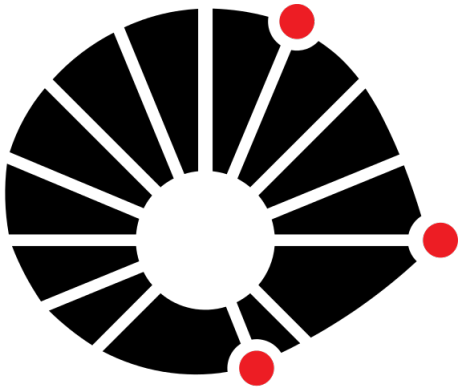
Universidade Estadual de Campinas

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

Aluno: João Victor Gitti Arêdes, RA: 170715

Orientadora: Profª Dra. Paula Dornhofer Paro Costa

CAMPINAS, JANEIRO DE 2021

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 grupo multidisciplinar de pesquisas em Clima&Saúde da Unicamp.

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 resultados *software* deste estudo podem ser encontrados no repositório do GitHub:

https://github.com/jvgitti/Clima-e-Saude-DS

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 [2].

Um dos componentes que caracterizam a poluição do ar é o material particulado em suspensão (ou PM, do inglês *Particulate Matter*), 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 - 2,5 μm. 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 ao 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 [3].

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 [4].

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 integrada 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 adota uma metodologia de Ciência de Dados, seguindo o processo ilustrado na Figura 1, tipicamente referenciado como KDD, do inglês *Knowledge Discovery in Databases* [5].

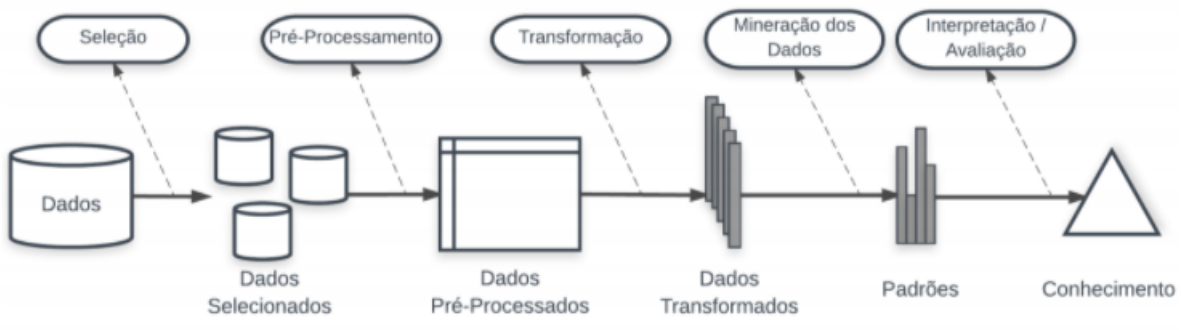


Figura 1: Processo típico de análise de dados [1]

.

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

- **Passo 1 - Entendimento do Problema:** Nessa etapa o objetivo é entender a motivação do estudo, 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 de estudo:** Seleção do conjunto de dados de interesse, com as variáveis a serem estudadas. Essa etapa é abordada com mais profundidade na Seção 4.1;

- **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 4.2;

- **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 4.3;

- **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, de forma a validá-las quantitativamente através de análises gráficas e com estatísticas descritivas. O conteúdo está na Seção 4.4.

- **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 também na Seção 4.4;

- **Passo 8 - Consolidação dos Resultados e Extração do Conhecimento:** A partir dos conhecimentos obtidos nos passos anteriores, analisar se os pontos levantados no primeiro passo foram validados ou não. As análises realizadas estão descritas na Seção 5.

Os processos apresentados nas próximas seções (4.2, 4.3 e 4.4) foram realizados no ambiente de programação Google Colab [6] através da ferramenta de programação Python [7].

4.1 Bases de Estudo

4.1.1 Poluentes

Utilizou-se dados provenientes das estações da CETESB [8], dos bairros Taquaral, Centro e Vila União da cidade de Campinas, disponibilizados pelo sistema Qualar [9].

Para as estações em questão, os parâmetros de qualidade do ar coletados são apresentados na Tabela 1, extraída do sistema 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.

A partir dos estudos realizados, escolheu-se trabalhar com os parâmetros PM10, PM2,5 e NO2. Os períodos de monitoramento de cada parâmetro são apresentados na Tabela 2, conforme disponibilidade dos dados.

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 período de monitoramento dos poluentes, até o dia 20/12/2020, no formato CSV (do inglês, *Comma Separated Values*).

Do arquivo extraído, utilizou-se as seguintes informações:

- Data da informação coletada;
- Hora da informação coletada;
- Média horária (µg/m³) da concentração dos poluentes.

4.1.2 Óbitos

Os dados de óbitos foram cedidos pela Secretaria de Saúde de Campinas, no contexto do Projeto Clima & Saúde, e conta com projeto aprovado no Comitê de Ética em Pesquisa sob CAAE:95503318.4.0000.5404. O acesso aos dados só foi permitido após assinatura de termo de confidencialidade e anuência aos termos de proteção aos dados.

O arquivo foi disponibilizado no formato CSV. Dentre as informações, selecionou-se as seguintes como relevantes para o estudo:

- Data do óbito;
- Idade;
- Código do município;
- Sexo;
- CID (Classificação Estatística Internacional de Doenças) de óbito.

4.2 Pré-processamento dos dados

4.2.1 Poluentes

Os arquivos extraídos 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.

4.2.2 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.

4.3 Transformação dos dados

4.3.1 Poluentes

Com os dados pré-processados, 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 PM10 (Taquaral), NO2 (Taquaral) e PM2,5 (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 PM10 (Centro), a normal de poluição foi gerada através do período 2002 - 2018.

Gerou-se 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 três dias críticos, em sequência, de máximas de poluição. Cogitou-se três maneiras de se determinar a criticidade da poluição, como descrito abaixo:

- **Percentil 90 de dia para dia:** considerou-se como onda de poluição três 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:** Considerou-se como onda de poluição três 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:** Considerou-se como onda de poluição três dias em sequência em que a máxima concentração do poluente, da base de dados, estivesse acima do nível considerado prejudicial para a saúde.

Então, 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 foram criados 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. Considerou-se de 1 até 10 dias de atraso na geração do conjunto de dados.

4.3.2 Óbitos

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

- CID = J (Apenas óbitos causados por doenças respiratórias);
- 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.

4.3.3 Óbitos e poluentes

As duas bases de dados foram integradas, e então transformadas, contendo todas as informações necessárias para se aplicar os métodos estatísticos, apresentados na Seção 4.4. 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.

Haviam sido criadas então, duas bases de dados integradas, uma para o poluente PM2,5, do período de 2015-2019 (devido à limitação dos dados do poluente) e uma para o o poluente PM10, do período de 2008-2019 (devido à limitação dos dados de óbitos).

4.4 Análise Exploratória e Mineração dos dados

Com os dados transformados para o modelo de interesse, aplicou-se técnicas de estatística para realizar a mineração dos dados, definidas pela análise exploratória.

Inicialmente, verificou-se a normalidade dos dados, através do teste de Shapiro-Wilk [10], do QQ-plot [11] para a distribuição dos dados e do QQ-plot para a distribuição de Poisson [12].

Em seguida, aplicou-se os testes de hipótese de Wann-Whitney/Wilcoxon rank-sum [13] e de Kolmogorov-Smirnov [14]. A hipótese nula considerada foi a de que a poluição não possui influência sobre o número de óbitos por doenças respiratórias, e a hipótese alternativa foi a de que possui sim influência.

Esses passos foram realizados tanto para o poluente PM10, quanto para o poluente PM2,5, através de funções disponibilizadas pela biblioteca, do Python, SciPy [15]. As devidas análises foram feitas em paralelo, que serão abordadas na Seção 4.5.

4.5 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.

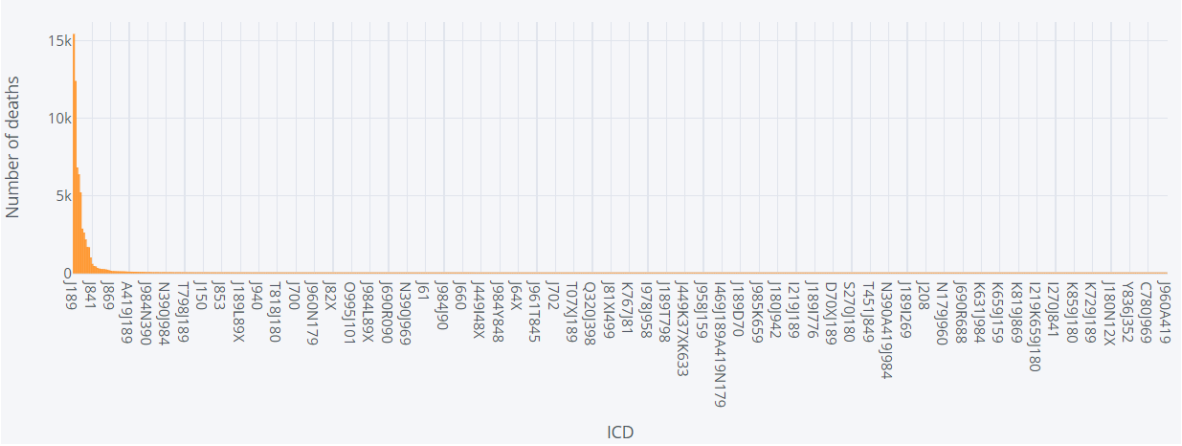


Figura 2: Número de óbitos por CID (2008-2019)

CIDs de óbitos mais frequentes:

- J189 - 24.1% - (Pneumonia não especificada)
- J180 - 19.4% - (Broncopneumonia não especificada)
- J969 - 10.6% - (Insuficiência respiratória não especificada)
- J81X - 9.9% - (Edema pulmonar não especificado)
- 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 PM2,5, da Vila União. Essas mesmas análises foram realizadas para o poluente PM10, 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 4.3.1.

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.

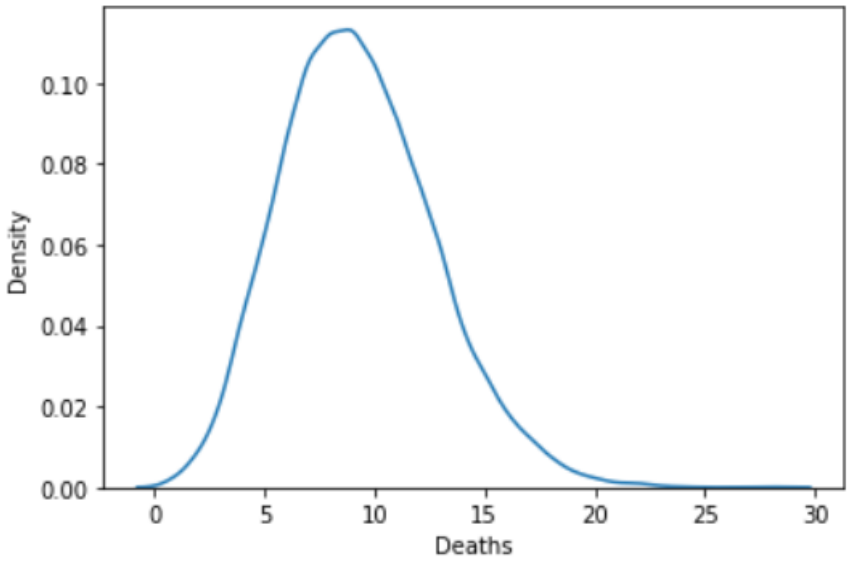


Figura 3: Histograma para os dados de óbitos

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.

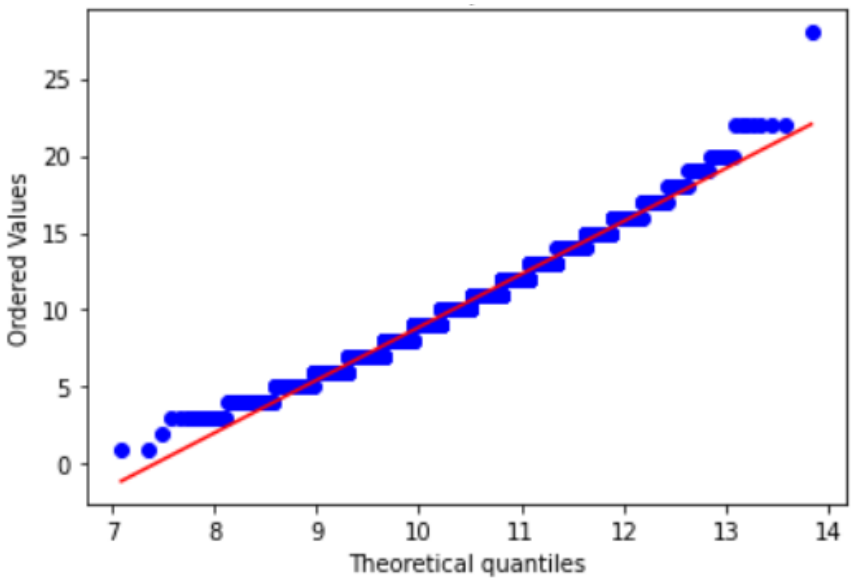


Figura 4: QQ-plot para os dados de óbitos

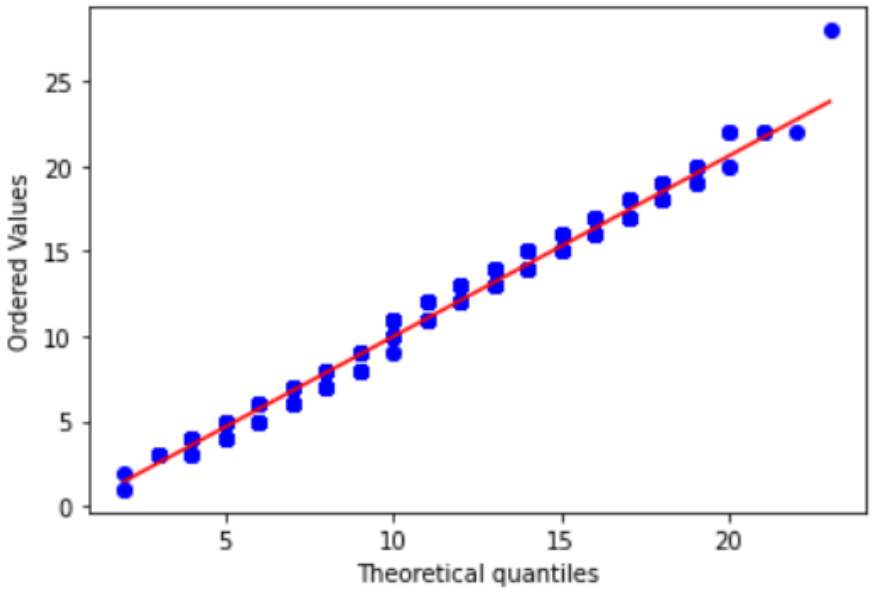


Figura 5: QQ-plot para a distribuição de Poisson para os dados de óbitos

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.

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.

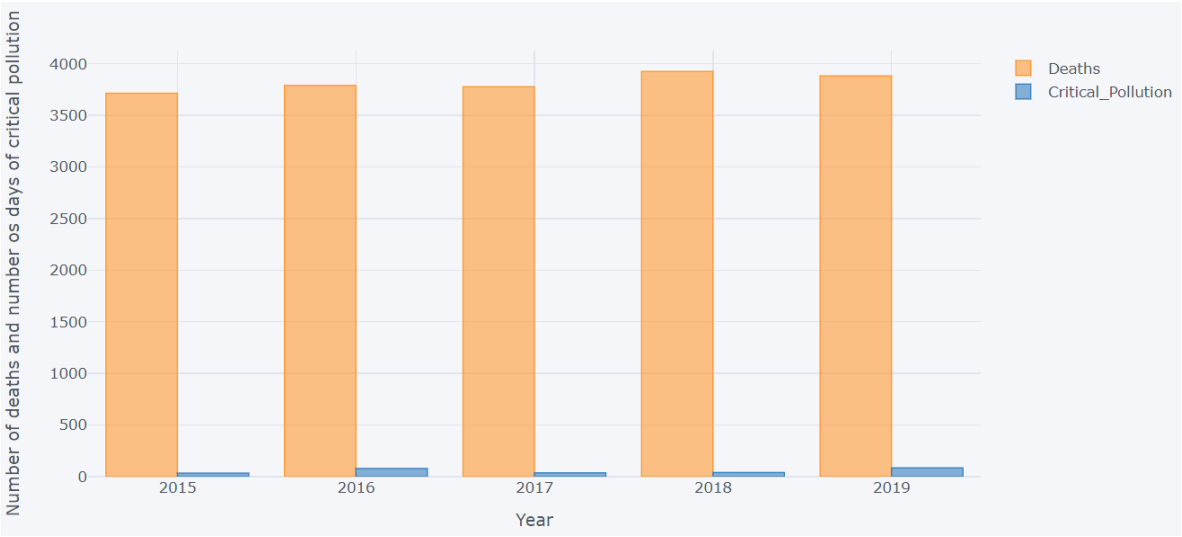


Figura 6: Número de óbitos/Número de dias de onda de poluição - PM2,5 (2015-2019)

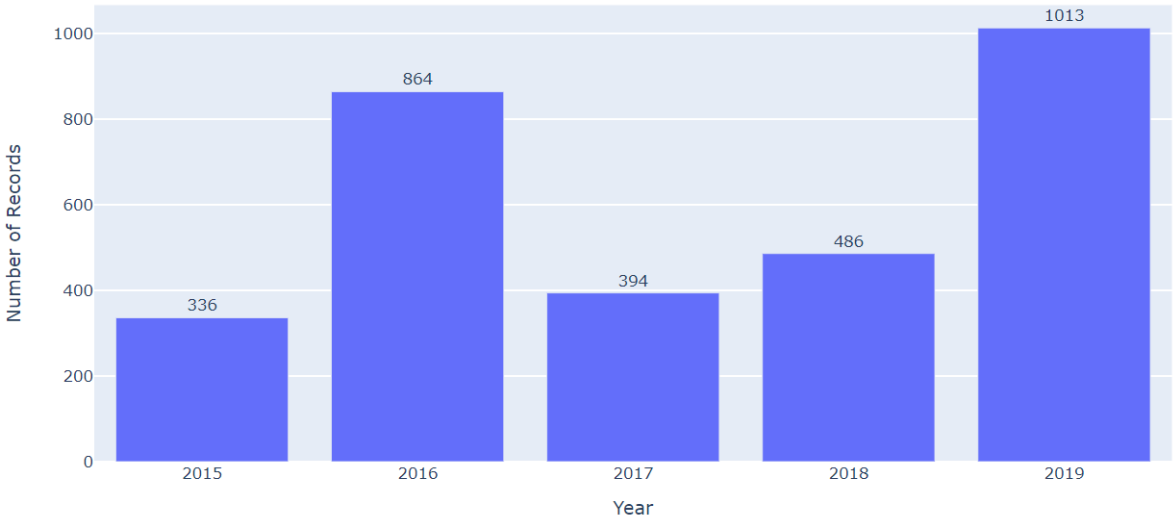


Figura 7: Número de óbitos em dias de onda de poluição - PM2,5 (2015-2019)

Após a validação de normalidade, os testes de hipótese para dados não normais poderiam ser aplicados, com o objetivo de se rejeitar a hipótese levantada.

Os 3 testes foram então aplicados, simultaneamente, para cada um dos filtros citados anteriormente. As funções utilizadas retornaram o p-valor dos testes, que é a probabilidade de se obter resultados, pelo menos, tão extremos quanto aos resultados obtidos pelo teste de hipótese [16].

Primeiramente, aplicou-se os testes para os CIDs iniciados por J (problemas respiratórios), obtendo-se os parâmetros abaixo (em negrito, os parâmetros abaixo de 0,05):

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

Filtro(s)	Mann-Whitney	Wilcoxon rank-sum	Kolmogorov-Smirnov
CID = J	**0,00**	**0,00**	**0,00**
CID = J Gênero: feminino	**0,01**	**0,01**	**0,02**
CID = J Gênero: masculino	**0,01**	**0,01**	**0,03**
CID = J Faixa etária: adolescente	0,53	0,85	1,00
CID = J Faixa etária: jovem	**0,28**	**0,48**	1,00
CID = J Faixa etária: adulto	**0,00**	**0,00**	0,07
CID = J Faixa etária: idoso	**0,00**	**0,00**	**0,00**
CID = J Faixa etária: idoso Gênero: feminino	0,07	0,07	0,74
CID = J Faixa etária: idoso Gênero: masculino	**0,00**	**0,00**	**0,02**

Por fim, foram aplicados os testes para os dias de atraso, como mostrado a seguir:

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

Dias de atraso (CID = J)	Mann-Whitney	Wilcoxon rank-sum	Kolmogorov-Smirnov
1	**0,00**	**0,00**	**0,01**
2	**0,00**	**0,00**	**0,00**
3	**0,01**	**0,01**	0,07
4	**0,00**	**0,00**	**0,01**
5	**0,00**	**0,00**	**0,03**
6	**0,01**	**0,01**	0,15
7	0,07	0,07	0,10
8	0,36	0,36	0,89
9	0,55	0,55	0,52
10	0,21	0,21	0,43

5 Resultados

O gráfico da Figura 2 revela que os CIDs mais frequentes são o J189 (Pneumonia não-especificada) e o J180 (Broncopneumonia não especificada), 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 não apresentavam distribuições normais, como explicado na Seção 4.5.

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 p-valor, maior é a evidência para se rejeitar a hipótese nula [16]. Definiu-se o valor de significância como 5%, então, para valores de p-valor 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 três 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	10,32	11,25
CID = J Gênero: feminino	5,00	5,40
CID = J Gênero: masculino	5,34	5,79
CID = J Faixa etária: adulto	2,62	2,92
CID = J Faixa etária: idoso	7,30	7,94
CID = J Faixa etária: idoso Gênero: masculino	3,62	4,03

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	10,37	11,30
2	10,37	11,36
3	10,40	11,08
4	10,39	11,18
5	10,38	11,23
6	10,38	11,22

6 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 PM2,5, 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 Tabela 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 Tabela 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 PM10 também não apresentou influência no nú-

mero 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 PM10, 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 relacionadas aos problemas existentes.

Referências

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

[2] 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.

[3] 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.

[4] 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.

[5] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, *et al.*, "Knowledge discovery and data mining: Towards a unifying framework.," in *KDD*, vol. 96, pp. 82–88, 1996.

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

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

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

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

[10] "Teste de shapiro-wilk," *Portal Action*, 2020.

[11] M. Wilk and R. Gnanadesikan, "Probability plotting methods for the analysis of data," *Biometrika Trust*, pp. 1–17, 1968.

[12] "DistribuiÇÃo de poisson," *Portal Action*, 2020.

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

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

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

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