Análise de Poluição Atmosférica Utilizando Modelos de Sensoriamento Virtual

Gabriel Oliveira Campos[1], **Felipe Domingos da Cunha**[2], **Leandro Aparecido Villas**[1]

[1]Instituto de Computação
Universidade Estadual de Campinas (Unicamp)
SP – Brasil

g265146@dac.unicamp.br,leandro@ic.unicamp.br

[2]Instituto de Ciências Exatas e Informática
Pontíficia Universidade Católica de Minas Gerais (PUC-MG)
MG – Brasil

felipe@pucminas.br

Abstract. *Virtual sensing models have been used to generate synthetic data and provide complementary information. This approach is important for analyzing aspects for which there are no physical sensors. One of the problems in the literature on air pollution analysis is the financial difficulty for the purchase of sensors, the loss of information, and the storage of invalid data. Therefore, this article analyzes several factors that contribute to air pollution and presents the comparison of different models for creating a virtual sensor for carbon monoxide. Finally, the Boosted Trees model produced better results, with an average of 5,078 RMSE, in the generation of synthetic data for each analyzed city.*

Resumo. *Modelos de sensoriamento virtuais têm sido utilizados para gerar dados sintéticos e fornecer informações complementares. Essa abordagem é importante para analisar aspectos para os quais não existem sensores físicos. Alguns dos problemas na literatura em análise de poluição atmosférica estão ligados à dificuldade financeira para a compra de sensores, à perda de informação e ao armazenamento de dados inválidos. Portanto, esse artigo analisa diversos fatores que contribuem para a poluição atmosférica e apresenta a comparação de diferentes modelos para criação de um sensor virtual para o monóxido de carbono. Por fim, o modelo de Boosted Trees produziu os melhores resultados, com média de 5.078 de RMSE, na geração dos dados sintéticos de cada cidade analisada.*

1. Introdução

O rápido crescimento da densidade populacional ao longo dos últimos anos ocasiona uma série de consequências na vida humana e na saúde ambiental. Uma delas é a qualidade do ar, a qual vem piorando gradativamente ano após ano [Moran 2020][1]. Além disso, a

[1]Disponível em: https://www.ecodebate.com.br/2020/01/21/poluicao-por-incendios-florestais-no-brasil-agrava-qualidade-do-ar-em-cidades-distantes/

grande emissão de gases provenientes das indústrias e o aumento do tráfego de veículos nas rodovias vem contribuindo também para o aumento da concentração de poluentes na atmosfera. Esse aumento influencia negativamente a qualidade de vida dentro dos centros urbanos e contribui para o agravamento de doenças respiratórias e danos ao meio ambiente.

Dentre os vários poluentes que são emitidos diariamente na natureza, seis são considerados como os mais prejudiciais à saúde humana, entre eles se encontram: o dióxido de enxofre (SO_2), o dióxido de nitrogênio (NO_2), o monóxido de carbono (CO), o ozônio (O_3), o material particulado de até 2.5 micrômetros de tamanho ($PM_{2.5}$) e o material particulado de 2.5 até 10 micrômetros de tamanho (PM_{10}) [Guarieiro et al. 2011]. A grande concentração desses poluentes no ar acarreta uma série de problemas para a saúde humana, além de graves consequências ao meio ambiente. Dentre esses problemas, estão inclusos: problemas no coração, problemas na visão, diminuição das capacidades físicas e mentais, inflamação e irritação na respiração, dificuldades respiratórias em geral, redução na capacidade do pulmão, batidas irregulares no coração e a agravação nos sintomas de asma [Amâncio and Nascimento 2012, Bucco 2010, Peres 2005, Twomey 1977].

Segundo a Organização Mundial de Saúde (OMS), a estimativa é que ocorram anualmente 4.2 milhões de mortes prematuras atribuídas à poluição do ar no mundo. A OMS também estima que a poluição do ar tenha sido responsável, no ano de 2016, por aproximadamente 58% de mortes prematuras por doenças cerebrovasculares (DCV) e doenças isquêmicas do coração (DIC), 18% por doença pulmonar obstrutiva crônica (DPOC) e infecção respiratória aguda baixa e 6% por câncer de pulmão, traqueia e brônquios [Mendes 2019][2].

No Brasil, o custo para o tratamento da asma, somente em 2019, foi de aproximadamente R$ 500 milhões para os cofres públicos segundo o Datasus [de Jaraguá 2019][3]. Também foi constatado que o número de óbitos por Doenças Respiratórias Crônicas não Transmissíveis aumentou em 14%, passando de 38.782 em 2006 para 44.228 em 2016. Em 2018, o custo com doenças respiratórias ultrapassou R$1,3 Bilhão, e é estimado que os gastos entre 2008 e 2019 chegaram a 14 bilhões [Mendes 2019].

Idealmente, todas as informações dos níveis de poluentes na atmosfera são capturadas por sensores físicos. As leituras desses sensores atmosféricos ajudam na descoberta de poluentes com alta concentração e, consequentemente, quais são as áreas de maior risco para a saúde humana. Esse monitoramento pode guiar o governo a tomar decisões que visam proporcionar melhorias na qualidade do ar com a diminuição da emissão de poluentes. Por exemplo, ações que motivem o uso do transporte público e bicicletas apresentam bons resultados em grandes centros urbanos. Em contrapartida, cidades com pouco desenvolvimento ou com escassez de recursos possuem grande dificuldade para a instalação destes sensores, o que inviabiliza a execução de um bom monitoramento da concentração desses poluentes. Neste contexto, uma possível solução é a utilização de sensores virtuais.

[2]Disponível em: `https://www.gov.br/saude/pt-br/assuntos/noticias/mortes-devido-a-poluicao-aumentam-14-em-dez-anos-no-brasil`

[3]Disponível em: `https://heja.org.br/noticias/6-milhoes-de-brasileros-sofrem-com-asma/`

Técnicas de sensoriamento virtual, também chamado de sensoriamento suave, sensoriamento proxy, sensoriamento inferencial ou sensor substituto, são utilizados para fornecer alternativas viáveis e econômicas para instrumentos de medição física caros ou impraticáveis [Zaidan et al. 2020]. Um sistema de detecção virtual usa informações disponíveis de outras medições e parâmetros de processo para calcular uma estimativa da quantidade de interesse [Liu et al. 2009].

A dificuldade financeira para a compra de sensores, a perda de informação e armazenamento de dados inválidos são um grande problema na literatura em contexto de qualidade do ar [Samal et al. 2019]. Isso se deve ao fato de que, para a população com a saúde mais sensível, a distribuição de dados corretos pode ajudar diariamente no monitoramento da qualidade do ar. Neste contexto, esse artigo tem o propósito de explorar a criação de um sensor virtual para o monóxido de carbono, um dos poluentes mais maléficos conhecidos na atmosfera, utilizando métodos de aprendizado de máquina e aprendizado profundo, com o intuito de atacar os problemas apresentados pelos sensores físicos. Aliado a isso, será feita a análise da importância das variáveis utilizadas, a fim de encontrar quais são os fatores mais impactantes na predição do poluente atmosférico. Será executado também uma análise de correlação entre as variáveis das diferentes localidades analisadas, para identificar se existe um comportamento em comum dessas variáveis nas diferentes cidades.

Este artigo está organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados; A Seção 3 descreve a metodologia do artigo; A Seção 4 detalha a avaliação dos resultados obtidos; Por fim, a Seção 5 apresenta a conclusão dessa pesquisa além de descrever alguns trabalhos futuros.

2. Trabalhos Relacionados

Trabalhos mais recentes já têm começado a investigar sensores virtuais para substituir a ausência de dados de sensores físicos. [Zaidan et al. 2020] desenvolveram um *Low-cost Sensor* (LCS) integrado com modelos de aprendizado de máquina baseado em modelos de calibração e sensores virtuais. Porém, para o dióxido de carbono (CO_2), não pode ser realizado o LCS, então os autores utilizaram de sensores virtuais para criação dos dados do CO_2 e também para o carbono negro. No artigo, os dados dos poluentes dos autores são obtidos através de duas estações de monitoramento somente em Helsinki, e para os dados faltantes e sensores com poucos valores foi implementada a estratégia da utilização de sensores virtuais para substituir esses dados.

Já em [Campolina et al. 2017], os autores investigam a aplicação de sensores virtuais em veículos para identificação do comportamento do motorista e identificação do motorista. Esse trabalho utiliza meios de medição física já presentes nos veículos e sensores também presentes nos celulares. Esse trabalho também apresenta as características necessárias para poder se criar um sensor virtual como a presença de múltiplos sensores físicos e a alta correlação entre os dados.

Em [Nguyen and Hoang 2020], os autores propõem um *software-defined virtual sensor* para solucionar o problema da rigidez dos sensores físicos presentes em dispositivos de *Internet of Things* (IoT). Com isso eles utilizam desses tipos de sensores virtuais que permitem a programabilidade de dispositivos de IoT que estejam de acordo com as aplicações IoT sob demanda. O artigo também demonstra como é feito o uso desses

sensores no cenário de IoT sob demanda.

Por fim, em [Oehmcke et al. 2018], os autores exploraram a utilização de combinação de métodos de *Deep Learning* para a criação de sensores virtuais que possam substituir dados faltantes em sensores marinhos como velocidade do vento, pressão, temperatura, condutividade e direção. Os resultados obtidos pelos autores foram melhores que o da linha base do trabalho, que foi utilizando somente o *bidirectional long short-term memory* (BLSTM). Como conclusão os autores confirmam que a utilização de sensores virtuais para substituir dados faltantes pode ser altamente benéfico quando utilizado de um bom modelo de aprendizado.

Devido à falta de modelos de sensoriamento virtual para análise de poluição atmosférica, esse trabalho aborda um caso de utilização diferente das aplicações dos sensores virtuais citados previamente. A Tabela 1 apresenta um resumo sobre os principais artigos relacionados e seus casos de utilização em comparação ao modelo proposto por esse artigo.

Artigo	Caso de utilização	Dados obtidos em
[Zaidan et al. 2020]	*Low-Cost-Sensor*	Helsinki
[Campolina et al. 2017]	Sensores veiculares	Veículos e Celulares
[Nguyen and Hoang 2020]	Sensores de IoT	Dispositivos IoT
[Oehmcke et al. 2018]	Sensores marinhos	Mar norte entre as ilhas Langeoog e Spiekeroog
Modelo proposto	Sensores de poluição atmosférica	Diversas cidades

Tabela 1. Principais características apresentadas pelos artigos sobre sensores virtuais.

3. Metodologia

Esta seção apresenta uma visão geral da utilização de sensores virtuais. Sendo assim, será demonstrado como são aplicados sensores virtuais para poluentes atmosféricos, como foi feito a aquisição e o pré-processamento dos dados e apresenta algumas análises estatísticas nos dados dos poluentes.

3.1. Modelos de Sensoriamento Virtual

Sensores virtuais são feitos a partir de um conjunto de sensores físicos, os quais são utilizados para o aprendizado do modelo de sensoriamento virtual. Para a a criação desse meio de medição virtual que analisa a concentração de poluentes, é utilizado de um ou mais dados de sensores físicos. Esses dados são obtidos a partir de sensores posicionados na cidade que analisam esses poluentes atmosféricos. Com a junção de cada uma dessas informações é montado uma base de dados com os valores da concentração desses poluentes.

Com a quantidade de dados nulos, causados pela perda de conexão que esses sensores físicos que analisam a concentração dos poluentes no ar apresentam, deve ser

realizado um pré-processamento nos dados para corrigir esse problema. Essa técnica pode ser realizada com métodos de tratamento de valores faltantes para esses dados nulos, ou com técnicas de detecção de outliers para os erros de gravação de dados.

3.2. Definição do Trabalho

Como foi descrito, o maior desafio ao se utilizar sensores físicos que analisam a concentração dos poluentes atmosféricos é a inviabilidade financeira para o estabelecimento desta rede de monitoramento. Aliado a isso, pode ocorrer também a perda de dados provocada por dificuldades na comunicação e algumas leituras incorretas pelos sensores.

Por esse motivo, cidades emergentes, ou até mesmo algumas grandes cidades, que não possuem toda infraestrutura necessária para o monitoramento, podem fazer o uso de modelos de sensoriamento virtual. Isso porque, esses modelos aproveitam da infraestrutura já existente e a complementam fazendo o enriquecimento do monitoramento através da predição de novos dados.

Outro uso para os sensores virtuais é a substituição de valores perdidos nos sensores físicos de um poluente. Os quais podem ocorrer quando este não coleta ou disponibiliza seus dados.

A Figura 1 mostra um fluxograma do uso de um sensor virtual para análise da concentração dos poluentes quando ocorre perda de dados por um dos sensores. Seguindo o processo, o sensor virtual é ativado quando o servidor não recebe os dados de uma ou mais leituras, ou quando o sensor físico não está presente nesta rede de monitoramento. Com isso, os dados não coletados, são preditos por meio de um sensor virtual através das informações fornecidas pelos equipamentos presentes na rede de monitoramento e dos dados meteorológicos. Por fim, as informações de todos os sensores (físicos e virtuais), são armazenadas no histórico presente no servidor.

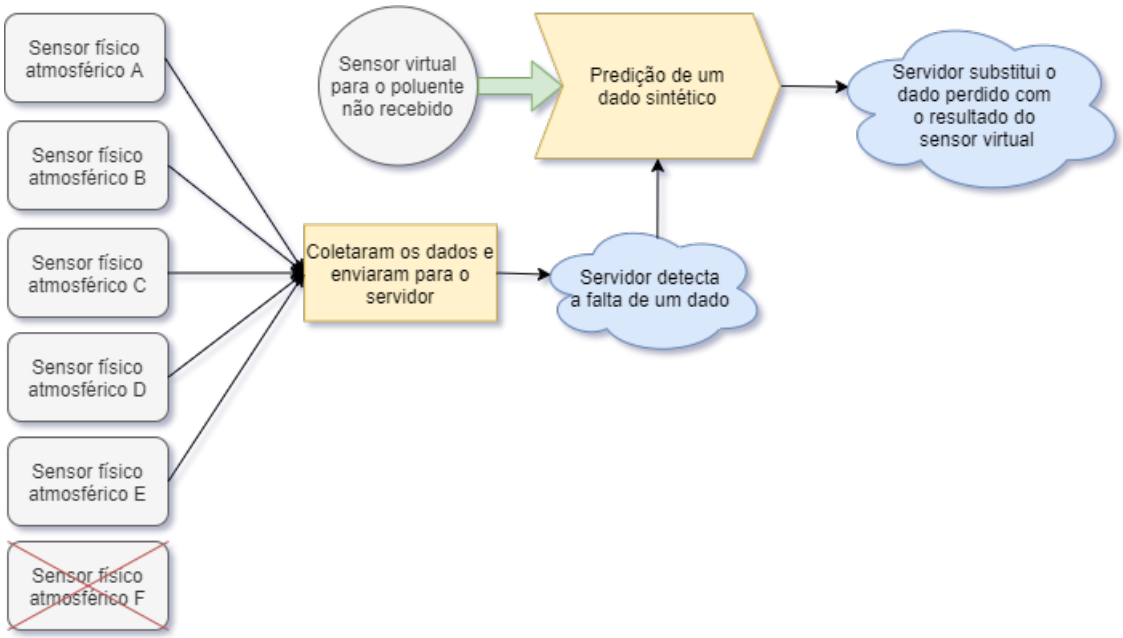

Figura 1. Fluxograma da utilização dos sensores virtuais dos poluentes atmosféricos.

3.3. Aquisição e pré-processamento dos dados

A avaliação da proposta será baseada nas cidades as quais já possuem uma rede de monitoramento completa com todos os sensores físicos necessários. Esses dados dos sensores foram adquiridos através do *The World Air Quality Project* (AQICN) [Kumar 2015], que é um projeto que analisa todos os sites de governos que possuem essas concentrações atmosféricas. É disponibilizado no AQICN dados de 2014 até o dia atual para cada uma das cidades. Vale ressaltar que a janela de tempo dos dados é diária, então os dados são sobre as médias diárias da concentração dos poluentes já transformados no *United States Enviroment Protection Agency Air Quality Index* (US EPA AQI), formato do índice da qualidade do ar nos Estados Unidos.

Das cidades disponibilizadas no *The World Air Quality Project*, poucas possuíam uma rede de monitoramento completo, com todos os sensores físicos para os seis poluentes primários da atmosfera. Portanto, foram selecionadas algumas grandes cidades que possuem essa rede de monitoramento completa para utilizar na base de dados. As cidades utilizadas na base são: Amsterdan, Beijing, Busan, Copenhague, Londres, Los Angeles, São Paulo, Shanghai e Tokyo.

Porém, nos dados disponibilizados pelos sites dos governos pode-se encontrar muitos dados nulos, que são as informações perdidas. No total, dos 150.000 dados que foram coletados, foi encontrado aproximadamente mais de 20.650 dados nulos, apresentando assim as informações perdidas nos dados coletados. Esses dados serão removidos das bases de dados utilizadas para o treino do modelo. Aliado a eles, foi utilizado um *web crawler* [Kausar et al. 2013] para adicionar ao modelo dados meteorológicos[4], os quais são recebidos por meio da API *CustomWeather* das diferentes cidades utilizadas. Foi adicionado também a informação sobre a estação do ano de cada uma das cidades utilizadas no modelo, devido ao fato que a sazonalidade das estações afeta diretamente na qualidade do ar.

Além da remoção dos dados nulos, foi utilizado de um modelo de remoção de outliers, utilizando o *InterQuartile Range* (IQR), que consiste em um modelo que seleciona todos os dados entre o 25^o e 75^o percentil. Isso se deve a uma pequena quantidade de dados que demonstram valores fora da realidade para alguma dessas variáveis meteorológicas armazenadas pelo *CustomWeather*.

Um fator que também será analisado por esse artigo é o custo de processamento de cada um dos modelos de aprendizado, pois somente ter o melhor resultado não oferece um melhor modelo de sensoriamento virtual. Ter um bom custo de processamento para substituir sensores físicos não presentes nas redes de monitoramento também possui uma grande importância de análise nesse artigo.

3.4. Análise Estatística dos Poluentes

Para analisar a correlação dos dados dos poluentes atmosféricos, será necessário explorar a correlação linear entre as variáveis nessa base de dados. Por tal motivo, foi utilizado o coeficiente de Pearson [Benesty et al. 2009], pois ele avalia a relação linear entre duas variáveis contínuas. Por fim, este será o método abordado neste trabalho, uma vez que

[4]Disponível em: `https://www.timeanddate.com/`

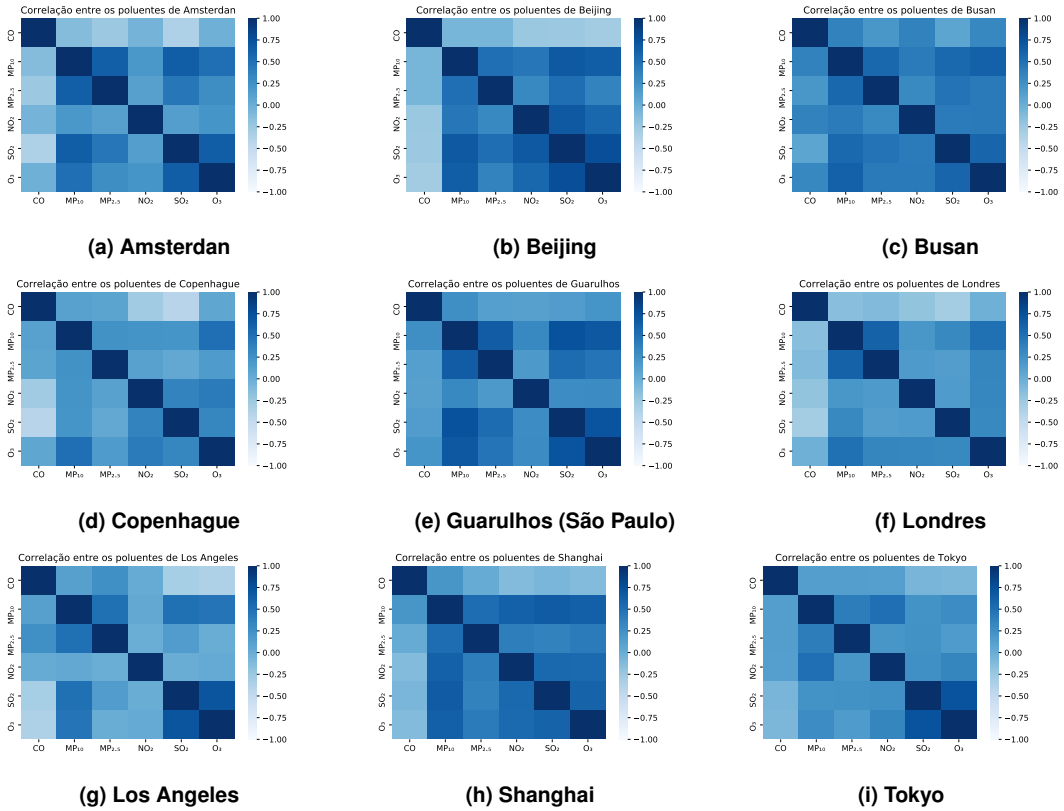

Figura 2. Correlação de Pearson entre os poluentes para cada cidade analisada.

ele apresenta características apropriadas para a avaliação dos resultados, dado que outros métodos, como o Spearman, que avalia a relação monotônica entre duas variáveis contínuas ou ordinais não é o método mais apropriado nesse caso.

Na Figura 2 pode-se ver que o comportamento dos poluentes é bem próximo entre as diferentes cidades, isso se deve a proximidade da coloração do mapa de calor, mostrando assim que os dados das diferentes cidades possuem um comportamento bem aproximado (possuem valores próximos), como pode-se ver através dos mapas de calor. Portanto, pode-se dizer então que esse modelo é replicável, isso se deve à essa proximidade da correlação entre as diferentes cidades.

4. Avaliação do Modelo e Resultados

Esta seção apresenta qual foi o ambiente de processamento e os resultados obtidos pelos modelos de sensoriamento virtual. Junto a eles, também são apresentadas as importâncias de cada uma das variáveis utilizadas no modelo.

4.1. Ambientes de processamento

Nesse trabalho é analisado diferentes ambientes para a simulação de um sensor virtual. Sendo assim, uma alta acurácia e precisão do modelo sozinhos não demonstram que o modelo de sensoriamento é eficiente: o tempo de processamento também tem grande importância na criação desse tipo de arquitetura. Isso porque, o atraso de processamento pode provocar a disponibilização de dados não vigentes sobre a concentração dos poluentes atmosféricos. Portanto serão analisados diferentes tipos de ambientes para o pro-

cessamento dos sensores virtuais, seja eles em nuvem ou em algum tipo de dispositivo de baixo poder de processamento, isso para obter a melhor combinação entre tempo de processamento e resultado.

Portanto, foram utilizados três ambientes de processamento para simulação de um sensor virtual. O primeiro é através do Google Colab, uma nuvem disponível gratuitamente pelo Google com especificações de 13 GB de memória RAM e 100 GB de espaço gratuito. O segundo é através de um RaspBerry Pi de 3^a versão, com 1 GB de memória RAM. Por último temos uma máquina física, um Alienware 17R4 com 16 GB de RAM.

Devido a utilização de alguns métodos de aprendizado profundo, o RaspBerry Pi de 3^a versão não consegue atualizar o *tensorflow* para a versão necessária pra execução. Por isso, alguns dos métodos não podem ser analisados através dele. Entretanto, RaspBerry Pi com versões superiores à 3^a pode executar os métodos normalmente.

4.2. Modelos Utilizados

Para criação dos sensores virtuais é necessário a utilização de métodos de aprendizado de máquina ou métodos de aprendizado profundo. Portanto, no modelo proposto foi utilizado de diversos tipos diferentes de métodos de aprendizado para verificação do melhor modelo de sensoriamento virtual do monóxido de carbono.

Os modelos de aprendizado de máquina utilizados nesse trabalho são: Regressão Linear (LR), Regressão Linear com regularização L1 (Lasso), Regressão Linear com regularização L2 (Ridge), Árvores de Decisão (DT), *k-Nearest Neighboors* (KNN), Vetores de Suporte a Regressão (SVR), Vetores de Suporte a Regressão com kernel RBF (RBF), Florestas Randômicas (RF), *Gradient Boosting Machine* (GBM) [Bonaccorso 2017].

Já os modelos de aprendizado profundo utilizadas nesse trabalho são: *Deep Belief Neural Networks* (DBN), *Boosted Trees* (BT), Perceptron Multicamadas (MLP), *Deep Neural Network (Linear Combined Regressor)* (DNN), *Gated Recurrent Units* (GRU), *Long-Short Term Memory* (LSTM), *Linear Estimator* (LE), Modelo Sequencial de 4 camadas (Seq) [LeCun et al. 2015, Hua et al. 2015].

Nesse trabalho foi selecionado uma diversa variedade de modelos de aprendizado de máquina e profundo. Isso se deve a falta de trabalhos realizados utilizando modelos de sensoriamento virtual para análise de poluição atmosférica. Portanto, a seleção desses modelos retratam a grande diversidade de métodos presentes na literatura que são utilizados e possuem resultados relevantes em diversos contextos.

4.3. Treino e Validação dos Modelos

Na etapa de treino e validação, para os métodos de aprendizado de máquina, foi utilizado o *KFolds* com 5 *Folds* (80% para treino e 20% para validação) e para cada um desses *folds*, foram feitas 10 repetições na busca de hiper parâmetros de cada um dos modelos. Foi selecionado o *KFolds* nos métodos de aprendizado de máquina devido a importância da combinação da execução do método várias vezes com diferentes bases para treino e validação com a busca por hiper parâmetros feitas em cada uma dos *folds*, isso auxilia o modelo a encontrar os melhores parâmetros possíveis para a execução final. Após essas repetições, foi computado uma média desses valores para se obter e minimizar o resultado médio de cada um dos *folds* através da raiz quadrada do erro médio (RMSE) do modelo.

Já para os métodos de aprendizado profundo, foi feita a divisão em 70% para treino, 15% para validação e 15% para teste. No caso dos métodos de aprendizado profundo, foi utilizado o *hold-out* para treino, validação e teste, pois na aplicação dos métodos é realizado uma análise através de épocas, onde é feito o refinamento do modelo para se obter melhores resultados. Nos modelos de aprendizado profundo também foi analisado o RMSE como métrica de resultado.

A Tabela 2 mostra os melhores parâmetros encontrados para cada um dos modelos.

Modelos	BoostedTrees	DNN	LSTM	GRU
Parâmetros (Deep Learning)	Batches = 32 Epochs = 10000 Learning rate = 0.1 n_trees = 100 max_depth = 6	Batches = 32 Epochs = 10000 Optimizer = Adam Activation = Relu	Batches = 32 Epochs = 200 Optimizer = Adam Activation = Relu	Batches = 32 Epochs = 200 Optimizer = Adam Activation = Relu
	Sequential	LinearEstimator	MLP	
	Batches = 32 Epochs = 200 Optimizer = Adam Activation = Relu	Batches = 32 Epochs = 10000 Optimizer = Adam	hidden_layer_sizes = 20 max_iter = 1000	
Modelos	**SVR**	**Ridge**	**Lasso**	**Random Forest**
Parâmetros (Machine Learning)	epsilon = 0.3 C = 5.25	alpha = -1.32	alpha = -1.79	n_estimators = 1000 max_features = 5
	DT	GBM	k-NN	
	ccp_alpha = 0.037	n_estimators = 79 learning_rate = 0.27 max_depth = 2	n_neighbors = 11	

Tabela 2. Parâmetros utilizados em cada um dos modelos

Outro caso de estudo foi realizado, onde ao invés de utilizar de maneira randômica o treino, validação e teste, foi utilizado dos dados de 8 das 9 cidades presentes na base para fazer a predição dos últimos 180 dias dessa 9^a cidade. Esse caso de estudo serve para simular como seria a aplicação real do modelo nessas cidades.

4.4. Resultados

Conforme a Figura 3, é possível analisar que o tempo de processamento é maior para o DBN e RBF quando analisados em todos os ambientes. Sendo assim, mesmo possuindo bons resultados esses modelos são classificados de maneira ruim, devido ao seu alto tempo de processamento. Outro fator a ser analisado é que o ambiente em nuvem (Google Colab) é o melhor local para implementar o modelo, devido ao seu custo/benefício.

Do modelo generalizado de treino e teste pode-se constatar que vários métodos de aprendizado profundo apresentaram os melhores resultados segundo a Figura 4, sendo eles os métodos de LSTM, GRU, *Boosted Trees*, Sequencial e *Deep Belief*. Porém, devido ao alto tempo de processamento do *Deep Belief*, esse método e o RBF já estão sendo desconsiderados em outras análises presentes nesse artigo. Dos métodos apresentados pelo modelo generalizado, será apresentado na Figura 5 as últimas 90 predições para esse modelo geral utilizando o método com um dos melhores resultados, o *Boosted Trees*, para demonstrar visualmente a proximidade dos dados preditos com os dados reais. O motivo da seleção do *Boosted trees* para essas verificações foi devido ao seu menor tempo de processamento nos ambientes analisados.

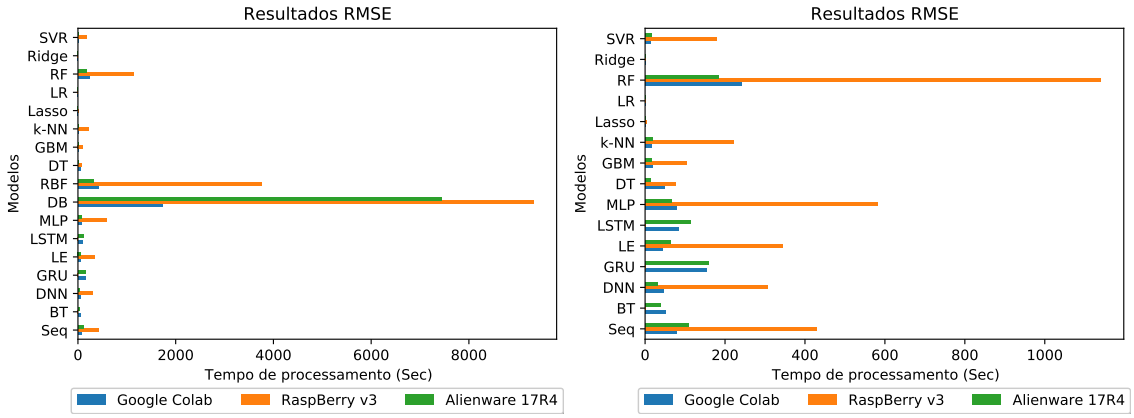

Figura 3. Tempo de processamento com e sem os métodos do *Deep Belief* e RBF

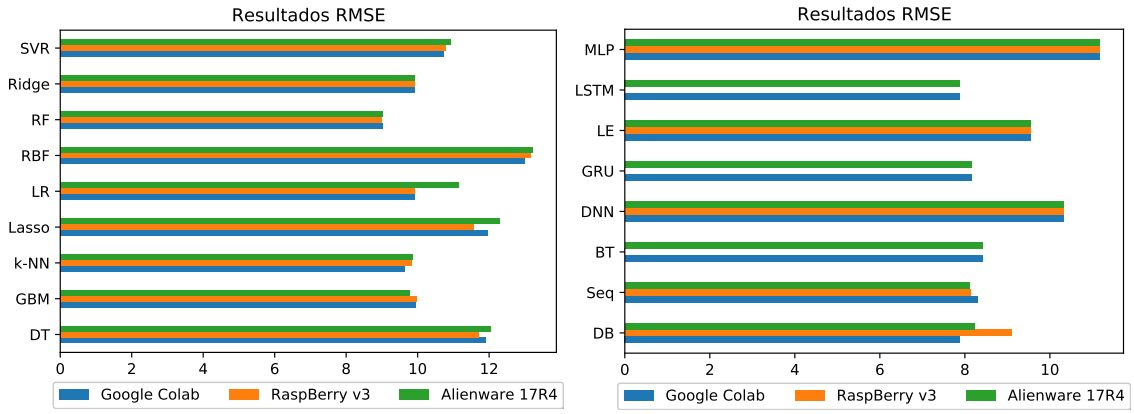

Figura 4. RMSE do modelo generalizado.

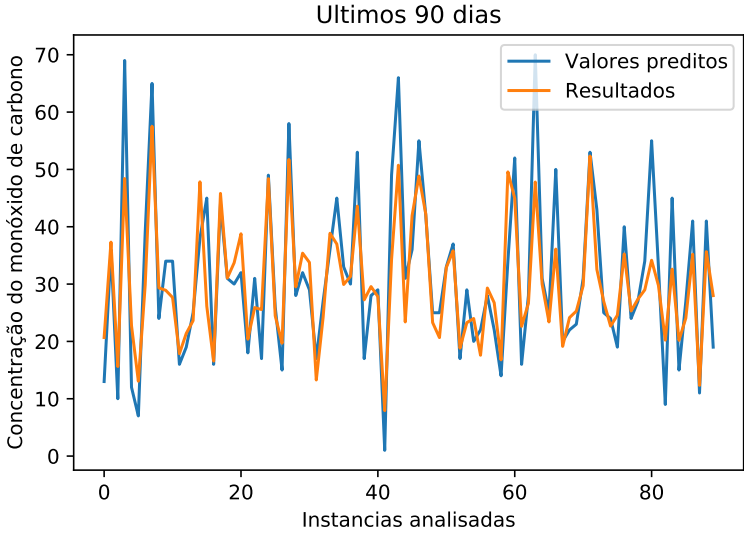

Figura 5. Valores preditos do modelo generalizado

Do modelo de cidades específicas, temos ainda que os melhores resultados são provenientes também do *Boosted Trees*, demonstrando resultados muito próximos dos valores reais conforme as Figuras 6 e 7, onde podemos ver que o menor valor do RMSE e a menor variação de RMSE através das cidades foi proveniente desse modelo. Além

disso é apresentado os últimos 90 dados previstos de cada uma das cidades utilizando ele, conforme a Figura 8, para demonstrar a proximidade dos dados previstos com os dados reais.

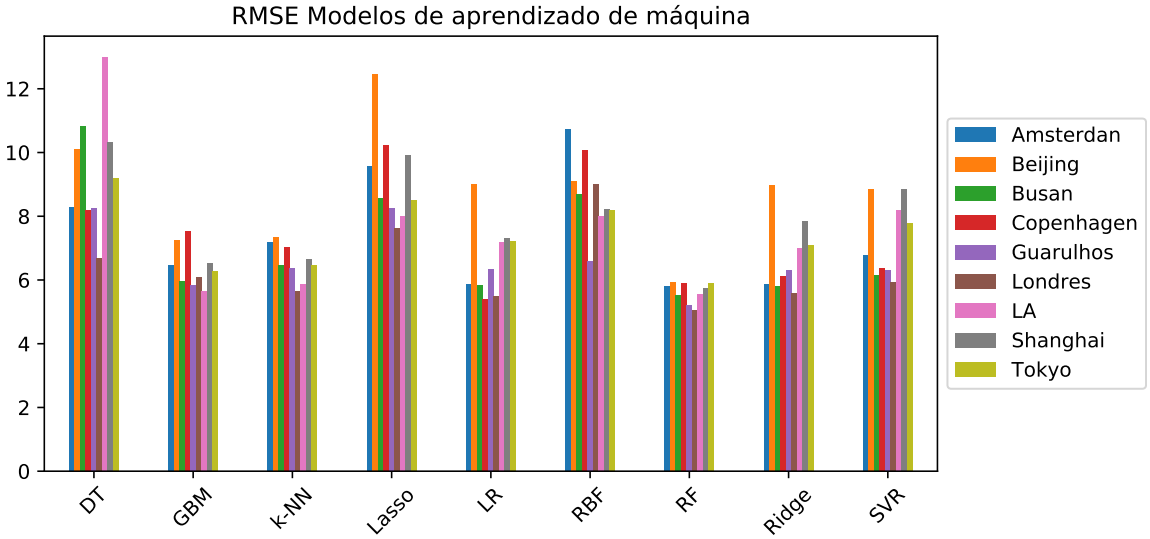

Figura 6. RMSE dos modelos específicos utilizando aprendizado de máquina.

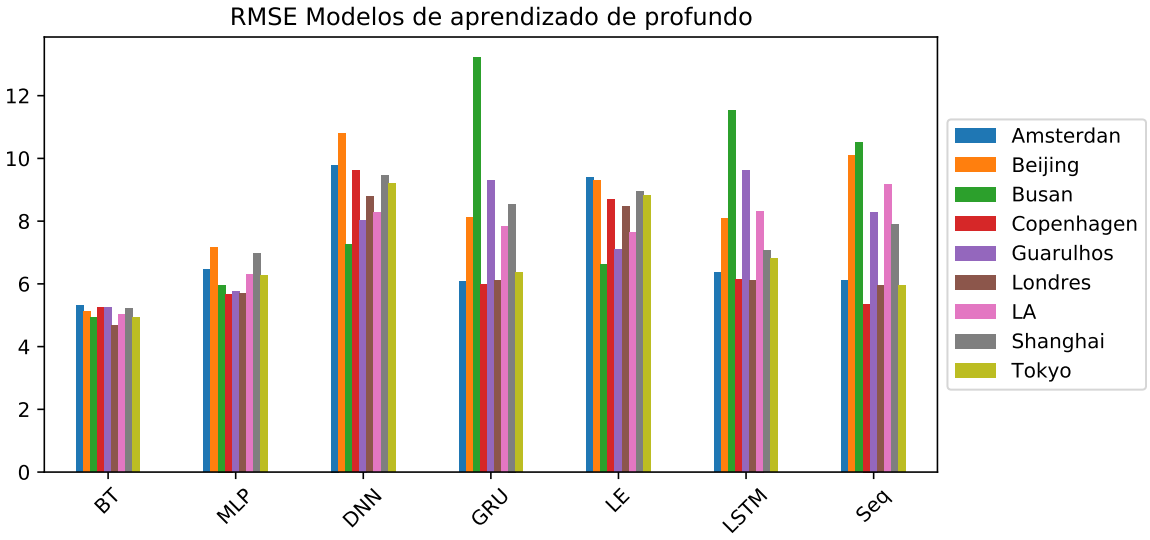

Figura 7. RMSE do modelo específico utilizando aprendizado de profundo.

Por último, pode ser analisado na Figura 9 as variáveis que mais impactaram no modelo de cidades específicas através do *Shapley Values* [Kalai and Samet 1987] do melhor método analisado, demonstrando assim os itens que mais influenciaram na predição desse dado sintético do monóxido de carbono. Isso serve principalmente para auxiliar e complementar o compreendimento do aumento da concentração desse poluente. E conforme demonstrado nas figuras, os mesmos fatores impactam na predição do monóxido de carbono nessas diferentes localidades.

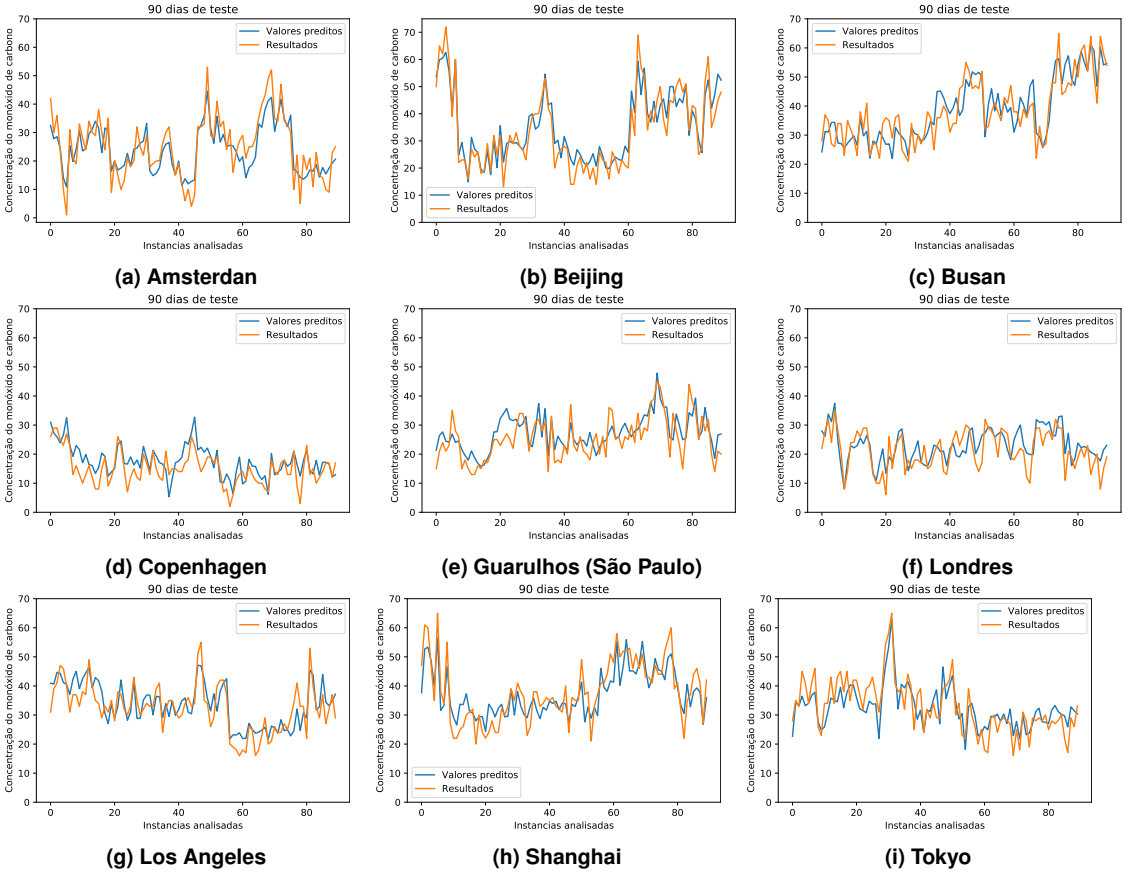

Figura 8. Simulação do modelo *Boosted Trees* em aplicações reais

5. Considerações Finais

Este trabalho apresentou modelos de sensoriamento virtual para o poluente atmosférico do monóxido de carbono, uma jeito viável, econômico e eficiente para complementar redes de monitoramento que não possuem uma infraestrutura completa para analisar todos os poluentes atmosféricos necessários dos grandes centros urbanos. Análises estatísticas através do coeficiente de Pearson revelam que os poluentes atmosféricos têm um comportamento semelhante nos diversos grandes centros urbanos analisados nesse trabalho, mostrando assim uma compatibilidade em se utilizar esse modelo em diferentes centros urbanos não analisados nesse artigo, isso sem ter um déficit tão grande na acurácia e precisão do modelo.

Outro fator analisado foi a raiz quadrada do erro médio do modelo generalizado desse trabalho, demonstrando que o modelo de *Boosted Trees* possui um melhor resultado, devido ao seu baixo tempo de processamento ao se criar um sensor virtual para poluição atmosférica nesse contexto. Aliado a isso também foi obtido resultados dos modelos específicos, tentando mostrar a utilização desses modelos mais próximo de contextos reais, e novamente o *Boosted Trees* demonstrou um melhor resultado para todas as cidades.

Como trabalhos futuros ainda serão investigados se os outros poluentes atmosféricos também produzem bons resultados utilizando esses modelos. Também será investigado se o modelo de *Boosted Trees* continua produzindo o melhor resultado. Além disso será avaliado a criação de um sensor virtual para dados com uma menor janela de

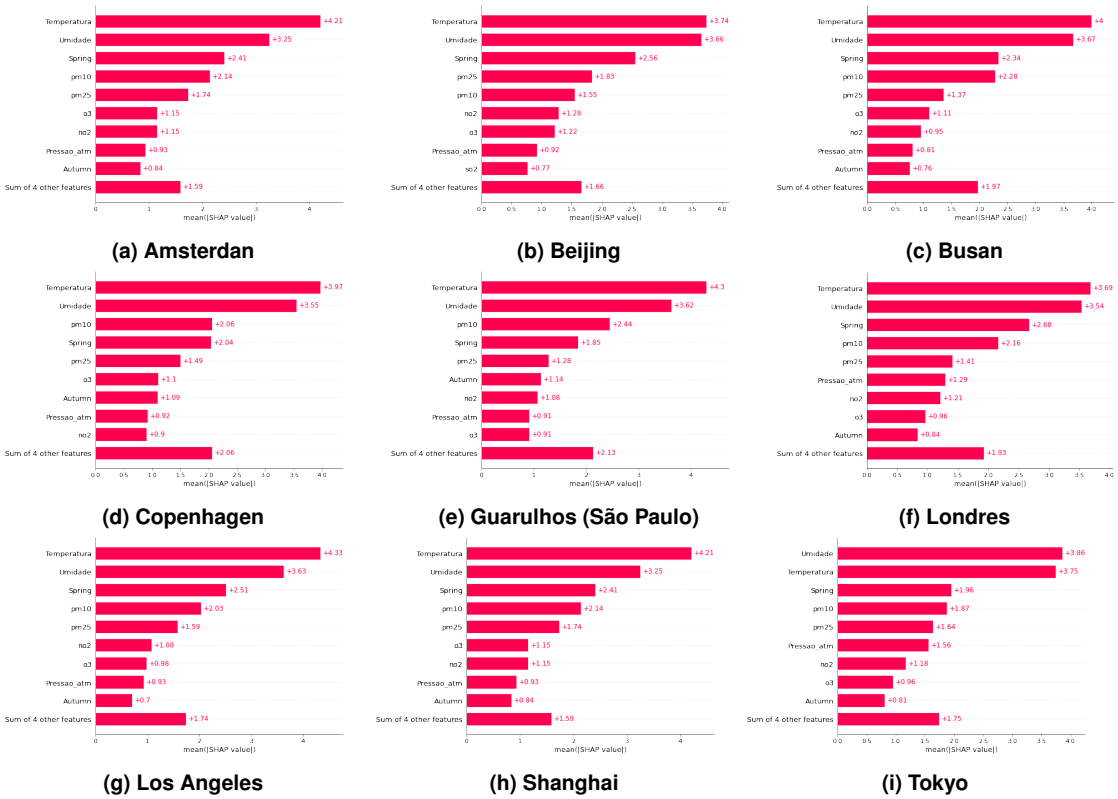

Figura 9. *Shapley Values* de cada cidade analisada pelo *Boosted Trees*

tempo, ou seja, dados horários ao invés de dados diários.

6. Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001

Referências

Amâncio, C. T. and Nascimento, L. F. C. (2012). Asma e poluentes ambientais: um estudo de séries temporais. *Revista da Associação Médica Brasileira*, 58(3):302–307.

Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer.

Bonaccorso, G. (2017). *Machine learning algorithms*. Packt Publishing Ltd.

Bucco, M. V. S. (2010). Construção e testes de validação de amostradores passivos para dióxido de nitrogênio e ozônio.

Campolina, A. B., Rettore, P. H. L., Machado, M. D. V., and Loureiro, A. A. (2017). On the design of vehicular virtual sensors. In *2017 13th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pages 134–141. IEEE.

de Jaraguá, H. E. (2019). 6,4 milhões de brasileiros acima de 18 anos sofrem com asma. [Online; accessed 16-february-2021].

Guarieiro, L. L., Vasconcellos, P. C., and Solci, M. C. (2011). Poluentes atmosféricos provenientes da queima de combustíveis fósseis e biocombustíveis: uma breve revisão. *Revista Virtual de Química*, 3(5):434–445.

Hua, Y., Guo, J., and Zhao, H. (2015). Deep belief networks and deep learning. In *Proceedings of 2015 International Conference on Intelligent Computing and Internet of Things*, pages 1–4.

Kalai, E. and Samet, D. (1987). On weighted shapley values. *International journal of game theory*, 16(3):205–222.

Kausar, M. A., Dhaka, V., and Singh, S. K. (2013). Web crawler: a review. *International Journal of Computer Applications*, 63(2).

Kumar, M. (2015). Real time air pollution map of the world: The world air quality index project.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Liu, L., Kuo, S. M., and Zhou, M. (2009). Virtual sensing techniques and their applications. In *2009 International Conference on Networking, Sensing and Control*, pages 31–36. IEEE.

Mendes, A. (2019). Mortes devido a poluição. [Online; accessed 13-July-2020].

Moran, T. (2020). Poluição por incêndios florestais no brasil agrava qualidade do ar em cidades distantes. [Online; accessed 17-December-2020].

Nguyen, C. and Hoang, D. (2020). Software-defined virtual sensors for provisioning iot services on demand. In *2020 5th International Conference on Computer and Communication Systems (ICCCS)*, pages 796–802. IEEE.

Oehmcke, S., Zielinski, O., and Kramer, O. (2018). Input quality aware convolutional lstm networks for virtual marine sensors. *Neurocomputing*, 275:2603–2615.

Peres, F. d. F. (2005). Meio ambiente e saúde: os efeitos fisiológicos da poluição do ar no desempenho físico-o caso do monóxido de carbono (co). *Arquivos em movimento*, 1(1):55–63.

Samal, K. K. R., Babu, K. S., Das, S. K., and Acharaya, A. (2019). Time series based air pollution forecasting using sarima and prophet model. In *Proceedings of the 2019 International Conference on Information Technology and Computer Communications*, pages 80–85.

Twomey, S. (1977). The Influence of Pollution on the Shortwave Albedo of Clouds. *Journal of the Atmospheric Sciences*, 34(7):1149–1152.

Zaidan, M. A., Motlagh, N. H., Fung, P. L., Lu, D., Timonen, H., Kuula, J., Niemi, J. V., Tarkoma, S., Petäjä, T., Kulmala, M., et al. (2020). Intelligent calibration and virtual sensing for integrated low-cost air quality sensors. *IEEE Sensors Journal*, 20(22):13638–13652.