MARCELO DÓRIA HILTNER ALMEIDA
MARCO ANTÔNIO DOS SANTOS

APLICAÇÃO DE ALGORITMOS DE CLASSIFICAÇÃO E REGRESSÃO MULTIVARIADOS NA ANÁLISE DE DADOS DE QUALIDADE DO AR

São Paulo
2020

MARCELO DÓRIA HILTNER ALMEIDA
MARCO ANTÔNIO DOS SANTOS

APLICAÇÃO DE ALGORITMOS DE CLASSIFICAÇÃO E REGRESSÃO MULTIVARIADOS NA ANÁLISE DE DADOS DE QUALIDADE DO AR

Dissertação apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do Título de Engenheiro em Engenharia Química.

Área de Concentração:

Análise de dados, poluição do ar.

Orientador:

Roberto Guardani

São Paulo
2020

RESUMO

Este trabalho estuda a viabilidade de aplicação de técnicas de análise de dados para fenômenos de poluição do ar no estado de São Paulo, com enfoque no município de Santa Gertrudes, que possui maior taxa de poluentes particulados do estado. O objetivo é compreender o perfil de poluição na cidade, que é fortemente marcado pela atividade de cerâmica, e desenvolver ferramentas que auxiliem no tratamento desses dados. Para isso, utilizam-se técnicas de regressão multivariadas, decomposição de séries históricas, análises de estatística descritiva e de componentes principais. Os poluentes foco deste trabalho foram MP10, MP2.5 e NOx, que são os maiores problemas da região.

Concluiu-se que a regressão multivariável pode ser utilizada para preenchimento de vazios nos dados (imputação) e que a decomposição da série histórica é útil para identificar dados anômalos. A análise de componentes principais reduziu em 50% a quantidade de variáveis que precisamos para descrever o sistema.

Palavras-Chave – Poluição, MP10, MP2.5, NOx,Santa Gertrudes, CETESB, regressão multivariável, machine learning, séries históricas.

ABSTRACT

This work consists in the evaluation of the applicability of data science techniques in the field of air pollution, focused mainly in the city of Santa Gertrudes, the most polluted city regarding level of particulates in the state of São Paulo, Brazil. The objective is to understand the city's unique pollution profile, strongly defined by the ceramic activity in the region, and to develop tools to support further data analysis. A few techniques were applied: multivariate regression, time series decomposition, descriptive statistics and principal component analysis. The pollutants considered in this study were: PM10, PM2.5 and NOx, which are the main pollutants observed in Santa Gertrudes.

It was concluded that multivariate regression may be used for data inputation purposes, and that time series decomposition is useful for outlier detection. The principal component analysis was able to reduce by half the number of variables needed to describe the system.

Keywords – Air pollution, PM10, PM2.5, Santa Gertrudes, NOx, CETESB, multivariate regression, machine learning, time series.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 Contexto

O problema da poluição do ar no estado de São Paulo – e em todo o mundo – vem se agravando há algumas décadas, tornando-se um dos principais problemas na capital e em outros pontos do estado. A poluição é um problema grave pois induz à incidência de doenças respiratórias e o número de mortes por câncer. Em escala global, a Organização Mundial da Saúde (OMS) atribui mais de 4,2 milhões de mortes ao problema da poluição do ar.

O efeito da poluição do ar também afeta zonas rurais. A Agência de Proteção Ambiental dos Estados Unidos (EPA,2006) estimou perdas agrícolas anuais da ordem de 500 milhões de dólares causadas pelo ozônio, sem incluir os danos à folhagens de árvores e outras plantas, que afetam a paisagem das cidades, áreas de recreação, parques urbanos e áreas de vegetação natural.

A qualidade do ar que se respira é diretamente ligada à distribuição e à intensidade das emissões de poluentes atmosféricos de origem veicular e industrial. Enquanto emissões veiculares se destacam nas grandes cidades, atividades industriais específicas marcam fortemente a qualidade do ar em regiões específicas.

Esse é o caso de Santa Gertrudes, pequena cidade na região de Rio Claro, a algumas horas da capital. Santa Gertrudes é, atualmente, o município com maior taxa de material particulado do estado, embora possua somente 25 mil habitantes. Trata-se de um polo cerâmico, com concentração da atividade ceramista de fabricação de pisos a partir de argila. A extração, a manipulação e o transporte da matéria prima constituem a principal fonte de poluentes.

Embora detenha, atualmente, a maior concentração de poluentes no estado, a região de Rio Claro e Santa Gertrudes ganhou mais atenção a partir de 2014, quando o problema já tinha grandes dimensões. As grandes fábricas de cerâmica atuam com filtros para os

particulados, mas a ampla movimentação de caminhões carregando argila por estradas não pavimentadas representa uma fonte significativa.

De acordo com a Companhia Ambiental do Estado de São Paulo (CETESB), algumas medidas já foram estabelecidas para diminuir o impacto da atividade na qualidade do ar. Há um plano – o Plano de Redução de Emissão de Fontes Estacionárias – que possui um conjunto de ações a serem implementadas pelos estabelecimentos, ligados à diminuição da emissão e propagação de poluentes.

Algoritmos de classificação e regressão vem se tornando cada vez mais importantes para diferentes áreas e setores da economia, seja no setor energético, na prevenção de doenças ou no mercado financeiro. As aplicações são diversas: busca de padrões a partir classificação em agrupamentos de observações semelhantes (clustering), previsão de séries históricas, regressão multivariável para estimar dados faltantes ou futuros, entre outras possibilidades.

A grande complexidade desses processos tem motivado, desde o final da década de 1990, estudos voltados à previsão da qualidade do ar na RMSP com base em modelos estatísticos multivariados, em estudos conjuntos entre a USP e a CETESB, na forma de programas específicos e em projetos maiores, com apoio da FAPESP (como o projeto de pesquisa em políticas públicas 1998/14157-7). Vários desses estudos objetivaram o ajuste de correlações e associações entre poluentes utilizando modelos de redes neurais (Guardani et al, 1999), identificação de fatores que afetam o comportamento observado em estações medidoras (Guardani et al, 2003), desenvolvimento de modelos preditivos para ozônio (Guardani e Nascimento, 2004, Borges et al, 2012) e, mais recentemente, a aplicação de técnicas de classificação, incluindo diferentes configurações de redes neurais e "random forests" (Paula e Guardani, 2016). Tais técnicas passaram a ser aplicadas pelas equipes da USP e da CETESB em vários estudos sobre qualidade do ar.

Este estudo se insere no escopo do convênio existente entre a Escola Politécnica da USP e a CETESB, coordenado pelo orientador deste trabalho.

1.2 Objetivo

O objetivo deste estudo é testar a aplicabilidade de diferentes algoritmos de classificação e discriminação, baseados em reconhecimento de padrões, na análise de variáveis e grupos de variáveis medidas em estações monitoras selecionadas, assim como entre as medidas dos níveis dos principais poluentes atmosféricos em Santa Gertrudes.

Tais algoritmos serão aplicados com o objetivo de identificar cenários específicos associados a altos níveis dos seguintes poluentes atmosféricos: NOx e material particulado. Assim, o objetivo geral do estudo é a identificação de algoritmos mais adequados para aplicação na previsão de padrões de distribuição dos valores das variáveis a serem consideradas no estudo.

Serão utilizadas técnicas estatísticas de decomposição de séries históricas, técnicas de classificação e testes de hipótese. Os algoritmos computacionais necessários ao estudo foram desenvolvidos e adaptados pelo departamento, com base em plataformas como MATLAB, SCILAB e R. A implementação deverá ser feita em Python e R.

O projeto possui como objetivo, também, a aplicação de conceitos matemáticos já aprendidos em curso, e apreensão de técnicas estatísticas e de reconhecimento de padrões que se constituem em ferramentas de ampla aplicação em engenharia.

2 MATERIAIS E MÉTODOS

Esta seção do trabalho apresenta os conceitos necessários à compreensão mais aprofundada do problema abordado, bem como as ferramentas utilizadas na modelagem e análise dos dados obtidos.

2.1 Química da troposfera e a dispersão de poluentes

A troposfera, a camada atmosférica mais próxima da superfície terrestre, possui uma alta complexidade de reações químicas e dinâmicas de dispersão, influenciadas pelos compostos emitidos na superfície terrestre, pela quantidade de vapor d'água, pela atuação da radiação e dos componentes meteorológicos.

Entre esses componentes estão os materiais particulados e os gases NOx, que serão estudados no presente trabalho. Os particulados serão divididos em dois grupos, baseados no tamanho.

MP10: São partículas inaláveis, de material sólido ou líquido que ficam suspensos no ar, com tamanho menor ou igual a dez micras. As principais fontes desse material particulado são: combustão (veicular e industrial) e aerossóis formados na atmosfera. Os principais efeitos estão ligados à deterioração da visibilidade, danos à vegetação e ao sistema respiratório.

MP2.5: Semelhante ao MP10, mas com tamanho menor, abaixo de 2.5 micras. É ainda mais perigoso por ser mais fina, alcançando os tecidos pulmonares com maior facilidade. As emissões e danos são os mesmos que o MP10.

NOx: São gases que podem ter odor forte e causar irritação. Também está associado à formação de chuva ácida. As principais emissões são combustão e atividade industrial.

Vale ressaltar que, embora trabalhemos somente com MP10 e MP2.5, os aerossóis primários (emitidos diretamente) e secundários (formados na atmosfera em conversão de gás-partícula) possuem uma distribuição de tamanho que varia de nanômetros a micro-

metros, que pode variar durante dispersão.

O movimento destes poluentes ocorre através de três princípios: transporte, dispersão e deposição. Dependendo das condições, alguns poluentes podem se deslocar em nível regional (2000 km do local de emissão), embora a maior parte se concentre no campo próximo (200 m do local de emissão) e campo urbano (até 20 km do local de emissão). Contudo, os efeitos da dispersão regional ainda são notáveis, influenciando áreas rurais com baixa emissão própria de poluentes. [1]

Esses poluentes se dispersam na atmosfera, como resultado de alguns fenômenos, ligados à turbulência. Aqui, consideram-se movimentações de ar, flutuações na velocidade do vento e difusão ligada ao gradiente de concentração. Essa turbulência pode estar ligada a diferentes fatores, mecânicos ou térmicos.

Independentemente da distância percorrida pelo poluente, a variação da sua concentração segue uma equação de conservação de massa [2]. Podemos expressá-la matematicamente como:

$$\frac{\partial \chi}{\partial t} + \frac{\partial (u\chi)}{\partial x} + \frac{\partial (v\chi)}{\partial y} + \frac{\partial (w\chi)}{\partial z} = Q + R + S \tag{2.1}$$

Onde X é a concentração em μg/m^3, u, v e w são as componentes do vento, (leste-oeste, norte-sul e vertical), dadas em m/s. Cada componente do vento pode ser representado por uma média do vento e um efeito de turbulência. Q é a taxa de emissão, R é a taxa de reação e S a taxa de remoção por deposição, todos em μg/m^3/s.

As condições meteorológicas estão diretamente ligadas à tais condições de dispersão, uma vez que temperatura, pressão atmosférica e umidade influenciam o vento, as taxas de reações e a emissão.

De forma genérica, as condições meteorológicas servem para indicar ou refletir instabilidade atmosférica. A CETESB classifica os dias como favoráveis ou desfavoráveis à dispersão de poluentes, onde os dias desfavoráveis são dias de atmosfera estável, baixa ventilação e calmaria. Esses fatores tendem a concentrar os poluentes no campo da emissão.

As duas principais formas de deposição, por gravidade (seca) ou pela chuva (molhada), conseguem garantir que os particulados não possuam tempos de residência muito longos no ar, variando de alguns dias a algumas semanas. Em termos de química da troposfera, considera-se esse período como curto, uma vez que alguns gases podem ficar até um século

antes de sua deposição.

A precipitação é um bom indicador de atmosfera favorável à dispersão de poluentes, uma vez que realiza a deposição dos particulados e é, por si só, um indicador de instabilidade atmosférica, o que aumenta a ocorrência dos fenômenos de turbulência. A temperatura também é um fator crítico no estudo dos poluentes, uma vez que dias frios normalmente causam maiores taxas de emissão de poluentes. [2]

Outro poluente importante é o ozônio. Em regiões urbanas, a concentração de ozônio (O_3) encontrada pode ser extremamente elevada, pela presença dos poluente emitidos pelos carros e outros meios. Essa concentração pode ser perigosa para a saúde, causando problemas aos pulmões.

O ozônio na troposfera é gerado de dois precursores: os componentes voláteis orgânicos (VOCs) e os óxidos de nitrogênio (NOx). Trata-se de uma reação complexa, iniciada pela reação do radical OH com compostos orgânicos. O NOx serve de catalisador, e ainda há influência da radiação solar, que pode causar a dissociação do O3 em outros compostos, como a hidroxila.. A hidroxila (OH) é o principal oxidante na troposfera, pois não reage com O_2 .

A formação de O_3 ainda possui grande complexidade por estar relacionada aos outros ciclos da troposfera. A concentração de NOx afeta diretamente o comportamento do monóxido de carbono, e a interação entre essas espécies altera o rendimento da produção de ozônio. [3]

Em baixas concentrações de NOx, a taxa de produção do O3 aumenta linearmente com o aumento de NO, e proporcionalmente à raiz quadrada da geração de HOx. Já em altas concentrações de NOx, a taxa de produção de O_3 aumenta linearmente com a concentração de monóxido de carbono e a geração de HOx. A taxa de produção de ozônio pode ser descrita por:

$$PO_3 \longrightarrow K_{HO2} + NO[HO_2][NO] \tag{2.2}$$

O HO_2 é uma espécie instável, um radical livre que reage rapidamente, e originado pela oxidação do monóxido de carbono pela hidroxila.

$$CO + OH \longrightarrow CO_2 + HO_2 \tag{2.3}$$

A concentração de NO é crítica para determinar se a atmosfera de uma região é uma

fonte de ozônio ou uma destruidora de ozônio. A figura 1 mostra como se comporta a troposfera no Havaí, para determinadas concentraçoes de NOx em partes por trilhão. Observa-se que, acima de 60 ppt, aproximadamente, o local torna-se uma fonte de O3, visto que a produção supera a perda (Loss) nesse ponto.

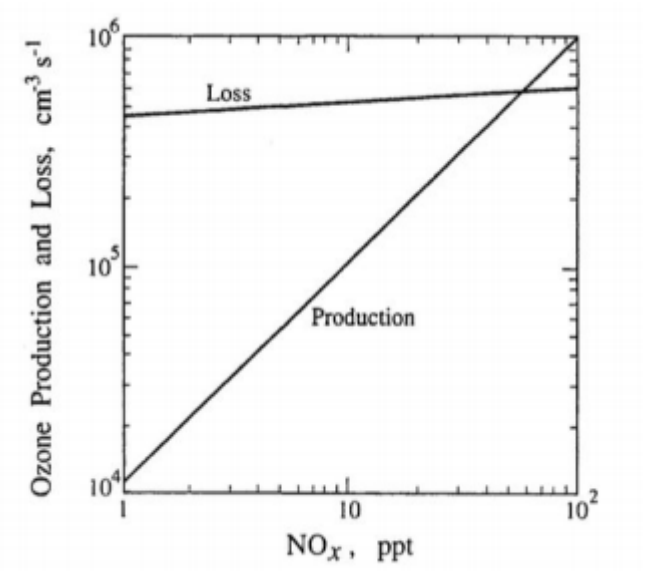


Figura 1: Relação entre NOx e produção de ozônio.

Embora o ozônio apresente um grande potencial para ser estudado por técnicas multivariáveis, não se tornou o foco deste trabalho, pois em Santa Gertrudes não foram realizadas medições desse poluente.

2.2 Conceitos básicos de Estatística descritiva e inferencial

Ciências experimentais e, de maneira mais geral, análise de dados requerem o planejamento otimizado e a execução cautelosa de uma série de experimentos, como forma de obter a matéria-prima para o trabalho de análise propriamente dita. Uma parte desde trabalho preparatório se baseia no conceito de Planejamento de Experimentos, proposto inicialmente em [4].

Esta etapa, no entanto, está fora do escopo deste trabalho, uma vez que este esforço já foi realizado pelas equipes da CETESB. Nosso ponto de partida são os dados sobre a qualidade do ar no município de Santa Gertrudes, estado de São Paulo, provindos das bases de dados da referida companhia.

O primeiro passo é a análise global dos dados quanto à presença de lacunas nas

observações ou existência de dados aberrantes (*outliers*). Para tanto, é necessário que se definam algum conceitos estatísticos usados nesse processamento de dados.

Um sistema físico pode ser estudado pela observação de certas variáveis ligadas a ele ao longo do tempo. A observação de uma variável consiste na medida de seu valor a um dado instante de tempo. Sua notação é $x_{i,k}$, que representa a i-ésima observação, $i = 1, \ldots, N$, da k-ésima variável, $k = 1, \ldots, p$.

A medição das variáveis é sujeita a perturbações e apresenta exatidão e precisão limitadas, pois não há aparelho de medida ideal. Assim, os valores das observações são modelados por variáveis aleatórias e várias medições são realizadas, de modo a se estimar a média (2.4) destas observações (estimativa do valor verdadeiro) e a variância (2.5) entre as observações (devida à precisão do método de medição ou à flutuação do valor verdadeiro entre as observações).

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2.4}$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{2.5}$$

Estas definições correspondem à média e à variância amostrais, que são os estimadores de máxima verossimilhança de seus equivalentes populacionais. Média (2.6) e variância (2.7) populacionais são atributos do conjunto de todos os valores assumidos pela variável de estudo. Ao realizar-se observações, o que se obtém é uma amostra da população, que contém uma fração do conjunto completo.

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{2.6}$$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2 \tag{2.7}$$

Outras características de uma amostra da dados que podem ser úteis ao analisá-las são:

- moda: valor que mais se repete em uma amostra, ou seja, que apresente a maior frequência em um histograma;
- mediana (x_{med}): valor que divide a amostra, com seus elementos ordenados, em dois

subconjuntos de mesmo tamanho. É equivalente ao segundo quartil da amostra.

2.2.1 Intervalos de confiança

A partir das observações feitas, há um interesse em se caracterizar a variável observada. Isso consiste em se estimar os parâmetros populacionais a partir dos amostrais. Pode-se demonstrar que, para amostras aleatórias (de tamanho n) de uma população, média e variância amostrais "orbitam"em torno dos valores populacionais, respeitando distribuições de probabilidade bem conhecidas.

A ligação entre as médias amostral e populacional se dá através da distribuição de Student: a partir destas médias, calcula-se o parâmetro t (2.8), que segue esta distribuição.

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \tag{2.8}$$

A variância amostral tem sua relação com a variância populacional mediada pela distribuição qui-quadrado (χ^2) (2.9).

$$\chi^2 = (n-1)\frac{s^2}{\sigma^2} \tag{2.9}$$

A partir destas relações, é possível percorrer o caminho inverso e calcular-se intervalos de confiança para média e variância populacionais. Para tanto, é necessário que se adote um valor para a confiança do intervalo. Pelo fato de a dstribuição de Student, bem como a Gaussiana, só se anularem no infinito, qualquer valor de t (ou z) tem probabilidade positiva de ocorrer. No entanto, é possível arbitrar limites mínimo e máximo para o intervalo de confiança, considerando que a probabilidade residual de o valor verdadeiro estar fora do intervalo (chamada significância e denotada α) é suficientemente baixa, podendo ser admitida.

Deste raciocínio, reorganizando (2.8) e (2.9), decorrem as correlações apresentadas em (2.10) e (2.11).

$$\bar{x} - t_{\nu,\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} - t_{\nu,1-\alpha/2}\frac{s}{\sqrt{n}} \tag{2.10}$$

$$\nu\frac{s^2}{\chi^2_{\nu,1-\alpha/2}} < \sigma^2 < \nu\frac{s^2}{\chi^2_{\nu,\alpha/2}} \tag{2.11}$$

Com:

$$\nu = n - 1 \tag{2.12}$$

Vale notar que as probabilidades de os valores populacionais serem maiores ou menores que os valores amostrais foram consideradas iguais, fato expresso pelas probabilidades cumuladas $\alpha/2$ e $1 - \alpha/2$ dadas como parâmetro das distribuições de Student e qui-quadrado. Além disso, o outro parâmetro destas distribuições, ν, é chamado de grau de liberdade e é uma função do tamanho da amostra (2.12).

2.2.2 Testes de Hipótese

Ao estabelecermos intervalos de confiança para média e variância populacionais, indiretamente indicamos que a hipótese nula (H_0) é o parâmetro populacional ser igual ao amostral e fixamos valores a partir dos quais essa hipótese seria rejeitada. Estes valores foram fixados ao admitirmos um valor máximo para a significância (ou probabilidade de occorência do erro tipo I, que é aquele cometido ao rejeitarmos H_0 sendo ela verdadeira).

Testes análogos podem ser estabelecidos com outras hipóteses nulas, admitindo-se por exemplo que médias de uma amostra seja maior que aquela de outra amostra.

$$\begin{cases} H_0 : \mu_1 > \mu_2 \\ H_1 : \mu_1 \leq \mu_2 \end{cases} \tag{2.13}$$

A aceitabilidade ou não da hipótese nula é definida através da comparação do parâmetro estatístico relativo ao teste em questão (t de Student, χ^2, etc.) e o valor limite, definido pela significância do teste (α), que pode corresponder à probabilidade residual de uma cauda única da função distribuição (para testes de superioridade/inferioridade) ou às duas caudas (testes de igualdade).

2.2.3 Aquisição e apresentação dos dados

Os dados utilizados nesse estudo foram fornecidos pela CETESB, que adquire dados sobre a qualidade do ar através de suas inúmeras estações medidoras (manuais ou automáticas), distribuídas por todo o estado de São Paulo. As variáveis medidas por cada estação variam, mas geralmente compreendem a temperatura, a concentração de material particulado (MP_{10} e $MP_{2,5}$), concentração de óxidos de enxofre, carbono e nitrogênio, intensidade e direção do vento, dentre outras. Estes valores são aferidos continuamente e uma média horária lhes é atribuída. Portanto, as observações apresentadas correspondem aos valores de média horária do período indicado.

Os dados são geralmente apresentados sob a forma de matrizes, cujas linhas apresentam as observações ordenadas temporalmente e as colunas, as variáveis observadas. Tais matrizes podem ser apresentadas sob a forma de séries temporais, nas quais uma variável individualizada é apresentada no eixo das ordenadas, e o valor de suas diferentes observações são apresentados no eixo das abscissas.

De uma maneira visual e intuitiva, esta representação permite ver a dispersão dos dados ao longo das observações, assim como deduzir seu valor médio e eventuais pontos aberrantes, chamados de *outliers*. Estes últimos podem ser devidos a erros grosseiros de medida e devem ser tratados em análises preliminares, para que não haja propagação deste erro sistemático não enviese análises posteriores.

À série temporal pode-se acrescentar duas retas, representando os limites superior e inferior do intervalo de confiança para a média dos valores. Se este intervalo tiver uma amplitude de 6σ, 99,73 % dos dados deveria estar entre estes limites, como exemplificado na Figura 2. Note que a variável apresentada foi padronizada, ou seja, subtraiu-se de todas as observações a média (centralização) e dividiu-se pelo desvio padrão.

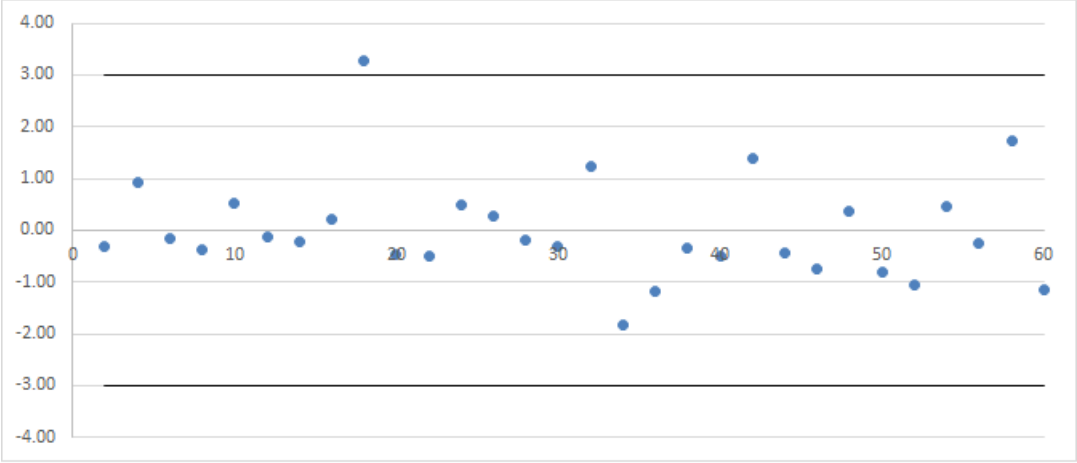


Figura 2: Exemplo de série temporal de observações de uma variável padronizada, com intervalo de confiança de 99,73%.

Nota-se que há um valor fora dos limites do intervalo de confiança no instante 18 deste exemplo. Este valor pode ser um indicativo de um dado anômalo e deve ser observado com atenção.

Para dados multivariados (observações que contenham mais de uma variável), a análise de dados anômalos deve ser mais cautelosa.

Deve-se primeiramente analisar as correlações entre as variáveis observadas, ou seja, a relação natural que eventualmente exista entre as duas. Esta etapa é importante pois

evidencia variabilidades das variáveis que não são aleatórias, mas sim dependentes da variação de outras variáveis do sistema, reduzindo assim os graus de liberdade do sistema. Assim, dados que poderiam ser considerados anômalos em análises individualizadas podem não o ser caso essa variabilidade seja explicada pela variação de uma segunda variável da qual esta dependa.

Uma grandeza que revela possível dependência entre variáveis observadas é sua covariância (2.14).

$$\begin{aligned} s_{jk} &= \frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k)}{\nu} \\ &= \frac{\sum_{i=1}^{n} X_{ij}X_{ik}}{\nu} \end{aligned} \tag{2.14}$$

Onde X_j é a variável x_j centrada na média.

A matriz de covariância **Cov** (2.15) apresenta as covariâncias de todas as combinações de variáveis, inclusive as variâncias de cada variável em sua diagonal principal.

$$\mathbf{Cov} = \begin{bmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & s_2^2 & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_p^2 \end{bmatrix} \tag{2.15}$$

Para problemas cujas variáveis tenham ordens de grandeza muito distintas, costuma-se preferir o emprego do coeficiente de correlação (2.16) e de sua matriz associada, análoga à matriz de covariância.

$$r_{jk} = \frac{s_{jk}}{s_j s_k} \tag{2.16}$$

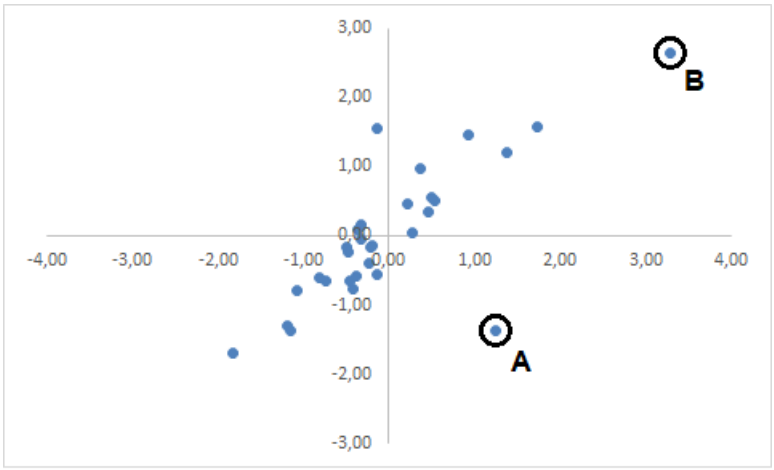
Onde $s_j = \sqrt{s_j^2}$ é o desvio padrão amostral da variável j.

Vale notar que o coeficiente de correlação varia, em módulo, de 0 a 1. Quando a correlação entre duas variáveis é nula, diz-se que são variáveis independentes.

Como citado anteriormente, a análise de correlações entre variáveis permite que se identifiquem pontos anômalos multivariados que não seriam identificados em análises individualizadas, como exemplificado na Figura 3.

Esta figura apresenta o conjunto de observações de duas variáveis, cada uma em um

Figura 3: Análise de correlação entre observações de duas variáveis



Fonte: Autoria Própria

dos eixos coordenados. Nota-se que os pontos se aglomeram em torno de uma reta de coeficiente angular positivo, o que evidencia uma correlação positiva ($r > 0$).

Além disso, nota-se que há 2 pontos que destoam do comportamento geral, identificados como A e B. O ponto A distancia-se do aglomerado de pontos mas provavelmente não seria considerado um *outlier* numa análise individualizada da variável do eixo das abscissas, pois encontra-se entro da projeção do aglomerado neste eixo. O fator que o diferencia é seu distanciamento com respeito à média da segunda variável, do eixo das ordenadas. O ponto B, por outro lado, é um *outlier* identificável em ambas as análises.

Uma análise rigorosa de *outliers* multivariados requer a utilização da distribuição normal multivariada, que definem intervalos de confiança com a forma de hiperelipsoides em espaços de p dimensões. A medida que permite avaliar o distanciamento entre observações multivariadas é a distância estatística (*statistical distance* - SD - em inglês) (2.17) que, diferentemente da distância euclideana clássica, leva em conta a variância de cada variável do sistema, dado que diferentes variáveis admitem diferentes dispersões com respeito à média.

$$SD_{ih} = \sqrt{\sum_{j=1}^{p} \left(\frac{x_{ij} - x_{hj}}{s_j} \right)^2} \tag{2.17}$$

2.3 Análise de Componentes Principais

Para sistemas em que se observam muitas variáveis, é possível que algumas delas expliquem melhor a variância total observada que outras. Além disso, em etapas subsequentes de modelagem, é de interesse que se otimize o número de variáveis levadas em conta, no sentido de escolher o menor número de variáveis capaz de explicar a maior parte de variância do sistema.

Um método que pode ser aplicado a esse tipo de problema é a Análise de Componentes Principais (ou, em inglês, *Principal Component Analysis - PCA*). O princípio deste método é de combinar linearmente as variáveis originais de modo a encontrar combinações independentes entre si para que se possa, a partir delas, descobrir qual nova variável é responsável pela maior contribuição à variância total.

O problema de se encontrar componentes principais e_j, $j = 1, \ldots, p$, a partir das variáveis originais X_j, $j = 1, \ldots, p$, pode ser descrito da seguinte forma:

$$\begin{cases} e_1 = w_{11}X_1 + w_{12}X_2 + \cdots + w_{1p}X_p \\ e_2 = w_{21}X_1 + w_{22}X_2 + \cdots + w_{2p}X_p \\ \quad \vdots \\ e_p = w_{p1}X_1 + w_{p2}X_2 + \cdots + w_{pp}X_p \end{cases} \tag{2.18}$$

Onde w_{jk} corresponde ao peso que a variável original k tem no componente j.

Impõe-se algumas condições suplementares, como a normalidade (2.19) e ortogonalidade (2.20) dos componentes.

$$\begin{aligned} w_{k1}^2 + w_{k2}^2 + \cdots + w_{kp}^2 = 1 \\ k = 1, ..., p \end{aligned} \tag{2.19}$$

$$\begin{aligned} w_{k1}w_{j1} + w_{k2}w_{j2} + \cdots + w_{kp}w_{jp} = 0 \\ k = 1, ..., p \\ j = 1, ..., p \\ k \neq j \end{aligned} \tag{2.20}$$

Sucintamente, o problema de se encontrar combinações independentes das variáveis originais é análogo a diagonalizar a matriz de covariância delas, umas vez que a matriz diagonal apresentaria covariâncias nulas (novas variáveis independentes). Tem-se que a nova matriz de covariância, diagonal e identificada por s_e^2, é calculada como em (2.21).

$$\begin{aligned} s_e^2 &= \mathbf{w}^T \mathbf{Cov} \mathbf{w} = \lambda \\ &= \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix} \end{aligned} \tag{2.21}$$

Logo, o primeiro passo desta análise é a identificação do polinômio característico da matriz de covariância (2.22).

$$p_{\mathbf{Cov}}(\lambda) = \det(\mathbf{Cov} - \lambda \mathbf{I}) \tag{2.22}$$

A seguir, calcula-se as raízes do polinômio característico, que são os autovalores da matriz de covariância, bem como seu autovetor associado.

Os autovalores são as variâncias das novas variáveis. Para facilitar a análise, costuma-se organizá-los em ordem decrescente, de modo a priorizar as variáveis que expliquem mais a variância total.

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0 \tag{2.23}$$

Os autovetores w_k correspondem aos pesos que constituem os componentes principais e são calculados a partir dos correspondentes autovalores como em (2.24).

$$(\mathbf{Cov} - \lambda_k \mathbf{I}) w_k = 0 \tag{2.24}$$

Assim, obtém-se como resultado desta análise a nova matriz de covariância (2.21), bem como a matriz $\mathbf{w}$ de autovetores, correspondendo aos pesos dos componentes principais, equivalente a uma matriz de mudança de base entre a base de variáveis originais e uma base ortogonal (variáveis independentes).

A redução de variáveis do sistema pode ser realizada ao se comparar os valores das novas variâncias. Pode-se fixar uma porcentagem mínima a da variância total a ser explicada pelos componentes principais e calcula-se a variância acumulada ao se acrescentar

sucessivamente os componentes principais ao modelo. Tão logo o valor mínimo é atingido, os componentes subsequentes podem ser desprezados.

No entanto, ainda falta exprimir esses componentes escolhidos em termos das variáveis originais. Para tanto, observa-se os autovetores desdes componentes e identifica-se a qual variável corresponde o maior peso dele, e adiciona-se esta variável ao modelo. Prossegue-se assim sucessivamente para todos os componentes escolhidos, acrescentando-se a cada vez uma variável ao modelo.

2.4 Regressão Linear Multivariável

2.4.1 Introdução à Regressão

A regressão linear é uma das ferramentas de base da aprendizagem estatística supervisionada, sendo amplamente utilizada para cálculo de saídas quantitativas. As técnicas de regressão são fundamentais para compreender a aplicação de métodos mais complexos.

A regressão linear simples é amplamente conhecida, onde se tenta prever uma variável dependente Y a partir de uma variável independente X, definindo-se um coeficiente angular e um ponto de intercepto.

$$Y = \beta_0 + \beta_1 x \tag{2.25}$$

Assim, para um conjunto de pontos, busca-se encontrar a reta que minimize a distância entre os valores previstos e os valores reais. O método mais comum baseia-se no Residual Sum of Squares (RSS), ou soma quadrática dos resíduos.

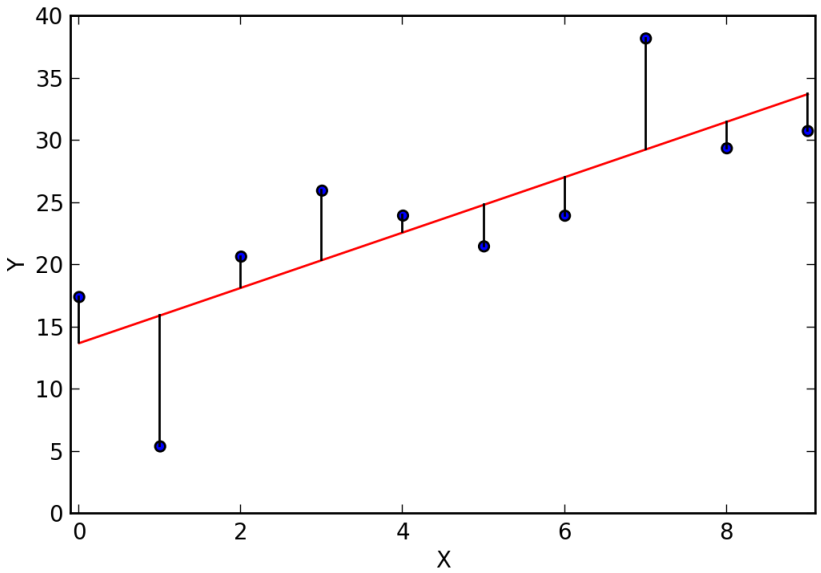
Com duas variáveis preditivas, ainda conseguimos ter uma percepção visual do fenômeno. Aqui, busca-se reduzir a distância entre o plano e os pontos, como mostrado na figura 5 .

Entretanto, para problemas do mundo real, trabalhamos com uma série de variáveis preditivas. Para se trabalhar com p variáveis, é necessário realizar alguns ajustes na regressão. Podemos escrevê-la como:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + u \tag{2.26}$$

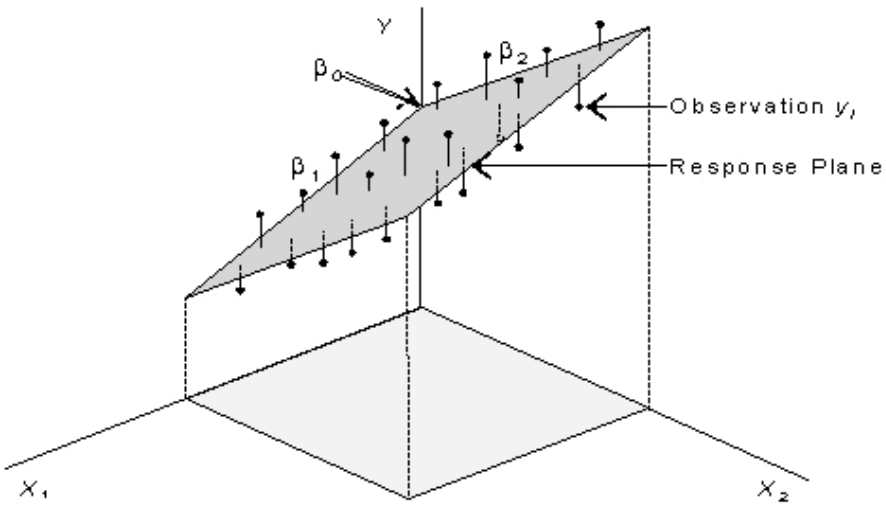
Onde β representam o intercepto (índice zero) e os coeficientes angulares. O valor u é um resíduo. Utiliza-se a mesma ideia de ajuste, reduzindo o RSS através da técnica de

Figura 4: Exemplo de Regressão simples com distância aos pontos



Fonte: Wikipedia

Figura 5: Exemplo de Regressão bivariável com distância entre plano e pontos



Fonte: Towards Data Science

mínimos quadrados, ou OLS, exemplificada abaixo. [5]

$$\begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & ... & x_{1p} \\ 1 & x_{21} & ... & x_{2p} \\ ... & ... & ... & ... \\ 1 & x_{n1} & ... & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ ... \\ \beta_p \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ ... \\ u_n \end{bmatrix} \tag{2.27}$$

Isso nos permite escrever:

$$Y = X\beta + u \tag{2.28}$$

A estratégia para minimizar o erro quadrático será de calcular a soma quadrática residual e, em seguida, encontrar um estimador que minimize a soma. Podemos escrever as seguintes transformações:

$$u'u = (Y - X\beta)'(Y - X\beta) \tag{2.29}$$

$$u'u = Y'Y - 2\beta'X'Y' + \beta'X'X\beta \tag{2.30}$$

Resta derivar e igualar a zero para encontrar o estimador de mínimo erro quadrático.

$$\frac{d(u'u)}{d\beta} = -2X'Y + 2X'X\beta \tag{2.31}$$

$$X'X\beta = X'Y \tag{2.32}$$

$$\beta = (X'X)^{-1}X'Y \tag{2.33}$$

Esse resultado permite calcular os valores de β_p respeitando o mínimo erro quadrático.

2.4.2 Métricas

Antes de se realizar um bom modelo de regressão múltipla, é fundamental entender o que caracteriza um bom modelo. Para isso, existem diferentes métricas que podem ser utilizadas. Para algoritmos de classificação, seria possível utilizar a quantidade de acertos sobre a quantidade de valores estimados. No caso da regressão, utilizam-se, majoritariamente, quatro técnicas. [6]

MSE: Trata-se do erro médio quadrático. Como eleva ao quadrado a diferença entre o valor previsto e o real, penaliza fortemente todos os erros. É utilizada por ser diferenciável e, portanto, otimizável. Para N observações, o MSE é calculado através da fórmula seguinte, onde y_r é o valor real, e y_p é o valor previsto.

$$MSE = \frac{1}{n}\sum_{i=1}^{N}(y_r - y_p)^2 \tag{2.34}$$

RMSE: Raiz do erro médio quadrático. É a métrica mais utilizada para problemas de regressão, e penaliza principalmente erros grandes, uma vez que realiza primeiro a potência e em seguida a raiz quadrada.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_r - y_p)^2}{n}} \tag{2.35}$$

MAE: Erro médio absoluto. É um método mais robusto a outliers, que não penaliza erros de forma tão extrema quando MSE. É a métrica que utilizaremos, para que os outliers não tenham um impacto tão grande no modelo.

$$MAE = \frac{\sum_{i=1}^{N}|y_r - y_p|}{n} \tag{2.36}$$

R^2: É o coeficiente de determinação. Compara o modelo ajustado com um modelo básico (a média) e mostrar quão melhor é o modelo. É um valor sempre menor que 1, onde coeficientes maiores indicam que o modelo ajustado representa bem os dados.

$$R^2 = 1 - \frac{MSE(modelo)}{MSE(base)} \tag{2.37}$$

2.4.3 Seleção de variáveis

A seleção das variáveis preditivas é muito importante para que se obtenha uma boa regressão. Incluir parâmetros desnecessários pode tornar o modelo mais complexo que o necessário, aumentar seu viés ou sua variância. Nesses casos, dizemos que o modelo está sendo "treinado" com parâmetros irrelevantes, o que normalmente resulta em pior desempenho. Por isso, é necessário analisar quais parâmetros (ou features) realmente ajudam a explicar a variável que se deseja estimar. A seleção de parâmetros também evita overfitting, que será descrito em seguida.

Existem diferentes métodos de seleção de features. Neste trabalho, utilizaremos um algoritmo de "backward elimination" que se baseia na relevância estatística de cada fator − isto é, avaliando se determinado componente contribui ou não para explicar a variância da variável desejada.

A eliminação backward é um método stepwise, o que quer dizer que se retira uma variável por vez, em vários passos consecutivos. Diz-se backward pois a eliminação é realizada de trás para frente. Os passos de eliminação do algoritmo são os seguintes:

1. Um modelo de regressão é estimado usando todos os parâmetros disponíveis.

2. Para cada variável, realiza-se um teste de hipótese no coeficiente angular. A hipótese nula é de que o coeficiente angular real da variável testada é igual a zero, sendo irrelevante no modelo.

3. Calcula-se o valor do teste t de Student (assumindo, portanto, normalidade dos dados) e calcula-se o p-valor para cada uma das variáveis.

4. A cada iteração, elimina-se a variável com maior p-valor.

5. Termina-se quando todas as variáveis restantes possuem p-valor abaixo de um nível de confiança (p-valor menor que 0.05 ou 0.01).

Vale ressaltar que um p-valor pequeno indica que a probabilidade da hipótese nula ocorrer é baixa, e por isso podemos descartá-la.

2.4.4 Multicolinearidade

A multicolinearidade é mais um ponto que deve ser levado em consideração quando se realiza a seleção das variáveis para o modelo. De modo grosseiro, multicolinearidade é redundância entre variáveis. Quando duas variáveis estão fortemente correlacionadas, o modelo é prejudicado, uma vez que cada uma passa a fornecer menos informação independente à regressão, diminuindo a significância de ambos parâmetros.

O resultado disso é a diminuição na confiabilidade da regressão, para os parâmetros envolvidos e para os demais. Isso significa que, se o objetivo for realizar a análise da influência individual de cada parâmetro, a multicolinearidade terá um impacto negativo na análise. Se a análise individual não for objetivada, o resultado geral da regressão não costuma ser afetado.

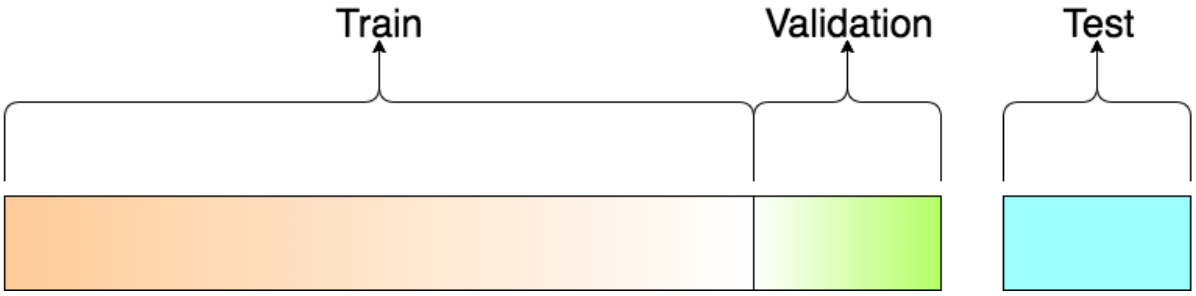
O modo mais simples para testar a multicolinearidade é identificando se as variáveis possuem um coeficiente de correlação alto entre elas. Caso esse coeficiente seja superior ao limite de 0.4 (variando de problema a problema), podemos considerar a exclusão de uma dessas variáveis. Em geral, opta-se por guardar aquela que explica melhor a variável que se deseja prever.

2.4.5 Overfit, treino e teste

Nas técnicas de Machine Learning, os modelos costumam ser ajustados utilizando RSS no conjunto de dados. Entretanto, uma parte do conjunto de dados deve ser separada, para que o modelo possa ser testado contra novos dados, sobre os quais ele não foi treinado. Chama-se isso de set de testes e set de treino. Aquelas métricas também são utilizadas sobre os sets de teste, além da acurácia, em alguns casos.

Em outros casos, antes de passar para o set de testes, pode-se passar pelo set de validação, que ajuda a identificar erros/viés no modelo, e corrige-os através de algumas iterações com o set de treino. Em seguida, passa-se para o set de testes.

Figura 6: Divisão do dataset em três partes



Fonte: Towards Data Science Website

Ao ajustar o modelo, é necessário ser cauteloso com a ocorrência de overfitting, isto é, quando o modelo se torna mais complexo e se ajusta exclusivamente aos pontos treinados, tornando-se menos eficaz na previsão de novos pontos. Matematicamente, é equivalente a realizar um ajuste polinomial de grau alto que obrigue seu polinômio a passar por todos os pontos mostrados. Em seguida, ao incluir um novo ponto, esse ajuste deixa de ser útil.

Para evitar overfitting, a validação com o set de teste é essencial, assim como a seleção de parâmetros. Outras técnicas também são possíveis, como verificar se a reta ajustada para o conjunto de testes está muito diferente da reta ajustada pelo treino. Caso esteja, isso pode ser um indício que a reta está ajustada demais para o set de treino. Outra forma é alterando-se a função de custo do ajuste, através de técnicas de regressão como Lasso ou Ridge.

2.4.6 Regressão Lasso

A regressão linear básica, tal como descrita até então, ajusta a reta baseada na RSS, a mínima soma de quadrados. Chama-se isso de função de custo do modelo. Isso significa

que, caso um parâmetro seja considerado relevante pelo modelo, receberá um coeficiente angular. Como não há penalização nos coeficientes, o ajuste pode incluir parâmetros em excesso, ou favorizar um certo parâmetro além do necessário para se ajustar melhor ao set de treino. Isso favorece o overfit. A função de custo para essa regressão seria:

$$Custo(RSS) = \sum_{i=1}^{N}(y_r - y_p)^2 \tag{2.38}$$

A regressão Lasso é um modelo que penaliza a escolha dos coeficiente, incluindo-a na função de custo, que passa a ser escrita como:

$$Custo = \sum_{i=1}^{N}(y_r - y_p)^2 + \alpha \sum_{i=0}^{p} |\beta_i| \tag{2.39}$$

Onde o α é um parâmetro de peso que o utilizador escolhe. A técnica padrão utiliza 0.

Na prática, isso implica que a regressão Lasso trabalhará para diminuir os coeficientes, exceto se o aumento destes realmente seja compensado por uma explicação maior da variável observada.

Por fim, a regressão Lasso também traz, por consequência, a seleção de features. Ao penalizar os coeficientes, essa regressão também pode estabelecer coeficiente zero para determinadas features, excluindo-as do modelo.

2.5 Decomposição de Séries Históricas

O tratamento de séries históricas é uma ferramenta útil quando se possui uma grande quantidade de dados onde o eixo temporal é uma variável significativa. As séries históricas são importantes para compreender o fenômeno estudado e possibilita a realização de previsões. As séries históricas são amplamente utilizadas para modelização de fenômenos estocásticos.

Para tratar uma série histórica de dados, costuma-se realizar a decomposição em três componentes dos valores obtidos: tendência, sazonalidade e ruído. A tendência, como o nome indica, mostra o comportamento geral da série, podendo mostrar crescimento, decrescimento ou lateralização, onde o modelo já se encontra em um estado estacionário. A sazonalidade traz à luz as questões cíclicas do fenômeno, podendo ser aplicada por dia,

por ano ou por estação, dependendo do que se observa. Finalmente, o ruído é tudo aquilo que escapa da explicação da tendência e da sazonalidade.

Um outro parâmetro que se utiliza para descrever séries históricas é o nível, que se relaciona ao valor numérico observado para a variável. Isso significa que uma tendência crescente, por exemplo, leva o nível das observações de um ponto inferior para um nível superior. Na maioria das vezes, a observação do nível é feita em conjunto com a observação da tendência, unificando os dois conceitos. Esses três componentes podem se combinar de diferentes maneiras, sendo as mais conhecidas a decomposição multiplicativa e a decomposição aditiva.

A decomposição aditiva é representada por:

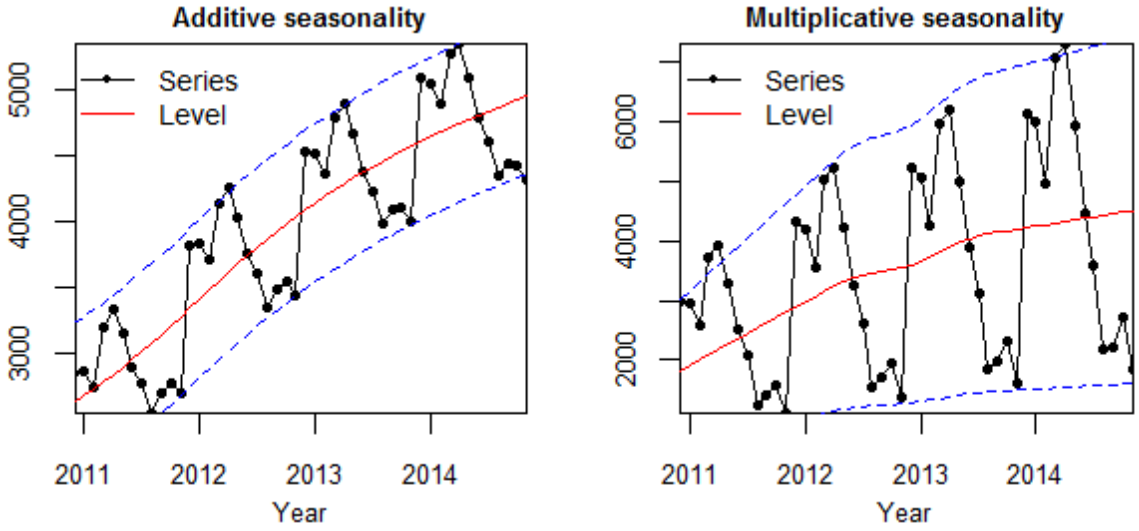
$$Y = Tendencia + Sazonalidade + Residuo \tag{2.40}$$

Enquanto a multiplicativa traz:

$$Y = Tendencia * Sazonalidade * Residuo \tag{2.41}$$

O fator principal para a escolha é a sazonalidade. Se a sazonalidade tem a amplitude aumentada com o aumento do nível, o fenômeno é representado de forma mais precisa pela decomposição multiplicativa. A figura abaixo representa essa relação. [7]

Figura 7: Comparação entre sazonalidade multiplicativa e aditiva



Fonte: Lancaster University - Time Series Decomposition

Para realizar a decomposição, existem diferentes técnicas de identificar os componentes. Os dois principais modelos utilizados são o clássico e o STL. Neste trabalho, utilizaremos e estudaremos o clássico. [8]

O método clássico é uma das primeiras técnicas de decomposição conhecidas, e é utilizado até hoje para modelização de diferentes fenômenos. O método depende da definição de um parâmetro de sazonalidade, que depende da distribuição dos dados (e.g., 4 para dados trimestrais, 12 para dados mensais, entre outros). Existem outros métodos que estimam esse parâmetro de sazonalidade a partir dos dados.

O método clássico assume que a sazonalidade se aplica de ano em ano de forma constante. A aplicação do método aditivo passa por cinco etapas. A componente tendência (T) é calculada usando uma média móvel central baseada no parâmetro de sazonalidade definido acima. Em seguida, calcula-se a série sem tendência, subtraindo-se a tendência do Y. No caso de uma série multiplicativa, divide-se pela tendência.

Para passar para o componente sazonal, é necessário calcular a média para cada período de sazonalidade – no caso das observações de particulado, esse período é anual. Depois, replica-se o padrão encontrado para todo o período observado. Isso fornece diretamente a componente Sazonal.

Finalmente, o cálculo do ruído vem da subtração das observações pelas componentes sazonalidade e tendência.

O procedimento é análogo para a decomposição multiplicativa, utilizando divisões em vez de subtrações.

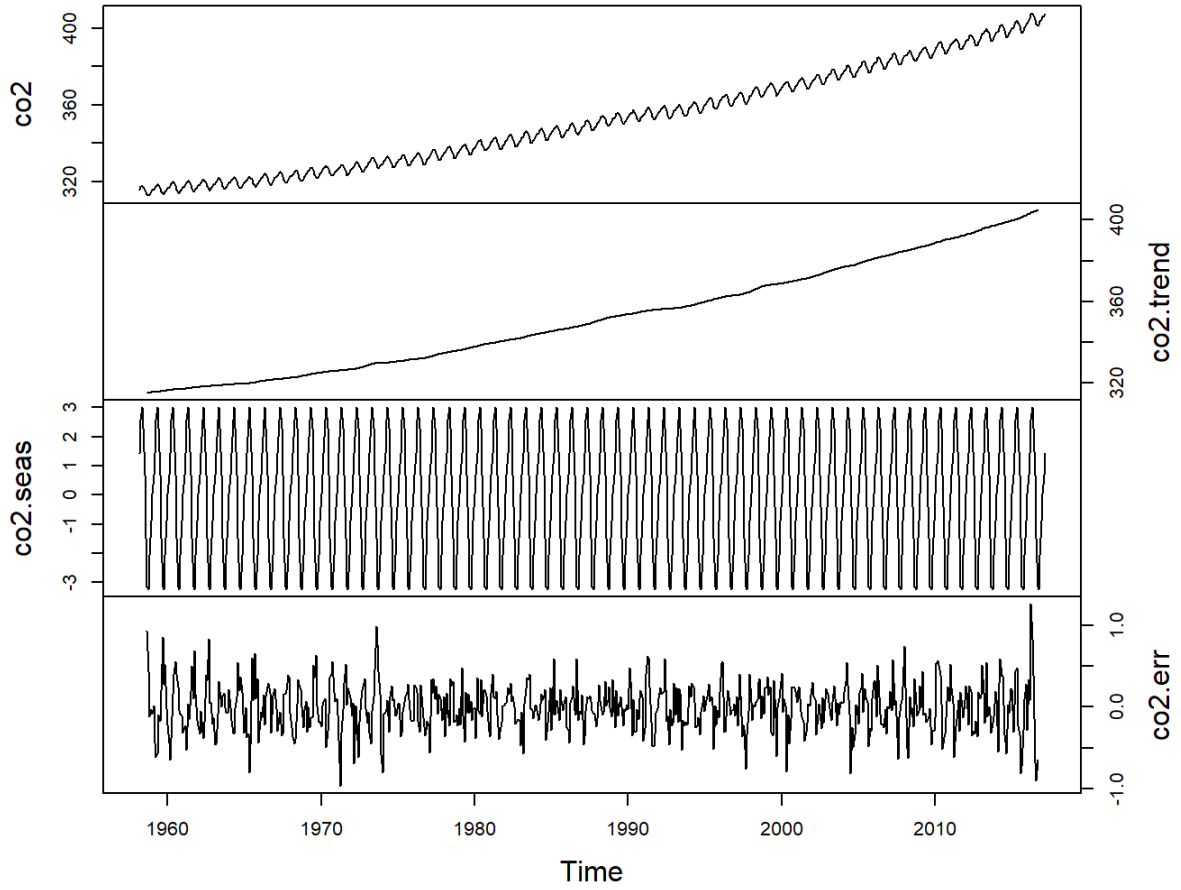
Uma das principais limitações do método clássico é que a média móvel centrada utilizada para calcular a tendência fica limitada no final da série, o que pode implicar uma aproximação mais grosseira nos extremos da série.

Além disso, assume-se que é possível isolar a sazonalidade através da média daquela estação através dos anos. Isso implica que, quanto maior a série, melhor essa aproximação, por considerar mais períodos e realizar uma aproximação mais confiável.

Por fim, pode-se analisar diretamente o residual ou a combinação do residual com a tendência, sem o efeito da sazonalidade. Isso permite identificar pontos anômalos mesmo em momentos de menor nível, quando a estação possui valores mais baixos de poluentes.

O exemplo abaixo, retirado do livro “Applied Time Series Analysis for Fisheries and Environmental Sciences” [9], ilustra bem os três componentes presentes nas séries temporais. Nesse caso, temos dados mensais de CO2 nos oceanos.

Figura 8: Exemplo de decomposição



Fonte: Applied Time Series Analysis for Fisheries and Environmental Sciences

A tendência fica clara, sendo de aumento, e a sazonalidade anual também é bem demarcada, com picos no início de cada ano. Por fim, o residual representa tudo aquilo que não foi explicado pelos outros componentes da decomposição.

3 ANÁLISES

3.1 Tratamento dos dados

Como citado anteriormente, todos os dados utilizados para as análises presentes neste relatório foram obtidos através do banco de dados QUALAR. A CETESB coleta e disponibiliza no QUALAR os dados de todas as estações de medição do estado de São Paulo. O acesso é público.

Os equipamentos da CETESB realizam medições de diferentes parâmetros, como material particulado, direção do vento, velocidade do vento e concentração de nitrogênio. A obtenção dos dados é feita em intervalos de 30 segundos a 5 minutos, dependendo do parâmetro. Em seguida, os dados são condensados e publicados pela CETESB no formato de médias horárias. Portanto, a estrutura do banco de dados é tal que cada linha é uma observação horária, e cada dia será composto por 24 observações dos diferentes parâmetros.

Para que a observação de uma hora seja considerada válida pela CETESB, é necessário que 75% dos dados que compõem a média horário sejam válidos.

Embora o QUALAR disponha de dados referentes a vários anos para a cidade de São Paulo, as medições de material particulado em Santa Gertrudes são mais recentes, com medições iniciadas no segundo semestre de 2014 para MP10 e em 2018 para MP2.5. As estações manuais realizavam medições antes desse período, mas com menos de uma medição por mês.

As variáveis trabalhadas foram a direção do vento (DV), que possui valores de 0 a 359 graus, a umidade relativa (UR), com valores de 0 a 100, a temperatura (Temp) indicada em graus Celsius, os poluentes particulados, MP10 e MP2.5, medidos em $\mu g/m^3$, e o NOx, dado em partes por bilhão (ppb).

Assim, os bancos de dados trabalhados estão expostos na tabela 1.

As cidades de São Paulo e Araçatuba foram utilizadas como comparação, sendo ana-

Refeência	Período	Parâmetros
Santa Gertrudes 1	01/2015 a 12/2019	DV, MP10
Santa Gertrudes 2	01/2018 a 12/2019	DV, MP10, MP2.5,UR, Temp
São Paulo	01/2018 a 01/2019	DV, MP10
Araçatuba	01/2018 a 01/2019	DV, MP10

Tabela 1: Bancos de dados utilizados

lisadas somente para um ano completo.

Entretanto, as estações CETESB possuem períodos de manutenção e falhas nos sensores, o que compromete a qualidade dos dados, especialmente para a medição do material particulado. A análise exploratória dos dados mostra certos períodos com falhas de medição – marcado como valor vazio – que passaram pelo tratamento de dados descrito a seguir.

3.1.1 Tratamento de dados para poluentes e fenômenos meteorológicos

O material particulado é um fenômeno que não é facilmente previsível, por depender, além das condições atmosféricas, da interação do ser humano com a natureza, variando muito de um dia para o outro. Portanto, optou-se por realizar a imputação de dados estimados somente para pequenos intervalos de falhas (menores do que 12 horas), enquanto intervalos maiores, que chegam a mais de uma quinzena, tinham as observações descartadas.

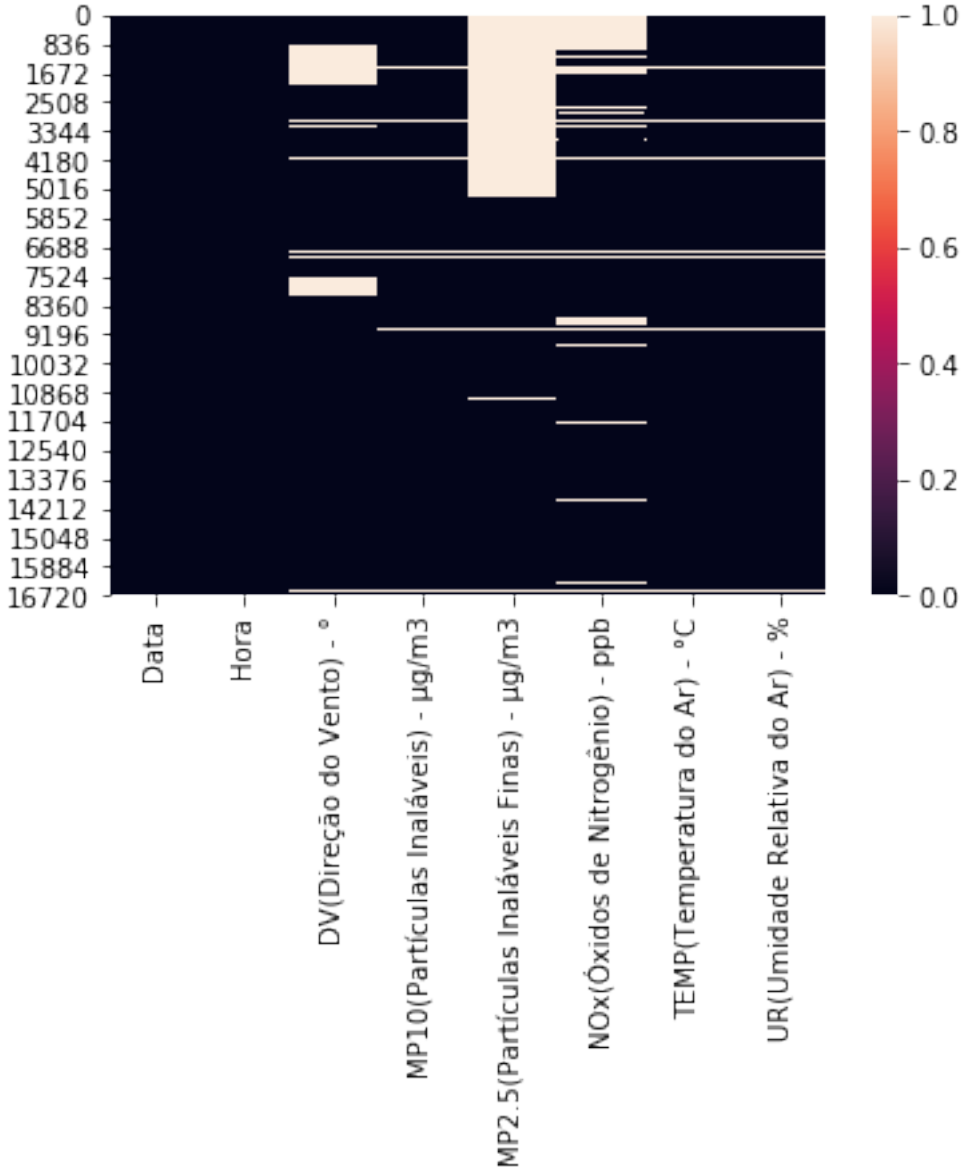
Ao preencher somente pequenos intervalos, consegue-se garantir maior coesão dos dados e diminuir a quantidade de informação perdida. Esses intervalos eram preenchidos combinando o perfil diário histórico e o intervalo de observações em que se encontra. Por exemplo, uma observação faltante ao meio-dia de um dia qualquer em janeiro seria preenchido considerando o perfil diário histórico dos dias de janeiro, indicando quanto acima ou abaixo da média diária se espera que a observação de meio-dia fique, e a média dos valores absolutos 12h antes e depois do vazio.

Vale ressaltar que se estudará, mais a frente, um método de imputar dados para MP10 através de regressão multivariável. O tratamento realizado aqui foi para obter o perfil dos poluentes.

Para a direção do vento, temperatura e umidade relativa, considerou-se que o efeito cíclico, por ser um fenômeno natural, era notável, possibilitando a estimativa de seus valores com base nas médias históricas para diferentes meses e horas do dia.

A imagem abaixo ilustra a densidade de vazios para as diferentes variáveis. Observa-se um período sem observações para MP2.5, o primeiro semestre de 2018, que precisou ser descartado. Existem alguns momentos em que temos a interrupção de todas as medições, e outros de erros pontuais, como erros no MP2.5 e no NOx.

Figura 9: Vazios



Fonte: Autoria própria

Seguindo o tratamento de dados descrito, obtivemos a tabela 2.

Os dados anômalos para material particulado não foram examinados no tratamento de dados, sendo analisados separadamente, uma vez que se consideraram válidos todos os dados (seguindo a classificação da CETESB). Isso indica que os valores anômalos observados possuem natureza física, não sendo necessariamente oriundos de erros de medição,

Refeência	No de Observações	No de Inputs	Descartados
Santa Gertrudes 1	43.848	535	1723
Santa Gertrudes 2	17.544	830	5774
São Paulo	8760	0	0
Araçatuba	8760	0	0

Tabela 2: Resultado do tratamento de dados

mas sim de ocorrências físicas.

3.2 Perfil dos poluentes em Santa Gertrudes

Para compreender a problemática da poluição do ar em Santa Gertrudes, buscou-se analisar o perfil de poluição no município, para os três principais poluentes analisados: MP10, MP2.5 e NOx.

As condições meteorológicas (temperatura e precipitação) em Santa Gertrudes se comportam tal como mostra a figura 10, com picos de precipitação no início e no final do ano, e com período de seca no meio do ano. A temperatura diminui durante os meses de inverno.

3.2.1 MP10

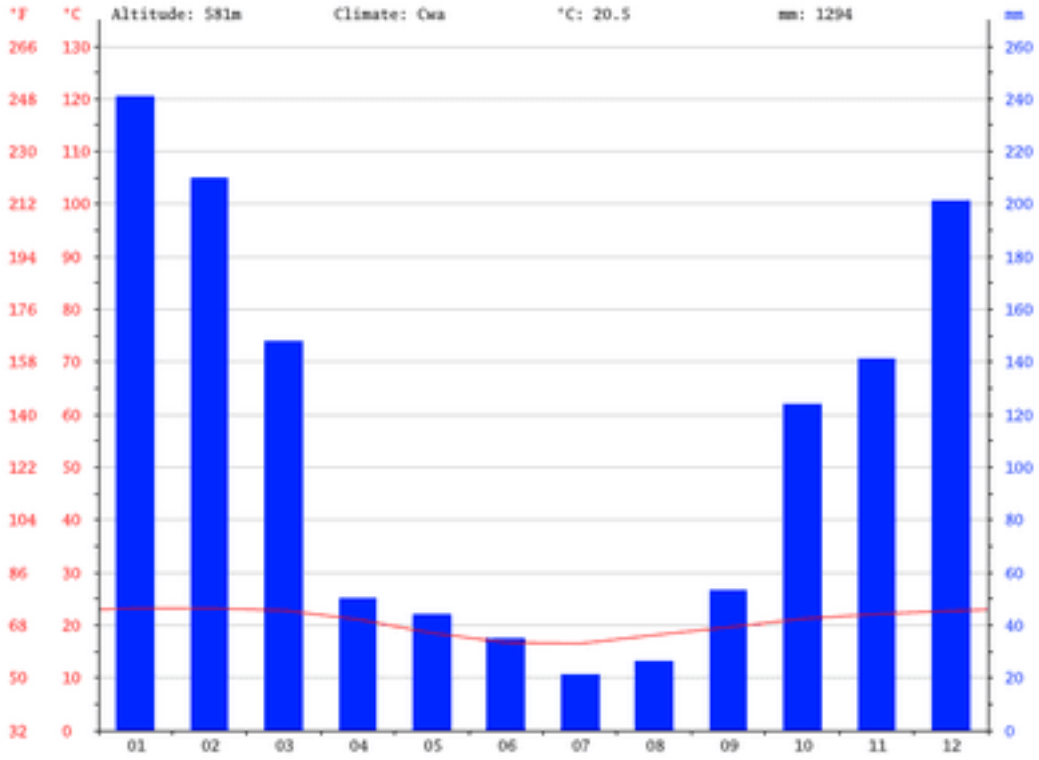
A análise realizada busca identificar o comportamento diário dos poluentes acima, nas diferentes estações do ano. A figura abaixo indica qual o valor esperado para um dia qualquer em cada estação, divido por hora.

O gráfico revela que existe uma grande diferença entre as estações do ano, com alta na concentração de poluentes durante inverno e outono, as estações mais frias e mais secas na região. Além disso, em todas as estações, observa-se um perfil diário marcado por dois picos, às 7h e às 19h.

Inicialmente, imaginou-se que tais picos poderiam estar ligados à hora de pico da circulação de carros. Entretanto, ao comparar o perfil diário anual de Santa Gertrudes com São Paulo, a cidade com maior frota veicular do país, observa-se que o comportamento de picos é único ao município de Santa Gertrudes, indicando uma correlação com o forte tráfego de caminhões carregando fontes de particulados na região, e a atividade econômica em si.

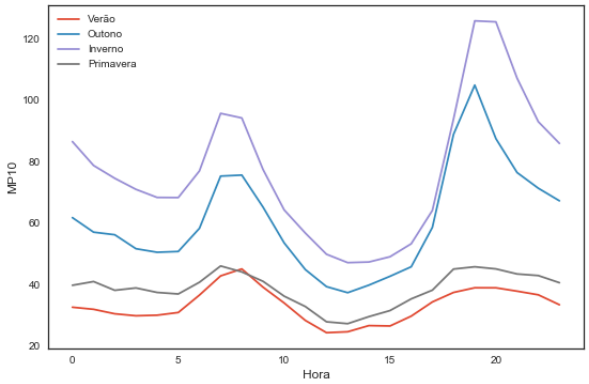
O gráfico acima também serve para colocar em perspectiva a magnitude da poluição

Figura 10: Precipitação e temperatura em Santa Gertrudes



Fonte: Climate Data Brasil

Figura 11: Perfil diário de MP10 em Santa Gertrudes

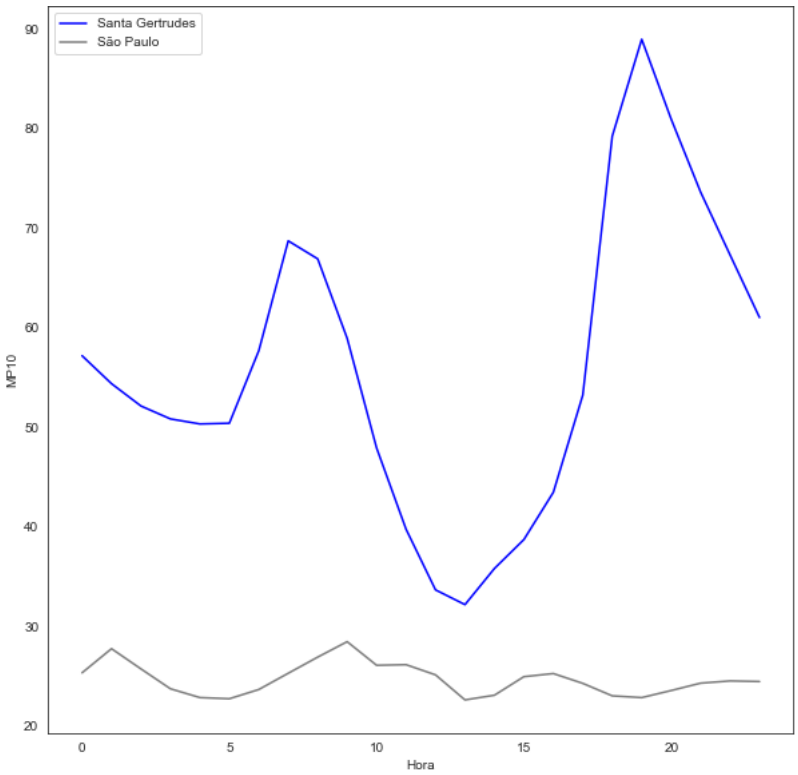


Fonte: Autoria própria

do município. Embora a população de Santa Gertrudes seja 500 vezes menor do que a de São Paulo, contando com 24 mil habitantes – segundo censo IBGE de 2015 – o nível de poluição por particulados é 2,5 vezes maior.

Considerando os limites impostos pela Organização Mundial da Saúde (OMS), o valor médio de concentração de MP10 apresentado em Santa Gertrudes é superior ao consi-

Figura 12: Comparação entre Santa Gertrudes e São Paulo



Fonte: Autoria própria

derável saudável (40 μg/m^3). Entre 2015 e 2020, o município de Santa Gertrudes obteve uma média diária superior ao recomendado pela OMS em 57% dos dias. Ao mesmo tempo, o pico diário superou o valor diário recomendado em 92% dos dias analisados. Se considerarmos o valor de curto prazo limite, de 120 μg/m^3, esse valor foi superado 32% das horas.

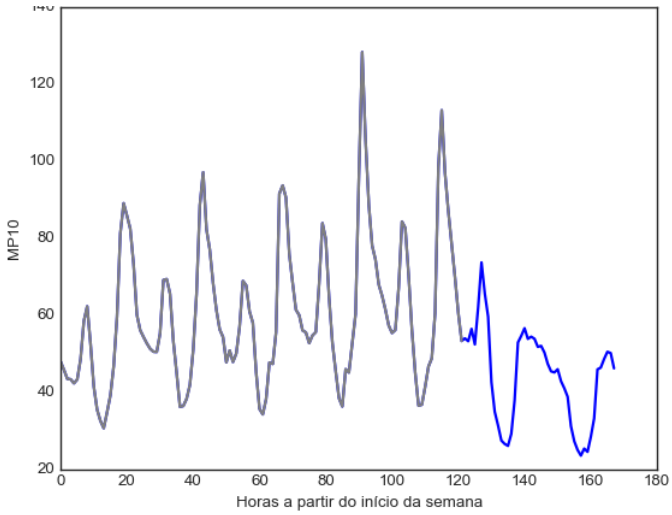
A influência da atividade econômica da região no nível de poluentes também fica explícita quando consideramos a distribuição de poluentes nos dias da semana. Em dias úteis, onde as fábricas e a distribuição de argila e outras matérias primas da cerâmica.

O decrescimento em finais de semana é notável, contando como a maior parte dos dias em que a média diária não supera o valor limite da OMS. Na imagem acima, cada conjunto de dois picos representa um dia, sendo o maior pico no momento da noite, seguindo o perfil diário estudado anteriormente.

O perfil anual está representado pela figura 14, indicando o comportamento de alta durante os meses secos, e baixa durante os meses úmidos.

Julgou-se pertinente analisar a correlação entre a concentração média de material particulado e a direção do vento. Para tanto, os dados de 2015 a 2019 foram separados

Figura 13: Dias da semana (cinza) e final de semana (azul)



Fonte: Autoria própria

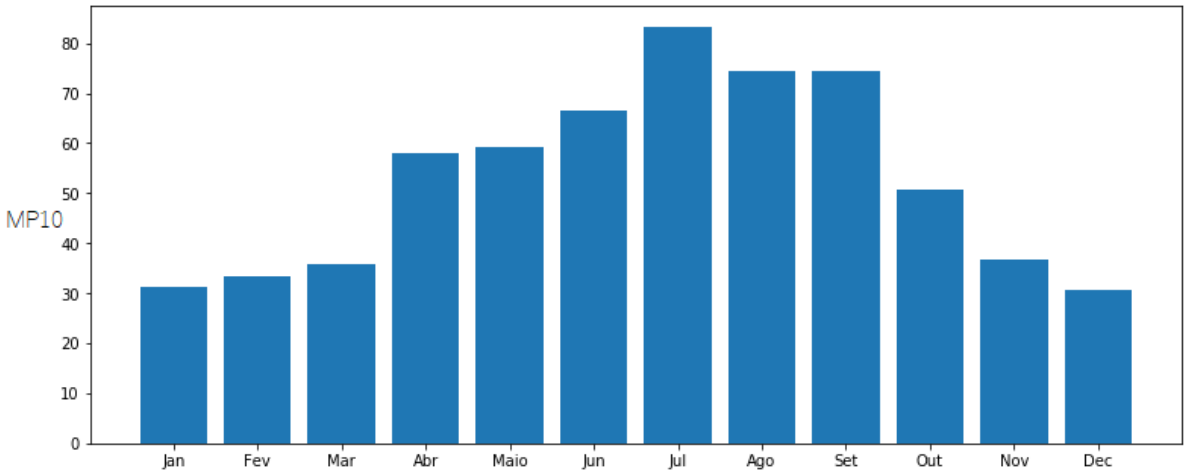
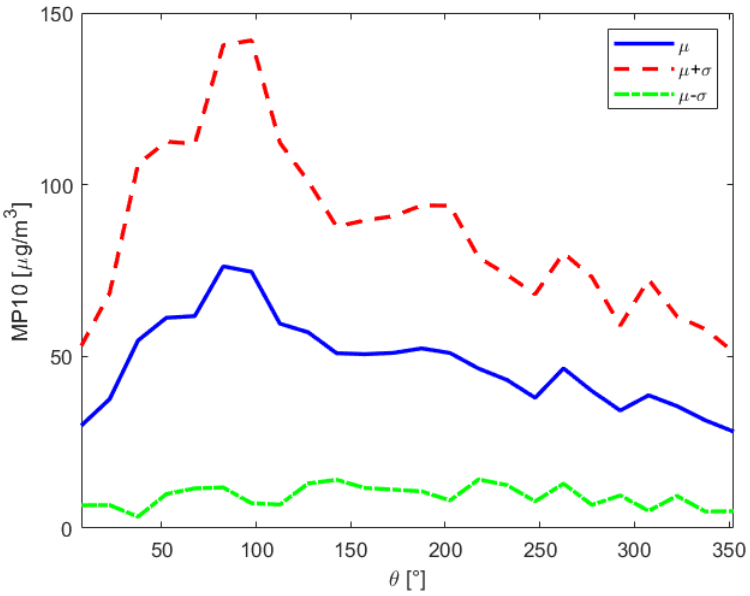


Figura 14: Perfil anual de MP10

em 24 classes, correspondendo a 24 subdivisões equivalentes da direção do vento, de 0° a 360°. Esta direção indica a proveniência dos ventos, sendo 0° o norte. Em seguida, calculou-se a média e o desvio padrão da concentração de MP_{10} de cada classe de direção dos ventos.

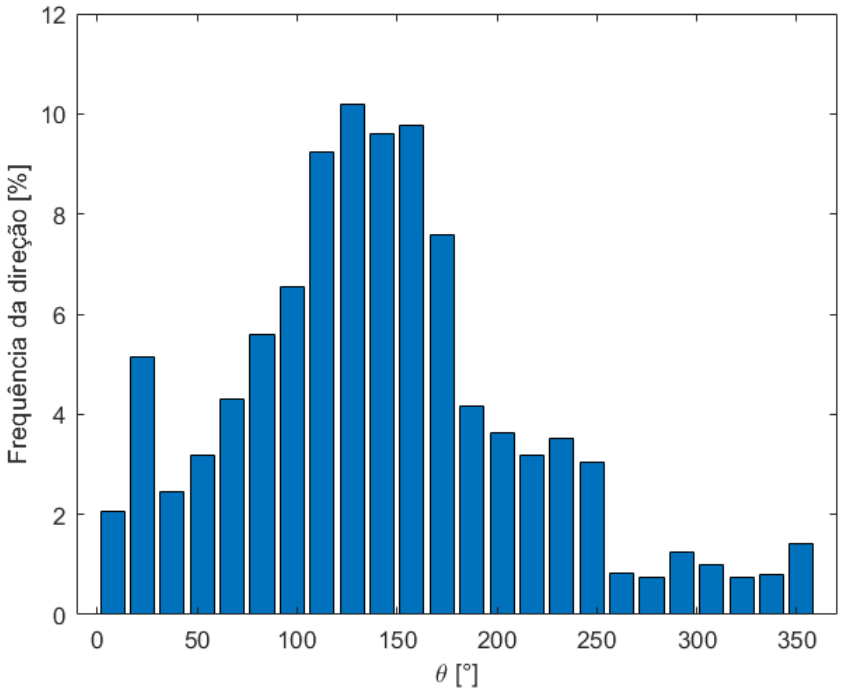
A Figura 15 apresenta curvas relativas à média da concentração para cada direção do vento, bem como uma margem de amplitude 2 σ. Nota-se que há um intervalo de ângulos que apresenta um pico na concentração média de MP_{10}, em torno de 100°, entre leste e lés-sudeste. Antes de se concluir que os ventos provenientes de tal direção induzam tal aumento na concentração de MP_{10}, deve-se avaliar a frequência de ocorrência de cada direção ao longo do tempo, análise que é apresentada na Figura 16.

Figura 15: Média das concentrações de MP_{10} em função da direção de origem dos ventos



Fonte: Autoria Própria

Figura 16: Distribuição da ocorrência de ventos em função da direção de proveniência



Fonte: Autoria Própria

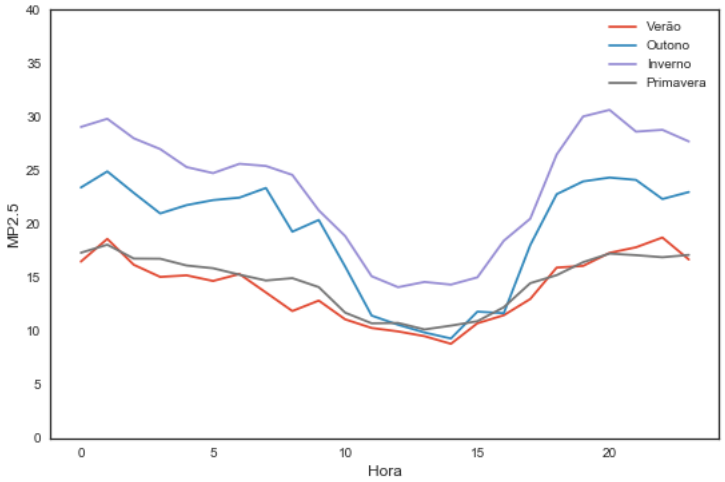
Constata-se que, ao longo dos anos de 2015 a 2019, as direções do vento compreendidas entre 100° e 160° apresentam a maior frequência de ocorrência, o que pode justificar em

parte o fato de ventos provindos dessa direção representarem o pico na concentração de MP_{10}, devido a um efeito cumulativo.

3.2.2 MP2.5

Realizando-se as mesmas análises para o material particulado de diâmetro de corte 2.5 μm, obtemos o gráfico abaixo. Tal como para o material particulado de maior diâmetro, o inverno e o outono destacam-se com maior concentração do poluente. Além disso, o perfil de menor concentração durante a tarde é comum aos dois.

Figura 17: Perfil de MP2.5 para Santa Gertrudes



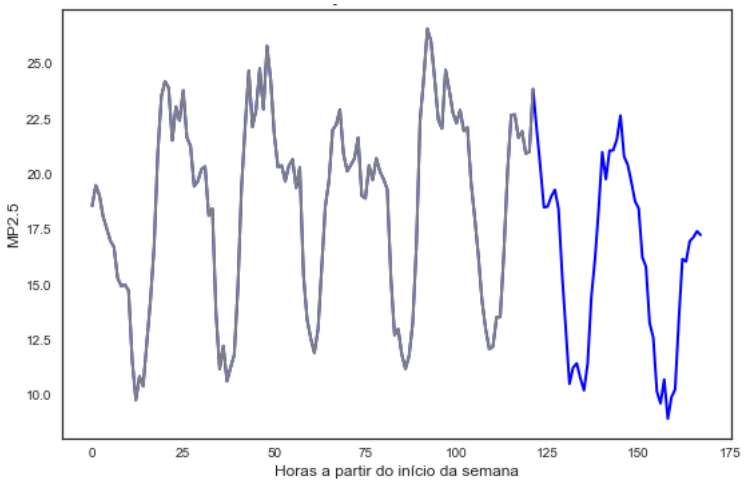
Fonte: Autoria própria

Ao analisarmos o perfil pelos dias da semana, não se observa um comportamento acentuado como no MP10. Nesse caso, a concentração aos sábados e domingos é muito semelhante à concentração durante os dias de semana. Uma hipótese para justificar tal diferença é o fato de que as partículas menores tem tendência a ficar mais tempo em suspensão, o que favorece a permanência de poluentes gerados durante os dias úteis.

Vale ressaltar que as partículas de menor diâmetro são ainda mais perigosas que as maiores, pois penetram nas vias respiratórias e atingem os alvéolos pulmonares. A recomendação da OMS para esse poluente é de 10 $\mu g/m^3$, embora as cidades em desenvolvimento apresentem, em geral, médias superiores. A CETESB estabelece como padrão anual 20 $\mu g/m^3$. No caso de Santa Gertrudes, dias acima do recomendado são a regra: 80,5% dos dias estão acima da recomendação da OMS.

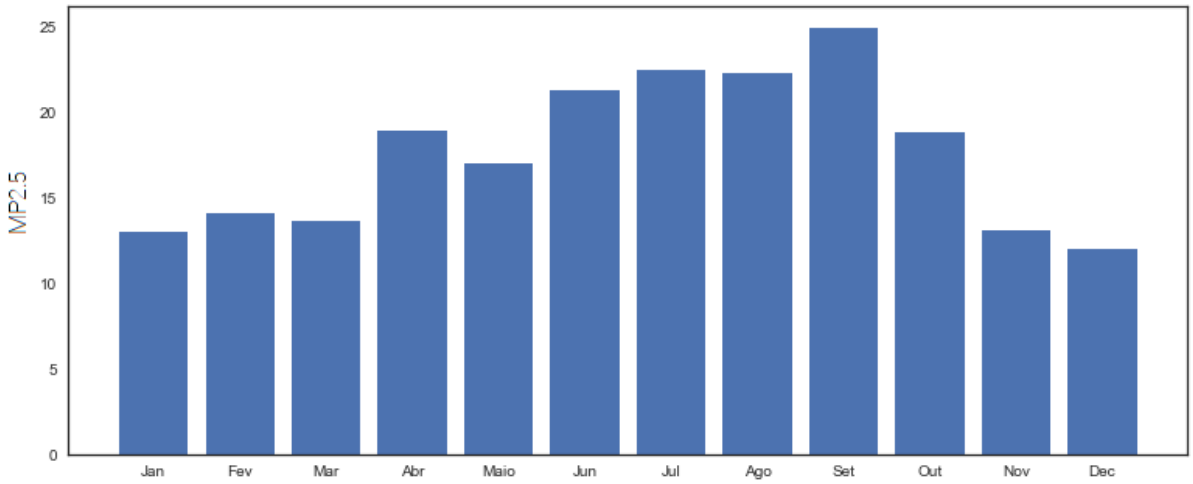
O perfil anual de MP2.5 está representado na figura 19. A concentração nos meses secos também é notável, mas de forma menos acentuada que para o MP10.

Figura 18: Perfil de MP2.5 para Santa Gertrudes



Fonte: Autoria própria

Figura 19: Perfil de MP2.5 anual para Santa Gertrudes



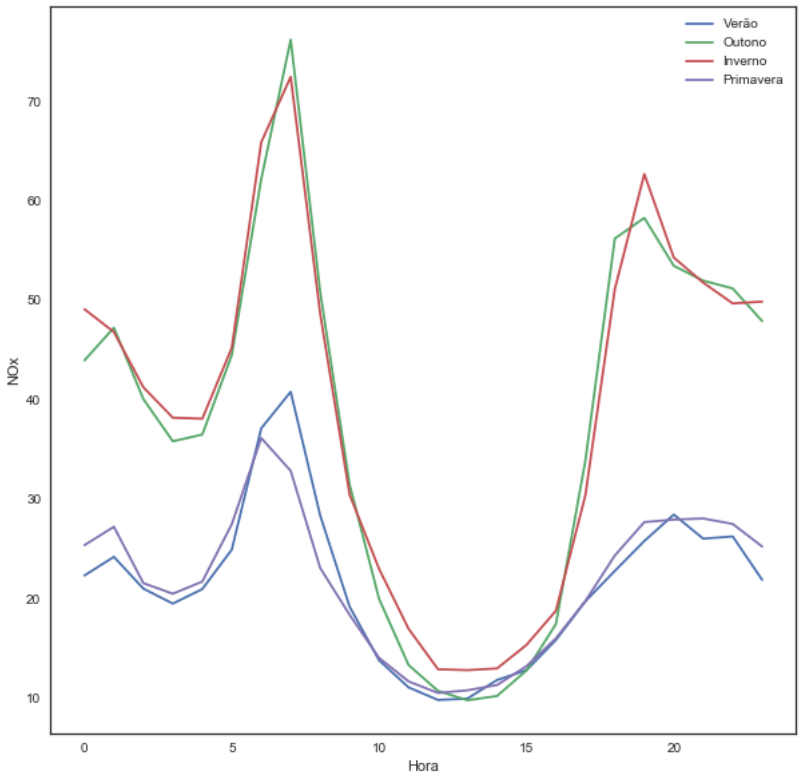
Fonte: Autoria própria

3.2.3 NOx

Como visto anteriormente, o NOx não é material particulado, trata-se um gás poluente. O seu perfil é um pouco mais complexo que o do material particulado, uma vez que pode passar por reações químicas com mais facilidade. Observamos dois picos, um próximo das 7h da manhã e outro próximo das 19h, o que provavelmente está ligado aos poluentes móveis na região (caminhões e carros). Além disso, é durante o inverno que possuímos os valores mais altos observados de NOx, o que se justifica pela maior quantidade de dias desfavoráveis a dispersão de poluentes e a maior emissão em dias frios.

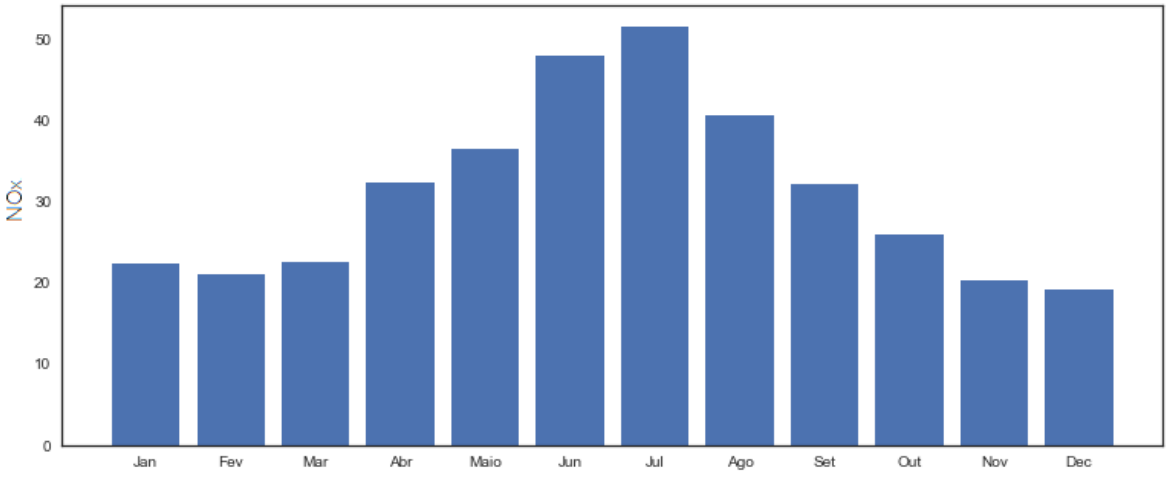
A distribuição anual é representada de forma mais clara na figura 21.

Figura 20: Perfil de NOx diário para Santa Gertrudes



Fonte: Autoria própria

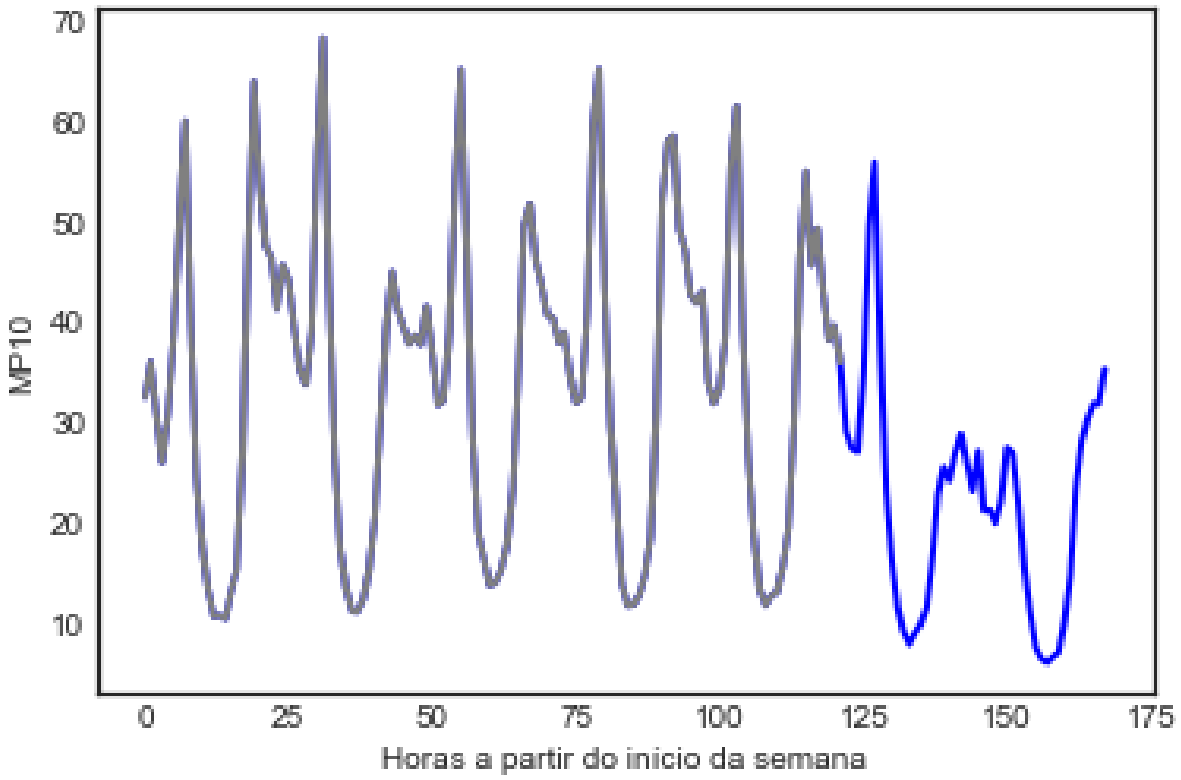
Figura 21: Perfil de NOx anual para Santa Gertrudes



Fonte: Autoria própria

Observando os dias da semana, o comportamento do NOx se assemelha ao comportamento do MP10, com diminuição abrupta no final de semana. Esse fenômeno está representado na figura 22.

Figura 22: NOx durante os dias da semana



Autoria própria

3.3 Decomposição de séries temporais para MP10

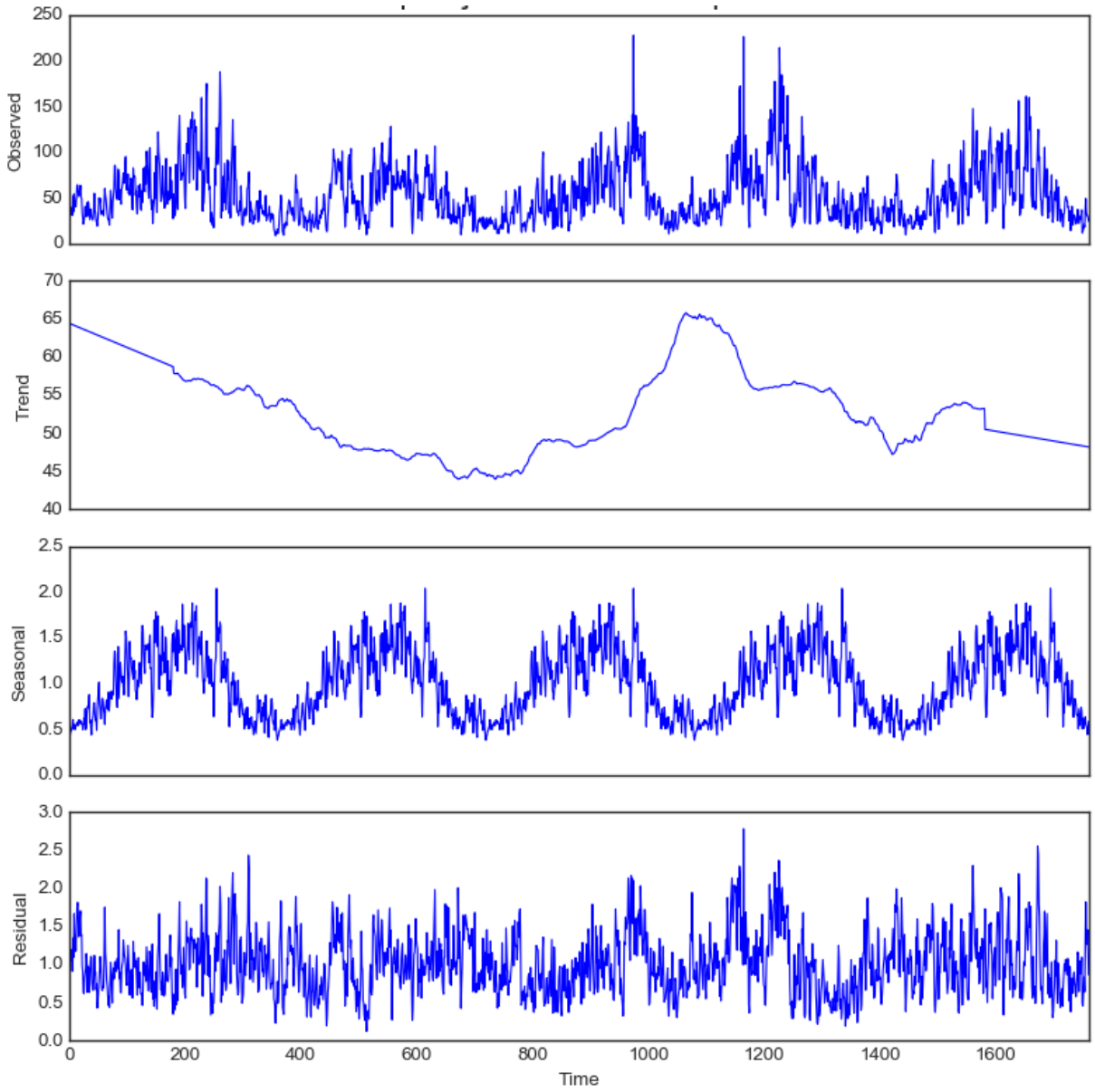
A técnica de detecção de outliers descrita neste trabalho não pode ser aplicada diretamente aos valores históricos de MP10. Uma vez que existe uma sazonalidade bem demarcada, a técnica sem desconsiderar esse efeito seria enviesada, apontando anomalias somente em momentos de maior concentração de poluentes, como no inverno e outono. Entretanto, é do nosso interesse identificar, também, momentos em que existiram anomalias dentro de estações com menor concentração de poluentes. [10]

Buscando compreender a distribuição histórica dos poluentes no tempo, estudou-se a decomposição da série temporal em seus três componentes: a tendência, a sazonalidade e o resíduo. O resíduo serve a entender melhor pontos de dados anômalos, uma vez que já não está influenciado pela sazonalidade.

Optou-se por realizar a decomposição multiplicativa, uma vez que o modelo se ajustou melhor aos dados, principalmente aos valores de 2016, onde uma pequena alteração na tendência (foi um ano de menor taxa de poluentes), também se refletiu em uma alteração da amplitude da sazonalidade. De todo modo, por ser uma variação muito pequena, os

resultados da decomposição aditiva e multiplicativa convergiram. Como a multiplicativa conseguiu identificar mais outliers, seguiu-se com esta.

Figura 23: Decomposição de série temporal de MP10



Fonte: Autoria própria

A aplicação foi feita atráves da ferramenta seasonal_decompose, que faz parte do pacote de modelos estatísticos do Python. A ferramenta baseia-se no método de média móvel, seguindo o algoritmo clássico descrito neste trabalho.Utilizou como período intervalos de um ano. Cada ponto da figura 23 representa um dia, percorrendo o período completo dos dados, de 2015 a 2019.

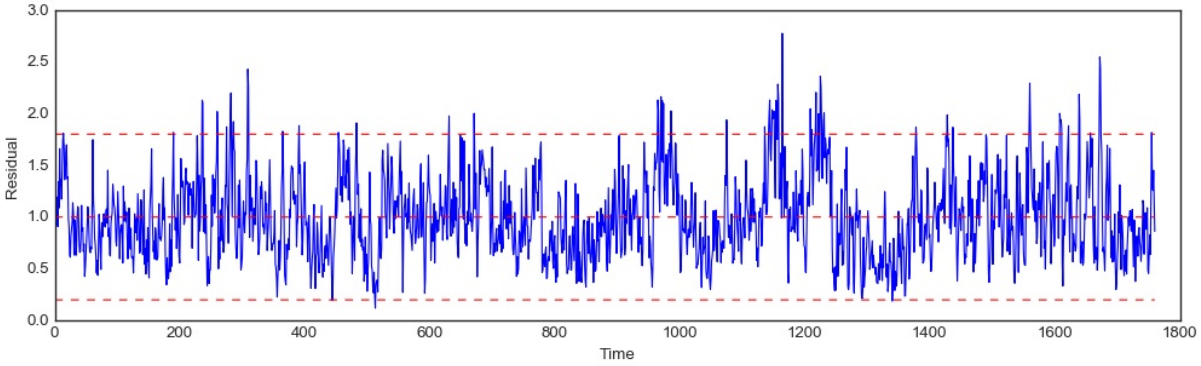
A análise de tendência é limitada pela escassez dos dados. Enquanto séries históricas

normalmente são compostas por mais de dez anos, dispomos somente de cinco anos de dados, para somente um dos poluentes. Por essa razão, a ocorrência de um ano com menor concentração (o segundo pico, ano de 2016) cria um grande impacto na tendência, que tem um declive e em seguida volta a seguir, oscilando em torno da média, por volta de 55 μg/m^3.

Ao mesmo tempo, a análise de sazonalidade deixa bem claro o perfil estudado anteriormente: o ano é marcado por um forte pico durante as estações secas e frias, entre abril e setembro. Por fim, a análise de resíduo representa quanto daquela variação não é explicada pela sazonalidade nem pela tendência. Com os resíduos analisaremos a ocorrência de outliers.

Optou-se por utilizar os resíduos para detecção de anomalias pois os resíduos permitem identificar outliers sem influência da tendência e sem influência da sazonalidade, o que permite a identificação de anomalias intra-anuais. Se houvéssemos considerado a sazonalidade, anomalias no verão não seriam identificadas. Se considerássemos a tendência, o ano de 2016, que mostrou uma leve baixa, também não teria pontos detectados. Os outliers dos resíduos mostram pontos que se sobressaem em relação ao esperado para aquele instante.

Figura 24: Análise de Outliers



Fonte: Autoria própria

Na figura 24, as linhas pontilhadas em vermelho representam o intervalo de valores que estão dois desvios padrões acima ou abaixo da média – representada pela linha central.

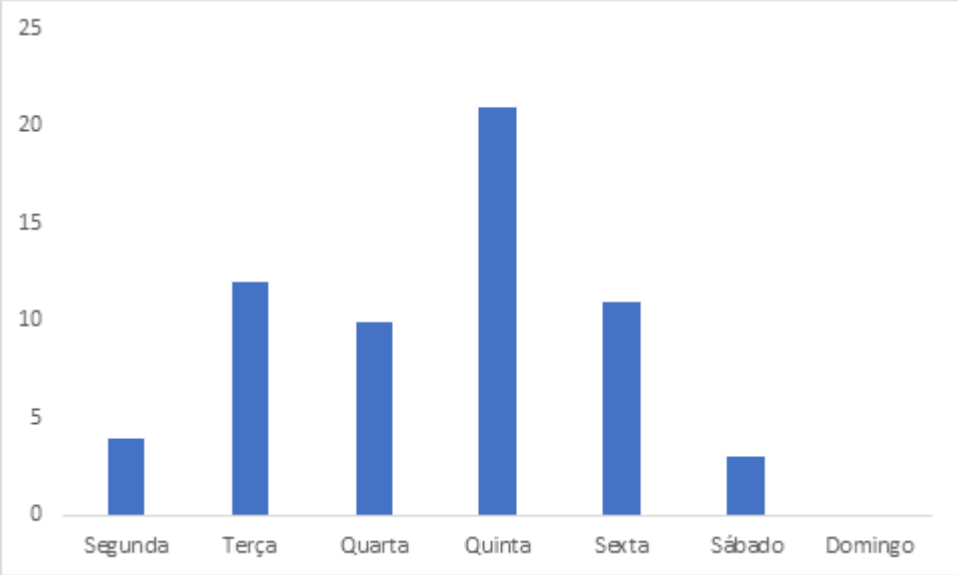
Em seguida, tratam-se os pontos considerando como outlier todos os valores superiores à linha da média com dois desvios padrões, e todos os valores inferiores à linha inferior do gráfico. Esses são valores que tiveram comportamentos muito diferente do valor esperado para determinado período no município, podendo estar relacionado a uma variação

meteorológica pontual combinada à atividade econômica da região, ou a acontecimentos externos, como uma queimada na região.

Analisando os pontos identificados como anômalos, observa-se que a maioria deles ocorreu durante dias de semana, com somente 4.9% das ocorrências aos sábados, e nenhuma aos domingos. Há uma forte indicação, portanto, que ocorrências externas representam uma parcela pouco significativa dos comportamentos anômalos, que são, em sua maioria, ligados à atividade econômica que ocorre durante a semana.

O gráfico abaixo indica a distribuição de ocorrências por dia da semana.

Figura 25: Ocorrência de outlier por dia da semana



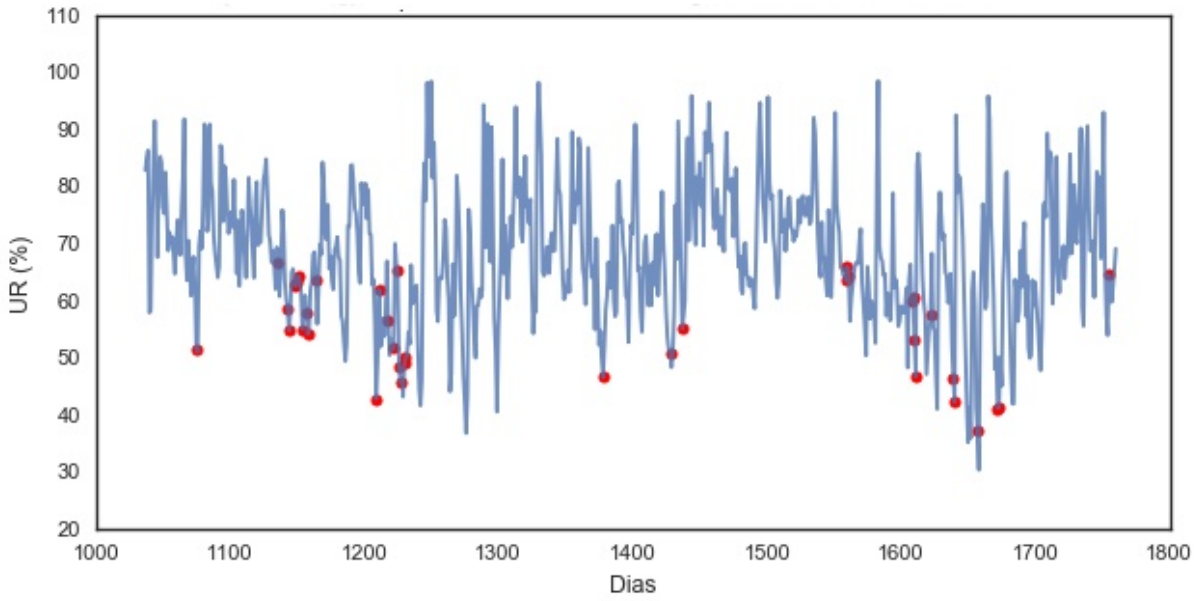
Fonte: Autoria própria

Além disso, a maior parte das ocorrências é durante a estação seca, quando há menor dispersão de poluentes. O comportamento de dispersão colabora para justificar, também, a menor taxa de ocorrência de outliers na segunda-feira: a quantidade de particulado acumulada durante o final de semana não é alta suficiente para gerar um ponto anômalo.

A figura 26 mostra que a maior parte da ocorrência de extremos de particulados ocorre em dias de baixa umidade, sendo todos os dias de ocorrência abaixo da média histórica de umidade relativa. Os dados acima correspondem ao período de 2 anos, limitados pela medição da umidade relativa, de 2018 a 2020, com os dias sendo contados a partir de 2015 (início da série de MP10). Cada ponto vermelho corresponde a um outlier identificado pelo tratamento da série residual.

É possível perceber, também, que em momentos em que a baixa umidade se prolonga por alguns dias, a ocorrência de extremos torna-se mais frequente, como ocorre entre os

Figura 26: Ocorrência de outlier na série de Umidade Relativa

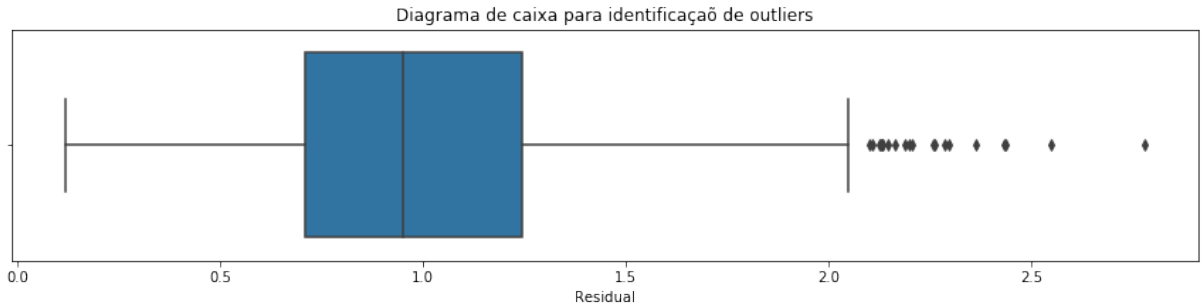


Fonte: Autoria própria

dias 1100 e 1200 no gráfico. Isso ocorre pois os dias secos dificultam a dispersão dos poluentes.

Outra forma de identificação de dados anômalos para uma única variável é realizar uma análise do diagrama de caixa (boxplot) que revela a distribuição dos pontos em quartis, assim como os pontos que não se encaixam nos quartis.

Figura 27: Ocorrência de outlier na série de MP10



Fonte: Autoria própria

Como se observa, o corte para detecção de outliers foi maior no segundo caso, considerando apenas os pontos residuais que superaram o índice 2.1. Vale ressaltar que esse índice multiplica a sazonalidade e a tendência, fornecendo o valor final do poluente. Quanto mais alto, maior em relação ao esperado para aquele momento.

Nesse caso, os dados revelados foram iguais aos da análise residual, com exceção de dois pontos que escaparam ao corte.

Como os dados para MP2.5 e NOx foram coletados somente a partir de 2018, não há sentido em realizar uma análise de série histórica para esses dois parâmetros. Mesmo para o MP10, observa-se que seria vantajoso dispor de mais alguns anos de dados para ter análises de tendência e sazonalidade mais precisas.

3.4 Regressão multivariável para MP10

Com os dados do banco Santa Gertrudes 2, excluíram-se os pontos faltantes e obteve-se um dataframe onde cada linha é a observação de uma hora. As 5 primeiras observações estão ilustradas na tabela 3.

Tabela 3: Observações banco de dados

ID	Data	Hora	DV	MP10	MP2.5	NOX	UR	Ano	Temp
3764	2018-08-10	1	179.0	27.0	8.0	38.6	74.0	2018	14.2
3765	2018-08-10	2	218.0	23.0	7.0	17.0	73.0	2018	13.0
3766	2018-08-10	3	202.0	22.0	9.0	24.0	75.0	2018	12.0
3767	2018-08-10	4	258.5	23.0	7.0	33.0	79.0	2018	10.9
3768	2018-08-10	5	277.3	25.0	12.0	27.0	86.0	2018	10.0

Em seguida, realizamos a padronização das variáveis, para que as diferentes ordens de grandeza não influenciem na escolha dos coeficientes angulares. O MP10, que é a variável target que se deseja prever, não será padronizado. A padronização está descrita na equação seguinte, e consiste em passar todas as variáveis para média 0 e desvio padrão 1.

$$z = \frac{x - \mu}{\sigma} \tag{3.1}$$

O próximo passo foi realizar a criação de novas variáveis que podem auxiliar na regressão. Nesse caso, a partir da data, geraram-se três novas variáveis: mês, estação e final de semana. A última é uma variável binária, que indica 1 se for final de semana e 0 se for dia de semana. As variáveis 'Ano' e 'Data' foram descartadas, pois não se pode levar em consideração valores específicos a uma observação.

A tabela passa a ser como a mostrada abaixo, que traz apenas as 2 primeiras observações.

Tabela 4: Dados após criação de novas features e padronização

ID	Hora	DV	MP10	MP2.5	NOX	UR	Mês	Temp	Weekday	Estação
0	1	0.14	27.0	-0.72	0.32	0.28	8	-1.60	0	3
1	2	0.57	23.0	-0.79	-0.47	0.24	8	-1.81	0	3

Realizou-se então a seleção de variáveis que seriam utilizadas no modelo. Para isso, aplicaram-se os dois métodos estudados neste trabalho: Lasso e OLS.

A aplicação da técnica OLS indica que todas as variáveis, exceto 'Estação' são significativas na explicabilidade da variável que se deseja prever. A escolha dos coeficientes por Lasso indicou que a variável 'Weekday', binária, também poderia ser excluída, junto com a estação do ano. Antes de passar para o ajuste do modelo de regressão, é preciso verificar se essas variáveis escolhidas satisfazem o critério de multicolinearidade. Para isso, olhamos a correlação entre as variáveis.

Tabela 5: Correlação entre variáveis

	Hora	DV	MP2.5	NOX	UR	Mês	Temp	Weekday	Estação
Hora	1.00	-0.04	-0.06	-0.05	**-0.45**	-0.00	**0.42**	-0.00	-0.00
DV	-0.04	1.00	-0.09	-0.13	-0.01	-0.03	0.08	0.00	-0.03
MP2.5	-0.06	-0.08	1.00	**0.54**	-0.02	-0.01	-0.14	-0.07	-0.00
NOX	-0.06	-0.12	0.54	1.00	0.17	-0.06	**-0.41**	-0.18	-0.05
UR	-0.45	-0.01	-0.02	0.17	1.00	-0.06	-0.74	-0.02	-0.06
Mês	-0.01	-0.03	-0.01	-0.06	-0.06	1.00	0.00	0.01	0.97
Temp	**0.42**	0.09	-0.15	**-0.41**	**-0.75**	0.00	1.00	-0.00	0.00
Weekday	-0.00	0.00	-0.08	-0.18	-0.02	0.01	-0.00	1.00	-0.00
Estação	-0.00	-0.03	-0.01	-0.05	-0.06	0.97	0.00	-0.00	1.00

Nota-se que a temperatura e a umidade relativa possuem correlação alta, e as duas possuem correlação média com a hora. A temperatura possui maior correlação com a variável que desejamos prever, portanto iremos descartar, inicialmente, a observação da umidade relativa. O MP2.5 e NOx também possuem alta correlação. Iremos estudar modelos com e sem o NOx (guardando o MP2.5 pois possui maior correlação com o MP10).

Para realizar a comparação dos testes, utilizaremos o Mean Squared Error (MAE) e a função score do Python, que fornece o valor R^2. A configuração de treino seleciona valores aleatórios, deixando 10% de não treinados que em seguida precisam ser previstos. Como os valores são aleatórios, o modelo se ajusta de forma diferente a cada vez, retornando acurácias diferentes a cada iteração. Por isso, para cada experimento, rodou-se mil vezes o código, obtendo, no final, uma acurácia média e score médio.

As tabelas a seguir descrevem os experimentos e os resultados obtidos.

Tabela 6: Descrição dos experimentos

Regressão OLS - Dados aleatórios									
Experimento	DV	MP2.5	NOX	UR	Temp	Weekday	Estação	Hora	Mês
1	X	X	X		X	X		X	X
2	X	X			X	X		X	X
3	X	X	X	X	X	X		X	X

Tabela 7: Resultados

	Experimento		
Métrica	1	2	3
Score R^2	0.68	0.56	0.71
MAE	15.19	17.19	14.41

Observa-se, com isso, que o melhor resultado foi obtido quando se ignorou os possíveis efeitos da multicolinearidade. Embora apresente um resultado melhor, com maior acurácia e menor erro absoluto médio, não seria possível se basear nos coeficientes dessa regressão para entender o impacto individual de cada parâmetro.

A regressão Lasso forneceu resultado semelhante ao do experimento 3, com MAE médio de 14.45 e score R^2 de 0.70. Os ajustes fornecidos para as regressões foram:

Tabela 8: Coeficientes angularer da regressão

	DV	MP2.5	NOX	UR	Temp	Weekday	Hora	Mês
OLS - Best	-1.35	20.81	19.22	-10.90	-4.35	-1.166	0.28	0.19
Lasso	-0.59	20.19	19.44	-7.13	-0.55	0.00	0.28	0.23

Como esperado, os maiores coeficientes, que influenciam mais na definição da variável preditiva, são os outros dois poluentes, MP2.5 e NOx, que possuíam a maior correlação com a variável que se desejava prever. Alguns pontos merecem ser discutidos: se a sazonalidade é tão marcada, como explicar mês e hora com baixa influência no ajuste, com coeficientes próximos de zero?

O que ocorre é que, embora esses dois parâmetros sejam importantes na definição do MP10, as outras variáveis meteorológicas e poluente presentes no modelo já possuem, de forma intríseca, o efeito da sazonalidade, que se reflete no ajuste final. Isso significa, de outra perspectiva, que seria possível estimar, a partir de uma medição qualquer, o mês e a hora desta.

Para avaliar se a técnica multivariável é uma opção válida para imputar dados, comparamos esse método com o método inicial do tratamento de dados, onde utilizávamos a

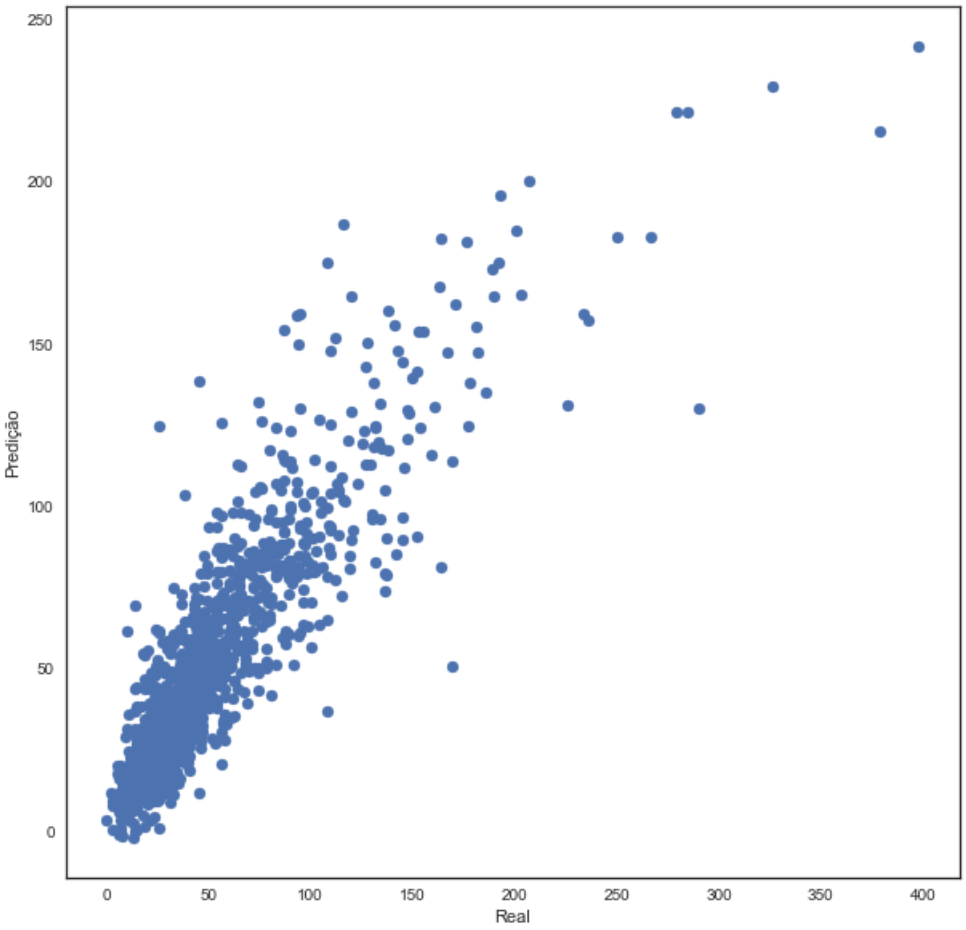
comparação com a média dos valores do dia. Esse método será identificado como "Interpolação", enquanto a regressão multivariável é a "Predição".

A técnica da interpolação utiliza uma matriz 12x24 para prever os valores. Cada elemento da matriz representa a proximidade da média diária para um mês específico. Assim, para um valor faltante, a interpolação verifica os outros valores do dia, realiza a média e em seguida multiplica pelo elemento correspondente da matriz 12x24. Não se utilizou a média direta entre os pontos mais próximos do vazio pois muitas vezes se observam vazios consecutivos.

Comparando os valores de MAE, a regressão apresentou um valor médio de erro 28% menor do que a interpolação, que foi de 18.50 $\mu g/m^3$.

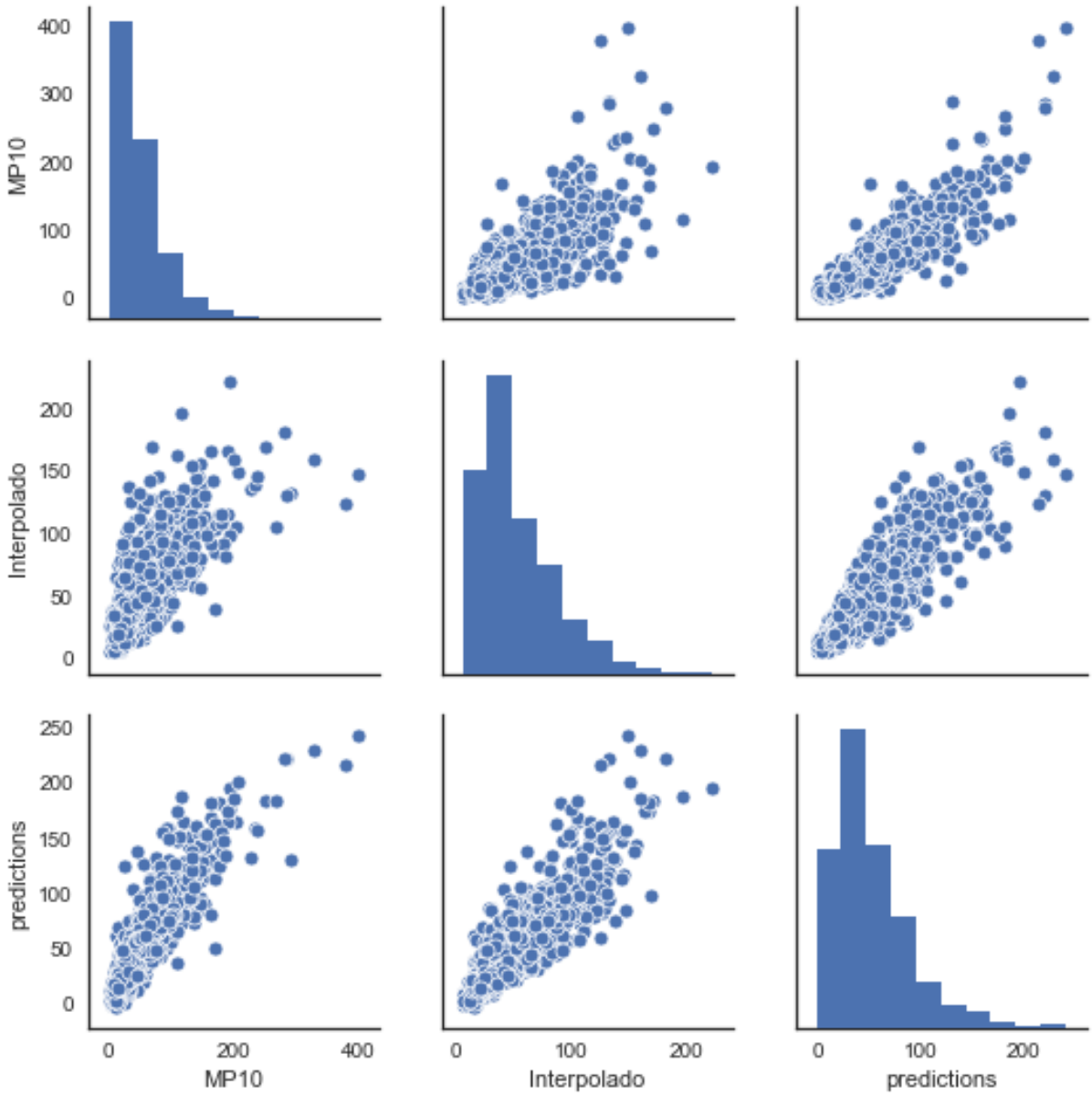
Observou-se que tanto a interpolação quando a predição se aproximam do perfil real de MP10, mas não conseguem realizar uma boa aproximação dos pontos de pico, onde ocorrem mais outliers. Valores de MP10 acima de 200 $\mu g/m^3$ começam a apresentar erros maiores. É possível observar, na primeira linha da figura 29, que a interpolação começa a mostrar seus limites de predição mais cedo do que as predições, que apresenta um formato eliptico mais alongado.

Figura 28: Distribuição entre previsto e realizado



Fonte: Autoria própria

Figura 29: Distribuição entre previsto, interpolado e realizado



Fonte: Autoria própria

3.5 Análise de componentes principais

O objetivo dessa análise é estabelecer quais são os principais fatores que seriam levados em conta para explicar a variabilidade do sistema como um todo, e realizar a fundação do que seria necessário para uma análise de clusters.

Como explorado neste trabalho, a técnica PCA tem a hipótese de normalidade dos componentes. Embora as variáveis tenham perfil próximo da normalidade, as distribuições são assimétricas e não correspondem a condições ótimas para a aplicação do método. De todo modo, a análise traz algumas informações importantes.

Realizando-se o PCA em Python, para as variáveis DV, MP10, MP2.5, NOX, UR e Temperatura, todas agrupadas por dia (cada dia é representado pela média dos valores observados de cada uma dessas variáveis) obteve-se que seriam necessários 3 componentes principais para alcançar 90% da variabilidade do sistema explicada. O primeiro componente explica 50%, enquanto o segundo e o terceiro explica 23% e 17%, respectivamente.

Isso significa que conseguimos reduzir a dimensão da nossa análise de 6 variáveis para 3 variáveis, guardando a maior parte da informação, cumprindo o objetivo principal da PCA. Diria-se que não houve ganhos com PCA se a análise resultasse no mesmo número de variáveis para explicar a maior parte da variância,o que indicaria que as variáveis já eram ortogonais entre si.

Observando-se os fatores (ou loadings) de cada componente, obtem-se o coeficiente linear das variáveis iniciais que formam a nova variável. De um ponto de vista numérico, os fatores são iguais às coordenandas das variáveis divididas pela raiz quadrada do autovalor associado ao componente principal. A tabela a seguir indica os fatores para cada um dos 3 componentes principais encontrados.

Tabela 9: Tabela de fatores PCA

	PC1	**PC2**	**PC3**
DV	-0.14	-0.20	-0.95
MP10	0.55	0.05	-0.11
MP25	0.51	-0.09	-0.15
NOX	0.47	0.36	-0.13
UR	-0.41	0.40	-0.15
Temp	0.02	-0.80	0.11

Como se pode observar, o MP10 é o de maior peso para o PC1, a temperatura é o de maior peso para o PC2 e a direção do vento é a principal para a PC3. Os efeitos do NOX e do MP2.5 são contabilizados majoritariamente no PC1, junto com o MP10.

4 CONSIDERAÇÕES FINAIS

4.1 Conclusões

A análise de dados de ar obtidos pela CETESB se mostram úteis para modelizar e compreender o perfil de poluição no município de Santa Gertrudes, em escala anual ou diária. As análises de perfil de poluentes mostram grandes picos durante o dia, causados pela maior circulação de veículos leves e pesados em horário de pico. A comparação com perfil de poluentes de São Paulo, que também possui horário de pico de circulação de veículos mas não possui picos tão marcados quanto Santa Gertrudes, mostra a unicidade da região, com nível de poluição muito maior que a capital, com picos nos horários de entrada e de saída dos caminhões.

Vale ressaltar que há uma limitação a nível de dados, pois embora conte com três estações CETESB, o município de Santa Gertrudes só possui uma estação automática, com alguns poluentes sendo medidos somente a partir de agosto de 2018. As outras estações são manuais e contam somente com medições espaçadas de 1 semana ou 1 mês, que não agregam valor à análise de dados.

A correlação da atividade econômica da região com a poluição também fica explícita ao examinar a poluição por dias de semana. Em finais de semana, os níveis de poluição são significativamente menores, alcançando níveis aceitáveis para OMS. Isso é válido principalmente para o MP10 e o NOX, enquanto o MP2.5 possui presença mais longa, sem ter uma diminuição tão marcada nos finais de semana.

A decomposição da série histórica mostrou-se uma ferramenta útil para a detecção de outliers, através da remoção da sazonalidade. Seria possível realizar previsões ou alertas a partir disso, informando a ocorrência de observação superior à esperada naquele momento do ano. É importante ressaltar que as séries históricas ganham robustez com o aumento do período observado. Embora tenha sido possível obter o perfil de sazonalidade e a tendência, seria fundamental ter maior quantidade de anos para realizar métodos confiáveis de predição (conhecidos como forecasting), como ARIMA.

A regressão multivariável mostrou-se válida para predição de valores horários de MP10, podendo se tornar uma ferramenta útil no preenchimento de dados faltantes. A técnica mostrou-se mais eficaz do que a técnica inicial de imputação a partir de modulação horária, sazonalidade e média das horas próximas ao ponto faltante. Ao mesmo tempo, sua utilização está limitada aos momentos de falha do parâmetro MP10 mas funcionamento dos outros. O que se observou é que, na maior parte das falhas, é uma falha geral do equipamento, perdendo-se todos os dados.

A análise de componentes principais estabelece uma base para aplicação de agrupamentos, que pode ser tornar uma ferramenta na compreensão dos fenômenos de poluição. As primeiras análises realizadas foram inconclusivas e, portanto, não fizeram parte desse trabalho.

4.2 Próximos passos

Esse trabalho serve de ponto de partida para mais estudos no munício de Santa Gertrudes. No escopo de aprofundamento dos estudos para o município, possíveis próximos passos seriam: realização de análises através de redes neurais, para substituir a regressão multivariável e obter método mais robusto de previsão, técnicas de agrupamentos para os dados coletados, utilização da série histórica para realizar previsões com meses de antecedência e, por fim, técnica de manutenção preventiva que informasse quando o aparelho da estação coletora possui maior chance de quebrar e perder observações.

No escopo geral, o campo é muito e existe uma série de testes a serem feitos. Algumas análises realizadas nesse trabalho poderiam ser ampliadas para outras cidades e regiões. Em locais sem medição de MP10, poder-se-ia estudar a estimativa desse parâmetro através de regressões multivariáveis aplicadas nos outros parâmetros medidos, utilizando ajustes de outros municípios com comportamentos semelhantes. Outros comparativos poderiam ser traçados, como por exemplo a influência das condições meteorológicas e os perfis diários entre cidades com diferentes atividades econômicas.

REFERÊNCIAS

[1] SEIGNEUR, C. Atmospheric dispersion. In: _____. *Air Pollution: Concepts, Theory, and Applications*. [S.l.]: Cambridge University Press, 2019. p. 95–124.

[2] WATSON, S. A. Y.; BATES, M. R. R.; KENNEDY, P. D. (Ed.). *Air Pollution, the Automobile, and Public Health*. Washington, DC: The National Academies Press, 1988. ISBN 978-0-309-08682-0. Disponível em: <https://www.nap.edu/catalog/1033/air-pollution-the-automobile-and-public-health>.

[3] SEINFELD, J.; PANDIS, S. *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*. Wiley, 2016. ISBN 9781119221166. Disponível em: <https://books.google.com.br/books?id=MfHbCwAAQBAJ>.

[4] FISHER, R. A. Design of experiments. *Br Med J*, British Medical Journal Publishing Group, v. 1, n. 3923, p. 554–554, 1936.

[5] YAMANO, T. Multivariate regression model in matrix form. *Advanced Econometrics class notes*, Unknown, 2012.

[6] JAMES, G.; WITTEN, D. (Ed.). *An Introduction to Statistical Learning*. Springer, 2013. ISBN 978-1-4614-7137-0. Disponível em: <http://faculty.marshall.usc.edu/gareth-james/ISL/ISLR

[7] KOURENTZES, N. Time series decomposition. *Kourentzes blog*, Nikolaos Kourentzes, v. 1, n. 1, p. 1, 2014.

[8] HYNDMAN, R. J.; ATHANASOPOULOS, G. Forecasting: Principles and practice. OTexts: Melbourne, Australia, 2018. Disponível em: <https://otexts.com/fpp2/>.

[9] HOLMES, E. E.; SCHEURELL, M. D.; WARD, E. J. *Applied Time Series Analysis for Fisheries and Environmental Sciences*. [S.l.]: NOAA Fisheries, 2020.

[10] ETIENNE, B. Time series in python: Dealing with seasonal data. *Towards Data Science*, 2019. Disponível em: <https://towardsdatascience.com/time-series-in-python-part-2-dealing-with-seasonal-data-397a65b74051>.