Universidade Estadual de Campinas

Faculdade de Engenharia Elétrica

EA006 - Trabalho de Fim de Curso

Estudo exploratório do impacto da Poluição na Saúde da População de Campinas: uma abordagem baseada em Data Science

Autor: João Victor Gitti Arêdes

Orientadora: Profª Dra. Paula Dornhofer Paro Costa

CAMPINAS, DEZEMBRO DE 2020

Resumo

A tecnologia está, cada vez mais, presente em todos os âmbitos de pesquisa e desenvolvimento. Conseguimos, através dela, chegar a resultados e conclusões que antes não se era possível, muito devido agora ao grande número de processamentos que conseguimos realizar e também ao grande volume de dados que conseguimos tratar atualmente.

Esse trabalho se utiliza de ferramentas de programação, atreladas a conhecimentos científicos, para lidar com um grande volume de dados, e assim gerar conhecimento e informação para a área da saúde.

Para se gerar esses resultados, cruzou-se dados de poluição provenientes da CETESB [1], de estações de Campinas, através do site da QUALAR [2], e dados de óbitos, também da cidade de Campinas, disponibilizados pelo governo para o projeto, de modo que esses dados foram tratados e analisados. Temos por resultado que o poluente MP 2,5 influencia na saúde de pessoas com doenças respiratórias, impactando no número de óbitos de CID J9.

Lista de Figuras

Lista de Tabelas

Sumário

1 Apresentação

O presente relatório visa atender os requisitos da disciplina “EA006 - Trabalho de fim de curso"e apresentar um resumo das principais atividades e resultados obtidos ao longo do ano de 2020.

As atividades conduzidas durante este período fazem parte do projeto Clima&Saúde.

As seções seguintes contemplam as etapas de descrição do projeto, a origem dos dados analisados, o preprocessamento dos dados, a análise dos dados e a discussão dos resultados.

Todos os códigos deste estudo podem ser encontrados no repositório do GitHub:

https://github.com/climate-and-health-datasci-Unicamp/air-pollution-respiratory-effects

2 Introdução

Estudos apontam que doenças cardiovasculares e respiratórias são responsáveis por uma grande porcentagem das mortes nas grandes regiões metropolitanas. Doenças respiratórias foram responsáveis por 18,6% das mortes em idosos e por 36,2% das mortes em crianças, em São Paulo. Doenças cardiovasculares representaram 47,3% das mortes em idosos [4].

Um componente da poluição do ar é o material particulado (PM) em suspensão, uma mistura de partículas sólidas e/ou líquidas. O PM2.5 é representado pelo sub grupo mais fino dessas partículas, que são respiráveis, possuindo diâmetros de 0.1 - 2,5µm. São compostas por carbono, produzido pela combustão de combustível fóssil, além de outros elementos, incluindo metais pesados e hidrocarbonetos. A exposição a componentes específicos da poluição podem ter efeitos diferenciais sobre a saúde humana. Estudos mostraram que a exposição a PM aumenta a mortalidade por causas cardiovasculares e respiratórias, e que dentre o PM10 (partículas inaláveis grossas) e o PM2.5, o PM2.5 apresenta um maior risco para a saúde [5].

Os perfis de poluição desses poluentes podem variar significativamente para diferentes cidades/regiões. Isso ocorre porque, além da diferença de intensidade de emissão de poluentes para o ar pelas diferentes regiões, as diferentes condições meteorológicas também influenciam nas concentrações do material particulado, como foi relatado por um estudo realizado na região metropolitana do Rio de Janeiro [6].

Pelos motivos citados acima, é importante e relevante estudar os impactos dos índices de poluição para a saúde na cidade de Campinas a fim de informar as autoridades de políticas públicas, bem como possivelmente gerar alertas para o sistema de saúde.

3 Objetivos

O objetivo geral deste trabalho é estudar as correlações existentes entre períodos nos quais os níveis de poluentes encontram-se com valores extremos na cidade de Campinas e o número

de óbitos associados, em comparação a dias nos quais os poluentes encontram-se abaixo dos valores considerados nocivos à saúde humana.

São objetivos específicos deste trabalho:

• Criar base de dados integrado com dados de poluentes e de saúde;
• Extrair das bases estatísticas descritivas que descrevam, incluindo visualizações, as características dos poluentes e número de óbitos por doenças respiratórias em Campinas ao longo dos anos;
• Conduzir análises estatísticas de correlação entre parâmetros de poluentes e de desfechos na saúde.

4 Metodologia

Este trabalho se utiliza de conceitos de estatística, aplicados através da ciência de dados, que por si é implementada por meio de ferramentas de programação, para se testar as hipóteses de interesse. Tal processo é descrito nos próximos tópicos.

4.1 Hipótese

• Questão: A poluição possui influência sobre o número de mortes por doenças respiratórias?
• Hipótese nula: A poluição não possui influência sobre o número de mortes por doenças respiratórias.
• Hipótese validada: A poluição possui influência sobre o número de mortes por doenças respiratórias.

4.2 Ciência de dados

Ciência de Dados é um estudo muito disciplinado com relação aos dados e demais informações que cercam um determinado assunto. Em resumo é uma ciência que visa estudar as informações, seu processo de captura, transformação, geração e, posteriormente, análise de dados [7], como é mostrado na figura abaixo.

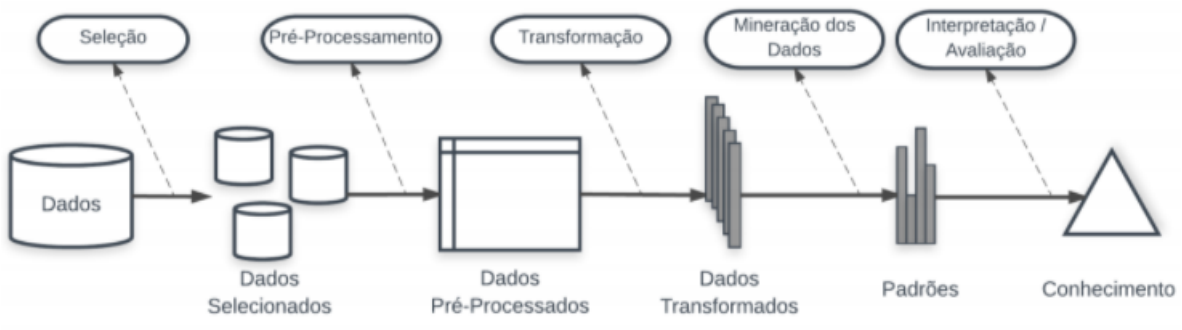

Figura 1: Processo completo da análise de dados [3]

.

O processo consiste nos seguintes passos que buscam extrair informação de base de dados:

• Passo 1 (Entendimento do Problema): Nessa etapa o objetivo é entender primordialmente o motivo que justifica o processo , ou as perguntas de pesquisa que guiarão o processo, do ponto de vista de quem ou o quê utilizará esse conhecimento. No contexto deste trabalho, esta etapa foi cumprida por meio do estudo de artigos relacionados à saúde da população, e de artigos relacionados aos poluentes presentes no ar;

• Passo 2 (Criação da base alvo): Selecionar as bases de dados de interesse, que possuam as variáveis a serem estudadas. Essa etapa é abordada com mais profundidade na Seção 5.1;

• Passo 3 (Limpeza e preprocessamento de dados): Retirar dados com ruídos ou inconsistentes, tratar valores faltantes e processar a base para uma forma utilizável. O processo é descrito na Seção 5.2;

• Passo 4 (Transformação dos dados): Encontrar ou transformar, se necessário, informações na sua base que melhor a representa, dependendo do seu objetivo. Processo abordado na Seção 5.3;

• Passo 5 (Checagem de objetivo): Checar se os tratamentos realizados até então estão de acordo com os objetivos do Passo 1. Podendo, em caso de negativa, repetir passos anteriores;

• Passo 6 (Análise exploratória): Definir metodologias de análises e de seleção para aplicar na base, visando o objetivo principal. No contexto desse projeto esta etapa consistiu em analisar hipóteses levantadas, descritas na Seção 4.1, de forma a validá-las quantitativamente através de análises gráficas e com estatísticas descritivas, descritas na Seção 4.3.

• Passo 7 (Mineração de dados): Explorar os dados à procura de padrões consistentes, para detectar relacionamentos sistemáticos entre variáveis, detectando assim novos subconjuntos de dados. Passo abordado na Seção 5.4;

• Passo 8 (Utilização do conhecimento): A partir dos conhecimentos obtidos nos passos anteriores, analisar se a hipótese foi validada ou não. As análises realizadas estão descritas na Seção 5.5;

4.3 Estatística

Por definição, a estatística é um conjunto de técnicas úteis para a tomada de decisão sobre um processo ou população, baseada na análise da informação contida em uma amostra desta população [8].

Algumas dessas técnicas, descritas mais detalhadamente abaixo, foram utilizadas para se testar se os dados eram normais, para então se testar a hipótese descrita na Seção 4.1.

• Testes de normalidade:

- Teste de Shapiro-Wilk: testa se uma distribuição é normal ou não normal, através da hipótese nula, considerando a normalidade dos dados. Assim, dado um nível de significância, geralmente 5%, se o teste rejeitar a hipótese, então temos indícios para rejeitar a normalidade dos dados.

- QQ Plot: testa se uma distribuição é normal ou não normal, graficamente. Se os dados, ao serem plotados, estiverem alinhados em torno de uma reta, são normais.

- QQ Plot para distribuição de Poisson: teste se uma distribuição é normal ou não, através do QQ Plot, porém utilizando-se a distribuição de Poisson. Nesse caso, se os dados plotados estiverem alinhados em torno de uma reta, então eles não são normais.

• Testes de hipótese para dados não normais:

- Mann-Whitney / Wilcoxon rank-sum: testes não paramétricos, baseados nas posições das observações e não em suas grandezas numéricas, utilizando-se das medianas [9]. Assim, testa a diferença entre duas amostras independentes.

- Kolmogorov-Smirnov: teste não não paramétrico, sobre a igualdade de distribuição de probabilidades contínuas e unidimensionais. Pode ser usado para comparar uma amostra com uma distribuição de probabilidade de referência ou duas amostras uma com a outra [10].

4.4 Dificuldades/Limitações

Encontrou-se dificuldade ao se trabalhar com dados faltantes ou pouco condizentes com a realidade. Tal problema foi contornado através do Passo 3, descrito na Seção 4.2.

A disponibilidade limitada de dados de poluentes apenas a partir de 2015 também foi algo que dificultou as análises.

5 Desenvolvimento

Os processos apresentados nas próximas seções (5.2, 5.3, 5.4 e 5.5) foram realizados no ambiente de programação Google Colab [11] através da ferramenta de programação Python [12].

5.1 Obtenção dos dados

5.1.1 Poluentes

Utilizou-se dados provenientes das estações da CETESB [1], dos bairros Taquaral, Centro e Vila União da cidade de Campinas, extraídos através do site Qualar [2].

Para as estações em questão, os poluentes coletados estão distribuidos conforme a tabela abaixo, disponibilizada pelo Qualar.

Tabela 1: Configuração das estações - CETESB

Estações	PARÂMETROS														
	CO	MP10	MP2.5	NO	NO2	NOx	O3	DV	DVG	PRESS	RADG	RADUV	TEMP	UR	VV
Campinas-Centro	X	X	--	*	*	*	--	--	--	--	--	--	X	X	--
Campinas-Taquaral	--	X	--	X	X	X	X	X	X	X	X	X	X	X	X
Campinas-V.União	--	--	X	X	X	X	X	X	X	X	X	X	X	X	X
Total de monitoramento	1	2	1	3	3	3	2	2	2	2	2	2	3	3	2

LEGENDA
(X) Parâmetro monitorado.
(*) Monitoramento desativado. Somente dados históricos.
(--) Parâmetro não monitorado.

Desse modo, escolheu-se trabalhar com os parâmetros MP10, MP2.5 e NO2, devido às pesquisas que foram feitas, assim como apontado na Seção 2.

Esses parâmetros estão sendo monitorados pelo período de tempo apresentado na tabela a seguir.

Tabela 2: Período de monitoramento dos poluentes

Poluente	Estação	Período
MP10	Centro	2000 - 2020
MP10	Taquaral	2015 - 2020
NO2	Taquaral	2015 - 2020
MP2.5	Vila União	2015 - 2020

Os dados foram extraídos para todo o tempo de monitoramento dos poluentes, até o dia 20/12/2020, no formato csv.

O arquivo extraído apresenta as seguintes informações:

- Tipo de Monitoramento
- Tipo de Rede
- Tipo de Dado
- Código da estação
- Nome da estação
- Data, hora e média horária (μg/m³) da concentração dos poluentes

5.1.2 Óbitos

Logo no início do projeto, assinou-se um termo de responsabilidade, para que o acesso aos dados de óbitos, disponibilizados pelo governo, através do Google Drive [13], fosse possível.

O arquivo foi disponibilizado no formato csv contendo as seguintes informações, no período de 2008 à 2019:

- Data do óbito
- Hora do óbito
- Idade
- Sexo
- Raça/cor
- Bairro
- Código do município
- Linha A (CID)
- Linha B (CID)
- Linha C (CID)
- Linha D (CID)
- Linha II (CID)
- Causa (CID)
- CID

5.2 Pré-processamento dos dados

5.2.1 Poluentes

O arquivo csv proveniente da extração foi transformado para o formato xlsx, e algumas das informações foram retiradas da base, restando apenas as de interesse para o projeto.

Informações retiradas:

- Tipo de Monitoramento
- Tipo de Rede

• Tipo de Dado
• Código da estação
• Nome da estação

Informações de interesse, que foram consideradas:

• Data, hora e média horária (μg/m³) da concentração dos poluentes

Em seguida, os arquivos xlsx foram lidos através da linguagem Python, no ambiente de programação Google Colab, e os devidos tratamentos foram realizados, como a limpeza de dados que não condiziam com a realidade, e a limpeza de dados nulos.

5.2.2 Óbitos

Os dados já haviam sidos pré-processados por outros integrantes do grupo Clima&Saúde anteriormente, em que se foi filtrado os óbitos por doença respiratória através do CID J e apenas para a cidade de Campinas, através do código do município. Por este motivo, foi necessário apenas retirar as informações que não seriam utilizadas, deixando-se apenas as seguintes:

• Data do óbito
• Idade
• Sexo
• CID

5.3 Transformação dos dados

5.3.1 Poluentes

Com os dados "limpos"e organizados, transformou-se o modo em que estavam dispostos. As médias horárias de concentração dos poluentes foram transformadas para máximas, e mínimas, diárias.

Em seguida, definiu-se o grupo de dados que seriam responsáveis pela normal de poluição, que mais para frente seria utilizada como comparação pra se encontrar as ondas de poluição. Para os poluentes MP10 (Taquaral), NO2(Taquaral) e MP2.5 (Vila União), escolheu-se trabalhar com todo o volume de dados, pois para eles não havia um volume de dados suficiente para se limitar. Para o poluente MP10(Centro), a normal de poluição foi gerada através do período 2002 - 2018.

Foi-se gerado então o grupo de dados que seria utilizado para realizar a análise. Para todos os poluentes, todo o volume de dados foi incluído.

Por fim, criou-se uma nova base de dados, que constavam as datas, e um indicador de onda, ou não, de poluição. A onda de poluição foi considerada como 3 dias críticos, em

sequência, de máximas de poluição. Foi-se cogitado 3 maneiras de se determinar a criticidade da poluição, como descrito abaixo:

• Percentil 90 de dia para dia: foi-se considerada como onda de poluição 3 dias em sequência, da base de dados, em que a máxima e a mínima concentração do poluente, estivesse acima de 90% desses mesmos determinados dias para a base normal de poluição.
• Percentil 90 para toda a base de normal de poluição: foi-se considerada como onda de poluição 3 dias em sequência, da base de dados, em que a máxima e a mínima concentração do poluente, estivesse acima de 90% de toda a base normal de poluição.
• Comparação com os níveis de poluição: foi-se considerada como onda de poluição 3 dias em sequência em que a máxima concentração do poluente, da base de dados, estivesse acima do nível considerado "ruim"para a saúde.

Por fim, escolheu-se a última maneira, visto que, como apresentado na Seção 2, algumas regiões podem apresentar um nível de poluição diferente das outras, devido também ao clima e outro fatores. Com isso, a região do Taquaral apresentou baixas concentrações de poluentes, a grande maioria dentro dos níveis saudáveis. Assim sendo, ao se comparar esses valores com os níveis críticos de poluição, não se encontrou ondas de poluição para a região do Taquaral, ou ondas inexpressivas.

Ao final do projeto, também se criou Datasets incluindo a informação de "lag day"para os poluentes. O lag day foi definido como o dia após a onda de poluição, baseando-se nos dias de atraso. Foi-se considerado de 1 até 10 dias de atrasado na geração do conjunto de dados.

5.3.2 Óbitos

Com o decorrer do projeto, foram-se feitas diversas análises. Para isso, os dados foram tratados também de diferentes maneiras. Foi-se aplicado os seguintes filtros na base de dados:

• CID = J

• Gênero:
- Masculino
- Feminino

• Faixa etária:
- Adolescente (13-19 anos)
- Jovem (20-39 anos)
- Adulto (40-64 anos)
- Idoso (mais de 64 anos)

Após a aplicação dos filtros, os números de óbitos foram somados por data.

5.3.3 Óbitos e poluentes

As duas bases de dados foram integradas, e então transformadas, contendo todas as informações necessárias para se aplicar os métodos estatísticos. O modelo final apresentava os seguintes parâmetros:

- Data
- Quantidade de óbitos para o determinado filtro aplicado para o dia em questão
- Indicador de onda de poluição para o dia em questão

Tinha-se em mão 2 base de dados integradas, uma para o poluente MP2.5, do período de 2015-2019 (devido à limitação dos dados do poluente) e uma para o o poluente MP10, do período de 2008-2019 (devido à limitação dos dados de óbitos).

5.4 Mineração dos dados

Com os dados transformados para o modelo de interesse, as técnicas de estatística (apresentadas na Seção 4.3) foram aplicadas para a mineração dos dados.

Inicialmente, implementou-se o teste de normalidade de Shapiro-Wilk, e foram plotados o QQ-plot, e o QQ-plot para a distribuição de Poisson.

Em seguida, aplicou-se os testes de hipótese, de Wann-Whitney/Wilcoxon rank-sum e de Kolmogorov-Smirnov.

Esses passos foram realizados tanto para o poluente MP10, quanto para o poluente MP2.5, através de funções disponibilizadas pela biblioteca, do Python, SciPy [14]. As devidas análises foram feitas em paralelo, que serão abordadas na próxima seção.

5.5 Análises

Iniciando-se as análises, plotou-se um gráfico de barras do número de óbitos pelo código CID dos óbitos, do período de 2008 à 2019, obtendo-se o resultado abaixo.

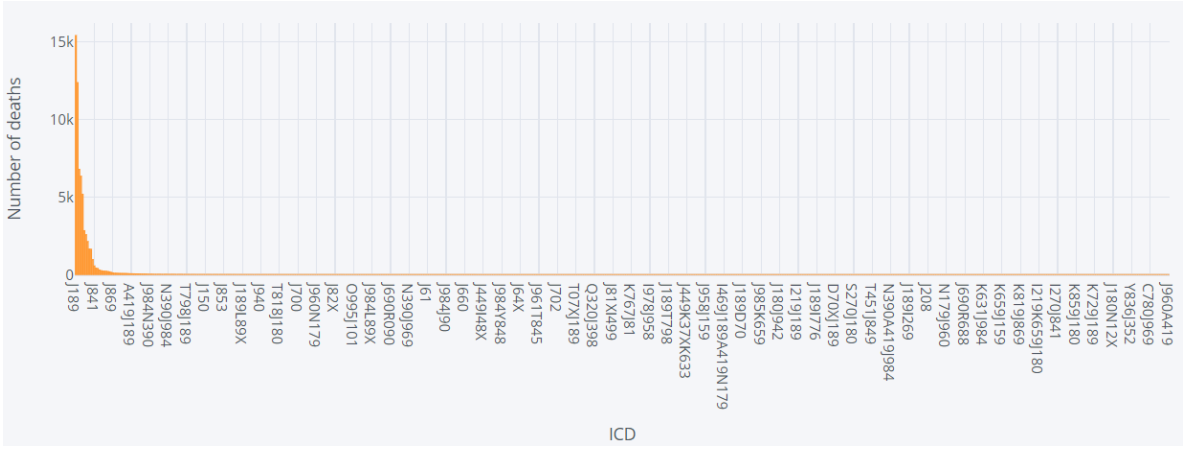

Figura 2: Número de óbitos por CID (2008-2019)

CIDs de óbitos mais frequentes:

1. J189 - 24.1% - (Pneumonia não especificada)
2. J180 - 19.4% - (Broncopneumonia não especificada)
3. J969 - 10.6% - (Insuficiência respiratória não especificada)
4. J81X - 9.9% - (Edema pulmonar não especificado)
5. J960 - 8.1% - (Insuficiência respiratória aguda)

Em seguida, iniciou-se as análises divergentes para os diferentes poluentes. A seguir, serão apresentadas as análises realizadas para o poluente MP2.5, da Vila União. Essas mesmas análises foram realizadas para o poluente MP10, porém não se obteve um resultado expressivo, como também se era esperado, como explicado na Seção 2. Os poluentes da estação do Taquaral não foram levados em conta, pois não foram encontradas ondas de poluição expressivas para o local, assim como apontado na Seção 5.3.1.

Para se estudar a normalidade, ou não, dos dados em questão, aplicou primeiramente o teste de Shapiro-Wilk, considerando-se 1% para o nível de significância. Obteve-se um resultado menor que 0,01. Logo, a hipótese nula foi rejeitada, de modo que se havia indícios para se rejeitar a normalidade dos dados. Neste mesmo passo, plotou-se o histograma dos dados, como mostrado na figura abaixo.

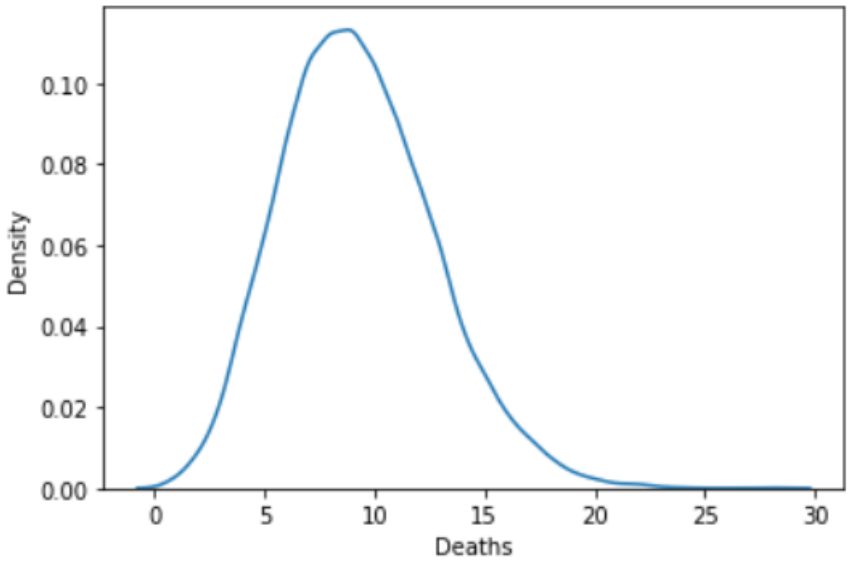

Figura 3: Histograma para os dados de óbitos

Para se comprovar a não normalidade dos dados, plotou-se o QQ-plot e o QQ-plot para a distribuição de Poisson, apresentados nas figuras 4 e 5, respectivamente.

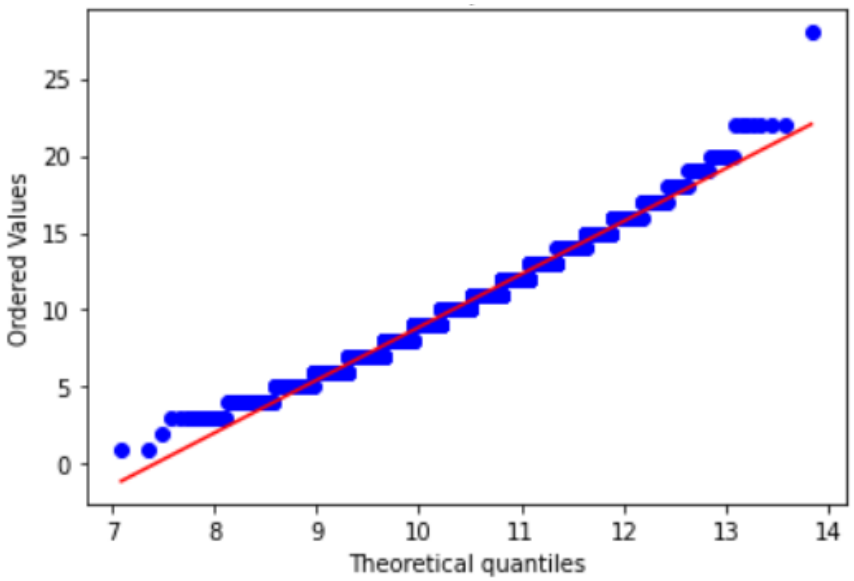

Figura 4: QQ-plot para os dados de óbitos

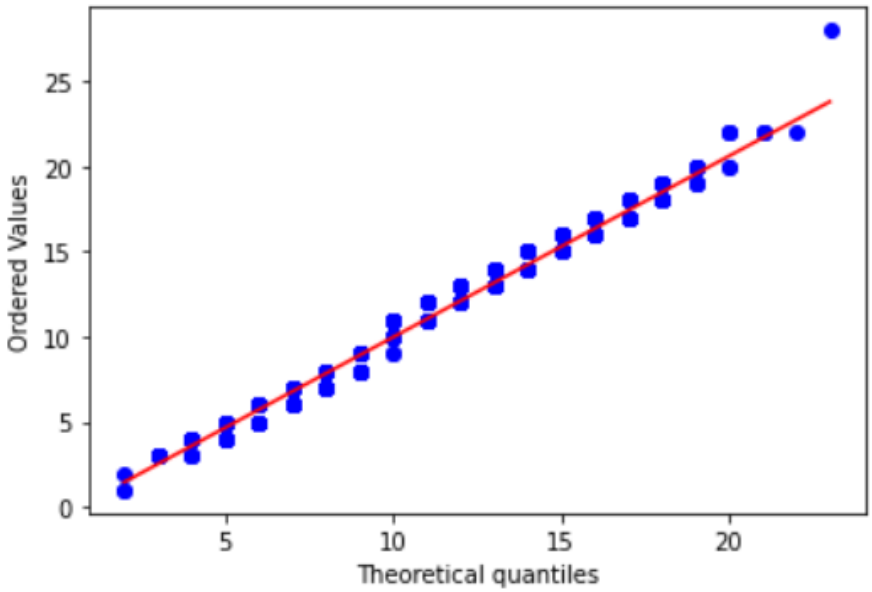

Figura 5: QQ-plot para a distribuição de Poisson para os dados de óbitos

Desse modo, validou-se que os dados não eram normais, pois, para o QQ-plot, os dados não estavam alinhados, e para a distribuição de Poisson, os dados estavam dispostos sobre a reta, assim como determinado na Seção 4.3.

Plotou-se então, dois gráficos que revelam a distribuição dos óbitos e das ondas de poluição com o decorrer dos anos. O primeiro gráfico (Figura 6) apresenta, lado a lado, o número de óbitos de cada ano, e o número de dias pertencentes à ondas de poluição. O segundo gráfico (Figura 7) apresenta o número de óbitos, por ano, que aconteceram em dias de onda de poluição.

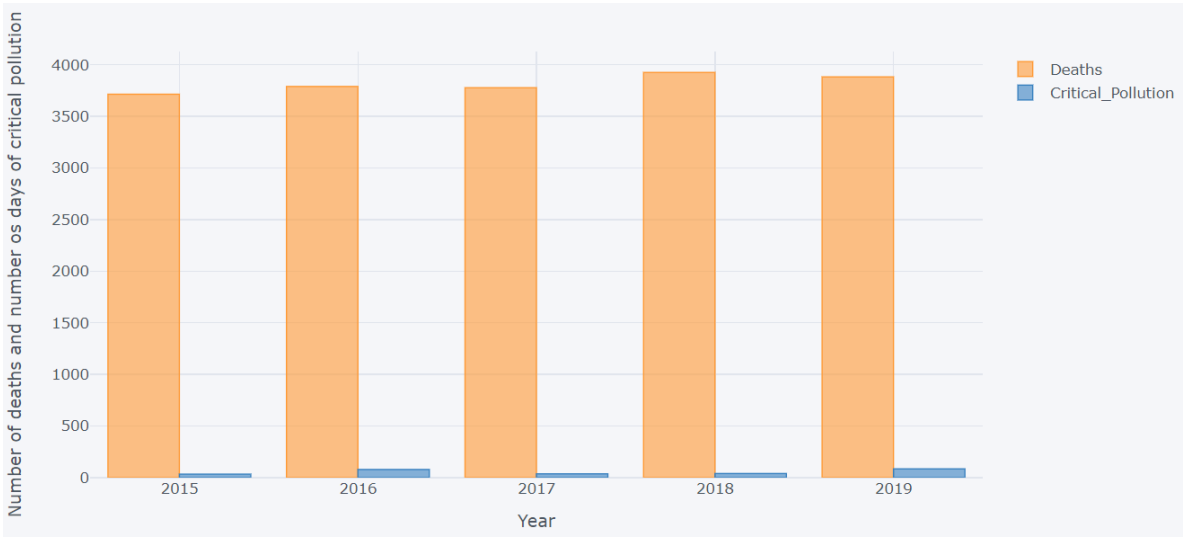

Figura 6: Número de óbitos/Número de dias de onda de poluição - MP 2.5 (2015-2019)

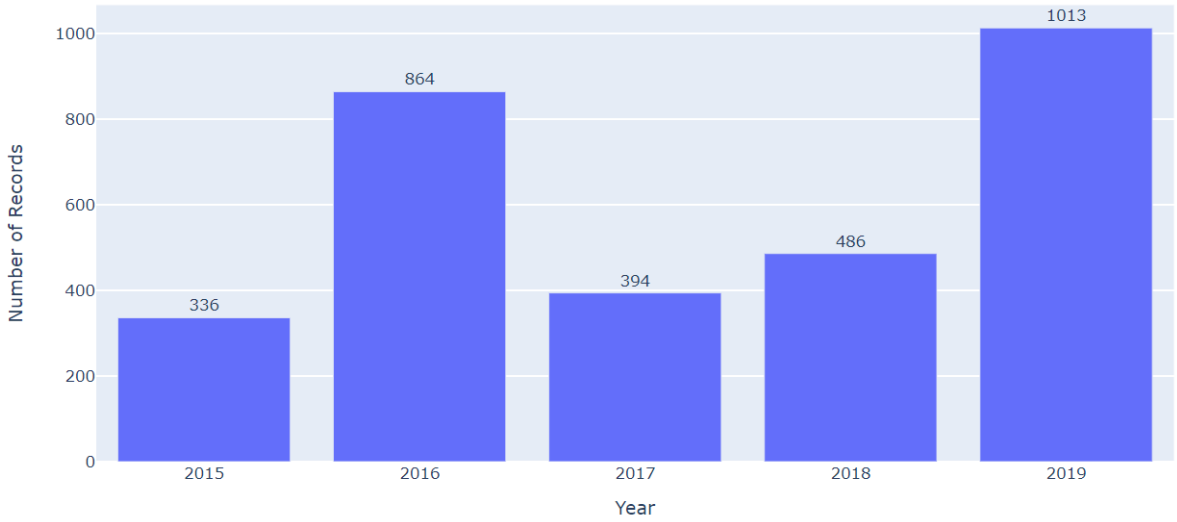

Figura 7: Número de óbitos em dias de onda de poluição - MP2.5 (2015-2019)

Após a validação de normalidade, os testes de hipótese para dados não normais poderiam ser aplicados, para se validar a questão apresentada na Seção 4.1.

Os 3 testes foram então aplicados, simultaneamente, para cada um dos filtros citados anteriormente. As funções utilizadas retornaram o Valor-p dos testes, que é a probabilidade de se obter resultados, pelo menos, tão extremos quanto aos resultados obtidos pelo teste de hipótese [15].

Primeiramente, aplicou-se os testes para os CIDs iniciados por J (problemas respiratórios), obtendo-se os parâmetros abaixo.

Tabela 3: Parâmetros (Valor-p) obtidos pelos testes de hipótese (CID = J)

Filtro(s)	Mann-Whitney	Wilcoxon rank-sum	Kolmogorov-Smirnov
CID = J	0.000046	0.000049	0.000325
CID = J Gênero: feminino	0.005694	0.006162	0.018070
CID = J Gênero: masculino	0.008789	0.009344	0.033258
CID = J Faixa etária: adolescente	0.532227	0.850156	1.000000
CID = J Faixa etária: jovem	0.282752	0.482113	0.997538
CID = J Faixa etária: adulto	0.000748	0.001034	0.067360
CID = J Faixa etária: idoso	0.000395	0.000425	0.001270
CID = J Faixa etária: idoso Gênero: feminino	0.066859	0.070517	0.744462
CID = J Faixa etária: idoso Gênero: masculino	0.002103	0.00242	0.017637

Por fim, foram aplicados os testes para os dias de atraso, como mostra a tabela abaixo.

Tabela 4: Parâmetros (Valor-p) obtidos pelos testes de hipótese para os dias de atraso

Dias de atraso (CID = J)	Mann-Whitney	Wilcoxon rank-sum	Kolmogorov-Smirnov
1	0.000925	0.000968	0.008597
2	0.000183	0.000194	0.000332
3	0.007787	0.008025	0.071865
4	0.002629	0.00273	0.011065
5	0.002752	0.002856	0.029718
6	0.006333	0.006536	0.149503
7	0.065928	0.066963	0.102315
8	0.362537	0.364344	0.894668
9	0.546832	0.548323	0.516796
10	0.211434	0.213166	0.426218

6 Resultados

O gráfico da Figura 2 revela que os CIDs mais frequentes são os J189 e o J180, Pneumonia não especificada e Broncopneumonia não especificada, respectivamente. Sendo eles responsáveis por 43,5% do número de óbitos por doenças respiratórias no período de 2008 à 2019.

Com os testes de normalidade aplicados, deduziu-se que os dados eram não normais, como explicado na Seção 5.5.

Ao se analisar o gráfico da Figura 6, percebe-se que houve um aumento do número de óbitos por doenças respiratórias no período de 2017-2019, se comparado ao período de 2015 - 2017. O mesmo ocorre para os dias pertencentes à ondas de poluição. Numericamente, o

primeiro período é responsável por 45,17% dos óbitos, contra 54.83% do segundo período.

Quanto aos testes de hipótese realizados, sabe-se que quanto menor for o Valor-p, maior é a evidência para se rejeitar a hipótese nula [15]. Definiu-se o valor de significância como 5%, então, para valores de Valor-p menores que 0,05, os grupos analisados não pertencem à mesma população. A Tabela 5 aponta todos os casos em que se rejeitou a hipótese nula por algum dos 3 testes realizados, juntamente com o valor da média do número de óbitos para dias de onda de poluição, e a média para dias não pertencentes à ondas de poluição. A tabela 6 aponta todos os dias após as ondas de poluição, que também se rejeitou a hipótese nula por algum dos 3 testes realizados, apresentando também o número de dias de atraso considerado.

Tabela 5: Média de óbitos por dia para hipótese nula rejeitada

Filtro	**Média de óbitos por dia**	
	Dias comuns	**Dias de onda de poluição**
CID = J	10.321935	11.247273
CID = J Gênero: feminino	4.995434	5.396364
CID = J Gênero: masculino	5.341352	5.794872
CID = J Faixa etária: adulto	2.615438	2.924303
CID = J Faixa etária: idoso	7.300000	7.938182
CID = J Faixa etária: idoso Gênero: masculino	3.624582	4.030418

Tabela 6: Média de óbitos por dia para hipótese nula rejeitada para dias atrasados de ondas de poluição

Dias de atraso (CID = J)	Média de óbitos por dia	
	Dias comuns	Dias após onda de poluição
1	10.374244	11.304094
2	10.368803	11.356725
3	10.397219	11.081871
4	10.386941	11.181287
5	10.381499	11.233918
6	10.383313	11.216374

7 Conclusão

Lidar com a base de dados foi a maior dificuldade encontrada no trabalho. Tratar os dados faltantes e trabalhar apenas com o período limitado de dados disponível foi algo que exigiu mais reflexão sobre como lidar com isso, e como fazer a melhor abordagem.

Ao se analisar os resultados, conclui-se que o poluente MP2.5, monitorado pela estação da Vila União, possui influência sobre a saúde de quem possui problemas respiratórios na cidade de Campinas, principalmente sobre quem possui idade mais avançada e é do gênero masculino (como mostra a Figura 5), o que acaba impactando no número de óbitos por doenças respiratórias. Conclui-se também que o efeito não ocorre apenas nos dias exatos de ondas de poluição, mas também nos dias posteriores ao ocorrido. A Figura 6 revela que, em até seis dias de atraso, percebe-se efeitos no número de óbitos.

Notou-se que os monitoramentos realizados pela estação localizada no Taquaral não apresentou indícios de poluição crítica, e que o MP10 também não apresentou influência no número de óbitos.

Os estudos realizados por esse projeto podem ser mais aprofundados por estudos futuros, para por exemplo se entender mais sobre as consequências do MP10, ou se estudar também

os impactos no número de internações por problemas respiratórios, visto a relevância de se utilizar de ferramentas tecnológicas para gerar informação e conhecimento para outras áreas. Tais informações podem ser repassadas para as autoridades públicas, de modo que se possa auxiliar na definição de políticas públicas para se trabalhar em cima do problema existente.

Referências

[1] CETESB, "Companhia ambiental do estados de são paulo." https://cetesb.sp.gov.br/. Acesso em: 23/12/2020.

[2] Qualar, "Sistema de informações da qualidade do ar. 3.83." https://qualar.cetesb.sp.gov.br/qualar/home.do. Acesso em: 22/12/2020.

[3] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI magazine*, pp. 37–37, 1996.

[4] K. Abe, G. Santos, M. Coêlho, and S. Miraglia, "Pm10 exposure and cardiorespiratory mortality – estimating the effects and economic losses in são paulo, brazil," *Aerosol and Air Quality Research*, pp. 3217–3133, 2018.

[5] D. Vilas Boas, M. Matsuda, O. Toffoletto, M. Garcia, P. Saldiva, and V. Marquezini, "Workers of são paulo city, brazil, exposed to air pollution: Assessment of genotoxicity," *Mutat Res Gen Tox En*, pp. 18–24, 2018.

[6] T. Santos, V. Carvalho, and M. Reboita, "Avaliação da influência das condições meteorológicas em dias com altas concentrações de material particulado na região metropolitana do rio de janeiro," *Eng Sanit Ambient*, pp. 307–313, 2016.

[7] L. Coelho, "Ciência de dados: O que é, conceito e definição," *Blog Cetax*, 2020.

[8] A. Santos, "Estatística: definição e conceitos básicos.".

[9] B. Oliveira, "Teste t e mann-whitney para amostras independentes," *Oper Data*, 2020.

[10] W. Daniel, *Applied Nonparametric Statistics*. Duxbury, 2000.

[11] Google, "Google colab." https://colab.research.google.com/. Acesso em: 23/12/2020.

[12] Python Software Foundation, "Python." https://www.python.org/. Acesso em: 23/12/2020.

[13] Google, "Google drive." https://www.google.com.br/drive/apps.html. Acesso em: 23/12/2020.

[14] SciPy, "Statistical functions." https://docs.scipy.org/doc/scipy/reference/stats.html. Acesso em: 24/12/2020.

[15] B. Beers, "P-value," *Investopedia*, 2020.