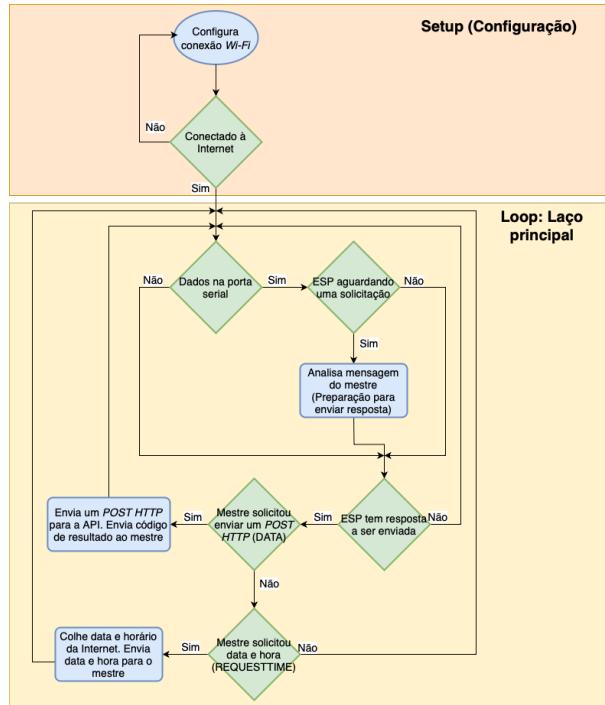
E.1 CONFIGURAÇÃO E CONEXÃO *WI-FI*

Conforme mostra o fluxograma da 82, as primeiras ações que o programa executa são aquelas relacionadas ao estabelecimento da conexão à Internet via rede *Wi-Fi* e a comunicação serial com o microcontrolador "mestre". Isso é realizado na função **setup()** conforme mostrado no código da Lista E.1.

Listing E.1 – Definição dos identificadores dos sensores de um dispositivo

```
void setup()
{
    Serial.begin(9600UL);
    Serial.print(F("+STARTESP;"));
    setup_wifi_connection<NUM_WIFIS>(wifiCreds);
    espHTTP.set_available(true);
    espSerial.set_status(WT_REQUEST);
    Serial.print(F("+ESPREADY;"));
}
```

Figura 82 – Fluxograma do firmware programado para o microcontrolador ESP8266



Fonte: Desenvolvido pelo autor (2023)

}

Primeiramente o programa inicializa a porta serial do ESP8266 a uma taxa de transmissão de 9600 bauds e imprime a mensagem '+STARTESP', indicando ao mestre que o programa foi inicializado. A função `setup_wifi_connection()` estabelece a conexão em uma rede *Wi-Fi* previamente armazenada na variável `wifiCreds`. Por fim, uma vez estabelecida a conexão, o objeto `espSerial` que armazena o estado da comunicação serial é setado para `WT_REQUEST` e a mensagem '+ESPREADY' é impressa, indicando que o ESP8266 está aguardando por uma solicitação do microcontrolador "mestre". A partir desse ponto o objeto `espHTTP` executa as requisições à API Renovar através do protocolo *HTTP*. A continuação são descritos os objetos, constantes e funções usados nesta parte do código.

E.1.1 NUM_WIFIS

Esta constante define a quantidade de redes *Wi-Fi* com as quais o ESP8266 tentará estabelecer uma conexão. Esta constante deve ser declarada antes do objeto `wifiCreds` e antes de invocar a função `setup_wifi_connection()`.

E.1.2 WiFiCredentials wifiCreds[]

Esta variável é uma lista de objetos do tipo `WiFiCredentials`. Esta matriz armazena o SSID da rede, a senha e o nome de usuário (este último apenas em redes *WPA3 ENTERPRISE*). A lista `wifiCreds` deve ser declarada antes da função `setup()`,

conforme se mostra no código da Lista E.1.2. O tamanho da lista vai depender do valor da constante `NUM_WIFIS`.

Listing E.2 – Definição dos identificadores dos sensores de um dispositivo

```
/* Credenciais de uma rede empresarial WPA3*/
const WiFiCredentials CRED_1( "ssid1" , "senha1" , ENTERPRISE, "nomedeu"

/* Credenciais das redes pessoais WPA3*/
const WiFiCredentials CRED_2( "ssid2" , "senha2" , PERSONAL);
const WiFiCredentials CRED_3( "ssid3" , "senha3" , PERSONAL);

/* Declara o array wifiCreds */
const WiFiCredentials wifiCreds [NUM_WIFIS] = { CRED_1, CRED_2, CRED_3}
```

E.1.3 `setup_wifi_connection<NUM_WIFIS>(wifiCreds)`

Esta função estabelece uma conexão a uma rede *Wi-Fi*. É definido como um *template* que recebe a quantidade de redes *Wi-Fi* armazenadas em `wifiCreds`.

E.1.4 `espHTTP`

Esta variável é um objeto da classe `HTTPHandler`, definida no arquivo `esp-iot.h`, cujo objetivo é encapsular as funcionalidades relacionadas às operações *HTTP*.

E.1.5 `espSerial`

Esta variável é um objeto da classe `ESPSerialHandler`, definida no arquivo `esp-serial-iot.h`. O objetivo deste objeto é encapsular as funcionalidades relacionadas à comunicação serial entre o ESP8266 e o microcontrolador mestre. Dependendo do estado deste objeto, o ESP8266 pode ler uma mensagem serial do mestre, ou executar uma determinada operação e enviar seu resultado de volta ao mestre. Para a versão atual do firmware, foram implementados dois estados:

`WT_REQUEST`: Indica que o ESP8266 não recebeu nenhuma solicitação do mestre e está aguardando até receber uma nova. `WT_RESPONSE`: Indica que o ESP8266 recebeu alguma nova solicitação do mestre e está realizando as operações para enviar uma resposta.

E.1.6 `Serial`

Este é um objeto do framework Arduino para controlar a comunicação serial.

E.2 O LAÇO PRINCIPAL

O laço principal do programa é executado dentro da função `loop()`, conforme mostrado no código abaixo. Esta função é responsável por monitorar a porta serial do ESP8266 e atender às solicitações do "mestre", conforme já ilustrado na Figura 82.

Listing E.3 – Laço principal do programa

```
void loop()
{
    static CommandTypes _cmdType = ERROR;

    if(Serial.available())
    {
        String serial_Str = Serial.readStringUntil(';');

        if(espSerial.get_status() == WT_REQUEST) _cmdType = espSerial.p
            Serial.flush();

        if(espSerial.get_status() == WT_RESPONSE)
        {
            espSerial.set_status(WT_REQUEST);
            switch (_cmdType)
            {
                case DATA:
                {
                    static uint8_t _numberOfPostTries = 0;
                    #define MAX_NUM_TRIES 3
                    int code = espHTTP.post(HOST, PORT, URL, espSerial.get_da

                    if(code <= 0 && WiFi.status() != WL_CONNECTED)
                    {
                        setup_wifi_connection<NUM_WIFIS>(wifiCreds);
                        if(++_numberOfPostTries >= MAX_NUM_TRIES-1) ESP.restart();
                    }
                    else espSerial.send_http_code(code, Serial);
                    break;
                }

                case REQUESTTIME:
                {
                    static uint8_t _numberOfTries = 0;
```

```

#define MAX_NUM_TRIES 3
time_t t = get_time(TIMEZONE_SEC, DAYLIGHTOFFSET_SEC);
if (!t && WiFi.status() != WL_CONNECTED)
{
    setup_wifi_connection<NUM_WIFIS>(wifiCreds);
    if (++_numberOfTries >= MAX_NUM_TRIES - 1) ESP.restart();
}
else espSerial.send_time(t, Serial);
break;
}

default:
{
    break;
}
}
}
}

```

Verificação das solicitações do "mestre"

Sempre que houver dados na porta serial e o estado da comunicação serial (armazenado no objeto `espSerial`) for “aguardando solicitação” (`WT_REQUEST`), o mesmo objeto `espSerial` analisará a cadeia de caracteres recebida do “mestre” para determinar o tipo de solicitação que recebeu. Isso é feito dentro do primeiro `if` da função `loop()`, conforme mostrado no código da Lista E.2.

Listing E.4 – Código para verificar as solicitações do mestre

```

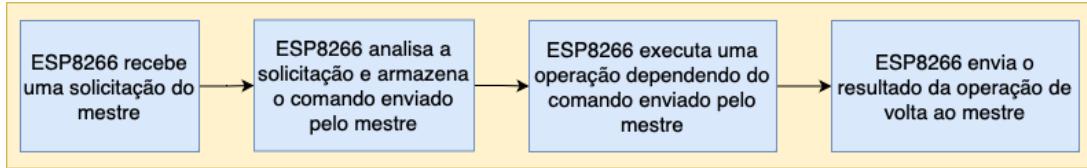
if (Serial.available())
{
    String serial_Str = Serial.readStringUntil(';' );
    if (espSerial.get_status() == WT_REQUEST) _cmdType = espSerial.parse(
        Serial.flush());
}

```

Uma vez analisada a mensagem do “mestre”, `espSerial` definirá seu estado para “aguardando resposta” (`WT_RESPONSE`), indicando que uma operação está em execução e uma resposta deve ser enviada de volta para o mestre (isso é feito internamente na função `parse`). Se o status de `espSerial` for `WT_RESPONSE`, o programa realizará uma operação `switch` para determinar qual operação o microcontrolador deve executar a seguir (fluxograma da Figura 82). A seleção da operação dependerá do tipo de solicitação enviado

pelo mestre. O tipo de solicitação é armazenado na variável `_cmdType` como resultado da operação `parse()` executada por `espSerial`. A Figura 83 mostra esse processo.

Figura 83 – Processo de atendimento a uma solicitação do mestre



Fonte: Desenvolvido pelo autor (2023)

Na versão atual do firmware, o ESP8266 aceita dois tipos de solicitações ou comandos do mestre: (1) uma solicitação para enviar uma requisição à API Renovar via um *POST HTTP* ou (2) uma solicitação de data e horário. A estrutura de uma mensagem de solicitação é mostrada no JSON da Lista E.2. Depois que o objeto `espSerial` processa a mensagem do mestre, ele armazena o tipo de solicitação na variável `_cmdType`. Esta variável é uma enumeração do tipo `CommandTypes` e, para a versão atual, pode conter os valores `ERROR`, `DATA` e `REQUESTTIME`, veja a Tabela 23.

Listing E.5 – O formato da string JSON trocada entre o ESP8266 e o mestre

```
{
    'type': // The type of the request. Could be 1 (DATA) or 2 (REQUESTTIME)
    'body': // The body of the request. Only used when the master is sending a POST request
}
```

Tabela 23 – Tipos de solicitações representadas no tipo `CommandTypes`

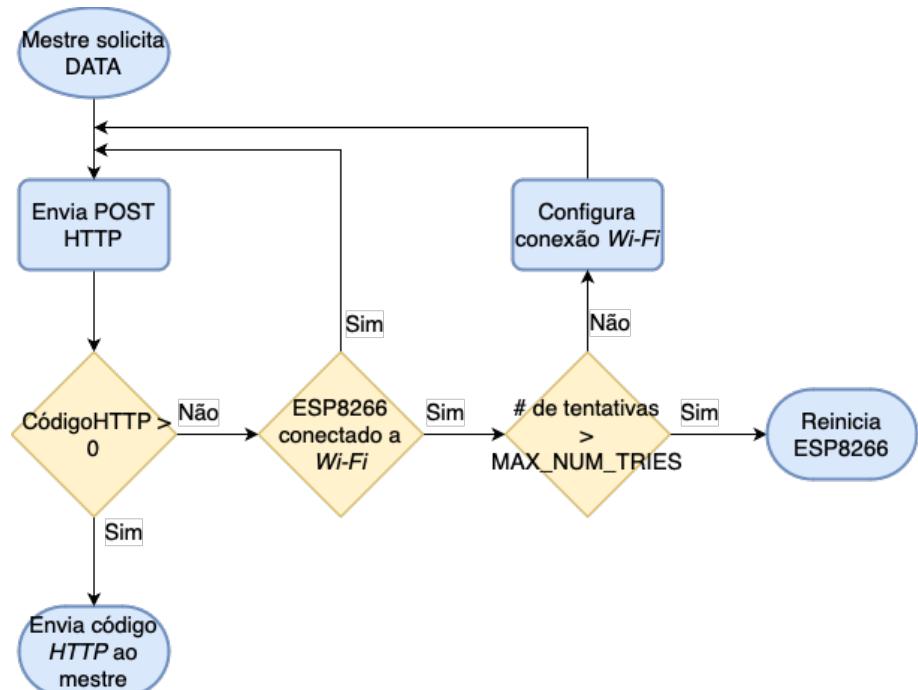
<code>CommandTypes</code>	Valor	Descrição	Resposta do mestre
<code>ERROR</code>	0	Este comando indica que ocorreu um erro na comunicação do mestre com o ESP8266. Normalmente usado do ESP8266 para o mestre	É a resposta em caso de erro de comunicação
<code>DATA</code>	1	Este comando indica que o mestre enviou os dados para uma postagem HTTP e está aguardando o código HTTP retornado do servidor Web como resposta à postagem	O código HTTP retornado pelo servidor Web após a postagem
<code>REQUESTTIME</code>	2	Este comando indica que o mestre está solicitando a hora de um servidor NTP	O carimbo de data/hora obtido pelo servidor NTP

Fonte: Desenvolvido pelo autor (2023)

E.2.1 O comando DATA: enviando um POST HTTP para a API Renovar

O código executado quando o "mestre" solicita o envio de um *POST HTTP* é mostrado na Lista E.2.1. O objeto `espHTTP` faz uma requisição tipo *POST* para a API hospedada em um servidor web identificado por um *HOST* (hospedeiro), uma Porta e uma URL específica. Os dados enviados no *POST* são os mesmos enviados anteriormente pelo mestre. Estes dados são acessados pelo método `get_data()` de `espSerial`. O código retornado dessa operação é enviado como resposta ao "mestre" para manter controle das suas operações. Caso o código for de falha e o ESP8266 não estiver conectado à rede *Wi-Fi*, o programa tentará se reconectar à rede e repassar os dados no máximo por três tentativas. Se, por outro lado, o código for de falha mas o ESP8266 estiver conectado a uma rede *Wi-Fi*, o programa tentará enviar a mesma requisição indefinidamente até que o mestre envie uma nova solicitação. A Figura 84 apresenta um fluxograma que representa esse processo.

Figura 84 – Fluxograma do processo após uma solicitação de DATA do mestre.



Fonte: Desenvolvido pelo autor (2023)

Listing E.6 – Sequencia de operações no comando DATA

```

case DATA:
{
    static uint8_t _numberOfPostTries = 0;
#define MAX_NUM_TRIES 3
    int code = espHTTP.post(HOST, PORT, URL, espSerial.get_data());
}

```

```

if( code <= 0 && WiFi. status () != WL_CONNECTED)
{
    setup_wifi_connection<NUM_WIFI>( wifiCreds );
    if(++_numberOfPostTries >= MAX_NUM_TRIES-1) ESP. restart ();
}
else espSerial.send_http_code( code , Serial );
break;
}

```

Os valores de HOST, PORT e URL são definidos no arquivo iot-generic.h conforme se mostra abaixo; eles representam o *endpoint* para enviar requisições à API Renovar.

```

#define HOST F( "renovar.lcqar.ufsc.br" )
#define PORT 8080UL
#define URL F( "/sample/" )

```

O comando REQUESTTIME: obtendo a hora de um servidor NTP

O código executado quando o ESP8266 recebe uma solicitação de horário é mostrado na Lista E.2.1. A função `get_time()` é definida no arquivo `esp-iot.h` para obter a data e hora atuais desde um servidor NTP. O valor retornado dessa operação é enviado como resposta ao "mestre". Caso o valor for zero e o ESP8266 não esteja conectado à rede Wi-Fi, o programa tentará se reconectar à rede e obter o horário da Internet por no máximo três tentativas. Por outro lado, se o valor for zero, mas o ESP8266 estiver conectado a uma rede Wi-Fi, o programa tentará obter o horário da Internet indefinidamente até receber uma nova solicitação. A Figura 85 apresenta um fluxograma que representa esse processo.

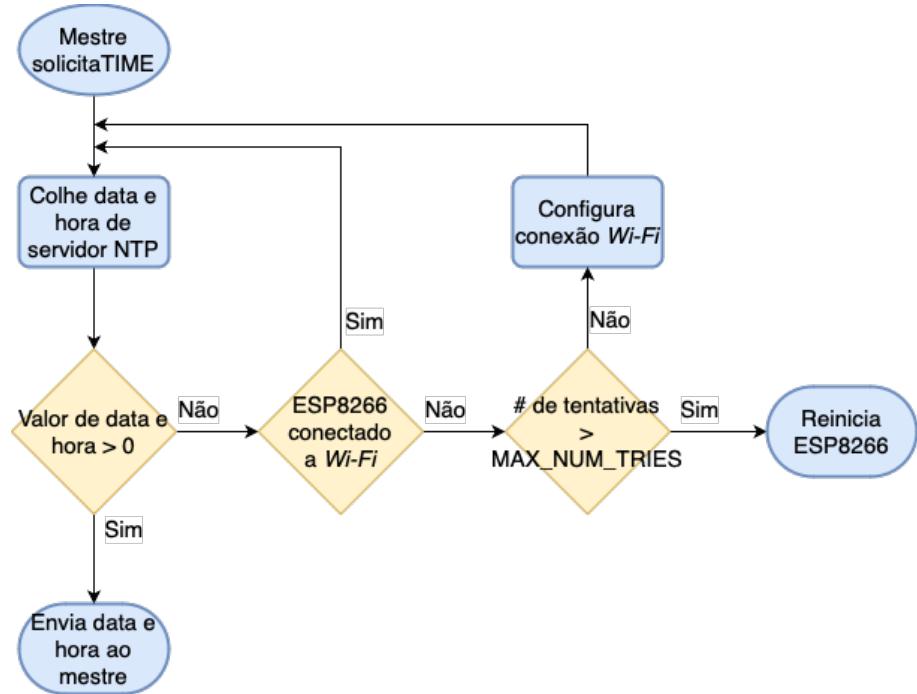
Listing E.7 – Sequencia de operações no comando REQUESTTIME

```

case REQUESTTIME:
{
    static uint8_t _numberOfTries = 0;
#define MAX_NUM_TRIES 3
    time_t t = get_time(TIMEZONE_SEC, DAYLIGHTOFFSET_SEC);
    if (!t && WiFi. status () != WL_CONNECTED)
    {
        setup_wifi_connection<NUM_WIFI>( wifiCreds );
        if(++_numberOfTries >= MAX_NUM_TRIES-1) ESP. restart ();
    }
    else espSerial.send_time( t , Serial );
    break;
}

```

Figura 85 – Fluxograma do processo após uma solicitação TIME do mestre



Fonte: Desenvolvido pelo autor (2023)

A função `get_time()` recebe os parâmetros `TIMEZONE_SEC` e `DAYLIGHTOFFSET_SEC`, definidos anteriormente no arquivo `main.cpp`. O exemplo abaixo mostra como definir esses dois parâmetros para a aplicação no Brasil. `TIMEZONE_SEC` é o fuso horário onde o monitor será instalado, convertido a segundos. `DAYLIGHTOFFSET_SEC` define o deslocamento, em segundos, para o horário de verão, nesse caso foi definido como zero.

```
#define TIMEZONE_SEC      -3*3600
#define DAYLIGHTOFFSET_SEC 0*3600
```

APÊNDICE F – LEITURAS DE SO₂ ADQUIRIDAS PELO SENSOR SO₂-B4

Para a medição de SO_2 foram utilizados dois sensores do modelo SO₂-B4. As Figuras 86a e 86c mostram as séries temporais dos sensores depois de removidos os valores fora de intervalo. Observa-se que ambos sensores sofreram seguidas alterações no valor de linha base (Figuras 86b e 86d), sendo difícil aproveitar algum período de dados úteis. As distribuições das leituras dos sensores também não apresentaram um formato log-normal como se observa nos histogramas das leituras dos sensores mostrados nas Figuras 87a e 87b. Por esses motivos os dados desses sensores não foram considerados para análise no trabalho.

Figura 86 – Série temporal dos sensores de SO_2 modelo SO₂-B4

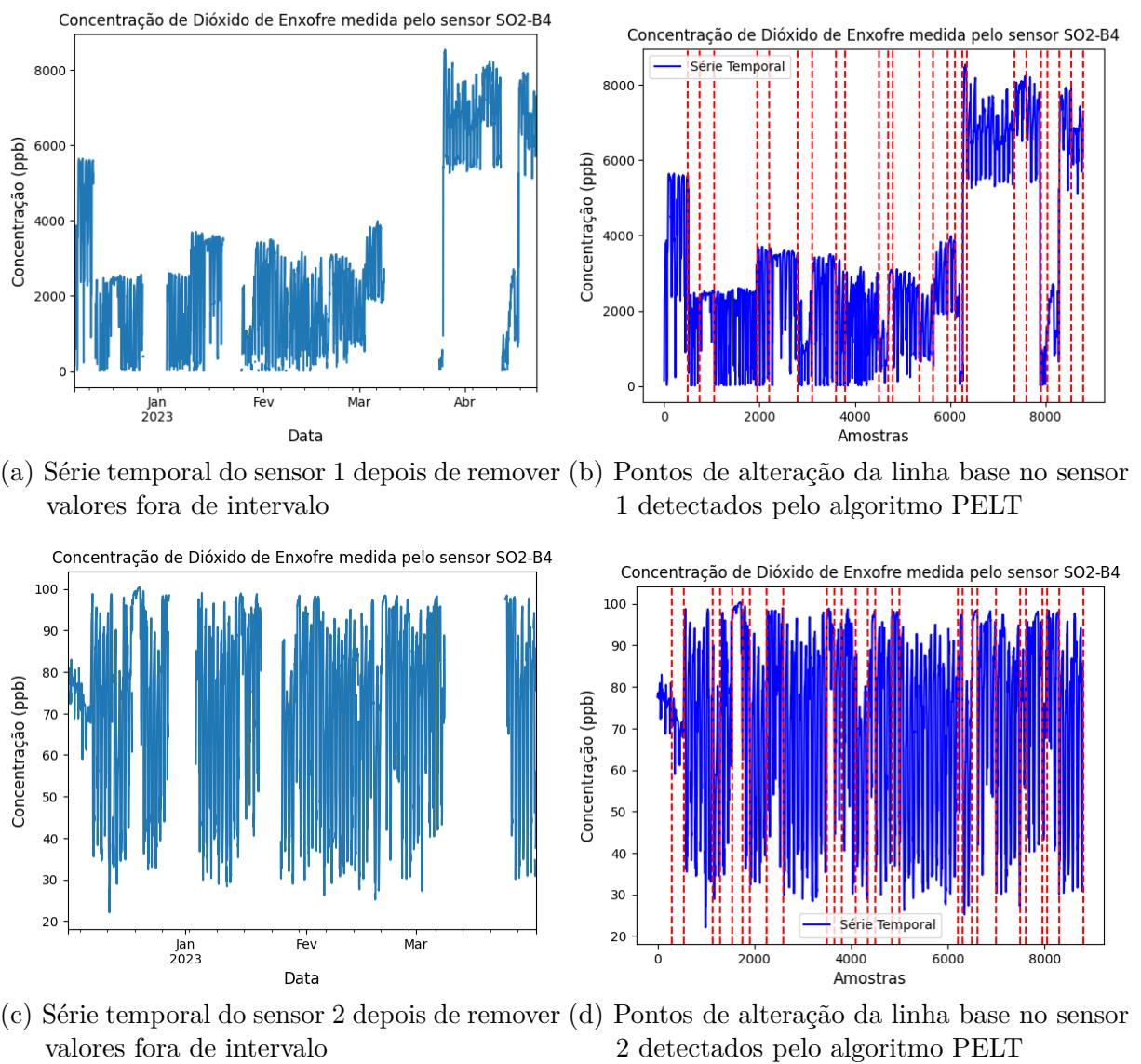
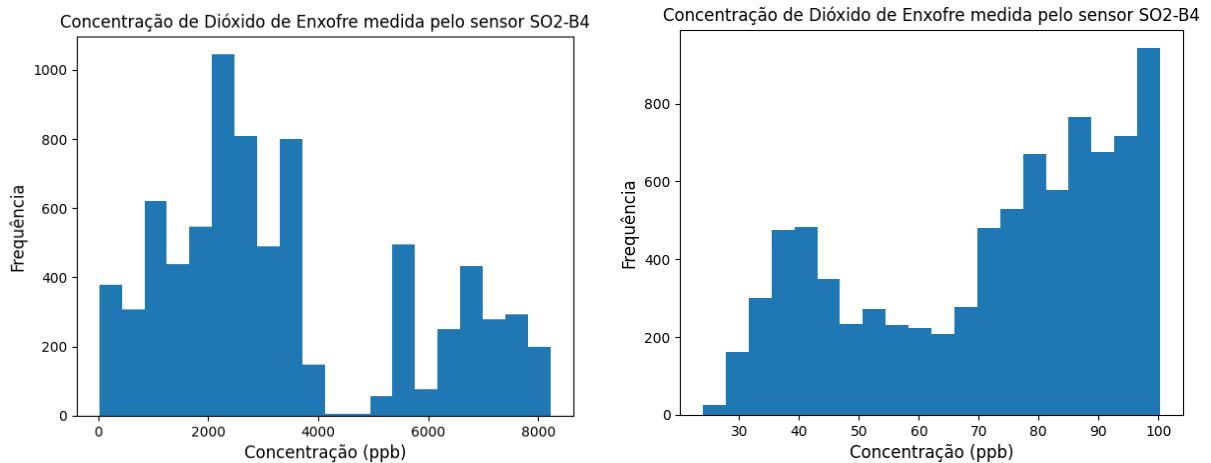


Figura 87 – Histogramas das leituras dos sensores SO₂-B4(a) Histograma das leituras do sensor 1 SO₂-B4 (b) Histograma das leituras do sensor 2 SO₂-B4

ANEXO A – DOCUMENTAÇÃO DA API RENOVAR: *ENDPOINTS E REQUISIÇÕES*

Api Documentation

Api Documentation

[Apache 2.0](#)

basic-error-controller : Basic Error Controller

DELETE /error

errorHtml

Response Class (Status 200)

OK

Example Value

```
{  
    "empty": true,  
    "model": {},  
    "modelMap": {},  
    "reference": true,  
    "status": "100",  
    "view": {  
        "contentType": "string"  
    },  
    "viewName": "string"  
}
```

Response Content Type

Response Messages

HTTP Status Code	Reason	Response Model	Headers
204	No Content		
401	Unauthorized		
403	Forbidden		

GET /error

errorHtml

Response Class (Status 200)

OK

Example Value

```
{  
    "empty": true,  
    "model": {},  
    "modelMap": {},  
    "reference": true,  
    "status": "100",  
    "view": {  
        "contentType": "string"  
    },  
    "viewName": "string"  
}
```

Response Content Type [text/html](#)

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

HEAD /error errorHtml

Response Class (Status 200)

OK

Example Value

```
{  
    "empty": true,  
    "model": {},  
    "modelMap": {},  
    "reference": true,  
    "status": "100",  
    "view": {  
        "contentType": "string"  
    },  
    "viewName": "string"  
}
```

Response Content Type [text/html](#)

Response Messages

HTTP Status Code	Reason	Response Model	Headers
204	No Content		
401	Unauthorized		
403	Forbidden		

OPTIONS /error

errorHtml

Response Class (Status 200)

OK

Example Value

```
{  
  "empty": true,  
  "model": {},  
  "modelMap": {},  
  "reference": true,  
  "status": "100",  
  "view": {  
    "contentType": "string"  
  },  
  "viewName": "string"  
}
```

Response Content Type

Response Messages

HTTP Status Code	Reason	Response Model	Headers
204	No Content		
401	Unauthorized		
403	Forbidden		

PATCH /error

errorHtml

Response Class (Status 200)

OK

Example Value

```
{  
  "empty": true,  
  "model": {},  
  "modelMap": {},  
  "reference": true,  
  "status": "100",  
  "view": {  
    "contentType": "string"  
  },  
  "viewName": "string"  
}
```

Response Content Type

Response Messages

HTTP Status	Reason	Response Model	Headers
-------------	--------	----------------	---------

Code

204 No Content

401 Unauthorized

403 Forbidden

Try it out!

POST /error

errorHtml

Response Class (Status 200)

OK

Example Value

```
{  
  "empty": true,  
  "model": {},  
  "modelMap": {},  
  "reference": true,  
  "status": "100",  
  "view": {  
    "contentType": "string"  
  },  
  "viewName": "string"  
}
```

Response Content Type

Response Messages

HTTP Status Code	Reason	Response Model	Headers
201	Created		
401	Unauthorized		
403	Forbidden		
404	Not Found		

Try it out!

PUT /error

errorHtml

Response Class (Status 200)

OK

Example Value

```
{
  "empty": true,
  "model": {},
  "modelMap": {},
  "reference": true,
  "status": "100",
  "view": {
    "contentType": "string"
  },
  "viewName": "string"
}
```

Response Content Type [text/html](#)

Response Messages

HTTP Status Code	Reason	Response Model	Headers
201	Created		
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

coordinate-resource : Coordinate Resource

GET	/coordinate	getAllCoordinates
-----	-------------	-------------------

Response Class (Status 200)

OK

Example Value

```
[
  {
    "id": 0,
    "latitude": 0,
    "longitude": 0
  }
]
```

Response Content Type [/*](#)

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		

403

Forbidden

404

Not Found

[Try it out!](#)

POST /coordinate

insertDevice

Parameters

Parameter	Value	Description	Parameter Type	Data Type
coordinate	<input type="text"/>	coordinate	body	Example Value <pre>{ "id": 0, "latitude": 0, "longitude": 0 }</pre>

Parameter content type:

Response Messages

HTTP Status Code	Reason	Response Model	Headers
200	OK		
201	Created		
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

GET /coordinate/check/{latitude}/{longitude}

getCoordinate

Response Class (Status 200)

boolean

Response Content Type

Parameters

Parameter	Value	Description	Parameter Type	Data Type
latitude	<input type="text"/>	latitude	path	double
longitude	<input type="text"/>	longitude	path	double

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

Try it out!

GET /coordinate/lat/{lat}

getCoordinatesByLatitude

Response Class (Status 200)

OK

Example Value

```
[  
  {  
    "id": 0,  
    "latitude": 0,  
    "longitude": 0  
  }  
]
```

Response Content Type `/*`

Parameters

Parameter	Value	Description	Parameter Type	Data Type
latitude	<input type="text"/>	latitude	query	double

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

Try it out!

GET /coordinate/long/{long}

getCoordinatesByLongitude

Response Class (Status 200)

OK

Example Value

```
[  
  {  
    "id": 0,  
    "latitude": 0,  
    "longitude": 0  
  }  
]
```

Response Content Type

Parameters

Parameter	Value	Description	Parameter Type	Data Type
longitude	<input type="text"/>	longitude	query	double

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

GET /coordinate/unique/{sensorID}

getUniqueCoordinateFromSamples

Response Class (Status 200)

OK

Example Value

```
[  
  {  
    "id": 0,  
    "latitude": 0,  
    "longitude": 0  
  }  
]
```

Response Content Type

Parameters

Parameter	Value	Description	Parameter Type	Data Type
sensorID	<input type="text"/>	sensorID	path	integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

DELETE /coordinate/{id}

deleteDevice

Parameters

Parameter	Value	Description	Parameter Type	Data Type
id	<input type="text"/>	id	path	integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
200	OK		
204	No Content		
401	Unauthorized		
403	Forbidden		

[Try it out!](#)

GET /coordinate/{id}

getCoordinate

Response Class (Status 200)

OK

Example Value

```
{  
  "id": 0,  
  "latitude": 0,  
  "longitude": 0  
}
```

Response Content Type `/*`

Parameters

Parameter	Value	Description	Parameter Type	Data Type
id	<input type="text"/>	id	path	integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

device-resource : Device Resource

GET `/device` [getAllDevices](#)

Response Class (Status 200)

OK

Example Value

```
[  
  {  
    "coordinate": {  
      "id": 0,  
      "latitude": 0,  
      "longitude": 0  
    },  
    "created_at": "2023-11-07T03:03:32.810Z",  
    "id": 0,  
    "name": "string",  
    "portable": true
```

Response Content Type `/*`

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

POST /device [insertDevice](#)

Parameters

Parameter	Value	Description	Parameter Type	Data Type
deviceDTO	<input type="text"/>	deviceDTO	body	Example Value

Parameter content type:
[application/json](#)

```
{
  "latitude": 0,
  "longitude": 0,
  "name": "string",
  "tracking": "string",
  "userID": 0
}
```

Response Messages

HTTP Status Code	Reason	Response Model	Headers
200	OK		
201	Created		
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

GET /device/active/{deviceid} [isDeviceActive](#)

Response Class (Status 200)

boolean

Response Content Type [/*](#)

Parameters

Parameter	Value	Description	Parameter Type	Data Type
-----------	-------	-------------	----------------	-----------

deviceid	<input type="text"/>	deviceid	path	integer
Response Messages				
HTTP Status Code	Reason	Response Model		Headers
401	Unauthorized			
403	Forbidden			
404	Not Found			
Try it out!				

GET	/device/coordid/{coordid}	getByCoordinateSample
Response Class (Status 200)		
OK		
Example Value <pre>[{ "coordinate": { "id": 0, "latitude": 0, "longitude": 0 }, "created_at": "2023-11-07T03:03:32.813Z", "id": 0, "name": "string", "portable": true }]</pre>		

Response Content Type

Parameters

Parameter	Value	Description	Parameter Type	Data Type
coordid	<input type="text"/>	coordid	path	integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		
Try it out!			

GET	/device/coordinate/{id}	getByCoordinate
-----	-------------------------	---------------------------------

Response Class (Status 200)

OK

Example Value

```
[  
 {  
   "latitude": 0,  
   "longitude": 0,  
   "name": "string",  
   "tracking": "string",  
   "userID": 0  
 }  
]
```

Response Content Type [/*](#)

Parameters

Parameter	Value	Description	Parameter Type	Data Type
id	<input type="text"/>	id	path	integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

GET /device/name/{name} [getDeviceByUser](#)

Response Class (Status 200)

OK

Example Value

```
[  
 {  
   "latitude": 0,  
   "longitude": 0,  
   "name": "string",  
   "tracking": "string",  
   "userID": 0  
 }  
]
```

Response Content Type `*/*`

Parameters

Parameter	Value	Description	Parameter Type	Data Type
name	<input type="text"/>	name	path	string

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

GET [/device/nonportables](#) [getNonPortableDevice](#)

Response Class (Status 200)

OK

Example Value

```
[  
  {  
    "latitude": 0,  
    "longitude": 0,  
    "name": "string",  
    "tracking": "string",  
    "userID": 0  
  }  
]
```

Response Content Type `*/*`

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

GET [/device/page](#) [buscarPagina](#)

Response Class (Status 200)

OK

Example Value

```
{
  "content": [
    {
      "latitude": 0,
      "longitude": 0,
      "name": "string",
      "tracking": "string",
      "userID": 0
    }
  ],
  "first": true
}
```

Response Content Type [/*](#)

Parameters

Parameter	Value	Description	Parameter Type	Data Type
page	<input type="text"/>	page	query	integer
linesPerPage	<input type="text"/> 10	linesPerPage	query	integer
orderBy	<input type="text"/> id	orderBy	query	string
direction	<input type="text"/> ASC	direction	query	string

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

Try it out!

GET /device/portables

getPortableDevice

Response Class (Status 200)

OK

Example Value

```
[
  {
    "latitude": 0,
    "longitude": 0,
    "name": "string",
    "tracking": "string",
    "userID": 0
  }
]
```

Response Content Type [/*](#)

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

GET /device/user/{id} [getDeviceByUser](#)

Response Class (Status 200)

OK

Example Value

```
[
  {
    "latitude": 0,
    "longitude": 0,
    "name": "string",
    "tracking": "string",
    "userID": 0
  }
]
```

Response Content Type [/*](#)

Parameters

Parameter	Value	Description	Parameter Type	Data Type
id	<input type="text"/>	id	path	integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
------------------	--------	----------------	---------

401 Unauthorized

403 Forbidden

404 Not Found

Try it out!

DELETE /device/{id}

deleteDevice

Parameters

Parameter	Value	Description	Parameter Type	Data Type
id	<input type="text"/>	id	path	integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
200	OK		
204	No Content		
401	Unauthorized		
403	Forbidden		

Try it out!

GET /device/{id}

getDevice

Response Class (Status 200)

OK

Example Value

```
{  
  "latitude": 0,  
  "longitude": 0,  
  "name": "string",  
  "tracking": "string",  
  "userID": 0  
}
```

Response Content Type

Parameters

Parameter	Value	Description	Parameter Type	Data Type
id	<input type="text"/>	id	path	integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

Try it out!

PUT /device/{id}

updateDevice

Parameters

Parameter	Value	Description	Parameter Type	Data Type	Example Value
device	<input type="text"/>	device	body		<pre>{ "coordinate": { "id": 0, "latitude": 0, "longitude": 0 }, "created_at": "2023-11-07T03:03:32.822Z", "id": 0, "name": "string", "portable": true, "tracking_id": "string" }</pre>
id	<input type="text"/>	id	path	integer	

Response Messages

HTTP Status Code	Reason	Response Model	Headers
200	OK		
201	Created		
401	Unauthorized		
403	Forbidden		
404	Not Found		

Try it out!

pollutant-resource : Pollutant Resource

GET	/pollutant		getAll
-----	------------	--	--------

Response Class (Status 200)

OK

Example Value

```
{}
```

Response Content Type `*/*`

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

POST `/pollutant`

addPollutant

Parameters

Parameter	Value	Description	Parameter Type	Data Type	Example Value
<code>pollutantDTO</code>	<input type="text"/>	<code>pollutantDTO</code>	body		<pre>{ "max": 0, "name": "string", "unitId": 0 }</pre>

Parameter content type:

`application/json`

Response Messages

HTTP Status Code	Reason	Response Model	Headers
200	OK		
201	Created		

401 Unauthorized

403 Forbidden

404 Not Found

Try it out!

GET /pollutant/name/{name}

getPollutantByUnit

Response Class (Status 200)

OK

Example Value

```
{}
```

Response Content Type `*/*`

Parameters

Parameter	Value	Description	Parameter Type	Data Type
name	<input type="text"/>	name	path	string

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

Try it out!

GET /pollutant/sensor/{sensorID}

getPollutantBySensor

Response Class (Status 200)

OK

Example Value

```
{}
```

Response Content Type

Parameters

Parameter	Value	Description	Parameter Type	Data Type
sensorID	<input type="text"/>	sensorID	path	integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

GET /pollutant/unit/{id} getPollutantByUnit

Response Class (Status 200)

OK

Example Value

```
{}
```

Response Content Type

Parameters

Parameter	Value	Description	Parameter Type	Data Type
-----------	-------	-------------	----------------	-----------

id**id**

path

integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

DELETE /pollutant/{id} deletePollutant

Parameters

Parameter	Value	Description	Parameter Type	Data Type
id	<input type="text"/>	id	path	integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
200	OK		
204	No Content		
401	Unauthorized		
403	Forbidden		

[Try it out!](#)

GET /pollutant/{id} getPollutant

Response Class (Status 200)

OK

Example Value

```
{}
```

Response Content Type `*/*`

Parameters

Parameter	Value	Description	Parameter Type	Data Type
id	<input type="text"/>	id	path	integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

PUT `/pollutant/{id}` updatePollutant

Response Class (Status 200)

OK

Example Value

```
{}
```

Response Content Type `*/*`

Parameters

Parameter	Value	Description	Parameter Type	Data Type
pollutant	<input type="text"/>	pollutant	body	Example Value

Parameter content type:

`application/json`

```
{
  "id": 0,
  "max": 0,
  "name": "string",
  "unit": {
    "id": 0,
    "name": "string",
    "type": {
      "id": 0,
      "name": "string"
    }
  }
},
```

id	<input type="text"/>	id	path	integer
-----------	----------------------	-----------	------	---------

Response Messages

HTTP Status Code	Reason	Response Model	Headers
201	Created		
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

sample-resource : Sample Resource

GET	/sample	getAll
-----	---------	--------

Response Class (Status 200)

OK

Example Value

```
[
  {
    "coordinate": {
      "id": 0,
      "latitude": 0,
      "longitude": 0
    },
    "date": 0,
    "device": "string",
    "id": 0,
    "measuring": 0
  }
],
```

Response Content Type [/*](#)

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		

403

Forbidden

404

Not Found

[Try it out!](#)

POST /sample

Adds a sample

Parameters

Parameter	Value	Description	Parameter Type	Data Type
sampleDTO	<input type="text"/>	sampleDTO	body	Example Value <pre>{ "latitude": 0, "longitude": 0, "measuring": 0, "sensorid": 0, "timestamp": 0 }</pre>

Parameter content type:

 application/json

Response Messages

HTTP Status Code	Reason	Response Model	Headers
200	OK		
201	Created		
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

GET /sample/boxplot/

getBoxPlotValues

Response Class (Status 200)

OK

Example Value

```
[
  {
    "group": "string",
    "values": [
      0
    ]
  }
]
```

Response Content Type [/*](#)

Parameters

Parameter	Value	Description	Parameter Type	Data Type
sensorID	<input type="text"/>	sensorID	query	integer
group	<input type="text"/>	group	query	string
start	<input type="text"/>	start	query	string
end	<input type="text"/>	end	query	string

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

GET </sample/coord/{id}> [getSamplesByCoordinate](#)

Response Class (Status 200)

OK

Example Value

```
[
  {
    "coordinate": {
      "id": 0,
      "latitude": 0,
      "longitude": 0
    },
    "date": 0,
    "device": "string",
    "id": 0,
    "measuring": 0
  }
]
```

Response Content Type [/*](#)

Parameters

Parameter	Value	Description	Parameter Type	Data Type
id	<input type="text"/>	id	path	integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

GET </sample/device/active/{deviceID}> checkActiveDevice

Response Class (Status 200)

boolean

Response Content Type [/*](#)

Parameters

Parameter	Value	Description	Parameter Type	Data Type
deviceID	<input type="text"/>	deviceID	path	integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

GET	/sample/device/desc/{deviceID}	getSamplesViaDeviceDESC
-----	--------------------------------	-------------------------

Response Class (Status 200)

OK

Example Value

```
[  
  {  
    "coordinate": {  
      "id": 0,  
      "latitude": 0,  
      "longitude": 0  
    },  
    "date": "2023-11-07T03:03:32.834Z",  
    "id": 0,  
    "latitude": 0,  
    "longitude": 0
```

Response Content Type `/*`

Parameters

Parameter	Value	Description	Parameter Type	Data Type
deviceID	<input type="text"/>	deviceID	path	integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

Try it out!

GET	/sample/device/last/{deviceID}	getLastViaDevice
-----	--------------------------------	------------------

Response Class (Status 200)

OK

Example Value

```
{
  "content": [
    {
      "coordinate": {
        "id": 0,
        "latitude": 0,
        "longitude": 0
      },
      "date": "2023-11-07T03:03:32.835Z",
      "id": 0,
      "latitude": 0
    }
  ]
}
```

Response Content Type `*/*`

Parameters

Parameter	Value	Description	Parameter Type	Data Type
deviceID	<input type="text"/>	deviceID	path	integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

GET `/sample/device/pollutant/{deviceID}/{pollutantID}`

[getLastSampleViaPollutant](#)

Response Class (Status 200)

OK

Example Value

```
{
  "coordinate": {
    "id": 0,
    "latitude": 0,
    "longitude": 0
  },
  "date": "2023-11-07T03:03:32.836Z",
  "id": 0,
  "latitude": 0,
  "longitude": 0,
  "measuring": 0
}
```

Response Content Type `*/*`

Parameters

Parameter	Value	Description	Parameter Type	Data Type
-----------	-------	-------------	----------------	-----------

deviceID	<input type="text"/>	deviceID	path	integer
pollutantID	<input type="text"/>	pollutantID	path	integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

GET	/sample/device/{deviceID}	getSamplesViaDevice
-----	---------------------------	---------------------

Response Class (Status 200)

OK

Example Value

```
[  
  {  
    "coordinate": {  
      "id": 0,  
      "latitude": 0,  
      "longitude": 0  
    },  
    "date": 0,  
    "device": "string",  
    "id": 0,  
    "measuring": 0
```

Response Content Type [/*](#)

Parameters

Parameter	Value	Description	Parameter Type	Data Type
deviceID	<input type="text"/>	deviceID	path	integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

GET /sample/exists/{deviceID}/{pollutantID}

checkSampleExist

Response Class (Status 200)

OK

Example Value

```
{}
```

Response Content Type `*/*`

Parameters

Parameter	Value	Description	Parameter Type	Data Type
deviceID	<input type="text"/>	deviceID	path	integer
pollutantID	<input type="text"/>	pollutantID	path	integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

GET /sample/page

searchPage

Response Class (Status 200)

OK

Example Value

```
{
  "content": [
    {
      "coordinate": {
        "id": 0,
        "latitude": 0,
        "longitude": 0
      },
      "date": "2023-11-07T03:03:32.840Z",
      "id": 0,
      "latitude": 0
    }
  ]
}
```

Response Content Type [/*](#)

Parameters

Parameter	Value	Description	Parameter Type	Data Type
deviceID	<input type="text"/>	deviceID	query	integer
page	<input type="text"/>	page	query	integer
linesPerPage	<input type="text"/> 5	linesPerPage	query	integer
orderBy	<input type="text"/> data	orderBy	query	string
direction	<input type="text"/> DESC	direction	query	string

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

GET /sample/range/ [getPollutantDates](#)

Response Class (Status 200)

OK

Example Value

```
{
  "content": [
    {
      "coordinate": {
        "id": 0,
        "latitude": 0,
        "longitude": 0
      },
      "date": "2023-11-07T03:03:32.842Z",
      "id": 0,
      "latitude": 0
    }
  ]
}
```

Response Content Type [/*](#)

Parameters

Parameter	Value	Description	Parameter Type	Data Type
deviceID	<input type="text"/>	deviceID	query	integer
pollutantID	<input type="text"/>	pollutantID	query	integer
startDate	<input type="text"/>	startDate	query	string
endDate	<input type="text"/>	endDate	query	string

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

GET </sample/sensor/all/{id}> [getAllSamplesBySensor](#)

Response Class (Status 200)

OK

Example Value

```
[
  {
    "coordinate": {
      "id": 0,
      "latitude": 0,
      "longitude": 0
    },
    "date": "2023-11-07T03:03:32.844Z",
    "id": 0,
    "latitude": 0,
    "longitude": 0
  }
]
```

Response Content Type `/*`

Parameters

Parameter	Value	Description	Parameter Type	Data Type
id	<input type="text"/>	id	path	integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

GET `/sample/sensor/last/{sensorID}`

[getLastSampleViaSensor](#)

Response Class (Status 200)

OK

[Example Value](#)

```
{
  "coordinate": {
    "id": 0,
    "latitude": 0,
    "longitude": 0
  },
  "date": "2023-11-07T03:03:32.845Z",
  "id": 0,
  "latitude": 0,
  "longitude": 0,
  "measuring": 0
}
```

Response Content Type `/*`

Parameters

Parameter	Value	Description	Parameter Type	Data Type
-----------	-------	-------------	----------------	-----------

sensorID**sensorID**

path

integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

GET /sample/sensor/range/ getSensorDates

Response Class (Status 200)

OK

Example Value

```
{
  "content": [
    {
      "coordinate": {
        "id": 0,
        "latitude": 0,
        "longitude": 0
      },
      "date": "2023-11-07T03:03:32.846Z",
      "id": 0,
      "latitude": 0
    }
  ]
}
```

Response Content Type /*

Parameters

Parameter	Value	Description	Parameter Type	Data Type
sensorID	<input type="text"/>	sensorID	query	integer
startDate	<input type="text"/>	startDate	query	string
endDate	<input type="text"/>	endDate	query	string

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

GET /sample/sensor/{id} getSamplesBySensor

Response Class (Status 200)

OK

Example Value

```
{  
  "content": [  
    {  
      "coordinate": {  
        "id": 0,  
        "latitude": 0,  
        "longitude": 0  
      },  
      "date": "2023-11-07T03:03:32.847Z",  
      "id": 0,  
      "latitude": 0  
    }  
  ]  
}
```

Response Content Type `/*`

Parameters

Parameter	Value	Description	Parameter Type	Data Type
id	<input type="text"/>	id	path	integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

Try it out!

GET /sample/sensor/{sensorID}/device/{deviceID} getLastViaSensorDevice

Response Class (Status 200)

OK

Example Value

```
[
  {
    "coordinate": {
      "id": 0,
      "latitude": 0,
      "longitude": 0
    },
    "date": 0,
    "device": "string",
    "id": 0,
    "measuring": 0
  }
]
```

Response Content Type `/*`

Parameters

Parameter	Value	Description	Parameter Type	Data Type
sensorID	<input type="text"/>	sensorID	path	integer
deviceID	<input type="text"/>	deviceID	path	integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

Try it out!

GET `/sample/{deviceID}/{pollutantID}`

getSamplesViaDevicePollutant

Response Class (Status 200)

OK

Example Value

```
[
  {
    "coordinate": {
      "id": 0,
      "latitude": 0,
      "longitude": 0
    },
    "date": "2023-11-07T03:03:32.849Z",
    "id": 0,
    "latitude": 0,
    "longitude": 0
  }
]
```

Response Content Type `/*`

Parameters

Parameter	Value	Description	Parameter Type	Data Type
deviceID	<input type="text"/>	deviceID	path	integer
pollutantID	<input type="text"/>	pollutantID	path	integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

GET /sample/{id}

getSample

Response Class (Status 200)

OK

Example Value

```
{
  "coordinate": {
    "id": 0,
    "latitude": 0,
    "longitude": 0
  },
  "date": "2023-11-07T03:03:32.851Z",
  "id": 0,
  "latitude": 0,
  "longitude": 0,
  "measuring": 0
}
```

Response Content Type [/*](#)

Parameters

Parameter	Value	Description	Parameter Type	Data Type
id	<input type="text"/>	id	path	integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

sensor-resource : Sensor Resource

GET /sensor

getAllSensors

Response Class (Status 200)

OK

Example Value

```
[  
 {  
   "brand": "string",  
   "created_at": "2023-11-07T03:03:32.852Z",  
   "device": {  
     "coordinate": {  
       "id": 0,  
       "latitude": 0,  
       "longitude": 0  
     },  
     "created_at": "2023-11-07T03:03:32.852Z"  
   }  
 }]
```

Response Content Type `/*`

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

Try it out!

POST /sensor

addSensor

Response Class (Status 200)

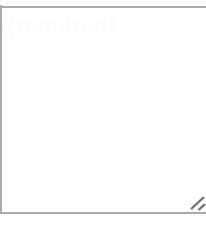
OK

Example Value

```
{  
   "brand": "string",  
   "created_at": "2023-11-07T03:03:32.853Z",  
   "device": {  
     "coordinate": {  
       "id": 0,  
       "latitude": 0,  
       "longitude": 0  
     },  
     "created_at": "2023-11-07T03:03:32.853Z",  
     "id": 0  
   }  
 }
```

Response Content Type `/*`

Parameters

Parameter	Value	Description	Parameter Type	Data Type
sensorDTO		sensorDTO	body	Example Value <pre>{ "brand": "string", "deviceId": 0, "model": "string", "name": "string", "pollutantId": 0 }</pre>

Parameter content type:

`application/json`

Response Messages

HTTP Status Code	Reason	Response Model	Headers
201	Created		
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

GET [/sensor/brand/{brand}](#) [getSensorsByBrand](#)

Response Class (Status 200)

OK

Example Value

```
[  
 {  
   "brand": "string",  
   "created_at": "2023-11-07T03:03:32.854Z",  
   "device": {  
     "coordinate": {  
       "id": 0,  
       "latitude": 0,  
       "longitude": 0  
     },  
     "created_at": "2023-11-07T03:03:32.851Z"  
   }  
 }]
```

Response Content Type `/*`

Parameters

Parameter	Value	Description	Parameter Type	Data Type
-----------	-------	-------------	----------------	-----------

brand	<input type="text"/>	brand	path	string
Response Messages				
HTTP Status Code	Reason	Response Model		Headers
401	Unauthorized			
403	Forbidden			
404	Not Found			
Try it out!				

GET	/sensor/device/{id}	getSensorsByDevice
Response Class (Status 200)		
OK		
Example Value <pre>[{ "brand": "string", "created_at": "2023-11-07T03:03:32.855Z", "device": { "coordinate": { "id": 0, "latitude": 0, "longitude": 0 }, "created_at": "2023-11-07T03:03:32.855Z" } }]</pre>		

Response Content Type [/*](#)

Parameters

Parameter	Value	Description	Parameter Type	Data Type
id	<input type="text"/>	id	path	integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		
Try it out!			

GET	/sensor/model/{model}	getSensorsByModel
-----	-----------------------	-----------------------------------

Response Class (Status 200)

OK

Example Value

```
[  
 {  
   "brand": "string",  
   "created_at": "2023-11-07T03:03:32.856Z",  
   "device": {  
     "coordinate": {  
       "id": 0,  
       "latitude": 0,  
       "longitude": 0  
     },  
     "created_at": "2023-11-07T03:03:32.856Z"  
   }  
 }]
```

Response Content Type

/*

Parameters

Parameter	Value	Description	Parameter Type	Data Type
model	<input type="text"/>	model	path	string

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

Try it out!

GET /sensor/pollutant/{id} [getSensorsByPollutant](#)

Response Class (Status 200)

OK

Example Value

```
[  
 {  
   "brand": "string",  
   "created_at": "2023-11-07T03:03:32.856Z",  
   "device": {  
     "coordinate": {  
       "id": 0,  
       "latitude": 0,  
       "longitude": 0  
     },  
     "created_at": "2023-11-07T03:03:32.856Z"  
   }  
 }]
```

Response Content Type `/*`

Parameters

Parameter	Value	Description	Parameter Type	Data Type
id	<input type="text"/>	id	path	integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

GET [/sensor/{id}](#) [getSensorsById](#)

Response Class (Status 200)

OK

Example Value

```
{  
  "brand": "string",  
  "created_at": "2023-11-07T03:03:32.857Z",  
  "device": {  
    "coordinate": {  
      "id": 0,  
      "latitude": 0,  
      "longitude": 0  
    },  
    "created_at": "2023-11-07T03:03:32.857Z",  
    "id": 0  
  }  
}
```

Response Content Type `/*`

Parameters

Parameter	Value	Description	Parameter Type	Data Type
id	<input type="text"/>	id	path	integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

type-resources : Type Resources

GET /type

getAllTypes

Response Class (Status 200)

OK

Example Value

```
[  
  {  
    "id": 0,  
    "name": "string"  
  }  
]
```

Response Content Type `/*`

Response Messages

HTTP Status
Code

Reason

Response Model

Headers

401 Unauthorized

403 Forbidden

404 Not Found

Try it out!

POST /type

insert

Parameters

Parameter

Value

Description

Parameter
Type

Data Type

type

type

body

Example Value

Parameter content type:

application/json

```
{  
  "id": 0,  
  "name": "string"  
}
```

Response Messages

HTTP Status Code	Reason	Response Model	Headers
200	OK		
201	Created		
401	Unauthorized		
403	Forbidden		
404	Not Found		

Try it out!

GET /type/name/{name}

getByName

Response Class (Status 200)

OK

Example Value

```
{  
  "id": 0,  
  "name": "string"  
}
```

Response Content Type

Parameters

Parameter	Value	Description	Parameter Type	Data Type
name	<input type="text"/>	name	path	string

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

GET	/type/{id}	getById
-----	------------	---------

Response Class (Status 200)

OK

Example Value

```
{
  "id": 0,
  "name": "string"
}
```

Response Content Type [/*](#)

Parameters

Parameter	Value	Description	Parameter Type	Data Type
id	<input type="text"/>	id	path	integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

unit-resources : Unit Resources

GET	/unit	getAllUnits
-----	-------	-------------

Response Class (Status 200)

OK

Example Value

```
[  
  {  
    "id": 0,  
    "name": "string",  
    "type": {  
      "id": 0,  
      "name": "string"  
    }  
  }  
]
```

Response Content Type `/*`

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

Try it out!

POST `/unit`

insert

Parameters

Parameter	Value	Description	Parameter Type	Data Type	Example Value
unitDTO	<input type="text"/>	unitDTO	body		<pre>{ "name": "string", "typeId": 0 }</pre>

Parameter content type:

`application/json`

Response Messages

HTTP Status Code	Reason	Response Model	Headers
200	OK		
201	Created		
401	Unauthorized		

403 Forbidden

404 Not Found

Try it out!

GET /unit/name/{name}

getByName

Response Class (Status 200)

OK

Example Value

```
{  
    "id": 0,  
    "name": "string",  
    "type": {  
        "id": 0,  
        "name": "string"  
    }  
}
```

Response Content Type `/*`

Parameters

Parameter	Value	Description	Parameter Type	Data Type
name	<input type="text"/>	name	path	string

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

Try it out!

GET /unit/type/{typeid}

getByType

Response Class (Status 200)

OK

Example Value

```
[
  {
    "id": 0,
    "name": "string",
    "type": {
      "id": 0,
      "name": "string"
    }
  }
]
```

Response Content Type `/*`

Parameters

Parameter	Value	Description	Parameter Type	Data Type
typeid	<input type="text"/>	typeid	path	integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

GET `/unit/{id}`

getUnit

Response Class (Status 200)

OK

Example Value

```
{
  "id": 0,
  "name": "string",
  "type": {
    "id": 0,
    "name": "string"
  }
}
```

Response Content Type `/*`

Parameters

Parameter	Value	Description	Parameter Type	Data Type
-----------	-------	-------------	----------------	-----------

id

[Required]

id

path

integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

user-resource : User Resource

GET	/user	getUsers
-----	-------	----------

Response Class (Status 200)

OK

[Example Value](#)

{}

[Response Content Type](#) [*/*](#)

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

POST /user

insertUser

Parameters

Parameter	Value	Description	Parameter Type	Data Type
-----------	-------	-------------	----------------	-----------

user	user	body	Example Value
			<pre>{ "created_at": "2023-11-07T03:03:32.868Z", "email": "string", "id": 0, "name": "string", "password": "string", "surname": "string" }</pre>

Response Messages

HTTP Status Code	Reason	Response Model	Headers
200	OK		
201	Created		
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

GET /user/page getPage

Response Class (Status 200)

OK

Example Value

```
{
  "content": [
    {
      "email": "string",
      "name": "string",
      "password": "string",
      "surname": "string"
    }
  ],
  "first": true,
  "last": true
}
```

Response Content Type [/*](#)

Parameters

Parameter	Value	Description	Parameter Type	Data Type
page	<input type="text"/>	page	query	integer

linesPerPage	<input type="text" value="24"/>	linesPerPage	query	integer
orderBy	<input type="text" value="name"/>	orderBy	query	string
direction	<input type="text" value="ASC"/>	direction	query	string

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

GET	/user/{id}	getUser
-----	------------	---------

Response Class (Status 200)

OK

Example Value

```
{}
```

Response Content Type

Parameters

Parameter	Value	Description	Parameter Type	Data Type
id	<input type="text"/>	id	path	integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

PUT

/user/{id}

updateUser

Response Class (Status 200)

OK

Example Value

{}

Response Content Type `/*`

Parameters

Parameter	Value	Description	Parameter Type	Data Type
user	<input type="text"/>	user	body	<h3>Example Value</h3> <pre>{ "created_at": "2023-11-07T03:03:32.870Z", "email": "string", "id": 0, "name": "string", "password": "string", "surname": "string" }</pre>
id	<input type="text"/>	id	path	integer

Response Messages

HTTP Status Code	Reason	Response Model	Headers
201	Created		
401	Unauthorized		
403	Forbidden		
404	Not Found		

[Try it out!](#)

ANEXO B – REGISTRO DE PROGRAMA DE COMPUTADOR CLEAN



REPÚBLICA FEDERATIVA DO BRASIL

MINISTÉRIO DA ECONOMIA

INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL

DIRETORIA DE PATENTES, PROGRAMAS DE COMPUTADOR E TOPOGRAFIAS DE CIRCUITOS INTEGRADOS

Certificado de Registro de Programa de Computador

Processo Nº: **BR512022001116-6**

O Instituto Nacional da Propriedade Industrial expede o presente certificado de registro de programa de computador, válido por 50 anos a partir de 1º de janeiro subsequente à data de 14/03/2020, em conformidade com o §2º, art. 2º da Lei 9.609, de 19 de Fevereiro de 1998.

Título: CLEAN

Data de publicação: 14/03/2020

Data de criação: 14/03/2020

Titular(es): UNIVERSIDADE FEDERAL DE SANTA CATARINA

Autor(es): LEONARDO HOINASKI; FERNANDO CAMPO GARCÍA

Linguagem: C++

Campo de aplicação: IN-02; MA-03; MA-04

Tipo de programa: DS-05; IT-02; TI-03

Algoritmo hash: SHA-512

Resumo digital hash:

b0072c9affcd1042202cf1e5e27bfd9237be9c90a39f31131bd6e9f95327f63dee611aa858cea3d06e1a4b94ddc51701bcd
9e8bb413d069319d1fe9c7e28149e

Expedido em: 17/05/2022

Aprovado por:

Joelson Gomes Pequeno

Chefe Substituto da DIPTO - PORTARIA/INPI/DIRPA Nº 02, DE 10 DE FEVEREIRO DE 2021

**ANEXO C – JUPYTER NOTEBOOKS COM METODOLOGIA DE
PRÉ-PROCESSAMENTO APLICADA ÀS LEITURAS DOS SENsoRES CO-B4,
OX-B431, NO2-B43F E DE MP10 DO OPC-N3**

CO-field-data-treatment-Diamante

February 13, 2024

1 Electrochemical CO sensor data preprocessing

- Pollutant: Carbon Monoxide
- Sensor: Alphasense CO-B4

1.1 Constants

1.1.1 Sensors IDs

1.1.2 API Constants

```
[ ]: HOST = "renovar.lcqar.ufsc.br"
PORT = 8080
RAW_DATA_DIR = "data/raw-data-monit-fix-2022-2023-Diamante/"
RAW_FILE_NAME = "ISB_CO.CSV"

SENSOR_FILE_DIR = 'data/input/'
SENSOR_NAME = 'alpha_co_conc'
SENSOR_FILE_NAME = SENSOR_NAME + 'web_dataframe.csv'
```

1.1.3 Sensor Constants

```
[ ]: M = 28.01
lower_limit = 4.0
upper_limit = 1e6
t_90 = 30      # sensor takes 30 seconds to reach a value of 10e3
t_90_value = 10e3
sampling_period = 15 * 60
```

2 Alphasense CO Sensor Data

```
[ ]: import locale
locale.setlocale(locale.LC_TIME, 'pt_BR')
```

```
[ ]: 'pt_BR'
```

```
[ ]: from GetSensorDataService import GetSensorDataService
```

```

get_sensor_data_service = GetSensorDataService(HOST, PORT)
sensor_data = get_sensor_data_service.
    ↪get_data_from_file(RAW_DATA_DIR+RAW_FILE_NAME, sensor_name=SENSOR_NAME)
sensor_data.head()

/Users/Fernando/Documents/Projects/Github/lcqar-low-cost-monit-proc/data-pre-
processing/GetSensorDataService.py:13: UserWarning: The argument
'infer_datetime_format' is deprecated and will be removed in a future version. A
strict version of it is now the default, see
https://pandas.pydata.org/pdps/0004-consistent-to-datetime-parsing.html. You
can safely remove this argument.
df['DateTime'] = (pd.to_datetime(df[date_time_col],
infer_datetime_format=False, format='%d/%m/%Y/%H/%M/%S'))

[ ]:   latitude longitude measuring           DateTime
0 -28.456899 -48.972999    125.92 2022-11-20 13:45:23
1 -28.456899 -48.972999     86.05 2022-11-21 10:37:45
2 -28.456899 -48.972999     86.38 2022-11-21 10:53:33
3 -28.456899 -48.972999     96.05 2022-11-21 11:09:20
4 -28.456899 -48.972999     84.43 2022-11-21 11:25:07

```

2.1 Upload Data from File

```

[ ]: import pandas as pd

df = pd.read_csv(SENSOR_FILE_DIR + SENSOR_FILE_NAME)
df.head()

[ ]:   Unnamed: 0   latitude   longitude   measuring           DateTime
0          0 -28.456899 -48.972999    125.92 2022-11-20 13:45:23
1          1 -28.456899 -48.972999     86.05 2022-11-21 10:37:45
2          2 -28.456899 -48.972999     86.38 2022-11-21 10:53:33
3          3 -28.456899 -48.972999     96.05 2022-11-21 11:09:20
4          4 -28.456899 -48.972999     84.43 2022-11-21 11:25:07

```

2.1.1 Create Sensor Dataframe as Pandas Series with a period of 15 mins

```

[ ]: # Remove the first column with the indexes and save data into web dataframe
web_dataframe = df.drop(df.columns[0], axis='columns')
web_dataframe['DateTime'] = (pd.to_datetime(df['DateTime']),
    ↪infer_datetime_format=True)

# Resample data with 15 mins period and create sensor dataframe
sensor_dataframe = web_dataframe.sort_values(by='DateTime', ascending=True).
    ↪reset_index().drop(columns='index')
sensor_dataframe.index = sensor_dataframe['DateTime']
sensor_dataframe = sensor_dataframe.drop(columns=['DateTime'])

```

```

sensor_dataframe = sensor_dataframe.resample('15T').mean()
sensor_dataframe

/var/folders/wc/_83zcrx913j1dqwg4g90kbhh0000gp/T/ipykernel_1994/166902210.py:3:
UserWarning: The argument 'infer_datetime_format' is deprecated and will be
removed in a future version. A strict version of it is now the default, see
https://pandas.pydata.org/pdps/0004-consistent-to-datetime-parsing.html. You
can safely remove this argument.
    web_dataframe['DateTime'] = (pd.to_datetime(df['DateTime'],
infer_datetime_format=True))

[ :]:          latitude  longitude  measuring
DateTime
2022-11-20 13:45:00 -28.456899 -48.972999      125.92
2022-11-20 14:00:00        NaN        NaN        NaN
2022-11-20 14:15:00        NaN        NaN        NaN
2022-11-20 14:30:00        NaN        NaN        NaN
2022-11-20 14:45:00        NaN        NaN        NaN
...
2023-05-23 08:15:00        ...        ...        ...
2023-05-23 08:30:00        NaN        NaN        NaN
2023-05-23 08:45:00        NaN        NaN        NaN
2023-05-23 09:00:00        NaN        NaN        NaN
2023-05-23 09:15:00 -28.456899 -48.972999      163.04

[17647 rows x 3 columns]

```

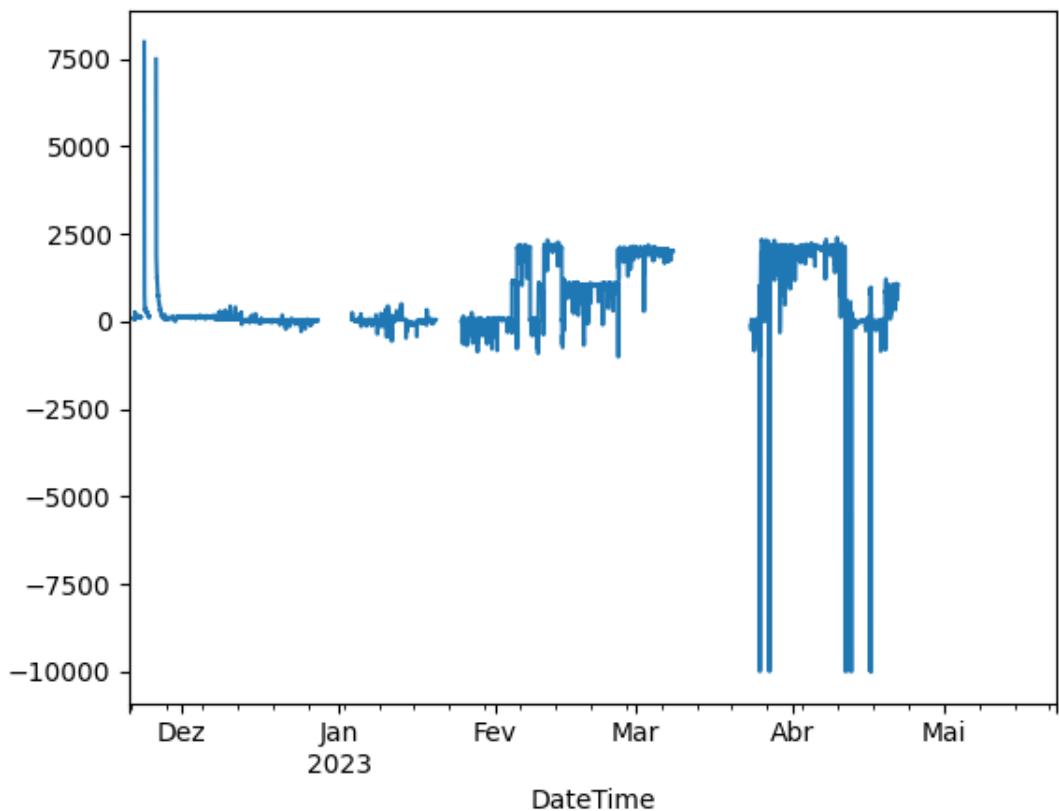
2.1.2 Plot raw data

```

[ :]: sensor_dataframe['measuring'].plot()

[ :]: <Axes: xlabel='DateTime'>

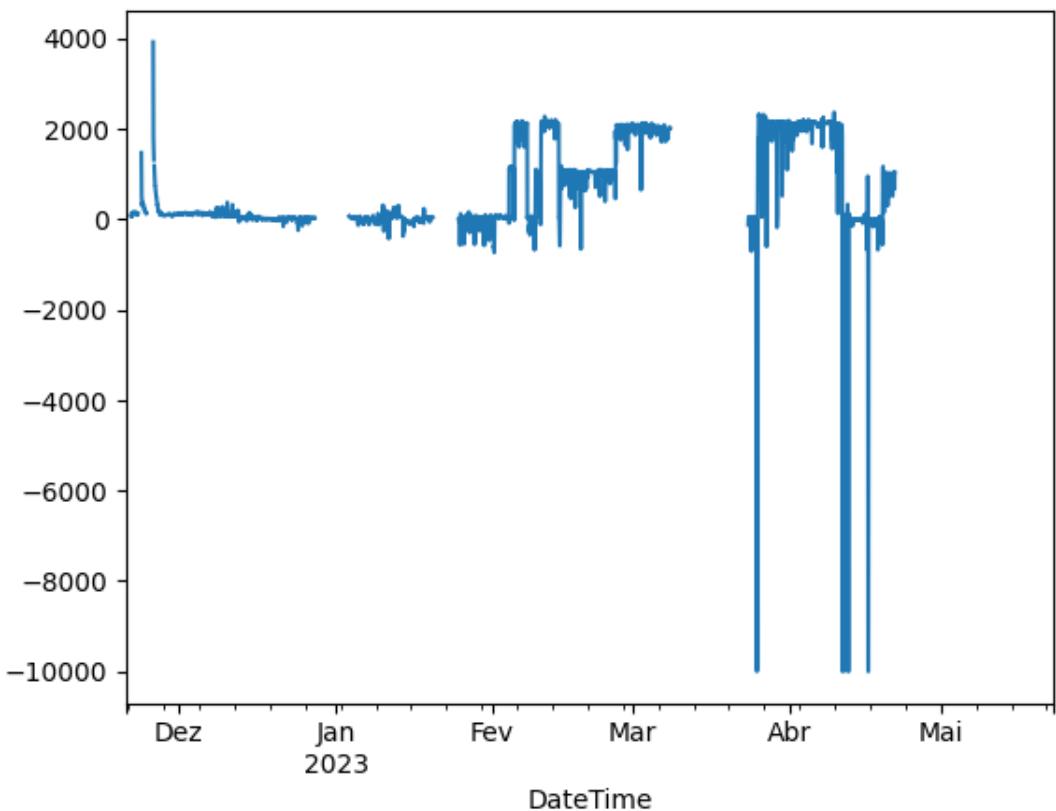
```



2.1.3 Smooth data with a window of 4 samples, i.e.: an hour of data

```
[ ]: sensor_dataframe['raw_measuring'] = sensor_dataframe['measuring']
sensor_dataframe['measuring'] = sensor_dataframe['measuring'].rolling(window=4).
    median()
sensor_dataframe['Hour'] = sensor_dataframe.index.hour
sensor_dataframe['measuring'].plot()
```

```
[ ]: <Axes: xlabel='DateTime'>
```



2.1.4 Convert to ppm

```
[ ]: sensor_dataframe['value'] = sensor_dataframe['measuring'].map(lambda v: v/1e3)
```

2.2 Tag data

Apply tags to the data according to the quality control processes described above

```
[ ]: sensor_dataframe['Tag'] = 'VALID'
```

Tag data during stabilizing period AQMesh recommends to considering the first two days after installation as a period for stabilization. In our case, a period of seven days was selected in order to remove outliers candidates detected during that period. For that purpose we detect the first 7 days of uninterrupted measurements

```
[ ]: # Identify consecutive measurements (15 mins between each)
consecutive_periods = sensor_dataframe.index.to_series().diff().dt.seconds == 900
```

```

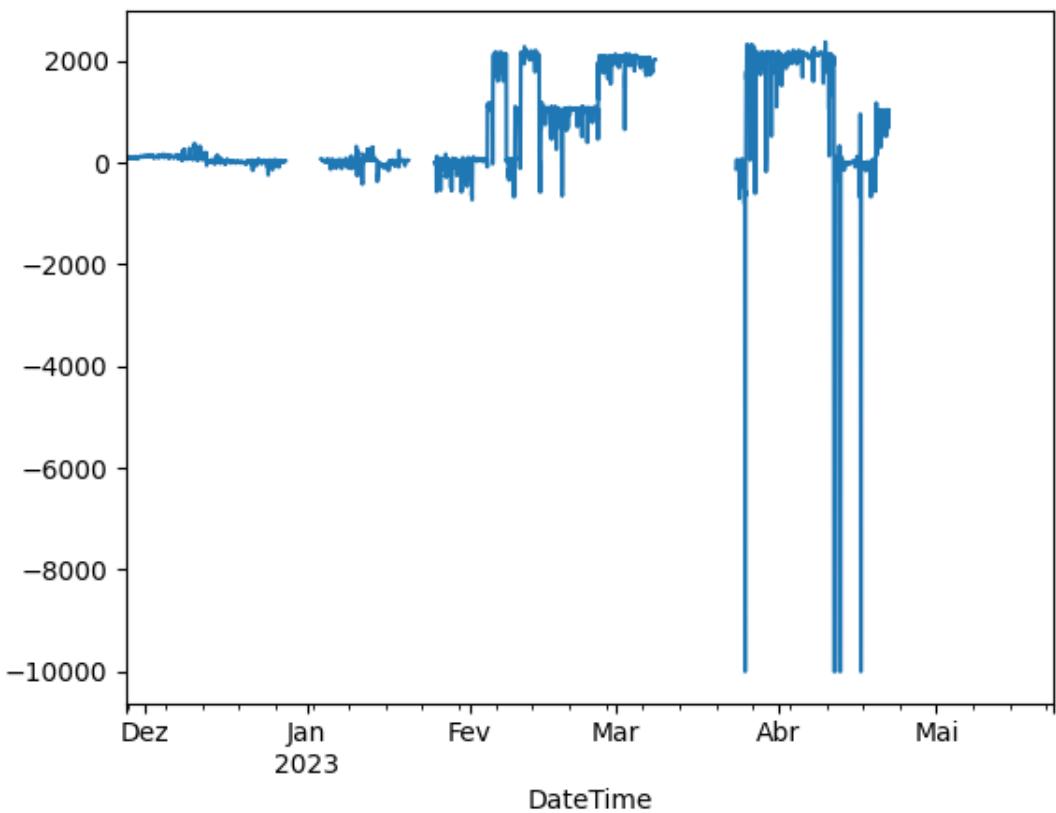
# Select the first seven days of consecutive measurements (96 = 24 * 60 mins / ↳
↪15 mins, amount of 15 mins periods in a day)
last_stabilizing_index = sensor_dataframe[consecutive_periods].head(96 * 7).
↪index[-1] # 7 days
sensor_dataframe.loc[sensor_dataframe.index[0]:last_stabilizing_index, 'Tag'] = ↳
↪'STABILIZING'
sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['measuring'].
↪resample('15T').mean().plot()
sensor_dataframe.head()

```

```
[ ]:          latitude longitude measuring raw measuring Hour \
DateTime
2022-11-20 13:45:00 -28.456899 -48.972999      NaN      125.92    13
2022-11-20 14:00:00      NaN      NaN      NaN      NaN      14
2022-11-20 14:15:00      NaN      NaN      NaN      NaN      14
2022-11-20 14:30:00      NaN      NaN      NaN      NaN      14
2022-11-20 14:45:00      NaN      NaN      NaN      NaN      14

           value        Tag
DateTime
2022-11-20 13:45:00      NaN  STABILIZING
2022-11-20 14:00:00      NaN  STABILIZING
2022-11-20 14:15:00      NaN  STABILIZING
2022-11-20 14:30:00      NaN  STABILIZING
2022-11-20 14:45:00      NaN  STABILIZING

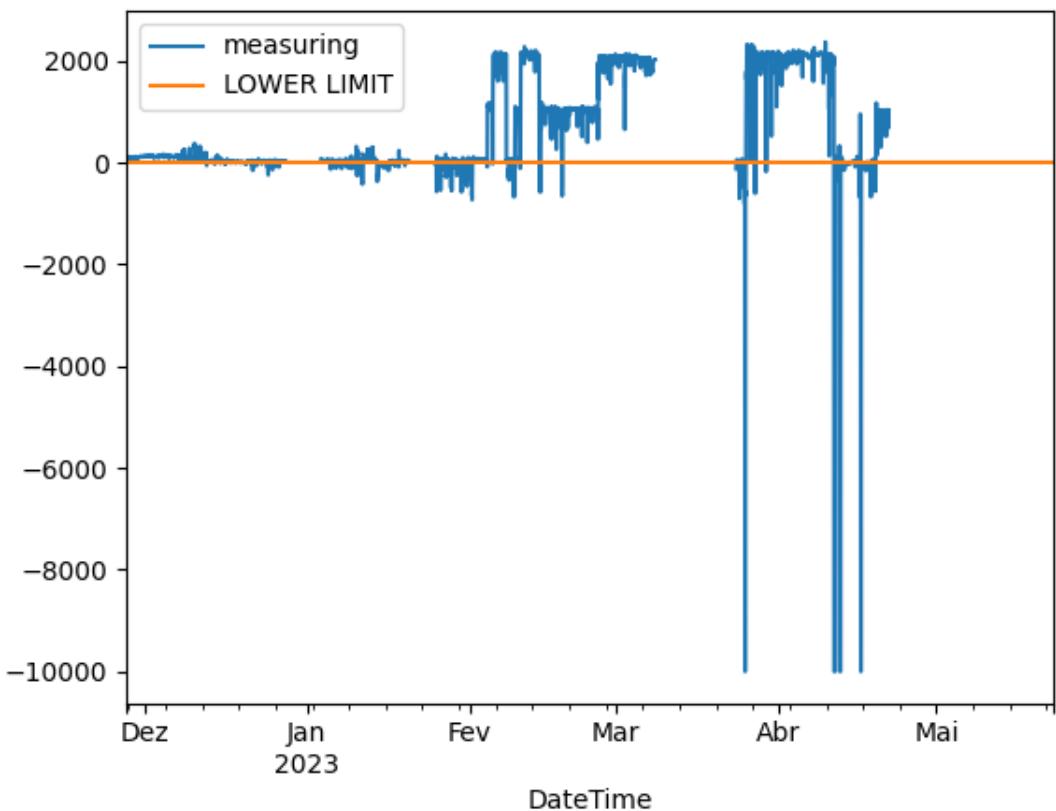
```



Tag values that are missing, are greater than sensor span or are lower than sensor resolution The upper limit was too high that affected the graph resolution, therefore it is not plotted

```
[ ]: sensor_dataframe['UPPER LIMIT'] = upper_limit
sensor_dataframe['LOWER LIMIT'] = lower_limit
sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'][['measuring', 'LOWER LIMIT']].resample('15T').mean().plot()
```

```
[ ]: <Axes: xlabel='DateTime'>
```



```
[ ]: import numpy as np
from SensorDataAnalysisService import SensorDataAnalysisService as sensor_analysis

def tag_by_sensor_limits(value, tag, lower_limit, upper_limit):
    if (tag != 'VALID'): return tag
    return sensor_analysis.get_tags_from_series(value=value,
                                                lower_limit=lower_limit,
                                                upper_limit=upper_limit)

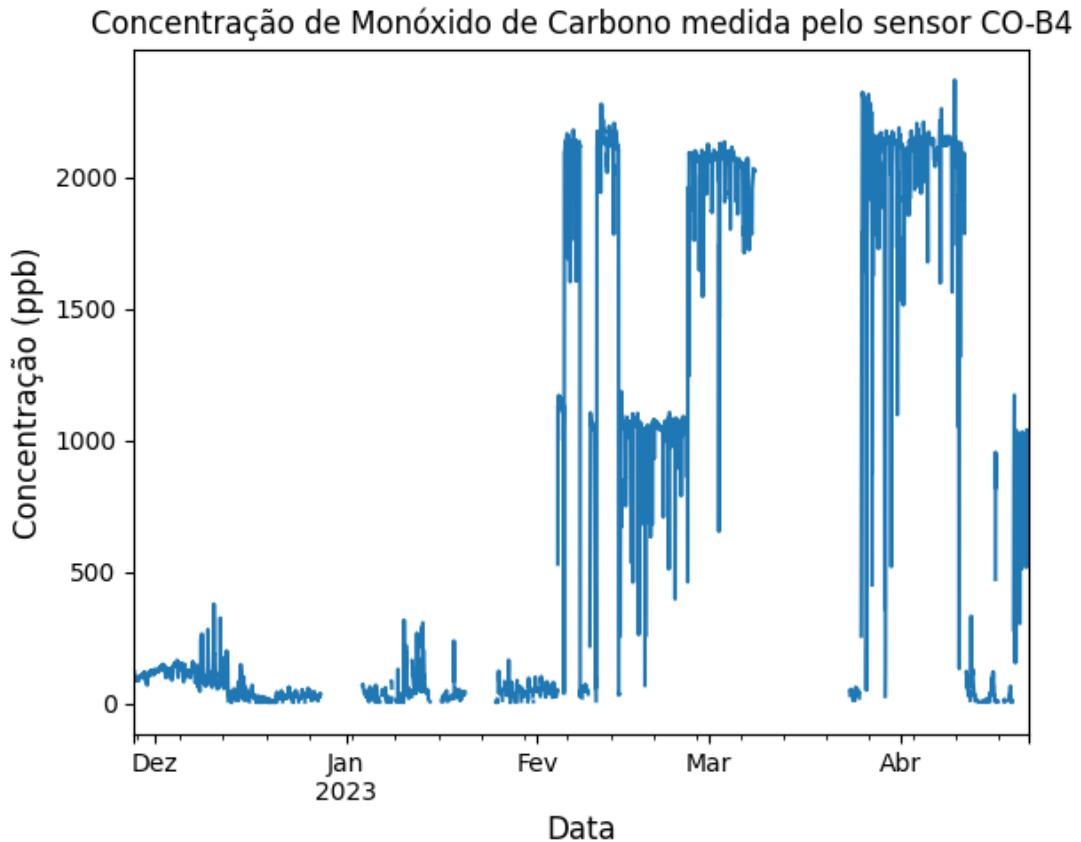
sensor_dataframe['Tag'] = (sensor_dataframe[['measuring', 'Tag']]
                           .apply(lambda df: tag_by_sensor_limits(value=df[0], tag=df[1],
                           lower_limit=lower_limit,
                           upper_limit=upper_limit),
                           axis=1))
```

Plot valid data after removing samples bellow lower limit

```
[ ]: import matplotlib.pyplot as plt

fig, ax = plt.subplots(figsize=(1.3*5,5))
sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['measuring'].
    ↪rename('Concentração de Monóxido de Carbono').resample('15T').mean().plot()
plt.title('Concentração de Monóxido de Carbono medida pelo sensor CO-B4')
ax.set_xlabel('Data', fontsize=12)
ax.set_ylabel('Concentração (ppb)', fontsize=12)
```

```
[ ]: Text(0, 0.5, 'Concentração (ppb)')
```



2.2.1 Data Tag contabilization

```
[ ]: tags = ['MISSING', 'LTLL', 'GTUL', 'STABILIZING', 'BADSPIKE', 'VALID']
data_contabilization = sensor_analysis.count_tags(tags, sensor_dataframe)
data_contabilization
```

```
[ ]:          #      %
MISSING      5756  32.617442
```

LTLL	1560	8.840029
GTUL	0	0.0
STABILIZING	673	3.813679
BADSPIKE	0	0.0
VALID	9658	54.728849
TOTAL	17647	100.0

2.3 Analyse valid data

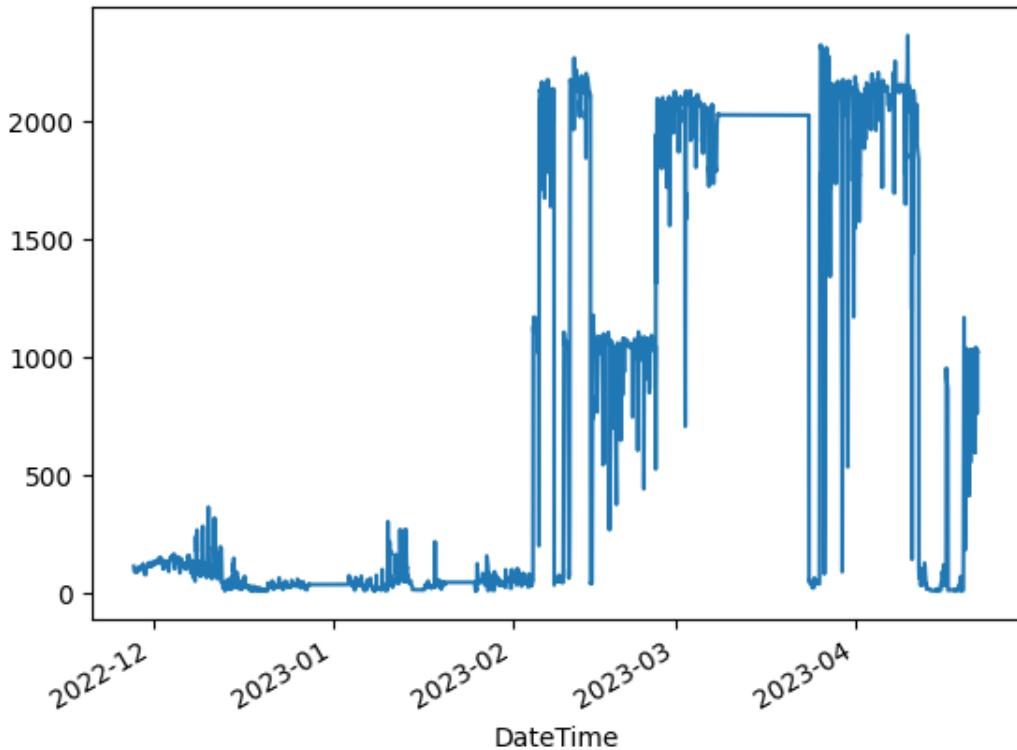
The graph shows the time series of the valid data, box plots of the data grouped by hour of the day, and histogram of the data.

```
[ ]: valid_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'].
    drop(columns=['Tag'])
```

2.3.1 Smooth data with a window of 4 samples, i.e.: an hour of data

```
[ ]: valid_dataframe['raw_measuring'] = valid_dataframe['measuring']
valid_dataframe['measuring'] = valid_dataframe['measuring'].rolling(window=4).
    median()
valid_dataframe['measuring'].plot()
```

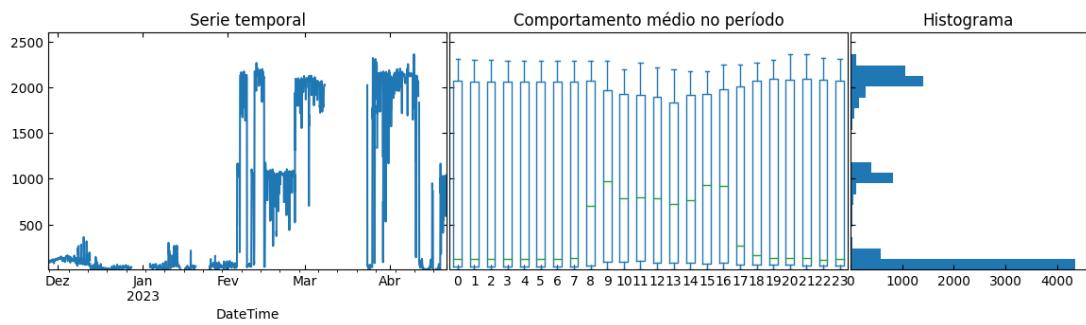
```
[ ]: <Axes: xlabel='DateTime'>
```



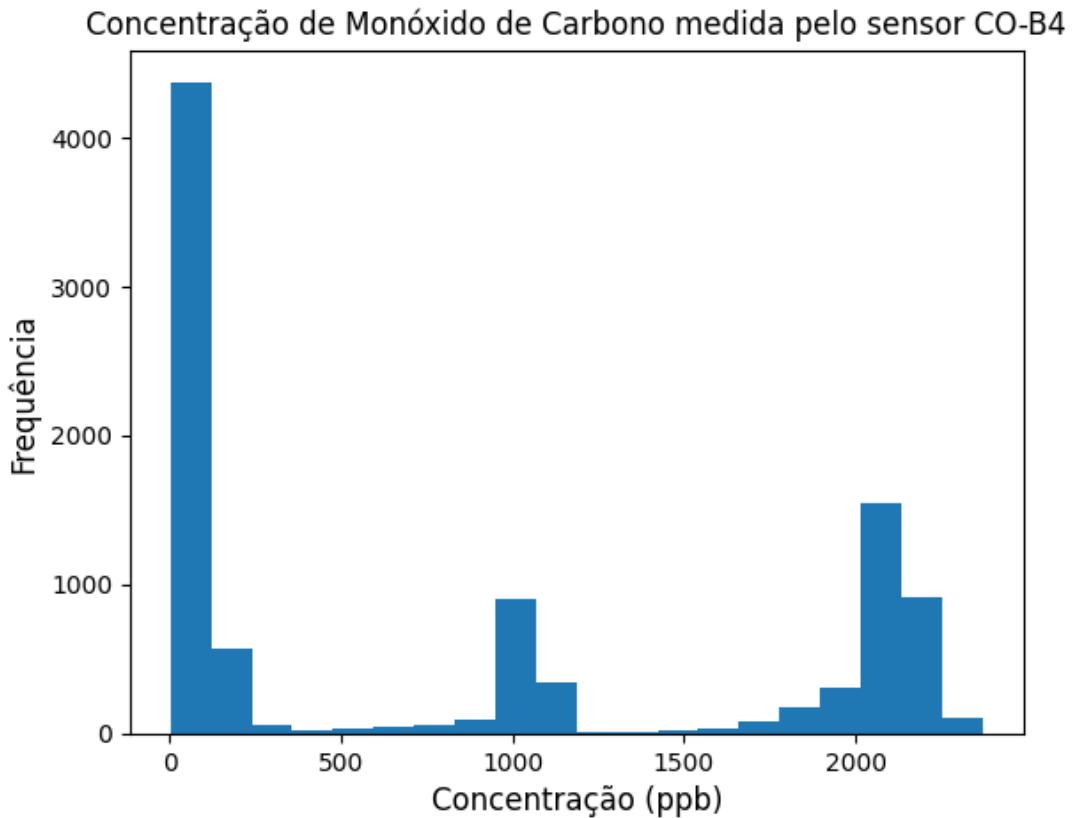
```
[ ]: valid_dataframe = valid_dataframe.resample('15T').mean()
sensor_analysis.plot_box_hist(df=valid_dataframe, bins=20)

/Users/Fernando/Documents/Projects/Github/lcqar-low-cost-monit-proc/data-pre-
processing/SensorDataAnalysisService.py:111: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
df['Hour'] = df['Hour'].astype('int64')
```



```
[ ]: fig, ax = plt.subplots(figsize=(1.3*5,5))
(sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['measuring'].
    rename('Concentração de Monóxido de Carbono')
    .resample('15T').mean().hist(bins=20))
plt.title('Concentração de Monóxido de Carbono medida pelo sensor CO-B4')
ax.set_xlabel('Concentração (ppb)', fontsize=12)
ax.set_ylabel('Frequência', fontsize=12)
ax.grid(False)
```



2.4 Change point Analysis

2.4.1 Change point analysis of 15 mins data

```
[ ]: import ruptures as rpt

series = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['measuring']
signal = np.array(series).reshape(-1, 1)
algo = rpt.Pelt(model="rbf", jump=50, min_size=100).fit(signal=signal)
result = algo.predict(pen=10)
result
```

```
[ ]: [4400, 4500, 4750, 4850, 4950, 5300, 6350, 7350, 7500, 9100, 9400, 9658]
```

```
[ ]: change_point_index = series[[x - 1 for x in result]].index
sensor_dataframe['CHANGE POINT'] = False
sensor_dataframe['CHANGE POINT'].loc[change_point_index] = True
```

```
/var/folders/wc/_83zcrx913j1dqwg4g90kbhh0000gp/T/ipykernel_1994/717246363.py:3:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame
```

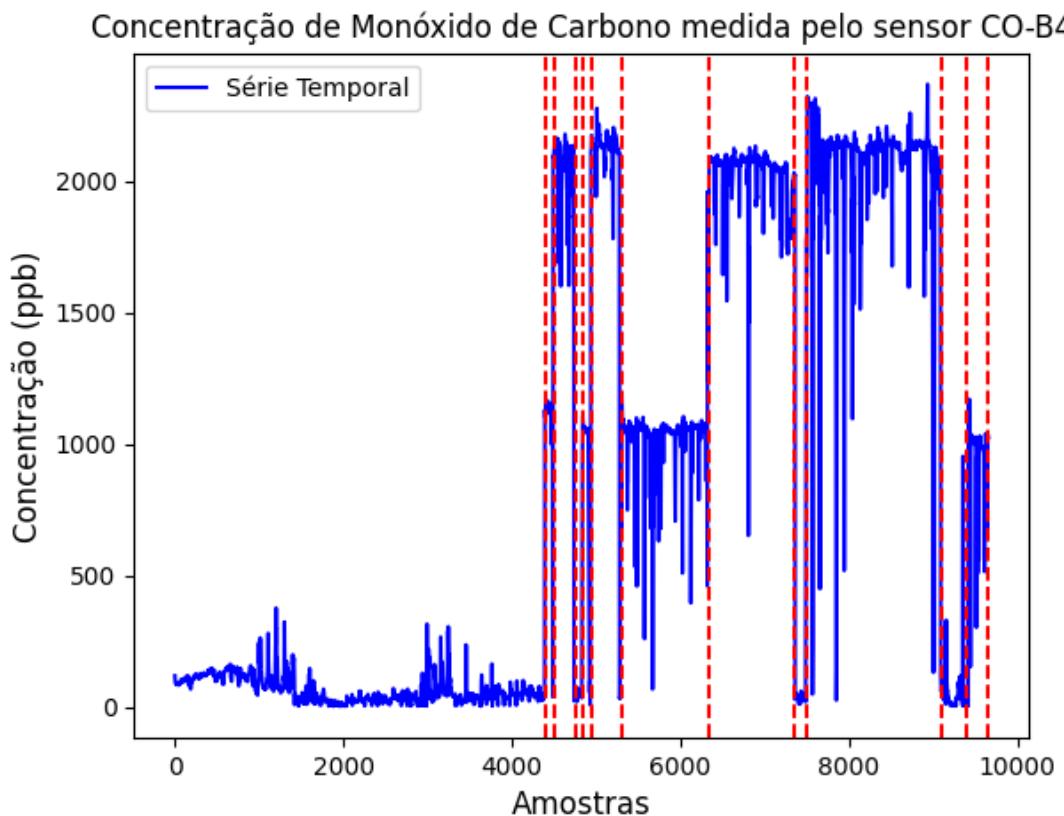
See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
sensor_dataframe['CHANGE POINT'].loc[change_point_index] = True

```
[ ]: import matplotlib.pyplot as plt

fig, ax = plt.subplots(figsize=(1.3*5,5))
plt.plot(signal, label='Série Temporal', color='blue')
plt.title('Concentração de Monóxido de Carbono medida pelo sensor CO-B4')
ax.set_xlabel('Amostras', fontsize=12)
ax.set_ylabel('Concentração (ppb)', fontsize=12)

# Plot change points
for point in result:
    plt.axvline(x=point, color='red', linestyle='--')

# Show the plot
plt.legend()
plt.show()
```



Tag dataframe with changepoints

```
[ ]: sensor_dataframe.loc[change_point_index[0]:, 'Tag'] = (sensor_dataframe.  
    →loc[change_point_index[0]:, 'Tag']  
        .apply(lambda t:  
            ↪'REBASE' if t == 'VALID' else t))  
sensor_dataframe[sensor_dataframe['Tag'] == 'REBASE'].head()
```

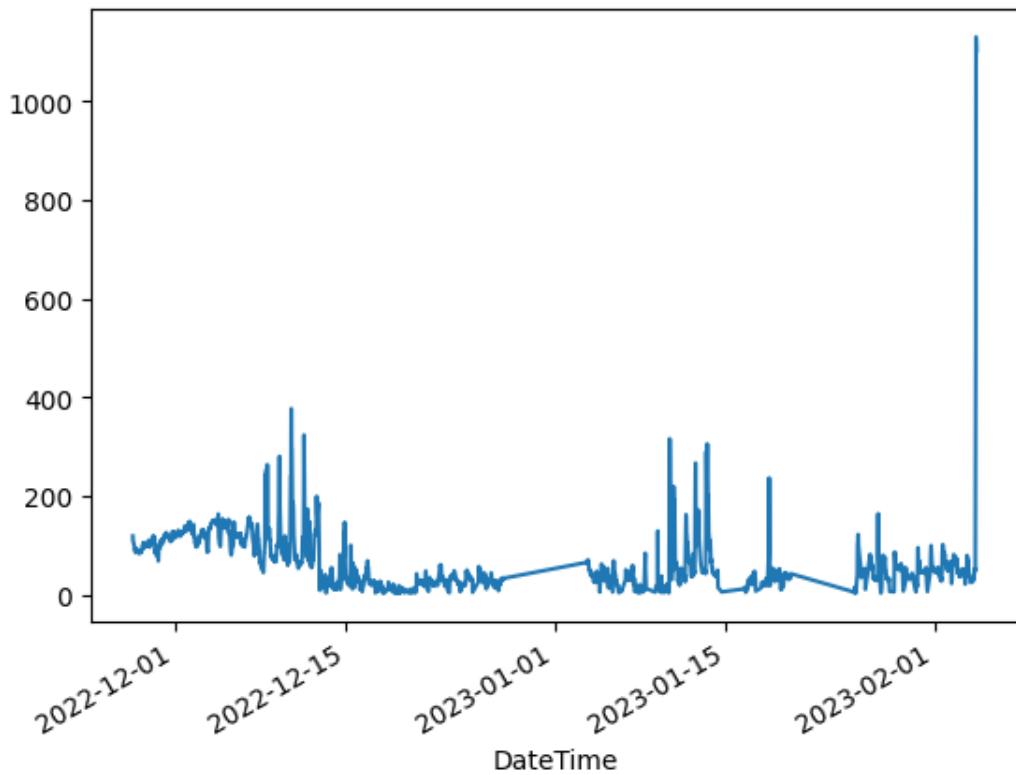
```
[ ]:          latitude  longitude  measuring  raw measuring  Hour \  
DateTime  
2023-02-04 11:15:00 -28.456899 -48.972999  1108.640      1109.40   11  
2023-02-04 11:30:00 -28.456899 -48.972999  1110.445      1112.44   11  
2023-02-04 11:45:00 -28.456899 -48.972999  1110.445      1104.39   11  
2023-02-04 12:00:00 -28.456899 -48.972999  1110.920      1113.96   12  
2023-02-04 12:15:00 -28.456899 -48.972999  1113.200      1121.73   12  
  
          value     Tag  UPPER LIMIT  LOWER LIMIT  CHANGE POINT  
DateTime  
2023-02-04 11:15:00  1.108640  REBASE  1000000.0       4.0      True  
2023-02-04 11:30:00  1.110445  REBASE  1000000.0       4.0     False  
2023-02-04 11:45:00  1.110445  REBASE  1000000.0       4.0     False  
2023-02-04 12:00:00  1.110920  REBASE  1000000.0       4.0     False  
2023-02-04 12:15:00  1.113200  REBASE  1000000.0       4.0     False
```

Data Tag contabilization

```
[ ]: tags = ['MISSING', 'LTLL', 'GTUL', 'STABILIZING', 'BADSPIKE', 'VALID', 'REBASE']  
data_contabilization = sensor_analysis.count_tags(tags, sensor_dataframe)  
data_contabilization
```

```
[ ]:          #      %  
MISSING      5756  32.617442  
LTLL         1560  8.840029  
GTUL          0    0.0  
STABILIZING   673  3.813679  
BADSPIKE      0    0.0  
VALID         4399  24.92775  
REBASE        5259  29.801099  
TOTAL         17647 100.0
```

```
[ ]: sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['measuring'].plot()  
valid_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']
```



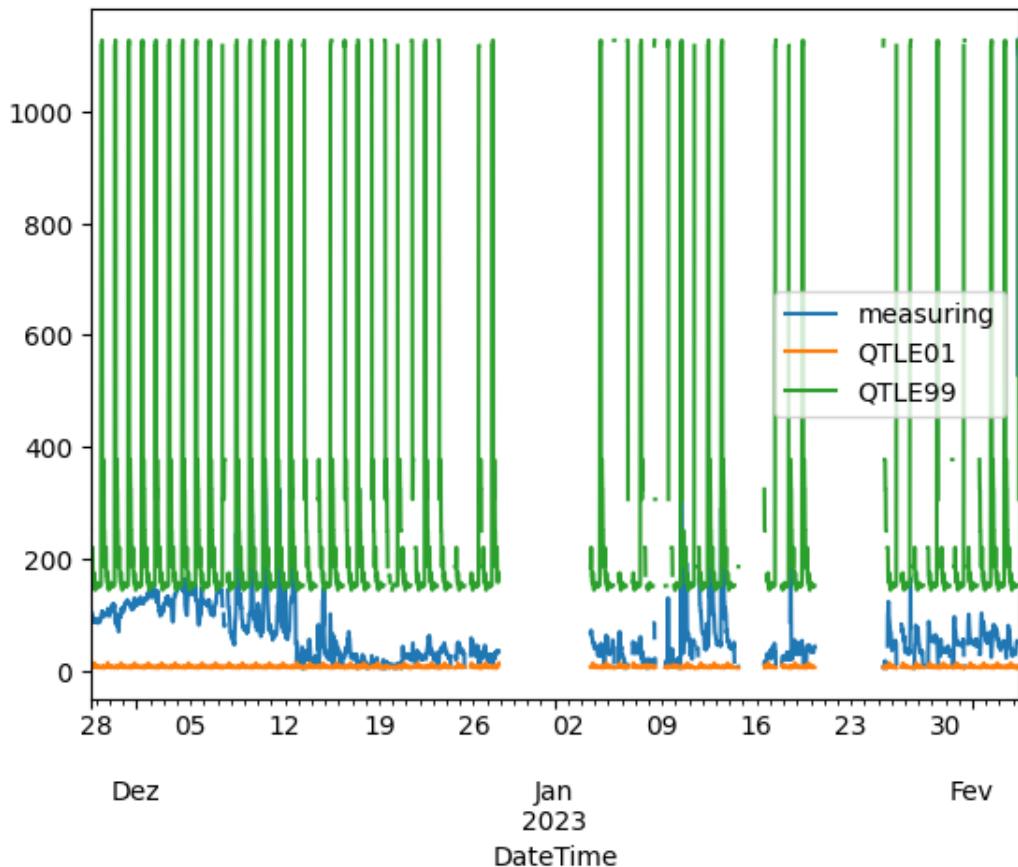
2.4.2 Calculate quantiles

We calculated the 1% and 99 % quantiles of every hour of the day. The values greater outside the 1 % - 99 % where tagged as Greater than Quantile 99 (GTQTL99) and Lower than Quantile 1 (LTQTL01)

```
[ ]: import numpy as np

global_qtle_01 = valid_dataframe.pivot(columns='Hour')['measuring'].
    ↪quantile(q=0.01, axis='index', interpolation='lower').dropna()
global_qtle_99 = valid_dataframe.pivot(columns='Hour')['measuring'].
    ↪quantile(q=0.99, axis='index', interpolation='higher').dropna()
sensor_dataframe['QTLE01'] = sensor_dataframe['Hour'].map(lambda hr: ↴
    ↪global_qtle_01[hr] if (not np.isnan(hr)) else np.nan)
sensor_dataframe['QTLE99'] = sensor_dataframe['Hour'].map(lambda hr: ↴
    ↪global_qtle_99[hr] if (not np.isnan(hr)) else np.nan)
sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'][['measuring', 'QTLE01', ↴
    ↪'QTLE99']].resample('15T').mean().plot()

[ ]: <Axes: xlabel='DateTime'>
```



2.4.3 Tag data according to quantiles

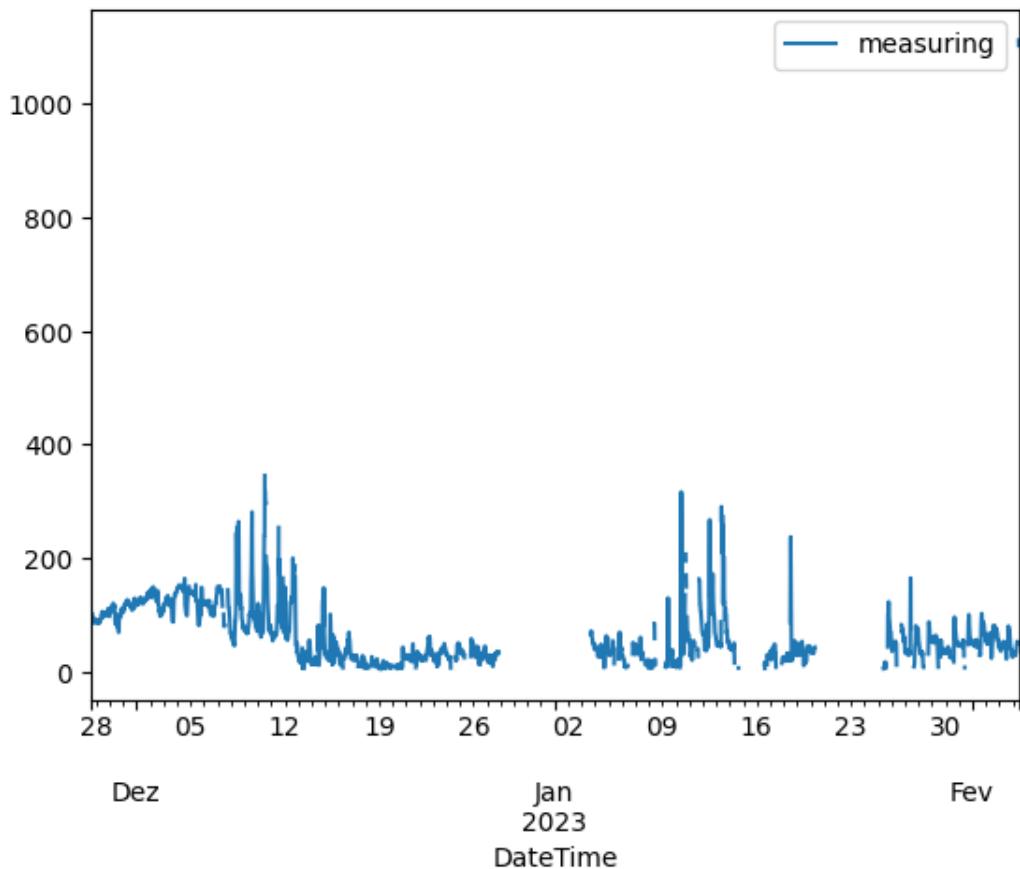
```
[ ]: sensor_dataframe['Tag'] = (sensor_dataframe[['Tag', 'measuring', 'QTLE01', 'QTLE99']]
    .apply(lambda df: sensor_analysis.
        tag_by_quantiles(current_tag=df[0],
        value=df[1],
        quantile_01=df[2],
        quantile_99=df[3]),
        axis=1))
tags = ['MISSING', 'LTLL', 'GTUL', 'BADSPIKE', 'VALID', 'LTQTLE01', 'GTQTLE99',
    'REBASE']
data_contabilization = sensor_analysis.count_tags(tags, sensor_dataframe)
data_contabilization
```

```
[ ]:          #      %
MISSING     5756  32.617442
LTLL        1560   8.840029
GTUL         0     0.0
BADSPIKE     0     0.0
VALID       4273  24.213747
LTQTLE01     63    0.357001
GTQTLE99     63    0.357001
REBASE      5259  29.801099
TOTAL       17647 100.0
```

Plot valid data

```
[ ]: sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'][['measuring']].  
    ↪resample('15T').mean().plot()
```

```
[ ]: <Axes: xlabel='DateTime'>
```

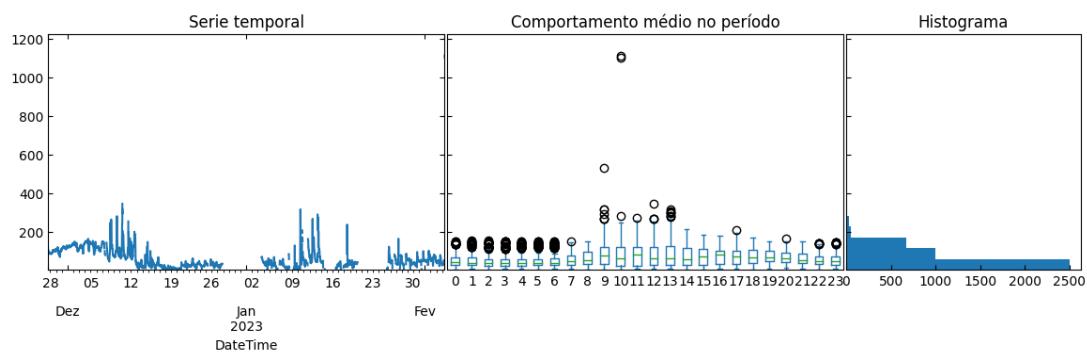


Analyse data after removing quantiles

```
[ ]: valid_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'].
    drop(columns=['Tag']).resample('15T').mean()
sensor_analysis.plot_box_hist(df=valid_dataframe, bins=20)

/Users/Fernando/Documents/Projects/Github/lcqar-low-cost-monit-proc/data-pre-
processing/SensorDataAnalysisService.py:111: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
`df['Hour'] = df['Hour'].astype('int64')`



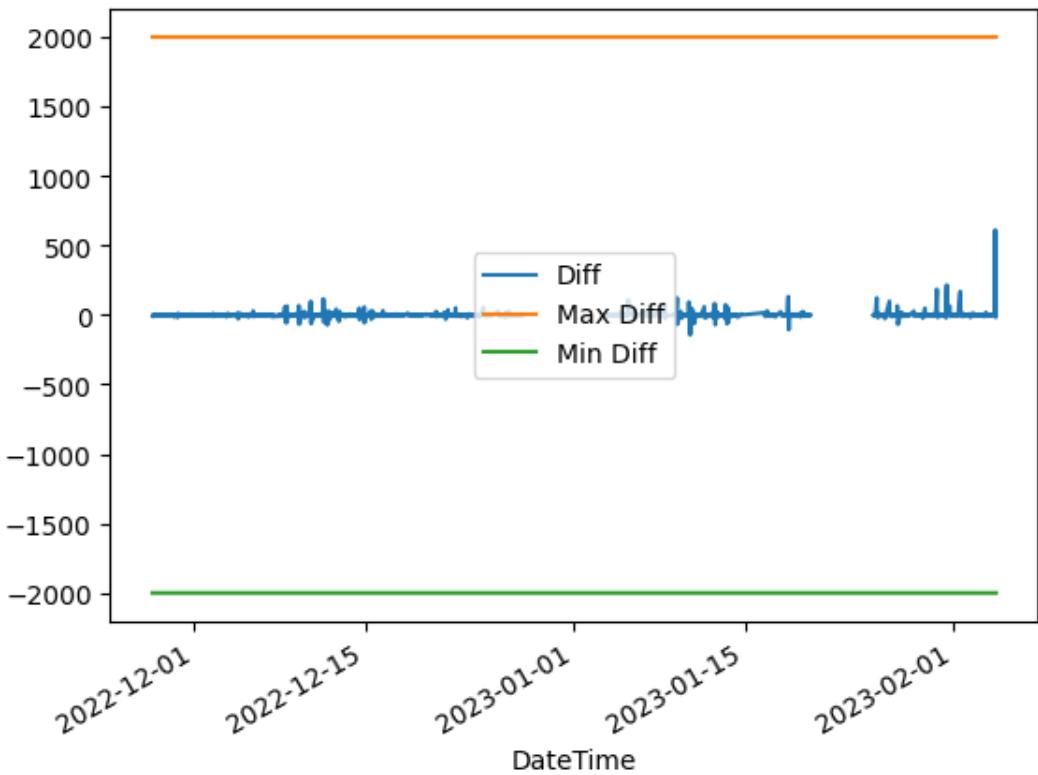
2.5 Analyse data derivatives

Analyse the derivatives of the data for removing abrupt changes in the data (BADSPIKE)

Tag values with derivatives faster than sensor response time The value of 2000 ppb was selected based on the maximum derivative found in the reference data. The maximum derivative (with measuring period of 1 hour) found was 4200 ppb. Therefore, a maximum of 4000 ppb was considered appropriate for a 15 mins period.

```
[ ]: max_diff_value = 2000
sensor_dataframe['Diff'] = sensor_dataframe['measuring'].resample('15T').mean().
    diff()
sensor_dataframe['Max Diff'] = max_diff_value
sensor_dataframe['Min Diff'] = -max_diff_value
sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'][['Diff', 'Max Diff', 'Min Diff']].plot()

[ ]: <Axes: xlabel='DateTime'>
```



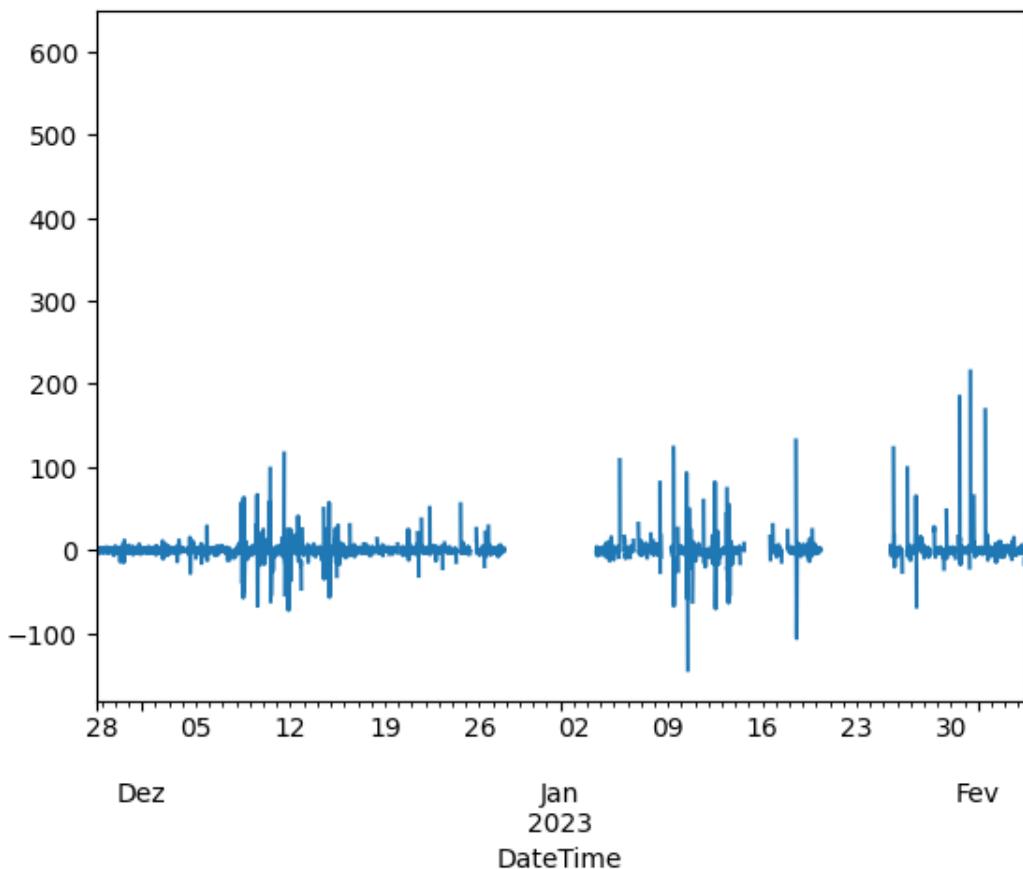
```
[ ]: import numpy as np

def tag_data_with_derivatives(tagged_df, max_diff_value):
    current_tag = tagged_df[0]
    value = tagged_df[1]
    if ((current_tag != 'VALID') or (np.isnan(value))): return current_tag
    if ((value > max_diff_value) or (value < -max_diff_value)): return u
    ↪'BADSPIKE'
    return 'VALID'

sensor_dataframe['Tag'] = (sensor_dataframe[['Tag', 'Diff', 'Max Diff']]
                           .apply(lambda df: u
    ↪tag_data_with_derivatives(tagged_df=df, max_diff_value=df[2]),
                           axis=1))

[ ]: sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['Diff'].resample('15T').
    ↪mean().plot()

[ ]: <Axes: xlabel='DateTime'>
```



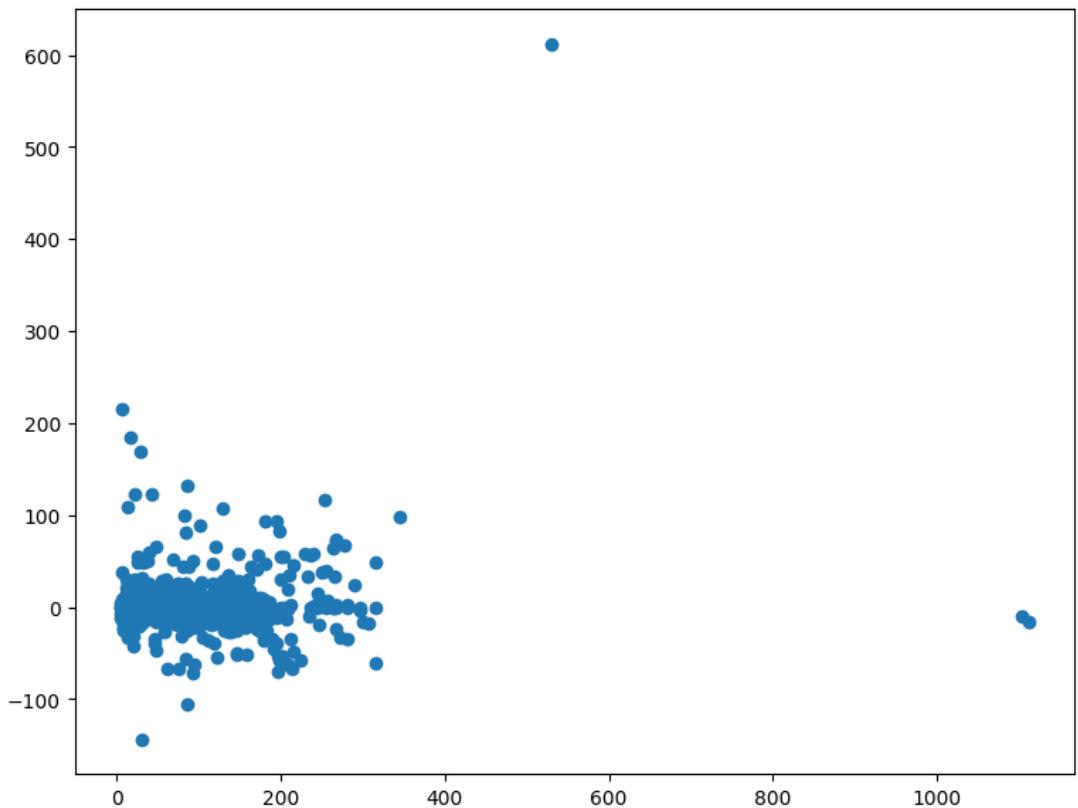
Plot data derivatives vs data

```
[ ]: valid_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']
```

```
[ ]: import matplotlib.pyplot as plt
```

```
fig = plt.figure(figsize=(1.3*7,7))
plt.scatter(valid_dataframe['measuring'], valid_dataframe['Diff'])
```

```
[ ]: <matplotlib.collections.PathCollection at 0x294b25fd0>
```

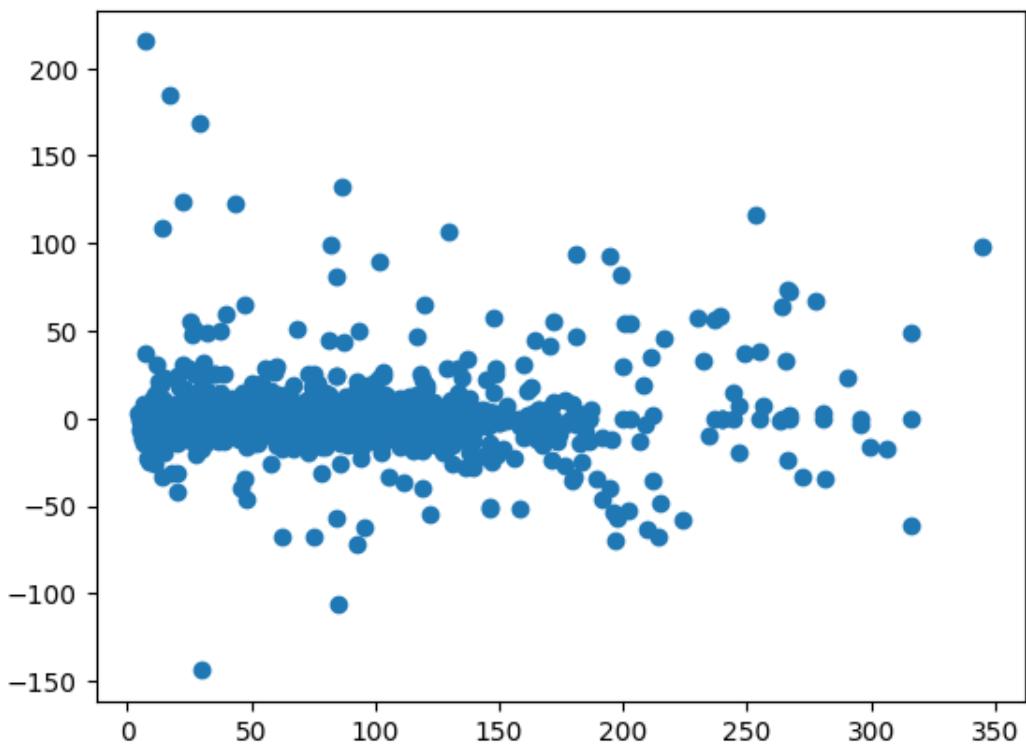


Remove sample with derivative 600 and value 1000

```
[ ]: sensor_dataframe['Tag'] = (sensor_dataframe[['measuring', 'Diff', 'Tag']]
                                .apply(lambda df: 'BADSPIKE' if df[2] == 'VALID' and
                                ↪(df[1] >= 600 or df[0] >= 1000) else df[2], axis=1))

# sensor_dataframe.loc[sensor_dataframe['Diff'] >= 600, 'Tag'] = 'BADSPIKE'
# sensor_dataframe.loc[sensor_dataframe['measuring'] >= 1000, 'Tag'] = ↪'BADSPIKE'

plt.scatter(sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['measuring'],
            sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['Diff'])
valid_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']
```



Data Tag contabilization

```
[ ]: tags = ['MISSING', 'LTLL', 'GTUL', 'STABILIZING', 'BADSPIKE', 'VALID',  
          'LTQTLE01', 'GTQTLE99', 'REBASE']  
data_contabilization = sensor_analysis.count_tags(tags, sensor_dataframe)  
data_contabilization
```

	#	%
MISSING	5756	32.617442
LTLL	1560	8.840029
GTUL	0	0.0
STABILIZING	673	3.813679
BADSPIKE	3	0.017
VALID	4270	24.196747
LTQTLE01	63	0.357001
GTQTLE99	63	0.357001
REBASE	5259	29.801099
TOTAL	17647	100.0

Plot valid data

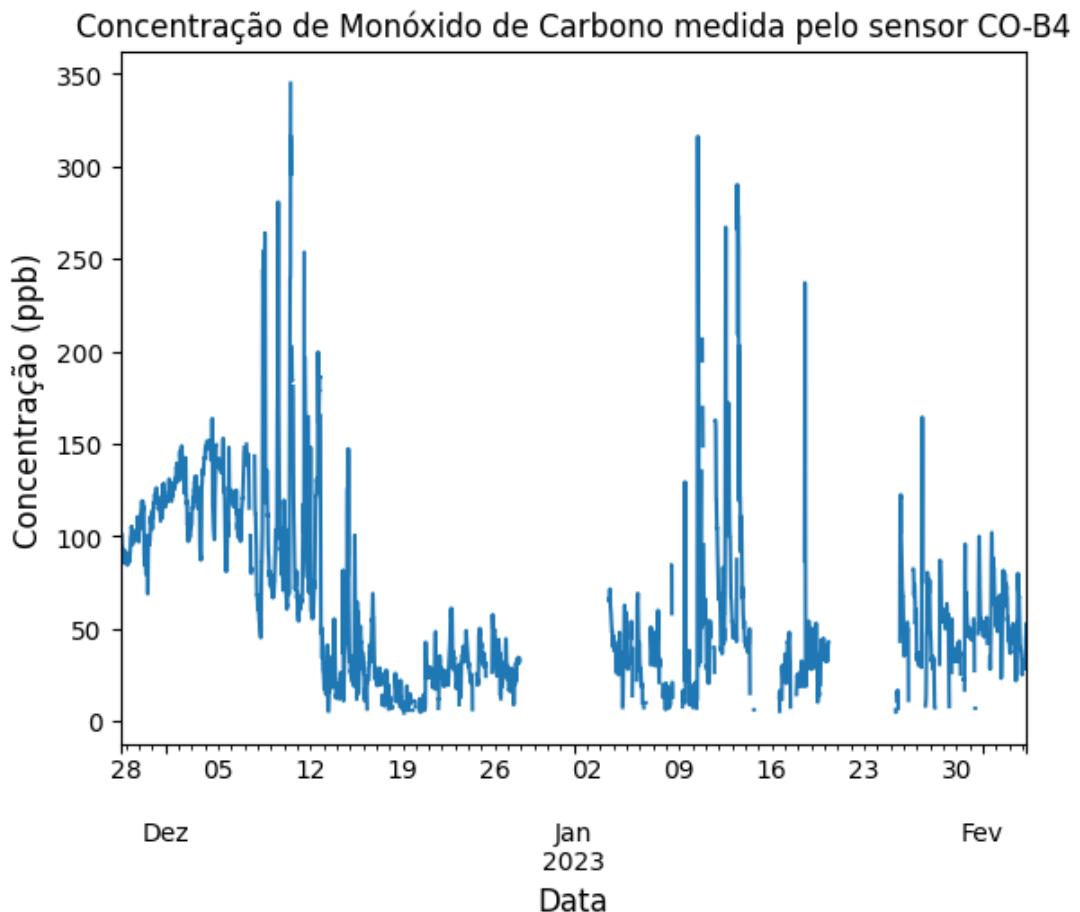
```
[ ]: fig, ax = plt.subplots(figsize=(1.3*5,5))
```

```

sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['measuring'].
    rename('Concentração de Monóxido de Carbono').resample('15T').mean().plot()
plt.title('Concentração de Monóxido de Carbono medida pelo sensor CO-B4')
ax.set_xlabel('Data', fontsize=12)
ax.set_ylabel('Concentração (ppb)', fontsize=12)

[ ]: Text(0, 0.5, 'Concentração (ppb)')

```



```

[ ]: valid_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'].
    drop(columns=['Tag']).resample('15T').mean()
sensor_analysis.plot_box(df=valid_dataframe)

```

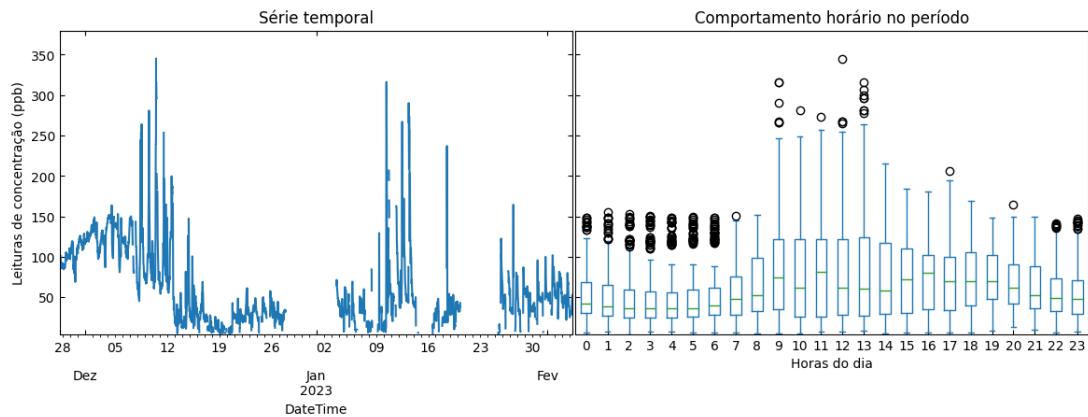
```

/Users/Fernando/Documents/Projects/Github/lcqar-low-cost-monit-proc/data-pre-
processing/SensorDataAnalysisService.py:143: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

```

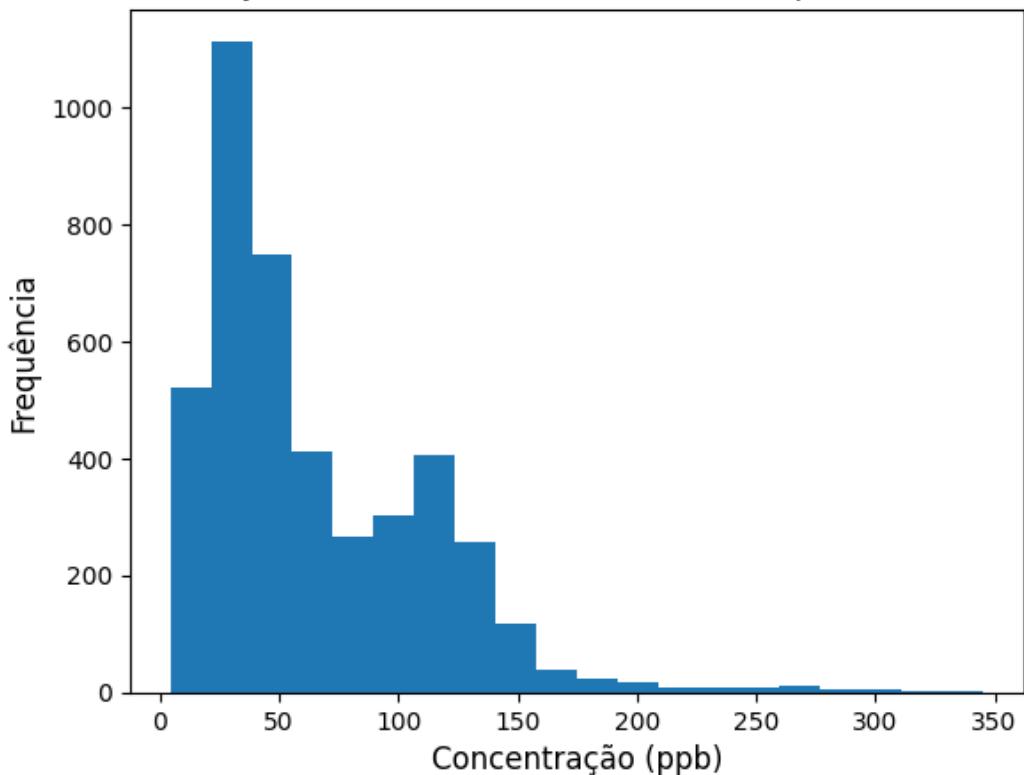
See the caveats in the documentation: <https://pandas.pydata.org/pandas->

```
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
df['Hour'] = df['Hour'].astype('int64')
```



```
[ ]: fig, ax = plt.subplots(figsize=(1.3*5,5))
(sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['measuring'].
    rename('Concentração de Monóxido de Carbono')
    .resample('15T').mean().hist(bins=20))
plt.title('Concentração de Monóxido de Carbono medida pelo sensor CO-B4')
ax.set_xlabel('Concentração (ppb)', fontsize=12)
ax.set_ylabel('Frequência', fontsize=12)
ax.grid(False)
```

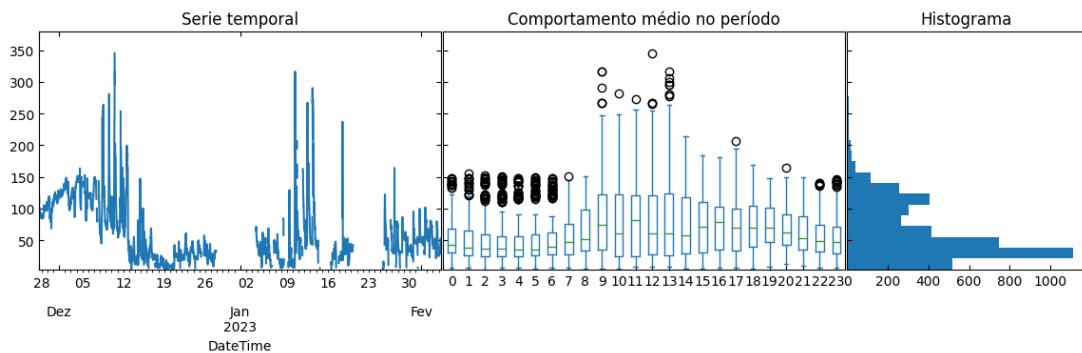
Concentração de Monóxido de Carbono medida pelo sensor CO-B4



```
[ ]: valid_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'].  
      .drop(columns=['Tag']).resample('15T').mean()  
sensor_analysis.plot_box_hist(df=valid_dataframe, bins=20)
```

```
/Users/Fernando/Documents/Projects/Github/lcqar-low-cost-monit-proc/data-pre-  
processing/SensorDataAnalysisService.py:111: SettingWithCopyWarning:  
A value is trying to be set on a copy of a slice from a DataFrame.  
Try using .loc[row_indexer,col_indexer] = value instead
```

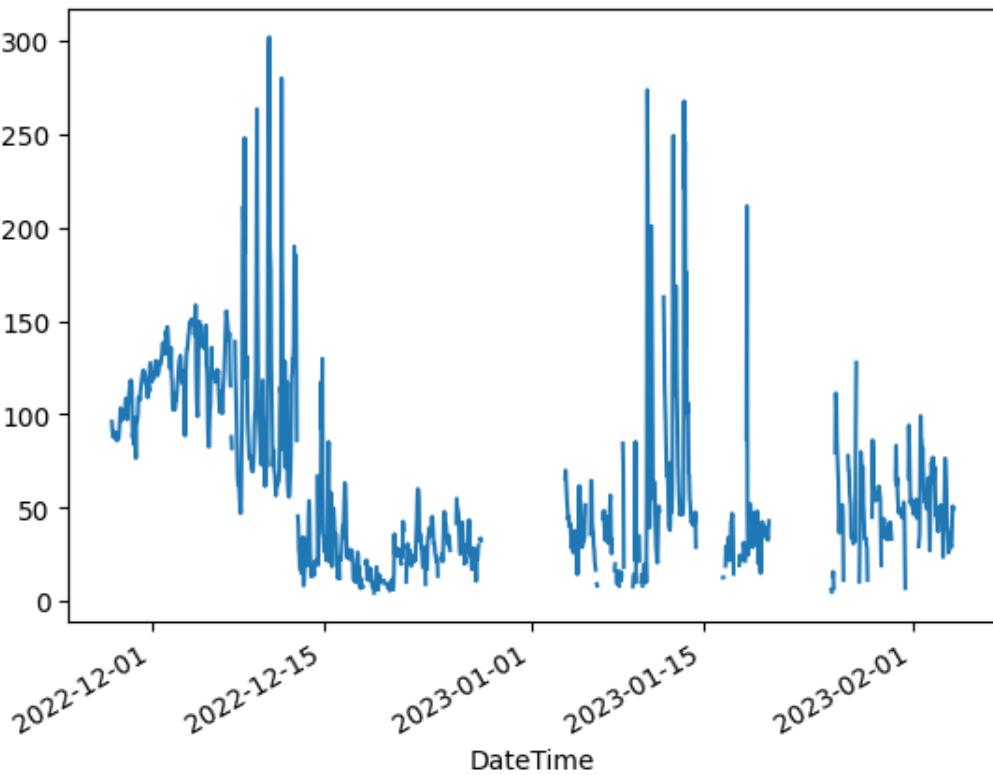
```
See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user\_guide/indexing.html#returning-a-view-versus-a-copy  
df['Hour'] = df['Hour'].astype('int64')
```



2.6 Resample valid data to 1 HR

```
[ ]: resampled_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'].
    ↪drop(columns=['Tag']).resample('H').mean()
resampled_dataframe['Hour'] = resampled_dataframe.index.hour
resampled_dataframe['Count'] = (valid_dataframe['measuring'].resample('H').
    ↪count())
resampled_dataframe['Std'] = (valid_dataframe['measuring'].resample('H').std())
resampled_dataframe.index = resampled_dataframe.index.map(lambda t: t.
    ↪replace(minute=30, second=0))
resampled_dataframe['measuring'].plot()
```

[]: <Axes: xlabel='DateTime'>

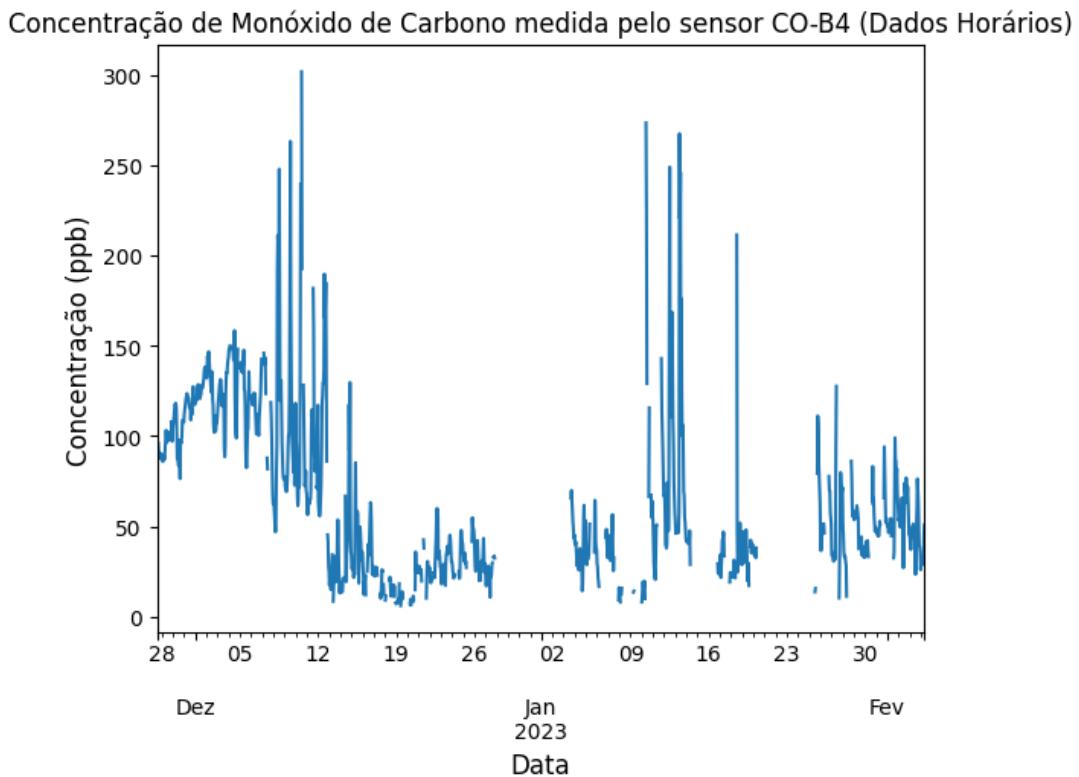


Tag hourly data according to the number of samples in an hour At least 3 samples must be valid in an hour (75 %) for the hourly data be considered as valid

```
[ ]: original_freq = sensor_dataframe.index.freq
resampled_dataframe['% valid'] = (resampled_dataframe['Count']
                                    .map(lambda c:
                                         c / (pd.Timedelta("1 hour")) /_
                                         original_freq) * 100)
resampled_dataframe['Tag'] = (resampled_dataframe['% valid']
                                .map(lambda c: 'VALID' if c >= 75 else_
                                'LOWSAMPLES'))

fig, ax = plt.subplots(figsize=(1.3*5,5))
resampled_dataframe[resampled_dataframe['Tag'] == 'VALID'][['measuring']]
    .rename('Concentração de Monóxido de Carbono').resample('1H').mean().plot()
plt.title('Concentração de Monóxido de Carbono medida pelo sensor C0-B4 (Dados Horários)')
ax.set_xlabel('Data', fontsize=12)
ax.set_ylabel('Concentração (ppb)', fontsize=12)

[ ]: Text(0, 0.5, 'Concentração (ppb)')
```



Valid data contabilization

```
[ ]: tags = ['LOWSAMPLES', 'VALID']
data_contabilization = sensor_analysis.count_tags(tags, resampled_dataframe)
data_contabilization
```

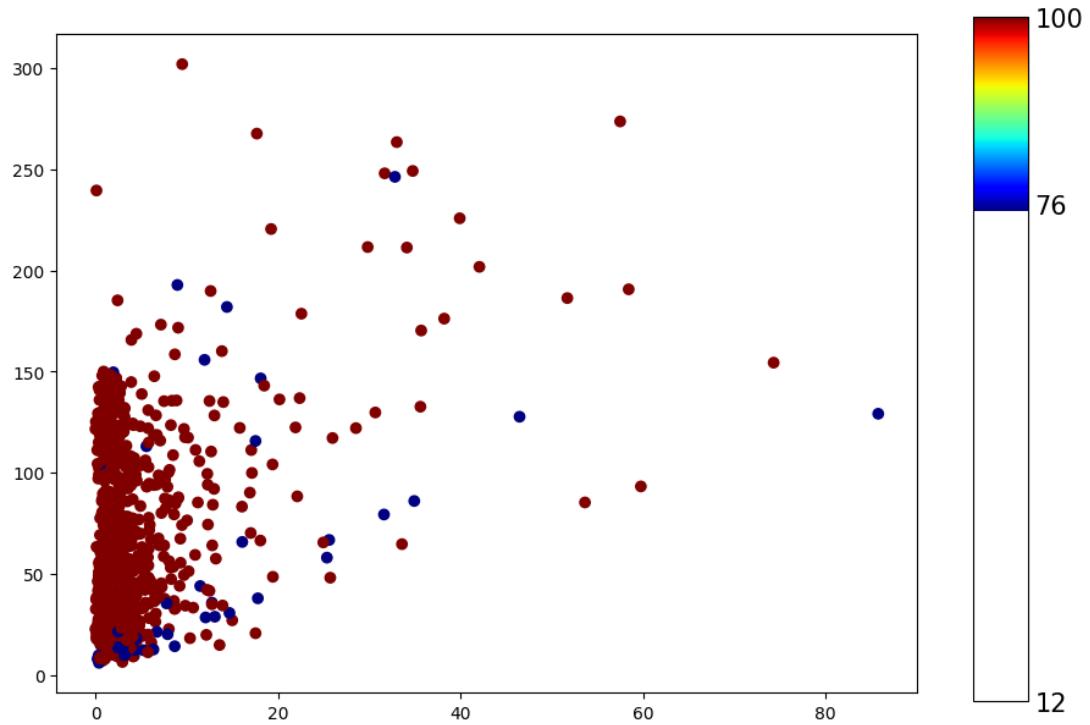
	#	%
LOWSAMPLES	603	36.523319
VALID	1048	63.476681
TOTAL	1651	100.0

2.7 Analyse the mean and standard deviation of the resampled data

The color of each data point represents the percentage of valid 15 mins samples found in an hour

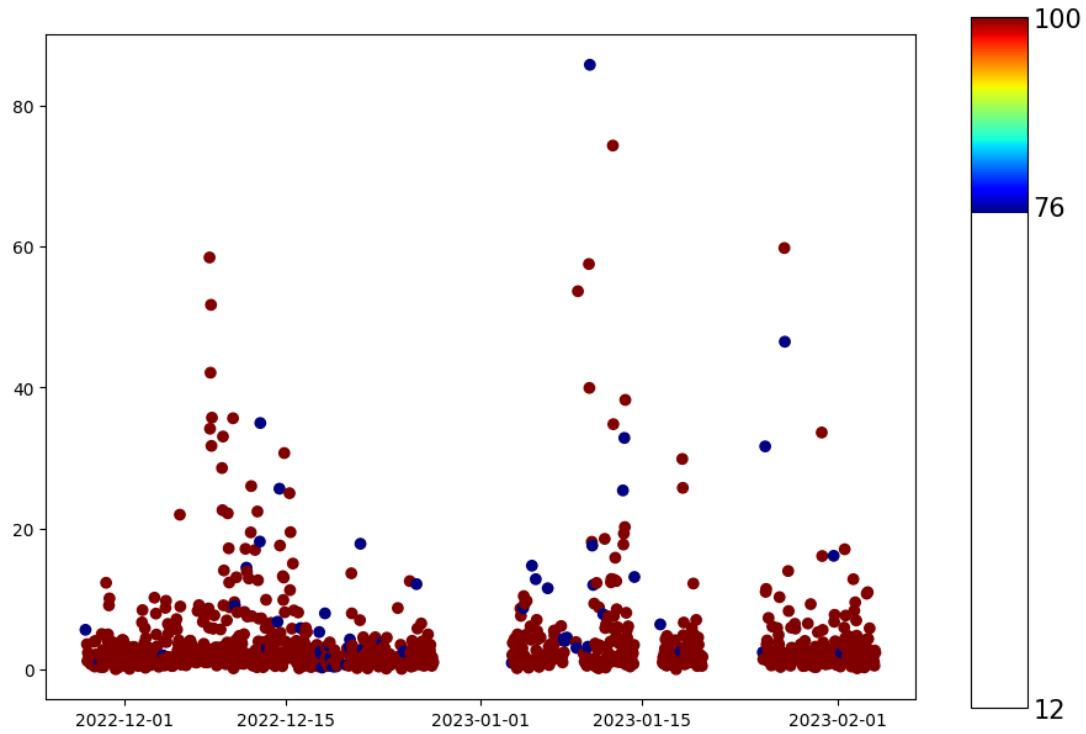
Plot mean (Y axis) vs. standard deviation (X axis) for valid hourly data colored by valid percentage of valid samples

```
[ ]: valid_resampled_dataframe = resampled_dataframe[resampled_dataframe['Tag'] == 'VALID']
sensor_analysis.plot_mean_vs_std(valid_resampled_dataframe)
```



2.7.1 Plot standard deviation in time colored according to the percentage of valid samples

```
[ ]: sensor_analysis.plot_std_in_time(valid_resampled_dataframe)
```



3 Save Data

```
[ ]: import os

output_directory_path = 'data/output/'
processing_directory_path = '../data-processing/input/'
valid_1HR_filename = '_valid_data_1HR.csv'
valid_filename = '_valid_data.csv'
dataframe_filename = '_dataframe.csv'
dataframe_1HR_filename = '_dataframe_1hr.csv'

valid_1HR_file_path_output = output_directory_path + SENSOR_NAME +_
    ↪valid_1HR_filename
valid_file_path_output = output_directory_path + SENSOR_NAME + valid_filename

dataframe_path_1HR_output = output_directory_path + SENSOR_NAME +_
    ↪dataframe_1HR_filename
dataframe_path_output = output_directory_path + SENSOR_NAME + dataframe_filename

valid_processing_1HR_file_path = processing_directory_path + SENSOR_NAME +_
    ↪valid_1HR_filename
```

```

valid_processing_file_path = processing_directory_path + SENSOR_NAME +_
    ↪valid_filename

processing_dataframe_path_1HR = processing_directory_path + SENSOR_NAME +_
    ↪dataframe_1HR_filename
processing_dataframe_path = processing_directory_path + SENSOR_NAME +_
    ↪dataframe_filename

if not os.path.exists(output_directory_path):
    os.makedirs(output_directory_path)

if not os.path.exists(processing_directory_path):
    os.makedirs(processing_directory_path)

sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'][['measuring', 'value']]._
    ↪to_csv(valid_file_path_output)
resampled_dataframe[resampled_dataframe['Tag'] == 'VALID'][['measuring', ↪
    'value']].to_csv(valid_1HR_file_path_output)
sensor_dataframe.to_csv(dataframe_path_output)
resampled_dataframe.to_csv(dataframe_path_1HR_output)

sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'][['measuring', 'value']]._
    ↪to_csv(valid_processing_file_path)
resampled_dataframe[resampled_dataframe['Tag'] == 'VALID'][['measuring', ↪
    'value']].to_csv(valid_processing_1HR_file_path)
sensor_dataframe.to_csv(processing_dataframe_path)
resampled_dataframe.to_csv(processing_dataframe_path_1HR)

```

O3_1-field-data-treatment-Diamante

February 13, 2024

1 Electrochemical O3 sensor data preprocessing

- Pollutant: Ozone
- Sensor: Alphasense OX-B4

1.1 Constants

1.1.1 Sensors IDs

1.1.2 API Constants

```
[ ]: HOST = "renovar.lcqar.ufsc.br"
PORT = 8080
RAW_DATA_DIR = "data/raw-data-monit-fix-2022-2023-Diamante/"
RAW_FILE_NAME = "ISB_031.CSV"

SENSOR_FILE_DIR = 'data/input/'
SENSOR_NAME = 'alpha_o3_1_conc'
SENSOR_FILE_NAME = SENSOR_NAME + 'web_dataframe.csv'
```

1.1.3 Sensor Constants

```
[ ]: M = 48
lower_limit=15.0
upper_limit=20e3
t_90 = 80      # sensor takes 30 seconds to reach a value of 10e3
t_90_value = 1e3
sampling_period = 15 * 60
```

2 Alphasense O3 Sensor Data

```
[ ]: import locale
locale.setlocale(locale.LC_TIME, 'pt_BR')
```

```
[ ]: 'pt_BR'
```

```
[ ]: from GetSensorDataService import GetSensorDataService
```

```

get_sensor_data_service = GetSensorDataService(HOST, PORT)
sensor_data = get_sensor_data_service.
    ↪get_data_from_file(RAW_DATA_DIR+RAW_FILE_NAME, sensor_name=SENSOR_NAME)
sensor_data

/Users/Fernando/Documents/Projects/Github/lcqar-low-cost-monit-proc/data-pre-
processing/GetSensorDataService.py:13: UserWarning: The argument
'infer_datetime_format' is deprecated and will be removed in a future version. A
strict version of it is now the default, see
https://pandas.pydata.org/pdps/0004-consistent-to-datetime-parsing.html. You
can safely remove this argument.
    df['DateTime'] = (pd.to_datetime(df[date_time_col],
infer_datetime_format=False, format='%d/%m/%Y/%H/%M/%S'))

[ ]:      latitude  longitude  measuring          DateTime
0     -28.456899 -48.972999    105.77 2022-11-20 13:48:30
1     -28.456899 -48.972999     92.35 2022-11-21 10:40:56
2     -28.456899 -48.972999     73.28 2022-11-21 10:56:42
3     -28.456899 -48.972999     69.64 2022-11-21 11:12:31
4     -28.456899 -48.972999     77.29 2022-11-21 11:28:16
...
       ...        ...
11671 -28.456899 -48.972999     2.38 2023-04-21 20:53:44
11672 -28.456899 -48.972999     0.40 2023-04-21 21:08:43
11673 -28.456899 -48.972999     1.69 2023-04-21 21:23:43
11674 -28.456899 -48.972999    33.83 2023-04-21 21:38:42
11675 -28.456899 -48.972999    20.38 2023-04-21 21:53:42

[11676 rows x 4 columns]

```

2.1 Upload Data from File

```

[ ]: import pandas as pd

df = pd.read_csv(SENSOR_FILE_DIR + SENSOR_FILE_NAME)
df.head()

[ ]:      Unnamed: 0  latitude  longitude  measuring          DateTime
0           0 -28.456899 -48.972999    105.77 2022-11-20 13:48:30
1           1 -28.456899 -48.972999     92.35 2022-11-21 10:40:56
2           2 -28.456899 -48.972999     73.28 2022-11-21 10:56:42
3           3 -28.456899 -48.972999     69.64 2022-11-21 11:12:31
4           4 -28.456899 -48.972999     77.29 2022-11-21 11:28:16

```

2.1.1 Create Sensor Dataframe as Pandas Series with a period of 15 mins

```
[ ]: # Remove the first column with the indexes and save data into web dataframe
web_dataframe = df.drop(df.columns[0], axis='columns')
web_dataframe['DateTime'] = (pd.to_datetime(df['DateTime'],  
    ↪infer_datetime_format=True))

# Resample data with 15 mins period and create sensor dataframe
sensor_dataframe = web_dataframe.sort_values(by='DateTime', ascending=True).  
    ↪reset_index().drop(columns='index')
sensor_dataframe.index = sensor_dataframe['DateTime']
sensor_dataframe = sensor_dataframe.drop(columns=['DateTime'])
sensor_dataframe = sensor_dataframe.resample('15T').mean()
sensor_dataframe
```

```
/var/folders/wc/_83zcrx913j1dqwg4g90kbhh0000gp/T/ipykernel_2699/166902210.py:3:
UserWarning: The argument 'infer_datetime_format' is deprecated and will be
removed in a future version. A strict version of it is now the default, see
https://pandas.pydata.org/pdeps/0004-consistent-to-datetime-parsing.html. You
can safely remove this argument.
    web_dataframe['DateTime'] = (pd.to_datetime(df['DateTime'],
infer_datetime_format=True))
```

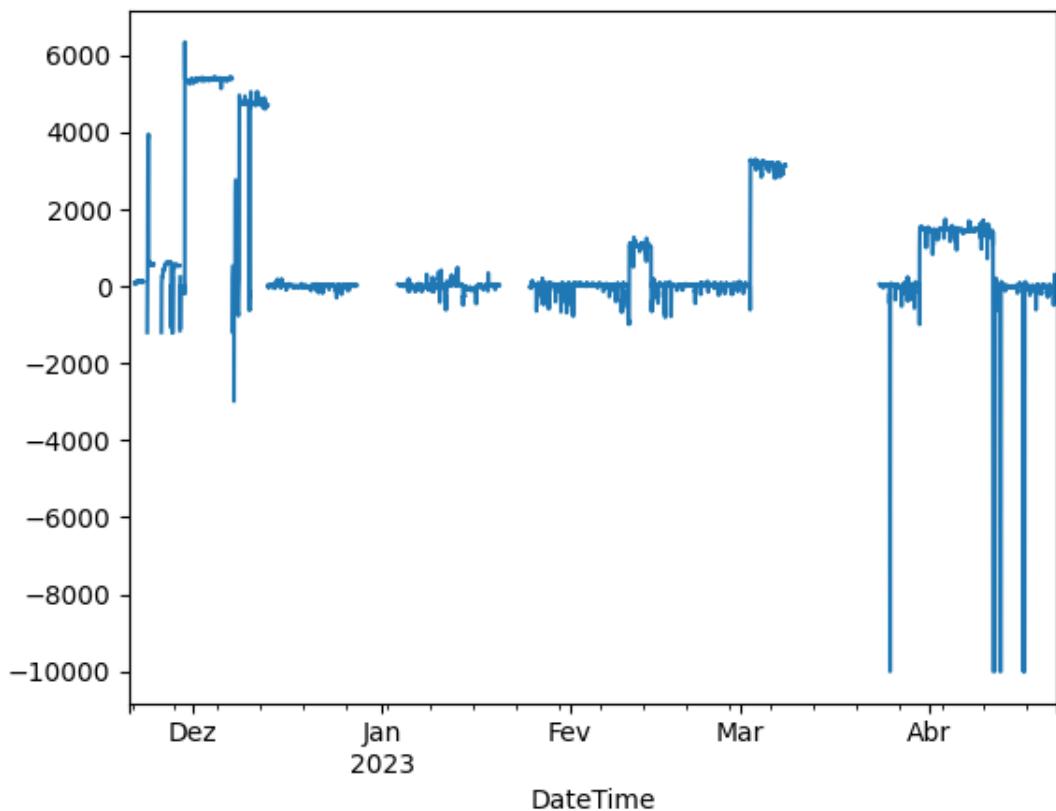
```
[ ]:          latitude  longitude  measuring
DateTime
2022-11-20 13:45:00 -28.456899 -48.972999      105.77
2022-11-20 14:00:00      NaN      NaN      NaN
2022-11-20 14:15:00      NaN      NaN      NaN
2022-11-20 14:30:00      NaN      NaN      NaN
2022-11-20 14:45:00      NaN      NaN      NaN
...
2023-04-21 20:45:00 -28.456899 -48.972999      2.38
2023-04-21 21:00:00 -28.456899 -48.972999      0.40
2023-04-21 21:15:00 -28.456899 -48.972999      1.69
2023-04-21 21:30:00 -28.456899 -48.972999     33.83
2023-04-21 21:45:00 -28.456899 -48.972999     20.38
```

[14625 rows x 3 columns]

2.1.2 Plot raw data

```
[ ]: sensor_dataframe['measuring'].plot()
```

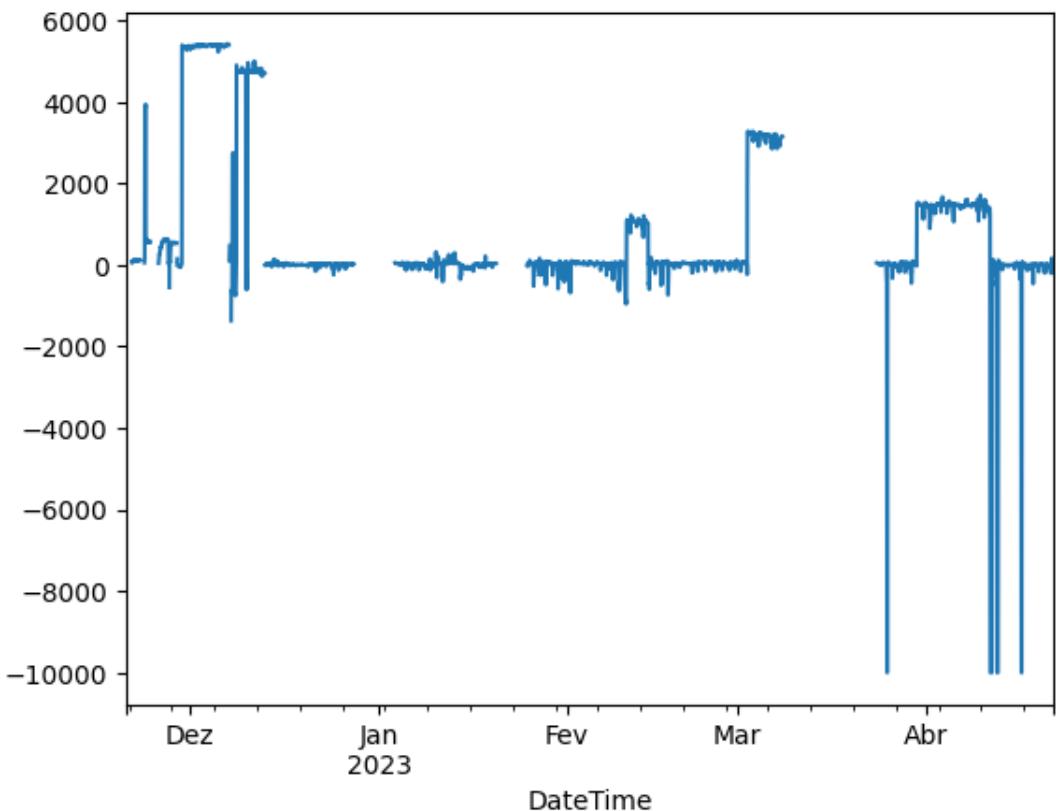
```
[ ]: <Axes: xlabel='DateTime'>
```



2.1.3 Smooth data with a window of 4 samples, i.e.: an hour of data

```
[ ]: sensor_dataframe['raw_measuring'] = sensor_dataframe['measuring']
sensor_dataframe['measuring'] = sensor_dataframe['measuring'].rolling(window=4).
    median()
sensor_dataframe['Hour'] = sensor_dataframe.index.hour
sensor_dataframe['measuring'].plot()
```

```
[ ]: <Axes: xlabel='DateTime'>
```



2.1.4 Convert to ug/m3

```
[ ]: sensor_dataframe['value'] = sensor_dataframe['measuring'].map(lambda v: 0.
    ↪0409*v*M)
```

2.2 Tag data

Apply tags to the data according to the quality control processes described above

```
[ ]: sensor_dataframe['Tag'] = 'VALID'
```

Tag data during stabilizing period AQMesh recommends to considering the first two days after installation as a period for stabilization. In our case, a period of seven days was selected in order to remove outliers candidates detected during that period. For that purpose we detect the first 7 days of uninterrupted measurements

```
[ ]: # Identify consecutive measurements (15 mins between each)
consecutive_periods = sensor_dataframe.index.to_series().diff().dt.seconds ==
    ↪900
```

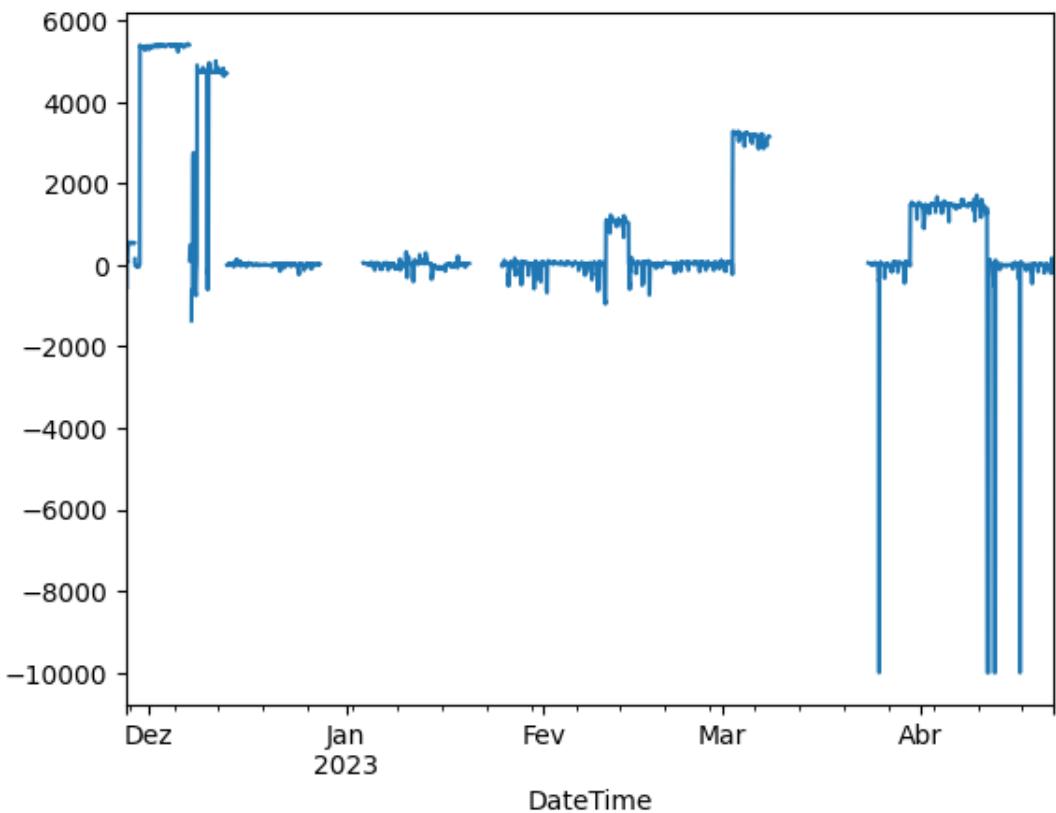
```

# Select the first seven days of consecutive measurements (96 = 24 * 60 mins / ↳
↪15 mins, amount of 15 mins periods in a day)
last_stabilizing_index = sensor_dataframe[consecutive_periods].head(96 * 7).
↪index[-1] # 7 days
sensor_dataframe.loc[sensor_dataframe.index[0]:last_stabilizing_index, 'Tag'] = ↳
↪'STABILIZING'
sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['measuring'].
↪resample('15T').mean().plot()
sensor_dataframe.head()

```

```
[ ]:          latitude longitude measuring raw measuring Hour \
DateTime
2022-11-20 13:45:00 -28.456899 -48.972999      NaN      105.77    13
2022-11-20 14:00:00      NaN      NaN      NaN      NaN      14
2022-11-20 14:15:00      NaN      NaN      NaN      NaN      14
2022-11-20 14:30:00      NaN      NaN      NaN      NaN      14
2022-11-20 14:45:00      NaN      NaN      NaN      NaN      14

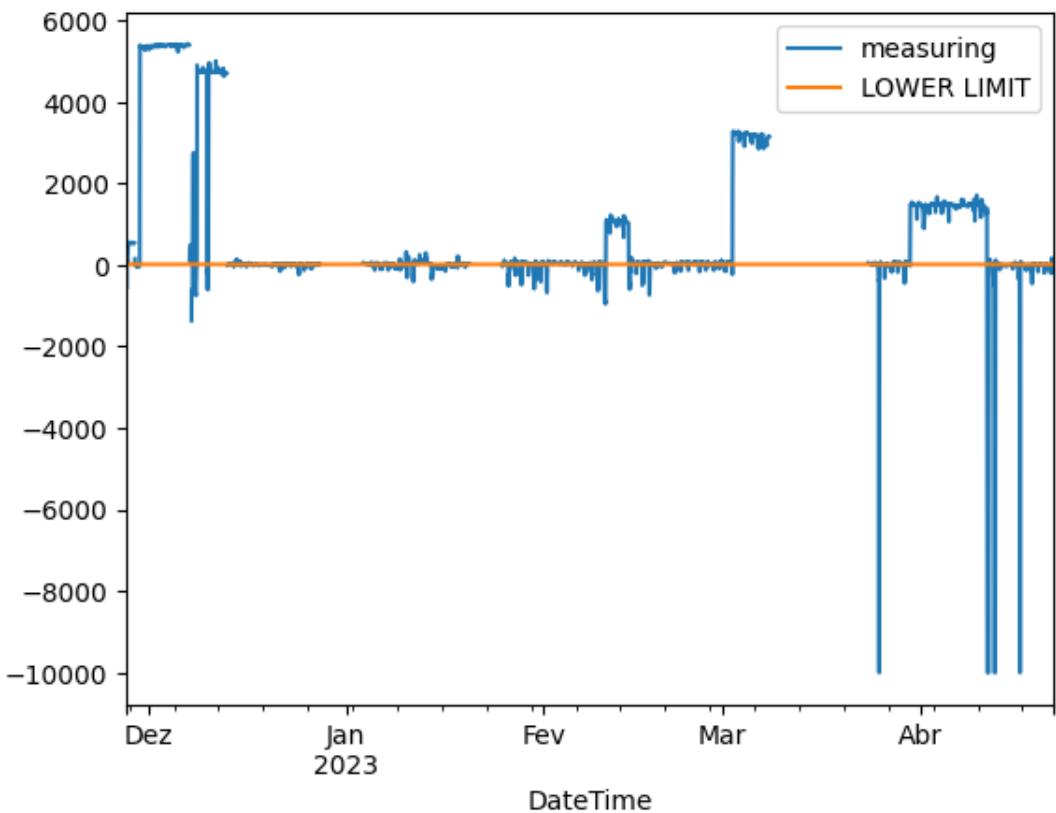
                           value        Tag
DateTime
2022-11-20 13:45:00      NaN  STABILIZING
2022-11-20 14:00:00      NaN  STABILIZING
2022-11-20 14:15:00      NaN  STABILIZING
2022-11-20 14:30:00      NaN  STABILIZING
2022-11-20 14:45:00      NaN  STABILIZING
```



Tag values that are missing, are greater than sensor span or are lower than sensor resolution The upper limit was too high that affected the graph resolution, therefore it is not plotted

```
[ ]: sensor_dataframe['UPPER LIMIT'] = upper_limit
sensor_dataframe['LOWER LIMIT'] = lower_limit
sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'][['measuring', 'LOWER LIMIT']].resample('15T').mean().plot()
```

```
[ ]: <Axes: xlabel='DateTime'>
```



```
[ ]: import numpy as np
from SensorDataAnalysisService import SensorDataAnalysisService as
    ↪sensor_analysis

def tag_by_sensor_limits(value, tag, lower_limit, upper_limit):
    if (tag != 'VALID'): return tag
    return sensor_analysis.get_tags_from_series(value=value,
                                                lower_limit=lower_limit,
                                                upper_limit=upper_limit)

sensor_dataframe['Tag'] = (sensor_dataframe[['measuring', 'Tag']]
                           .apply(lambda df: ↪
    ↪tag_by_sensor_limits(value=df[0], tag=df[1],
    ↪lower_limit=lower_limit,
    ↪upper_limit=upper_limit),
                           axis=1))
```

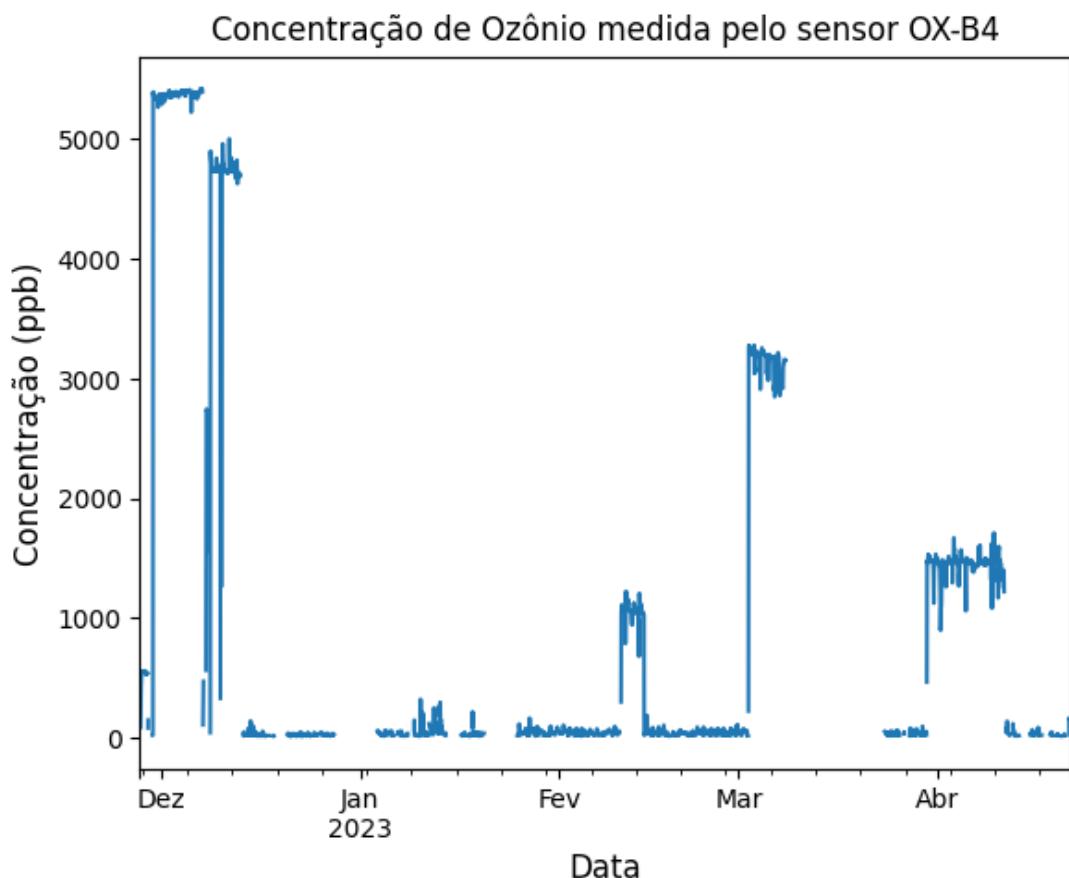
Plot valid data after removing samples bellow lower limit

```
[ ]: import matplotlib.pyplot as plt

fig, ax = plt.subplots(figsize=(1.3*5,5))
sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['measuring'].  

    ↪rename('Concentração de Ozônio').resample('15T').mean().plot()
plt.title('Concentração de Ozônio medida pelo sensor OX-B4')
ax.set_xlabel('Data', fontsize=12)
ax.set_ylabel('Concentração (ppb)', fontsize=12)
```

```
[ ]: Text(0, 0.5, 'Concentração (ppb)')
```



2.3 Change point Analysis

2.3.1 Change point analysis of 15 mins data

```
[ ]: import ruptures as rpt

series = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['measuring']
```

```
signal = np.array(series).reshape(-1, 1)
algo = rpt.Pelt(model="rbf", jump=50, min_size=100).fit(signal=signal)
result = algo.predict(pen=10)
result
```

```
[ ]: [100, 850, 950, 1350, 4000, 4350, 5650, 6200, 6600, 7750, 8068]
```

```
[ ]: change_point_index = series[[x - 1 for x in result]].index
sensor_dataframe['CHANGE POINT'] = False
sensor_dataframe['CHANGE POINT'].loc[change_point_index] = True
```

```
/var/folders/wc/_83zcrx913j1dqwg4g90kbhh0000gp/T/ipykernel_2699/717246363.py:3:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame
```

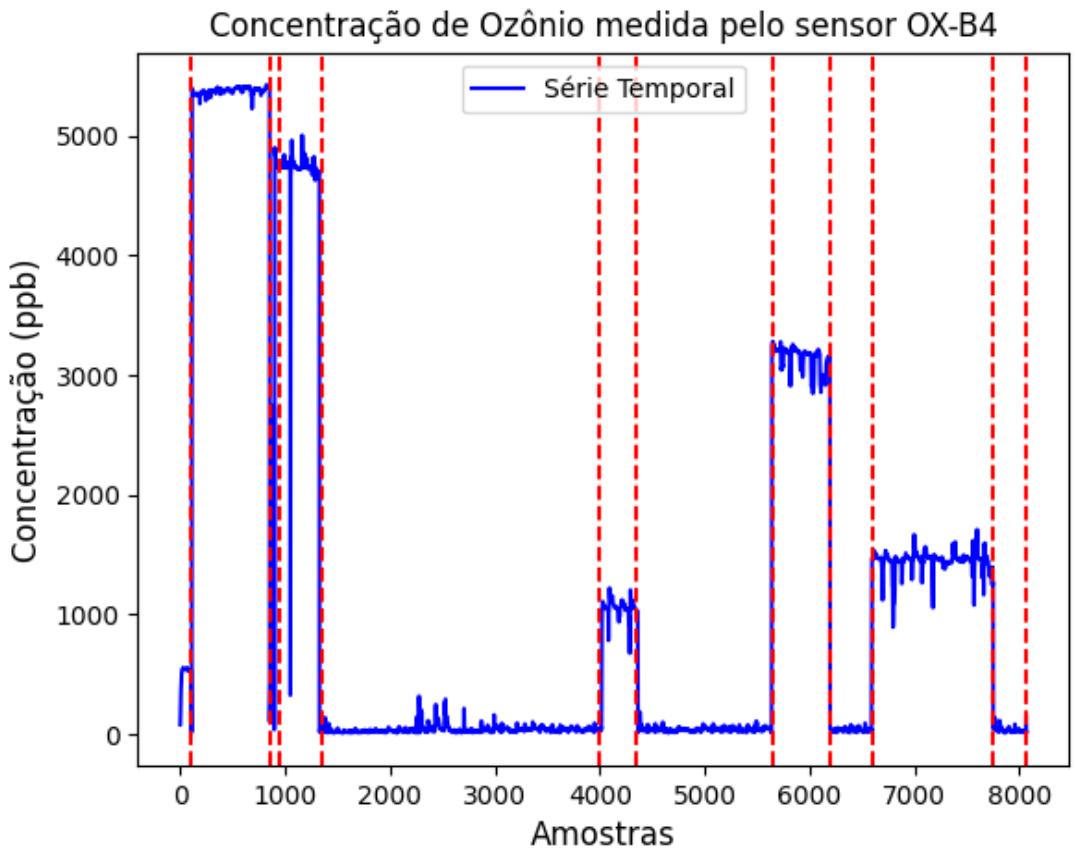
```
See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
    sensor_dataframe['CHANGE POINT'].loc[change_point_index] = True
```

```
[ ]: import matplotlib.pyplot as plt

fig, ax = plt.subplots(figsize=(1.3*5,5))
plt.plot(signal, label='Série Temporal', color='blue')
plt.title('Concentração de Ozônio medida pelo sensor OX-B4')
ax.set_xlabel('Amostras', fontsize=12)
ax.set_ylabel('Concentração (ppb)', fontsize=12)

# Plot change points
for point in result:
    plt.axvline(x=point, color='red', linestyle='--')

# Show the plot
plt.legend()
plt.show()
```



```
[ ]: # 100, 850, 950, 1350, 4000, 4350, 5650, 6200, 6600, 7750, 8068
rebase_df = sensor_dataframe[:change_point_index[3]]
rebase_df = rebase_df + sensor_dataframe[change_point_index[4] :
    ↪change_point_index[5]]
rebase_df = rebase_df + sensor_dataframe[change_point_index[6] :
    ↪change_point_index[7]]
rebase_df = rebase_df + sensor_dataframe[change_point_index[8] :
    ↪change_point_index[9]]
```

Tag dataframe with changepoints

```
[ ]: sensor_dataframe.loc[rebase_df.index, 'Tag'] = (sensor_dataframe.loc[rebase_df .
    ↪index, 'Tag']
                                .apply(lambda t: 'REBASE' if t ↪
    ↪== 'VALID' else t) )
sensor_dataframe[sensor_dataframe['Tag'] == 'REBASE'].head()
```

```
[ ]:          latitude longitude measuring raw measuring Hour \
DateTime
```

```

2022-11-27 14:45:00 -28.456899 -48.972999      81.460      197.19     14
2022-11-27 17:15:00 -28.456899 -48.972999      84.785      171.96     17
2022-11-27 17:30:00 -28.456899 -48.972999     147.105      214.13     17
2022-11-27 17:45:00 -28.456899 -48.972999     193.045      256.06     17
2022-11-27 18:00:00 -28.456899 -48.972999     235.095      289.57     18

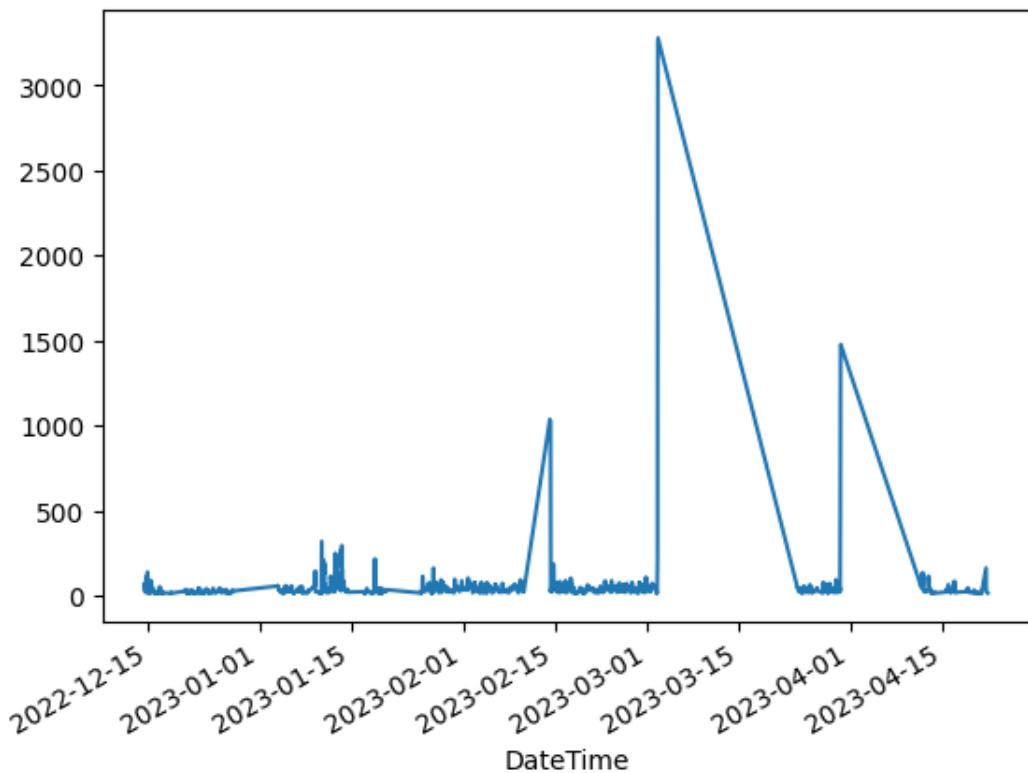
```

	value	Tag	UPPER LIMIT	LOWER LIMIT	\
DateTime					
2022-11-27 14:45:00	159.922272	REBASE	20000.0	15.0	
2022-11-27 17:15:00	166.449912	REBASE	20000.0	15.0	
2022-11-27 17:30:00	288.796536	REBASE	20000.0	15.0	
2022-11-27 17:45:00	378.985944	REBASE	20000.0	15.0	
2022-11-27 18:00:00	461.538504	REBASE	20000.0	15.0	

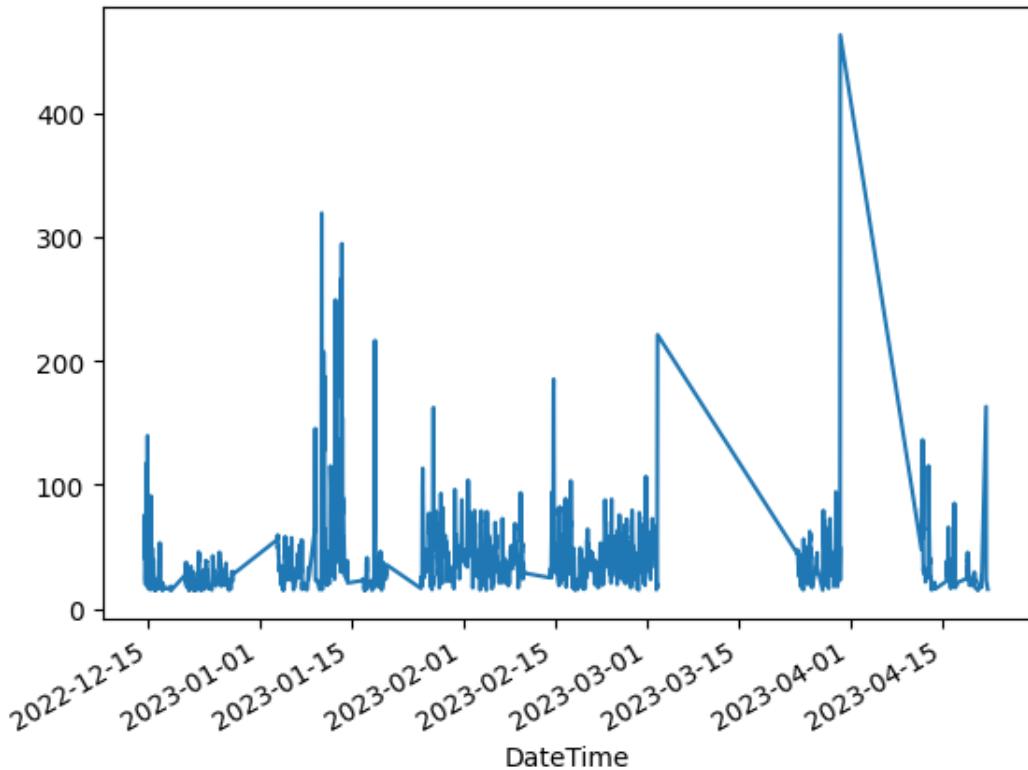
CHANGE POINT

DateTime	
2022-11-27 14:45:00	False
2022-11-27 17:15:00	False
2022-11-27 17:30:00	False
2022-11-27 17:45:00	False
2022-11-27 18:00:00	False

```
[ ]: sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['measuring'].plot()
valid_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']
```



```
[ ]: sensor_dataframe.loc[sensor_dataframe['measuring'] > 500, 'Tag'] = 'REBASE'
sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['measuring'].plot()
valid_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']
```



Data Tag contabilization

```
[ ]: tags = ['MISSING', 'LTLL', 'GTUL', 'STABILIZING', 'BADSPIKE', 'VALID', 'REBASE']
data_contabilization = sensor_analysis.count_tags(tags, sensor_dataframe)
data_contabilization
```

	#	%
MISSING	2750	18.803419
LTLL	3134	21.42906
GTUL	0	0.0
STABILIZING	514	3.51453
BADSPIKE	0	0.0
VALID	4635	31.692308
REBASE	3592	24.560684
TOTAL	14625	100.0

2.4 Analyse valid data

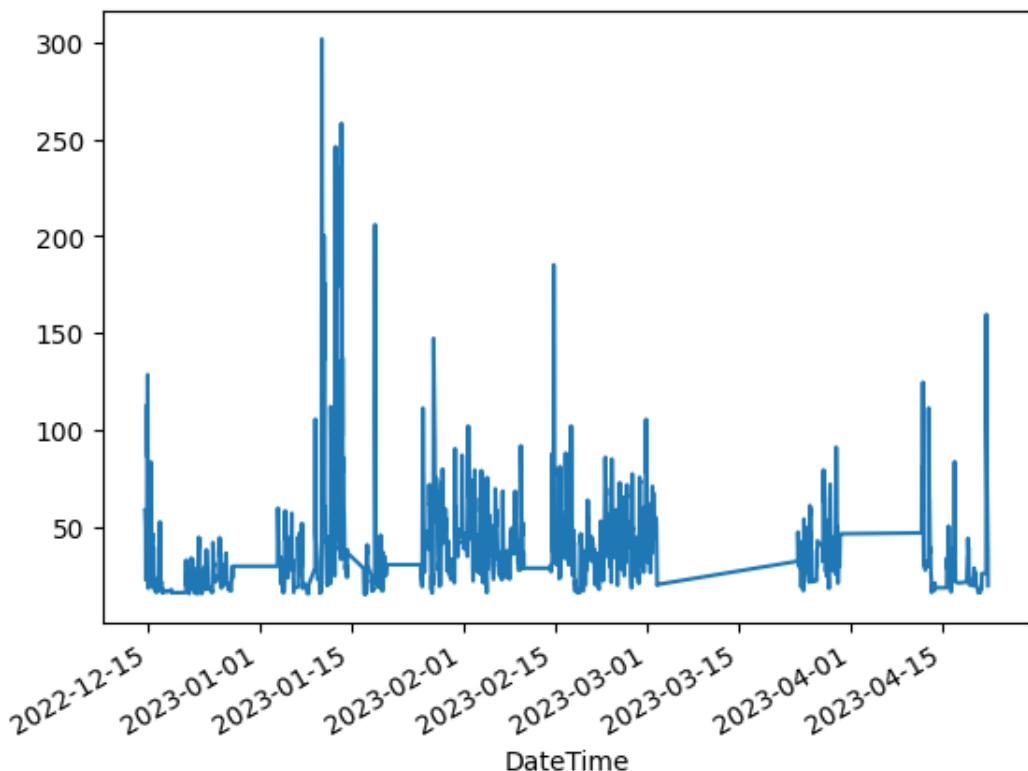
The graph shows the time series of the valid data, box plots of the data grouped by hour of the day, and histogram of the data.

```
[ ]: valid_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'].  
      .drop(columns=['Tag'])
```

2.4.1 Smooth data with a window of 4 samples, i.e.: an hour of data

```
[ ]: valid_dataframe['raw measuring'] = valid_dataframe['measuring']  
valid_dataframe['measuring'] = valid_dataframe['measuring'].rolling(window=4).  
    .median()  
valid_dataframe['measuring'].plot()
```

```
[ ]: <Axes: xlabel='DateTime'>
```



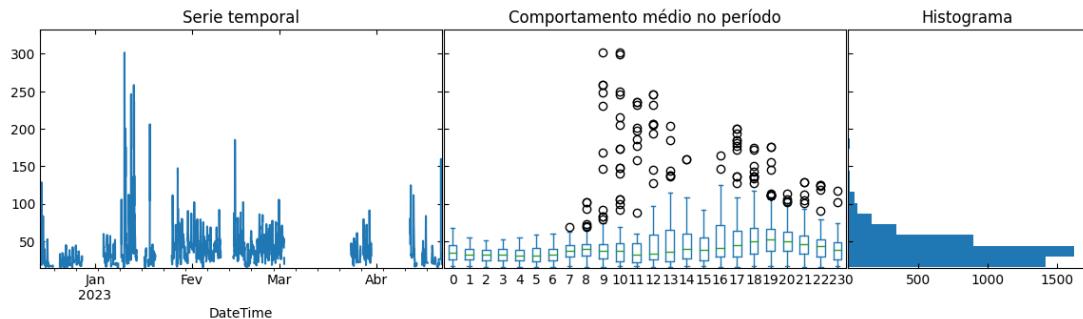
```
[ ]: valid_dataframe = valid_dataframe.resample('15T').mean()  
sensor_analysis.plot_box_hist(df=valid_dataframe, bins=20)
```

```
/Users/Fernando/Documents/Projects/Github/lcqar-low-cost-monit-proc/data-pre-  
processing/SensorDataAnalysisService.py:111: SettingWithCopyWarning:  
A value is trying to be set on a copy of a slice from a DataFrame.
```

```
Try using .loc[row_indexer,col_indexer] = value instead
```

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

```
df['Hour'] = df['Hour'].astype('int64')
```



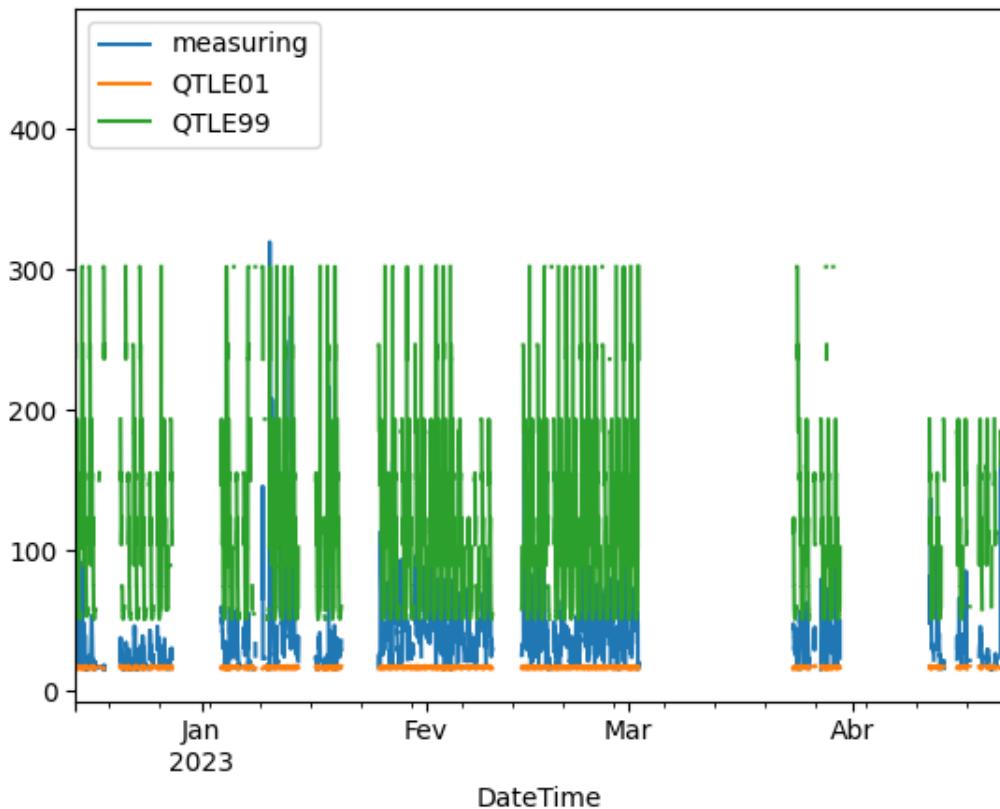
2.4.2 Calculate quantiles

We calculated the 1% and 99 % quantiles of every hour of the day. The values greater outside the 1 % - 99 % were tagged as Greater than Quantile 99 (GTQTL99) and Lower than Quantile 1 (LTQTL01)

```
[ ]: import numpy as np

global_qtle_01 = valid_dataframe.pivot(columns='Hour')['measuring'].
    ↪quantile(q=0.01, axis='index', interpolation='lower').dropna()
global_qtle_99 = valid_dataframe.pivot(columns='Hour')['measuring'].
    ↪quantile(q=0.99, axis='index', interpolation='higher').dropna()
sensor_dataframe['QTLE01'] = sensor_dataframe['Hour'].map(lambda hr: ↪
    ↪global_qtle_01[hr] if (not np.isnan(hr)) else np.nan)
sensor_dataframe['QTLE99'] = sensor_dataframe['Hour'].map(lambda hr: ↪
    ↪global_qtle_99[hr] if (not np.isnan(hr)) else np.nan)
sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'][['measuring', 'QTLE01', ↪
    ↪'QTLE99']].resample('15T').mean().plot()

[ ]: <Axes: xlabel='DateTime'>
```



2.4.3 Tag data according to quantiles

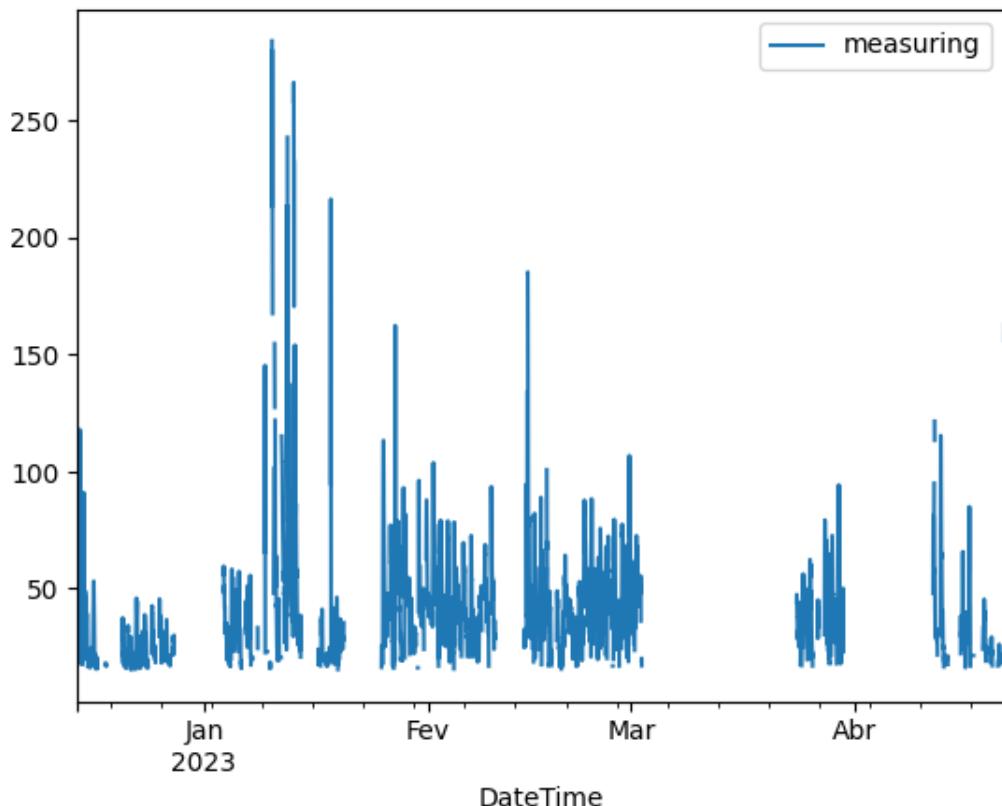
```
[ ]: sensor_dataframe['Tag'] = (sensor_dataframe[['Tag', 'measuring', 'QTLE01',
    ↪'QTLE99']] .apply(lambda df: sensor_analysis.
    ↪tag_by_quantiles(current_tag=df[0],
    ↪value=df[1],
    ↪quantile_01=df[2],
    ↪quantile_99=df[3]),
    ↪axis=1))
tags = ['MISSING', 'LTLL', 'GTUL', 'BADSPIKE', 'VALID', 'LTQTLE01', 'GTQTLE99',
    ↪'REBASE']
data_contabilization = sensor_analysis.count_tags(tags, sensor_dataframe)
data_contabilization
```

```
[ ]:          #      %
MISSING    2750  18.803419
LTLL       3134  21.42906
GTUL        0     0.0
BADSPIKE    0     0.0
VALID      4469  30.557265
LTQTL01     102   0.697436
GTQTL99      64   0.437607
REBASE      3592  24.560684
TOTAL      14625  100.0
```

Plot valid data

```
[ ]: sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'][['measuring']].  
      ↪resample('15T').mean().plot()
```

```
[ ]: <Axes: xlabel='DateTime'>
```

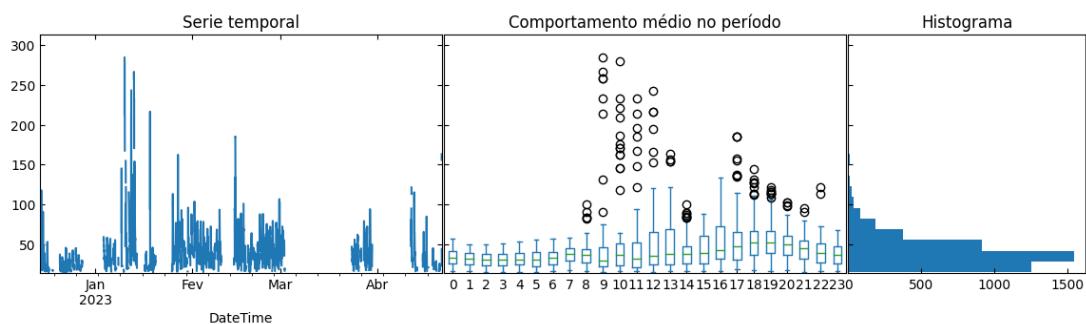


Analyse data after removing quantiles

```
[ ]: valid_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'].
    drop(columns=['Tag']).resample('15T').mean()
sensor_analysis.plot_box_hist(df=valid_dataframe, bins=20)
```

/Users/Fernando/Documents/Projects/Github/lcqar-low-cost-monit-proc/data-pre-processing/SensorDataAnalysisService.py:111: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
`df['Hour'] = df['Hour'].astype('int64')`



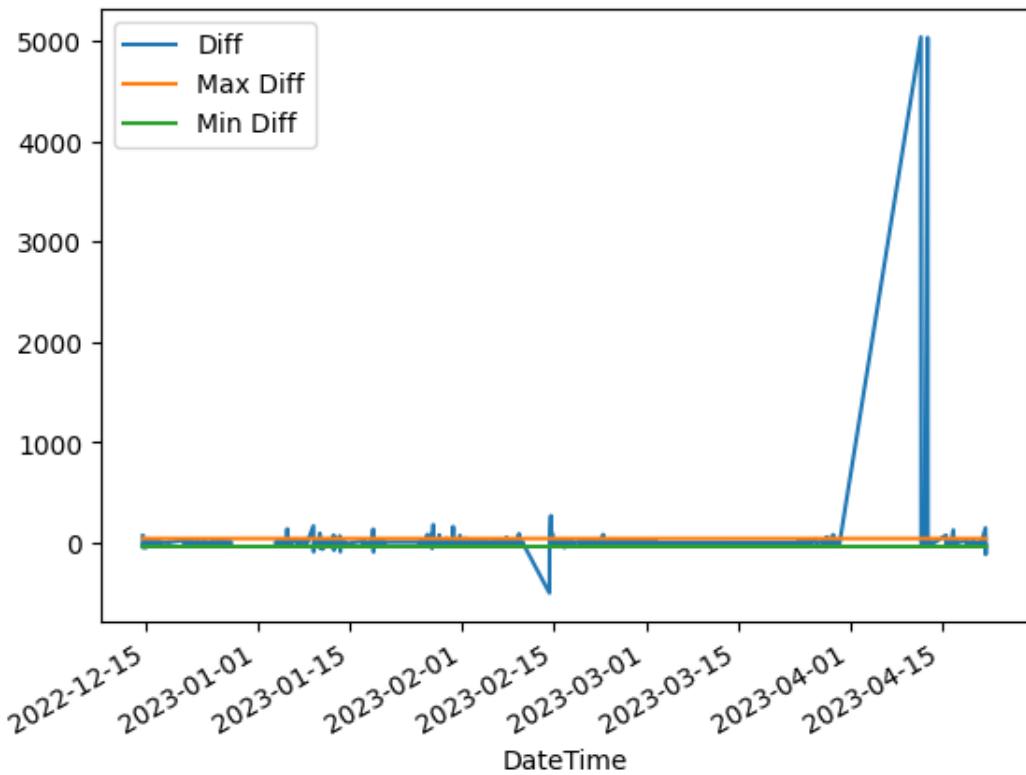
2.5 Analyse data derivatives

Analyse the derivatives of the data for removing abrupt changes in the data (BADSPIKE)

Tag values with derivatives faster than sensor response time The value of 40 ppb was selected based on the maximum derivative found in the reference data. The maximum derivative (with measuring period of 1 hour) found was -45 ppb. Therefore, a maximum of 40 ppb was considered appropriate for a 15 mins period.

```
[ ]: max_diff_value = 40
sensor_dataframe['Diff'] = sensor_dataframe['measuring'].resample('15T').mean().
    diff()
sensor_dataframe['Max Diff'] = max_diff_value
sensor_dataframe['Min Diff'] = -max_diff_value
sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'][['Diff', 'Max Diff', 'Min Diff']].plot()
```

```
[ ]: <Axes: xlabel='DateTime'>
```



```
[ ]: import numpy as np

def tag_data_with_derivatives(tagged_df, max_diff_value):
    current_tag = tagged_df[0]
    value = tagged_df[1]
    if ((current_tag != 'VALID') or (np.isnan(value))): return current_tag
    if ((value > max_diff_value) or (value < -max_diff_value)): return u
    ↪'BADSPIKE'
    return 'VALID'

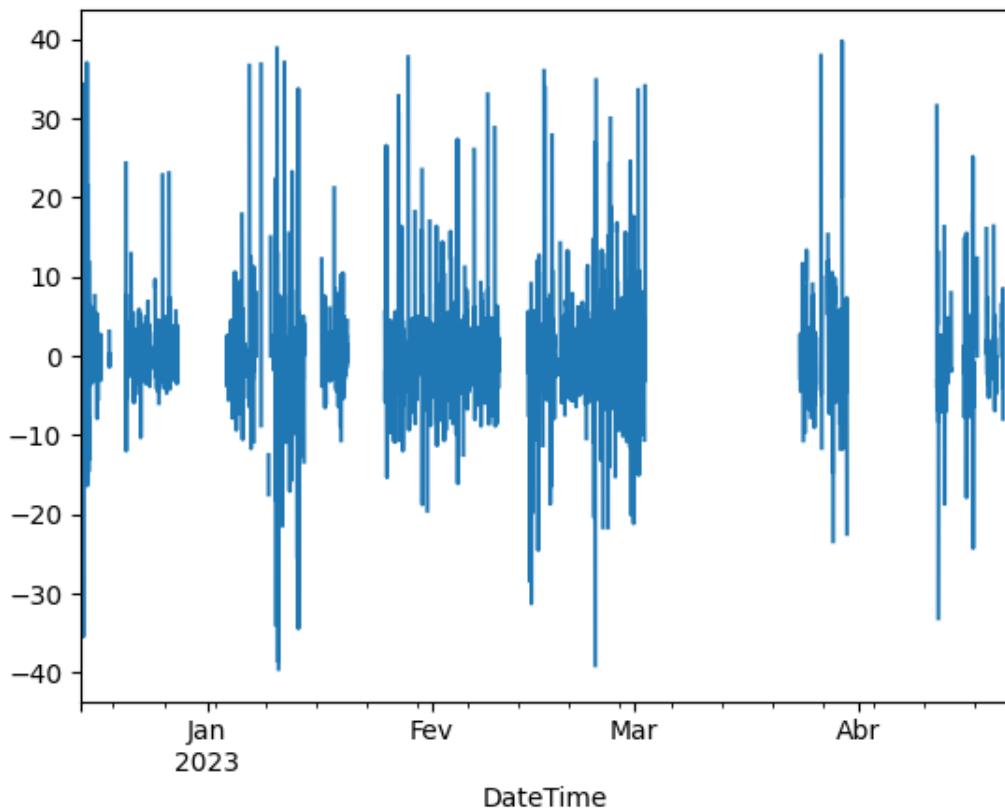
sensor_dataframe['Tag'] = (sensor_dataframe[['Tag', 'Diff', 'Max Diff']]
                           .apply(lambda df: u
    ↪tag_data_with_derivatives(tagged_df=df, max_diff_value=df[2]),
                           axis=1))

tags = ['MISSING', 'LTLL', 'GTUL', 'STABILIZING', 'BADSPIKE', 'VALID', u
    ↪'LTQTLE01', 'GTQTLE99', 'REBASE']
data_contabilization = sensor_analysis.count_tags(tags, sensor_dataframe)
data_contabilization
```

```
[ ]: # %  
MISSING      2750  18.803419  
LTLL         3134  21.42906  
GTUL          0    0.0  
STABILIZING   514   3.51453  
BADSPIKE     56   0.382906  
VALID        4413  30.174359  
LTQTLE01     102   0.697436  
GTQTLE99     64   0.437607  
REBASE       3592  24.560684  
TOTAL        14625 100.0
```

```
[ ]: sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['Diff'].resample('15T').mean().plot()
```

```
[ ]: <Axes: xlabel='DateTime'>
```



Plot data derivatives vs data

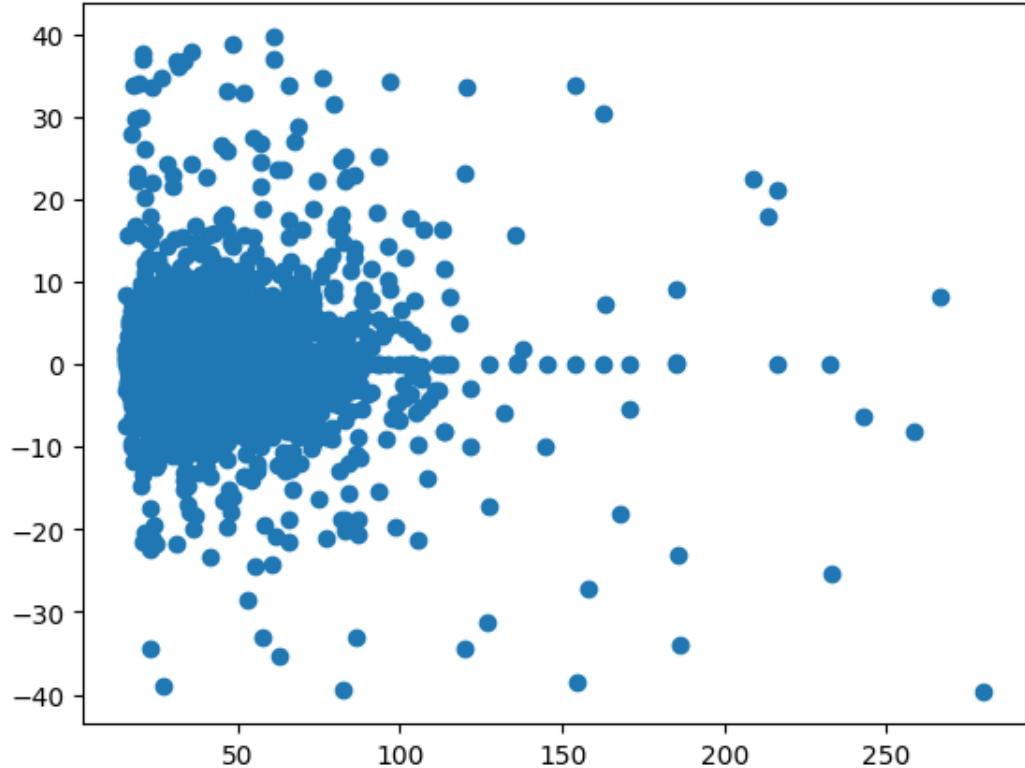
```
[ ]: import matplotlib.pyplot as plt
```

```

valid_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']
fig = plt.figure(figsize=(1.3*5,5))
plt.scatter(valid_dataframe['measuring'], valid_dataframe['Diff'])

```

[]: <matplotlib.collections.PathCollection at 0x158b33580>



Data Tag contabilization

```

tags = ['MISSING', 'LTLL', 'GTUL', 'STABILIZING', 'BADSPIKE', 'VALID', 'LTQTLE01', 'GTQTLE99', 'REBASE']
data_contabilization = sensor_analysis.count_tags(tags, sensor_dataframe)
data_contabilization

```

	#	%
MISSING	2750	18.803419
LTLL	3134	21.42906
GTUL	0	0.0
STABILIZING	514	3.51453
BADSPIKE	56	0.382906
VALID	4413	30.174359
LTQTLE01	102	0.697436
GTQTLE99	64	0.437607

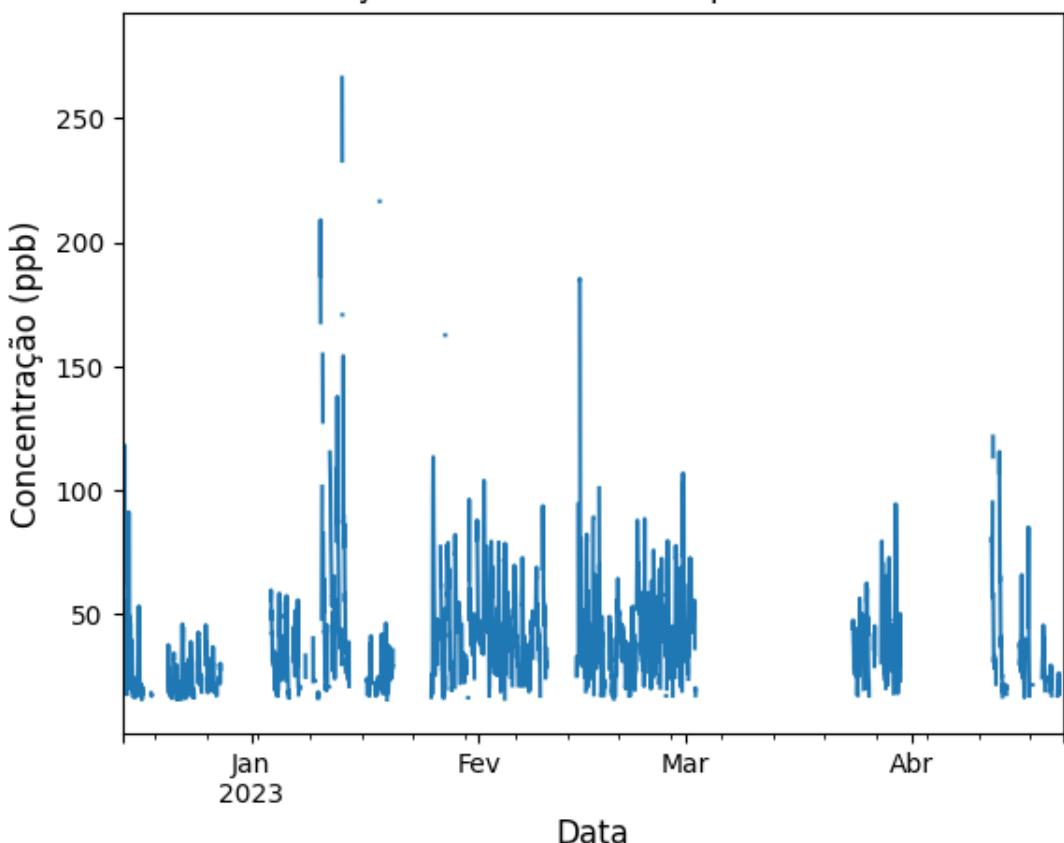
```
REBASE      3592  24.560684
TOTAL       14625   100.0
```

Plot valid data

```
[ ]: fig, ax = plt.subplots(figsize=(1.3*5,5))
sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'][['measuring']].
    ↪rename('Concentração de Ozônio').resample('15T').mean().plot()
plt.title('Concentração de Ozônio medida pelo sensor OX-B4')
ax.set_xlabel('Data', fontsize=12)
ax.set_ylabel('Concentração (ppb)', fontsize=12)

[ ]: Text(0, 0.5, 'Concentração (ppb)')
```

Concentração de Ozônio medida pelo sensor OX-B4



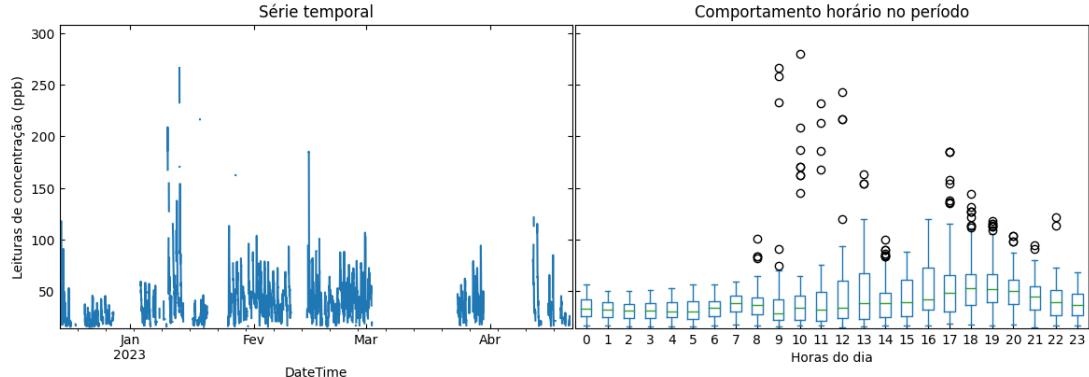
```
[ ]: valid_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'].
    ↪drop(columns=['Tag']).resample('15T').mean()
sensor_analysis.plot_box(df=valid_dataframe)
```

```
/Users/Fernando/Documents/Projects/Github/lcqar-low-cost-monit-proc/data-pre-
```

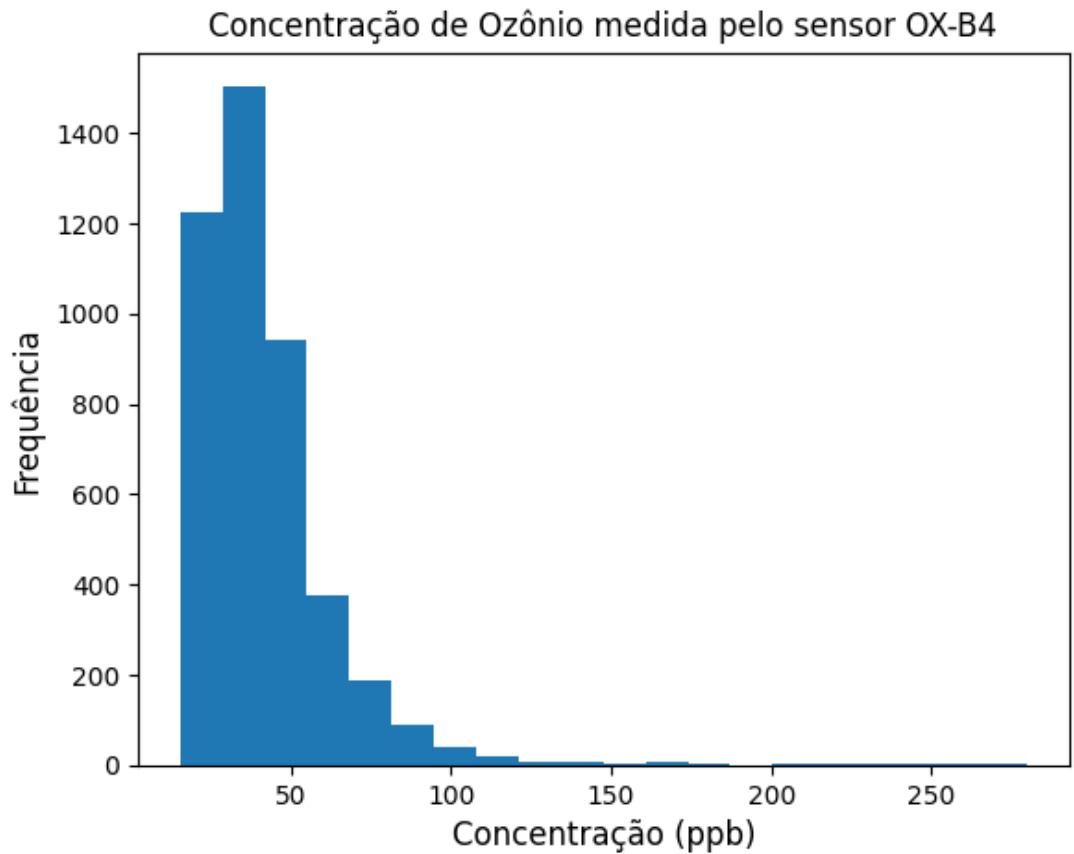
```
processing/SensorDataAnalysisService.py:143: SettingWithCopyWarning:  
A value is trying to be set on a copy of a slice from a DataFrame.  
Try using .loc[row_indexer,col_indexer] = value instead
```

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

```
df['Hour'] = df['Hour'].astype('int64')
```



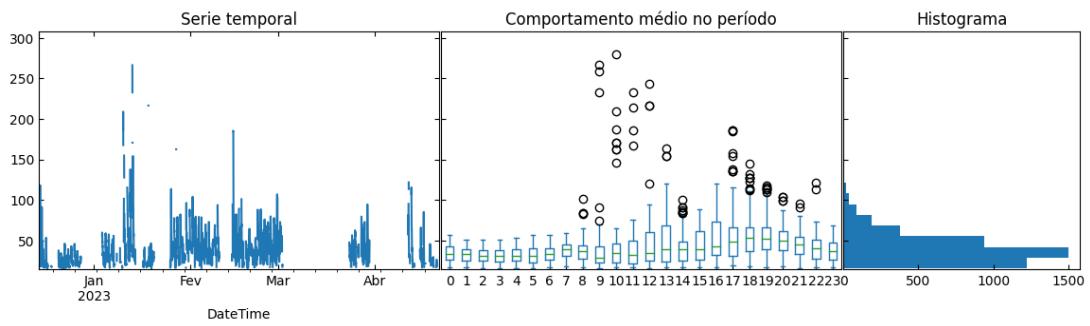
```
[ ]: fig, ax = plt.subplots(figsize=(1.3*5,5))  
(sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['measuring']  
 .rename('Concentração de Ozônio')  
 .resample('15T').mean().hist(bins=20))  
plt.title('Concentração de Ozônio medida pelo sensor OX-B4')  
ax.set_xlabel('Concentração (ppb)', fontsize=12)  
ax.set_ylabel('Frequência', fontsize=12)  
ax.grid(False)
```



```
[ ]: valid_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'].
    drop(columns=['Tag']).resample('15T').mean()
sensor_analysis.plot_box_hist(df=valid_dataframe, bins=20)
```

```
/Users/Fernando/Documents/Projects/Github/lcqar-low-cost-monit-proc/data-pre-
processing/SensorDataAnalysisService.py:111: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

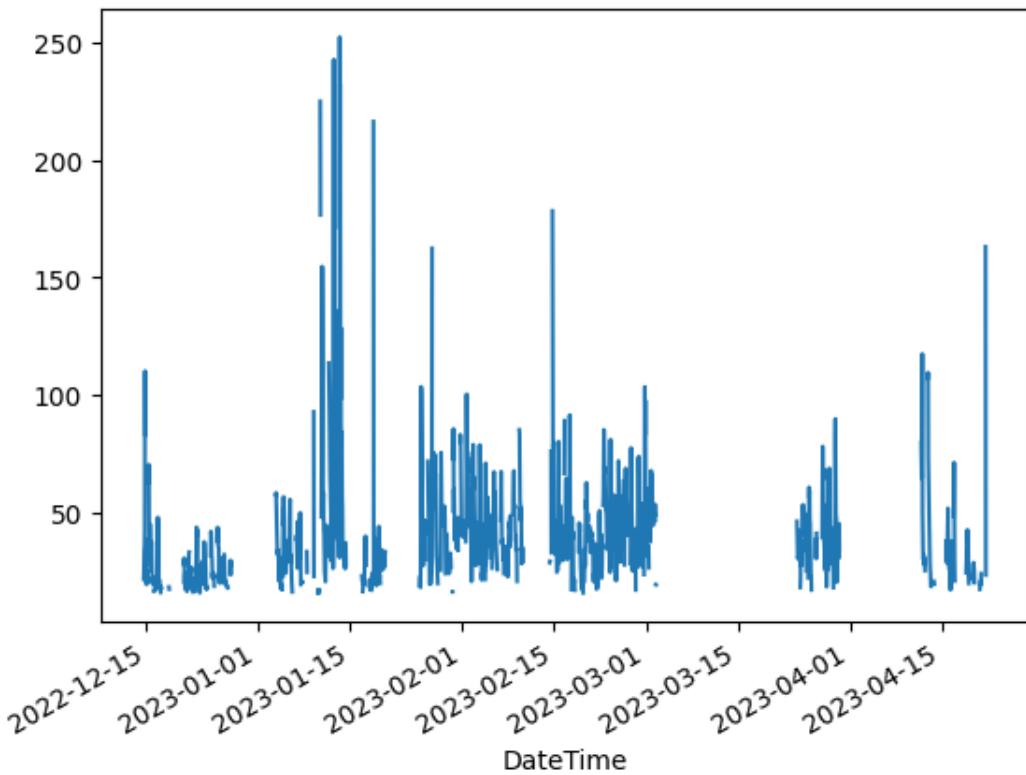
```
See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
df['Hour'] = df['Hour'].astype('int64')
```



2.6 Resample valid data to 1 HR

```
[ ]: resampled_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'].
    ↪drop(columns=['Tag']).resample('H').mean()
resampled_dataframe['Hour'] = resampled_dataframe.index.hour
resampled_dataframe['Count'] = (valid_dataframe['measuring'].resample('H').
    ↪count())
resampled_dataframe['Std'] = (valid_dataframe['measuring'].resample('H').std())
resampled_dataframe.index = resampled_dataframe.index.map(lambda t: t.
    ↪replace(minute=30, second=0))
resampled_dataframe['measuring'].plot()
```

```
[ ]: <Axes: xlabel='DateTime'>
```

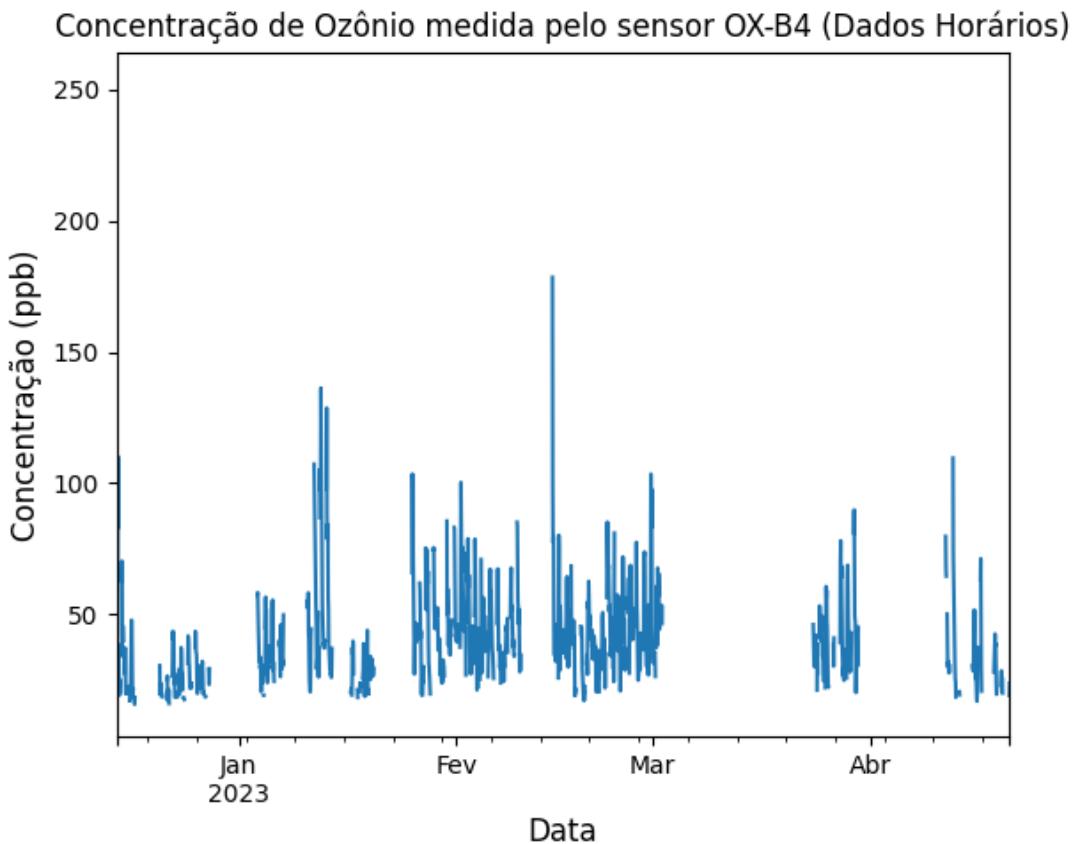


Tag hourly data according to the number of samples in an hour At least 3 samples must be valid in an hour (75 %) for the hourly data be considered as valid

```
[ ]: original_freq = sensor_dataframe.index.freq
resampled_dataframe['% valid'] = (resampled_dataframe['Count']
                                    .map(lambda c:
                                         c / (pd.Timedelta("1 hour")) /_
                                         original_freq) * 100)
resampled_dataframe['Tag'] = (resampled_dataframe['% valid']
                                .map(lambda c: 'VALID' if c >= 75 else_
                                'LOWSAMPLES'))

fig, ax = plt.subplots(figsize=(1.3*5,5))
resampled_dataframe[resampled_dataframe['Tag'] == 'VALID']['measuring']._
    rename('Concentração de Ozônio').resample('1H').mean().plot()
plt.title('Concentração de Ozônio medida pelo sensor OX-B4 (Dados Horários)')
ax.set_xlabel('Data', fontsize=12)
ax.set_ylabel('Concentração (ppb)', fontsize=12)

[ ]: Text(0, 0.5, 'Concentração (ppb)')
```



Valid data contabilization

```
[ ]: tags = ['LOWSAMPLES', 'VALID']
data_contabilization = sensor_analysis.count_tags(tags, resampled_dataframe)
data_contabilization
```

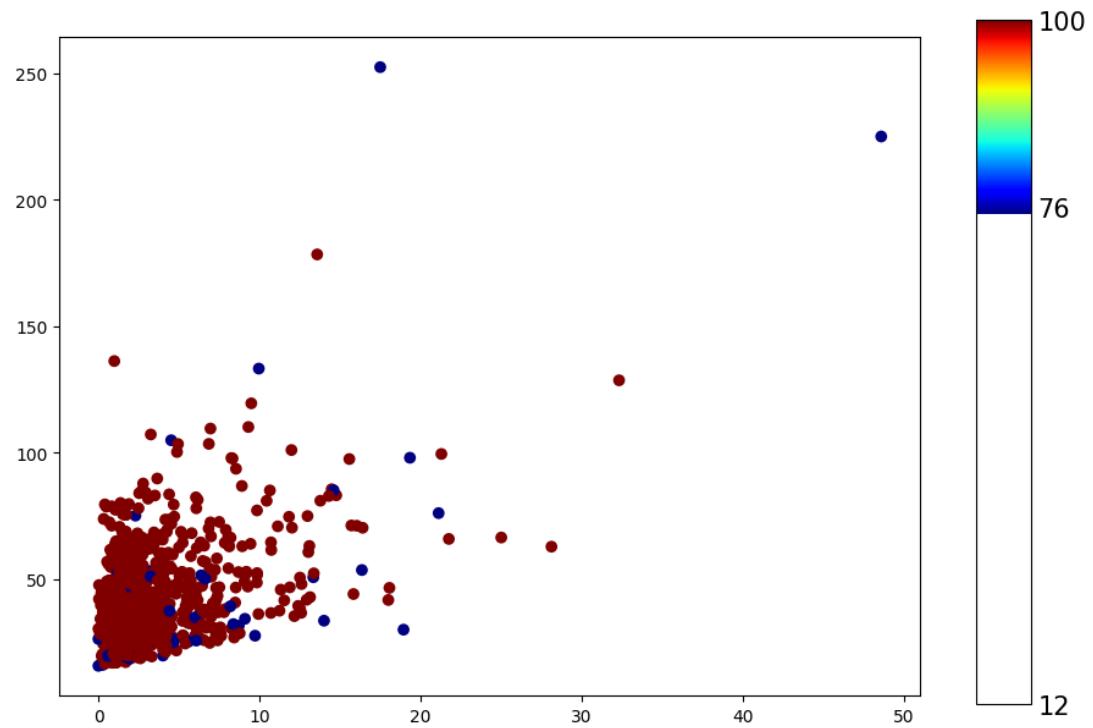
```
[ ]: # %
LOWSAMPLES 2020 65.669701
VALID 1056 34.330299
TOTAL 3076 100.0
```

2.7 Analyse the mean and standard deviation of the resampled data

The color of each data point represents the percentage of valid 15 mins samples found in an hour

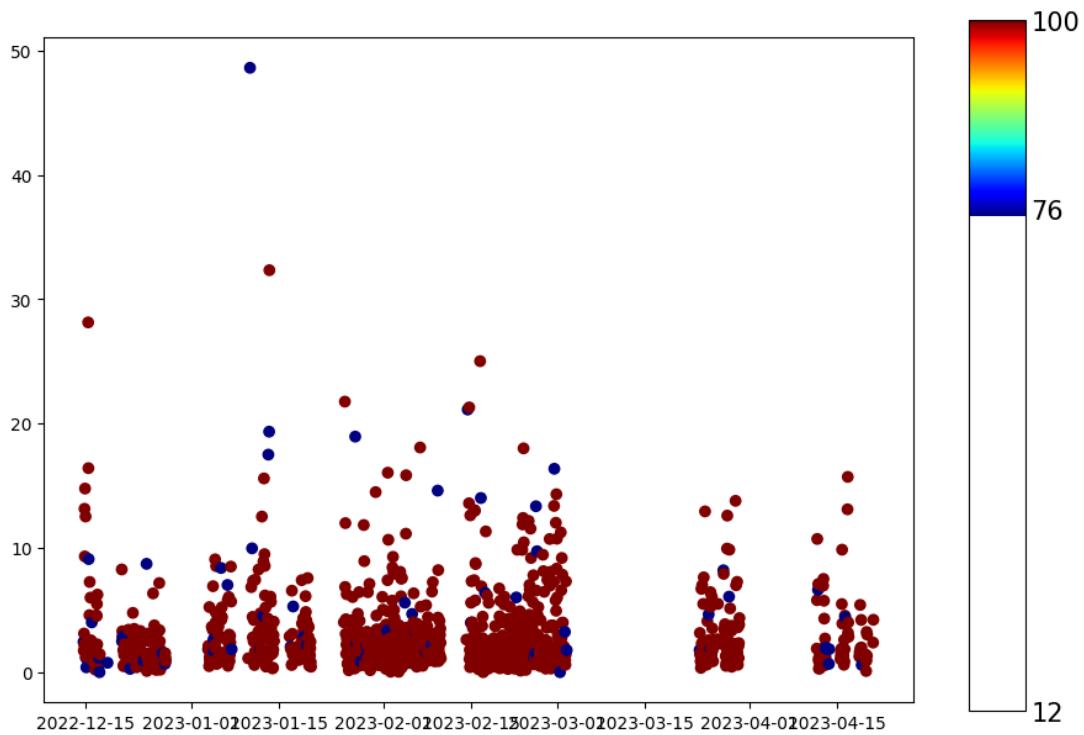
Plot mean (Y axis) vs. standard deviation (X axis) for valid hourly data colored by valid percentage of valid samples

```
[ ]: valid_resampled_dataframe = resampled_dataframe[resampled_dataframe['Tag'] == 'VALID']
sensor_analysis.plot_mean_vs_std(valid_resampled_dataframe)
```



2.7.1 Plot standard deviation in time colored according to the percentage of valid samples

```
[ ]: sensor_analysis.plot_std_in_time(valid_resampled_dataframe)
```



3 Save Data

```
[ ]: import os

output_directory_path = 'data/output/'
processing_directory_path = '../data-processing/input/'
valid_1HR_filename = '_valid_data_1HR.csv'
valid_filename = '_valid_data.csv'
dataframe_filename = '_dataframe.csv'
dataframe_1HR_filename = '_dataframe_1hr.csv'

valid_1HR_file_path_output = output_directory_path + SENSOR_NAME +_
    ↪valid_1HR_filename
valid_file_path_output = output_directory_path + SENSOR_NAME + valid_filename

dataframe_path_1HR_output = output_directory_path + SENSOR_NAME +_
    ↪dataframe_1HR_filename
dataframe_path_output = output_directory_path + SENSOR_NAME + dataframe_filename

valid_processing_1HR_file_path = processing_directory_path + SENSOR_NAME +_
    ↪valid_1HR_filename
```

```

valid_processing_file_path = processing_directory_path + SENSOR_NAME +_
    ↪valid_filename

processing_dataframe_path_1HR = processing_directory_path + SENSOR_NAME +_
    ↪dataframe_1HR_filename
processing_dataframe_path = processing_directory_path + SENSOR_NAME +_
    ↪dataframe_filename

if not os.path.exists(output_directory_path):
    os.makedirs(output_directory_path)

if not os.path.exists(processing_directory_path):
    os.makedirs(processing_directory_path)

sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'][['measuring', 'value']]._
    ↪to_csv(valid_file_path_output)
resampled_dataframe[resampled_dataframe['Tag'] == 'VALID'][['measuring', ↪
    'value']].to_csv(valid_1HR_file_path_output)
sensor_dataframe.to_csv(dataframe_path_output)
resampled_dataframe.to_csv(dataframe_path_1HR_output)

sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'][['measuring', 'value']]._
    ↪to_csv(valid_processing_file_path)
resampled_dataframe[resampled_dataframe['Tag'] == 'VALID'][['measuring', ↪
    'value']].to_csv(valid_processing_1HR_file_path)
sensor_dataframe.to_csv(processing_dataframe_path)
resampled_dataframe.to_csv(processing_dataframe_path_1HR)

```

O3_2-field-data-treatment-Diamante

February 13, 2024

1 Electrochemical O3 sensor data preprocessing

- Pollutant: Ozone
- Sensor: Alphasense OX-B4

1.1 Constants

1.1.1 Sensors IDs

1.1.2 API Constants

```
[ ]: HOST = "renovar.lcqar.ufsc.br"
PORT = 8080
RAW_DATA_DIR = "data/raw-data-monit-fix-2022-2023-Diamante/"
RAW_FILE_NAME = "ISB_032.CSV"

SENSOR_FILE_DIR = 'data/input/'
SENSOR_NAME = 'alpha_o3_2_conc'
SENSOR_FILE_NAME = SENSOR_NAME + 'web_dataframe.csv'
```

1.1.3 Sensor Constants

```
[ ]: M = 48
lower_limit=15.0
upper_limit=20e3
t_90 = 80      # sensor takes 30 seconds to reach a value of 10e3
t_90_value = 1e3
sampling_period = 15 * 60
```

2 Alphasense O3 (2) Sensor Data

```
[ ]: import locale
locale.setlocale(locale.LC_TIME, 'pt_BR')
```

```
[ ]: 'pt_BR'
```

```
[ ]: from GetSensorDataService import GetSensorDataService
```

```

get_sensor_data_service = GetSensorDataService(HOST, PORT)
sensor_data = get_sensor_data_service.
    ↪get_data_from_file(RAW_DATA_DIR+RAW_FILE_NAME, sensor_name=SENSOR_NAME)
sensor_data.head()

/Users/Fernando/Documents/Projects/Github/lcqar-low-cost-monit-proc/data-pre-
processing/GetSensorDataService.py:13: UserWarning: The argument
'infer_datetime_format' is deprecated and will be removed in a future version. A
strict version of it is now the default, see
https://pandas.pydata.org/pdps/0004-consistent-to-datetime-parsing.html. You
can safely remove this argument.
df['DateTime'] = (pd.to_datetime(df[date_time_col],
infer_datetime_format=False, format='%d/%m/%Y/%H/%M/%S'))

[ ]:   latitude longitude measuring           DateTime
0 -28.456899 -48.972999    27.48 2022-11-21 10:41:59
1 -28.456899 -48.972999    27.61 2022-11-21 10:57:45
2 -28.456899 -48.972999    27.92 2022-11-21 11:13:34
3 -28.456899 -48.972999    28.17 2022-11-21 11:29:19
4 -28.456899 -48.972999    28.34 2022-11-21 11:45:05

```

2.1 Upload Data from File

```

[ ]: import pandas as pd

df = pd.read_csv(SENSOR_FILE_DIR + SENSOR_FILE_NAME)
df.head()

[ ]:   Unnamed: 0   latitude   longitude   measuring           DateTime
0          0 -28.456899 -48.972999    27.48 2022-11-21 10:41:59
1          1 -28.456899 -48.972999    27.61 2022-11-21 10:57:45
2          2 -28.456899 -48.972999    27.92 2022-11-21 11:13:34
3          3 -28.456899 -48.972999    28.17 2022-11-21 11:29:19
4          4 -28.456899 -48.972999    28.34 2022-11-21 11:45:05

```

2.1.1 Create Sensor Dataframe as Pandas Series with a period of 15 mins

```

[ ]: # Remove the first column with the indexes and save data into web dataframe
web_dataframe = df.drop(df.columns[0], axis='columns')
web_dataframe['DateTime'] = (pd.to_datetime(df['DateTime']), ↪
    infer_datetime_format=True)

# Resample data with 15 mins period and create sensor dataframe
sensor_dataframe = web_dataframe.sort_values(by='DateTime', ascending=True).
    ↪reset_index().drop(columns='index')
sensor_dataframe.index = sensor_dataframe['DateTime']
sensor_dataframe = sensor_dataframe.drop(columns=['DateTime'])

```

```

sensor_dataframe = sensor_dataframe.resample('15T').mean()
sensor_dataframe

/var/folders/wc/_83zcrx913j1dqwg4g90kbhh0000gp/T/ipykernel_2814/166902210.py:3:
UserWarning: The argument 'infer_datetime_format' is deprecated and will be
removed in a future version. A strict version of it is now the default, see
https://pandas.pydata.org/pdeps/0004-consistent-to-datetime-parsing.html. You
can safely remove this argument.

    web_dataframe['DateTime'] = (pd.to_datetime(df['DateTime'],
infer_datetime_format=True))

[ :]:          latitude  longitude  measuring
DateTime
2022-11-21 10:30:00 -28.456899 -48.972999      27.48
2022-11-21 10:45:00 -28.456899 -48.972999      27.61
2022-11-21 11:00:00 -28.456899 -48.972999      27.92
2022-11-21 11:15:00 -28.456899 -48.972999      28.17
2022-11-21 11:30:00           NaN         NaN       NaN
...
          ...   ...
2023-04-21 20:45:00 -28.456899 -48.972999      18.83
2023-04-21 21:00:00 -28.456899 -48.972999      18.74
2023-04-21 21:15:00 -28.456899 -48.972999      18.77
2023-04-21 21:30:00 -28.456899 -48.972999      19.23
2023-04-21 21:45:00 -28.456899 -48.972999      19.48

[14542 rows x 3 columns]

```

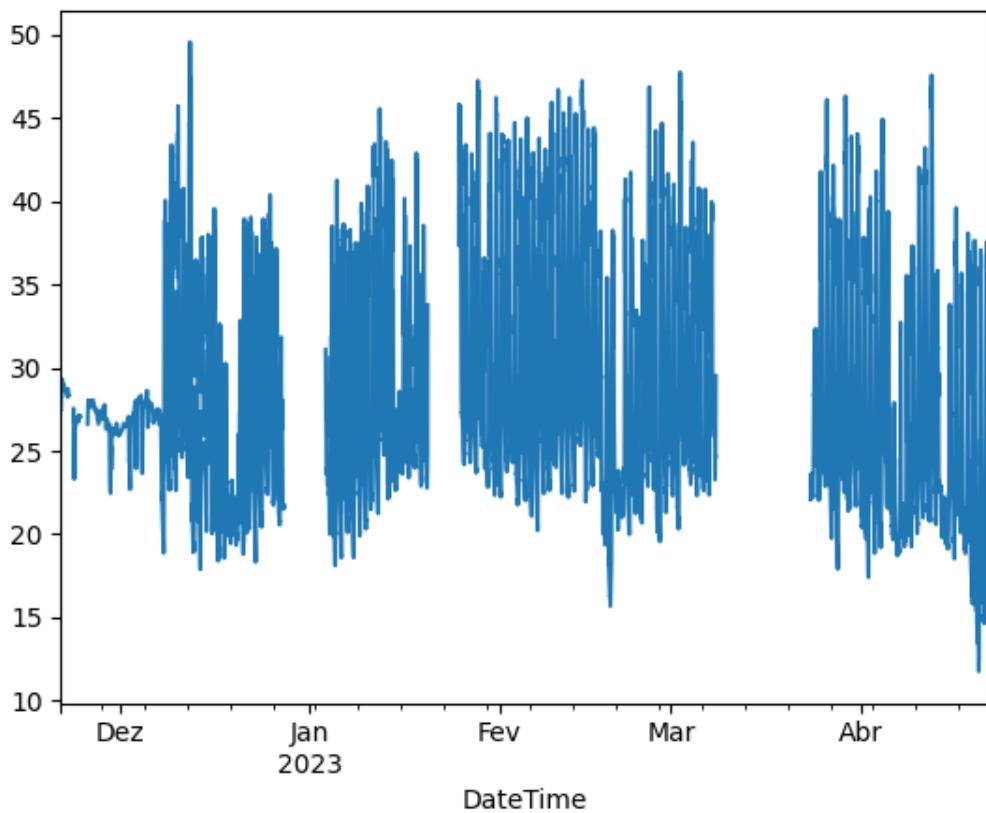
2.1.2 Plot raw data

```

[ :]: sensor_dataframe['measuring'].plot()

[ :]: <Axes: xlabel='DateTime'>

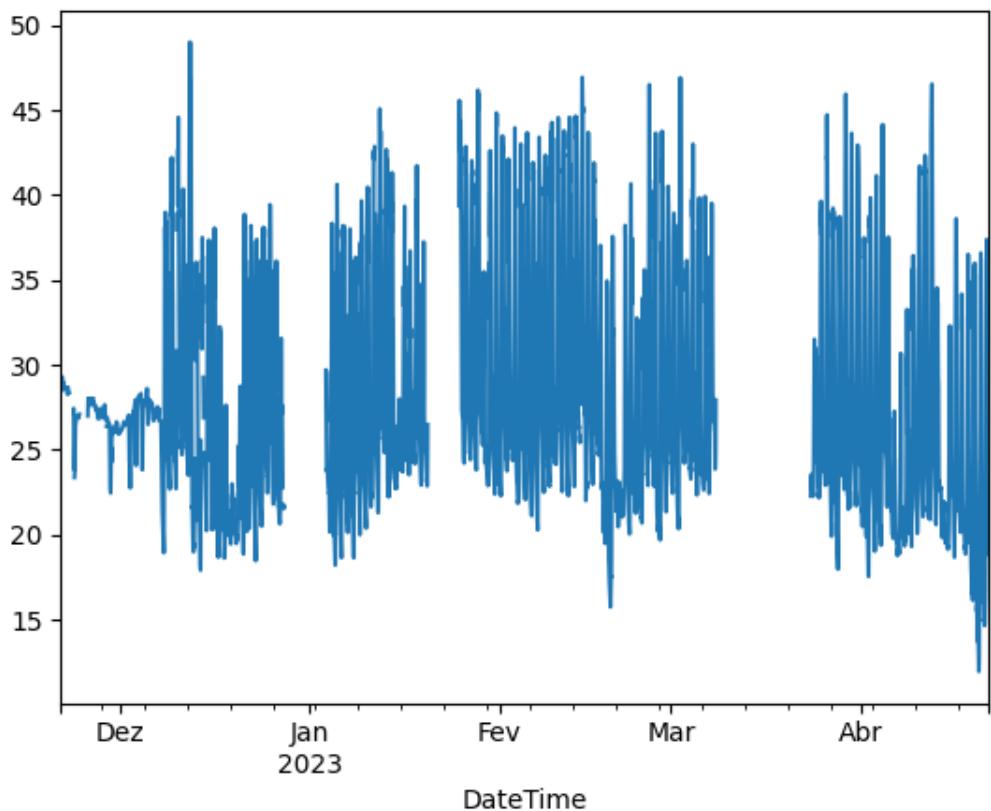
```



2.1.3 Smooth data with a window of 4 samples, i.e.: an hour of data

```
[ ]: sensor_dataframe['raw_measuring'] = sensor_dataframe['measuring']
sensor_dataframe['measuring'] = sensor_dataframe['measuring'].rolling(window=4).median()
sensor_dataframe['Hour'] = sensor_dataframe.index.hour
sensor_dataframe['measuring'].plot()
```

```
[ ]: <Axes: xlabel='DateTime'>
```



2.1.4 Convert to ug/m³

```
[ ]: sensor_dataframe['value'] = sensor_dataframe['measuring'].map(lambda v: 0.0409*v*M)
```

2.2 Tag data

Apply tags to the data according to the quality control processes described above

```
[ ]: sensor_dataframe['Tag'] = 'VALID'
```

Tag data during stabilizing period AQMesh recommends to considering the first two days after installation as a period for stabilization. In our case, a period of seven days was selected in order to remove outliers candidates detected during that period. For that purpose we detect the first 7 days of uninterrupted measurements

```
[ ]: # Identify consecutive measurements (15 mins between each)
consecutive_periods = sensor_dataframe.index.to_series().diff().dt.seconds == 900
```

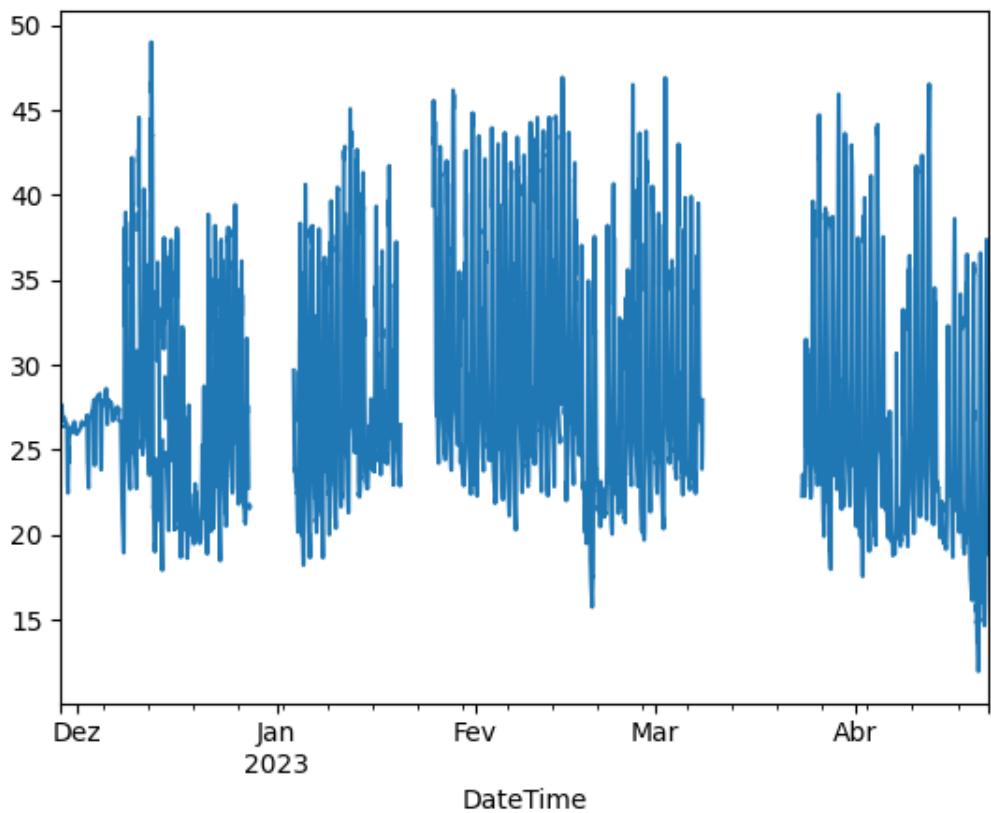
```

# Select the first seven days of consecutive measurements (96 = 24 * 60 mins / ↳
↪15 mins, amount of 15 mins periods in a day)
last_stabilizing_index = sensor_dataframe[consecutive_periods].head(96 * 7).
↪index[-1] # 7 days
sensor_dataframe.loc[sensor_dataframe.index[0]:last_stabilizing_index, 'Tag'] = ↳
↪'STABILIZING'
sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['measuring'].
↪resample('15T').mean().plot()
sensor_dataframe.head()

```

```
[ ]:          latitude longitude measuring raw measuring Hour \
DateTime
2022-11-21 10:30:00 -28.456899 -48.972999      NaN     27.48    10
2022-11-21 10:45:00 -28.456899 -48.972999      NaN     27.61    10
2022-11-21 11:00:00 -28.456899 -48.972999      NaN     27.92    11
2022-11-21 11:15:00 -28.456899 -48.972999  27.765     28.17    11
2022-11-21 11:30:00         NaN        NaN      NaN     NaN     11

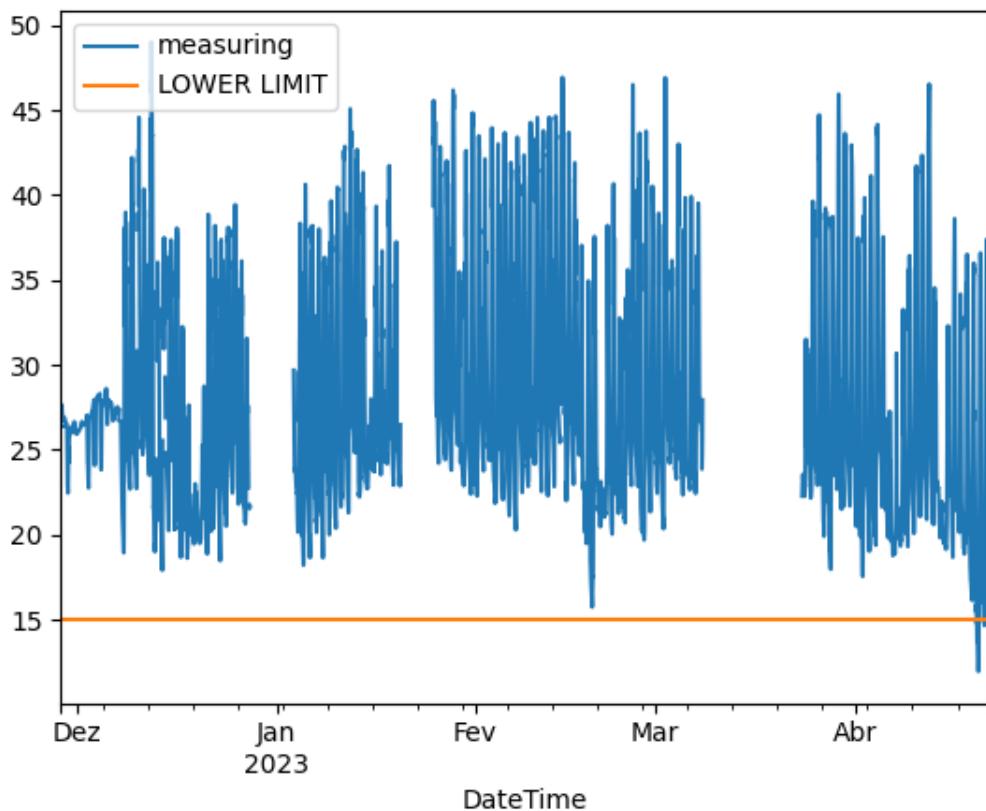
                           value      Tag
DateTime
2022-11-21 10:30:00      NaN STABILIZING
2022-11-21 10:45:00      NaN STABILIZING
2022-11-21 11:00:00      NaN STABILIZING
2022-11-21 11:15:00  54.508248 STABILIZING
2022-11-21 11:30:00      NaN STABILIZING
```



Tag values that are missing, are greater than sensor span or are lower than sensor resolution The upper limit was too high that affected the graph resolution, therefore it is not plotted

```
[ ]: sensor_dataframe['UPPER LIMIT'] = upper_limit
sensor_dataframe['LOWER LIMIT'] = lower_limit
sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'][['measuring', 'LOWER LIMIT']].resample('15T').mean().plot()
```

```
[ ]: <Axes: xlabel='DateTime'>
```



```
[ ]: import numpy as np
from SensorDataAnalysisService import SensorDataAnalysisService as
    ↪sensor_analysis

def tag_by_sensor_limits(value, tag, lower_limit, upper_limit):
    if (tag != 'VALID'): return tag
    return sensor_analysis.get_tags_from_series(value=value,
                                                lower_limit=lower_limit,
                                                upper_limit=upper_limit)

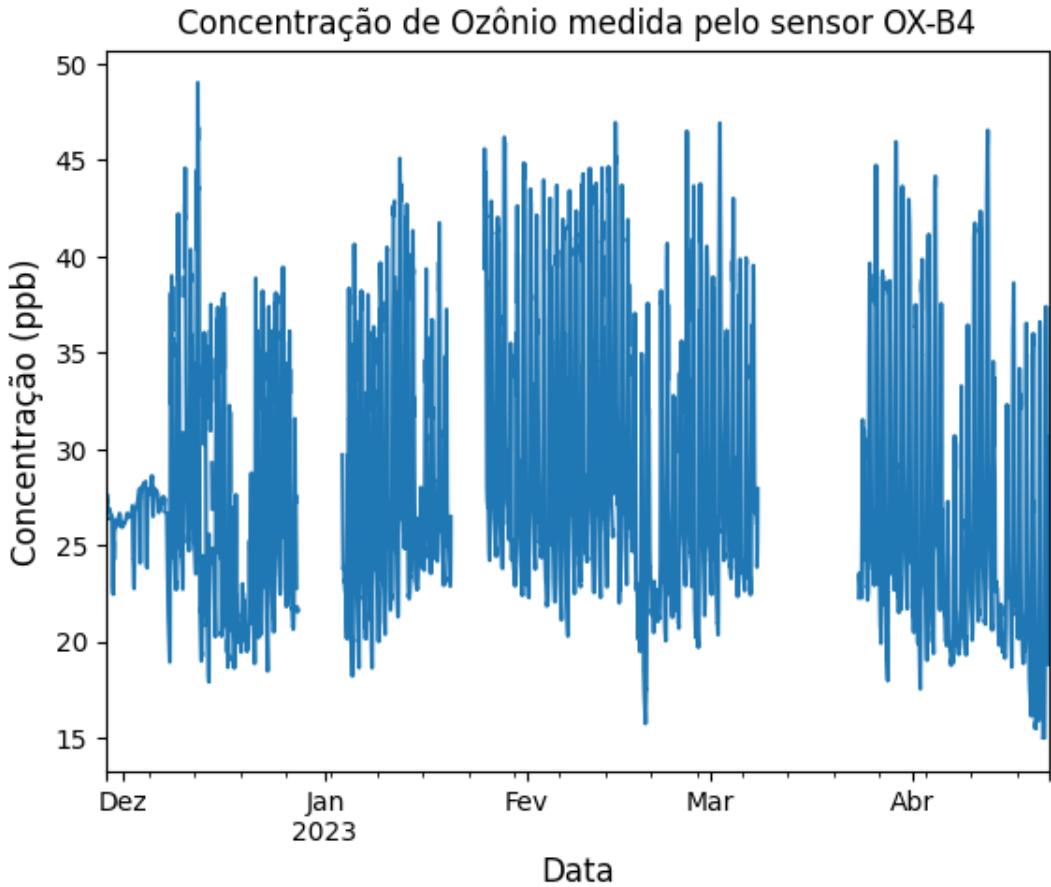
sensor_dataframe['Tag'] = (sensor_dataframe[['measuring', 'Tag']]
                           .apply(lambda df: ↪
                                 ↪tag_by_sensor_limits(value=df[0], tag=df[1],
                                 ↪lower_limit=lower_limit,
                                 ↪upper_limit=upper_limit),
                                 axis=1))
```

Plot valid data after removing samples bellow lower limit

```
[ ]: import matplotlib.pyplot as plt

fig, ax = plt.subplots(figsize=(1.3*5,5))
sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['measuring'].  
    ↪rename('Concentração de Ozônio').resample('15T').mean().plot()
plt.title('Concentração de Ozônio medida pelo sensor OX-B4')
ax.set_xlabel('Data', fontsize=12)
ax.set_ylabel('Concentração (ppb)', fontsize=12)
```

```
[ ]: Text(0, 0.5, 'Concentração (ppb)')
```



2.2.1 Data Tag contabilization

```
[ ]: tags = ['MISSING', 'LTLL', 'GTUL', 'STABILIZING', 'BADSPIKE', 'VALID']
data_contabilization = sensor_analysis.count_tags(tags, sensor_dataframe)
data_contabilization
```

```
[ ]: # %
MISSING      2734  18.800715
LTLL         49   0.336955
GTUL         0    0.0
STABILIZING  673   4.627974
BADSPIKE     0    0.0
VALID        11086 76.234356
TOTAL        14542 100.0
```

2.3 Analyse valid data

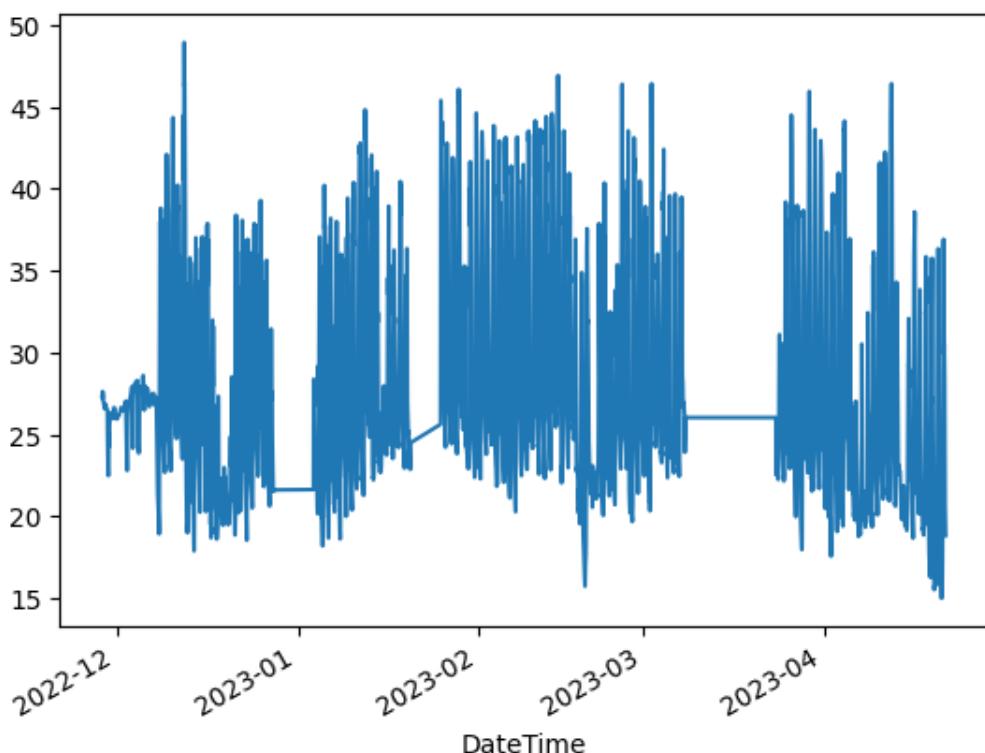
The graph shows the time series of the valid data, box plots of the data grouped by hour of the day, and histogram of the data.

```
[ ]: valid_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'] .
    .drop(columns=['Tag'])
```

2.3.1 Smooth data with a window of 4 samples, i.e.: an hour of data

```
[ ]: valid_dataframe['raw measuring'] = valid_dataframe['measuring']
valid_dataframe['measuring'] = valid_dataframe['measuring'].rolling(window=4) .
    .median()
valid_dataframe['measuring'].plot()
```

```
[ ]: <Axes: xlabel='DateTime'>
```

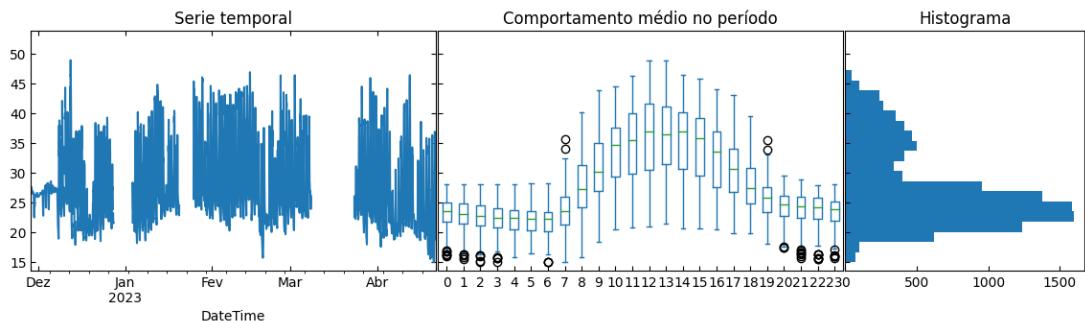


2.3.2 Analyse data

```
[ ]: valid_dataframe = valid_dataframe.resample('15T').mean()
sensor_analysis.plot_box_hist(df=valid_dataframe, bins=20)
```

/Users/Fernando/Documents/Projects/Github/lcqar-low-cost-monit-proc/data-pre-processing/SensorDataAnalysisService.py:111: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
df['Hour'] = df['Hour'].astype('int64')



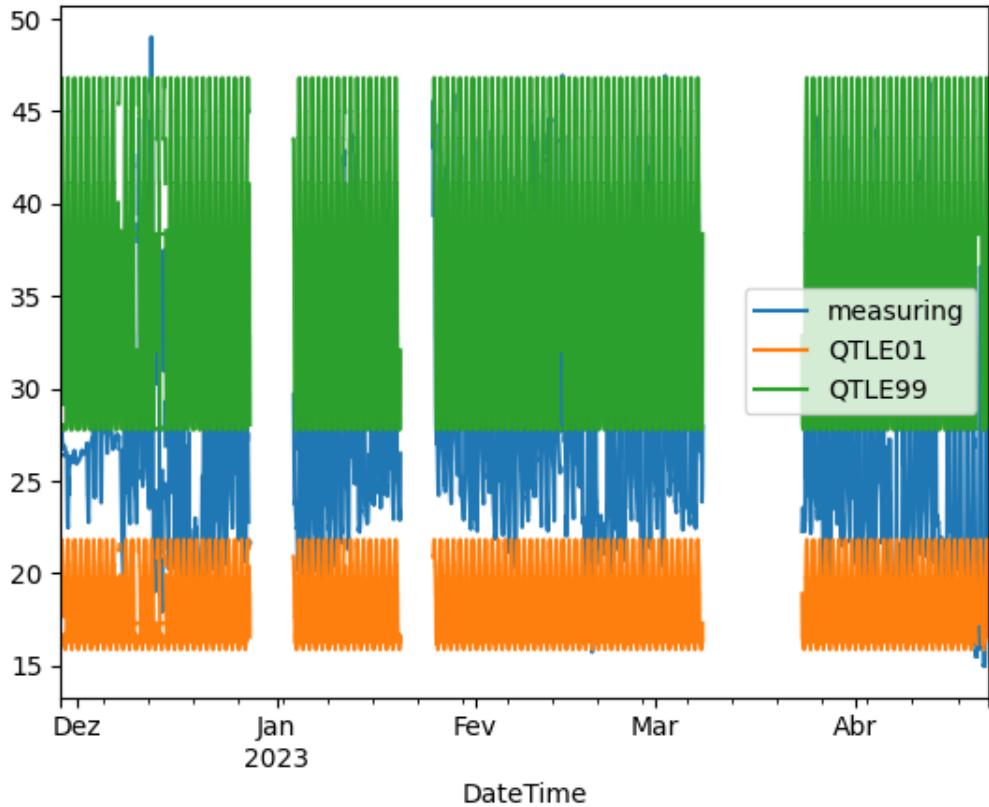
2.3.3 Calculate quantiles

We calculated the 1% and 99 % quantiles of every hour of the day. The values greater outside the 1 % - 99 % were tagged as Greater than Quantile 99 (GTQTL99) and Lower than Quantile 1 (LTQTL01)

```
[ ]: import numpy as np

global_qtle_01 = valid_dataframe.pivot(columns='Hour')['measuring'].
    ↪quantile(q=0.01, axis='index', interpolation='lower').dropna()
global_qtle_99 = valid_dataframe.pivot(columns='Hour')['measuring'].
    ↪quantile(q=0.99, axis='index', interpolation='higher').dropna()
sensor_dataframe['QTLE01'] = sensor_dataframe['Hour'].map(lambda hr: ↪
    ↪global_qtle_01[hr] if (not np.isnan(hr)) else np.nan)
sensor_dataframe['QTLE99'] = sensor_dataframe['Hour'].map(lambda hr: ↪
    ↪global_qtle_99[hr] if (not np.isnan(hr)) else np.nan)
sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'][['measuring', 'QTLE01', ↪
    ↪'QTLE99']].resample('15T').mean().plot()
```

```
[ ]: <Axes: xlabel='DateTime'>
```



2.3.4 Tag data according to quantiles

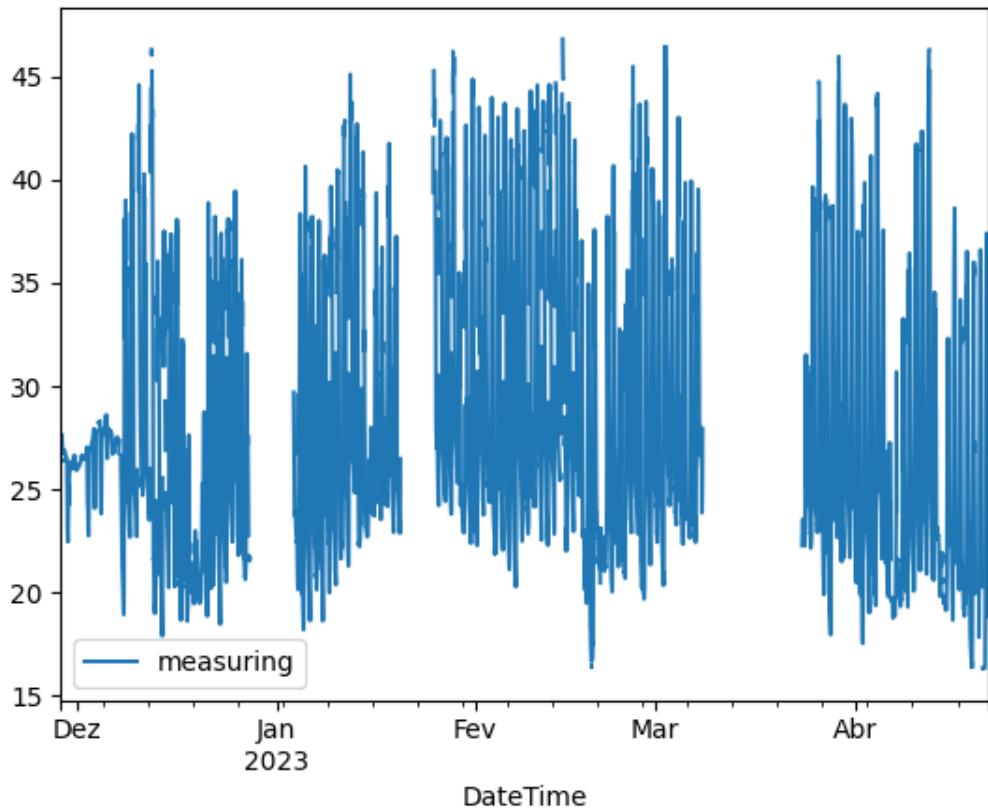
```
[ ]: sensor_dataframe['Tag'] = (sensor_dataframe[['Tag', 'measuring', 'QTLE01',  
        ↪ 'QTLE99']]  
        .apply(lambda df: sensor_analysis.  
        ↪ tag_by_quantiles(current_tag=df[0],  
        ↪ value=df[1],  
        ↪ quantile_01=df[2],  
        ↪ quantile_99=df[3]),  
        ↪ axis=1))  
tags = ['MISSING', 'LTLL', 'GTUL', 'BADSPIKE', 'VALID', 'LTQTLE01', 'GTQTLE99']  
data_contabilization = sensor_analysis.count_tags(tags, sensor_dataframe)  
data_contabilization
```

```
[ ]: # %
MISSING    2734  18.800715
LTLL        49   0.336955
GTUL        0    0.0
BADSPIKE    0    0.0
VALID      10814 74.363911
LTQTL01     125   0.859579
GTQTL99     147   1.010865
TOTAL      14542  100.0
```

Plot valid data

```
[ ]: sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'][['measuring']].  
    ↪resample('15T').mean().plot()
```

```
[ ]: <Axes: xlabel='DateTime'>
```



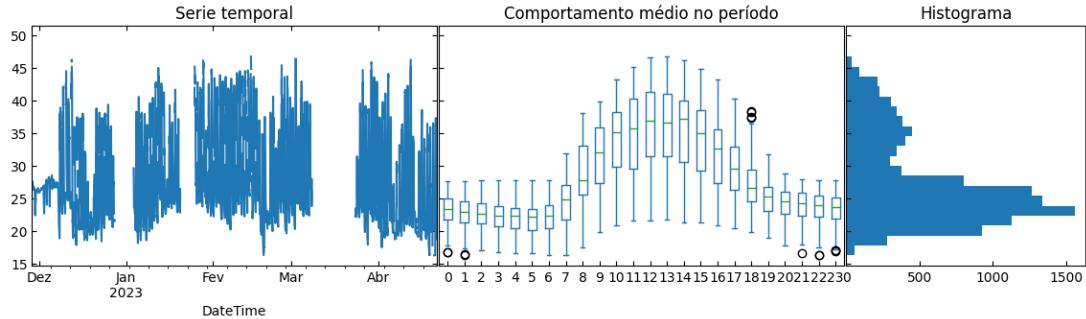
Analyse data after removing quantiles

```
[ ]: valid_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'].  
    ↪drop(columns=['Tag']).resample('15T').mean()  
sensor_analysis.plot_box_hist(df=valid_dataframe, bins=20)
```

```
/Users/Fernando/Documents/Projects/Github/lcqar-low-cost-monit-proc/data-pre-processing/SensorDataAnalysisService.py:111: SettingWithCopyWarning:  
A value is trying to be set on a copy of a slice from a DataFrame.  
Try using .loc[row_indexer,col_indexer] = value instead
```

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

```
df['Hour'] = df['Hour'].astype('int64')
```



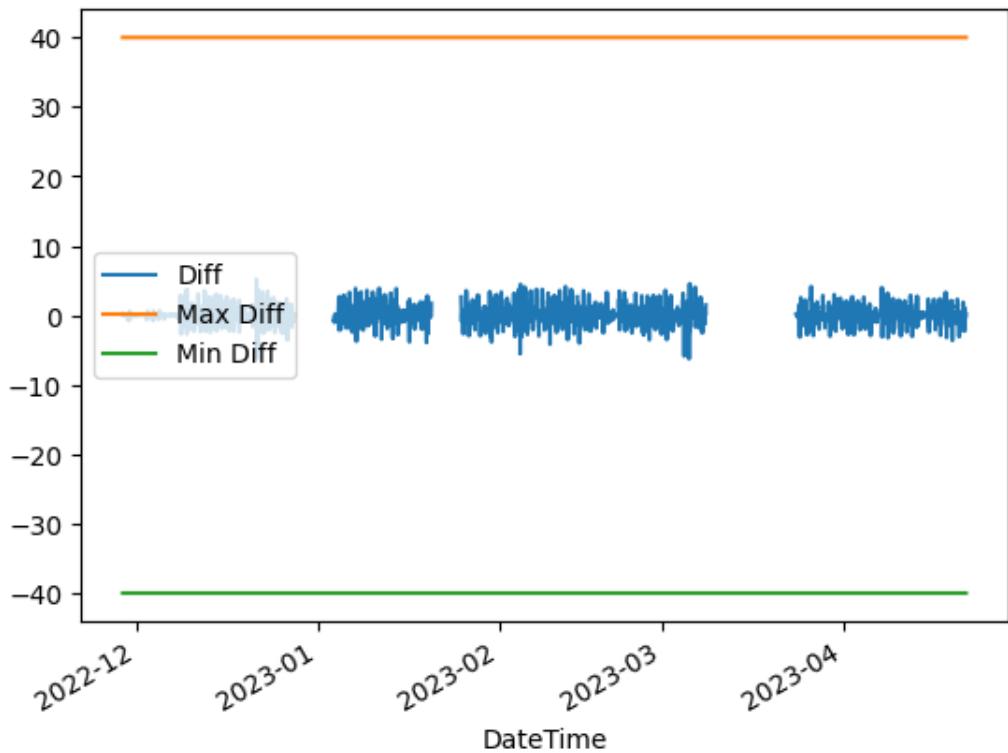
2.4 Analyse data derivatives

Analyse the derivatives of the data for removing abrupt changes in the data (BADSPIKE)

Tag values with derivatives faster than sensor response time The value of 40 ppb was selected based on the maximum derivative found in the reference data. The maximum derivative (with measuring period of 1 hour) found was -45 ppb. Therefore, a maximum of 40 ppb was considered appropriate for a 15 mins period.

```
[ ]: max_diff_value = 40  
sensor_dataframe['Diff'] = sensor_dataframe['measuring'].resample('15T').mean().  
    diff()  
sensor_dataframe['Max Diff'] = max_diff_value  
sensor_dataframe['Min Diff'] = -max_diff_value  
sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'][['Diff', 'Max Diff', 'Min Diff']].plot()
```

```
[ ]: <Axes: xlabel='DateTime'>
```



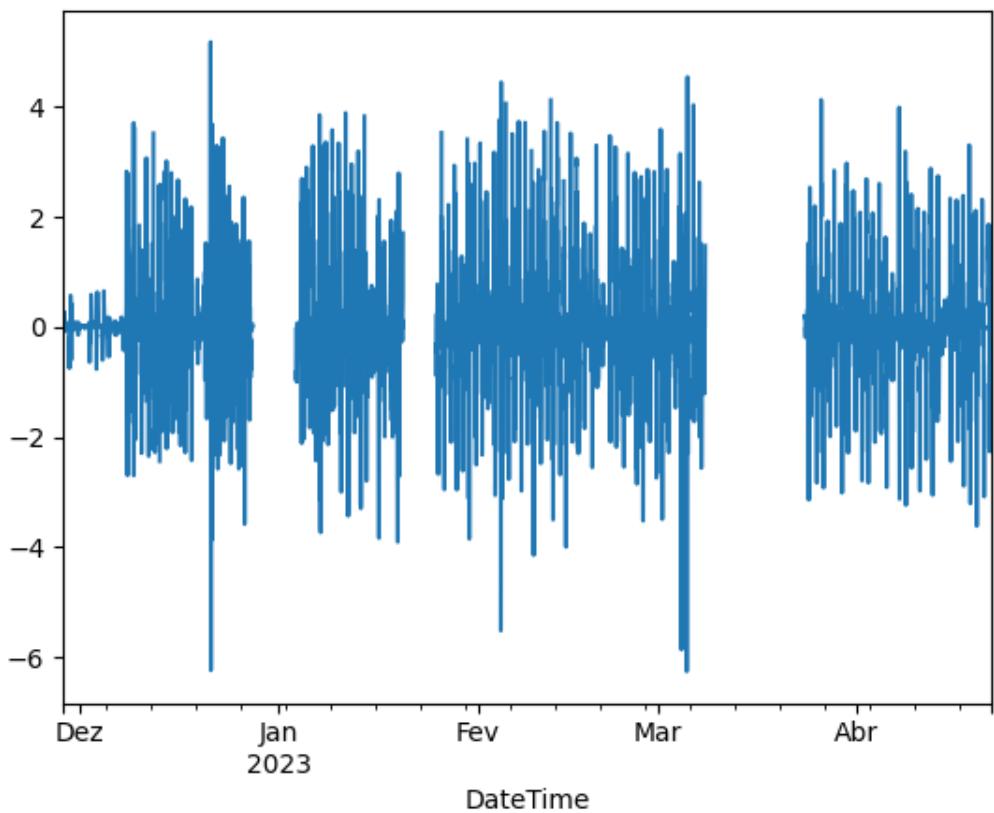
```
[ ]: import numpy as np

def tag_data_with_derivatives(tagged_df, max_diff_value):
    current_tag = tagged_df[0]
    value = tagged_df[1]
    if ((current_tag != 'VALID') or (np.isnan(value))): return current_tag
    if ((value > max_diff_value) or (value < -max_diff_value)): return
    ↵'BADSPIKE'
    return 'VALID'

sensor_dataframe['Tag'] = (sensor_dataframe[['Tag', 'Diff', 'Max Diff']]
                           .apply(lambda df:_
                                 tag_data_with_derivatives(tagged_df=df, max_diff_value=df[2]),
                                 axis=1))

[ ]: sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['Diff'].resample('15T').
    ↵mean().plot()

[ ]: <Axes: xlabel='DateTime'>
```

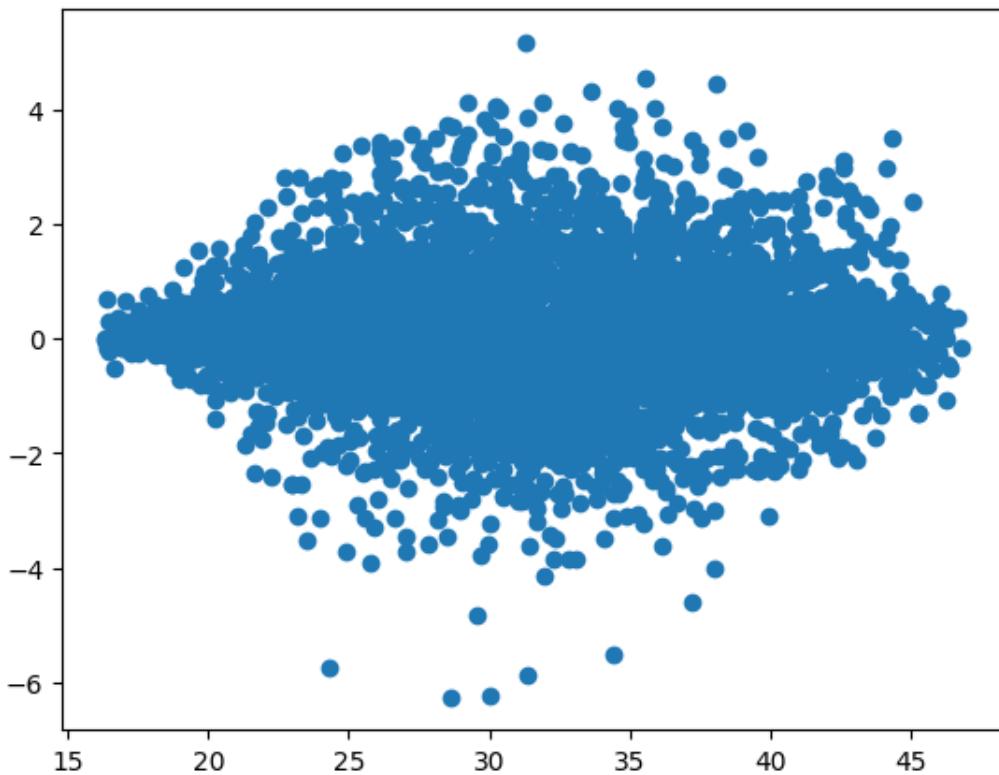


Plot data derivatives vs data

```
[ ]: import matplotlib.pyplot as plt

valid_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']
fig = plt.figure(figsize=(1.3*5,5))
plt.scatter(valid_dataframe['measuring'], valid_dataframe['Diff'])
```

```
[ ]: <matplotlib.collections.PathCollection at 0x16a977df0>
```



Data Tag contabilization

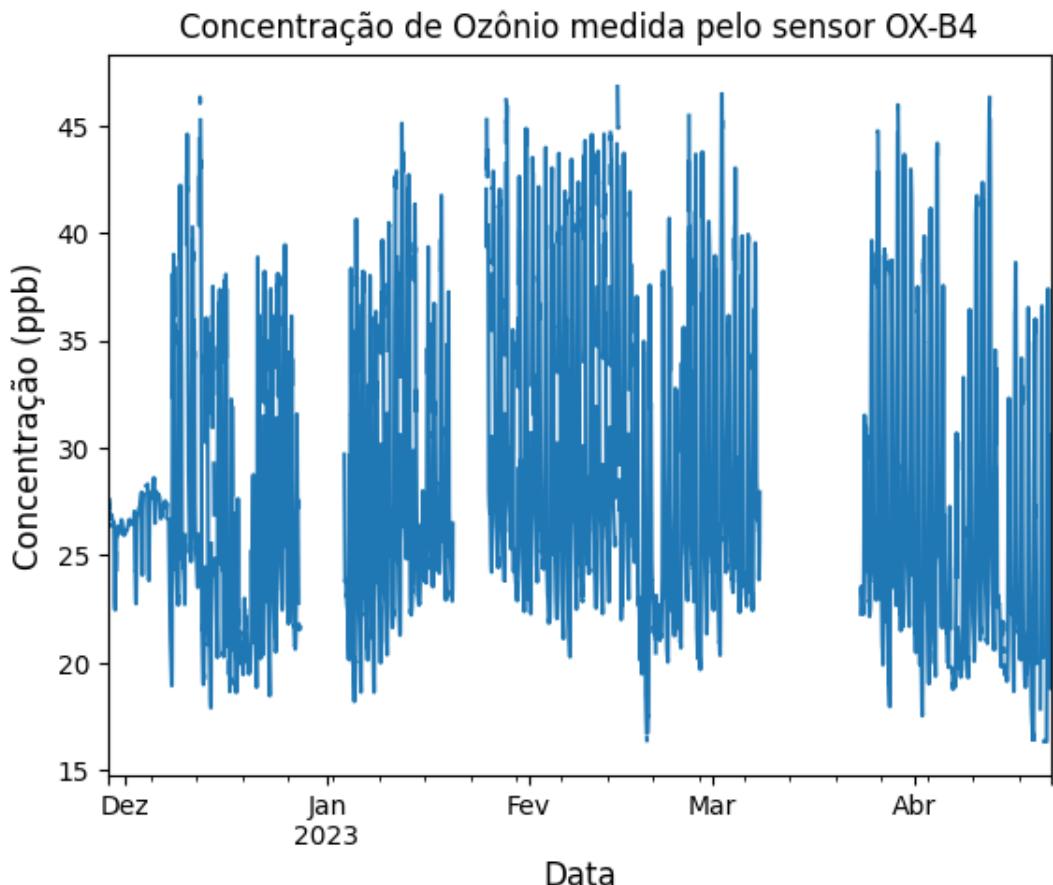
```
[ ]: tags = ['MISSING', 'LTLL', 'GTUL', 'STABILIZING', 'BADSPIKE', 'VALID',  
          'LTQTLE01', 'GTQTLE99', 'REBASE']  
data_contabilization = sensor_analysis.count_tags(tags, sensor_dataframe)  
data_contabilization
```

	#	%
MISSING	2734	18.800715
LTLL	49	0.336955
GTUL	0	0.0
STABILIZING	673	4.627974
BADSPIKE	0	0.0
VALID	10814	74.363911
LTQTLE01	125	0.859579
GTQTLE99	147	1.010865
REBASE	0	0.0
TOTAL	14542	100.0

Plot valid data

```
[ ]: fig, ax = plt.subplots(figsize=(1.3*5,5))
sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'][['measuring']]
    .rename('Concentração de Ozônio').resample('15T').mean().plot()
plt.title('Concentração de Ozônio medida pelo sensor OX-B4')
ax.set_xlabel('Data', fontsize=12)
ax.set_ylabel('Concentração (ppb)', fontsize=12)
```

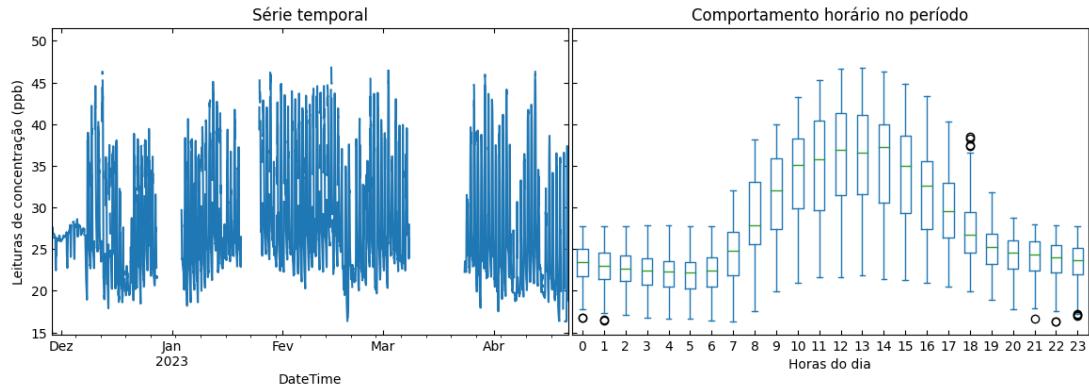
```
[ ]: Text(0, 0.5, 'Concentração (ppb)')
```



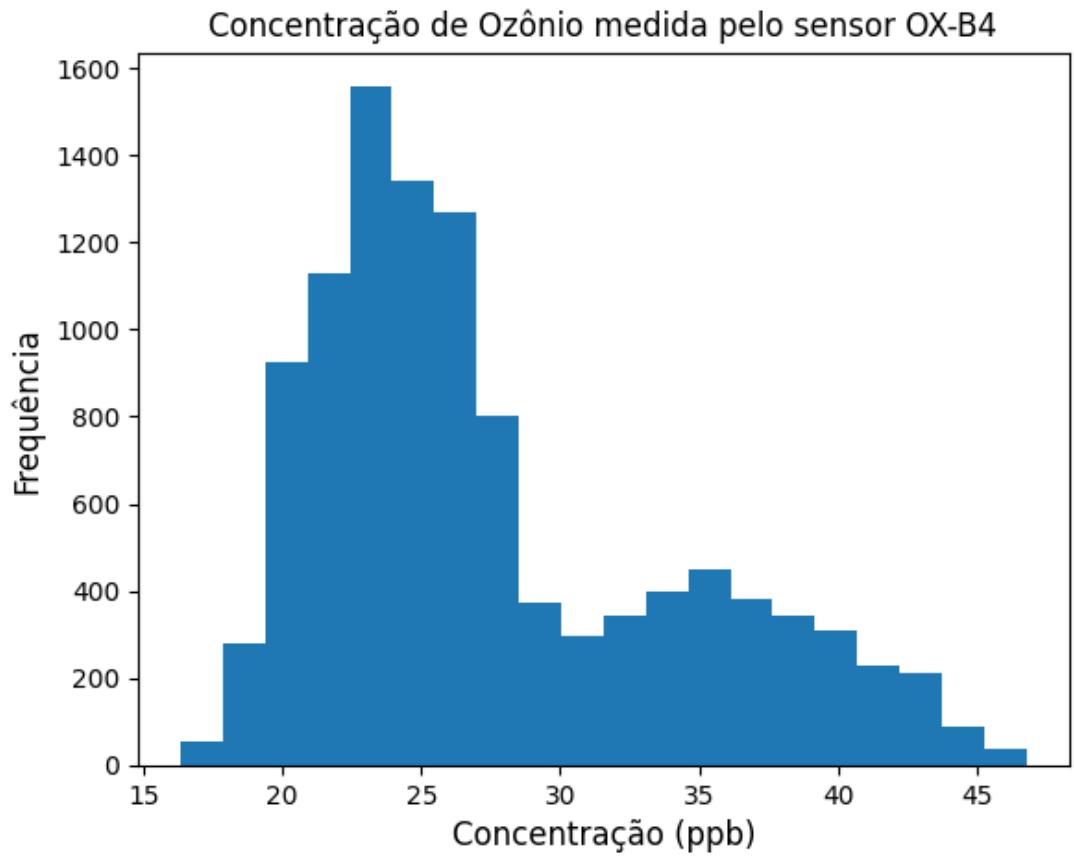
```
[ ]: valid_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'].
    .drop(columns=['Tag']).resample('15T').mean()
sensor_analysis.plot_box(df=valid_dataframe)
```

```
/Users/Fernando/Documents/Projects/Github/lcqar-low-cost-monit-proc/data-pre-
processing/SensorDataAnalysisService.py:143: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
`df['Hour'] = df['Hour'].astype('int64')`



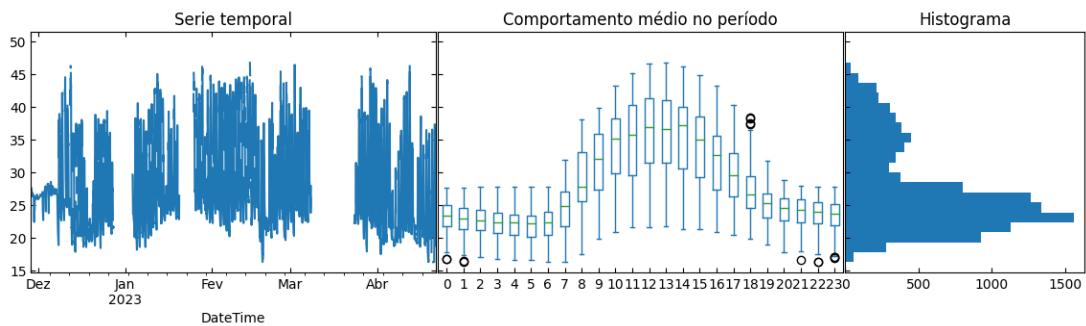
```
[ ]: fig, ax = plt.subplots(figsize=(1.3*5,5))
(sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['measuring'].
    rename('Concentração de Ozônio')
    .resample('15T').mean().hist(bins=20))
plt.title('Concentração de Ozônio medida pelo sensor OX-B4')
ax.set_xlabel('Concentração (ppb)', fontsize=12)
ax.set_ylabel('Frequência', fontsize=12)
ax.grid(False)
```



```
[ ]: valid_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'].
      drop(columns=['Tag']).resample('15T').mean()
sensor_analysis.plot_box_hist(df=valid_dataframe, bins=20)
```

```
/Users/Fernando/Documents/Projects/Github/lcqar-low-cost-monit-proc/data-pre-
processing/SensorDataAnalysisService.py:111: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

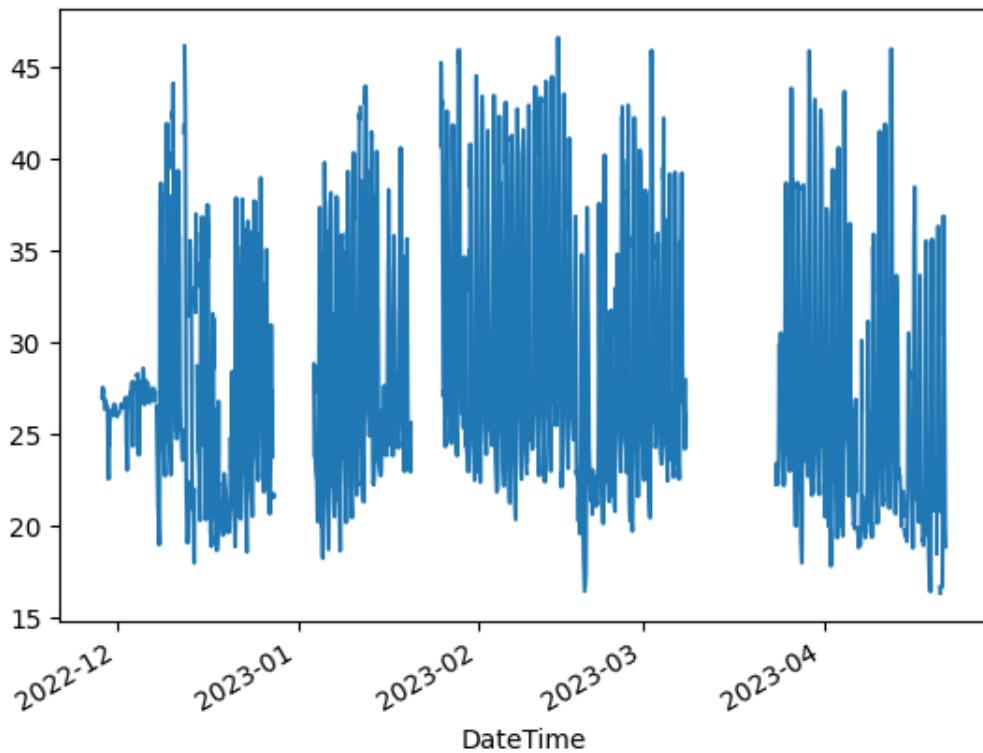
```
See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user\_guide/indexing.html#returning-a-view-versus-a-copy
df['Hour'] = df['Hour'].astype('int64')
```



2.5 Resample valid data to 1 HR

```
[ ]: resampled_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'].
    ↪drop(columns=['Tag']).resample('H').mean()
resampled_dataframe['Hour'] = resampled_dataframe.index.hour
resampled_dataframe['Count'] = (valid_dataframe['measuring'].resample('H').
    ↪count())
resampled_dataframe['Std'] = (valid_dataframe['measuring'].resample('H').std())
resampled_dataframe.index = resampled_dataframe.index.map(lambda t: t.
    ↪replace(minute=30, second=0))
resampled_dataframe['measuring'].plot()
```

```
[ ]: <Axes: xlabel='DateTime'>
```



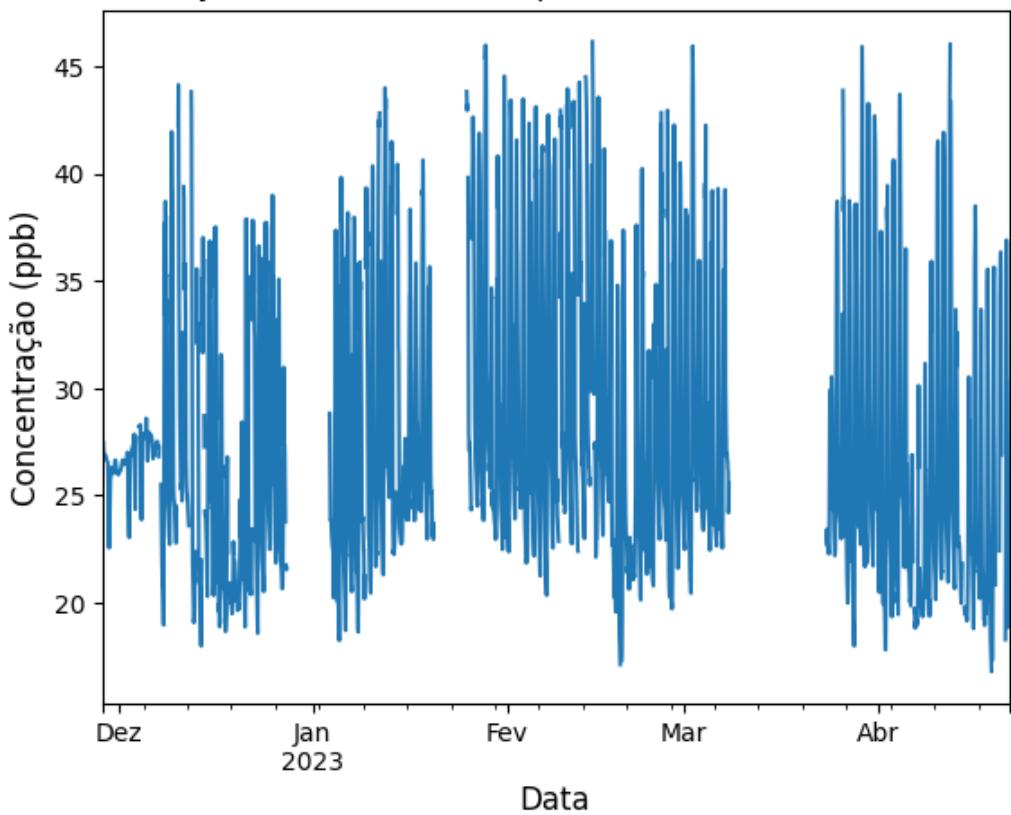
Tag hourly data according to the number of samples in an hour At least 3 samples must be valid in an hour (75 %) for the hourly data be considered as valid

```
[ ]: original_freq = sensor_dataframe.index.freq
resampled_dataframe['% valid'] = (resampled_dataframe['Count']
.map(lambda c:
c / (pd.Timedelta("1 hour") / ↴
original_freq) * 100))
resampled_dataframe['Tag'] = (resampled_dataframe['% valid']
.map(lambda c: 'VALID' if c >= 75 else ↴
'LOWSAMPLES'))

fig, ax = plt.subplots(figsize=(1.3*5,5))
resampled_dataframe[resampled_dataframe['Tag'] == 'VALID']['measuring'].\
rename('Concentração de Ozônio').resample('1H').mean().plot()
plt.title('Concentração de Ozônio medida pelo sensor OX-B4 (Dados Horários)')
ax.set_xlabel('Data', fontsize=12)
ax.set_ylabel('Concentração (ppb)', fontsize=12)
```

```
[ ]: Text(0, 0.5, 'Concentração (ppb)')
```

Concentração de Ozônio medida pelo sensor OX-B4 (Dados Horários)



Valid data contabilization

```
[ ]: tags = ['LOWSAMPLES', 'VALID']
data_contabilization = sensor_analysis.count_tags(tags, resampled_dataframe)
data_contabilization
```

```
[ ]: # %
LOWSAMPLES    783  22.577855
VALID        2685  77.422145
TOTAL        3468   100.0
```

3 Save Data

```
[ ]: import os

output_directory_path = 'data/output/'
processing_directory_path = '../data-processing/input/'
valid_1HR_filename = '_valid_data_1HR.csv'
valid_filename = '_valid_data.csv'
```

```

dataframe_filename = '_dataframe.csv'
dataframe_1HR_filename = '_dataframe_1hr.csv'

valid_1HR_file_path_output = output_directory_path + SENSOR_NAME +_
    ↪valid_1HR_filename
valid_file_path_output = output_directory_path + SENSOR_NAME + valid_filename

dataframe_path_1HR_output = output_directory_path + SENSOR_NAME +_
    ↪dataframe_1HR_filename
dataframe_path_output = output_directory_path + SENSOR_NAME + dataframe_filename

valid_processing_1HR_file_path = processing_directory_path + SENSOR_NAME +_
    ↪valid_1HR_filename
valid_processing_file_path = processing_directory_path + SENSOR_NAME +_
    ↪valid_filename

processing_dataframe_path_1HR = processing_directory_path + SENSOR_NAME +_
    ↪dataframe_1HR_filename
processing_dataframe_path = processing_directory_path + SENSOR_NAME +_
    ↪dataframe_filename

if not os.path.exists(output_directory_path):
    os.makedirs(output_directory_path)

if not os.path.exists(processing_directory_path):
    os.makedirs(processing_directory_path)

sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'][[['measuring', 'value']]]._
    ↪to_csv(valid_file_path_output)
resampled_dataframe[resampled_dataframe['Tag'] == 'VALID'][[['measuring', ↴
    ↪'value']]].to_csv(valid_1HR_file_path_output)
sensor_dataframe.to_csv(dataframe_path_output)
resampled_dataframe.to_csv(dataframe_path_1HR_output)

sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'][[['measuring', 'value']]]._
    ↪to_csv(valid_processing_file_path)
resampled_dataframe[resampled_dataframe['Tag'] == 'VALID'][[['measuring', ↴
    ↪'value']]].to_csv(valid_processing_1HR_file_path)
sensor_dataframe.to_csv(processing_dataframe_path)
resampled_dataframe.to_csv(processing_dataframe_path_1HR)

```

NO2-field-data-treatment-Diamante

February 13, 2024

1 Electrochemical NO₂ sensor data preprocessing

- Pollutant: Nitrogen Dioxide
- Sensor: Alphasense NO-B43F

1.1 Constants

1.1.1 Sensors IDs

1.1.2 API Constants

```
[ ]: HOST = "renovar.lcqar.ufsc.br"
PORT = 8080
RAW_DATA_DIR = "data/raw-data-monit-fix-2022-2023-Diamante/"
RAW_FILE_NAME = "ISB_NO2.CSV"

SENSOR_FILE_DIR = 'data/input/'
SENSOR_NAME = 'alpha_no2_conc'
SENSOR_FILE_NAME = SENSOR_NAME + 'web_dataframe.csv'
```

1.1.3 Sensor Constants

```
[ ]: M = 46.0055
lower_limit=15.0
upper_limit=20e3
t_90 = 80      # sensor takes 30 seconds to reach a value of 10e3
t_90_value = 2e3
sampling_period = 15 * 60
```

2 Alphasense NO₂ Sensor Data

```
[ ]: import locale
locale.setlocale(locale.LC_TIME, 'pt_BR')

[ ]: 'pt_BR'

[ ]: from GetSensorDataService import GetSensorDataService
```

```

get_sensor_data_service = GetSensorDataService(HOST, PORT)
sensor_data = get_sensor_data_service.
    ↪get_data_from_file(RAW_DATA_DIR+RAW_FILE_NAME, sensor_name=SENSOR_NAME)
sensor_data.head()

/Users/Fernando/Documents/Projects/Github/lcqar-low-cost-monit-
proc/venv/lib/python3.9/site-packages/urllib3/_init__.py:34: NotOpenSSLWarning:
urllib3 v2.0 only supports OpenSSL 1.1.1+, currently the 'ssl' module is
compiled with 'LibreSSL 2.8.3'. See:
https://github.com/urllib3/urllib3/issues/3020
    warnings.warn(
/Users/Fernando/Documents/Projects/Github/lcqar-low-cost-monit-proc/data-pre-
processing/GetSensorDataService.py:13: UserWarning: The argument
'infer_datetime_format' is deprecated and will be removed in a future version. A
strict version of it is now the default, see
https://pandas.pydata.org/pdps/0004-consistent-to-datetime-parsing.html. You
can safely remove this argument.
    df['DateTime'] = (pd.to_datetime(df[date_time_col],
infer_datetime_format=False, format='%d/%m/%Y/%H/%M/%S'))


[ ]:   latitude  longitude  measuring          DateTime
0 -28.456899 -48.972999    101.39 2022-11-20 13:46:27
1 -28.456899 -48.972999    103.08 2022-11-21 10:38:49
2 -28.456899 -48.972999    104.34 2022-11-21 10:54:36
3 -28.456899 -48.972999    100.55 2022-11-21 11:10:24
4 -28.456899 -48.972999     89.67 2022-11-21 11:26:10

```

2.1 Upload Data from File

```
[ ]: import pandas as pd

df = pd.read_csv(SENSOR_FILE_DIR + SENSOR_FILE_NAME)
df.head()
```

```
[ ]:   Unnamed: 0  latitude  longitude  measuring          DateTime
0           0 -28.456899 -48.972999    101.39 2022-11-20 13:46:27
1           1 -28.456899 -48.972999    103.08 2022-11-21 10:38:49
2           2 -28.456899 -48.972999    104.34 2022-11-21 10:54:36
3           3 -28.456899 -48.972999    100.55 2022-11-21 11:10:24
4           4 -28.456899 -48.972999     89.67 2022-11-21 11:26:10
```

2.1.1 Create Sensor Dataframe as Pandas Series with a period of 15 mins

```
[ ]: # Remove the first column with the indexes and save data into web dataframe
web_dataframe = df.drop(df.columns[0], axis='columns')
web_dataframe['DateTime'] = (pd.to_datetime(df['DateTime'],
    ↪infer_datetime_format=True))
```

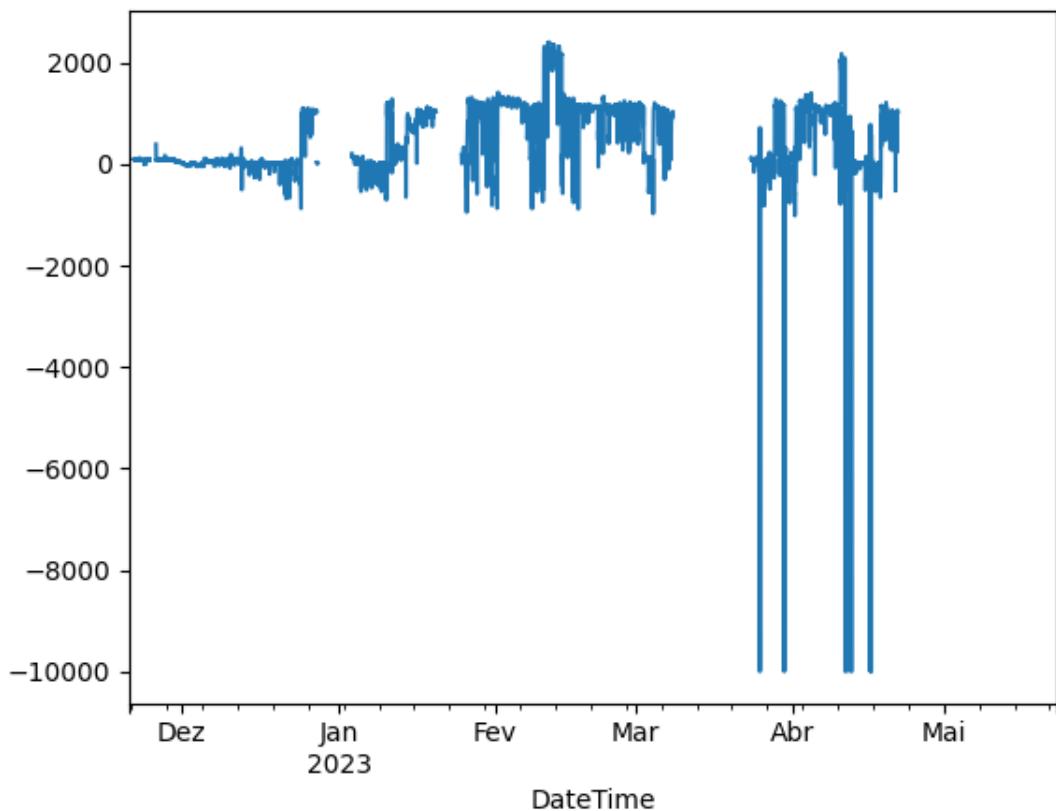
```
# Resample data with 15 mins period and create sensor dataframe
sensor_dataframe = web_dataframe.sort_values(by='DateTime', ascending=True).
    ↪reset_index().drop(columns='index')
sensor_dataframe.index = sensor_dataframe['DateTime']
sensor_dataframe = sensor_dataframe.drop(columns=['DateTime'])
sensor_dataframe = sensor_dataframe.resample('15T').mean()
sensor_dataframe.head()
```

```
/var/folders/wc/_83zcrx913j1dqwg4g90kbhh0000gp/T/ipykernel_4206/2771449038.py:3:
UserWarning: The argument 'infer_datetime_format' is deprecated and will be
removed in a future version. A strict version of it is now the default, see
https://pandas.pydata.org/pdeps/0004-consistent-to-datetime-parsing.html. You
can safely remove this argument.
    web_dataframe['DateTime'] = (pd.to_datetime(df['DateTime'],
infer_datetime_format=True))
```

```
[ ]:      latitude longitude measuring
DateTime
2022-11-20 13:45:00 -28.456899 -48.972999      101.39
2022-11-20 14:00:00      NaN      NaN      NaN
2022-11-20 14:15:00      NaN      NaN      NaN
2022-11-20 14:30:00      NaN      NaN      NaN
2022-11-20 14:45:00      NaN      NaN      NaN
```

2.1.2 Plot raw data

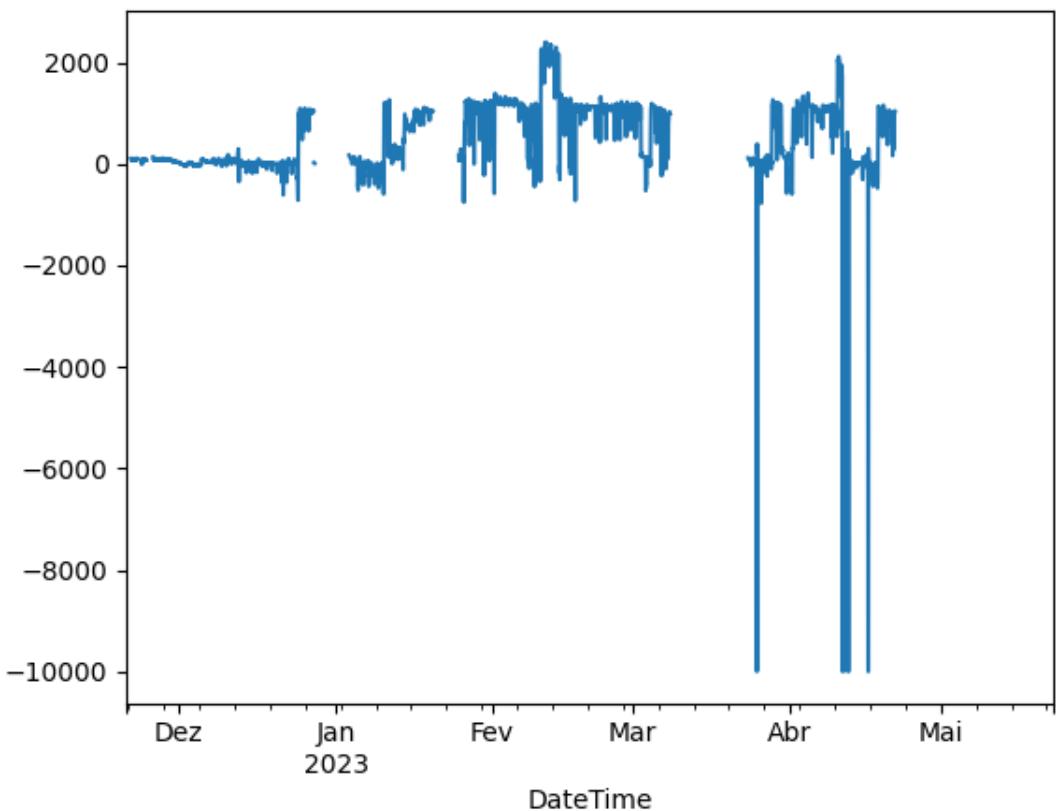
```
[ ]: sensor_dataframe['measuring'].plot()
[ ]: <Axes: xlabel='DateTime'>
```



2.1.3 Smooth data with a window of 4 samples, i.e.: an hour of data

```
[ ]: sensor_dataframe['raw_measuring'] = sensor_dataframe['measuring']
sensor_dataframe['measuring'] = sensor_dataframe['measuring'].rolling(window=4).
    median()
sensor_dataframe['Hour'] = sensor_dataframe.index.hour
sensor_dataframe['measuring'].plot()
```

```
[ ]: <Axes: xlabel='DateTime'>
```



2.1.4 Convert to ug/m3

```
[ ]: sensor_dataframe['value'] = sensor_dataframe['measuring'].map(lambda v: 0.
    ↪0409*v*M)
```

2.2 Tag data

Apply tags to the data according to the quality control processes described above

```
[ ]: sensor_dataframe['Tag'] = 'VALID'
```

Tag data during stabilizing period AQMesh recommends to considering the first two days after installation as a period for stabilization. In our case, a period of seven days was selected in order to remove outliers candidates detected during that period. For that purpose we detect the first 7 days of uninterrupted measurements

```
[ ]: # Identify consecutive measurements (15 mins between each)
consecutive_periods = sensor_dataframe.index.to_series().diff().dt.seconds ==
    ↪900
```

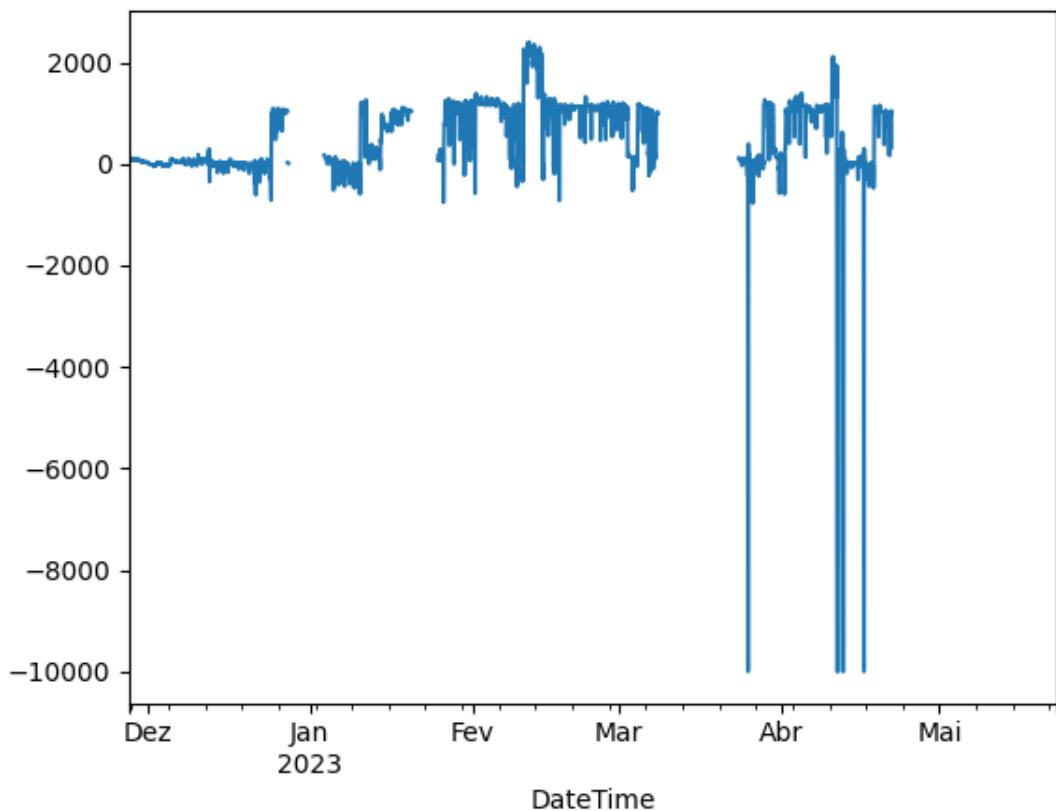
```

# Select the first seven days of consecutive measurements (96 = 24 * 60 mins / ↳
↪15 mins, amount of 15 mins periods in a day)
last_stabilizing_index = sensor_dataframe[consecutive_periods].head(96 * 7).
↪index[-1] # 7 days
sensor_dataframe.loc[sensor_dataframe.index[0]:last_stabilizing_index, 'Tag'] = ↳
↪'STABILIZING'
sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['measuring'].
↪resample('15T').mean().plot()
sensor_dataframe.head()

```

```
[ ]:          latitude longitude measuring raw measuring Hour \
DateTime
2022-11-20 13:45:00 -28.456899 -48.972999      NaN      101.39    13
2022-11-20 14:00:00      NaN      NaN      NaN      NaN      14
2022-11-20 14:15:00      NaN      NaN      NaN      NaN      14
2022-11-20 14:30:00      NaN      NaN      NaN      NaN      14
2022-11-20 14:45:00      NaN      NaN      NaN      NaN      14

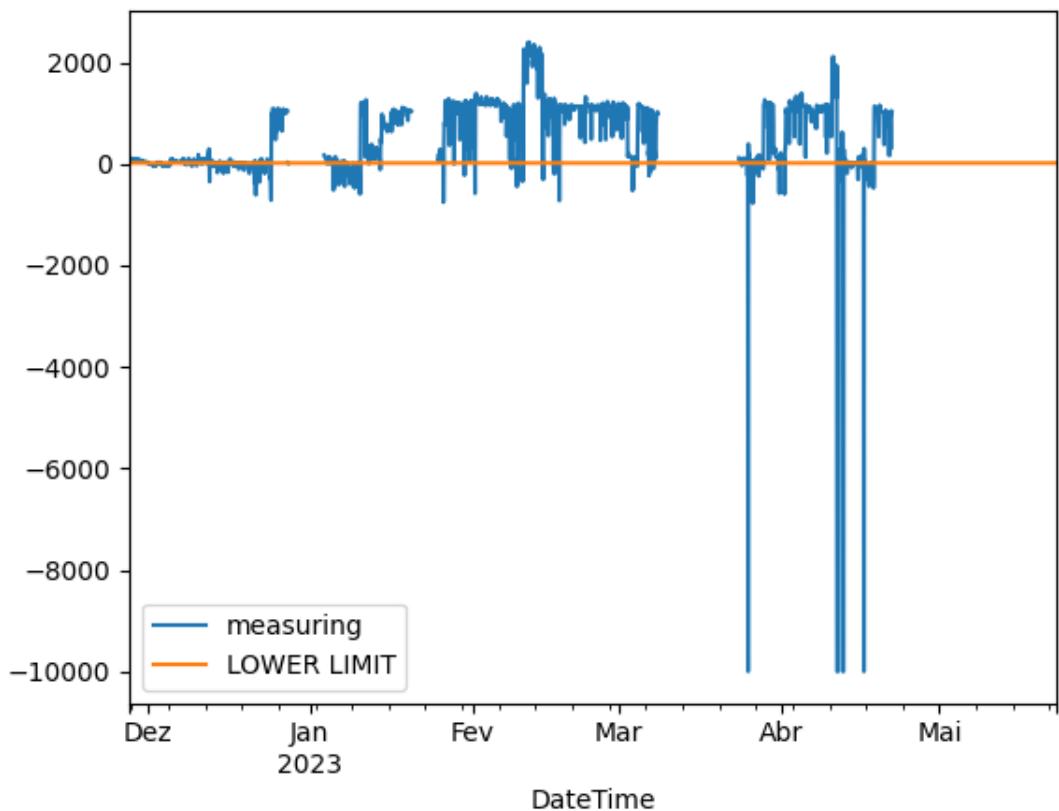
                           value        Tag
DateTime
2022-11-20 13:45:00      NaN  STABILIZING
2022-11-20 14:00:00      NaN  STABILIZING
2022-11-20 14:15:00      NaN  STABILIZING
2022-11-20 14:30:00      NaN  STABILIZING
2022-11-20 14:45:00      NaN  STABILIZING
```



Tag values that are missing, are greater than sensor span or are lower than sensor resolution The upper limit was too high that affected the graph resolution, therefore it is not plotted

```
[ ]: sensor_dataframe['UPPER LIMIT'] = upper_limit
sensor_dataframe['LOWER LIMIT'] = lower_limit
sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'][['measuring', 'LOWER LIMIT']].resample('15T').mean().plot()
```

```
[ ]: <Axes: xlabel='DateTime'>
```



```
[ ]: import numpy as np
from SensorDataAnalysisService import SensorDataAnalysisService as sensor_analysis

def tag_by_sensor_limits(value, tag, lower_limit, upper_limit):
    if (tag != 'VALID'): return tag
    return sensor_analysis.get_tags_from_series(value=value,
                                                lower_limit=lower_limit,
                                                upper_limit=upper_limit)

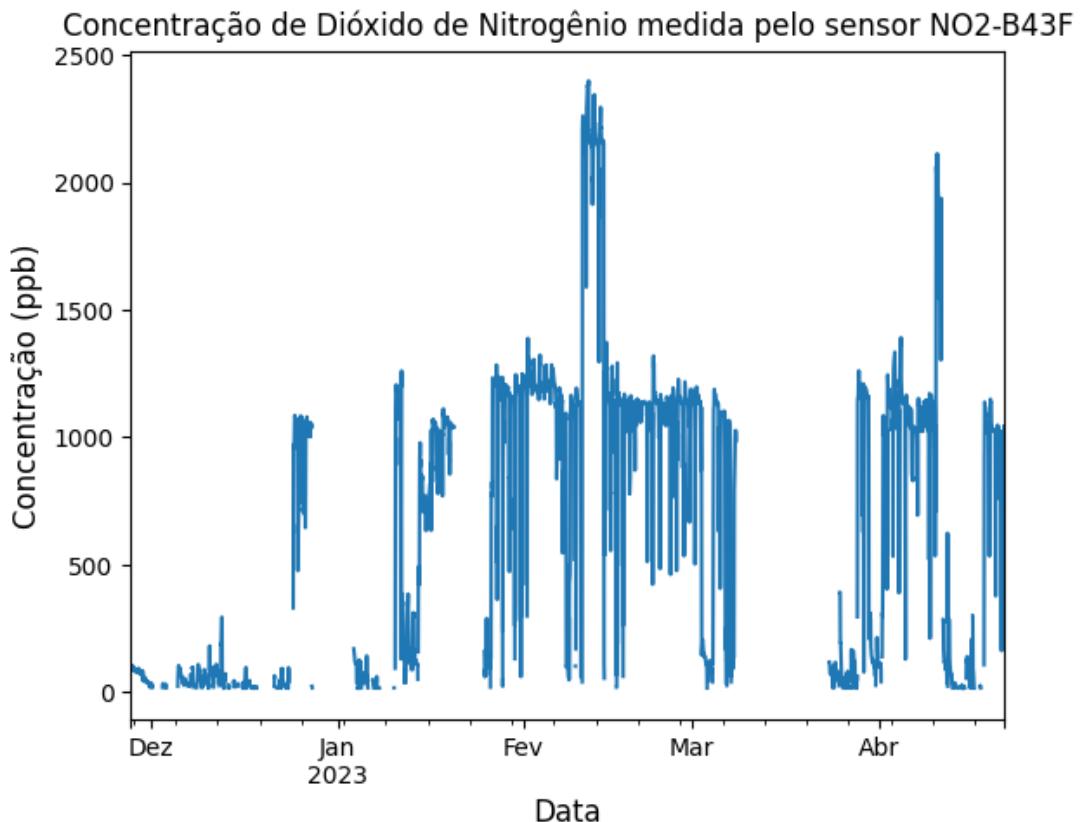
sensor_dataframe['Tag'] = (sensor_dataframe[['measuring', 'Tag']]
                           .apply(lambda df: tag_by_sensor_limits(value=df[0], tag=df[1],
                           lower_limit=lower_limit,
                           upper_limit=upper_limit),
                           axis=1))
```

Plot valid data after removing samples bellow lower limit

```
[ ]: import matplotlib.pyplot as plt

fig, ax = plt.subplots(figsize=(1.3*5,5))
sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['measuring'].
    ↪rename('Concentração de Dióxido de Nitrogênio').resample('15T').mean().plot()
plt.title('Concentração de Dióxido de Nitrogênio medida pelo sensor NO2-B43F')
ax.set_xlabel('Data', fontsize=12)
ax.set_ylabel('Concentração (ppb)', fontsize=12)
```

```
[ ]: Text(0, 0.5, 'Concentração (ppb)')
```



2.3 Change point Analysis

2.3.1 Change point analysis of 15 mins data

```
[ ]: import ruptures as rpt

series = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['measuring']
signal = np.array(series).reshape(-1, 1)
algo = rpt.Pelt(model="rbf", jump=50, min_size=100).fit(signal=signal)
```

```
result = algo.predict(pen=10)

[ ]: change_point_index = series[[x - 1 for x in result]].index
sensor_dataframe['CHANGE POINT'] = False
sensor_dataframe['CHANGE POINT'].loc[change_point_index] = True

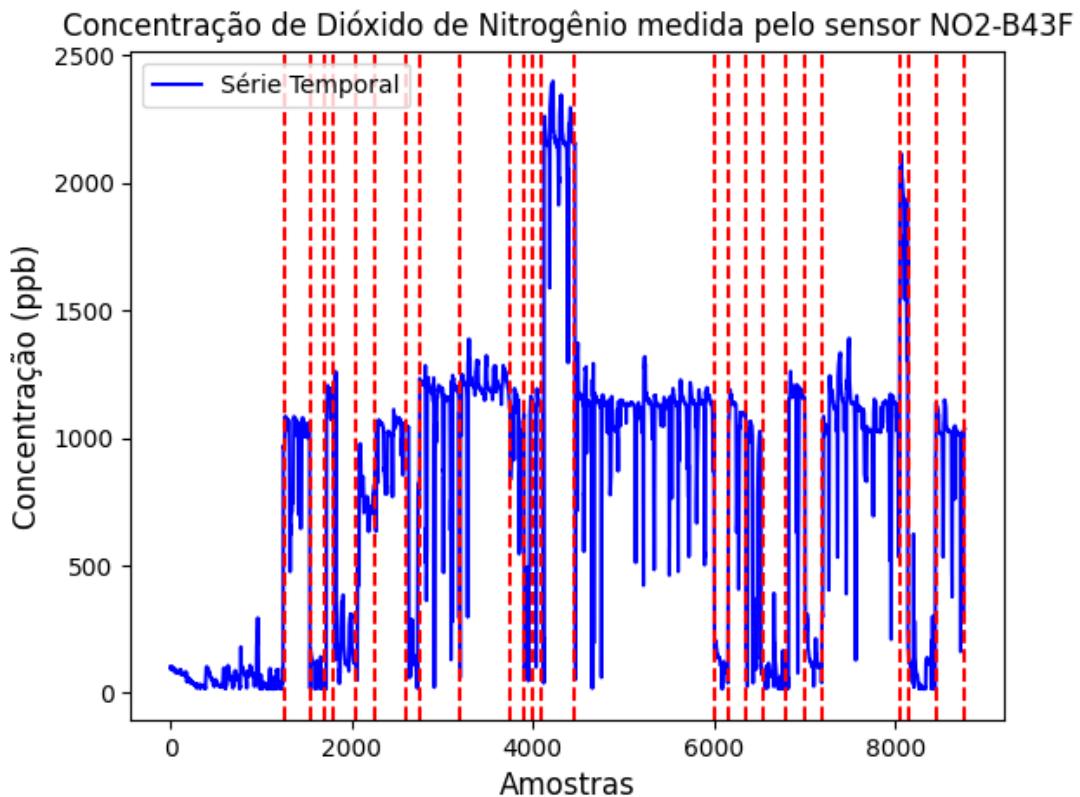
/var/folders/wc/_83zcrx913j1dqwg4g90kbhh0000gp/T/ipykernel_4206/717246363.py:3:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
    sensor_dataframe['CHANGE POINT'].loc[change_point_index] = True

[ ]: fig, ax = plt.subplots(figsize=(1.3*5,5))
plt.plot(signal, label='Série Temporal', color='blue')
plt.title('Concentração de Dióxido de Nitrogênio medida pelo sensor N02-B43F')
ax.set_xlabel('Amostras', fontsize=12)
ax.set_ylabel('Concentração (ppb)', fontsize=12)

# Plot change points
for point in result:
    plt.axvline(x=point, color='red', linestyle='--')

# Show the plot
plt.legend()
plt.show()
```



Tag dataframe with changepoints

```
[ ]: sensor_dataframe.loc[change_point_index[0]:, 'Tag'] = (sensor_dataframe.
    ↪ loc[change_point_index[0]:, 'Tag']
    .apply(lambda t: ↪
        'REBASE' if t == 'VALID' else t))
sensor_dataframe[sensor_dataframe['Tag'] == 'REBASE'].head()
```

```
[ ]:          latitude  longitude  measuring  raw measuring  Hour \
DateTime
2022-12-24 16:00:00 -28.456899 -48.972999    961.770      960.63    16
2022-12-24 16:15:00 -28.456899 -48.972999    961.770      970.24    16
2022-12-24 16:30:00 -28.456899 -48.972999    966.575      972.52    16
2022-12-24 16:45:00 -28.456899 -48.972999    968.470      966.70    16
2022-12-24 17:00:00 -28.456899 -48.972999    971.380      972.78    17
```

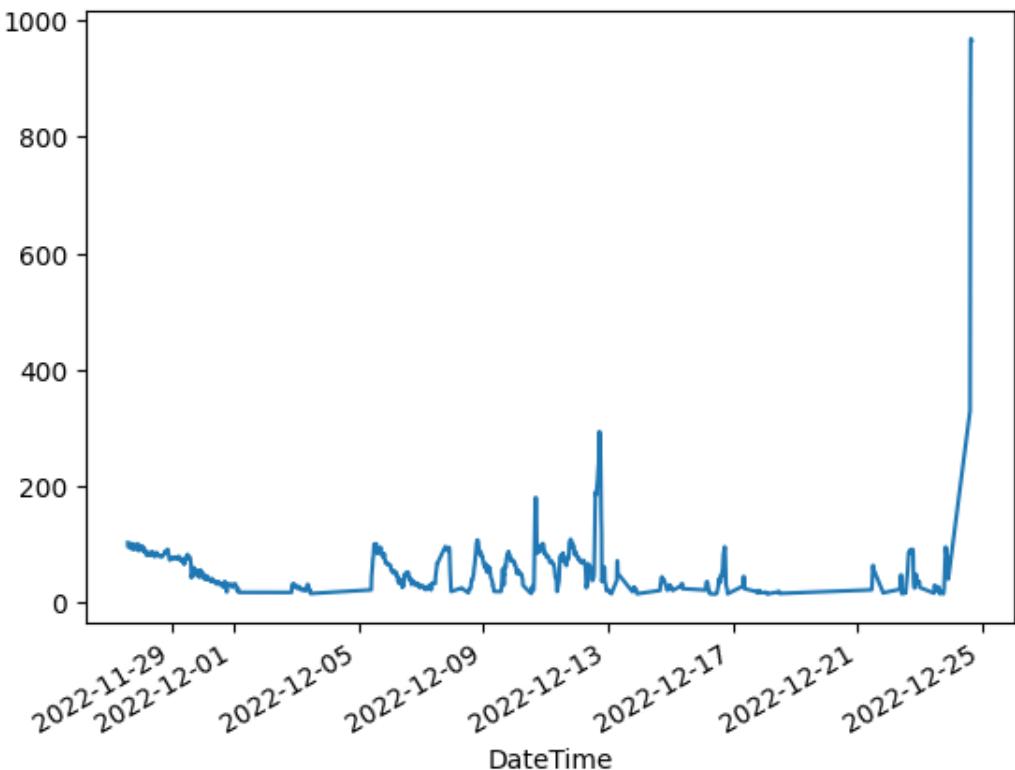
```
          value     Tag  UPPER LIMIT  LOWER LIMIT \
DateTime
2022-12-24 16:00:00  1809.690428  REBASE      20000.0      15.0
2022-12-24 16:15:00  1809.690428  REBASE      20000.0      15.0
2022-12-24 16:30:00  1818.731636  REBASE      20000.0      15.0
```

```
2022-12-24 16:45:00 1822.297315 REBASE      20000.0      15.0
2022-12-24 17:00:00 1827.772844 REBASE      20000.0      15.0
```

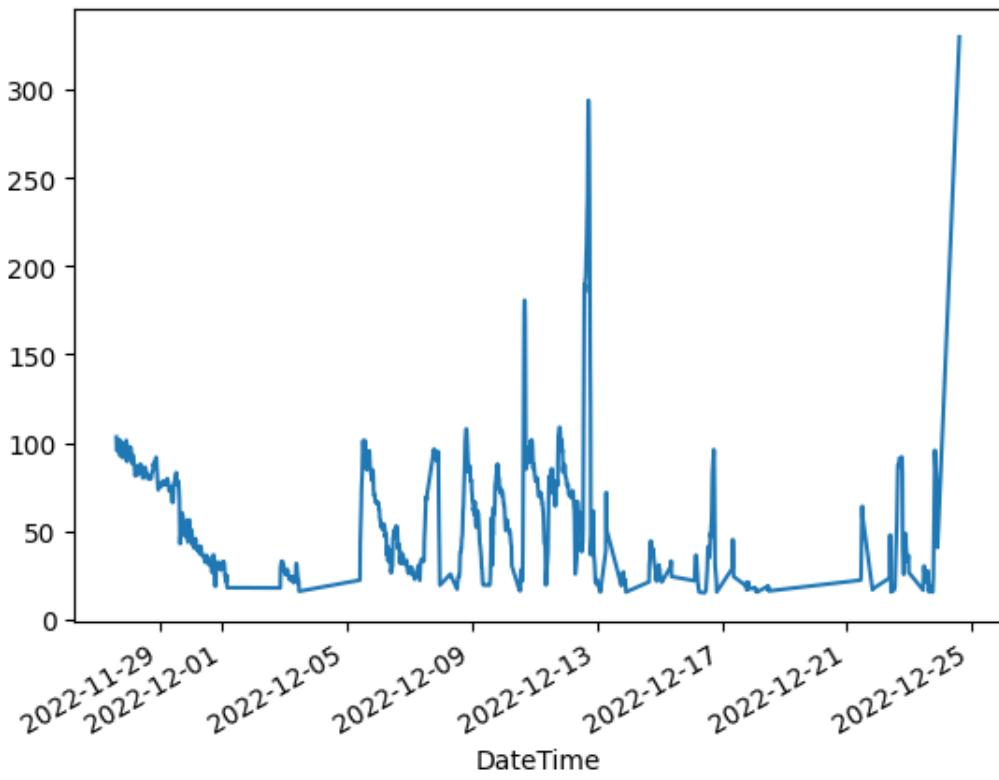
CHANGE POINT

```
DateTime
2022-12-24 16:00:00      True
2022-12-24 16:15:00     False
2022-12-24 16:30:00     False
2022-12-24 16:45:00     False
2022-12-24 17:00:00     False
```

```
[ ]: sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['measuring'].plot()
valid_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']
```



```
[ ]: sensor_dataframe['Tag'] = (sensor_dataframe[['Tag', 'measuring']]
                                .apply(lambda df: 'REBASE' if df[0] == 'VALID' and
                                     df[1] > 500 else df[0], axis=1))
sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['measuring'].plot()
valid_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']
```



Data Tag contabilization

```
[ ]: tags = ['MISSING', 'LTLL', 'GTUL', 'STABILIZING', 'BADSPIKE', 'VALID', 'REBASE']
data_contabilization = sensor_analysis.count_tags(tags, sensor_dataframe)
data_contabilization
```

	#	%
MISSING	5767	32.679776
LTLL	2438	13.815379
GTUL	0	0.0
STABILIZING	673	3.813679
BADSPIKE	0	0.0
VALID	1244	7.049357
REBASE	7525	42.641809
TOTAL	17647	100.0

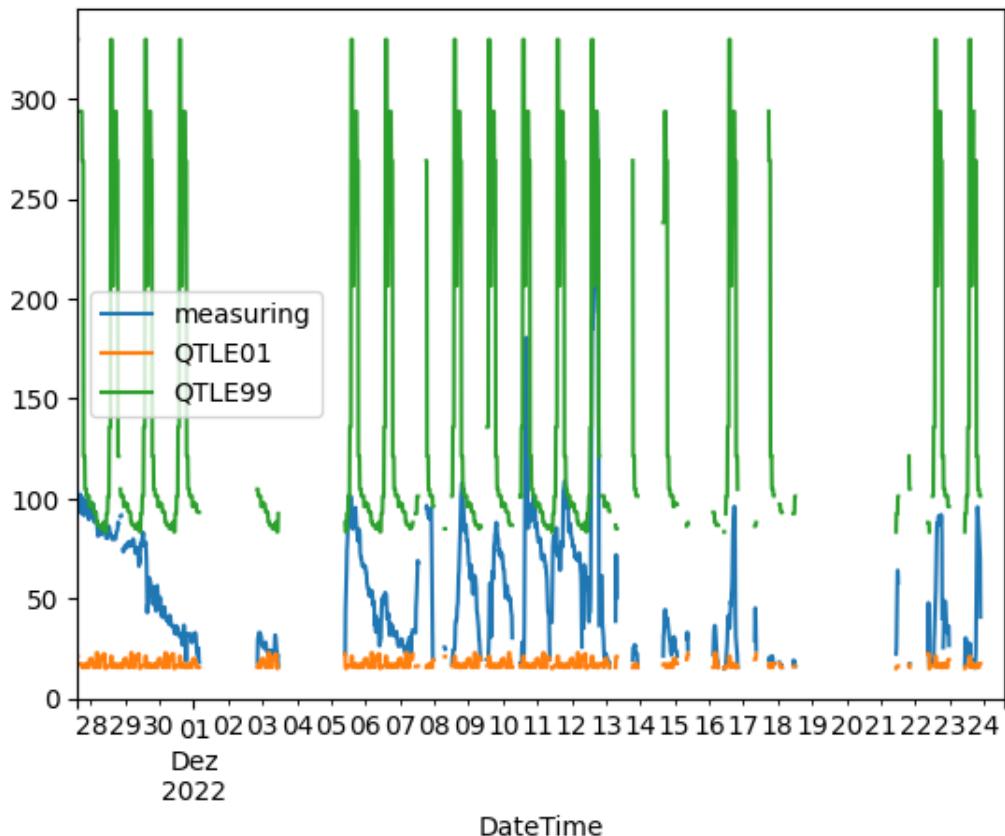
2.3.2 Calculate quantiles

We calculated the 1% and 99 % quantiles of every hour of the day. The values greater outside the 1 % - 99 % where tagged as Greater than Quantile 99 (GTQTL99) and Lower than Quantile 1 (LTQTL01)

```
[ ]: import numpy as np

global_qtle_01 = valid_dataframe.pivot(columns='Hour')['measuring'].
    ↪quantile(q=0.01, axis='index', interpolation='lower').dropna()
global_qtle_99 = valid_dataframe.pivot(columns='Hour')['measuring'].
    ↪quantile(q=0.99, axis='index', interpolation='higher').dropna()
sensor_dataframe['QTLE01'] = sensor_dataframe['Hour'].map(lambda hr: ↪
    ↪global_qtle_01[hr] if (not np.isnan(hr)) else np.nan)
sensor_dataframe['QTLE99'] = sensor_dataframe['Hour'].map(lambda hr: ↪
    ↪global_qtle_99[hr] if (not np.isnan(hr)) else np.nan)
sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'][['measuring', 'QTLE01', ↪
    ↪'QTLE99']].resample('15T').mean().plot()
```

```
[ ]: <Axes: xlabel='DateTime'>
```



2.3.3 Tag data according to quantiles

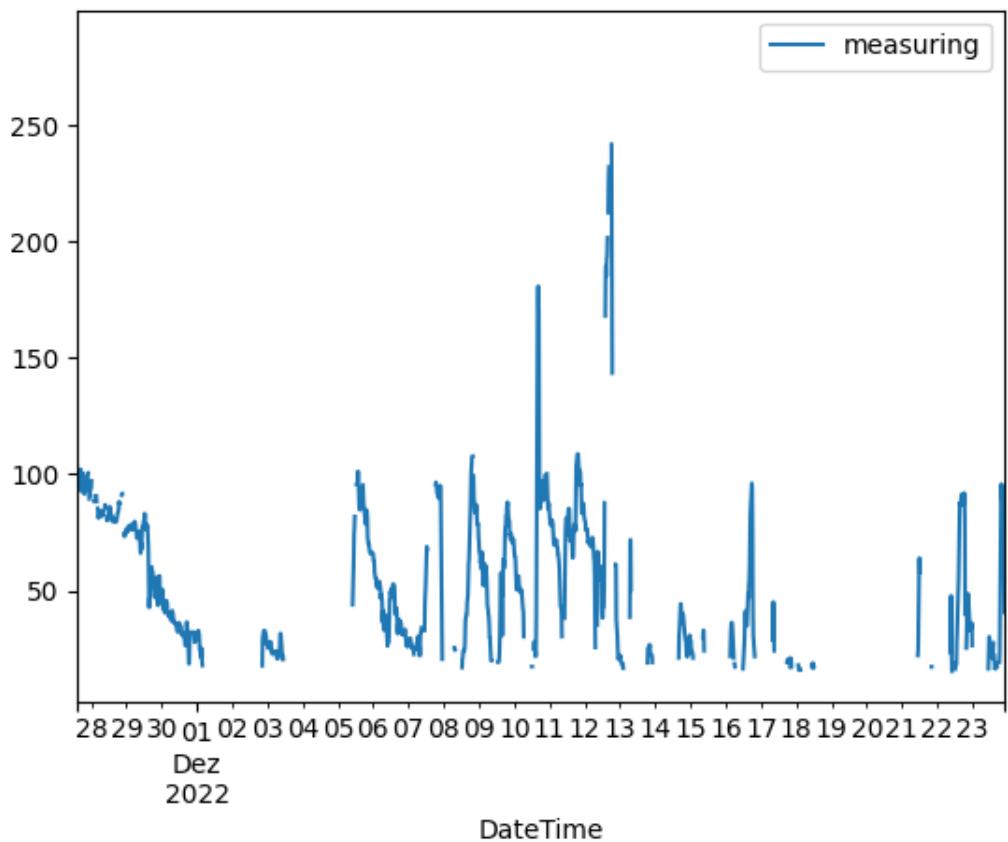
```
[ ]: sensor_dataframe['Tag'] = (sensor_dataframe[['Tag', 'measuring', 'QTLE01', 'QTLE99']]  
    .apply(lambda df: sensor_analysis.  
        tag_by_quantiles(current_tag=df[0],  
        value=df[1],  
        quantile_01=df[2],  
        quantile_99=df[3]),  
        axis=1))  
tags = ['MISSING', 'LTLL', 'GTUL', 'BADSPIKE', 'VALID', 'LTQTLE01', 'GTQTLE99']  
data_contabilization = sensor_analysis.count_tags(tags, sensor_dataframe)  
data_contabilization
```

```
[ ]: # %  
MISSING      5767  32.679776  
LTLL         2438  13.815379  
GTUL          0     0.0  
BADSPIKE      0     0.0  
VALID         1176  6.664022  
LTQTLE01      32    0.181334  
GTQTLE99      36    0.204001  
TOTAL         17647 100.0
```

Plot valid data

```
[ ]: sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'][['measuring']].  
    resample('15T').mean().plot()
```

```
[ ]: <Axes: xlabel='DateTime'>
```

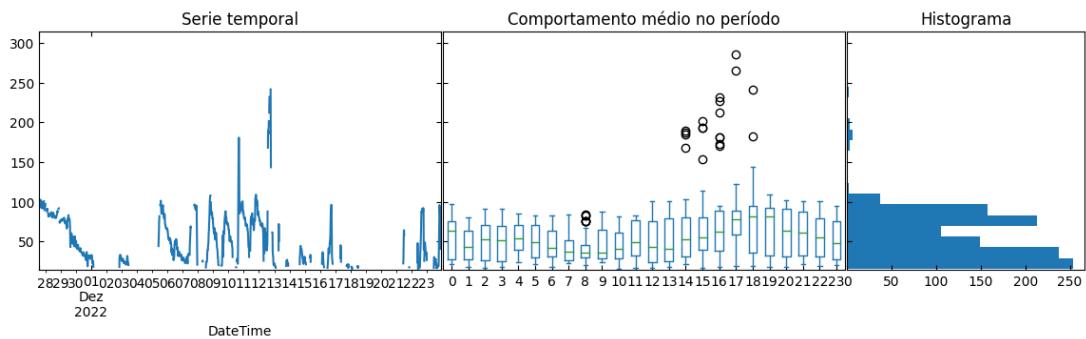


Analyse data after removing quantiles

```
[ ]: valid_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'].  
      .drop(columns=['Tag']).resample('15T').mean()  
sensor_analysis.plot_box_hist(df=valid_dataframe, bins=20)
```

/Users/Fernando/Documents/Projects/Github/lcqar-low-cost-monit-proc/data-pre-processing/SensorDataAnalysisService.py:111: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
df['Hour'] = df['Hour'].astype('int64')



2.4 Analyse valid data

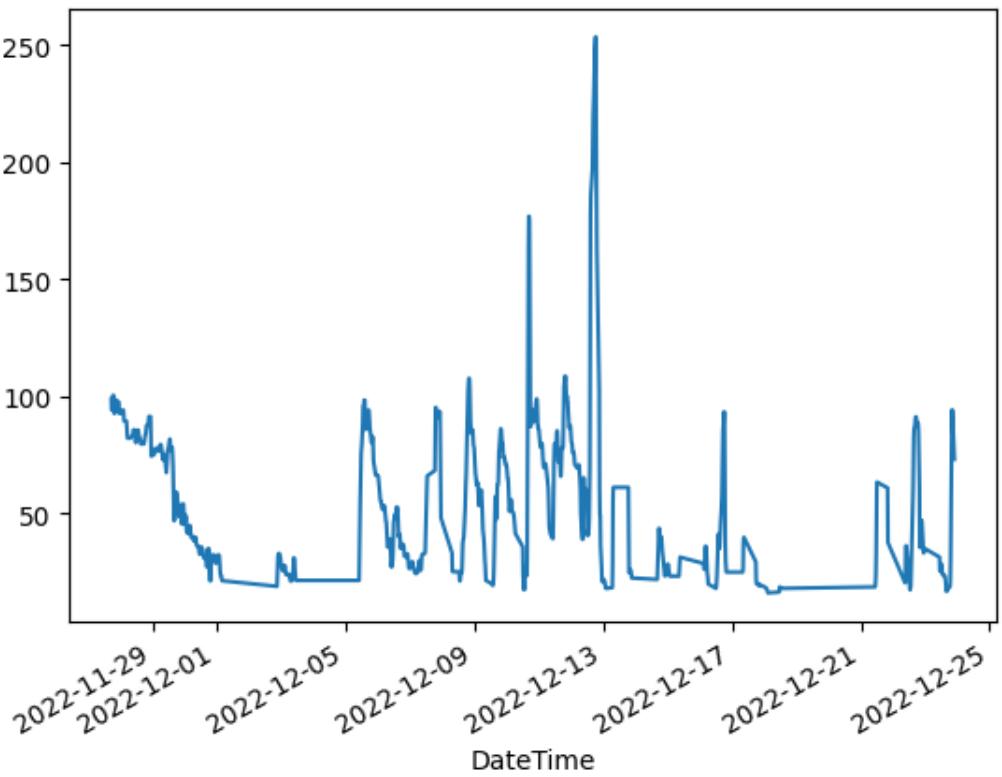
The graph shows the time series of the valid data, box plots of the data grouped by hour of the day, and histogram of the data.

```
[ ]: valid_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'].
      ↪drop(columns=['Tag'])
```

2.4.1 Smooth data with a window of 4 samples, i.e.: an hour of data

```
[ ]: valid_dataframe['raw_measuring'] = valid_dataframe['measuring']
valid_dataframe['measuring'] = valid_dataframe['measuring'].rolling(window=4).
    ↪median()
valid_dataframe['measuring'].plot()
```

```
[ ]: <Axes: xlabel='DateTime'>
```

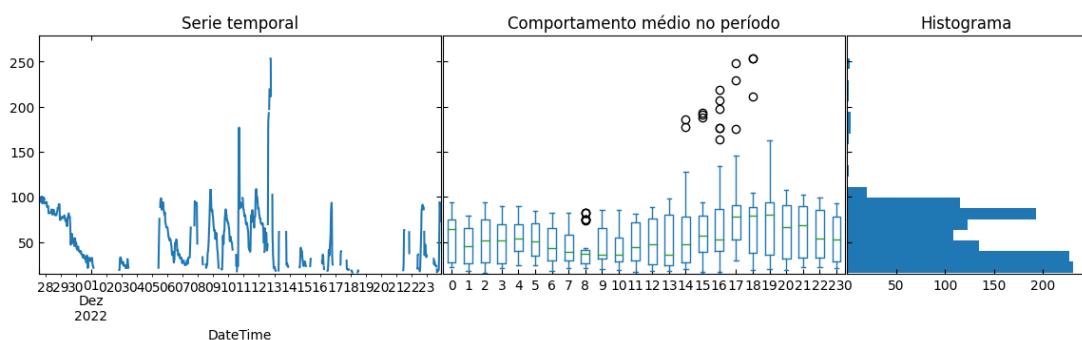


```
[ ]: valid_dataframe = valid_dataframe.resample('15T').mean()
sensor_analysis.plot_box_hist(df=valid_dataframe, bins=20)
```

```
/Users/Fernando/Documents/Projects/Github/lcqar-low-cost-monit-proc/data-pre-
processing/SensorDataAnalysisService.py:111: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

```
df['Hour'] = df['Hour'].astype('int64')
```



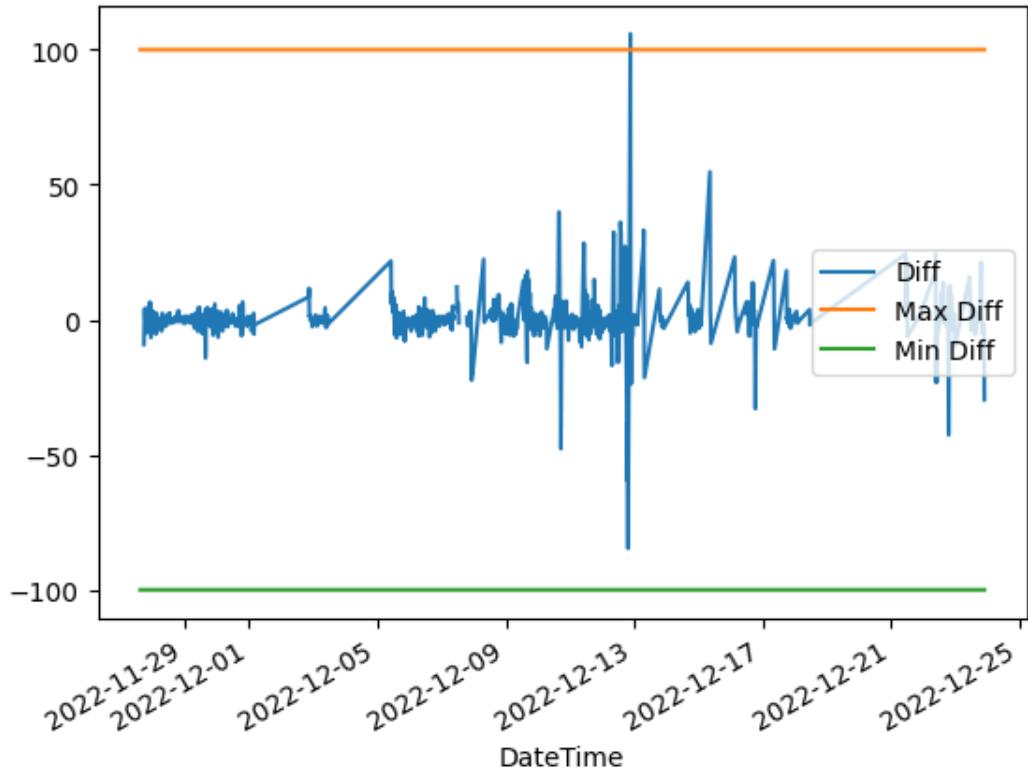
2.5 Analyse data derivatives

Analyse the derivatives of the data for removing abrupt changes in the data (BADSPIKE)

Tag values with derivatives faster than sensor response time The value of 100 ppb was selected based on the maximum derivative found in the reference data. The maximum derivative (with measuring period of 1 hour) found was 30 ppb. Therefore, a maximum of 100 ppb was considered appropriate for a 15 mins period.

```
[ ]: max_diff_value = 100
sensor_dataframe['Diff'] = sensor_dataframe['measuring'].resample('15T').mean().
    diff()
sensor_dataframe['Max Diff'] = max_diff_value
sensor_dataframe['Min Diff'] = -max_diff_value
sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'][['Diff', 'Max Diff', 'Min Diff']].plot()
```

```
[ ]: <Axes: xlabel='DateTime'>
```



```
[ ]: import numpy as np

def tag_data_with_derivatives(tagged_df, max_diff_value):
    current_tag = tagged_df[0]
    value = tagged_df[1]
    if ((current_tag != 'VALID') or (np.isnan(value))): return current_tag
    if ((value > max_diff_value) or (value < -max_diff_value)): return ↵
        ↵'BADSPIKE'
    return 'VALID'

sensor_dataframe['Tag'] = (sensor_dataframe[['Tag', 'Diff', 'Max Diff']]
                           .apply(lambda df: ↵
                                  ↵tag_data_with_derivatives(tagged_df=df, max_diff_value=df[2]),
                                  axis=1))
```

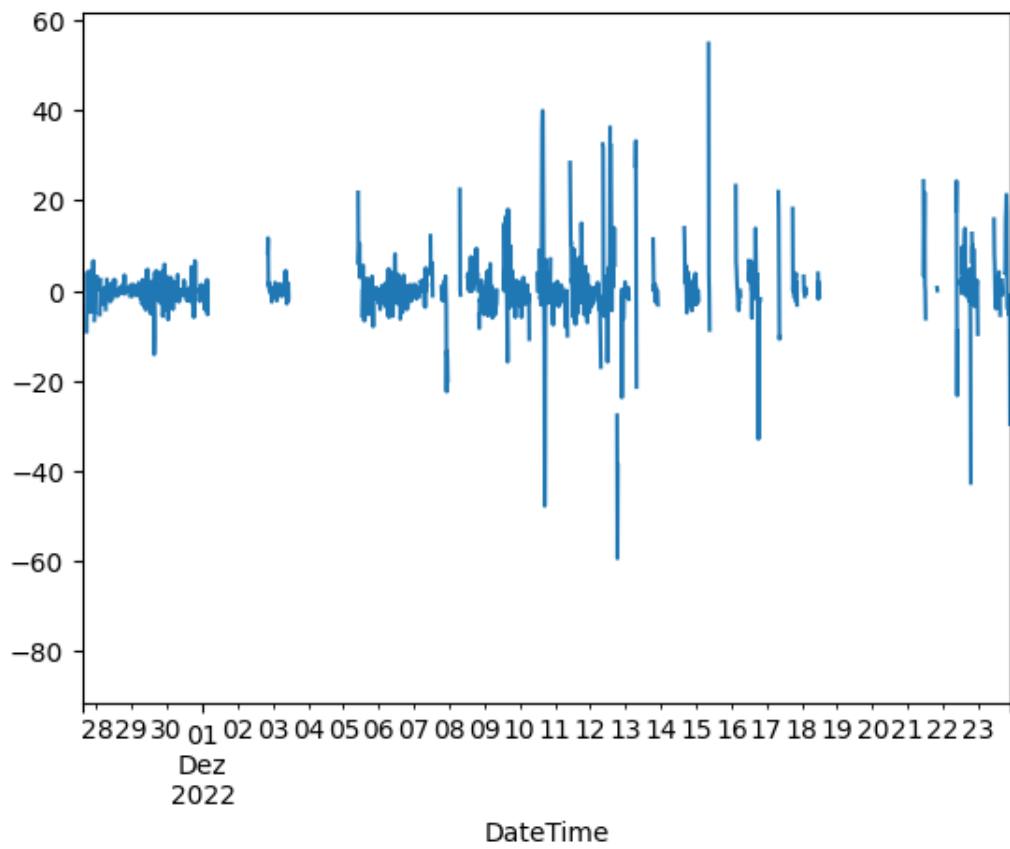
2.5.1 Data Tag contabilization

```
[ ]: tags = ['MISSING', 'LTLL', 'GTUL', 'STABILIZING', 'BADSPIKE', 'VALID']
data_contabilization = sensor_analysis.count_tags(tags, sensor_dataframe)
data_contabilization
```

```
[ ]:          #      %
MISSING      5767  32.679776
LTLL         2438  13.815379
GTUL          0     0.0
STABILIZING   673   3.813679
BADSPIKE      1     0.005667
VALID         1175  6.658356
TOTAL         17647 100.0
```

```
[ ]: sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['Diff'].resample('15T').
    ↵mean().plot()
```

```
[ ]: <Axes: xlabel='DateTime'>
```



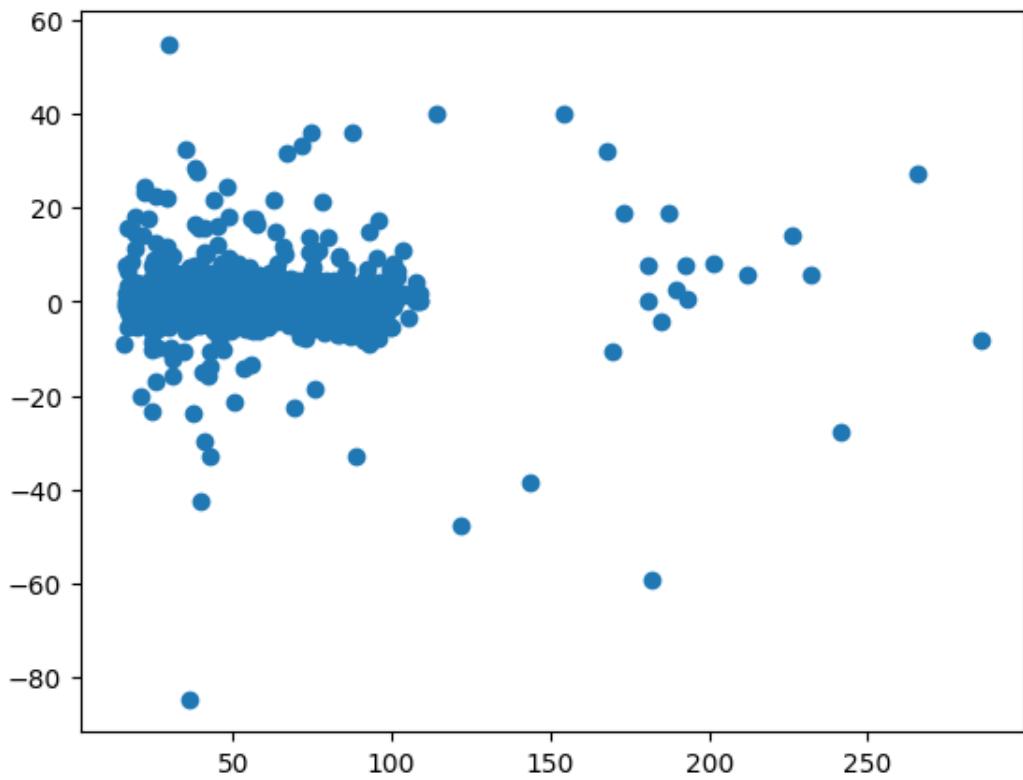
Plot data derivatives vs data

```
[ ]: valid_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']
```

```
[ ]: import matplotlib.pyplot as plt
```

```
fig = plt.figure(figsize=(1.3*5,5))
plt.scatter(valid_dataframe['measuring'], valid_dataframe['Diff'])
```

```
[ ]: <matplotlib.collections.PathCollection at 0x1477f6970>
```



Data Tag contabilization

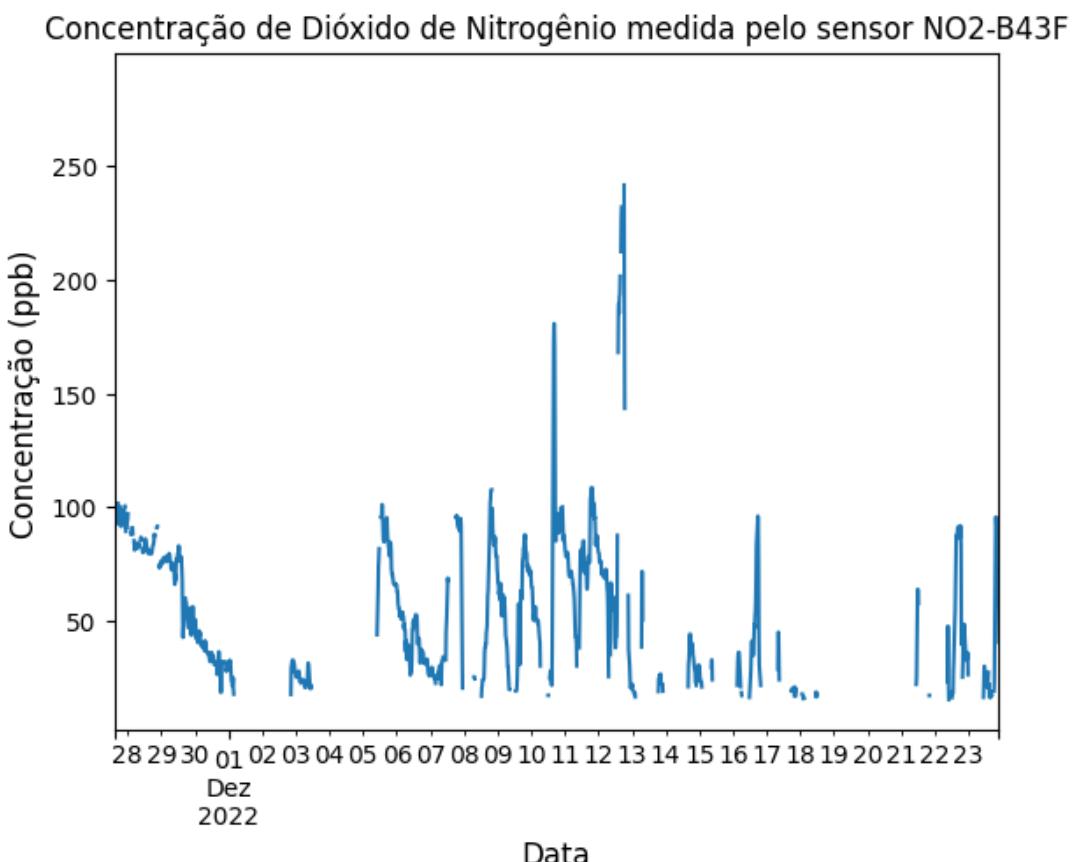
```
[ ]: tags = ['MISSING', 'LTLL', 'GTUL', 'STABILIZING', 'BADSPIKE', 'VALID', ▾
    ↵'LTQTLE01', 'GTQTLE99', 'REBASE']
data_contabilization = sensor_analysis.count_tags(tags, sensor_dataframe)
data_contabilization
```

	#	%
MISSING	5767	32.679776
LTLL	2438	13.815379
GTUL	0	0.0
STABILIZING	673	3.813679
BADSPIKE	1	0.005667
VALID	1175	6.658356
LTQTLE01	32	0.181334
GTQTLE99	36	0.204001
REBASE	7525	42.641809
TOTAL	17647	100.0

Plot valid data

```
[ ]: fig, ax = plt.subplots(figsize=(1.3*5,5))
(sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['measuring']
 .rename('Concentração de Dióxido de Nitrogênio').resample('15T').mean().plot())
plt.title('Concentração de Dióxido de Nitrogênio medida pelo sensor NO2-B43F')
ax.set_xlabel('Data', fontsize=12)
ax.set_ylabel('Concentração (ppb)', fontsize=12)
```

```
[ ]: Text(0, 0.5, 'Concentração (ppb)')
```

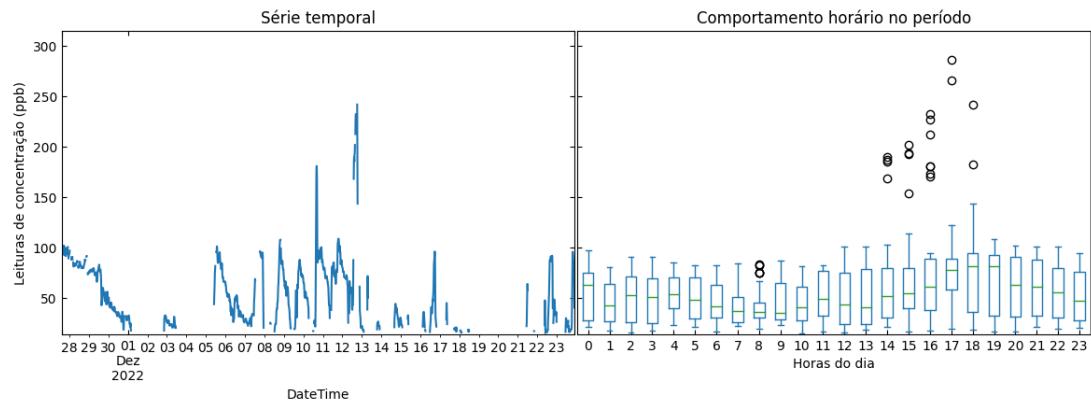


```
[ ]: valid_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'].
    drop(columns=['Tag']).resample('15T').mean()
sensor_analysis.plot_box(df=valid_dataframe)
```

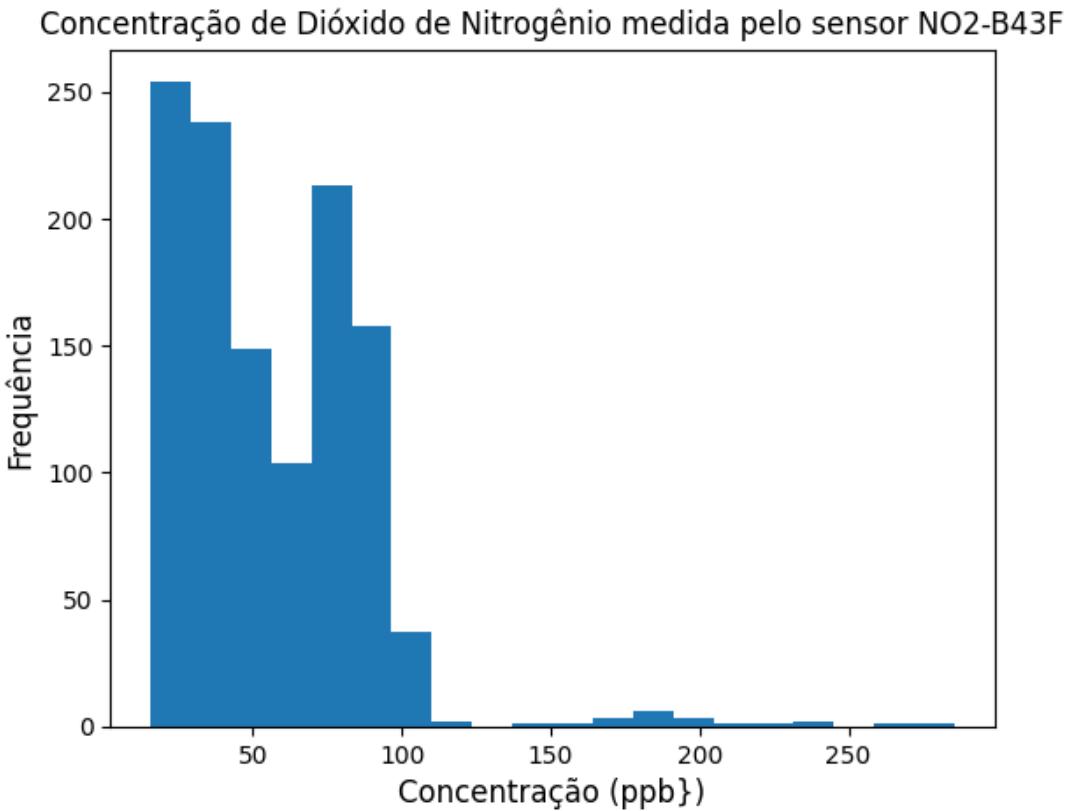
```
/Users/Fernando/Documents/Projects/Github/lcqar-low-cost-monit-proc/data-pre-
processing/SensorDataAnalysisService.py:143: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

See the caveats in the documentation: <https://pandas.pydata.org/pandas->

```
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
df['Hour'] = df['Hour'].astype('int64')
```



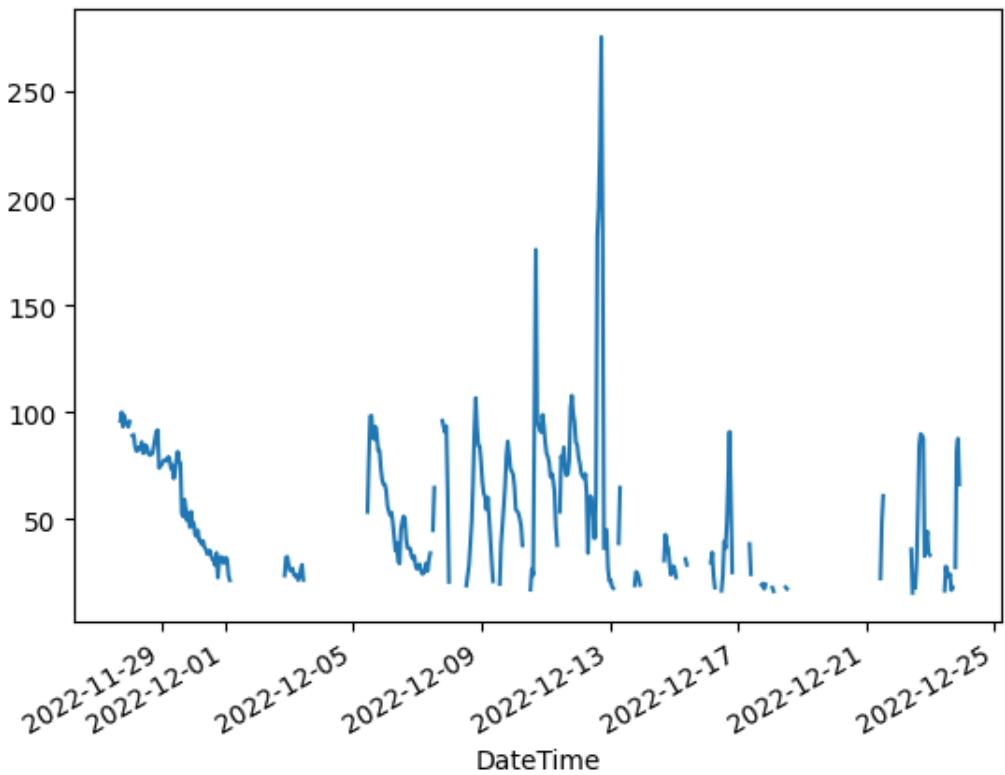
```
[ ]: fig, ax = plt.subplots(figsize=(1.3*5,5))
(sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['measuring'].
    rename('Concentração de Dióxido de Nitrogênio')
    .resample('15T').mean().hist(bins=20))
plt.title('Concentração de Dióxido de Nitrogênio medida pelo sensor N02-B43F')
ax.set_xlabel('Concentração (ppb)', fontsize=12)
ax.set_ylabel('Frequência', fontsize=12)
ax.grid(False)
```



2.6 Resample valid data to 1 HR

```
[ ]: resampled_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'].
    ↪drop(columns=['Tag']).resample('H').mean()
resampled_dataframe['Hour'] = resampled_dataframe.index.hour
resampled_dataframe['Count'] = (valid_dataframe['measuring'].resample('H').
    ↪count())
resampled_dataframe['Std'] = (valid_dataframe['measuring'].resample('H').std())
resampled_dataframe.index = resampled_dataframe.index.map(lambda t: t.
    ↪replace(minute=30, second=0))
resampled_dataframe['measuring'].plot()
```

```
[ ]: <Axes: xlabel='DateTime'>
```



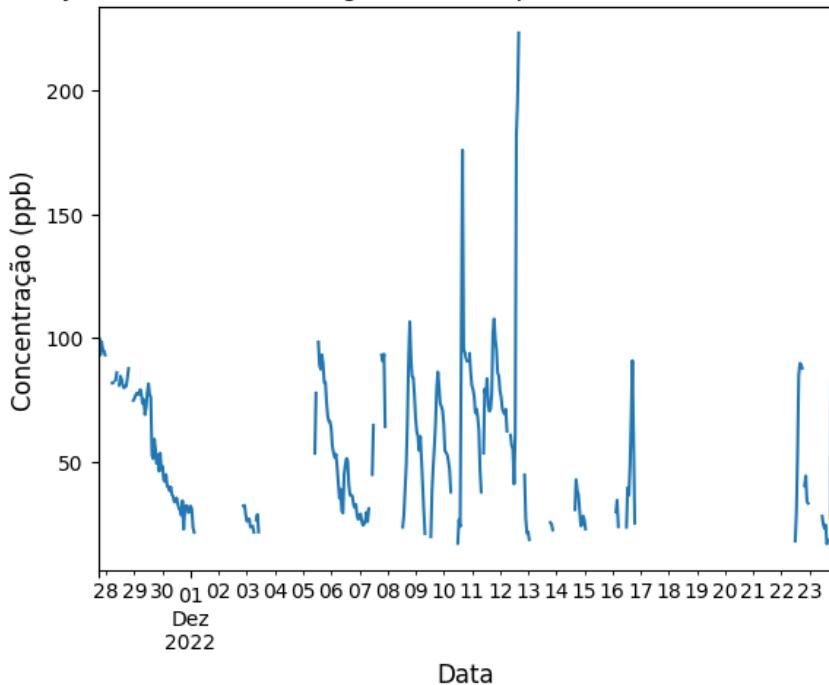
Tag hourly data according to the number of samples in an hour At least 3 samples must be valid in an hour (75 %) for the hourly data be considered as valid

```
[ ]: original_freq = sensor_dataframe.index.freq
resampled_dataframe['% valid'] = (resampled_dataframe['Count']
                                    .map(lambda c:
                                         c / (pd.Timedelta("1 hour")) /_
                                         original_freq) * 100)
resampled_dataframe['Tag'] = (resampled_dataframe['% valid']
                                .map(lambda c: 'VALID' if c >= 75 else_
                                'LOWSAMPLES'))

fig, ax = plt.subplots(figsize=(1.3*5,5))
(resampled_dataframe[resampled_dataframe['Tag'] == 'VALID'][['measuring']]
 .rename('Concentração de Dióxido de Nitrogênio').resample('1H').mean().plot())
plt.title('Concentração de Dióxido de Nitrogênio medida pelo sensor NO2-B43F_
(Dados Horários)')
ax.set_xlabel('Data', fontsize=12)
ax.set_ylabel('Concentração (ppb)', fontsize=12)

[ ]: Text(0, 0.5, 'Concentração (ppb)')
```

Concentração de Dióxido de Nitrogênio medida pelo sensor NO2-B43F (Dados Horários)



Valid data contabilization

```
[ ]: tags = ['LOWSAMPLES', 'VALID']
data_contabilization = sensor_analysis.count_tags(tags, resampled_dataframe)
data_contabilization
```

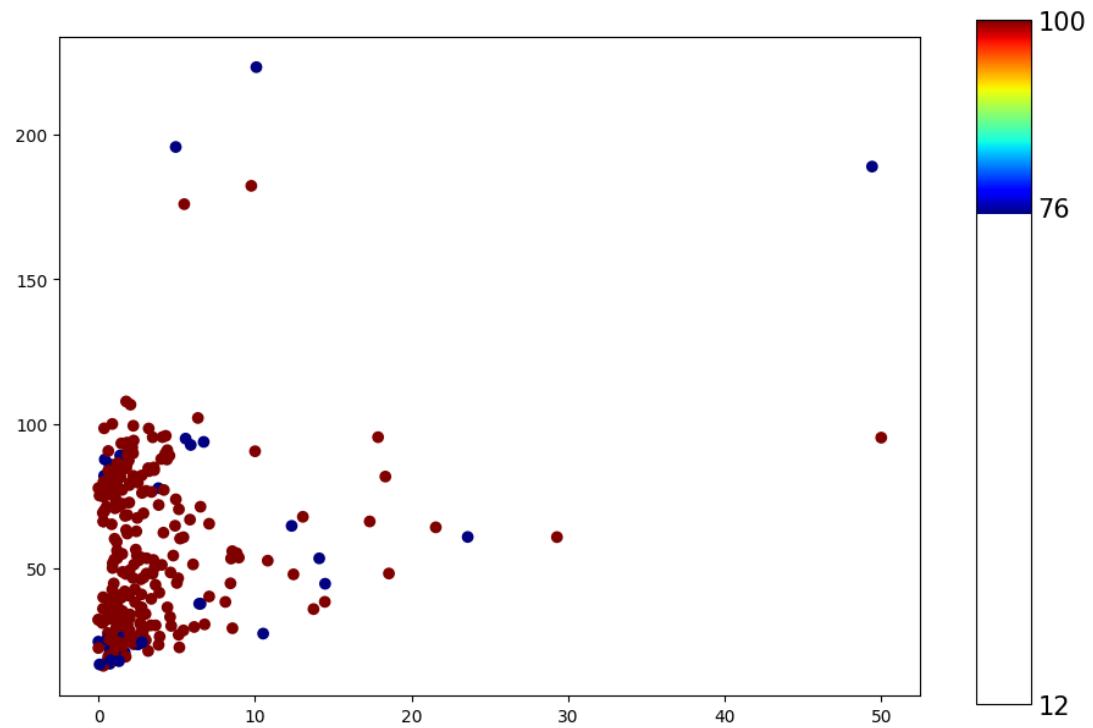
	#	%
LOWSAMPLES	347	54.905063
VALID	285	45.094937
TOTAL	632	100.0

2.7 Analyse the mean and standard deviation of the resampled data

The color of each data point represents the percentage of valid 15 mins samples found in an hour

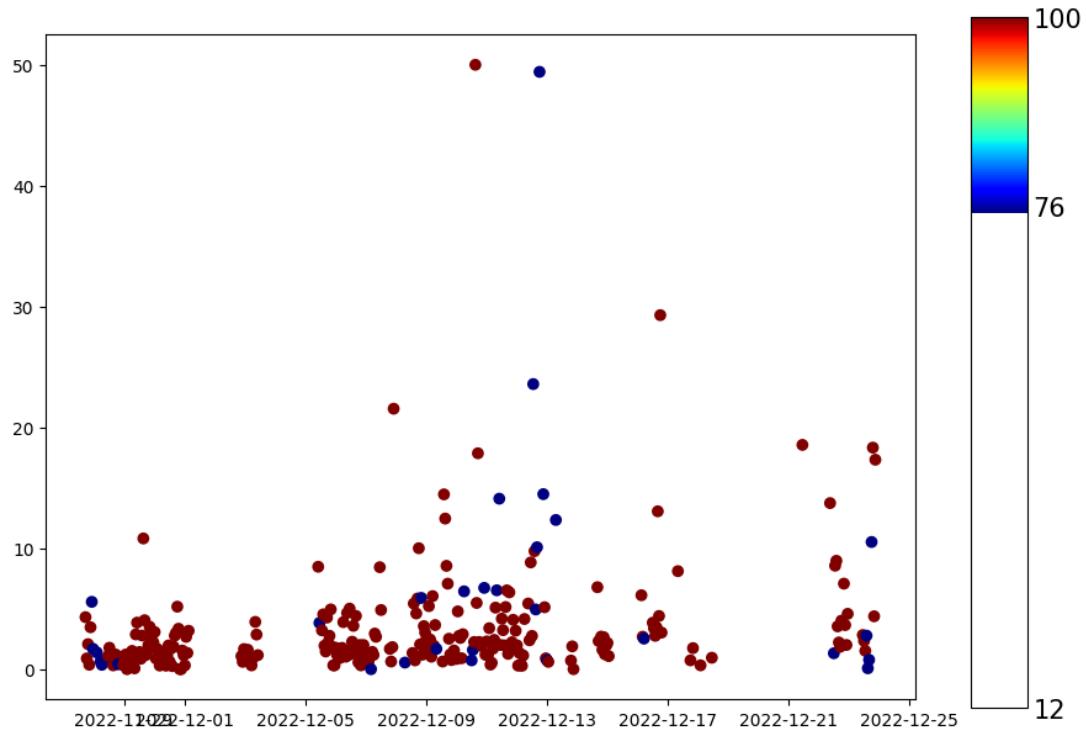
Plot mean (Y axis) vs. standard deviation (X axis) for valid hourly data colored by valid percentage of valid samples

```
[ ]: valid_resampled_dataframe = resampled_dataframe[resampled_dataframe['Tag'] == 'VALID']
sensor_analysis.plot_mean_vs_std(valid_resampled_dataframe)
```



2.7.1 Plot standard deviation in time colored according to the percentage of valid samples

```
[ ]: sensor_analysis.plot_std_in_time(valid_resampled_dataframe)
```



3 Save Data

```
[ ]: import os

output_directory_path = 'data/output/'
processing_directory_path = '../data-processing/input/'
valid_1HR_filename = '_valid_data_1HR.csv'
valid_filename = '_valid_data.csv'
dataframe_filename = '_dataframe.csv'
dataframe_1HR_filename = '_dataframe_1hr.csv'

valid_1HR_file_path_output = output_directory_path + SENSOR_NAME +_
    ↪valid_1HR_filename
valid_file_path_output = output_directory_path + SENSOR_NAME + valid_filename

dataframe_path_1HR_output = output_directory_path + SENSOR_NAME +_
    ↪dataframe_1HR_filename
dataframe_path_output = output_directory_path + SENSOR_NAME + dataframe_filename

valid_processing_1HR_file_path = processing_directory_path + SENSOR_NAME +_
    ↪valid_1HR_filename
```

```

valid_processing_file_path = processing_directory_path + SENSOR_NAME +_
    ↪valid_filename

processing_dataframe_path_1HR = processing_directory_path + SENSOR_NAME +_
    ↪dataframe_1HR_filename
processing_dataframe_path = processing_directory_path + SENSOR_NAME +_
    ↪dataframe_filename

if not os.path.exists(output_directory_path):
    os.makedirs(output_directory_path)

if not os.path.exists(processing_directory_path):
    os.makedirs(processing_directory_path)

sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'][['measuring', 'value']]._
    ↪to_csv(valid_file_path_output)
resampled_dataframe[resampled_dataframe['Tag'] == 'VALID'][['measuring', ↪
    'value']].to_csv(valid_1HR_file_path_output)
sensor_dataframe.to_csv(dataframe_path_output)
resampled_dataframe.to_csv(dataframe_path_1HR_output)

sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'][['measuring', 'value']]._
    ↪to_csv(valid_processing_file_path)
resampled_dataframe[resampled_dataframe['Tag'] == 'VALID'][['measuring', ↪
    'value']].to_csv(valid_processing_1HR_file_path)
sensor_dataframe.to_csv(processing_dataframe_path)
resampled_dataframe.to_csv(processing_dataframe_path_1HR)

```

MP10-field-data-treatment-Diamante

February 13, 2024

1 Electrochemical PM10 sensor data preprocessing

- Pollutant: Particulate Matter
- Sensor: Alphasense OPC-N3

1.1 Constants

1.1.1 Sensors IDs

1.1.2 API Constants

```
[ ]: HOST = "renovar.lcqar.ufsc.br"
PORT = 8080
RAW_DATA_DIR = "data/raw-data-monit-fix-2022-2023-Diamante/"
RAW_FILE_NAME = "OPC_PM10.CSV"

SENSOR_FILE_DIR = 'data/input/'
SENSOR_NAME = 'alpha_pm_10_conc'
SENSOR_FILE_NAME = SENSOR_NAME + 'web_dataframe.csv'
```

1.1.3 Sensor Constants

```
[ ]: lower_limit=0.0
upper_limit=2e3
sampling_period = 15 * 60
```

2 Alphasense PM10 Sensor Data

```
[ ]: import locale
locale.setlocale(locale.LC_TIME, 'pt_BR')

[ ]: 'pt_BR'

[ ]: from GetSensorDataService import GetSensorDataService

get_sensor_data_service = GetSensorDataService(HOST, PORT)
sensor_data = get_sensor_data_service.
    ↪get_data_from_file(RAW_DATA_DIR+RAW_FILE_NAME, sensor_name=SENSOR_NAME)
```

```
sensor_data.head()
```

```
/Users/Fernando/Documents/Projects/Github/lcqar-low-cost-monit-proc/data-pre-
processing/GetSensorDataService.py:13: UserWarning: The argument
'infer_datetime_format' is deprecated and will be removed in a future version. A
strict version of it is now the default, see
https://pandas.pydata.org/pdps/0004-consistent-to-datetime-parsing.html. You
can safely remove this argument.
df['DateTime'] = (pd.to_datetime(df[date_time_col],
infer_datetime_format=False, format='%d/%m/%Y/%H/%M/%S'))
```

```
[ ]:   latitude  longitude  measuring          DateTime
0 -28.456899 -48.972999 -9999.99 2022-11-21 10:44:06
1 -28.456899 -48.972999 -9999.99 2022-11-21 10:59:51
2 -28.456899 -48.972999 -9999.99 2022-11-21 11:15:41
3 -28.456899 -48.972999 -9999.99 2022-11-21 11:31:25
4 -28.456899 -48.972999 -9999.99 2022-11-21 11:47:05
```

2.1 Upload Data from File

```
[ ]: import pandas as pd

df = pd.read_csv(SENSOR_FILE_DIR + SENSOR_FILE_NAME)
df.head()
```

```
[ ]:   Unnamed: 0  latitude  longitude  measuring          DateTime
0          0 -28.456899 -48.972999 -9999.99 2022-11-21 10:44:06
1          1 -28.456899 -48.972999 -9999.99 2022-11-21 10:59:51
2          2 -28.456899 -48.972999 -9999.99 2022-11-21 11:15:41
3          3 -28.456899 -48.972999 -9999.99 2022-11-21 11:31:25
4          4 -28.456899 -48.972999 -9999.99 2022-11-21 11:47:05
```

2.1.1 Create Sensor Dataframe as Pandas Series with a period of 15 mins

```
[ ]: # Remove the first column with the indexes and save data into web dataframe
web_dataframe = df.drop(df.columns[0], axis='columns')
web_dataframe['DateTime'] = (pd.to_datetime(df['DateTime'],
                                             infer_datetime_format=True))

# Resample data with 15 mins period and create sensor dataframe
sensor_dataframe = web_dataframe.sort_values(by='DateTime', ascending=True).
    reset_index().drop(columns='index')
sensor_dataframe.index = sensor_dataframe['DateTime']
sensor_dataframe = sensor_dataframe.drop(columns=['DateTime'])
sensor_dataframe = sensor_dataframe.resample('15T').mean()
sensor_dataframe
```

```
/var/folders/wc/_83zcrx913j1dqwg4g90kbhh0000gp/T/ipykernel_4959/166902210.py:3:
```

```
UserWarning: The argument 'infer_datetime_format' is deprecated and will be removed in a future version. A strict version of it is now the default, see https://pandas.pydata.org/pdeps/0004-consistent-to-datetime-parsing.html. You can safely remove this argument.
```

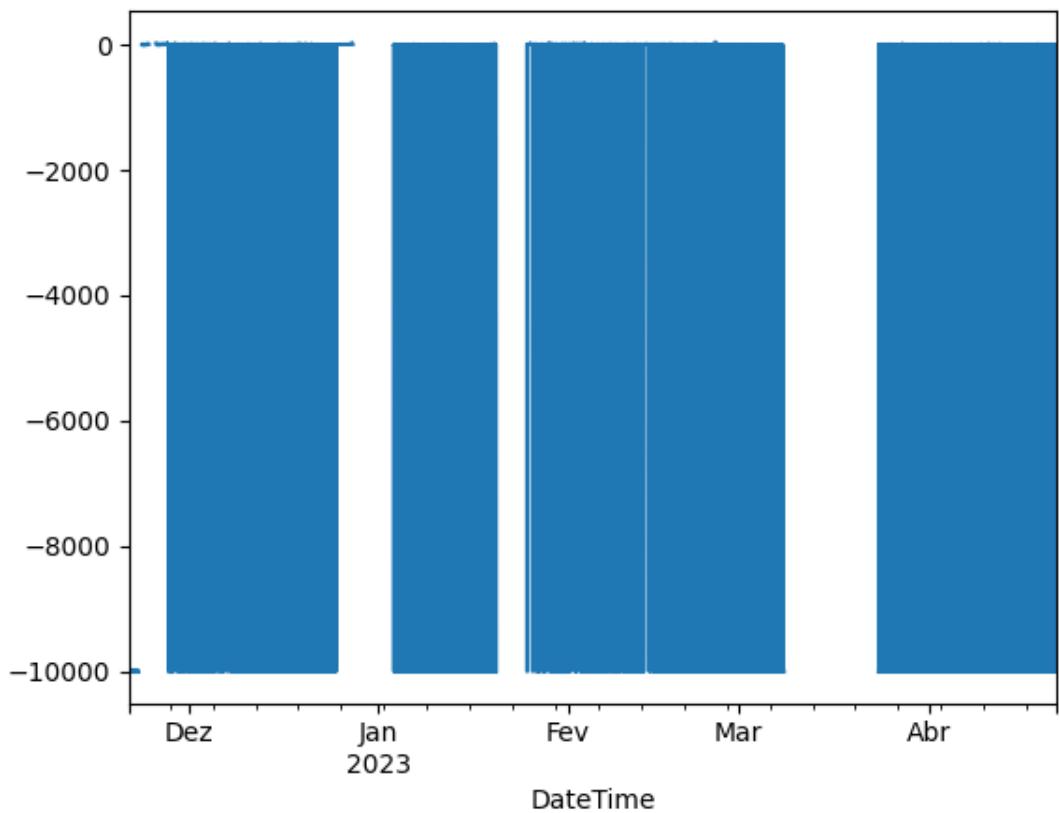
```
web_dataframe['DateTime'] = (pd.to_datetime(df['DateTime']),  
infer_datetime_format=True))
```

```
[ ]:          latitude  longitude  measuring  
DateTime  
2022-11-21 10:30:00 -28.456899 -48.972999 -9999.99  
2022-11-21 10:45:00 -28.456899 -48.972999 -9999.99  
2022-11-21 11:00:00      NaN        NaN        NaN  
2022-11-21 11:15:00 -28.456899 -48.972999 -9999.99  
2022-11-21 11:30:00 -28.456899 -48.972999 -9999.99  
...           ...       ...       ...  
2023-04-21 20:30:00 -28.456899 -48.972999 -9999.99  
2023-04-21 20:45:00 -28.456899 -48.972999 -9999.99  
2023-04-21 21:00:00 -28.456899 -48.972999 -9999.99  
2023-04-21 21:15:00 -28.456899 -48.972999 -9999.99  
2023-04-21 21:30:00 -28.456899 -48.972999 -9999.99  
  
[14541 rows x 3 columns]
```

2.1.2 Plot raw data

```
[ ]: sensor_dataframe['measuring'].plot()
```

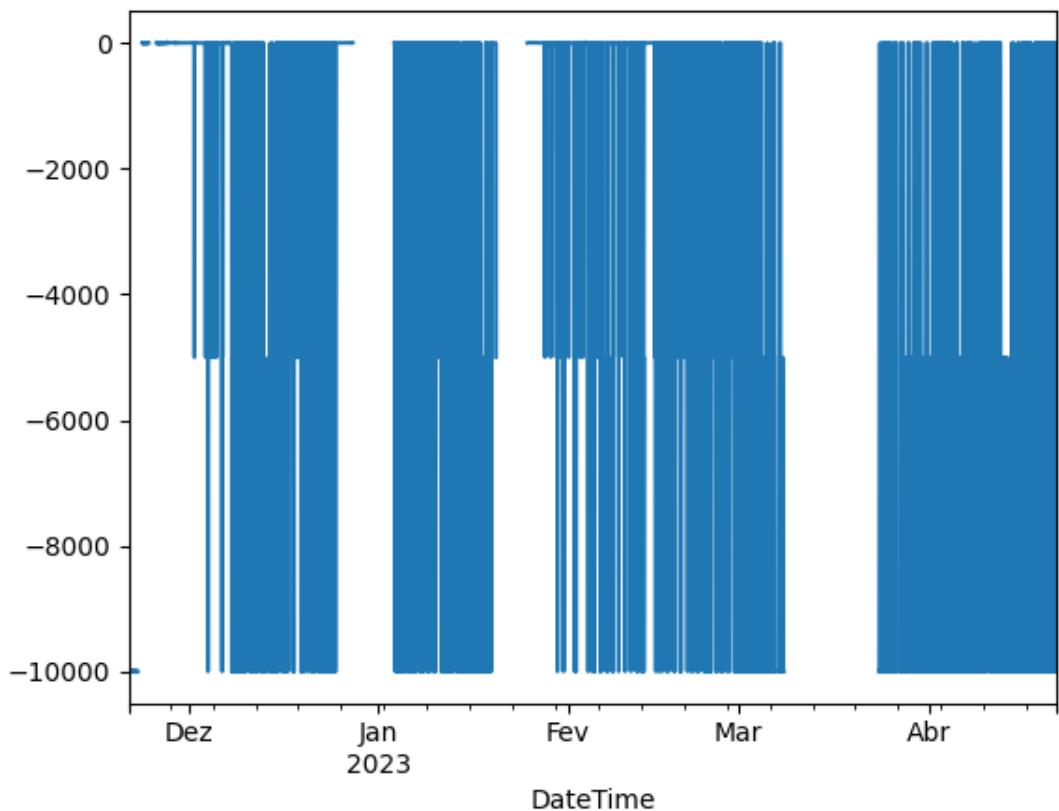
```
[ ]: <Axes: xlabel='DateTime'>
```



2.1.3 Smooth data with a window of 4 samples, i.e.: an hour of data

```
[ ]: sensor_dataframe['raw_measuring'] = sensor_dataframe['measuring']
sensor_dataframe['measuring'] = sensor_dataframe['measuring'].rolling(window=4).
    median()
sensor_dataframe['Hour'] = sensor_dataframe.index.hour
sensor_dataframe['measuring'].plot()
```

```
[ ]: <Axes: xlabel='DateTime'>
```



2.2 Tag data

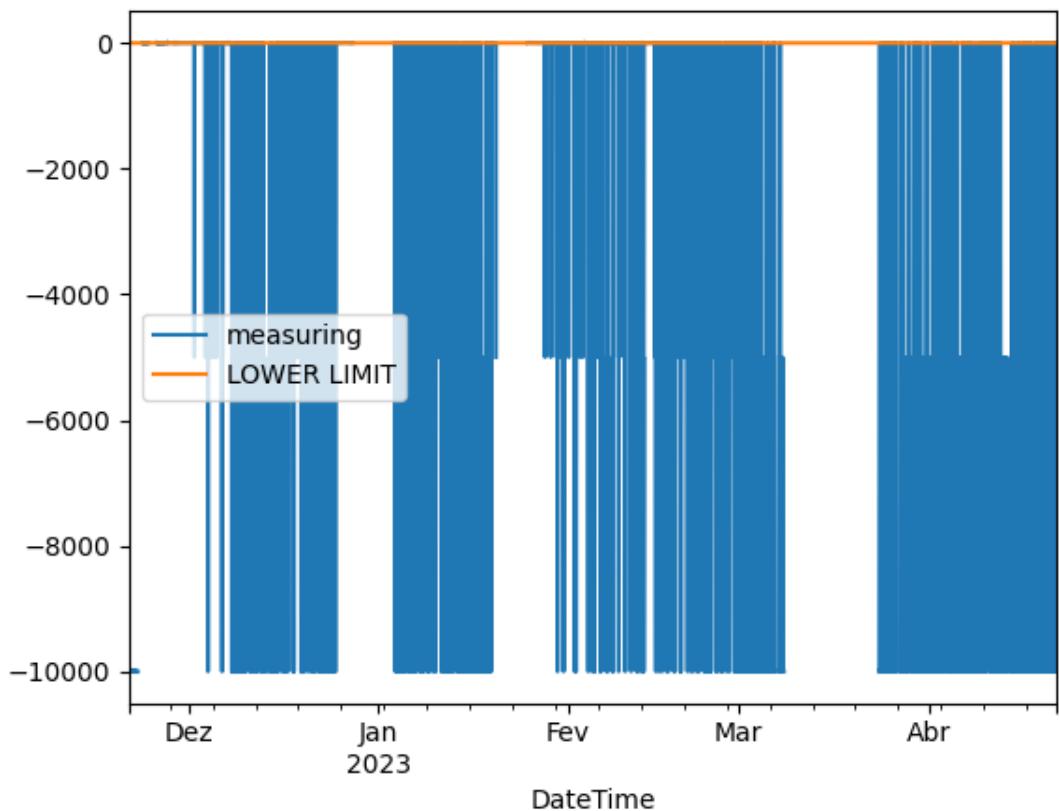
Apply tags to the data according to the quality control processes described above

```
[ ]: sensor_dataframe['Tag'] = 'VALID'
```

Tag values that are missing, are greater than sensor span or are lower than sensor resolution The upper limit was too high that affected the graph resolution, therefore it is not plotted

```
[ ]: sensor_dataframe['UPPER LIMIT'] = upper_limit
sensor_dataframe['LOWER LIMIT'] = lower_limit
sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'][['measuring', 'LOWERLIMIT']].resample('15T').mean().plot()
```

```
[ ]: <Axes: xlabel='DateTime'>
```



```
[ ]: import numpy as np
from SensorDataAnalysisService import SensorDataAnalysisService as
    ↪sensor_analysis

def tag_by_sensor_limits(value, tag, lower_limit, upper_limit):
    if (tag != 'VALID'): return tag
    return sensor_analysis.get_tags_from_series(value=value,
                                                lower_limit=lower_limit,
                                                upper_limit=upper_limit)

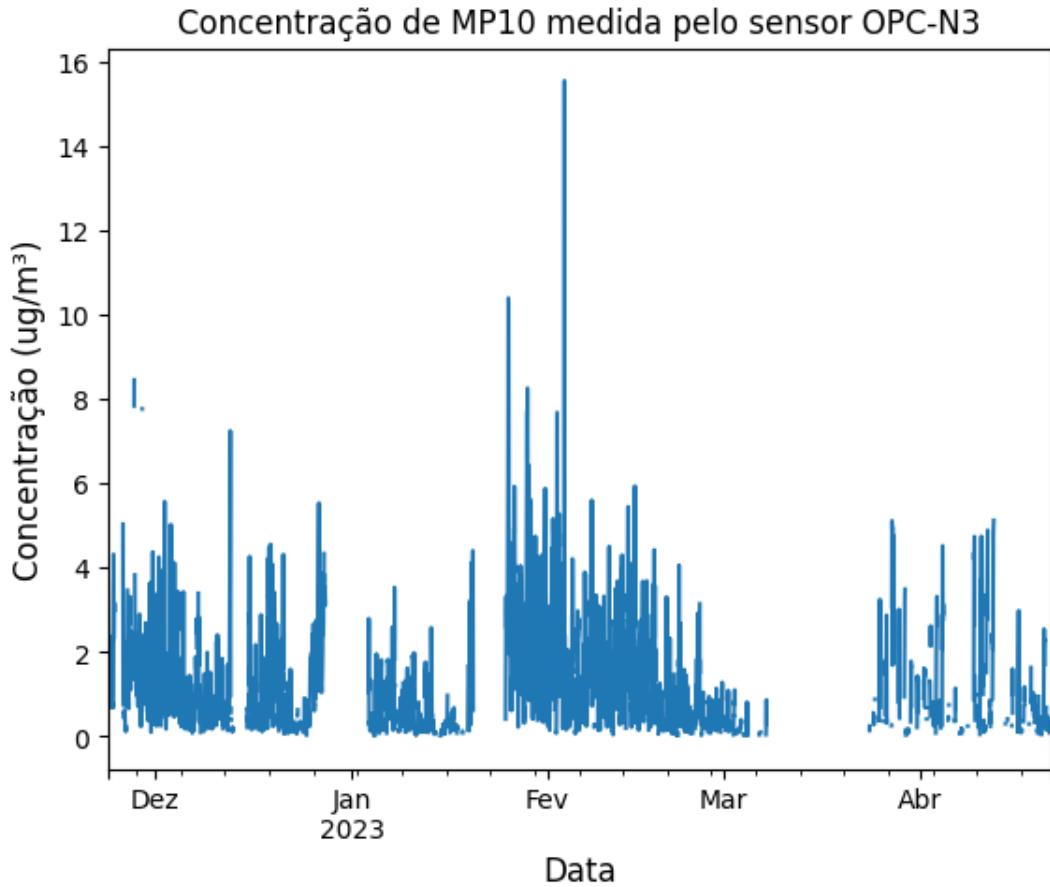
sensor_dataframe['Tag'] = (sensor_dataframe[['measuring', 'Tag']]
                           .apply(lambda df: ↪
                                 ↪tag_by_sensor_limits(value=df[0], tag=df[1],
                                 ↪lower_limit=lower_limit,
                                 ↪upper_limit=upper_limit),
                                 axis=1))
```

Plot valid data after removing samples bellow lower limit

```
[ ]: import matplotlib.pyplot as plt

fig, ax = plt.subplots(figsize=(1.3*5,5))
sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'][['measuring']]
    .rename('Concentração de MP10').resample('15T').mean().plot()
plt.title('Concentração de MP10 medida pelo sensor OPC-N3')
ax.set_xlabel('Data', fontsize=12)
ax.set_ylabel('Concentração (ug/m3)', fontsize=12)
```

```
[ ]: Text(0, 0.5, 'Concentração (ug/m3)')
```



2.3 Analyse valid data

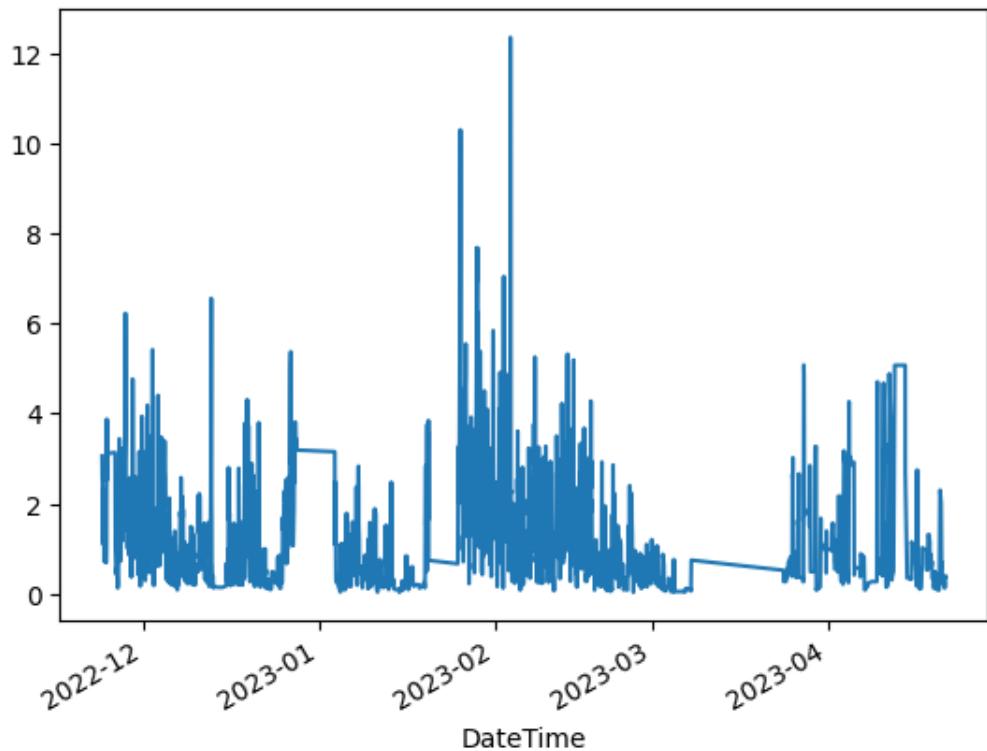
The graph shows the time series of the valid data, box plots of the data grouped by hour of the day, and histogram of the data.

```
[ ]: valid_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'].
    .drop(columns=['Tag'])
```

2.3.1 Smooth data with a window of 4 samples, i.e.: an hour of data

```
[ ]: valid_dataframe['raw measuring'] = valid_dataframe['measuring']
      valid_dataframe['measuring'] = valid_dataframe['measuring'].rolling(window=4).
        median()
      valid_dataframe['measuring'].plot()
```

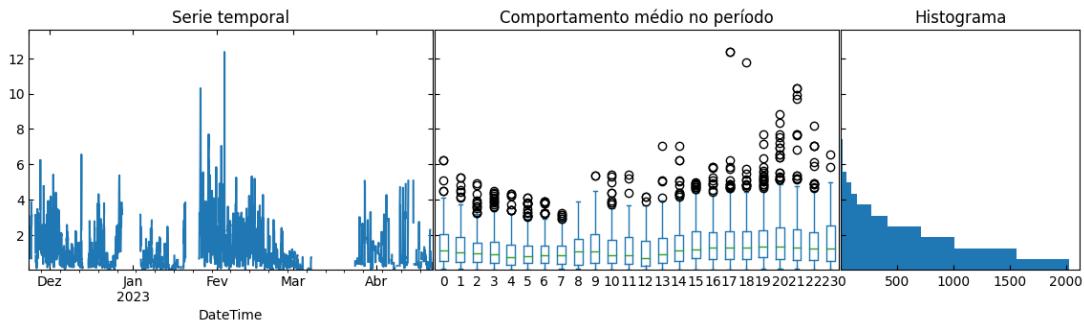
```
[ ]: <Axes: xlabel='DateTime'>
```



```
[ ]: valid_dataframe = valid_dataframe.resample('15T').mean()
      sensor_analysis.plot_box_hist(df=valid_dataframe, bins=20)
```

```
/Users/Fernando/Documents/Projects/Github/lcqar-low-cost-monit-proc/data-pre-
processing/SensorDataAnalysisService.py:111: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

```
See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
df['Hour'] = df['Hour'].astype('int64')
```



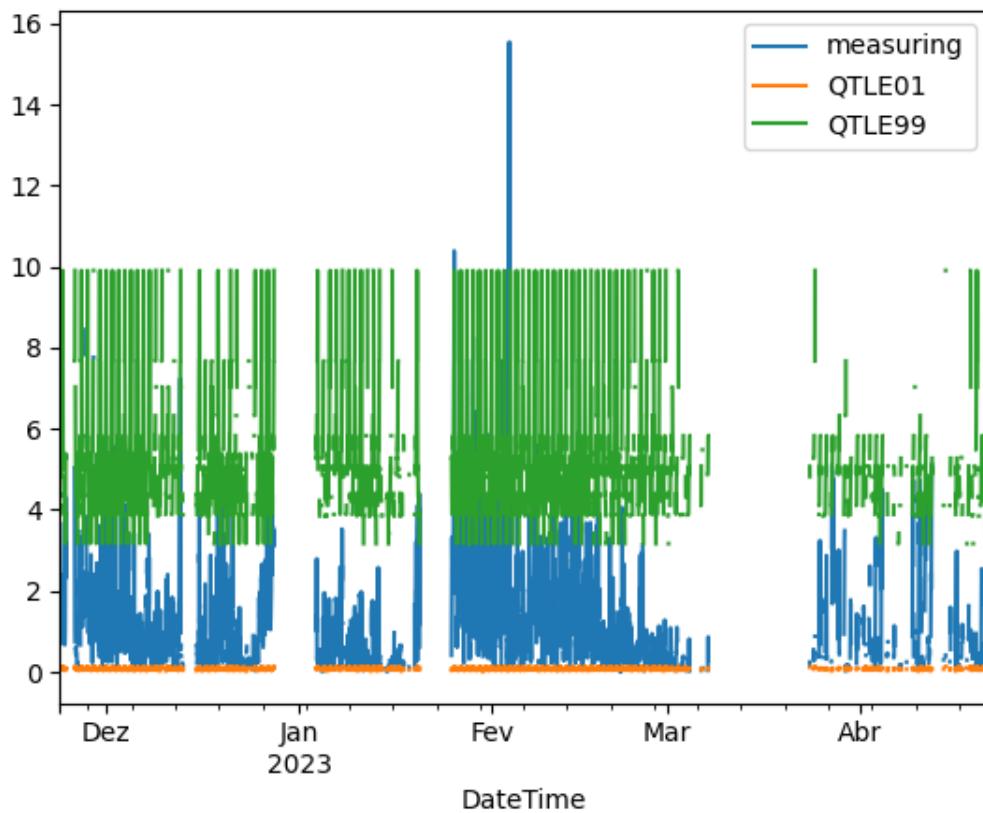
2.3.2 Calculate quantiles

We calculated the 1% and 99 % quantiles of every hour of the day. The values greater outside the 1 % - 99 % were tagged as Greater than Quantile 99 (GTQTL99) and Lower than Quantile 1 (LTQTL01)

```
[ ]: import numpy as np

global_qtle_01 = valid_dataframe.pivot(columns='Hour')['measuring'].
    ↪quantile(q=0.01, axis='index', interpolation='lower').dropna()
global_qtle_99 = valid_dataframe.pivot(columns='Hour')['measuring'].
    ↪quantile(q=0.99, axis='index', interpolation='higher').dropna()
sensor_dataframe['QTLE01'] = sensor_dataframe['Hour'].map(lambda hr: ↴
    ↪global_qtle_01[hr] if (not np.isnan(hr)) else np.nan)
sensor_dataframe['QTLE99'] = sensor_dataframe['Hour'].map(lambda hr: ↴
    ↪global_qtle_99[hr] if (not np.isnan(hr)) else np.nan)
sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'][['measuring', 'QTLE01', ↴
    ↪'QTLE99']].resample('15T').mean().plot()

[ ]: <Axes: xlabel='DateTime'>
```



2.3.3 Tag data according to quantiles

```
[ ]: sensor_dataframe['Tag'] = (sensor_dataframe[['Tag', 'measuring', 'QTLE01', 'QTLE99']]
    .apply(lambda df: sensor_analysis.
        tag_by_quantiles(current_tag=df[0],
            value=df[1],
            quantile_01=df[2],
            quantile_99=df[3]),
            axis=1))
tags = ['MISSING', 'LTLL', 'GTUL', 'BADSPIKE', 'VALID', 'LTQTLE01', 'GTQTLE99']
data_contabilization = sensor_analysis.count_tags(tags, sensor_dataframe)
data_contabilization
```

```
[ ]:          #      %
MISSING     6481  44.570525
```

```

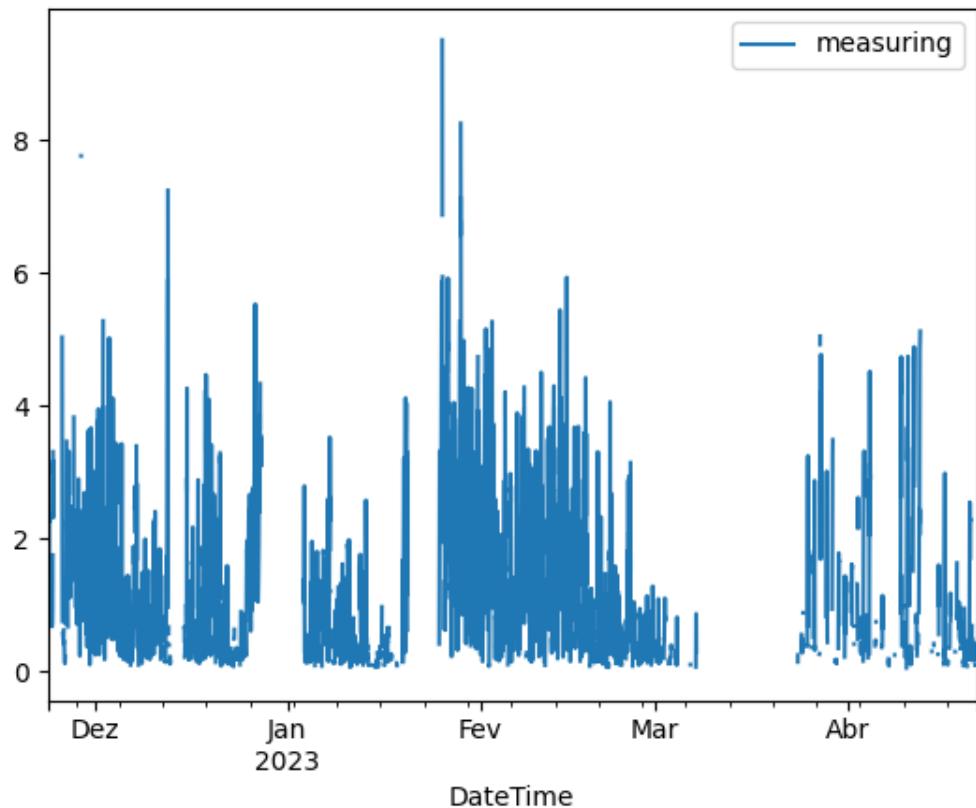
LTLL      1759   12.09683
GTUL       0     0.0
BADSPIKE   0     0.0
VALID     6098  41.936593
LTQTLE01   117   0.804621
GTQTLE99    86   0.591431
TOTAL    14541   100.0

```

Plot valid data

```
[ ]: sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'][['measuring']].  
    ↪resample('15T').mean().plot()
```

```
[ ]: <Axes: xlabel='DateTime'>
```



Analyse data after removing quantiles

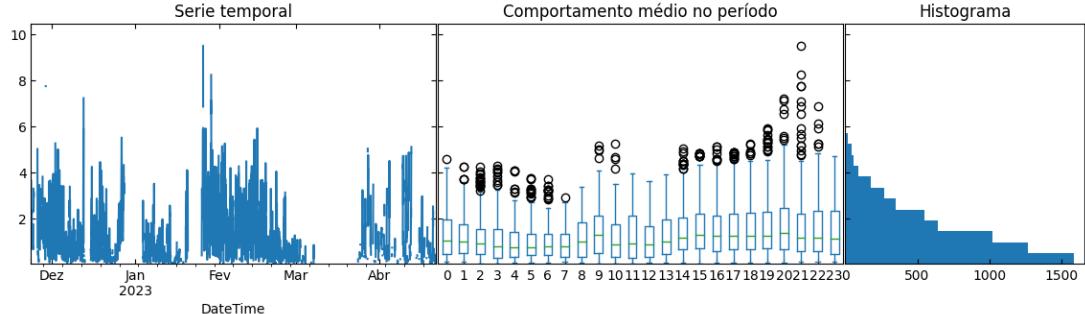
```
[ ]: valid_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'].  
    ↪drop(columns=['Tag']).resample('15T').mean()  
sensor_analysis.plot_box_hist(df=valid_dataframe, bins=20)
```

```
/Users/Fernando/Documents/Projects/Github/lcqar-low-cost-monit-proc/data-pre-
```

```
processing/SensorDataAnalysisService.py:111: SettingWithCopyWarning:  
A value is trying to be set on a copy of a slice from a DataFrame.  
Try using .loc[row_indexer,col_indexer] = value instead
```

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

```
df['Hour'] = df['Hour'].astype('int64')
```



2.4 Analyse data derivatives

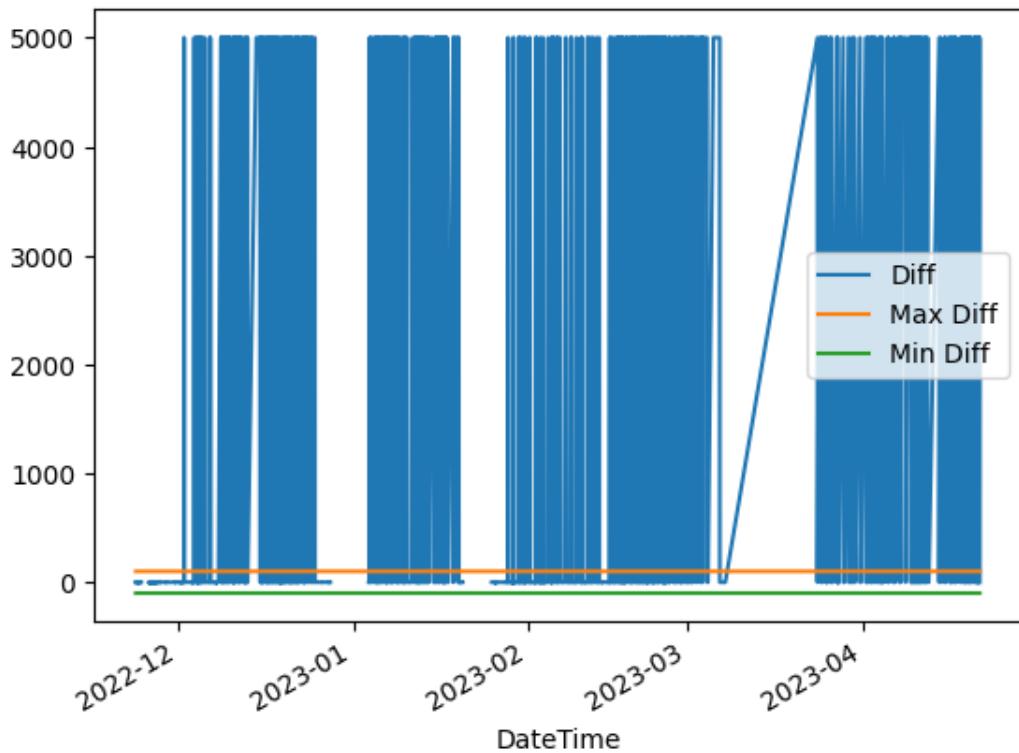
Analyse the derivatives of the data for removing abrupt changes in the data (BADSPIKE)

Tag values with derivatives faster than sensor response time The value of 100 ug/m³ was selected based on the maximum derivative found in the reference data. The maximum derivative (with measuring period of 1 hour) found was 150 ug/m³ ppb. Therefore, a maximum of 100 ug/m³ was considered appropriate for a 15 mins period.

```
[ ]: max_diff_value = 100  
sensor_dataframe['Diff'] = sensor_dataframe['measuring'].resample('15T').mean()  
    .diff()  
sensor_dataframe['Max Diff'] = max_diff_value  
sensor_dataframe['Min Diff'] = -max_diff_value  
sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'][['Diff', 'Max Diff', 'Min Diff']].plot()
```



```
[ ]: <Axes: xlabel='DateTime'>
```



```
[ ]: import numpy as np

def tag_data_with_derivatives(tagged_df, max_diff_value):
    current_tag = tagged_df[0]
    value = tagged_df[1]
    if ((current_tag != 'VALID') or (np.isnan(value))): return current_tag
    if ((value > max_diff_value) or (value < -max_diff_value)): return
    ↵'BADSPIKE'
    return 'VALID'

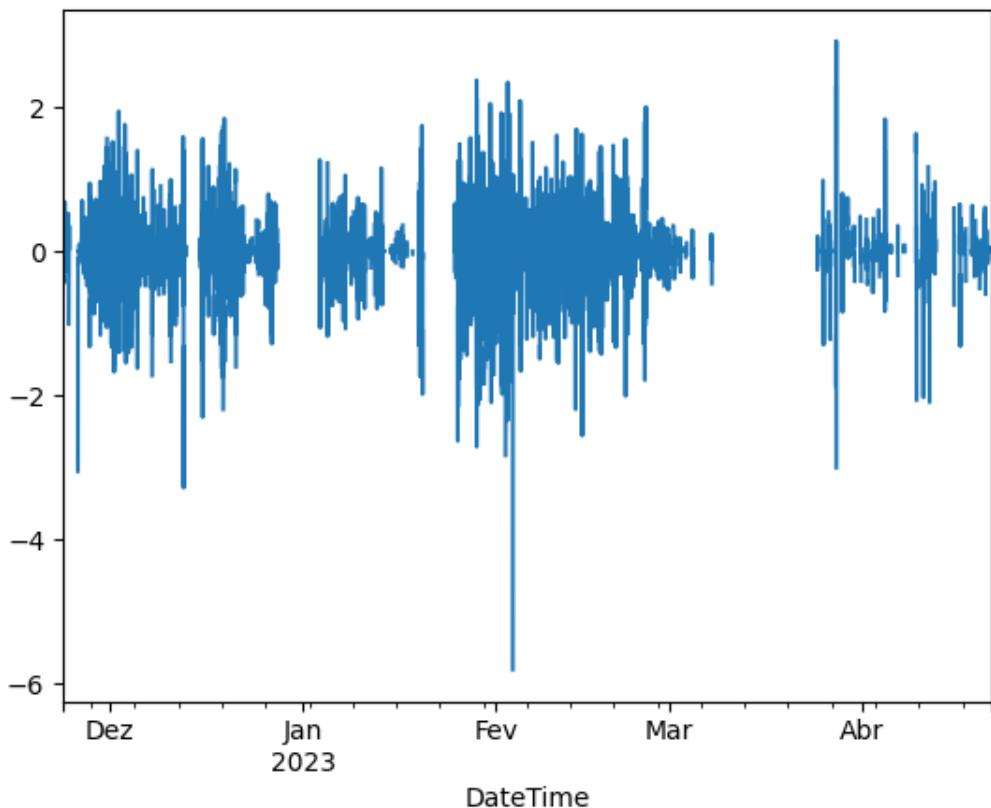
sensor_dataframe['Tag'] = (sensor_dataframe[['Tag', 'Diff', 'Max Diff']]
                           .apply(lambda df:
    ↵tag_data_with_derivatives(tagged_df=df, max_diff_value=df[2]),
                           axis=1))

tags = ['MISSING', 'LTLL', 'GTUL', 'STABILIZING', 'BADSPIKE', 'VALID',
    ↵'LTQTLE01', 'GTQTLE99', 'REBASE']
data_contabilization = sensor_analysis.count_tags(tags, sensor_dataframe)
data_contabilization
```

```
[ ]: # %  
MISSING      6481  44.570525  
LTLL         1759   12.09683  
GTUL          0     0.0  
STABILIZING   0     0.0  
BADSPIKE     427    2.936524  
VALID        5671  39.000069  
LTQTLE01     117    0.804621  
GTQTLE99     86    0.591431  
REBASE        0     0.0  
TOTAL       14541  100.0
```

```
[ ]: sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['Diff'].resample('15T').  
mean().plot()
```

```
[ ]: <Axes: xlabel='DateTime'>
```



Plot data derivatives vs data

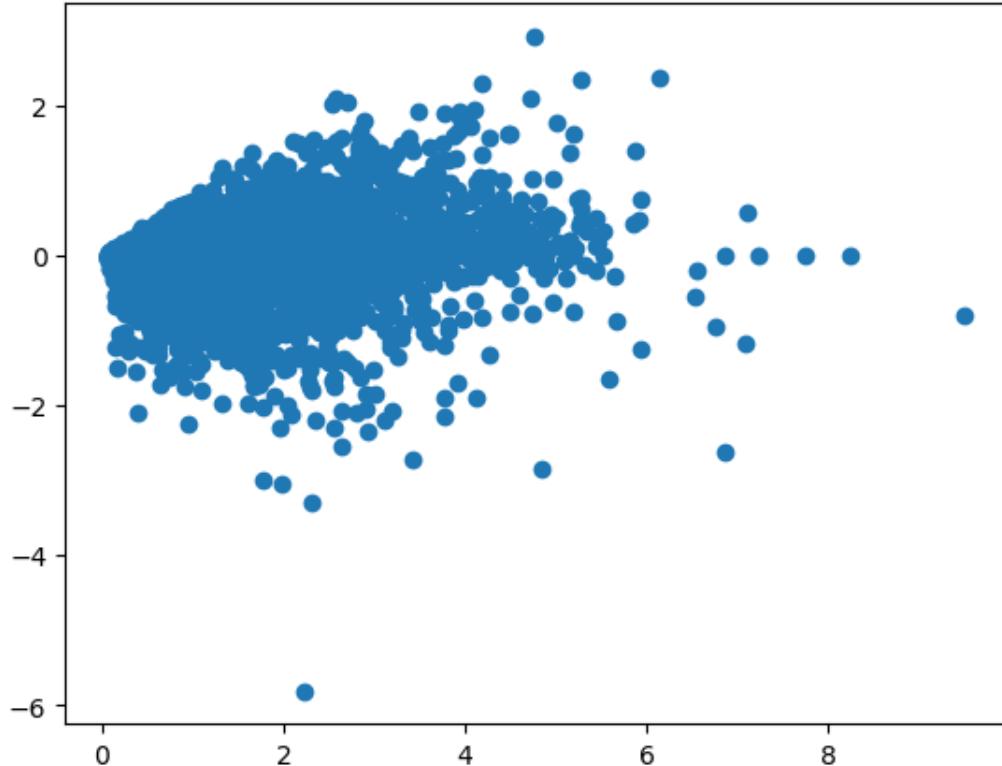
```
[ ]: import matplotlib.pyplot as plt
```

```

valid_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']
fig = plt.figure(figsize=(1.3*5,5))
plt.scatter(valid_dataframe['measuring'], valid_dataframe['Diff'])

```

[]: <matplotlib.collections.PathCollection at 0x168e1c220>

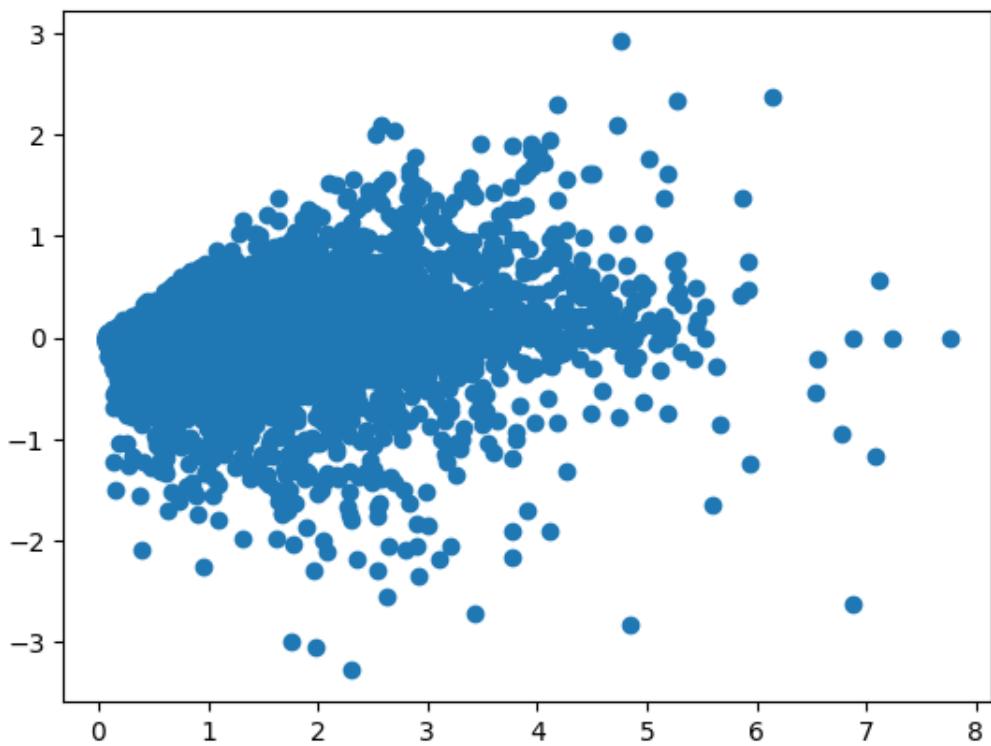


Remove sample with derivative < -4 and value 8

```

[ ]: sensor_dataframe['Tag'] = (sensor_dataframe[['measuring', 'Diff', 'Tag']])
      .apply(lambda df: 'BADSPIKE' if df[2] == 'VALID' and
      ↪(df[0] > 8 or df[1] <= -4) else df[2], axis=1)
plt.scatter(sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['measuring'],
           sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['Diff'])
valid_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']

```



Data Tag contabilization

```
[ ]: tags = ['MISSING', 'LTLL', 'GTUL', 'STABILIZING', 'BADSPIKE', 'VALID',  
          'LTQTLE01', 'GTQTLE99', 'REBASE']  
data_contabilization = sensor_analysis.count_tags(tags, sensor_dataframe)  
data_contabilization
```

	#	%
MISSING	6481	44.570525
LTLL	1759	12.09683
GTUL	0	0.0
STABILIZING	0	0.0
BADSPIKE	430	2.957156
VALID	5668	38.979437
LTQTLE01	117	0.804621
GTQTLE99	86	0.591431
REBASE	0	0.0
TOTAL	14541	100.0

Plot valid data

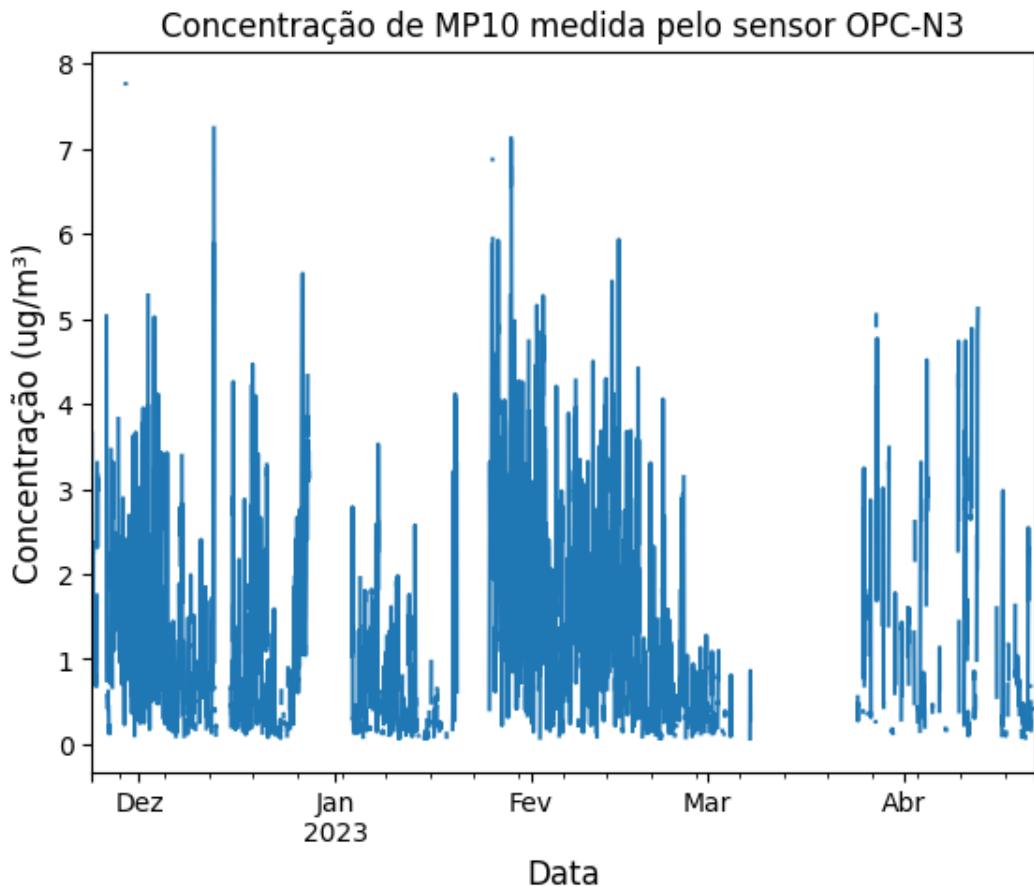
```
[ ]: fig, ax = plt.subplots(figsize=(1.3*5,5))
```

```

sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['measuring'].
    rename('Concentração de MP10').resample('15T').mean().plot()
plt.title('Concentração de MP10 medida pelo sensor OPC-N3')
ax.set_xlabel('Data', fontsize=12)
ax.set_ylabel('Concentração (ug/m3)', fontsize=12)

```

[]: Text(0, 0.5, 'Concentração (ug/m³)')



[]: valid_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'].
 drop(columns=['Tag']).resample('15T').mean()
sensor_analysis.plot_box(df=valid_dataframe)

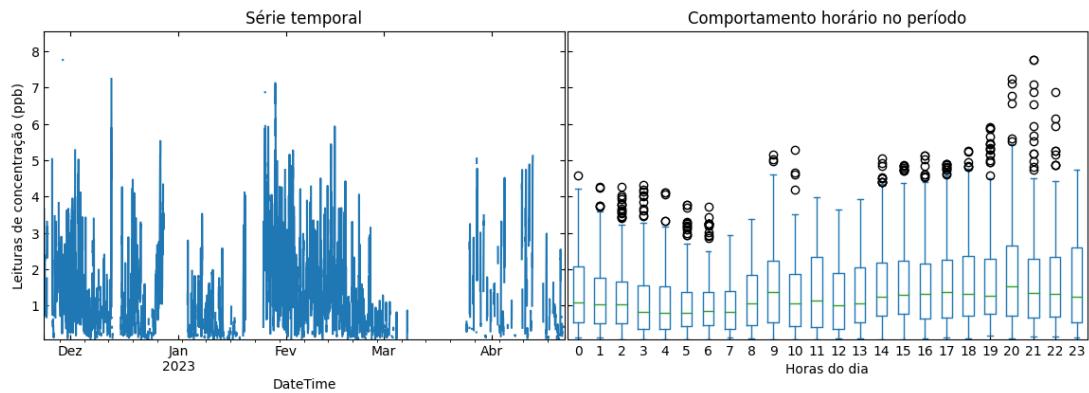
```

/Users/Fernando/Documents/Projects/Github/lcqar-low-cost-monit-proc/data-pre-
processing/SensorDataAnalysisService.py:143: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

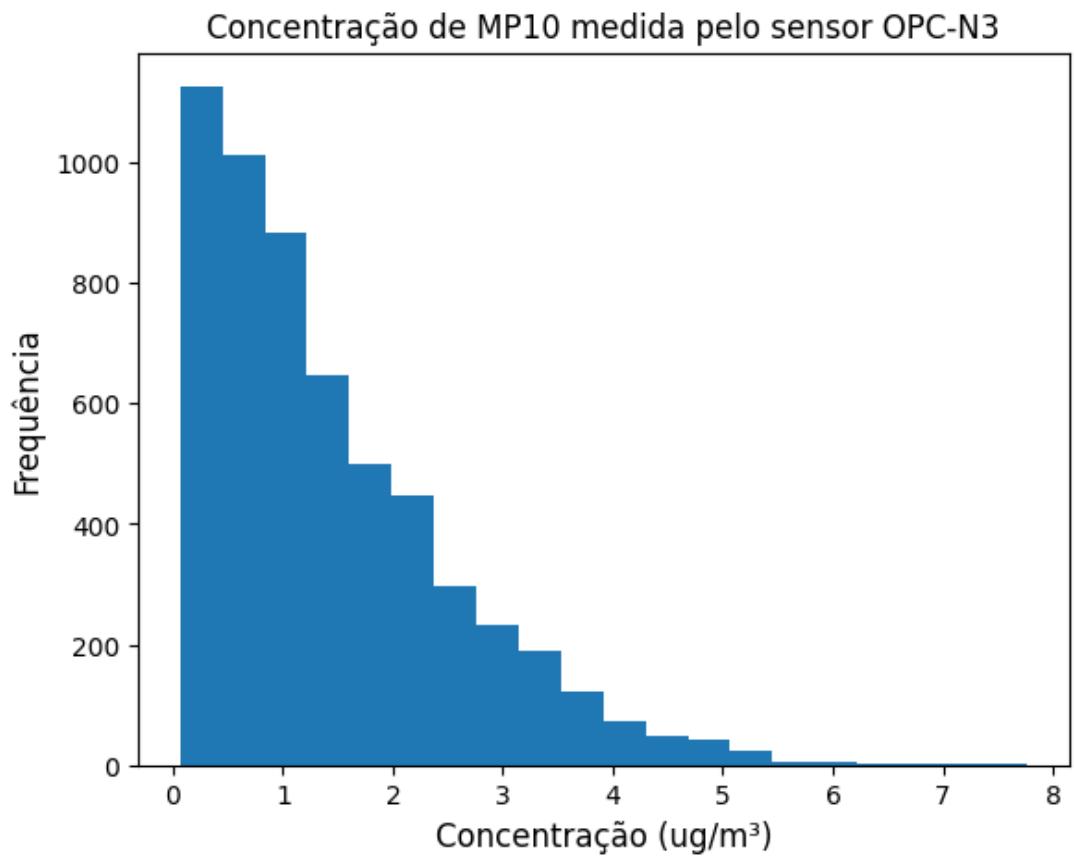
```

See the caveats in the documentation: <https://pandas.pydata.org/pandas->

```
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
df['Hour'] = df['Hour'].astype('int64')
```

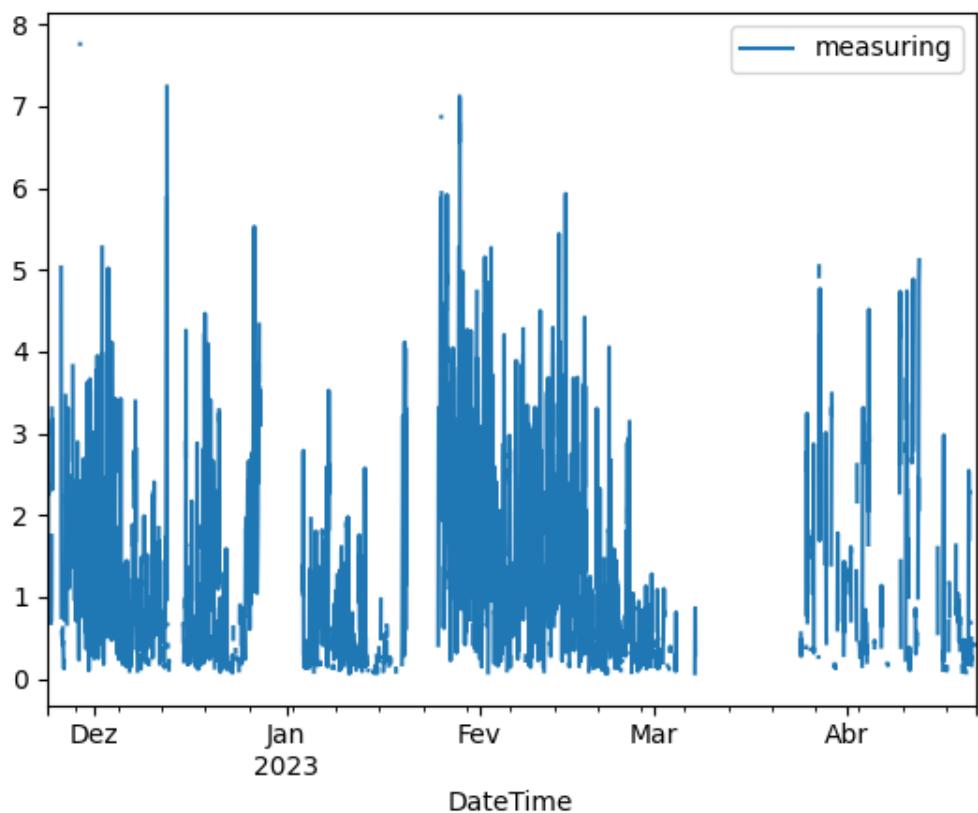


```
[ ]: fig, ax = plt.subplots(figsize=(1.3*5,5))
(sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['measuring'].
    rename('Concentração de MP10')
    .resample('15T').mean().hist(bins=20))
plt.title('Concentração de MP10 medida pelo sensor OPC-N3')
ax.set_xlabel('Concentração (ug/m\N{SUPERSCRIPT THREE})', fontsize=12)
ax.set_ylabel('Frequência', fontsize=12)
ax.grid(False)
```



```
[ ]: sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'][['measuring']].  
    resample('15T').mean().plot()
```

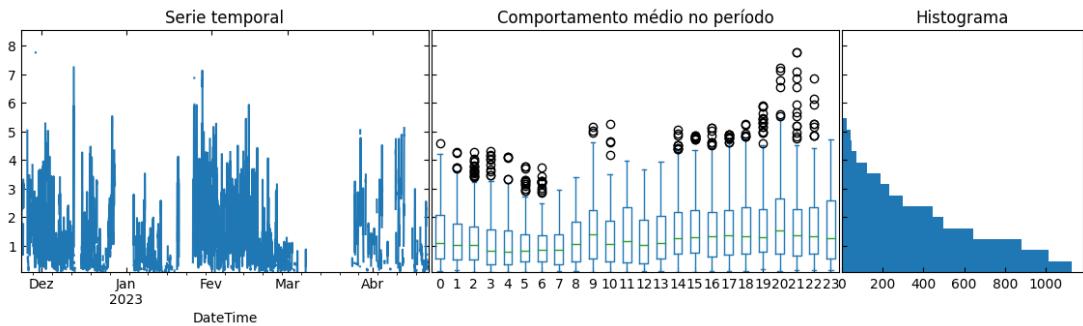
```
[ ]: <Axes: xlabel='DateTime'>
```



```
[ ]: valid_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'].
    drop(columns=['Tag']).resample('15T').mean()
sensor_analysis.plot_box_hist(df=valid_dataframe, bins=20)
```

/Users/Fernando/Documents/Projects/Github/lcqar-low-cost-monit-proc/data-pre-processing/SensorDataAnalysisService.py:111: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

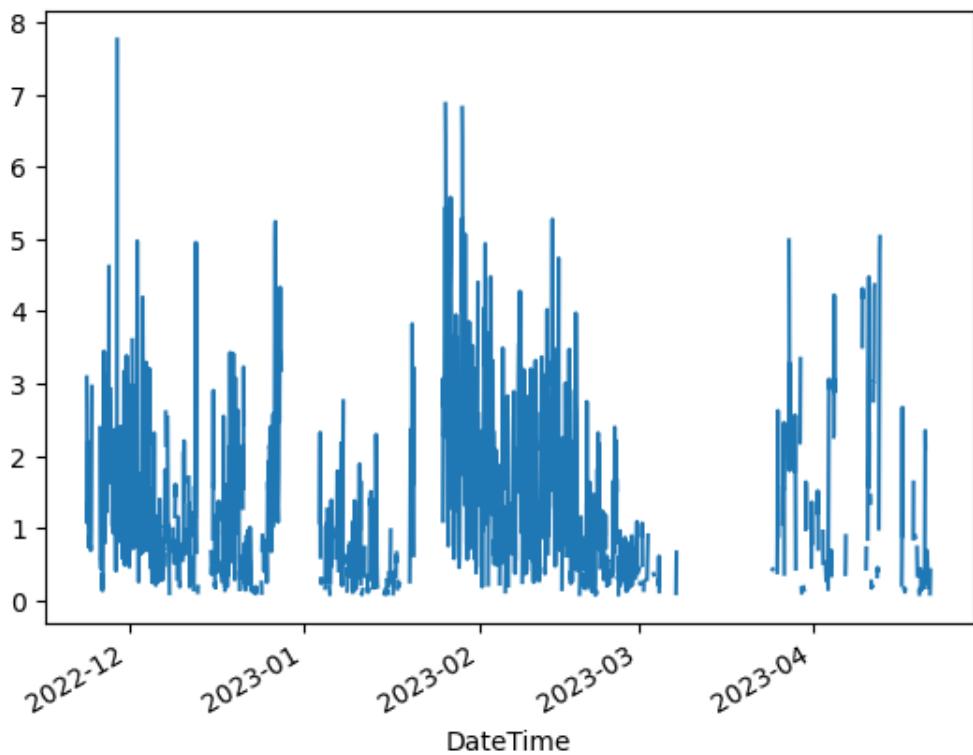
See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
df['Hour'] = df['Hour'].astype('int64')



2.5 Resample valid data to 1 HR

```
[ ]: resampled_dataframe = sensor_dataframe[sensor_dataframe['Tag'] == 'VALID'].
    ↪drop(columns=['Tag']).resample('H').mean()
resampled_dataframe['Hour'] = resampled_dataframe.index.hour
resampled_dataframe['Count'] = (valid_dataframe['measuring'].resample('H').
    ↪count())
resampled_dataframe['Std'] = (valid_dataframe['measuring'].resample('H').std())
resampled_dataframe.index = resampled_dataframe.index.map(lambda t: t.
    ↪replace(minute=30, second=0))
resampled_dataframe['measuring'].plot()
```

```
[ ]: <Axes: xlabel='DateTime'>
```



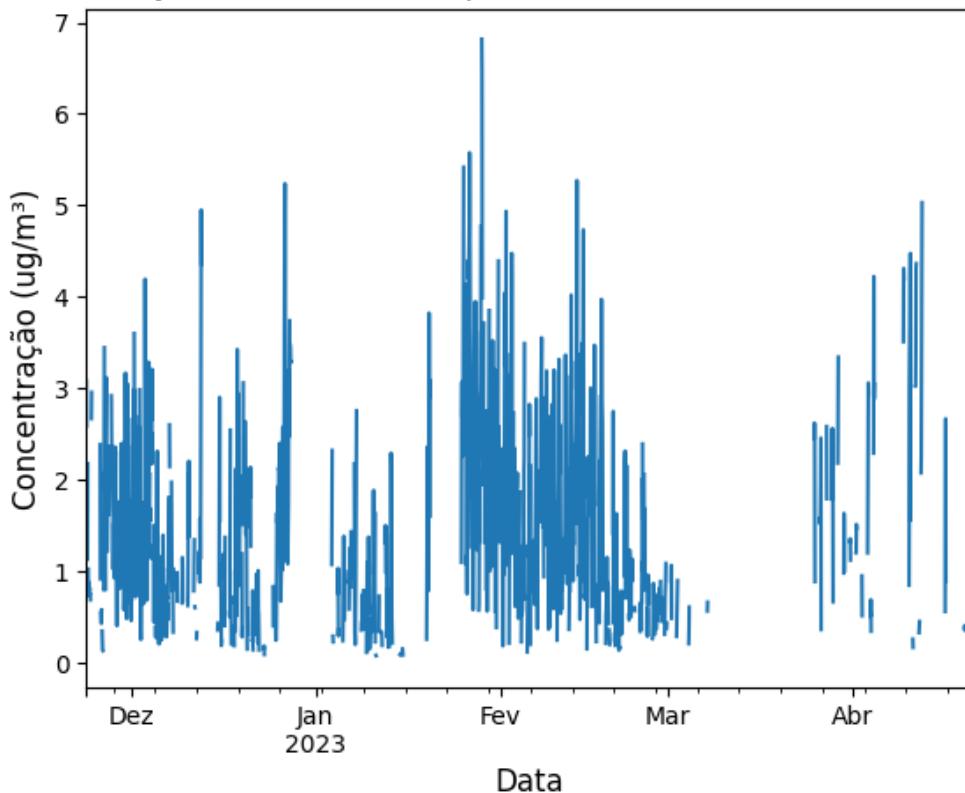
Tag hourly data according to the number of samples in an hour At least 3 samples must be valid in an hour (75 %) for the hourly data be considered as valid

```
[ ]: original_freq = sensor_dataframe.index.freq
resampled_dataframe['% valid'] = (resampled_dataframe['Count']
.map(lambda c:
c / (pd.Timedelta("1 hour") / ↴
original_freq) * 100))
resampled_dataframe['Tag'] = (resampled_dataframe['% valid']
.map(lambda c: 'VALID' if c >= 75 else ↴
'LOWSAMPLES'))

fig, ax = plt.subplots(figsize=(1.3*5,5))
resampled_dataframe[resampled_dataframe['Tag'] == 'VALID']['measuring'].\
rename('Concentração de MP10').resample('1H').mean().plot()
plt.title('Concentração de MP10 medida pelo sensor OPC-N3-B4 (Dados Horários)')
ax.set_xlabel('Data', fontsize=12)
ax.set_ylabel('Concentração (ug/m3)', fontsize=12)

[ ]: Text(0, 0.5, 'Concentração (ug/m3)')
```

Concentração de MP10 medida pelo sensor OPC-N3-B4 (Dados Horários)



Valid data contabilization

```
[ ]: tags = ['LOWSAMPLES', 'VALID']
data_contabilization = sensor_analysis.count_tags(tags, resampled_dataframe)
data_contabilization
```

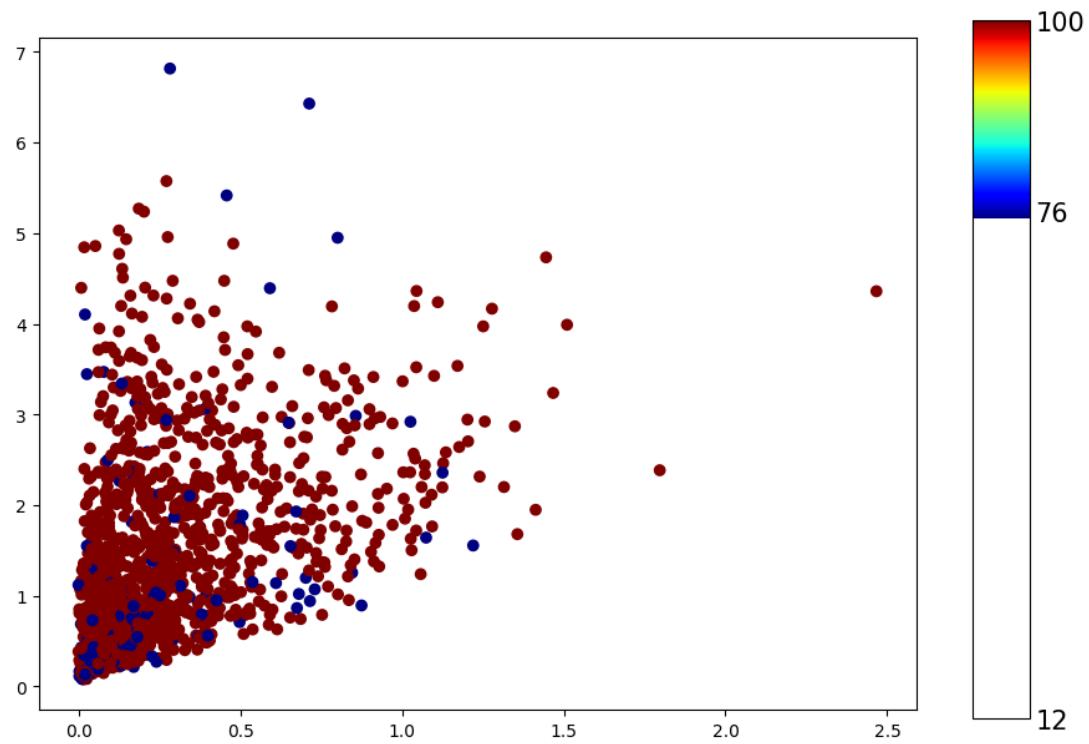
	#	%
LOWSAMPLES	2269	63.3445
VALID	1313	36.6555
TOTAL	3582	100.0

2.6 Analyse the mean and standard deviation of the resampled data

The color of each data point represents the percentage of valid 15 mins samples found in an hour

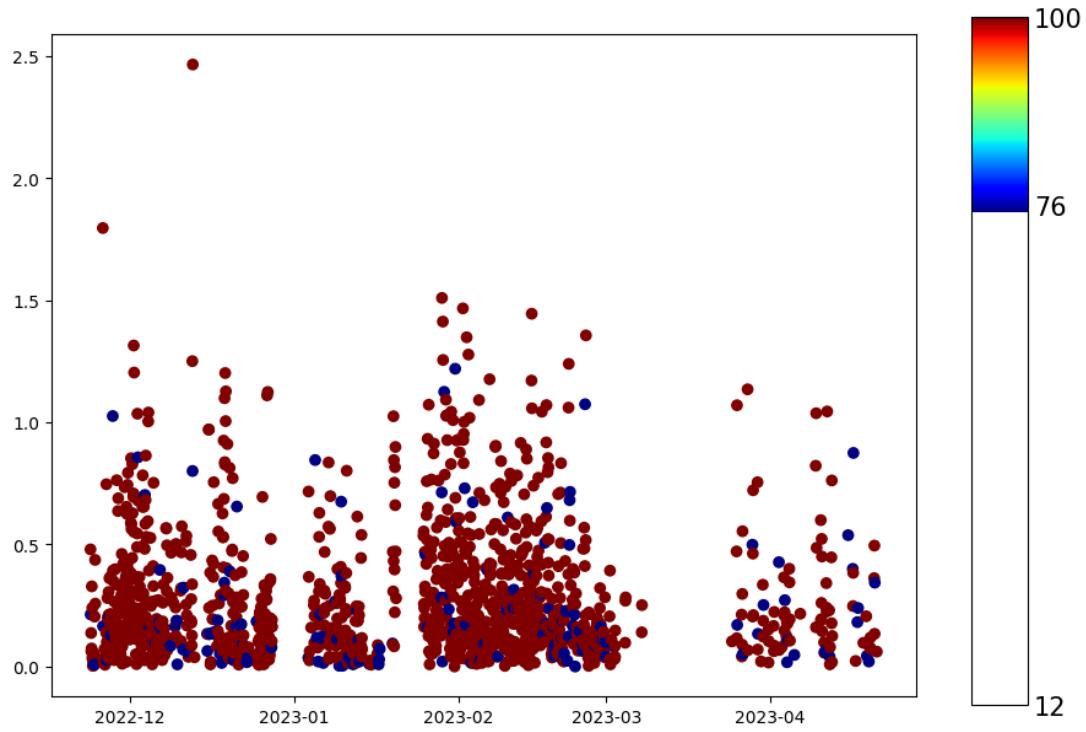
Plot mean (Y axis) vs. standard deviation (X axis) for valid hourly data colored by valid percentage of valid samples

```
[ ]: valid_resampled_dataframe = resampled_dataframe[resampled_dataframe['Tag'] == 'VALID']
sensor_analysis.plot_mean_vs_std(valid_resampled_dataframe)
```



2.6.1 Plot standard deviation in time colored according to the percentage of valid samples

```
[ ]: sensor_analysis.plot_std_in_time(valid_resampled_dataframe)
```



3 Save Data

```
[ ]: import os

output_directory_path = 'data/output/'
processing_directory_path = '../data-processing/input/'
valid_1HR_filename = '_valid_data_1HR.csv'
valid_filename = '_valid_data.csv'
dataframe_filename = '_dataframe.csv'
dataframe_1HR_filename = '_dataframe_1hr.csv'

valid_1HR_file_path_output = output_directory_path + SENSOR_NAME + \
    ↪valid_1HR_filename
valid_file_path_output = output_directory_path + SENSOR_NAME + valid_filename

dataframe_path_1HR_output = output_directory_path + SENSOR_NAME + \
    ↪dataframe_1HR_filename
dataframe_path_output = output_directory_path + SENSOR_NAME + dataframe_filename

valid_processing_1HR_file_path = processing_directory_path + SENSOR_NAME + \
    ↪valid_1HR_filename
```

```

valid_processing_file_path = processing_directory_path + SENSOR_NAME +_
    ↪valid_filename

processing_dataframe_path_1HR = processing_directory_path + SENSOR_NAME +_
    ↪dataframe_1HR_filename
processing_dataframe_path = processing_directory_path + SENSOR_NAME +_
    ↪dataframe_filename

if not os.path.exists(output_directory_path):
    os.makedirs(output_directory_path)

if not os.path.exists(processing_directory_path):
    os.makedirs(processing_directory_path)

sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['measuring'].
    ↪to_csv(valid_file_path_output)
resampled_dataframe[resampled_dataframe['Tag'] == 'VALID']['measuring'].
    ↪to_csv(valid_1HR_file_path_output)
sensor_dataframe.to_csv(dataframe_path_output)
resampled_dataframe.to_csv(dataframe_path_1HR_output)

sensor_dataframe[sensor_dataframe['Tag'] == 'VALID']['measuring'].
    ↪to_csv(valid_processing_file_path)
resampled_dataframe[resampled_dataframe['Tag'] == 'VALID']['measuring'].
    ↪to_csv(valid_processing_1HR_file_path)
sensor_dataframe.to_csv(processing_dataframe_path)
resampled_dataframe.to_csv(processing_dataframe_path_1HR)

```

ANEXO D – JUPYTER NOTEBOOKS COM METODOLOGIAS DE CORREÇÃO APLICADAS ÀS LEITURAS DOS SENSORES OX-B431

D.1 METODOLOGIA CONSIDERANDO APENAS LEITURAS DOS SENSORES OX-B431 E A TEMPERATURA

O3_reference_and_sensor_data_with_temp

February 14, 2024

0.0.1 Sensor Constants

```
[ ]: sensor_name = 'alpha_o3_conc'  
sensor_1_name = 'alpha_o3_1_conc'  
sensor_2_name = 'alpha_o3_2_conc'
```

0.1 Upload Data from File

0.1.1 Sensor 1

```
[ ]: import pandas as pd  
  
directory_path = 'input/'  
file_name = sensor_1_name + '_and_temp_valid_1HR.csv'  
df_1 = pd.read_csv(directory_path + file_name)  
df_1.head()
```

```
[ ]:          DateTime measuring 1 temperature Hour measuring 1 no Temp \  
0 2022-12-14 14:30:00    42.267696   31.52000  14           56.369783  
1 2022-12-14 15:30:00    50.822340   30.56750  15           69.328964  
2 2022-12-14 16:30:00    67.516902   28.82875  16           94.063829  
3 2022-12-14 17:30:00    68.069052   27.91125  17           98.858670  
4 2022-12-14 18:30:00    84.294900   27.16250  18          118.546877  
  
Count 1      Tag  
0        3  VALID  
1        4  VALID  
2        4  VALID  
3        4  VALID  
4        4  VALID
```

```
[ ]: import locale  
locale.setlocale(locale.LC_TIME, 'pt_BR')
```

```
[ ]: 'pt_BR'
```

0.1.2 Create Sensor Dataframe as Pandas Series

```
[ ]: # Remove the first column with the indexes and save data into web dataframe
dataframe = df_1.drop(df_1.columns[0], axis='columns')
dataframe['DateTime'] = (pd.to_datetime(df_1['DateTime']),  

    ↪infer_datetime_format=True))
```

```
# Resample data with 15 mins period and create sensor dataframe
sensor_1_dataframe = dataframe.sort_values(by='DateTime', ascending=True).  

    ↪reset_index().drop(columns='index')
sensor_1_dataframe.index = sensor_1_dataframe['DateTime']
sensor_1_dataframe = sensor_1_dataframe.drop(columns=['DateTime', 'Hour'])
sensor_1_dataframe = sensor_1_dataframe.rename(columns={'temperature':  

    ↪'temperature 1'})
sensor_1_dataframe
```

```
/var/folders/wc/_83zcrx913j1dqwg4g90kbhh0000gp/T/ipykernel_6358/1180347280.py:3:
UserWarning: The argument 'infer_datetime_format' is deprecated and will be
removed in a future version. A strict version of it is now the default, see
https://pandas.pydata.org/pdps/0004-consistent-to-datetime-parsing.html. You
can safely remove this argument.
    dataframe['DateTime'] = (pd.to_datetime(df_1['DateTime']),
infer_datetime_format=True))
```

	measuring	1	temperature	1	measuring	1	no	Temp	Count	1	\
DateTime											
2022-12-14 14:30:00	42.267696		31.52000		56.369783			3			
2022-12-14 15:30:00	50.822340		30.56750		69.328964			4			
2022-12-14 16:30:00	67.516902		28.82875		94.063829			4			
2022-12-14 17:30:00	68.069052		27.91125		98.858670			4			
2022-12-14 18:30:00	84.294900		27.16250		118.546877			4			
...			
2023-04-19 16:30:00	50.962218		28.84750		77.422442			4			
2023-04-19 17:30:00	56.130342		27.83625		87.266774			4			
2023-04-19 18:30:00	40.233330		25.86125		80.502529			4			
2023-04-19 19:30:00	51.222342		23.64875		101.722552			4			
2023-04-20 20:30:00	38.675040		22.74000		93.377479			3			
		Tag									
DateTime											
2022-12-14 14:30:00	VALID										
2022-12-14 15:30:00	VALID										
2022-12-14 16:30:00	VALID										
2022-12-14 17:30:00	VALID										
2022-12-14 18:30:00	VALID										
...	...										
2023-04-19 16:30:00	VALID										
2023-04-19 17:30:00	VALID										

```
2023-04-19 18:30:00  VALID  
2023-04-19 19:30:00  VALID  
2023-04-20 20:30:00  VALID
```

[1021 rows x 5 columns]

0.1.3 Sensor 2

```
[ ]: import pandas as pd  
  
directory_path = 'input/'  
file_name = sensor_2_name + '_and_temp_valid_1HR.csv'  
df_2 = pd.read_csv(directory_path + file_name)  
df_2.head()  
  
[ ]:          DateTime  measuring 2  temperature  Hour  measuring 2 no Temp  \\\n0  2022-11-28 11:30:00      53.759778    30.10750    11           11.239952  
1  2022-11-28 12:30:00      53.445666    29.88250    12           11.476654  
2  2022-11-28 13:30:00      54.100884    30.24125    13           11.253630  
3  2022-11-28 14:30:00      53.921742    30.13250    14           11.340715  
4  2022-11-28 15:30:00      53.494746    29.89875    15           11.485953  
  
      Count 2     Tag  
0        4  VALID  
1        4  VALID  
2        4  VALID  
3        4  VALID  
4        4  VALID
```

0.1.4 Create Sensor Dataframe as Pandas Series

```
[ ]: # Remove the first column with the indexes and save data into web dataframe  
dataframe = df_2.drop(df_2.columns[0], axis='columns')  
dataframe['DateTime'] = (pd.to_datetime(df_2['DateTime'],  
                                         infer_datetime_format=True))  
  
# Resample data with 15 mins period and create sensor dataframe  
sensor_2_dataframe = dataframe.sort_values(by='DateTime', ascending=True).  
    .reset_index().drop(columns='index')  
sensor_2_dataframe.index = sensor_2_dataframe['DateTime']  
sensor_2_dataframe = sensor_2_dataframe.drop(columns=['DateTime', 'Hour'])  
sensor_2_dataframe = sensor_2_dataframe.rename(columns={'temperature':  
                                         'temperature 2'})  
sensor_2_dataframe
```

```
/var/folders/wc/_83zcrx913j1dqwg4g90kbhh0000gp/T/ipykernel_6358/3647602612.py:3:  
UserWarning: The argument 'infer_datetime_format' is deprecated and will be  
removed in a future version. A strict version of it is now the default, see
```

<https://pandas.pydata.org/pdeps/0004-consistent-to-datetime-parsing.html>. You can safely remove this argument.

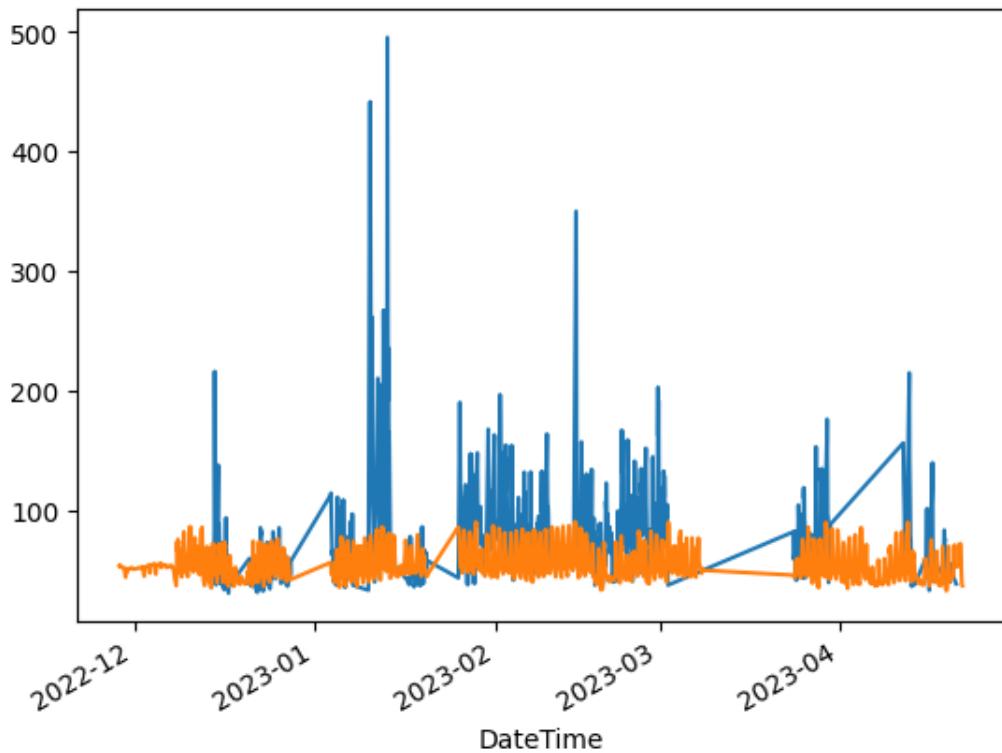
```
dataframe['DateTime'] = (pd.to_datetime(df_2['DateTime']),  
infer_datetime_format=True))
```

```
[ ]:               measuring 2   temperature 2   measuring 2 no Temp   Count 2  \  
DateTime  
2022-11-28 11:30:00    53.759778    30.10750    11.239952    4  
2022-11-28 12:30:00    53.445666    29.88250    11.476654    4  
2022-11-28 13:30:00    54.100884    30.24125    11.253630    4  
2022-11-28 14:30:00    53.921742    30.13250    11.340715    4  
2022-11-28 15:30:00    53.494746    29.89875    11.485953    4  
...  
2023-04-21 17:30:00    50.964672    30.53625    7.395241    4  
2023-04-21 18:30:00    42.360948    27.13750    7.111857    4  
2023-04-21 19:30:00    39.303264    25.32750    8.485160    4  
2023-04-21 20:30:00    37.688532    24.45875    8.997179    4  
2023-04-21 21:30:00    37.048038    23.90625    9.709237    4  
  
Tag  
DateTime  
2022-11-28 11:30:00  VALID  
2022-11-28 12:30:00  VALID  
2022-11-28 13:30:00  VALID  
2022-11-28 14:30:00  VALID  
2022-11-28 15:30:00  VALID  
...  
2023-04-21 17:30:00  VALID  
2023-04-21 18:30:00  VALID  
2023-04-21 19:30:00  VALID  
2023-04-21 20:30:00  VALID  
2023-04-21 21:30:00  VALID  
  
[2603 rows x 5 columns]
```

0.1.5 Plot raw data

```
[ ]: sensor_1_dataframe['measuring 1'].plot()  
sensor_2_dataframe['measuring 2'].plot()
```

```
[ ]: <Axes: xlabel='DateTime'>
```



0.2 Compare with original data

```
[ ]: input_data_directory = 'input/'
reference_data_path = input_data_directory + 'ref_air_quality_data_Vila_Moema.
˓→csv'
reference_column_name = 'Ozônio'
```

0.3 Load reference and sensor data

```
[ ]: import pandas as pd

reference_data = pd.read_csv(reference_data_path)
reference_data['DateTime'] = (pd.to_datetime(reference_data['DateTime']), 
˓→infer_datetime_format=True))
reference_data = reference_data.sort_values(by='DateTime', ascending=True).
˓→reset_index().drop(columns='index')
reference_data.index = reference_data['DateTime']
reference_data = reference_data.drop(columns='DateTime')[reference_column_name]

reference_data
```

/var/folders/wc/_83zcrx913j1dqwg4g90kbhh0000gp/T/ipykernel_6358/2986665073.py:4:

```
UserWarning: The argument 'infer_datetime_format' is deprecated and will be removed in a future version. A strict version of it is now the default, see https://pandas.pydata.org/pdeps/0004-consistent-to-datetime-parsing.html. You can safely remove this argument.
```

```
reference_data['DateTime'] = (pd.to_datetime(reference_data['DateTime'], infer_datetime_format=True))
```

```
[ ]: DateTime
2022-01-01 02:30:00    13.23
2022-01-01 03:30:00    12.07
2022-01-01 04:30:00    13.24
2022-01-01 05:30:00    14.42
2022-01-01 06:30:00    13.30
...
2023-02-08 12:30:00    50.01
2023-02-08 13:30:00    67.43
2023-02-08 14:30:00    72.46
2023-02-08 15:30:00    59.65
2023-02-08 16:30:00    NaN
Name: Ozônio, Length: 9687, dtype: float64
```

0.4 Merge sensor and reference data

```
[ ]: import numpy as np

sensor_data = pd.concat([sensor_1_dataframe, sensor_2_dataframe], join='outer', axis=1)
sensor_data['temperature'] = sensor_data[['temperature 1', 'temperature 2']].apply(lambda df: df[1] if np.isnan(df[0]) else df[0], axis=1)
sensor_data = sensor_data.drop(columns=['temperature 1', 'Count 1', 'Tag', 'temperature 2', 'Count 2'])
sensor_data
```

```
[ ]:               measuring 1   measuring 1 no Temp   measuring 2 \
DateTime
2022-11-28 11:30:00      NaN                 NaN  53.759778
2022-11-28 12:30:00      NaN                 NaN  53.445666
2022-11-28 13:30:00      NaN                 NaN  54.100884
2022-11-28 14:30:00      NaN                 NaN  53.921742
2022-11-28 15:30:00      NaN                 NaN  53.494746
...
...                   ...
2023-04-21 17:30:00      NaN                 NaN  50.964672
2023-04-21 18:30:00      NaN                 NaN  42.360948
2023-04-21 19:30:00      NaN                 NaN  39.303264
2023-04-21 20:30:00      NaN                 NaN  37.688532
2023-04-21 21:30:00      NaN                 NaN  37.048038
```

	measuring 2	no Temp	temperature
DateTime			
2022-11-28 11:30:00	11.239952	30.10750	
2022-11-28 12:30:00	11.476654	29.88250	
2022-11-28 13:30:00	11.253630	30.24125	
2022-11-28 14:30:00	11.340715	30.13250	
2022-11-28 15:30:00	11.485953	29.89875	
...	
2023-04-21 17:30:00	7.395241	30.53625	
2023-04-21 18:30:00	7.111857	27.13750	
2023-04-21 19:30:00	8.485160	25.32750	
2023-04-21 20:30:00	8.997179	24.45875	
2023-04-21 21:30:00	9.709237	23.90625	

[2616 rows x 5 columns]

```
[ ]: sensor_data = pd.concat([sensor_data, reference_data], axis=1, join='inner')
sensor_data = sensor_data.rename(columns={'Ozônio': 'reference'})
sensor_data
```

	measuring 1	measuring 1	no Temp	measuring 2	\
DateTime					
2022-11-28 11:30:00	NaN		NaN	53.759778	
2022-11-28 12:30:00	NaN		NaN	53.445666	
2022-11-28 13:30:00	NaN		NaN	54.100884	
2022-11-28 14:30:00	NaN		NaN	53.921742	
2022-11-28 15:30:00	NaN		NaN	53.494746	
...	
2023-02-08 12:30:00	NaN		NaN	80.255616	
2023-02-08 13:30:00	NaN		NaN	81.669120	
2023-02-08 14:30:00	NaN		NaN	80.498562	
2023-02-08 15:30:00	NaN		NaN	79.146408	
2023-02-08 16:30:00	NaN		NaN	71.617536	
	measuring 2	no Temp	temperature	reference	
DateTime					
2022-11-28 11:30:00	11.239952	30.10750	21.49		
2022-11-28 12:30:00	11.476654	29.88250	NaN		
2022-11-28 13:30:00	11.253630	30.24125	NaN		
2022-11-28 14:30:00	11.340715	30.13250	20.56		
2022-11-28 15:30:00	11.485953	29.89875	21.15		
...	
2023-02-08 12:30:00	15.204469	39.31125	50.01		
2023-02-08 13:30:00	15.351102	39.82875	67.43		
2023-02-08 14:30:00	13.929618	39.93125	72.46		
2023-02-08 15:30:00	12.788609	39.84500	59.65		
2023-02-08 16:30:00	9.874328	37.96000	NaN		

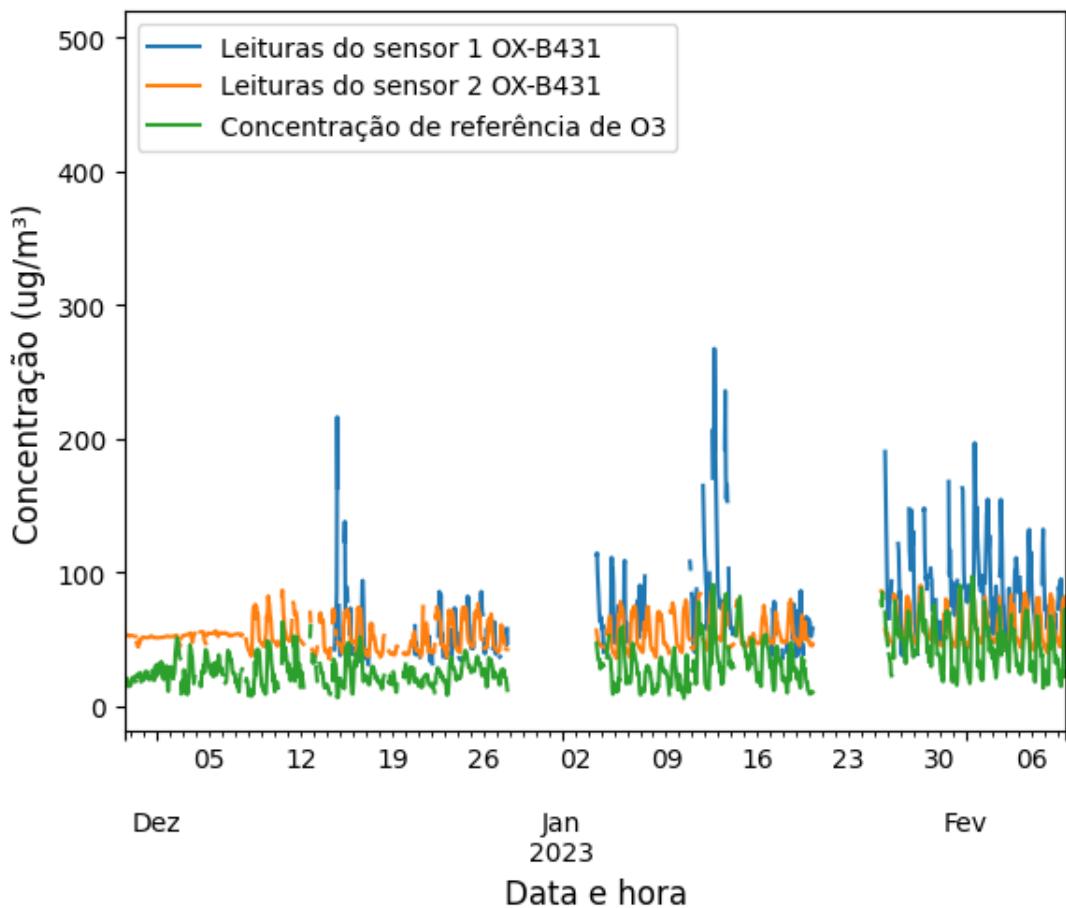
[1345 rows x 6 columns]

0.4.1 Plot reference and sensor data

```
[ ]: import matplotlib.pyplot as plt

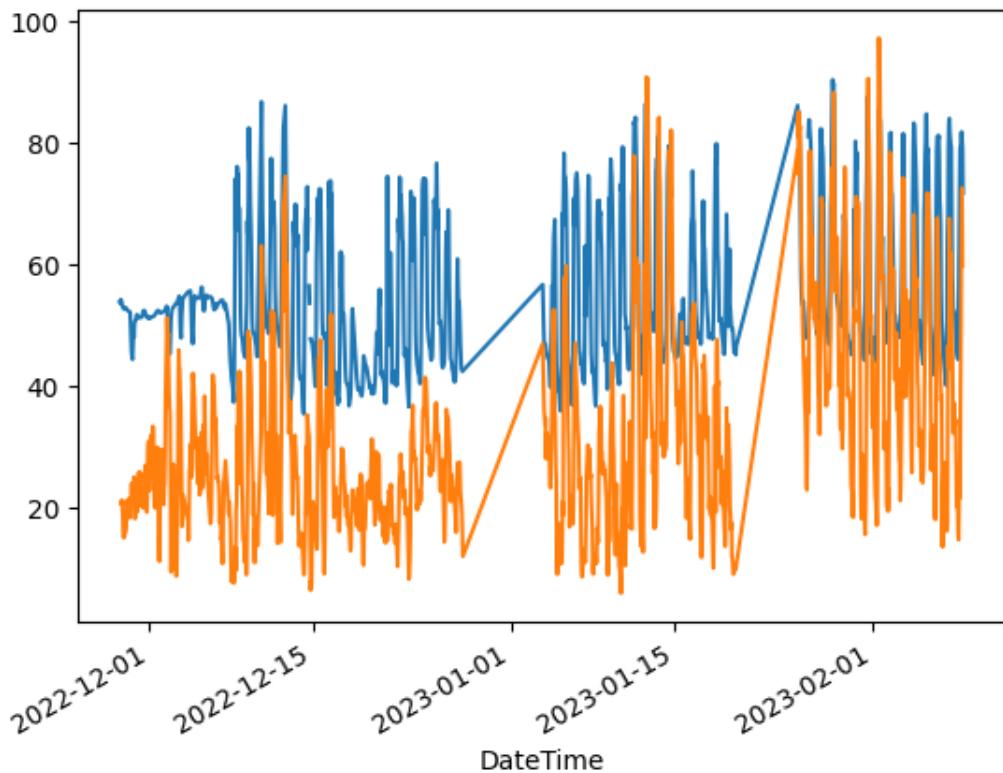
fig, ax = plt.subplots(figsize=(1.3*5,5))
sensor_data['measuring 1'].rename('Leituras do sensor 1 OX-B431').resample('H').
    mean().plot()
sensor_data['measuring 2'].rename('Leituras do sensor 2 OX-B431').resample('H').
    mean().plot()
sensor_data['reference'].rename('Concentração de referência de O3').
    resample('H').mean().plot()
plt.legend()
ax.set_xlabel('Data e hora', fontsize=12)
ax.set_ylabel('Concentração (ug/m3)', fontsize=12)
```

[]: Text(0, 0.5, 'Concentração (ug/m³)')



```
[ ]: sensor_data['measuring 2'].plot()  
sensor_data['reference'].plot()
```

```
[ ]: <Axes: xlabel='DateTime'>
```



0.5 Plot sensor vs. reference

0.5.1 Sensor 1

```
[ ]: from scipy.stats import kendalltau, spearmanr, gaussian_kde  
import matplotlib.pyplot as plt  
import numpy as np  
  
median_reference = sensor_data['reference'].median()  
median_measuring = sensor_data['measuring 1'].median()  
  
fig, ax = plt.subplots(figsize=(1.3*5,5))  
xy = np.vstack([sensor_data['reference'].fillna(value=median_reference),  
    sensor_data['measuring 1'].fillna(value=median_measuring)])
```

```

z = gaussian_kde(xy)(xy)

plt.scatter(sensor_data['reference'], sensor_data['measuring 1'], c=z, cmap='jet', s=15, alpha=.5)

plt.title('Leituras do sensor 1 OX-B431 vs Concentração de referência de O3 \n',
          fontdict={'fontsize':15})
plt.xlabel('Concentração de referência de O3 (ug/m3)',
           fontsize=12)
plt.ylabel('Leituras do sensor OX-B431 (ug/m3)',
           fontsize=12)
coef, p = spearmanr(sensor_data['reference'], sensor_data['measuring 1'],
                      nan_policy='omit')

print('Spearmans correlation coefficient: %.2f' % coef)
# interpret the significance
alpha = 0.05
if p > alpha:
    print('Samples are uncorrelated (fail to reject H0) p=%.2f' % p)
else:
    print('Samples are correlated (reject H0) p=%.2f' % p)

kendall, pken = kendalltau(sensor_data['reference'], sensor_data['measuring 1'],
                            nan_policy='omit')
print('Kendall correlation coefficient: %.2f' % coef)
# interpret the significance
alpha = 0.05
if p > alpha:
    print('Samples are uncorrelated (fail to reject H0) p=%.2f' % p)
else:
    print('Samples are correlated (reject H0) p=%.2f' % p)

cax = plt.axes([0.95, 0.1, 0.05, 0.8])
cbar = plt.colorbar(orientation='vertical', cax=cax, label="Densidade de kernel")
cbar.ax.tick_params(labelsize=11, length=0)

string = 'Coeficiente de Spearman=%.2f, p<0.05' % coef
plt.text(0.3, 0.95, string, horizontalalignment='left', fontsize=12,
         verticalalignment='center', transform=ax.transAxes)
string = 'Coeficiente de Kendall=%.2f, p<0.05' % kendall
plt.text(0.3, 0.90, string, horizontalalignment='left', fontsize=12,
         verticalalignment='center', transform=ax.transAxes)

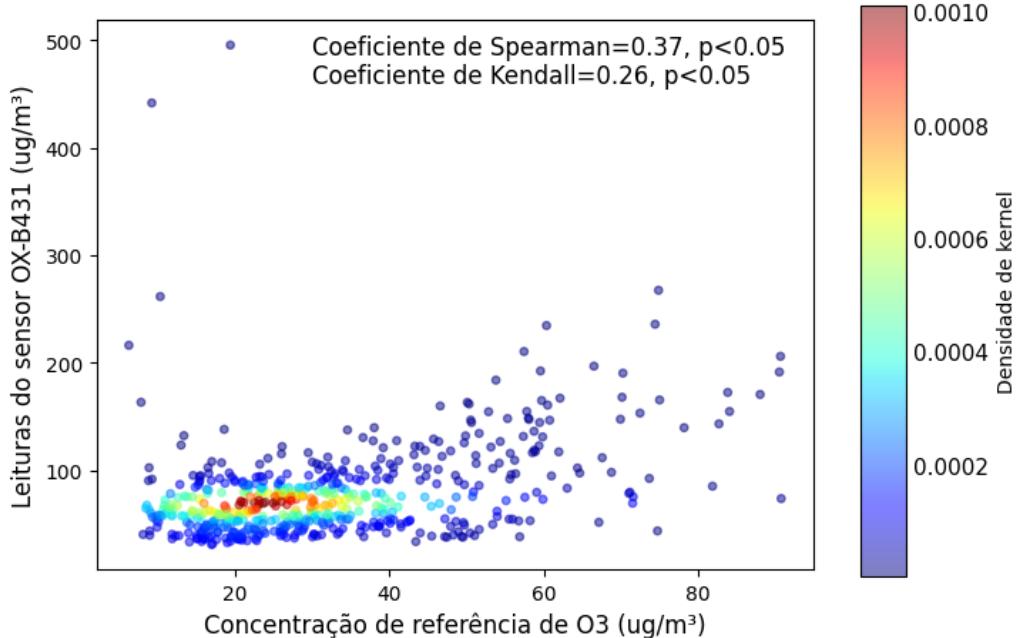
```

Spearmans correlation coefficient: 0.37
 Samples are correlated (reject H0) p=0.00
 Kendall correlation coefficient: 0.37

```
Samples are correlated (reject H0) p=0.00
```

```
[ ]: Text(0.3, 0.9, 'Coeficiente de Kendall=0.26, p<0.05')
```

Leituras do sensor 1 OX-B431 vs Concentração de referência de O3



0.5.2 Sensor 2

```
[ ]: from scipy.stats import kendalltau, spearmanr, gaussian_kde
import matplotlib.pyplot as plt
import numpy as np

median_reference = sensor_data['reference'].median()
median_measuring = sensor_data['measuring 2'].median()

fig, ax = plt.subplots(figsize=(1.3*5,5))
xy = np.vstack([sensor_data['reference'].fillna(value=median_reference),sensor_data['measuring 2'].fillna(value=median_measuring)])
z = gaussian_kde(xy)(xy)

plt.scatter(sensor_data['reference'], sensor_data['measuring 2'], c=z, cmap='jet', s=15, alpha=.5)

plt.title('Leituras do sensor 2 OX-B431 vs Concentração de referência de O3 \n', fontdict={'fontsize':15})
plt.xlabel('Concentração de referência de O3 (ug/m\N{SUPERSCRIPT THREE})',
```

```

    fontsize=12)
plt.ylabel('Leituras do sensor OX-B431 (ug/m\N{SUPERSCRIPT THREE})',
            fontsize=12)
coef, p = spearmanr(sensor_data['reference'], sensor_data['measuring 2'],
                     nan_policy='omit')

print('Spearmans correlation coefficient: %.2f' % coef)
# interpret the significance
alpha = 0.05
if p > alpha:
    print('Samples are uncorrelated (fail to reject H0) p=%.2f' % p)
else:
    print('Samples are correlated (reject H0) p=%.2f' % p)

kendall, pken = kendalltau(sensor_data['reference'], sensor_data['measuring 2'],
                            nan_policy='omit')
print('Kendall correlation coefficient: %.2f' % coef)
# interpret the significance
alpha = 0.05
if p > alpha:
    print('Samples are uncorrelated (fail to reject H0) p=%.2f' % p)
else:
    print('Samples are correlated (reject H0) p=%.2f' % p)

cax = plt.axes([0.95, 0.1, 0.05, 0.8])
cbar = plt.colorbar(orientation='vertical', cax=cax, label="Densidade de kernel")
cbar.ax.tick_params(labelsize=11, length=0)

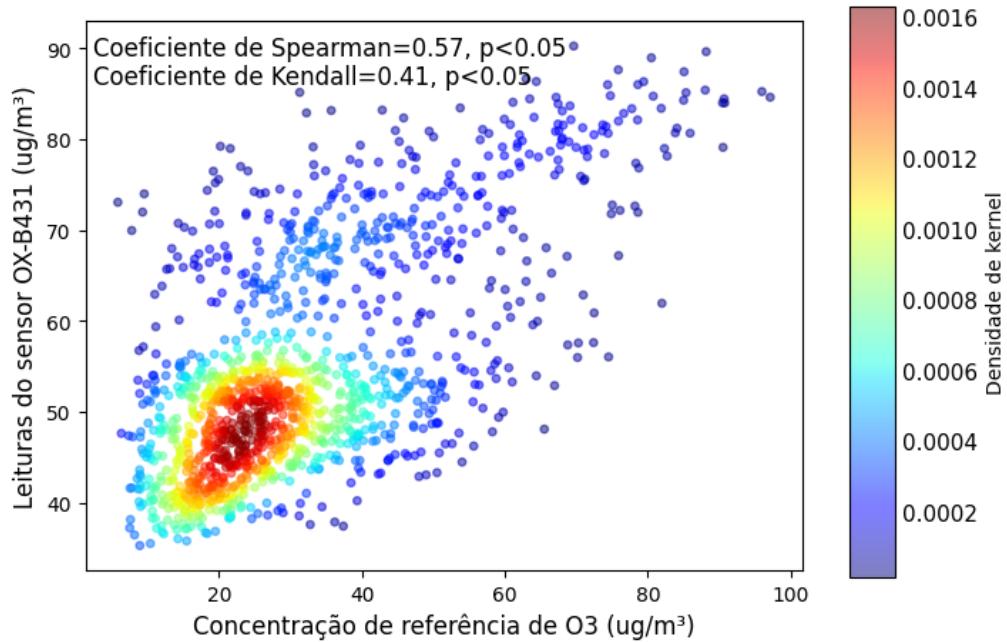
string = 'Coeficiente de Spearman=%.2f, p<0.05' % coef
plt.text(0.01, 0.95, string, horizontalalignment='left', fontsize=12,
         verticalalignment='center', transform=ax.transAxes)
string = 'Coeficiente de Kendall=%.2f, p<0.05' % kendall
plt.text(0.01, 0.90, string, horizontalalignment='left', fontsize=12,
         verticalalignment='center', transform=ax.transAxes)

```

Spearmans correlation coefficient: 0.57
 Samples are correlated (reject H0) p=0.00
 Kendall correlation coefficient: 0.57
 Samples are correlated (reject H0) p=0.00

[]: Text(0.01, 0.9, 'Coeficiente de Kendall=0.41, p<0.05')

Leituras do sensor 2 OX-B431 vs Concentração de referência de O3



0.6 Plot reference vs. temperature

```
[ ]: from scipy.stats import kendalltau, spearmanr, gaussian_kde
import matplotlib.pyplot as plt
import numpy as np

median_reference = sensor_data['reference'].median()
median_temperature = sensor_data['temperature'].median()

fig, ax = plt.subplots(figsize=(1.3*5,5))
xy = np.vstack([sensor_data['temperature'].fillna(value=median_temperature), ↪
    sensor_data['reference'].fillna(value=median_reference)])
z = gaussian_kde(xy)(xy)

plt.scatter(sensor_data['temperature'], sensor_data['reference'], c=z, ↪
    cmap='jet', s=15, alpha=.5)

plt.title('Concentração O3 de referência vs Temperatura\n',
          fontdict={'fontsize':15})
plt.xlabel('Temperatura (C)', fontsize=12)
plt.ylabel('Concentração O3 de referência (ug/m\N{SUPERSCRIPT THREE})', fontsize=12)
```

```

coef, p = spearmanr(sensor_data['temperature'], sensor_data['reference'],
                     nan_policy='omit')

print('Spearmans correlation coefficient: %.2f' % coef)
# interpret the significance
alpha = 0.05
if p > alpha:
    print('Samples are uncorrelated (fail to reject H0) p=%.2f' % p)
else:
    print('Samples are correlated (reject H0) p=%.2f' % p)

kendall, pken = kendalltau(sensor_data['temperature'], sensor_data['reference'],
                            nan_policy='omit')
print('Kendall correlation coefficient: %.2f' % coef)
# interpret the significance
alpha = 0.05
if p > alpha:
    print('Samples are uncorrelated (fail to reject H0) p=%.2f' % p)
else:
    print('Samples are correlated (reject H0) p=%.2f' % p)

cax = plt.axes([0.95, 0.1, 0.05, 0.8])
cbar = plt.colorbar(orientation='vertical', cax=cax, label="Densidade de kernel")
cbar.ax.tick_params(labelsize=11, length=0)

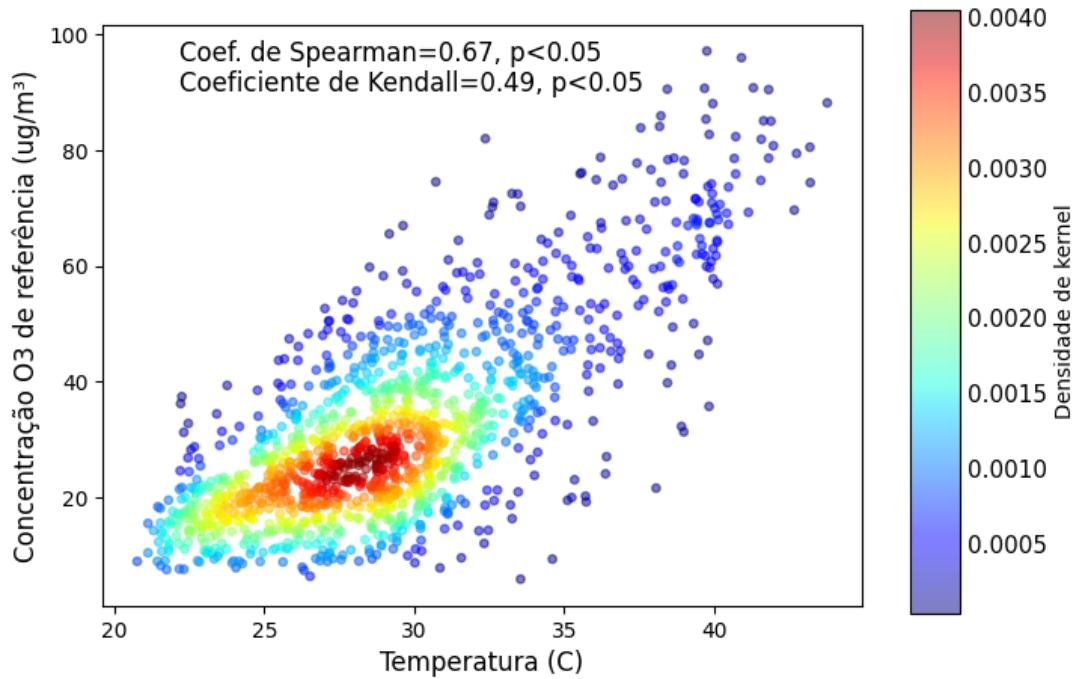
string = 'Coef. de Spearman=%.2f, p<0.05' % coef
plt.text(0.1, 0.95, string, horizontalalignment='left', fontsize=12,
         verticalalignment='center', transform=ax.transAxes)
string = 'Coeficiente de Kendall=%.2f, p<0.05' % kendall
plt.text(0.1, 0.90, string, horizontalalignment='left', fontsize=12,
         verticalalignment='center', transform=ax.transAxes)

```

Spearmans correlation coefficient: 0.67
 Samples are correlated (reject H0) p=0.00
 Kendall correlation coefficient: 0.67
 Samples are correlated (reject H0) p=0.00

[]: Text(0.1, 0.9, 'Coeficiente de Kendall=0.49, p<0.05')

Concentração O3 de referência vs Temperatura



```
[ ]: def plot_box(df):
    bottom, height = 0.1, 0.65
    left, width = bottom, height*1.3
    spacing = 0.005

    rect_ser = [left-width-spacing, bottom, width, height]
    rect_box = [left, bottom, width, height]

    plt.figure(figsize=(1.3*5,5))

    ax_ser = plt.axes(rect_ser)
    ax_ser.tick_params(direction='in', top=True, right=True)
    ax_ser.set_title('Série temporal')
    ax_ser.set_xlabel("Data e hora")
    ax_ser.set_ylabel("Leituras de concentração (ug/m\N{SUPERSCRIPT THREE})")

    ax_box = plt.axes(rect_box)
    ax_box.tick_params(direction='in', labelleft=False)

    lim_max = df['reference'].max()+df['reference'].max()*10/100
    lim_min = df['reference'].min()-df['reference'].min()*10/100
```

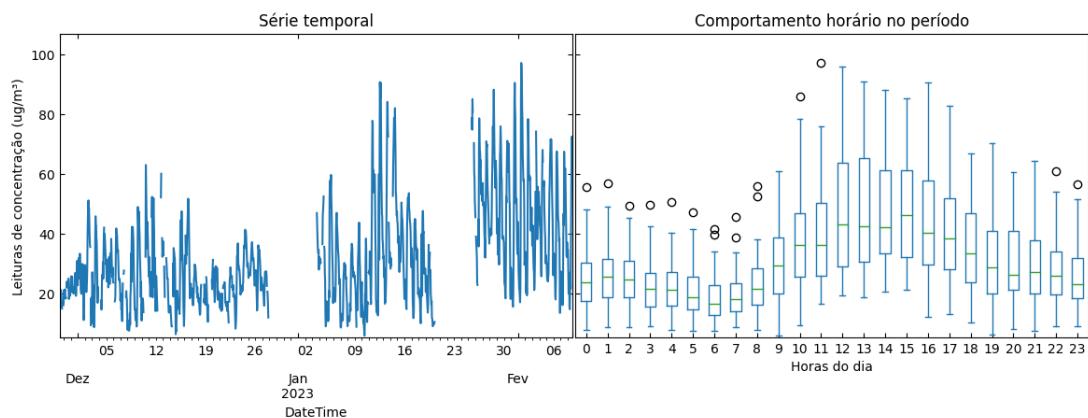
```

df['reference'].plot(ax=ax_ser)
ax_ser.set_ylim(lim_min, lim_max)

df = df.dropna(axis='index', how='all', subset=['Hour'])
df['Hour'] = df['Hour'].astype('int64')
df.pivot(columns='Hour')['reference'].dropna(
    axis='columns', how='all').plot.box(
        ax=ax_box,title='Comportamento horário no período')
ax_box.set_ylim(lim_min, lim_max)
ax_box.set_xlabel("Horas do dia")

valid_dataframe = sensor_data.resample('H').mean()
valid_dataframe['Hour'] = valid_dataframe.index.hour
plot_box(df=valid_dataframe)

```



0.7 Calibrate data

0.7.1 Prepare training and test sets

```
[ ]: from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_validate

reference_median = sensor_data['reference'].median()
sensor_1_median = sensor_data['measuring 1'].median()
sensor_2_median = sensor_data['measuring 2'].median()
temperature_mean = sensor_data['temperature'].mean()
trend_1_median = sensor_data['measuring 1 no Temp'].median()
trend_2_median = sensor_data['measuring 2 no Temp'].median()

variables_names = ['measuring 1', 'measuring 2', 'temperature']

y = sensor_data['reference'].fillna(value=reference_median)
```

```

X = (sensor_data[variables_names].fillna(value={
    'measuring 1': sensor_1_median,
    'measuring 2': sensor_2_median,
    'temperature': temperature_mean}).values.reshape(-1,3))

X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=42)

```

0.7.2 Grid search with different models and variables combinations

```

[ ]: from itertools import combinations

def check_if_list_contains(list1, list2):
    return [element for element in list1 if element in list2]

indexes = []
reference_indexes = [0, 1]
num_variables = len(variables_names)
for num_combinations in list(range(num_variables)):
    contains_reference = False
    index_list = [list(index_tuple) for index_tuple in
    ↪list(combinations(list(range(num_variables)), r=num_combinations+1))]
    for sublist in index_list:
        contains_reference = check_if_list_contains(sublist, reference_indexes)
        if contains_reference:
            indexes.append(sublist)
            contains_reference = False

feature_subsets = { }
for index_list in indexes:
    key = ""
    new_index_list = []
    for index in index_list:
        if len(key) < 1:
            new_index_list.append(index)
            key = key + variables_names[index] + " | "
        elif not ('measuring 1' in key and 'measuring 1' in
        ↪variables_names[index]) and not ('measuring 2' in key and 'measuring 2' in
        ↪variables_names[index]):
            new_index_list.append(index)
            key = key + variables_names[index] + " | "
    feature_subsets[key] = new_index_list

feature_subsets

```

[]: {'measuring 1 | ': [0],
 'measuring 2 | ': [1],
 'measuring 1 | measuring 2 | ': [0, 1],

```
'measuring 1 | temperature | ': [0, 2],
'measuring 2 | temperature | ': [1, 2],
'measuring 1 | measuring 2 | temperature | ': [0, 1, 2]}
```

Function for plotting observations vs. predictions

```
[ ]: import matplotlib.pyplot as plt
from scipy.stats import spearmanr, kendalltau, gaussian_kde
import numpy as np
import os

def plot_predictions_and_observations(X, y, r2, rmse, mae, file_name):
    fig, ax = plt.subplots(figsize=(1.3*5,5))
    xy = np.vstack([X, y])
    z = gaussian_kde(xy)(xy)
    ax.scatter(X, y, c=z,s=15,alpha=.5)
    spear_corr, p_value = spearmanr(y, X)
    spearman_text = ''
    alpha = 0.05
    if p_value > alpha:
        spearman_text = 'Coeficiente de Spearman: {:.2f}'.format(spear_corr) +\
        ', p>0.05'
    else:
        spearman_text = 'Coeficiente de Spearman: {:.2f}'.format(spear_corr) +\
        ', p<0.05'

    kendall_corr, p_value = kendalltau(y, X)
    alpha = 0.05
    kendall_text = ''
    if p_value > alpha:
        kendall_text = 'Coeficiente de Kendall: {:.2f}'.format(kendall_corr) +\
        ', p>0.05'
    else:
        kendall_text = 'Coeficiente de Kendall: {:.2f}'.format(kendall_corr) +\
        ', p<0.05'

    plt.text(0.02, 0.95, spearman_text, ha='left', va='center', transform=plt.\
    gca().transAxes, fontsize=12)
    plt.text(0.02, 0.90, kendall_text, ha='left', va='center', transform=plt.\
    gca().transAxes, fontsize=12)
    r2_text = 'R2 = {:.2f} ± {:.2f}'.format(r2.mean(), r2.\
    std())
    rmse_text = 'RMSE = {:.2f} ± {:.2f}'.format(rmse.mean(), rmse.std())
    mae_text = 'MAE = {:.2f} ± {:.2f}'.format(mae.mean(), mae.std())
    plt.text(0.02, 0.85, r2_text, ha='left', va='center', transform=plt.gca().\
    transAxes, fontsize=12)
```

```

plt.text(0.02, 0.80, rmse_text, ha='left', va='center', transform=plt.gca() .
        transAxes, fontsize=12)
plt.text(0.02, 0.75, mae_text, ha='left', va='center', transform=plt.gca() .
        transAxes, fontsize=12)

ax.set_xlim([np.min([y,X]),np.max([y,X])])
ax.set_ylim([np.min([y,X]),np.max([y,X])])
ax.set_aspect('equal')

ax.plot([xy.min(), xy.max()], [xy.min(), xy.max()], 'k-', lw=1,dashes=[2,_
    2])
ax.fill_between(np.linspace(xy.min(), xy.max(),y.shape[0]),
                np.linspace(xy.min(), xy.max(),y.shape[0])*0.5,
                alpha=0.2,facecolor='gray',edgecolor=None)
ax.fill_between(np.linspace(xy.min(),xy.max(),y.shape[0]),
                np.linspace(xy.max(),xy.max(),y.shape[0]),
                np.linspace(xy.min(),xy.max(),y.shape[0])*2,
                alpha=0.2,facecolor='gray',edgecolor=None)

ax.set_xlabel('Concentração de O3 observada (ug/m\N{SUPERSCRIPT THREE})',_
    fontsize=12)
ax.set_ylabel('Concentração de O3 inferida (ug/m\N{SUPERSCRIPT_
    THREE})',fontsize=12)

if not os.path.exists('images/'):
    os.makedirs('images/')

plt.savefig('images/' + '03_UNI_' + file_name + '.png')

```

```

[ ]: from sklearn.linear_model import LinearRegression
from sklearn.neural_network import MLPRegressor
from sklearn.neighbors import KNeighborsRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import GridSearchCV
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import Pipeline
from sklearn.metrics import mean_squared_error
import numpy as np

models = {
    'MLP Regression': (
        ('mlp_regressor', MLPRegressor(solver="lbfgs", max_iter=1000,_
            random_state=42)), {
            'mlp_regressor__hidden_layer_sizes': [
                (4,), (10,), (50,), (100,), (200,),
                (4,4), (4,10), (4,50), (4,100), (4,200),
                (10,4), (10,10), (10,50), (10,100),(10,200),

```

```

        (50,4), (50,10), (50,50), (50,100),(50,200),
        (100,4),(100,10),(100,50),(100,100),(100,200),
        (200,4),(200,10),(200,50),(200,100),(200,200)],
    'mlp_regressor__alpha': [0.0001, 0.001, 0.01, 0.1, 1, 10]
}
),
'Multilinear Regression': (
    ('linear_regressor', LinearRegression()), {}
),
'KNN Regression': (
    ('knn_regressor', KNeighborsRegressor()), {
        'knn_regressor__n_neighbors': [3, 5, 7, 9, 11, 13, 15, 17, 20],
        'knn_regressor__weights': ['uniform', 'distance'],
        'knn_regressor__p': [1, 2] # 1 for Manhattan distance, 2 for
        ↪Euclidean distance'
    }
),
'Random Forests Regression': (
    ('random_forest_regressor', RandomForestRegressor()), {
        'random_forest_regressor__n_estimators': [50, 100, 150],
        'random_forest_regressor__max_depth': [None, 10, 20, 30],
        'random_forest_regressor__min_samples_split': [2, 5, 10],
        'random_forest_regressor__min_samples_leaf': [1, 2, 4]
    }
)
}

# Perform grid search for each feature subset
results = {}
rmse_by_features = {}
r2_by_features = {}
mae_by_features = {}
for features_set, subset in feature_subsets.items():
    X_subset = X[:, subset]
    X_train_subset = X_train[:, subset]
    X_test_subset = X_test[:, subset]

    model_results = {}
    model_rmse = {}
    model_r2 = {}
    model_mae = {}
    for model_name, (model, param_grid) in models.items():
        print(f"Grid search for features: {features_set} with model: {model_name}...")
        pipeline = Pipeline([
            ('scaler', StandardScaler()),

```

```

        model
    ]))

# Perform grid search with cross-validation
grid_search = GridSearchCV(pipeline, param_grid, cv=3, u
scoring='neg_root_mean_squared_error', n_jobs=-1)
grid_search.fit(X_train_subset, y_train)

# Print the best parameters and best score
best_params = grid_search.best_params_

# Evaluate the best model on the test set
best_model = grid_search.best_estimator_
cross_validation = cross_validate(best_model, X_subset, y, cv=3, u
scoring=['r2', 'neg_root_mean_squared_error', 'neg_mean_absolute_error'])
y_pred = best_model.predict(X_test_subset)

# Evaluate the model
r2 = cross_validation['test_r2']
rmse = cross_validation['test_neg_root_mean_squared_error']
mae = cross_validation['test_neg_mean_absolute_error']

plot_predictions_and_observations(y_test, y_pred, r2=r2, rmse=rmse, u
mae=mae, file_name=model_name+features_set)

model_results[model_name] = {
    'Best Model': best_model,
    'Best Parameters': best_params,
    'Test R2': r2,
    'Test RMSE': rmse,
    'Test MAE': mae
}
model_rmse[model_name] = {
    'Mean': rmse.mean(),
    'Std': rmse.std()
}
model_r2[model_name] = {
    'Mean': r2.mean(),
    'Std': r2.std()
}
model_mae[model_name] = {
    'Mean': mae.mean(),
    'Std': mae.std()
}

results[features_set] = model_results
rmse_by_features[features_set] = model_rmse

```

```

r2_by_features[features_set] = model_r2
mae_by_features[features_set] = model_mae

for feature_set, models in results.items():
    for model_name, result in models.items():
        print(f"\nResults for features: {feature_set} with model: {model_name}:")
        print(f"Best Parameters: {result['Best Parameters']} ")
        print(f"Test RMSE: {result['Test RMSE'].mean()} +/- {result['Test RMSE'].std()}")
        print(f"Test R2: {result['Test R2'].mean()} +/- {result['Test R2'].std()}")
        print(f"Test MAE: {result['Test MAE'].mean()} +/- {result['Test MAE'].std()}")

```

1 Save Results

```

[ ]: output_directory_path = 'output/'
rmse_file_name = output_directory_path + sensor_name + '_rmse.csv'
r2_file_name = output_directory_path + sensor_name + '_r2.csv'
mae_file_name = output_directory_path + sensor_name + '_mae.csv'
results_file_name = output_directory_path + sensor_name + '_results.csv'

pd.DataFrame(rmse_by_features).transpose().to_csv(rmse_file_name)
pd.DataFrame(r2_by_features).transpose().to_csv(r2_file_name)
pd.DataFrame(mae_by_features).transpose().to_csv(mae_file_name)
pd.DataFrame(results).transpose().to_csv(results_file_name)

```

1.1 Plot Results

```

[ ]: import matplotlib.pyplot as plt
import numpy as np

def plot_metrics(features, r2_list, r2_error_list, rmse_list, rmse_error_list, mae_list, mae_error_list):
    bottom, height = 0.1, 0.65
    left, width = bottom, height*1.3
    spacing = 0.03

    rect_r2 = [left-width-spacing, bottom, width, height]
    rect_rmse = [left, bottom, width, height]
    rect_mae = [left + width + spacing, bottom, height/1.3, height]

    plt.figure(figsize=(1.3*5,8))

    ax_r2 = plt.axes(rect_r2)

```

```

ax_r2.tick_params(direction='in', top=True, right=True, labelsize=14)
ax_r2.set_title('R2')

ax_rmse = plt.axes(rect_rmse)
ax_rmse.tick_params(direction='in', labelleft=False, labelsize=14)
ax_rmse.set_title('RMSE')

ax_mae = plt.axes(rect_mae)
ax_mae.tick_params(direction='in', labelleft=False, labelsize=14)
ax_mae.set_title('MAE')

y_pos = np.arange(len(features))

ax_r2.banh(y_pos, r2_list, xerr=r2_error_list, align='center')
min_r2 = r2_list.min() - r2_error_list.max()
ax_r2.set_xlim([min_r2 - 0.05, 1.0 + 0.05])
ax_r2.set_yticks(y_pos, labels=features, fontsize=14)
ax_r2.invert_yaxis() # labels read top-to-bottom
ax_r2.set_xlabel('R2', fontsize=14)

ax_rmse.banh(y_pos, rmse_list, xerr=rmse_error_list, align='center')
max_rmse = rmse_list.max() + rmse_error_list.max()
min_rmse = rmse_list.min() - rmse_error_list.max()
if max_rmse <= 0: max_rmse = -min_rmse
ax_rmse.set_xlim([min_rmse - 0.05, max_rmse + 0.05])
ax_rmse.set_yticks(y_pos, labels=features, fontsize=14)
ax_rmse.invert_yaxis() # labels read top-to-bottom
ax_rmse.set_xlabel('RMSE', fontsize=14)

ax_mae.banh(y_pos, mae_list, xerr=mae_error_list, align='center')
max_mae = mae_list.max() + mae_error_list.max()
min_mae = mae_list.min() - mae_error_list.max()
if max_mae <= 0: max_mae = -min_mae
ax_mae.set_xlim([min_mae - 0.05, max_mae + 0.05])
ax_mae.set_yticks(y_pos, labels=features, fontsize=14)
ax_mae.invert_yaxis() # labels read top-to-bottom
ax_mae.set_xlabel('MAE', fontsize=14)

```

```
[ ]: mean_r2_by_features_dataframe = pd.DataFrame()
std_r2_by_features_dataframe = pd.DataFrame()

mean_rmse_by_features_dataframe = pd.DataFrame()
std_rmse_by_features_dataframe = pd.DataFrame()

mean_mae_by_features_dataframe = pd.DataFrame()
std_mae_by_features_dataframe = pd.DataFrame()
```

```

for key in list(feature_subsets.keys()):
    feature_dict = r2_by_features[key]
    for model in list(feature_dict.keys()):
        colum_name = key.replace('measuring', '03')
        colum_name = colum_name.replace(' |', ',')
        colum_name += f': {model[:-11]}'
        mean_r2_by_features_dataframe[colum_name] = [
            [feature_dict[model]['Mean']]]
        std_r2_by_features_dataframe[colum_name] = [feature_dict[model]['Std']]

for key in list(feature_subsets.keys()):
    feature_dict = rmse_by_features[key]
    for model in list(feature_dict.keys()):
        colum_name = key.replace('measuring', '03')
        colum_name = colum_name.replace(' |', ',')
        colum_name += f': {model[:-11]}'
        mean_rmse_by_features_dataframe[colum_name] = [
            [feature_dict[model]['Mean']]]
        std_rmse_by_features_dataframe[colum_name] = [
            [feature_dict[model]['Std']]]

for key in list(feature_subsets.keys()):
    feature_dict = mae_by_features[key]
    for model in list(feature_dict.keys()):
        colum_name = key.replace('measuring', '03')
        colum_name = colum_name.replace(' |', ',')
        colum_name += f': {model[:-11]}'
        mean_mae_by_features_dataframe[colum_name] = [
            [feature_dict[model]['Mean']]]
        std_mae_by_features_dataframe[colum_name] = [feature_dict[model]['Std']]

```

```

[ ]: r2_sorted_dataframe = (mean_r2_by_features_dataframe.
    sort_values(by=mean_r2_by_features_dataframe.index[0], axis=1,
    ascending=False))
features = r2_sorted_dataframe.columns

mean_r2 = r2_sorted_dataframe.values.flatten()
error_r2 = std_r2_by_features_dataframe[r2_sorted_dataframe.columns].values.
    flatten()

mean_rmse = mean_rmse_by_features_dataframe[r2_sorted_dataframe.columns].values.
    flatten()
error_rmse = std_rmse_by_features_dataframe[r2_sorted_dataframe.columns].values.
    flatten()

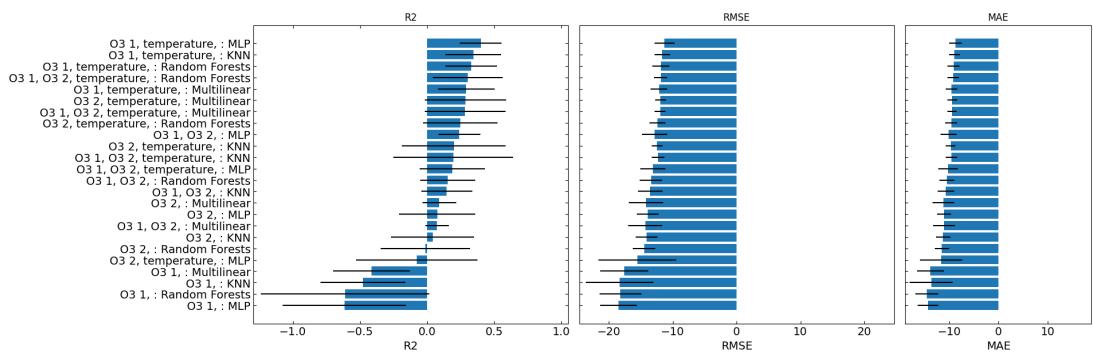
```

```

mean_mae = mean_mae_by_features_dataframe[r2_sorted_dataframe.columns].values.
    ↪flatten()
error_mae = std_mae_by_features_dataframe[r2_sorted_dataframe.columns].values.
    ↪flatten()

plot_metrics(features, r2_list=mean_r2, r2_error_list=error_r2, ↪
    ↪rmse_list=mean_rmse,
    ↪rmse_error_list=error_rmse, mae_list=mean_mae, ↪
    ↪mae_error_list=error_mae)

```



D.2 METODOLOGIA CONSIDERANDO AS LEITURAS DE TODOS OS SENORES E A TEMPERATURA

O3_reference_and_all_sensor_data_with_temp

February 14, 2024

0.0.1 Sensor Constants

```
[ ]: sensor_name = 'o3_all_sensors'  
sensor_co_name = 'alpha_co_conc'  
sensor_no2_name = 'alpha_no2_conc'  
sensor_o3_1_name = 'alpha_o3_1_conc'  
sensor_o3_2_name = 'alpha_o3_2_conc'  
sensor_so2_1_name = 'alpha_so2_1_conc'  
sensor_so2_2_name = 'alpha_so2_2_conc'  
sensor_pm_10_name = 'alpha_pm_10_conc'
```

0.1 Upload Data from File

0.1.1 CO Sensor

```
[ ]: import pandas as pd  
  
directory_path = 'input/'  
file_name = sensor_co_name + '_and_temp_valid_1HR.csv'  
df_co = pd.read_csv(directory_path + file_name)  
df_co.head()
```

```
[ ]:          DateTime  measuring  Hour  temperature  measuring no Temp  \  
0  2022-11-27 17:30:00    0.095855    17    29.78500      0.159694  
1  2022-11-27 18:30:00    0.091372    18    30.13125      0.153195  
2  2022-11-27 19:30:00    0.088210    19    30.09375      0.150251  
3  2022-11-27 20:30:00    0.087858    20    30.03750      0.150226  
4  2022-11-27 21:30:00    0.090610    21    29.96875      0.153379  
  
   Count    Tag  
0     3  VALID  
1     4  VALID  
2     4  VALID  
3     4  VALID  
4     4  VALID
```

0.1.2 Create Sensor Dataframe as Pandas Series

```
[ ]: # Remove the first column with the indexes and save data into web dataframe
dataframe = df_co.drop(df_co.columns[0], axis='columns')
dataframe['DateTime'] = (pd.to_datetime(df_co['DateTime']),  
    ↪infer_datetime_format=True))

# Resample data with 15 mins period and create sensor dataframe
sensor_co_dataframe = dataframe.sort_values(by='DateTime', ascending=True).  
    ↪reset_index().drop(columns='index')
sensor_co_dataframe.index = sensor_co_dataframe['DateTime']
sensor_co_dataframe = sensor_co_dataframe.drop(columns=['DateTime', 'Hour',  
    ↪'Count', 'Tag'])
sensor_co_dataframe = sensor_co_dataframe.rename(columns={'measuring':  
    ↪'measuring CO', 'measuring no Temp': 'measuring no Temp CO', 'temperature':  
    ↪'temperature CO'})
sensor_co_dataframe
```

/var/folders/wc/_83zcrx913j1dqwg4g90kbhh0000gp/T/ipykernel_8846/554760737.py:3:
UserWarning: The argument 'infer_datetime_format' is deprecated and will be
removed in a future version. A strict version of it is now the default, see
<https://pandas.pydata.org/pdeps/0004-consistent-to-datetime-parsing.html>. You
can safely remove this argument.
 dataframe['DateTime'] = (pd.to_datetime(df_co['DateTime'],
 infer_datetime_format=True))

```
[ ]:               measuring CO  temperature CO  measuring no Temp CO
DateTime
2022-11-27 17:30:00      0.095855     29.78500      0.159694
2022-11-27 18:30:00      0.091372     30.13125      0.153195
2022-11-27 19:30:00      0.088210     30.09375      0.150251
2022-11-27 20:30:00      0.087858     30.03750      0.150226
2022-11-27 21:30:00      0.090610     29.96875      0.153379
...
2023-02-04 03:30:00      ...          27.14750      0.109628
2023-02-04 04:30:00      0.028936     27.26000      0.107478
2023-02-04 05:30:00      0.033254     27.31750      0.111460
2023-02-04 06:30:00      0.044496     27.24875      0.123103
2023-02-04 07:30:00      0.050885     27.93250      0.125510
```

[1010 rows x 3 columns]

0.1.3 NO2 Sensor

```
[ ]: import pandas as pd

directory_path = 'input/'
file_name = sensor_no2_name + '_and_temp_valid_1HR.csv'
```

```
df_no2 = pd.read_csv(directory_path + file_name)
df_no2.head()
```

```
[ ]:          DateTime   measuring   Hour   temperature   Count   Tag
0  2022-11-27 17:30:00  182.373362    17     29.78500    3  VALID
1  2022-11-27 18:30:00  188.127215    18     30.13125    4  VALID
2  2022-11-27 19:30:00  175.393318    19     30.09375    4  VALID
3  2022-11-27 20:30:00  185.269497    20     30.03750    4  VALID
4  2022-11-27 21:30:00  179.436459    21     29.96875    4  VALID
```

0.1.4 Create Sensor Dataframe as Pandas Series

```
[ ]: # Remove the first column with the indexes and save data into web dataframe
dataframe = df_no2.drop(df_no2.columns[0], axis='columns')
dataframe['DateTime'] = (pd.to_datetime(df_no2['DateTime']),
                        infer_datetime_format=True))

# Resample data with 15 mins period and create sensor dataframe
sensor_no2_dataframe = dataframe.sort_values(by='DateTime', ascending=True).
    reset_index().drop(columns='index')
sensor_no2_dataframe.index = sensor_no2_dataframe['DateTime']
sensor_no2_dataframe = sensor_no2_dataframe.drop(columns=['DateTime', 'Hour',
    'Count', 'Tag'])
sensor_no2_dataframe = sensor_no2_dataframe.rename(columns={'measuring': 'measuring NO2',
    'measuring no Temp': 'measuring no Temp NO2', 'temperature': 'temperature NO2'})
sensor_no2_dataframe
```

```
/var/folders/wc/_83zcrx913j1dqwg4g90kbhh0000gp/T/ipykernel_8846/4266076299.py:3:
UserWarning: The argument 'infer_datetime_format' is deprecated and will be
removed in a future version. A strict version of it is now the default, see
https://pandas.pydata.org/pdeps/0004-consistent-to-datetime-parsing.html. You
can safely remove this argument.
    dataframe['DateTime'] = (pd.to_datetime(df_no2['DateTime'],
    infer_datetime_format=True))
```

```
[ ]:          measuring NO2   temperature NO2
DateTime
2022-11-27 17:30:00      182.373362      29.785000
2022-11-27 18:30:00      188.127215      30.131250
2022-11-27 19:30:00      175.393318      30.093750
2022-11-27 20:30:00      185.269497      30.037500
2022-11-27 21:30:00      179.436459      29.968750
...
2022-12-23 16:30:00      34.320839      31.335000
2022-12-23 18:30:00      51.619244      29.111667
2022-12-23 19:30:00      153.853416      26.977500
```

```
2022-12-23 20:30:00      164.976172      25.810000
2022-12-23 21:30:00      124.629429      25.195000
```

```
[278 rows x 2 columns]
```

0.1.5 O3

Sensor 1

```
[ ]: import pandas as pd

directory_path = 'input/'
file_name = sensor_o3_1_name + '_and_temp_valid_1HR.csv'
df_o3_1 = pd.read_csv(directory_path + file_name)
df_o3_1.head()
```



```
[ ]:          DateTime  measuring 1  temperature  Hour  measuring 1 no Temp \
0  2022-12-14 14:30:00      42.267696     31.52000    14              56.369783
1  2022-12-14 15:30:00      50.822340     30.56750    15              69.328964
2  2022-12-14 16:30:00      67.516902     28.82875    16             94.063829
3  2022-12-14 17:30:00      68.069052     27.91125    17             98.858670
4  2022-12-14 18:30:00      84.294900     27.16250    18            118.546877

      Count 1    Tag
0        3  VALID
1        4  VALID
2        4  VALID
3        4  VALID
4        4  VALID
```

0.1.6 Create Sensor Dataframe as Pandas Series

```
[ ]: # Remove the first column with the indexes and save data into web dataframe
dataframe = df_o3_1.drop(df_o3_1.columns[0], axis='columns')
dataframe['DateTime'] = (pd.to_datetime(df_o3_1['DateTime'],
                                         infer_datetime_format=True))

# Resample data with 15 mins period and create sensor dataframe
sensor_o3_1_dataframe = dataframe.sort_values(by='DateTime', ascending=True).
    reset_index().drop(columns='index')
sensor_o3_1_dataframe.index = sensor_o3_1_dataframe['DateTime']
sensor_o3_1_dataframe = sensor_o3_1_dataframe.drop(columns=['DateTime', 'Hour',
    'Count 1', 'Tag'])
sensor_o3_1_dataframe = sensor_o3_1_dataframe.rename(columns={'measuring 1': 'measuring O3 1',
    'measuring 1 no Temp': 'measuring no Temp O3 1',
    'temperature': 'temperature O3 1'})
sensor_o3_1_dataframe
```

```
/var/folders/wc/_83zcrx913j1dqwg4g90kbhh0000gp/T/ipykernel_8846/4189083599.py:3:  
UserWarning: The argument 'infer_datetime_format' is deprecated and will be  
removed in a future version. A strict version of it is now the default, see  
https://pandas.pydata.org/pdps/0004-consistent-to-datetime-parsing.html. You  
can safely remove this argument.  
    dataframe['DateTime'] = (pd.to_datetime(df_o3_1['DateTime'],  
    infer_datetime_format=True))
```

```
[ ]:               measuring 03 1  temperature 03 1  measuring no Temp 03 1  
DateTime  
2022-12-14 14:30:00      42.267696      31.52000      56.369783  
2022-12-14 15:30:00      50.822340      30.56750      69.328964  
2022-12-14 16:30:00      67.516902      28.82875      94.063829  
2022-12-14 17:30:00      68.069052      27.91125      98.858670  
2022-12-14 18:30:00      84.294900      27.16250     118.546877  
...                   ...          ...          ...  
2023-04-19 16:30:00      50.962218      28.84750      77.422442  
2023-04-19 17:30:00      56.130342      27.83625      87.266774  
2023-04-19 18:30:00      40.233330      25.86125      80.502529  
2023-04-19 19:30:00      51.222342      23.64875     101.722552  
2023-04-20 20:30:00      38.675040      22.74000     93.377479  
  
[1021 rows x 3 columns]
```

0.1.7 Sensor 2

```
[ ]: import pandas as pd  
  
directory_path = 'input/'  
file_name = sensor_o3_2_name + '_and_temp_valid_1HR.csv'  
df_o3_2 = pd.read_csv(directory_path + file_name)  
df_o3_2.head()  
  
[ ]:               DateTime  measuring 2  temperature  Hour  measuring 2 no Temp  \  
0  2022-11-28 11:30:00      53.759778      30.10750     11      11.239952  
1  2022-11-28 12:30:00      53.445666      29.88250     12      11.476654  
2  2022-11-28 13:30:00      54.100884      30.24125     13      11.253630  
3  2022-11-28 14:30:00      53.921742      30.13250     14      11.340715  
4  2022-11-28 15:30:00      53.494746      29.89875     15      11.485953  
  
Count 2   Tag  
0        4  VALID  
1        4  VALID  
2        4  VALID  
3        4  VALID  
4        4  VALID
```

0.1.8 Create Sensor Dataframe as Pandas Series

```
[ ]: # Remove the first column with the indexes and save data into web dataframe
dataframe = df_o3_2.drop(df_o3_2.columns[0], axis='columns')
dataframe['DateTime'] = (pd.to_datetime(df_o3_2['DateTime']),  
    ↪infer_datetime_format=True))

# Resample data with 15 mins period and create sensor dataframe
sensor_o3_2_dataframe = dataframe.sort_values(by='DateTime', ascending=True).  
    ↪reset_index().drop(columns='index')
sensor_o3_2_dataframe.index = sensor_o3_2_dataframe['DateTime']
sensor_o3_2_dataframe = sensor_o3_2_dataframe.drop(columns=['DateTime', 'Hour',  
    ↪'Count 2', 'Tag'])
sensor_o3_2_dataframe = sensor_o3_2_dataframe.rename(columns={'measuring 2':  
    ↪'measuring 03 2', 'measuring 2 no Temp': 'measuring no Temp 03 2',  
        ↪'temperature':  
            ↪'temperature 03 2'})
sensor_o3_2_dataframe
```

```
/var/folders/wc/_83zcrx913j1dqwg4g90kbhh0000gp/T/ipykernel_8846/2290086274.py:3:  
UserWarning: The argument 'infer_datetime_format' is deprecated and will be  
removed in a future version. A strict version of it is now the default, see  
https://pandas.pydata.org/pdps/0004-consistent-to-datetime-parsing.html. You  
can safely remove this argument.
```

```
    dataframe['DateTime'] = (pd.to_datetime(df_o3_2['DateTime'],  
    infer_datetime_format=True))
```

```
[ ]:               measuring 03 2   temperature 03 2   measuring no Temp 03 2
DateTime
2022-11-28 11:30:00      53.759778      30.10750      11.239952
2022-11-28 12:30:00      53.445666      29.88250      11.476654
2022-11-28 13:30:00      54.100884      30.24125      11.253630
2022-11-28 14:30:00      53.921742      30.13250      11.340715
2022-11-28 15:30:00      53.494746      29.89875      11.485953
...
2023-04-21 17:30:00      50.964672      30.53625      7.395241
2023-04-21 18:30:00      42.360948      27.13750      7.111857
2023-04-21 19:30:00      39.303264      25.32750      8.485160
2023-04-21 20:30:00      37.688532      24.45875      8.997179
2023-04-21 21:30:00      37.048038      23.90625      9.709237
```

[2603 rows x 3 columns]

0.1.9 SO2

Sensor 1

```
[ ]: import pandas as pd
```

```

directory_path = 'input/'
file_name = sensor_so2_1_name + '_and_temp_valid_1HR.csv'
df_so2_1 = pd.read_csv(directory_path + file_name)
df_so2_1.head()

```

```

[ ]:          DateTime measuring 1 temperature Hour measuring 1 no Temp \
0 2022-12-07 19:30:00 2334.044792    27.39125   19      2603.851233
1 2022-12-08 13:30:00 408.498126     34.00500   13      1809.151682
2 2022-12-10 08:30:00 1925.520463    32.23500    8      3023.531889
3 2022-12-13 20:30:00 3721.797934    24.45875   20      3490.193048
4 2022-12-13 21:30:00 4488.438582    23.87375   21      4156.807906

Count 1 Tag
0      4 VALID
1      4 VALID
2      4 VALID
3      4 VALID
4      4 VALID

```

0.1.10 Create Sensor Dataframe as Pandas Series

```

[ ]: # Remove the first column with the indexes and save data into web dataframe
dataframe = df_so2_1.drop(df_so2_1.columns[0], axis='columns')
dataframe['DateTime'] = (pd.to_datetime(df_so2_1['DateTime'], \
                                         infer_datetime_format=True))

# Resample data with 15 mins period and create sensor dataframe
sensor_so2_1_dataframe = dataframe.sort_values(by='DateTime', ascending=True). \
    reset_index().drop(columns='index')
sensor_so2_1_dataframe.index = sensor_so2_1_dataframe['DateTime']
sensor_so2_1_dataframe = sensor_so2_1_dataframe.drop(columns=['DateTime', \
    'Hour', 'Count 1', 'Tag'])
sensor_so2_1_dataframe = sensor_so2_1_dataframe.rename(columns={'measuring 1': \
    'measuring S02 1', 'measuring 1 no Temp': 'measuring no Temp S02 1', \
    'temperature': \
    'temperature S02 1'})
sensor_so2_1_dataframe

```

```

/var/folders/wc/_83zcrx913j1dqwg4g90kbhh0000gp/T/ipykernel_8846/3688823847.py:3:
UserWarning: The argument 'infer_datetime_format' is deprecated and will be
removed in a future version. A strict version of it is now the default, see
https://pandas.pydata.org/pdeps/0004-consistent-to-datetime-parsing.html. You
can safely remove this argument.
dataframe['DateTime'] = (pd.to_datetime(df_so2_1['DateTime'],
                                         infer_datetime_format=True))

```

```
[ ]:               measuring S02 1  temperature S02 1 \
DateTime
2022-12-07 19:30:00      2334.044792      27.39125
2022-12-08 13:30:00      408.498126      34.00500
2022-12-10 08:30:00     1925.520463      32.23500
2022-12-13 20:30:00     3721.797934      24.45875
2022-12-13 21:30:00     4488.438582      23.87375
...
          ...
2023-04-16 01:30:00     3447.282268      26.69000
2023-04-16 02:30:00     3646.048354      26.38125
2023-04-16 03:30:00     4043.577251      25.72250
2023-04-16 09:30:00     2506.453942      30.98125
2023-04-16 10:30:00     1947.160861      32.96375

               measuring no Temp S02 1
DateTime
2022-12-07 19:30:00      2603.851233
2022-12-08 13:30:00      1809.151682
2022-12-10 08:30:00      3023.531889
2022-12-13 20:30:00      3490.193048
2022-12-13 21:30:00      4156.807906
...
          ...
2023-04-16 01:30:00      3597.186000
2023-04-16 02:30:00      3743.160697
2023-04-16 03:30:00      4028.053717
2023-04-16 09:30:00      3390.093858
2023-04-16 10:30:00      3169.777062

[570 rows x 3 columns]
```

0.1.11 Sensor 2

```
[ ]: import pandas as pd

directory_path = 'input/'
file_name = sensor_so2_2_name + '_and_temp_valid_1HR.csv'
df_so2_2 = pd.read_csv(directory_path + file_name)
df_so2_2.head()
```

	DateTime	measuring 2	temperature	Hour	measuring 2 no Temp	\
0	2022-12-01 21:30:00	203.905149	29.25375	21	155.913165	
1	2022-12-01 22:30:00	205.143240	29.23250	22	156.961135	
2	2022-12-01 23:30:00	205.382342	29.23250	23	157.200237	
3	2022-12-02 00:30:00	205.765561	29.26375	0	157.863047	
4	2022-12-02 01:30:00	205.179269	29.28125	1	157.433325	

Count 2 Tag

```

0      4  VALID
1      4  VALID
2      4  VALID
3      4  VALID
4      4  VALID

```

0.1.12 Create Sensor Dataframe as Pandas Series

```
[ ]: # Remove the first column with the indexes and save data into web dataframe
dataframe = df_so2_2.drop(df_so2_2.columns[0], axis='columns')
dataframe['DateTime'] = (pd.to_datetime(df_so2_2['DateTime'], u
    ↪infer_datetime_format=True))

# Resample data with 15 mins period and create sensor dataframe
sensor_so2_2_dataframe = dataframe.sort_values(by='DateTime', ascending=True).
    ↪reset_index().drop(columns='index')
sensor_so2_2_dataframe.index = sensor_so2_2_dataframe['DateTime']
sensor_so2_2_dataframe = sensor_so2_2_dataframe.drop(columns=['DateTime', u
    ↪'Hour', 'Count 2', 'Tag'])
sensor_so2_2_dataframe = sensor_so2_2_dataframe.rename(columns={'measuring 2':u
    ↪'measuring SO2 2', 'measuring 2 no Temp': 'measuring no Temp SO2 2',
                           'temperature':u
    ↪'temperature SO2 2'})
sensor_so2_2_dataframe
```

```
/var/folders/wc/_83zcrx913j1dqwg4g90kbhh0000gp/T/ipykernel_8846/377440947.py:3:
UserWarning: The argument 'infer_datetime_format' is deprecated and will be
removed in a future version. A strict version of it is now the default, see
https://pandas.pydata.org/pdeps/0004-consistent-to-datetime-parsing.html. You
can safely remove this argument.
    dataframe['DateTime'] = (pd.to_datetime(df_so2_2['DateTime'],
infer_datetime_format=True))
```

```
[ ]:               measuring SO2 2   temperature SO2 2  \
DateTime
2022-12-01 21:30:00      203.905149      29.25375
2022-12-01 22:30:00      205.143240      29.23250
2022-12-01 23:30:00      205.382342      29.23250
2022-12-02 00:30:00      205.765561      29.26375
2022-12-02 01:30:00      205.179269      29.28125
...
          ...
2023-03-31 13:30:00      85.397741      41.20625
2023-03-31 15:30:00      99.663088      41.13750
2023-03-31 16:30:00     109.607124      39.82500
2023-03-31 17:30:00     137.731671      36.72125
2023-03-31 19:30:00     213.030341      30.76375
```

```

measuring no Temp S02 2
DateTime
2022-12-01 21:30:00      155.913165
2022-12-01 22:30:00      156.961135
2022-12-01 23:30:00      157.200237
2022-12-02 00:30:00      157.863047
2022-12-02 01:30:00      157.433325
...
2023-03-31 13:30:00      143.870195
2023-03-31 15:30:00      157.993882
2023-03-31 16:30:00      156.195106
2023-03-31 17:30:00      160.222656
2023-03-31 19:30:00      178.548184

```

[2034 rows x 3 columns]

0.1.13 PM Sensor

```
[ ]: import pandas as pd

directory_path = 'input/'
file_name = sensor_pm_10_name + '_and_temp_valid_1HR.csv'
df_pm_10 = pd.read_csv(directory_path + file_name)
df_pm_10.head()

[ ]:          DateTime  measuring  Hour  temperature  measuring no Temp \
0  2022-11-23 13:30:00    2.588333    13     28.635000      3.261152
1  2022-11-23 15:30:00    1.341250    15     26.161250      2.150700
2  2022-11-23 16:30:00    1.335000    16     26.441250      2.128985
3  2022-11-23 18:30:00    2.338333    18     28.901667      2.996423
4  2022-11-23 20:30:00    1.025000    20     29.287500      1.661779

      Count    Tag
0      3  VALID
1      4  VALID
2      4  VALID
3      3  VALID
4      4  VALID
```

0.1.14 Create Sensor Dataframe as Pandas Series

```
[ ]: # Remove the first column with the indexes and save data into web dataframe
dataframe = df_pm_10.drop(df_pm_10.columns[0], axis='columns')
dataframe['DateTime'] = (pd.to_datetime(df_pm_10['DateTime'], u
→infer_datetime_format=True))

# Resample data with 15 mins period and create sensor dataframe
```

```

sensor_pm_10_dataframe = dataframe.sort_values(by='DateTime', ascending=True) .
    ↪reset_index().drop(columns='index')
sensor_pm_10_dataframe.index = sensor_pm_10_dataframe['DateTime']
sensor_pm_10_dataframe = sensor_pm_10_dataframe.drop(columns=['DateTime', ↪
    'Hour', 'Count', 'Tag'])
sensor_pm_10_dataframe = sensor_pm_10_dataframe.rename(columns={'measuring': ↪
    'measuring PM10', 'measuring no Temp': 'measuring no Temp PM10', ↪
    'temperature': 'temperature PM10'})
sensor_pm_10_dataframe

```

```

/var/folders/wc/_83zcrx913j1dqwg4g90kbhh0000gp/T/ipykernel_8846/2521444390.py:3:
UserWarning: The argument 'infer_datetime_format' is deprecated and will be
removed in a future version. A strict version of it is now the default, see
https://pandas.pydata.org/pdps/0004-consistent-to-datetime-parsing.html. You
can safely remove this argument.
    dataframe['DateTime'] = (pd.to_datetime(df_pm_10['DateTime'],
infer_datetime_format=True))

```

```

[ ]:               measuring PM10  temperature PM10  measuring no Temp PM10
DateTime
2022-11-23 13:30:00      2.588333     28.635000      3.261152
2022-11-23 15:30:00      1.341250     26.161250      2.150700
2022-11-23 16:30:00      1.335000     26.441250      2.128985
2022-11-23 18:30:00      2.338333     28.901667      2.996423
2022-11-23 20:30:00      1.025000     29.287500      1.661779
...
...                      ...
...                      ...
...                      ...
2023-04-19 19:30:00      0.375000     23.648750      1.323222
2023-04-20 14:30:00      0.698750     35.017500      1.019047
2023-04-20 15:30:00      1.735000     35.582500      2.024091
2023-04-20 16:30:00      2.346250     34.355000      2.703139
2023-04-20 17:30:00      2.101667     32.158333      2.579883

[1229 rows x 3 columns]

```

0.2 Compare with original data

```

[ ]: input_data_directory = 'input/'
reference_data_path = input_data_directory + 'ref_air_quality_data_Vila_Moema.
    ↪csv'
reference_column_name = 'Ozônio'

```

0.3 Load reference and sensor data

```

[ ]: import pandas as pd

reference_data = pd.read_csv(reference_data_path)

```

```

reference_data['DateTime'] = (pd.to_datetime(reference_data['DateTime'],
                                             infer_datetime_format=True))
reference_data = reference_data.sort_values(by='DateTime', ascending=True).
    reset_index().drop(columns='index')
reference_data.index = reference_data['DateTime']
reference_data = reference_data.drop(columns='DateTime')[reference_column_name]

reference_data

```

```

/var/folders/wc/_83zcrx913j1dqwg4g90kbhh0000gp/T/ipykernel_8846/2986665073.py:4:
UserWarning: The argument 'infer_datetime_format' is deprecated and will be
removed in a future version. A strict version of it is now the default, see
https://pandas.pydata.org/pdeps/0004-consistent-to-datetime-parsing.html. You
can safely remove this argument.
    reference_data['DateTime'] = (pd.to_datetime(reference_data['DateTime'],
                                                infer_datetime_format=True))

```

```

[ ]: DateTime
2022-01-01 02:30:00    13.23
2022-01-01 03:30:00    12.07
2022-01-01 04:30:00    13.24
2022-01-01 05:30:00    14.42
2022-01-01 06:30:00    13.30
...
2023-02-08 12:30:00    50.01
2023-02-08 13:30:00    67.43
2023-02-08 14:30:00    72.46
2023-02-08 15:30:00    59.65
2023-02-08 16:30:00    NaN
Name: Ozônio, Length: 9687, dtype: float64

```

0.4 Merge sensor and reference data

```

[ ]: import numpy as np

def merge_temperatures(df):
    # df[0]: CO, df[1]: NO2, df[2]: O3
    # df[3]: O32, df[4]: SO21, df[5]: SO22
    if not np.isnan(df[0]): return df[0]
    elif not np.isnan(df[1]): return df[1]
    elif not np.isnan(df[2]): return df[2]
    elif not np.isnan(df[3]): return df[3]
    elif not np.isnan(df[4]): return df[4]
    elif not np.isnan(df[5]): return df[5]
    return df[6]

sensor_data = pd.concat([sensor_co_dataframe, sensor_no2_dataframe,

```

```

        sensor_o3_1_dataframe, sensor_o3_2_dataframe,
        sensor_so2_1_dataframe, sensor_so2_2_dataframe,
        sensor_pm_10_dataframe], join='outer', axis=1)

sensor_data['temperature'] = (sensor_data[['temperature CO', 'temperature NO2',
                                          'temperature O3 1',
                                          'temperature O3 2', 'temperature SO2 1',
                                          'temperature SO2 2', 'temperature PM10']]
                               .apply(lambda df: df
                                     .merge_temperatures(df), axis=1))
sensor_data = sensor_data.drop(columns=['temperature CO', 'temperature NO2',
                                         'temperature O3 1',
                                         'temperature O3 2', 'temperature SO2 1',
                                         'temperature SO2 2', 'temperature PM10'])
sensor_data

```

	measuring CO	measuring no Temp CO	measuring NO2	\
DateTime				
2022-11-23 13:30:00	NaN	NaN	NaN	
2022-11-23 15:30:00	NaN	NaN	NaN	
2022-11-23 16:30:00	NaN	NaN	NaN	
2022-11-23 18:30:00	NaN	NaN	NaN	
2022-11-23 20:30:00	NaN	NaN	NaN	
...	
2023-04-21 17:30:00	NaN	NaN	NaN	
2023-04-21 18:30:00	NaN	NaN	NaN	
2023-04-21 19:30:00	NaN	NaN	NaN	
2023-04-21 20:30:00	NaN	NaN	NaN	
2023-04-21 21:30:00	NaN	NaN	NaN	
	measuring O3 1	measuring no Temp O3 1	measuring O3 2	\
DateTime				
2022-11-23 13:30:00	NaN	NaN	NaN	
2022-11-23 15:30:00	NaN	NaN	NaN	
2022-11-23 16:30:00	NaN	NaN	NaN	
2022-11-23 18:30:00	NaN	NaN	NaN	
2022-11-23 20:30:00	NaN	NaN	NaN	
...	
2023-04-21 17:30:00	NaN	NaN	50.964672	
2023-04-21 18:30:00	NaN	NaN	42.360948	
2023-04-21 19:30:00	NaN	NaN	39.303264	
2023-04-21 20:30:00	NaN	NaN	37.688532	
2023-04-21 21:30:00	NaN	NaN	37.048038	
	measuring no Temp O3 2	measuring SO2 1	\	
DateTime				
2022-11-23 13:30:00	NaN	NaN		

2022-11-23 15:30:00	NaN	NaN
2022-11-23 16:30:00	NaN	NaN
2022-11-23 18:30:00	NaN	NaN
2022-11-23 20:30:00	NaN	NaN
...
2023-04-21 17:30:00	7.395241	NaN
2023-04-21 18:30:00	7.111857	NaN
2023-04-21 19:30:00	8.485160	NaN
2023-04-21 20:30:00	8.997179	NaN
2023-04-21 21:30:00	9.709237	NaN

measuring no Temp S02 1 measuring S02 2 \

Date	Time	Temp	S02 1	S02 2
2022-11-23	13:30:00	NaN	NaN	
2022-11-23	15:30:00	NaN	NaN	
2022-11-23	16:30:00	NaN	NaN	
2022-11-23	18:30:00	NaN	NaN	
2022-11-23	20:30:00	NaN	NaN	
...
2023-04-21	17:30:00	NaN	NaN	
2023-04-21	18:30:00	NaN	NaN	
2023-04-21	19:30:00	NaN	NaN	
2023-04-21	20:30:00	NaN	NaN	
2023-04-21	21:30:00	NaN	NaN	

measuring no Temp S02 2 measuring PM10 \

Date	Time	Temp	PM10
2022-11-23	13:30:00	NaN	2.588333
2022-11-23	15:30:00	NaN	1.341250
2022-11-23	16:30:00	NaN	1.335000
2022-11-23	18:30:00	NaN	2.338333
2022-11-23	20:30:00	NaN	1.025000
...
2023-04-21	17:30:00	NaN	NaN
2023-04-21	18:30:00	NaN	NaN
2023-04-21	19:30:00	NaN	NaN
2023-04-21	20:30:00	NaN	NaN
2023-04-21	21:30:00	NaN	NaN

measuring no Temp PM10 temperature

Date	Time	Temp	PM10	temperature
2022-11-23	13:30:00	3.261152	28.635000	
2022-11-23	15:30:00	2.150700	26.161250	
2022-11-23	16:30:00	2.128985	26.441250	
2022-11-23	18:30:00	2.996423	28.901667	
2022-11-23	20:30:00	1.661779	29.287500	
...

```

2023-04-21 17:30:00      NaN    30.536250
2023-04-21 18:30:00      NaN    27.137500
2023-04-21 19:30:00      NaN    25.327500
2023-04-21 20:30:00      NaN    24.458750
2023-04-21 21:30:00      NaN    23.906250

```

[2711 rows x 14 columns]

```
[ ]: sensor_data = pd.concat([sensor_data, reference_data], axis=1, join='inner')
sensor_data = sensor_data.rename(columns={'Ozônio': 'reference'})
sensor_data
```

	measuring CO	measuring no Temp CO	measuring NO2	\
DateTime				
2022-11-23 13:30:00	NaN	NaN	NaN	
2022-11-23 15:30:00	NaN	NaN	NaN	
2022-11-23 16:30:00	NaN	NaN	NaN	
2022-11-23 18:30:00	NaN	NaN	NaN	
2022-11-23 20:30:00	NaN	NaN	NaN	
...	
2023-02-08 12:30:00	NaN	NaN	NaN	
2023-02-08 13:30:00	NaN	NaN	NaN	
2023-02-08 14:30:00	NaN	NaN	NaN	
2023-02-08 15:30:00	NaN	NaN	NaN	
2023-02-08 16:30:00	NaN	NaN	NaN	
	measuring O3 1	measuring no Temp O3 1	measuring O3 2	\
DateTime				
2022-11-23 13:30:00	NaN	NaN	NaN	
2022-11-23 15:30:00	NaN	NaN	NaN	
2022-11-23 16:30:00	NaN	NaN	NaN	
2022-11-23 18:30:00	NaN	NaN	NaN	
2022-11-23 20:30:00	NaN	NaN	NaN	
...	
2023-02-08 12:30:00	NaN	NaN	80.255616	
2023-02-08 13:30:00	NaN	NaN	81.669120	
2023-02-08 14:30:00	NaN	NaN	80.498562	
2023-02-08 15:30:00	NaN	NaN	79.146408	
2023-02-08 16:30:00	NaN	NaN	71.617536	
	measuring no Temp O3 2	measuring S02 1	\	
DateTime				
2022-11-23 13:30:00	NaN	NaN		
2022-11-23 15:30:00	NaN	NaN		
2022-11-23 16:30:00	NaN	NaN		
2022-11-23 18:30:00	NaN	NaN		
2022-11-23 20:30:00	NaN	NaN		

	
2023-02-08 12:30:00		15.204469		NaN
2023-02-08 13:30:00		15.351102		NaN
2023-02-08 14:30:00		13.929618		NaN
2023-02-08 15:30:00		12.788609		NaN
2023-02-08 16:30:00		9.874328		NaN

	measuring no	Temp S02 1	measuring S02 2	\
DateTime				
2022-11-23 13:30:00		NaN	NaN	
2022-11-23 15:30:00		NaN	NaN	
2022-11-23 16:30:00		NaN	NaN	
2022-11-23 18:30:00		NaN	NaN	
2022-11-23 20:30:00		NaN	NaN	
...	
2023-02-08 12:30:00		NaN	105.037977	
2023-02-08 13:30:00		NaN	101.837936	
2023-02-08 14:30:00		NaN	106.043517	
2023-02-08 15:30:00		NaN	110.596287	
2023-02-08 16:30:00		NaN	136.193337	

	measuring no	Temp S02 2	measuring PM10	\
DateTime				
2022-11-23 13:30:00		NaN	2.588333	
2022-11-23 15:30:00		NaN	1.341250	
2022-11-23 16:30:00		NaN	1.335000	
2022-11-23 18:30:00		NaN	2.338333	
2022-11-23 20:30:00		NaN	1.025000	
...	
2023-02-08 12:30:00		147.029487	1.288333	
2023-02-08 13:30:00		148.459469	1.476250	
2023-02-08 14:30:00		153.582107	1.775000	
2023-02-08 15:30:00		157.363207	2.096250	
2023-02-08 16:30:00		166.095342	2.875000	

	measuring no	Temp PM10	temperature	reference
DateTime				
2022-11-23 13:30:00		3.261152	28.635000	37.76
2022-11-23 15:30:00		2.150700	26.161250	33.53
2022-11-23 16:30:00		2.128985	26.441250	30.82
2022-11-23 18:30:00		2.996423	28.901667	21.18
2022-11-23 20:30:00		1.661779	29.287500	20.67
...	
2023-02-08 12:30:00		1.362800	39.311250	50.01
2023-02-08 13:30:00		1.530810	39.828750	67.43
2023-02-08 14:30:00		1.823899	39.931250	72.46
2023-02-08 15:30:00		2.149913	39.845000	59.65

```
2023-02-08 16:30:00      3.032776      37.960000      NaN
```

```
[1430 rows x 15 columns]
```

0.5 Calibrate data

0.5.1 Prepare training and test sets

```
[ ]: from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_validate

reference_median = sensor_data['reference'].median()
temperature_mean = sensor_data['temperature'].mean()
sensor_co_median = sensor_data['measuring CO'].median()
trend_co_median = sensor_data['measuring no Temp CO'].median()
sensor_no2_median = sensor_data['measuring NO2'].median()
sensor_o3_1_median = sensor_data['measuring O3 1'].median()
trend_o3_1_median = sensor_data['measuring no Temp O3 1'].median()
sensor_o3_2_median = sensor_data['measuring O3 2'].median()
trend_o3_2_median = sensor_data['measuring no Temp O3 2'].median()
sensor_so2_1_median = sensor_data['measuring SO2 1'].median()
trend_so2_1_median = sensor_data['measuring no Temp SO2 1'].median()
sensor_so2_2_median = sensor_data['measuring SO2 2'].median()
trend_so2_2_median = sensor_data['measuring no Temp SO2 2'].median()
sensor_pm_10_median = sensor_data['measuring PM10'].median()

variables_names = ['measuring CO', 'measuring NO2', 'measuring O3 1',
                   'measuring O3 2', 'measuring PM10', 'temperature']

y = sensor_data['reference'].fillna(value=reference_median)
X = (sensor_data[variables_names].fillna(value={
                    variables_names[0]: sensor_co_median,
                    variables_names[1]: sensor_no2_median,
                    variables_names[2]: sensor_o3_1_median,
                    variables_names[3]: sensor_o3_2_median,
                    variables_names[4]: sensor_pm_10_median,
                    variables_names[5]: temperature_mean}).values.
    ↪reshape(-1,6))

X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=42)
```

0.5.2 Grid search with different models and variables combinations

```
[ ]: from itertools import combinations

def check_if_list_contains(list1, list2):
    return [element for element in list1 if element in list2]
```

```

indexes = []
reference_indexes = [2, 3]
num_variables = len(variables_names)
for num_combinations in list(range(num_variables)):
    contains_reference = False
    index_list = [list(index_tuple) for index_tuple in
    ↪list(combinations(list(range(num_variables)), r=num_combinations+1))]
    for sublist in index_list:
        contains_reference = check_if_list_contains(sublist, reference_indexes)
        if contains_reference:
            indexes.append(sublist)
            contains_reference = False

feature_subsets = { }
for index_list in indexes:
    key = ""
    trends_subset = ""
    for index in index_list:
        key = key + variables_names[index] + ","
    feature_subsets[key] = index_list
feature_subsets

```

```

[ ]: {'measuring O3 1': [2],
'measuring O3 2': [3],
'measuring CO,measuring O3 1': [0, 2],
'measuring CO,measuring O3 2': [0, 3],
'measuring NO2,measuring O3 1': [1, 2],
'measuring NO2,measuring O3 2': [1, 3],
'measuring O3 1,measuring O3 2': [2, 3],
'measuring O3 1,measuring PM10': [2, 4],
'measuring O3 1,temperature': [2, 5],
'measuring O3 2,measuring PM10': [3, 4],
'measuring O3 2,temperature': [3, 5],
'measuring CO,measuring NO2,measuring O3 1': [0, 1, 2],
'measuring CO,measuring NO2,measuring O3 2': [0, 1, 3],
'measuring CO,measuring O3 1,measuring O3 2': [0, 2, 3],
'measuring CO,measuring O3 1,measuring PM10': [0, 2, 4],
'measuring CO,measuring O3 1,temperature': [0, 2, 5],
'measuring CO,measuring O3 2,measuring PM10': [0, 3, 4],
'measuring CO,measuring O3 2,temperature': [0, 3, 5],
'measuring NO2,measuring O3 1,measuring O3 2': [1, 2, 3],
'measuring NO2,measuring O3 1,measuring PM10': [1, 2, 4],
'measuring NO2,measuring O3 1,temperature': [1, 2, 5],
'measuring NO2,measuring O3 2,measuring PM10': [1, 3, 4],
'measuring NO2,measuring O3 2,temperature': [1, 3, 5],
'measuring O3 1,measuring O3 2,measuring PM10': [2, 3, 4],

```

```
'measuring O3 1,measuring O3 2,temperature,' : [2, 3, 5],  
'measuring O3 1,measuring PM10,temperature,' : [2, 4, 5],  
'measuring O3 2,measuring PM10,temperature,' : [3, 4, 5],  
'measuring CO,measuring NO2,measuring O3 1,measuring O3 2,' : [0, 1, 2, 3],  
'measuring CO,measuring NO2,measuring O3 1,measuring PM10,' : [0, 1, 2, 4],  
'measuring CO,measuring NO2,measuring O3 1,temperature,' : [0, 1, 2, 5],  
'measuring CO,measuring NO2,measuring O3 2,measuring PM10,' : [0, 1, 3, 4],  
'measuring CO,measuring NO2,measuring O3 2,temperature,' : [0, 1, 3, 5],  
'measuring CO,measuring O3 1,measuring O3 2,measuring PM10,' : [0, 2, 3, 4],  
'measuring CO,measuring O3 1,measuring O3 2,temperature,' : [0, 2, 3, 5],  
'measuring CO,measuring O3 1,measuring PM10,temperature,' : [0, 2, 4, 5],  
'measuring CO,measuring O3 2,measuring PM10,temperature,' : [0, 3, 4, 5],  
'measuring NO2,measuring O3 1,measuring O3 2,measuring PM10,' : [1, 2, 3, 4],  
'measuring NO2,measuring O3 1,measuring O3 2,temperature,' : [1, 2, 3, 5],  
'measuring NO2,measuring O3 1,measuring PM10,temperature,' : [1, 2, 4, 5],  
'measuring NO2,measuring O3 2,measuring PM10,temperature,' : [1, 3, 4, 5],  
'measuring O3 1,measuring O3 2,measuring PM10,temperature,' : [2, 3, 4, 5],  
'measuring CO,measuring NO2,measuring O3 1,measuring O3 2,measuring PM10,' : [0,  
1,  
2,  
3,  
4],  
'measuring CO,measuring NO2,measuring O3 1,measuring O3 2,temperature,' : [0,  
1,  
2,  
3,  
5],  
'measuring CO,measuring NO2,measuring O3 1,measuring PM10,temperature,' : [0,  
1,  
2,  
4,  
5],  
'measuring CO,measuring NO2,measuring O3 2,measuring PM10,temperature,' : [0,  
1,  
3,  
4,  
5],  
'measuring CO,measuring O3 1,measuring O3 2,measuring PM10,temperature,' : [0,  
2,  
3,  
4,  
5],  
'measuring NO2,measuring O3 1,measuring O3 2,measuring PM10,temperature,' : [1,  
2,  
3,  
4,  
5],
```

```
'measuring CO,measuring NO2,measuring O3 1,measuring O3 2,measuring
PM10,temperature,' : [0,
1,
2,
3,
4,
5]}
```

Function for plotting observations vs. predictions

```
[ ]: import matplotlib.pyplot as plt
from scipy.stats import spearmanr, kendalltau, gaussian_kde
import numpy as np
import os

def plot_predictions_and_observations(X, y, r2, rmse, mae, aic, bic, file_name):
    fig, ax = plt.subplots(figsize=(1.3*5,5))
    xy = np.vstack([X, y])
    z = gaussian_kde(xy)(xy)
    ax.scatter(X, y, c=z,s=15,alpha=.5)
    spear_corr, p_value = spearmanr(y, X)
    spearman_text = ''
    alpha = 0.05
    if p_value > alpha:
        spearman_text = 'Coeficiente de Spearman: {:.2f}'.format(spear_corr) + u
        ' , p>0.05'
    else:
        spearman_text = 'Coeficiente de Spearman: {:.2f}'.format(spear_corr) + u
        ' , p<0.05'

    kendall_corr, p_value = kendalltau(y, X)
    alpha = 0.05
    kendall_text = ''
    if p_value > alpha:
        kendall_text = 'Coeficiente de Kendall: {:.2f}'.format(kendall_corr) + u
        ' , p>0.05'
    else:
        kendall_text = 'Coeficiente de Kendall: {:.2f}'.format(kendall_corr) + u
        ' , p<0.05'

    plt.text(0.02, 0.95, spearman_text, ha='left', va='center', transform=plt.
   .gca().transAxes, fontsize=12)
    plt.text(0.02, 0.90, kendall_text, ha='left', va='center', transform=plt.
    gca().transAxes, fontsize=12)
    r2_text = 'R2 = {:.2f} ± {:.2f}'.format(r2.mean(), r2.
    std())
    rmse_text = 'RMSE = {:.2f} ± {:.2f}'.format(rmse.mean(), rmse.std())
```

```

mae_text = 'MAE = {:.2f} ± {:.2f}'.format(mae.mean(), mae.std())
aic_text = 'AIC = {:.2f} ± {:.2f}'.format(aic.mean(), aic.std())
bic_text = 'BIC = {:.2f} ± {:.2f}'.format(bic.mean(), bic.std())
plt.text(0.02, 0.85, r2_text, ha='left', va='center', transform=plt.gca().transAxes, fontsize=12)
plt.text(0.02, 0.80, rmse_text, ha='left', va='center', transform=plt.gca().transAxes, fontsize=12)
plt.text(0.02, 0.75, mae_text, ha='left', va='center', transform=plt.gca().transAxes, fontsize=12)
plt.text(0.02, 0.70, aic_text, ha='left', va='center', transform=plt.gca().transAxes, fontsize=12)
plt.text(0.02, 0.65, bic_text, ha='left', va='center', transform=plt.gca().transAxes, fontsize=12)

ax.set_xlim([np.min([y,X]),np.max([y,X])])
ax.set_ylim([np.min([y,X]),np.max([y,X])])
ax.set_aspect('equal')

ax.plot([xy.min(), xy.max()], [xy.min(), xy.max()], 'k-', lw=1,dashes=[2,2])
ax.fill_between(np.linspace(xy.min(), xy.max(),y.shape[0]),
                np.linspace(xy.min(), xy.max(),y.shape[0])*0.5,
                alpha=0.2,facecolor='gray',edgecolor=None)
ax.fill_between(np.linspace(xy.min(),xy.max(),y.shape[0]),
                np.linspace(xy.max(),xy.max(),y.shape[0]),
                np.linspace(xy.min(),xy.max(),y.shape[0])*2,
                alpha=0.2,facecolor='gray',edgecolor=None)

ax.set_xlabel('Concentração de O3 observada (ug/m\N{SUPERSCRIPT THREE})',fontsize=12)
ax.set_ylabel('Concentração de O3 inferida (ug/m\N{SUPERSCRIPT THREE})',fontsize=12)

if not os.path.exists('images/'):
    os.makedirs('images/')

plt.savefig('images/' + 'O3_ALL_' + file_name + '.png')

```

```

[ ]: def calculate_bic(n, mse, num_params):
    bic = n * np.log(mse) + num_params * np.log(n)
    return bic

def calculate_aic(n, mse, num_params):
    aic = n * np.log(mse) + 2 * num_params
    return aic

```

```
[ ]: from sklearn.linear_model import LinearRegression
from sklearn.neural_network import MLPRegressor
from sklearn.neighbors import KNeighborsRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import GridSearchCV
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import Pipeline
import numpy as np

models = {
    'MLP Regression': (
        ('mlp_regressor', MLPRegressor(solver="lbfgs", max_iter=1000, random_state=42)), {
            'mlp_regressor__hidden_layer_sizes': [(4,50), (10,10), (200,), (200,4), (200,10), (200,50)],
            'mlp_regressor__alpha': [0.001, 0.01, 0.1, 1, 10]
        }
    ),
    'Multilinear Regression': (
        ('linear_regressor', LinearRegression()), {}
    ),
    'KNN Regression': (
        ('knn_regressor', KNeighborsRegressor()), {
            'knn_regressor__n_neighbors': [13, 15, 17, 20]
        }
    ),
    'Random Forests Regression': (
        ('random_forest_regressor', RandomForestRegressor()), {
            'random_forest_regressor__n_estimators': [100, 150],
            'random_forest_regressor__max_depth': [None, 10],
            'random_forest_regressor__min_samples_split': [2, 10],
            'random_forest_regressor__min_samples_leaf': [1, 2, 4]
        }
    )
}

# Perform grid search for each feature subset
results = {}
rmse_by_features = {}
r2_by_features = {}
mae_by_features = {}
mse_by_features = {}
aic_by_features = {}
bic_by_features = {}

for features_set, subset in feature_subsets.items():
    X_subset = X[:, subset]
    X_train_subset = X_train[:, subset]
```

```

X_test_subset = X_test[:, subset]

model_results = {}
model_rmse = {}
model_r2 = {}
model_mae = {}
model_mse = {}
model_aic = {}
model_bic = {}

for model_name, (model, param_grid) in models.items():
    print(f"Grid search for features: {features_set} with model: {model_name}...")

    pipeline = Pipeline([
        ('scaler', StandardScaler()),
        model
    ])

    # Perform grid search with cross-validation
    grid_search = GridSearchCV(pipeline, param_grid, cv=3,
                                scoring='neg_root_mean_squared_error', n_jobs=-1)
    grid_search.fit(X_train_subset, y_train)

    # Print the best parameters and best score
    best_params = grid_search.best_params_

    # Evaluate the best model on the test set
    best_model = grid_search.best_estimator_
    cross_validation = cross_validate(best_model, X_subset, y, cv=3,
                                      scoring=['r2', 'neg_root_mean_squared_error',
                                      'neg_mean_absolute_error', 'neg_mean_squared_error'])

    y_pred = best_model.predict(X_test_subset)

    # Evaluate the model
    r2 = cross_validation['test_r2']
    rmse = cross_validation['test_neg_root_mean_squared_error']
    mae = cross_validation['test_neg_mean_absolute_error']
    mse = cross_validation['test_neg_mean_squared_error']
    num_params = best_model.n_features_in_ + 1
    print(f"Number of parameters: {num_params}")
    aic = calculate_aic(len(y_pred), mse=abs(mse), num_params=num_params)
    bic = calculate_bic(len(y_pred), mse=abs(mse), num_params=num_params)

    plot_predictions_and_observations(y_test, y_pred, r2=r2, rmse=rmse,
                                     mae=mae,

```

```

    aic=aic, bic=bic,□
→file_name=model_name+features_set)

model_results[model_name] = {
    'Best Model': best_model,
    'Best Parameters': best_params,
    'Test R2': r2,
    'Test RMSE': rmse,
    'Test MAE': mae,
    'Test MSE': mse,
    'AIC': aic,
    'BIC': bic
}
model_rmse[model_name] = {
    'Mean': rmse.mean(),
    'Std': rmse.std()
}
model_r2[model_name] = {
    'Mean': r2.mean(),
    'Std': r2.std()
}
model_mae[model_name] = {
    'Mean': mae.mean(),
    'Std': mae.std()
}
model_mse[model_name] = {
    'Mean': mse.mean(),
    'Std': mse.std()
}
model_aic[model_name] = {
    'Mean': aic.mean(),
    'Std': aic.std()
}
model_bic[model_name] = {
    'Mean': bic.mean(),
    'Std': bic.std()
}

results[features_set] = model_results
rmse_by_features[features_set] = model_rmse
r2_by_features[features_set] = model_r2
mae_by_features[features_set] = model_mae
mse_by_features[features_set] = model_mse
aic_by_features[features_set] = model_aic
bic_by_features[features_set] = model_bic

for feature_set, models in results.items():

```

```

for model_name, result in models.items():
    print(f"\nResults for features: {feature_set} with model: {model_name}:
        ")
    print(f"Best Parameters: {result['Best Parameters']} ")
    print(f"Test RMSE: {result['Test RMSE'].mean()} +/- {result['Test RMSE'].std()}")
    print(f"Test R2: {result['Test R2'].mean()} +/- {result['Test R2'].std()}")
    print(f"Test MAE: {result['Test MAE'].mean()} +/- {result['Test MAE'].std()}")
    print(f"Test MSE: {result['Test MSE'].mean()} +/- {result['Test MSE'].std()}")
    print(f"Test AIC: {result['AIC'].mean()} +/- {result['AIC'].std()}")
    print(f"Test BIC: {result['BIC'].mean()} +/- {result['BIC'].std()}")

```

1 Save Results

```

[ ]: output_directory_path = 'output/'
rmse_file_name = output_directory_path + sensor_name + '_rmse.csv'
r2_file_name = output_directory_path + sensor_name + '_r2.csv'
mae_file_name = output_directory_path + sensor_name + '_mae.csv'
mse_file_name = output_directory_path + sensor_name + '_mse.csv'
aic_file_name = output_directory_path + sensor_name + '_aic.csv'
bic_file_name = output_directory_path + sensor_name + '_bic.csv'
results_file_name = output_directory_path + sensor_name + '_results.csv'

pd.DataFrame(rmse_by_features).transpose().to_csv(rmse_file_name)
pd.DataFrame(r2_by_features).transpose().to_csv(r2_file_name)
pd.DataFrame(mae_by_features).transpose().to_csv(mae_file_name)
pd.DataFrame(mse_by_features).transpose().to_csv(mse_file_name)
pd.DataFrame(aic_by_features).transpose().to_csv(aic_file_name)
pd.DataFrame(bic_by_features).transpose().to_csv(bic_file_name)
pd.DataFrame(results).transpose().to_csv(results_file_name)

```

1.1 Plot Results

```

[ ]: mean_r2_by_features_dataframe = pd.DataFrame()
std_r2_by_features_dataframe = pd.DataFrame()

mean_rmse_by_features_dataframe = pd.DataFrame()
std_rmse_by_features_dataframe = pd.DataFrame()

mean_mae_by_features_dataframe = pd.DataFrame()
std_mae_by_features_dataframe = pd.DataFrame()

mean_mse_by_features_dataframe = pd.DataFrame()

```

```

std_mse_by_features_dataframe = pd.DataFrame()

mean_aic_by_features_dataframe = pd.DataFrame()
std_aic_by_features_dataframe = pd.DataFrame()

mean_bic_by_features_dataframe = pd.DataFrame()
std_bic_by_features_dataframe = pd.DataFrame()

for key in list(feature_subsets.keys()):
    feature_dict = r2_by_features[key]
    for model in list(feature_dict.keys()):
        colum_name = key.replace('measuring', '')
        colum_name = colum_name.replace(' | ', ',')
        colum_name += f': {model[:-11]}'
        mean_r2_by_features_dataframe[colum_name] = \
            [feature_dict[model]['Mean']]
        std_r2_by_features_dataframe[colum_name] = [feature_dict[model]['Std']]

for key in list(feature_subsets.keys()):
    feature_dict = rmse_by_features[key]
    for model in list(feature_dict.keys()):
        colum_name = key.replace('measuring', '')
        colum_name = colum_name.replace(' | ', ',')
        colum_name += f': {model[:-11]}'
        mean_rmse_by_features_dataframe[colum_name] = \
            [feature_dict[model]['Mean']]
        std_rmse_by_features_dataframe[colum_name] = \
            [feature_dict[model]['Std']]

for key in list(feature_subsets.keys()):
    feature_dict = mae_by_features[key]
    for model in list(feature_dict.keys()):
        colum_name = key.replace('measuring', '')
        colum_name = colum_name.replace(' | ', ',')
        colum_name += f': {model[:-11]}'
        mean_mae_by_features_dataframe[colum_name] = \
            [feature_dict[model]['Mean']]
        std_mae_by_features_dataframe[colum_name] = [feature_dict[model]['Std']]

for key in list(feature_subsets.keys()):
    feature_dict = mse_by_features[key]
    for model in list(feature_dict.keys()):
        colum_name = key.replace('measuring', '')
        colum_name = colum_name.replace(' | ', ',')
        colum_name += f': {model[:-11]}'
        mean_mse_by_features_dataframe[colum_name] = \
            [feature_dict[model]['Mean']]

```

```

    std_mse_by_features_dataframe[colum_name] = [feature_dict[model]['Std']]

for key in list(feature_subsets.keys()):
    feature_dict = aic_by_features[key]
    for model in list(feature_dict.keys()):
        colum_name = key.replace('measuring', '')
        colum_name = colum_name.replace(' | ', ',')
        colum_name += f': {model[:-11]}'
        mean_aic_by_features_dataframe[colum_name] = \
            [feature_dict[model]['Mean']]
    std_aic_by_features_dataframe[colum_name] = [feature_dict[model]['Std']]

for key in list(feature_subsets.keys()):
    feature_dict = bic_by_features[key]
    for model in list(feature_dict.keys()):
        colum_name = key.replace('measuring', '')
        colum_name = colum_name.replace(' | ', ',')
        colum_name += f': {model[:-11]}'
        mean_bic_by_features_dataframe[colum_name] = \
            [feature_dict[model]['Mean']]
    std_bic_by_features_dataframe[colum_name] = [feature_dict[model]['Std']]

```

```

[ ]: import matplotlib.pyplot as plt
import numpy as np

def plot_metrics(features, r2_list, r2_error_list, rmse_list, rmse_error_list, \
                 mae_list, mae_error_list):
    bottom, height = 0.1, 0.65
    left, width = bottom, height*1.3
    spacing = 0.03

    rect_r2 = [left-width-spacing, bottom, width, height]
    rect_rmse = [left, bottom, width, height]
    rect_mae = [left + width + spacing, bottom, height/1.3, height]

    plt.figure(figsize=(1.3*5,5))

    ax_r2 = plt.axes(rect_r2)
    ax_r2.tick_params(direction='in', top=True, right=True, labelsize=14)
    ax_r2.set_title('R2')

    ax_rmse = plt.axes(rect_rmse)
    ax_rmse.tick_params(direction='in', labelleft=False, labelsize=14)
    ax_rmse.set_title('RMSE')

    ax_mae = plt.axes(rect_mae)
    ax_mae.tick_params(direction='in', labelleft=False, labelsize=14)

```

```

ax_mae.set_title('MAE')

y_pos = np.arange(len(features))

ax_r2.banh(y_pos, r2_list, xerr=r2_error_list, align='center')
min_r2 = r2_list.min() - r2_error_list.max()
ax_r2.set_xlim([min_r2 - 0.05, 1.0 + 0.05])
ax_r2.set_yticks(y_pos, labels=features, fontsize=14)
ax_r2.invert_yaxis() # labels read top-to-bottom
ax_r2.set_xlabel('R2', fontsize=14)

ax_rmse.banh(y_pos, rmse_list, xerr=rmse_error_list, align='center')
max_rmse = rmse_list.max() + rmse_error_list.max()
min_rmse = rmse_list.min() - rmse_error_list.max()
if max_rmse <= 0: max_rmse = -min_rmse
ax_rmse.set_xlim([min_rmse - 0.05, max_rmse + 0.05])
ax_rmse.set_yticks(y_pos, labels=features, fontsize=14)
ax_rmse.invert_yaxis() # labels read top-to-bottom
ax_rmse.set_xlabel('RMSE', fontsize=14)

ax_mae.banh(y_pos, mae_list, xerr=mae_error_list, align='center')
max_mae = mae_list.max() + mae_error_list.max()
min_mae = mae_list.min() - mae_error_list.max()
if max_mae <= 0: max_mae = -min_mae
ax_mae.set_xlim([min_mae - 0.05, max_mae + 0.05])
ax_mae.set_yticks(y_pos, labels=features, fontsize=14)
ax_mae.invert_yaxis() # labels read top-to-bottom
ax_mae.set_xlabel('MAE', fontsize=14)

```

```

[ ]: r2_sorted_dataframe = (mean_r2_by_features_dataframe.
    ↪sort_values(by=mean_r2_by_features_dataframe.index[0], axis=1, ↪
    ↪ascending=False))
features = r2_sorted_dataframe.columns

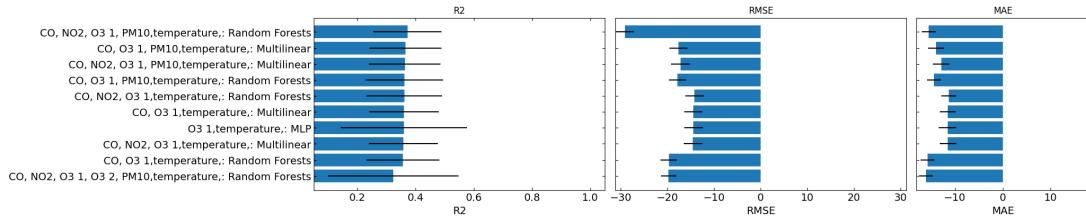
mean_r2 = r2_sorted_dataframe.values.flatten()
error_r2 = std_r2_by_features_dataframe[r2_sorted_dataframe.columns].values.
    ↪flatten()

mean_rmse = mean_rmse_by_features_dataframe.values.flatten()
error_rmse = std_rmse_by_features_dataframe[r2_sorted_dataframe.columns].values.
    ↪flatten()

mean_mae = mean_mae_by_features_dataframe.values.flatten()
error_mae = std_mae_by_features_dataframe[r2_sorted_dataframe.columns].values.
    ↪flatten()

```

```
plot_metrics(features=features[:10], r2_list=mean_r2[:10],  
            r2_error_list=error_r2[:10],  
            rmse_list=mean_rmse[:10], rmse_error_list=error_rmse[:10],  
            mae_list=mean_mae[:10], mae_error_list=error_mae[:10])
```



```
[ ]: import matplotlib.pyplot as plt  
import numpy as np  
  
def plot_other_metrics(features, first_list, first_error_list, first_title,  
                      second_list, second_error_list, second_title,  
                      third_list, third_error_list, third_title):  
    bottom, height = 0.1, 0.65  
    left, width = bottom, height*1.3  
    spacing = 0.005  
  
    rect_r2 = [left-width-spacing, bottom, width, height]  
    rect_rmse = [left, bottom, width, height]  
    rect_mae = [left + width + spacing, bottom, height/1.3, height]  
  
    plt.figure(figsize=(1.3*5,5))  
  
    ax_r2 = plt.axes(rect_r2)  
    ax_r2.tick_params(direction='in', top=True, right=True, labelsize=14)  
    ax_r2.set_title(first_title)  
  
    ax_rmse = plt.axes(rect_rmse)  
    ax_rmse.tick_params(direction='in', labelleft=False, labelsize=14)  
    ax_rmse.set_title(second_title)  
  
    ax_mae = plt.axes(rect_mae)  
    ax_mae.tick_params(direction='in', labelleft=False, labelsize=14)  
    ax_mae.set_title(third_title)  
  
    y_pos = np.arange(len(features))  
  
    # lim_max = df['measuring'].max()+df['measuring'].max()*10/100  
    # lim_min = df['measuring'].min()-df['measuring'].min()*10/100
```

```

ax_r2.banh(y_pos, first_list, xerr=first_error_list, align='center')
ax_r2.set_yticks(y_pos, labels=features, fontsize=14)
ax_r2.invert_yaxis() # labels read top-to-bottom
ax_r2.set_xlabel(first_title, fontsize=14)

ax_rmse.banh(y_pos, second_list, xerr=second_error_list, align='center')
ax_rmse.set_yticks(y_pos, labels=features, fontsize=14)
ax_rmse.invert_yaxis() # labels read top-to-bottom
ax_rmse.set_xlabel(second_title, fontsize=14)

ax_mae.banh(y_pos, third_list, xerr=third_error_list, align='center')
ax_mae.set_yticks(y_pos, labels=features, fontsize=14)
ax_mae.invert_yaxis() # labels read top-to-bottom
ax_mae.set_xlabel(third_title, fontsize=14)

```

```

[ ]: aic_sorted_dataframe = (mean_aic_by_features_dataframe.
    ↪sort_values(by=mean_aic_by_features_dataframe.index[0], axis=1,
    ↪ascending=True))
other_features = aic_sorted_dataframe.columns

mean_mse = mean_mse_by_features_dataframe[aic_sorted_dataframe.columns].values.
    ↪flatten()
error_mse = std_mse_by_features_dataframe[aic_sorted_dataframe.columns].values.
    ↪flatten()

mean_aic = mean_aic_by_features_dataframe[aic_sorted_dataframe.columns].values.
    ↪flatten()
error_aic = std_aic_by_features_dataframe[aic_sorted_dataframe.columns].values.
    ↪flatten()

mean_bic = mean_bic_by_features_dataframe[aic_sorted_dataframe.columns].values.
    ↪flatten()
error_bic = std_bic_by_features_dataframe[aic_sorted_dataframe.columns].values.
    ↪flatten()

plot_other_metrics(features=other_features[:10], first_list=mean_aic[:10],
    ↪first_error_list=error_aic[:10], first_title='AIC',
        second_list=mean_bic[:10], second_error_list=error_bic[:10],
    ↪second_title='BIC',
        third_list=mean_mse[:10], third_error_list=error_mse[:10],
    ↪third_title='MSE')

```

