UM ESTUDO SOBRE A RELAÇÃO ENTRE POLUIÇÃO ATMOSFÉRICA E INTERNAÇÕES HOSPITALARES VIA REGRAS DE ASSOCIAÇÃO

Palavras-Chave: Poluição atmosférica, internações hospitalares, regras de associação

Autores/as:
MATEUS MOLINARI FERRAZ - Faculdade de Tecnologia (FT), Universidade Estadual de Campinas (UNICAMP)
Prof. Dr. GUILHERME PALERMO COELHO (orientador) - Faculdade de Tecnologia (FT), Universidade Estadual de Campinas (UNICAMP)
Profa. Dra. ANA ESTELA ANTUNES DA SILVA (coorientadora) - Faculdade de Tecnologia (FT), Universidade Estadual de Campinas (UNICAMP)

1 INTRODUÇÃO

A preocupação da população mundial com a poluição do ar tem aumentado nos últimos anos, visto que a poluição atmosférica é responsável por cerca de 4,2 milhões de mortes por ano devido a acidentes vasculares cerebrais, doenças cardíacas, câncer de pulmão, doenças respiratórias agudas e crônicas. Além disso, cerca de 91% da população mundial vive em lugares onde os níveis de qualidade do ar excedem os limites definidos pela OMS (Organização Mundial da Saúde) (WHO, 2021a). Assim, estudos relacionados à poluição atmosférica têm se tornado cada vez mais importantes, uma vez que permitem um melhor entendimento da poluição do ar e seus impactos na saúde humana, possibilitando a proposição de estratégias mais eficazes para mitigar seus efeitos.

Poluição atmosférica pode ser entendida como a presença de químicos ou compostos na atmosfera que, geralmente, degradam a qualidade do ar ou causam mudanças prejudiciais na qualidade de vida (SARLA, 2020).

A partir de uma revisão bibliográfica, percebeu-se que uma possível maneira de se encontrar associações entre poluição atmosférica e enfermidades e que é pouco utilizada neste contexto é a extração de regras de associação, que são padrões que refletem quais itens estão frequentemente associados em um determinado contexto (HAN; KAMBER; PEI, 2012).

Um dos poucos estudos encontrados que utilizou esta estratégia a fim de se achar os mesmos tipos de associações foi o de Payus et al. (2013), que realizou tal mineração a partir de dados da cidade de Kuala Lumpur, Malásia, considerando todo o ano de 2008 e os poluentes O_3 (Ozônio), MP_{10} (Material Particulado com diâmetro aerodinâmico entre 2,5 e 10 µm), CO (Monóxido de Carbono), SO_2 (Dióxido de Enxofre) e NO_2 (Dióxido de Nitrogênio). Como resultado, esse estudo indicou que várias combinações entre os atributos considerados possuem forte influência em doenças respiratórias, como a combinação MP_{10}, CO e temperatura.

Diante disso, o estudo atual se propôs a avaliar as relações existentes entre a concentração de poluentes atmosféricos e o número de internações hospitalares por doenças respiratórias registradas na cidade de São Paulo, utilizado o algoritmo Apriori, uma estratégia de extração de regras de associação (HAN; KAMBER; PEI, 2012).

Para isso, os dados relacionados às concentrações dos poluentes atmosféricos foram coletados a partir da plataforma online QUALAR (CETESB, 2021a). Nesta plataforma, estações automáticas que monitoram a concentração de poluentes atmosféricos, coletando dados a cada hora, e que foram instaladas em várias cidades pela Companhia Ambiental do Estado de São Paulo (CETESB), responsável pelo monitoramento ambiental no Estado de São Paulo, disponibilizam publicamente esses dados.

Já os dados relacionados às internações hospitalares realizadas no Sistema Único de Saúde (SUS) brasileiro também são disponibilizados publicamente, através do sistema TABNET (DATASUS, 2021). O TABNET é gerenciado pelo Departamento de Informática do SUS (DATASUS).

Além disso, este estudo decidiu focar nos seguintes poluentes: MP_{10}, $MP_{2,5}$ (Material Particulado com diâmetro aerodinâmico inferior a 2,5 µm), NO_x (Óxidos de Nitrogênio) e O_3. Estes poluentes foram escolhidos por serem os que mais foram relacionados a problemas de saúde na literatura revisada, por não terem muitos dados ausentes nas estações de monitoramento e por não apresentarem somente concentrações classificadas como de boa qualidade. É importante deixar claro que há um certo destaque envolvendo o

poluente MP_{10} na literatura, pois foi o mais citado nas referências encontradas (AGUIAR, 2015; BRAGA et al., 2001; PAYUS et al., 2013).

Por outro lado, as seguintes enfermidades respiratórias foram escolhidas para serem consideradas no presente estudo, devido a relações entre elas e a poluição atmosférica apontadas em diversos estudos analisados durante o levantamento bibliográfico: asma, bronquiolite, bronquite, edema pulmonar, enfisema, gripe, pneumonia e rinite (AGUIAR, 2015; BRAGA et al., 2001; SARLA, 2020).

Além disso, a cidade de São Paulo foi escolhida por apresentar muitas estações de monitoramento da qualidade do ar, o que supre uma possível situação de dados faltantes, um grande número de internações e uma boa variação nos índices de qualidade do ar, pelo fato de ser um grande município.

Por fim, tal análise foi realizada em dois períodos diferentes: antes (ano de 2019) e durante a pandemia de COVID (ano de 2020), o que permitiu realizar comparações entre as regras extraídas referentes a cada período.

2 METODOLOGIA

A fim de se realizar a extração de regras de associação proposta, foi necessário preparar os dados relacionados à poluição atmosférica e os relacionados às internações hospitalares, construir a base de transações e definir uma metodologia para os experimentos, conforme a Figura 1. Todo este processo é descrito, mais detalhadamente, nesta seção.

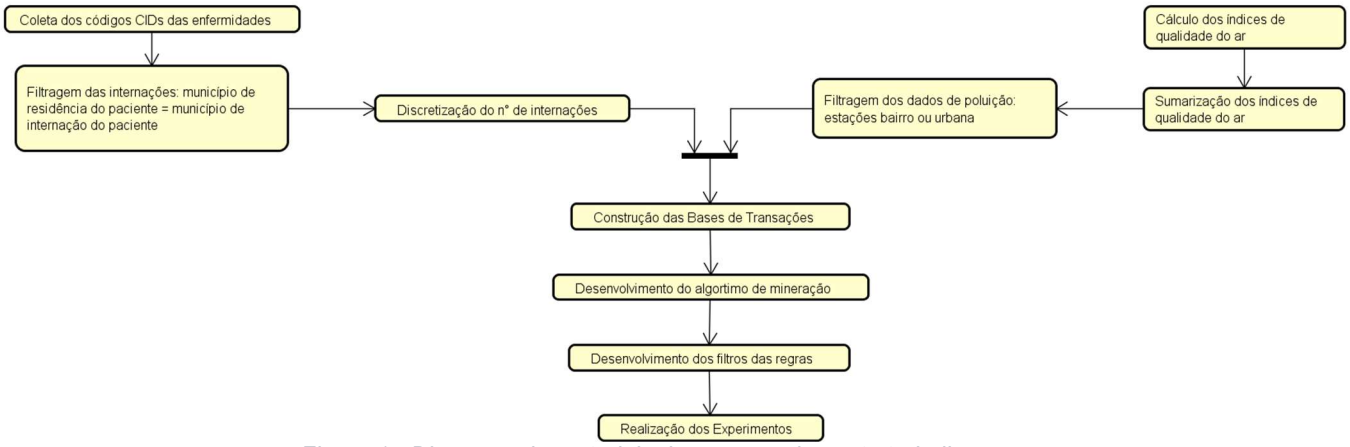

Figura 1 - Diagrama da metodologia empregada neste trabalho.

2.1 PREPARAÇÃO DE DADOS

Como os dados de internações hospitalares, contidos no TABNET, identificam o diagnóstico da internação por meio da Classificação Internacional de Doenças (CID-10) (WHO, 2021b), os seguintes códigos CID-10 foram escolhidos, correspondentes às enfermidades citadas anteriormente: de J09 a J11 para gripe, de J12 a J18 para pneumonia, J20, J40 e J42 para bronquite, J21 para bronquiolite, J30 para rinite, J43 para enfisema, J45 para asma e J81 para edema pulmonar.

Além disso, é importante destacar que foram utilizados apenas os registros em que o município de residência do paciente é igual ao município do estabelecimento onde este paciente foi internado. Isso foi feito porque, como a cidade de São Paulo está localizada em uma região metropolitana com um grande fluxo de pessoas de vários municípios, há uma grande possibilidade de que uma pessoa que more em outra cidade e que trabalhe em São Paulo adoeça devido à poluição de São Paulo e acabe sendo internada em um hospital em sua cidade de residência.

Por outro lado, com relação aos dados da poluição atmosférica, decidimos considerar apenas as combinações ‘estação de monitoramento x poluente’ que são classificadas em uma escala de bairro (relativa à representação espacial de áreas de vizinhança urbana com atividade uniforme e dimensões entre 501 e 4.000 metros) ou em uma escala urbana (relativa à representação espacial das cidades ou regiões metropolitanas, na ordem de 4 a 50 km), pois queríamos realizar a análise proposta de forma mais abrangente e não queríamos verificar a região específica onde cada paciente foi internado.

2.2 BASE DE TRANSAÇÕES

Como os dados de qualidade do ar correspondem a concentrações horárias de cada poluente e é necessário avaliar o dia como um todo, foi necessário calcular os índices da qualidade do ar para cada dia. Estes índices foram desenvolvidos para simplificar o processo de divulgação da qualidade do ar (CETESB, 2021b) e atribuem, para um determinado dia e um determinado poluente, uma das seguintes classificações: 'N1 - Boa', 'N2 - Moderada', 'N3 - Ruim', 'N4 - Muito Ruim' e 'N5 - Péssima'. Os índices de qualidade do ar para os poluentes MP_{10}, $MP_{2,5}$, NO_x e O_3 são calculados a partir da concentração média avaliada em 24, 24, 1 e 8 horas, respectivamente, conforme descrito em CETESB (2021b).

Além disso, como São Paulo monitora o mesmo poluente com diferentes estações, foi necessário resumir o índice diário de qualidade do ar da cidade. Para isso, as classificações 'N1 - Boa', 'N2 - Moderada', 'N3 - Ruim', 'N4 - Muito Ruim' e 'N5 - Péssima' foram convertidas para os respectivos valores inteiros: 1, 2, 3, 4 e 5. Assim, para cada poluente e para cada dia, uma média simples dos valores reportados pelas estações foi calculada e então esse valor foi arredondado. Por fim foi feito o processo de conversão reversa, onde o valor inteiro calculado foi convertido para uma das classificações, mencionadas anteriormente.

Em relação às internações, como o algoritmo Apriori não trabalha com valores contínuos, foi necessário discretizar o número diário de internações na cidade de São Paulo (HAN; KAMBER; PEI, 2012). Assim, com base no estudo de Payus et al. (2013), o método de discretização escolhido foi o *Equal Frequency Binning* (HAN; KAMBER; PEI, 2012). O *Equal Frequency Binning* é um algoritmo que divide os dados em alguns grupos em que cada grupo contém aproximadamente o mesmo número de valores (SAYAD, 2021).

Com isso, o número de internações diárias foi classificado nos seguintes grupos: 'BAIXA' para dias com até 72 internações, 'MEDIA-BAIXA' para dias com 73 a 89 internações, 'MEDIA' para dias com 90 a 105 internações, 'MEDIA-ALTA' para dias com 106 a 127 admissões e 'ALTA' para dias com mais de 127 internações.

Após todo esse processo, foi possível desenvolver a base de transações, que, neste caso, é uma base de dados onde cada registro, ou transação, corresponde a um dia, e cada coluna, exceto a coluna dia, equivale a um atributo a ser usado ao extrair regras de associação.

Como o número de internações em um determinado dia pode estar relacionado às concentrações de poluentes de um dia anterior, decidimos realizar experimentos considerando o aspecto 'temporal' dos poluentes. Para tanto, para cada experimento que considera um diferente aspecto 'temporal' dos poluentes, uma nova base de transações foi desenvolvida, com base na que já foi construída e mencionada acima. Com esta nova base, o algoritmo Apriori pode ser executado a fim de coletar os resultados do experimento em consideração. A Tabela 1 ilustra essas novas bases.

Tabela 1 - Exemplo de base de transações com poluição de até 2 dias antes das internações. Na tabela, T-i indica o índice de um dado poluente avaliado i dias antes da avaliação do número de internações

MP_{10} (T-2)	$MP_{2,5}$ (T-2)	NO_x (T-2)	O_3 (T-2)	MP_{10} (T-1)	$MP_{2,5}$ (T-1)	NO_x (T-1)	O_3 (T-1)	Internações
N1 - Boa	N1 - Boa	N1 - Boa	N1 - Boa	N1 - Boa	N2 - Moderada	N1 - Boa	N1 - Boa	MEDIA-BAIXA
N1 - Boa	N2 - Moderada	N1 - Boa	N1 - Boa	N1 - Boa	N1 – Boa	N1 - Boa	N1 - Boa	BAIXA
N1 - Boa	N1 - Boa	N1 - Boa	N1 - Boa	N1 - Boa	N1 – Boa	N1 - Boa	N1 - Boa	BAIXA
N1 - Boa	N1 - Boa	N1 - Boa	N1 - Boa	N1 - Boa	N1 – Boa	N1 - Boa	N1 - Boa	BAIXA
N1 - Boa	N1 - Boa	N1 - Boa	N1 - Boa	N1 - Boa	N1 - Boa	N1 - Boa	N1 - Boa	BAIXA

Dito isso, foram feitos experimentos que consideram as concentrações dos poluentes até três dias antes. Desse modo, foram desenvolvidas 3 novas bases de transações: uma para experimentos que consideram as concentrações de poluentes até 1 dia antes, uma para 2 dias antes e uma para 3 dias antes.

2.3 ALGORITMO E TÉCNICAS DE VALIDAÇÃO

Para realizar os experimentos de extração de regras, foi desenvolvido um *script* que utiliza um módulo com uma implementação pronta do algoritmo Apriori (PYPI, 2021), utilizando a linguagem de programação Python. Este módulo fornece uma função que executa a extração de regras a partir de uma lista de conjunto de itens e de valores mínimos de suporte e confiança, que são parâmetros utilizados pelo algoritmo Apriori a fim de se extrair as regras mais interessantes. Portanto, o *script* desenvolvido teve como objetivo carregar a base de transações, convertê-la em uma lista de conjunto de itens, chamar a função, passando esta lista e os valores mínimos de suporte e confiança como parâmetros, e exportar as regras resultantes para um arquivo CSV, nesta ordem.

Além disso, é importante deixar claro que os parâmetros suporte mínimo e confiança mínima foram definidos empiricamente. Para cada experimento, tentou-se atribuir o maior valor possível para o parâmetro de confiança mínima e, após esse valor já definido, o maior valor possível para o parâmetro de suporte mínimo.

Porém, mesmo com um ajuste cuidadoso dos parâmetros muitas regras foram geradas nos experimentos e muitas delas não foram interessantes para o presente estudo. Para tanto, foram aplicados alguns filtros às regras resultantes de cada experimento, com o objetivo de obter apenas aquelas que contribuam para a análise dos impactos dos poluentes nas internações. Esses filtros são descritos abaixo:

- Somente regras onde apenas o índice de hospitalização está no consequente.
- Apenas regras onde há pelo menos um poluente com índice igual a 'N3 – Ruim', 'N4 – Muito Ruim' ou 'N5 – Péssima'.
- Somente regras onde o índice de internações é igual a 'MEDIA-ALTA' ou 'ALTA'.

Por fim, mesmo com as métricas e os filtros, muitas regras foram geradas. Portanto, a análise das regras resultantes de cada experimento foi feita a partir de estatísticas sobre essas mesmas regras, ao invés da análise regra por regra. Dentre essas estatísticas, encontram-se, por exemplo, os pares índice-poluente e as combinações entre pares índice-poluente e entre os índices de internação que mais apareceram nos resultados.

3 RESULTADOS E DISCUSSÃO

Esta seção apresenta os resultados das extrações de regras para as bases de transações de cada ano.

3.1 RESULTADOS DE 2019

Para o ano de 2019, o experimento que considerou as concentrações de poluentes até 1 dia antes extraiu 8 regras de associação, todas com métrica de suporte igual a aproximadamente 0,003 e métrica de confiança igual a 1,00. A associação mais comum encontrada entre essas regras foi a associação entre a concentração do poluente O_3 na véspera da internação e com índice igual a 'N4 – Muito Ruim' com internações classificadas como 'ALTA'.

Ainda considerando o ano de 2019, o experimento que considerou concentrações de poluentes até 2 dias antes extraiu 364 regras de associação, todas com os mesmos valores de suporte e confiança mencionados acima. Com essas regras, foi encontrada a mesma associação citada acima, e, além disso, também foi encontrada associação entre a concentração do poluente O_3 dois dias antes da internação e com índice igual a 'N4 – Muito Ruim' com internações classificadas como 'ALTA'.

Por fim, o experimento que considerou as concentrações de poluentes até 3 dias antes, considerando o ano de 2019, extraiu 12.520 regras de associação, todas com os mesmos valores de suporte e confiança mencionados acima. Com essas regras, foram encontradas as mesmas duas associações citadas acima, e, além disso, também foi encontrada associação entre a concentração do poluente O_3 dois dias antes da internação e com um índice igual a 'N3 – Ruim' com internações classificadas como 'ALTA'.

Todos esses resultados podem ser vistos na Tabela 2.

Tabela 2 - Resultados de 2019

Poluentes Até	N° Regras Geradas	Suporte	Confiança	Associações Mais Comuns	Ocorrência das Associações mais comuns dentre as Regras Geradas
1 dia antes	8	0,003	1,00	O_3 (T-1) (N4) → ALTA	100,00%
2 dias antes	364	0,003	1,00	O_3 (T-1) (N4) → ALTA	35,16%
				O_3 (T-2) (N4) → ALTA	35,16%
3 dias antes	12.520	0,003	1,00	O_3 (T-1) (N4) → ALTA	16,35%
				O_3 (T-2) (N4) → ALTA	16,35%
				O_3 (T-2) (N3) → ALTA	10,73%

3.2 RESULTADOS DE 2020

Para o ano de 2020, o experimento que considerou as concentrações de poluentes até 1 dia antes extraiu 4 regras de associação, todas com suporte igual a aproximadamente 0,005 e confiança igual a 0,33. A associação mais comum encontrada entre essas regras foi a associação entre a concentração do poluente O_3 na véspera da internação e com índice igual a 'N3 – Ruim' com internações classificadas como 'ALTA'.

Ainda considerando o ano de 2020, o experimento que considerou concentrações de poluentes até 2 dias antes extraiu 32 regras de associação, todas com suporte igual a aproximadamente 0,005 e confiança igual a 0,66. Com essas regras, foi encontrada a mesma associação mencionada acima.

Por fim, o experimento que considerou concentrações de poluentes até 3 dias antes, considerando o ano de 2020, extraiu 1984 regras de associação, todas com suporte igual a aproximadamente 0,003 e confiança igual a 1,00. Com essas regras, também foi encontrada a mesma associação mencionada acima.

Todos esses resultados podem ser vistos na Tabela 3.

Tabela 3 - Resultados de 2020

Poluentes Até	N° Regras Geradas	Suporte	Confiança	Associações Mais Comuns	Ocorrência das Associações mais comuns dentre as Regras Geradas
1 dia antes	4	0,005	0,33	O_3 (T-1) (N3) → ALTA	100,00%
2 dias antes	32	0,005	0,66	O_3 (T-1) (N3) → ALTA	100,00%
3 dias antes	1984	0,003	1,00	O_3 (T-1) (N3) → ALTA	51,61%

4 CONCLUSÕES

Considerando o ano de 2019, concluímos que episódios do poluente O_3 com concentração classificada como 'N3 – Ruim' ou 'N4 – Muito Ruim' esteve associado a um 'ALTO' número de internações por doenças respiratórias até 2 dias depois.

Por outro lado, considerando o ano de 2020, concluímos que episódios do poluente O_3 com concentração classificada como 'N3 – Ruim' esteve associado a um 'ALTO' número de internações por doenças respiratórias até 2 dias depois.

Por fim, comparando as conclusões dos dois anos, percebemos que há mais semelhanças do que diferenças. Para ambos, o poluente O_3 com concentração classificada como 'N3 – Ruim' impactou em um 'ALTO' número de internações por doenças respiratórias até 2 dias. A única diferença é que, em 2019, o poluente O_3 com concentração classificada como 'N4 – Muito Ruim' também impactou em um 'ALTO' número de internações por doenças respiratórias até 2 dias depois.

Conclui-se então que, para o período estudado, O_3 foi o poluente atmosférico que mais provocou internações hospitalares por doenças respiratórias e que não houve diferenças significantes no período pandêmico (ano de 2020).

REFERÊNCIAS BIBLIOGRÁFICAS

AGUIAR, Laís Senhorini. Estudo da relação da qualidade do ar e variáveis meteorológicas na ocorrência de morbidade respiratória e circulatória na região metropolitana de São Paulo. 2015. 105 f. Dissertação (Mestrado em Engenharia Ambiental) - Universidade Tecnológica Federal do Paraná, Londrina, 2015.

BRAGA, Alfesio et al. Poluição atmosférica e saúde humana. **Revista USP**, n. 51, p. 58, 2001. DOI: 10.11606/issn.2316-9036.v0i51p58-71. Disponível em: https://www.revistas.usp.br/revusp/article/view/35099. Acesso em: 8 fev. 2021.

CETESB - Companhia Ambiental do Estado de São Paulo. **QUALAR – Sistema de Informações da Qualidade do Ar**. Disponível em: https://cetesb.sp.gov.br/ar/qualar/. Acesso em: 20 jan. 2021a.

CETESB - Companhia Ambiental do Estado de São Paulo. **Padrões de Qualidade do Ar**. Disponível em: https://cetesb.sp.gov.br/ar/padroes-de-qualidade-do-ar/. Acesso em: 21 jan. 2021b.

DATASUS – Departamento de Informática do SUS. **TABNET**. Disponível em: https://datasus.saude.gov.br/informacoes-de-saude-tabnet/. Acesso em: 20 jan. 2021.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data Mining : Concepts and Techniques**, 3rd. ed., Morgan Kaufman, 2012.

PAYUS, Carolyn et al. Association rules of data mining application for respiratory illness by air pollution database. **Int J Basic Appl Sci**, v. 13, n. 3, p. 11–16, jun. 2013.

PYPI - The Python Package Index. **apriori-python 1.0.4**. Disponível em: https://pypi.org/project/apriori-python/. Acesso em: 14 aug. 2021

SARLA, Gurmeet Singh. Air pollution : Health effects. **Medicina Legal de Costa Rica**, Heredia, v. 37, n. 1, p. 33–38, mar. 2020.

SAYAD, Saed. **Unsupervised Binning**. Disponível em: https://www.saedsayad.com/unsupervised_binning.htm. Acesso em: 29 jul. 2021

WHO - World Health Organization. **Air Pollution**. Disponível em: https://www.who.int/health-topics/air-pollution. Acesso em: 19 jul. 2021a.

WHO - World Health Organization. **Classification of Diseases (ICD)**. Disponível em: https://www.who.int/standards/classifications/classification-of-diseases. Acesso em: 27 jul. 2021b.