UNIVERSIDADE FEDERAL DE MINAS GERAIS
Departamento de Ciência da Computação

Introdução a Banco de Dados

TRABALHO PRÁTICO 2

Acesso, coleta, gerenciamento e análise de conjuntos de dados públicos

GIOVANNI RUSSO PASCHOAL
LUCAS DA SILVA SANTOS
RAFAEL PRADO PANIAGO

Belo Horizonte
Junho, 2024

Introdução

Quando se fala de saúde pública, entender as relações intrínsecas entre fatores ambientais e o bem-estar respiratório é altamente importante. Como parte da disciplina “Introdução a Banco de Dados (IBD)” e o tópico que trata sobre a Lei de Acesso à Informação (Lei nº 12.527/2011, ou LAI), este relatório detalha as decisões tomadas na utilização de dados públicos do governo federal, entendendo como esses dados são coletados e o que pode ser melhorado. Exploramos, com base nos conjuntos de dados escolhidos, a conexão crucial entre a qualidade do ar e sua influência na incidência de várias doenças e infecções.

O presente documento está dividido em partes que relatam a nossa jornada de acesso, coleta, gerenciamento, tratamento, normalização, e análise dos dados. Na primeira parte, está descrito o processo de aquisição e curadoria dos bancos de dados de diferentes fontes. O processo de limpeza dos dados também é abordado, sendo usado para garantir uma melhor organização e confiabilidade quando se trata de bancos de dados relacionais. Com os conjuntos de dados meticulosamente preparados, partimos para a análise exploratória de dados, descrita na segunda parte do documento. Usamos técnicas estatísticas e visualização de dados para descobrir padrões e resultados, a partir de nossas hipóteses e objetivos. Na terceira parte, discorremos acerca das limitações e potenciais enviesamentos presentes nos conjuntos de dados analisados, com uma avaliação crítica. Essa seção serve como resumo dos pontos que foram tratados nas seções anteriores no que se diz respeito às escolhas feitas para conciliar a insuficiência, escassez, ou restrições dos dados. Por fim, apresentamos uma conclusão com considerações adicionais e recomendações para trabalhos futuros.

O resultado desse processo pode ser consultado na página https://github.com/GiorussoP/TP2-IBD-UFMG, referente ao repositório onde estão armazenados os dados originais, os dados tratados durante a produção do trabalho, e os documentos gerados a partir das melhorias realizadas.

1. Obtenção e Correção dos Dados

Iniciamos nosso trabalho com a busca por dados que fizessem uma ligação entre aspectos ambientais e saúde, dentre as coleções disponibilizadas pelo governo federal. O primeiro conjunto de dados que encontramos foi o *Ar Puro*[1], na plataforma MonitorAr, um aplicativo mantido pelo Ministério do Meio Ambiente e Mudança do Clima (MMA). Essa plataforma foi crucial ao mostrar-nos importantes indicadores de qualidade do ar, incluindo concentrações de materiais particulados, níveis de ozônio, e fatores meteorológicos. Além disso, obtivemos também informações acerca das estações de monitoramento dos indicadores. Os dados cobrem o período entre

janeiro e novembro de 2022, providenciando uma visão detalhada – mas extremamente desnormalizada – sobre a qualidade do ar durante aquele ano.

Em seguida, usamos o *Sistema de Informação de Doenças e Agravos de Notificação*[1], na plataforma DataSUS do Ministério da Saúde, para captar dados acerca da incidência de asma, influenza, intoxicações respiratórias, meningite, e pneumoconioses nos municípios brasileiros. Em suma, doenças relacionadas com o ar ou transmitidas através dele. Ao incorporar tais informações em nosso banco de dados, nosso objetivo foi estabelecer potenciais correlações entre os índices de qualidade de ar e a ocorrência e agravamento dessas enfermidades. Por fim, decidimos utilizar também dados do Instituto Brasileiro de Geografia e Estatística (IBGE) sobre os municípios[1] do país e suas respectivas populações[1]. Isso foi importante para que pudéssemos relacionar as diferentes tabelas entre si, além de levar em consideração o tamanho da população ao analisar as incidências das doenças.

Com todos os conjuntos de dados em mãos, foi fácil perceber como nenhum deles estava propício para análises e manipulações em um sistema gerenciador de bancos de dados relacionais. Listamos abaixo as principais ações tomadas pelo grupo para unir e transformar os dados, garantindo compatibilidade entre diferentes fontes e formando uma base sólida para nossas análises posteriores.

1. **Remoção de redundâncias de nomes de municípios nas tabelas de doenças e infecções.** Para relacionar as doenças com os municípios onde foram notificadas, usamos apenas os códigos dos municípios.
2. **Conversão dos códigos de municípios para o formato de 6 dígitos.** O Código dos Municípios do IBGE é composto por 7 dígitos, sendo os dois primeiros referentes ao código da Unidade da Federação e o último, um verificador. Decidimos usar somente os 6 primeiros dígitos, sem o verificador, para manter consistência com os municípios que não possuem verificador informado.
3. **Conversão dos nomes de municípios para uso com letras maiusculas.** Isso garante consistência entre os conjuntos de dados, possibilitando relacionar estações de monitoramento da qualidade do ar, municípios, e doenças.
4. **Exclusão de descrições de dentro das tabelas.** A maioria das tabelas de dados que conseguimos obter estão no formato CSV. Muitas delas, em especial as referentes ao sistema MonitorAr, possuíam longas frases descritivas dentro das colunas, o que é péssimo para uma análise integrada e manipulação do banco de dados.
5. **Retirada da coluna redundante *TOTAL* da tabela de intoxicações.** Essa tabela contém diferentes tipos de intoxicações e a coluna referenciada mostrava a soma dos casos individuais de cada um desses tipos. Em um sistema

gerenciador de bancos de dados relacionais, esse cálculo é facilmente realizado.

6. **Reorganização das tabelas da qualidade do ar e adição de valores extras para medidas da mesma grandeza.** As tabelas provenientes da plataforma MonitorAr contavam com muitas informações separadas que, afinal, eram redundantes ou não seriam necessárias para o tipo de análise proposto na introdução deste documento. Logo, muitas colunas foram removidas e as linhas foram reorganizadas. Por outro lado, adicionamos medidas de média e desvio padrão para cada indicador, com o intuito de visualizar melhor os indicadores de qualidade do ar durante o ano de 2022 como um todo. Com isso, conseguimos remover linhas duplicadas que, muitas vezes, traziam a medição de um mesmo poluente em múltiplos horários em um mesmo dia. Agora, cada linha representa um conjunto de medidas de uma estação de monitoramento específica.

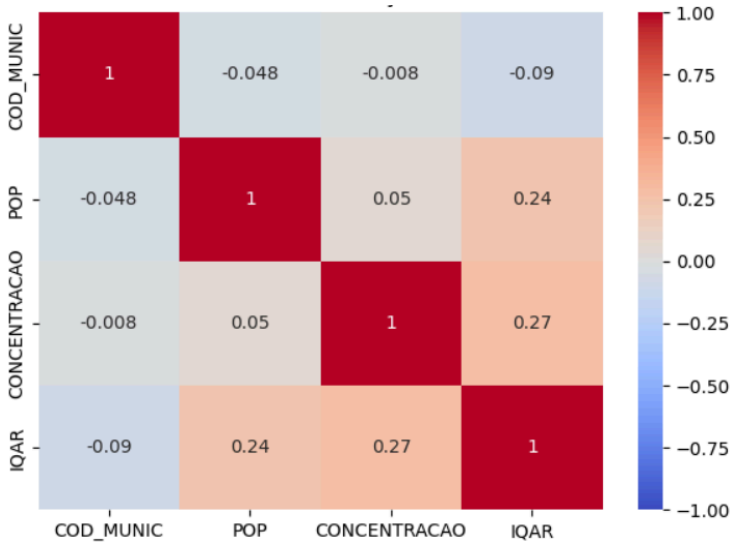
Por fim, com as novas tabelas organizadas a partir dos dados originais, foram desenvolvidos alguns documentos importantes para uma posterior publicação do banco de dados. Em primeiro lugar, o diagrama UML representa visualmente as relações entre as tabelas e os conceitos de chaves como identificadores únicos no modelo relacional. Além disso, um dicionário de dados completo define cada tabela e suas colunas, incluindo os significados, tipos de dados, e unidades de medida (quando aplicável) para cada atributo. Uma reorganização dos metadados obtidos nas fontes primárias dos dados também foi produzido para auxiliar a localização de informações como o órgão responsável, data, forma de coleta dos dados, limitações, e abrangência geográfica. Os três documentos podem ser consultados diretamente no repositório dedicado ao atual trabalho.

(1) Os links para essas fontes e plataformas encontram-se nas referências, ao fim do documento.

2. Análise Exploratória

Uma vez que nosso banco de dados foi corrigido e melhorado, possibilitando uma melhor interação e integração entre as tabelas existentes em um modelo relacional, seguimos para uma análise exploratória dos dados, com base nos requerimentos do trabalho e naquilo que foi explicado em aula pelo professor Clodoveu Davis Jr. Neste ponto, deparamo-nos fortemente com as limitações existentes na coleta e publicação de dados públicos, em especial nas fontes consultadas para o projeto. Isso é explorado em mais detalhes no próximo capítulo deste relatório, porém é importante dizer que esse problema gerou dificuldades para que encontrássemos conclusões e analisássemos até onde elas eram válidas.

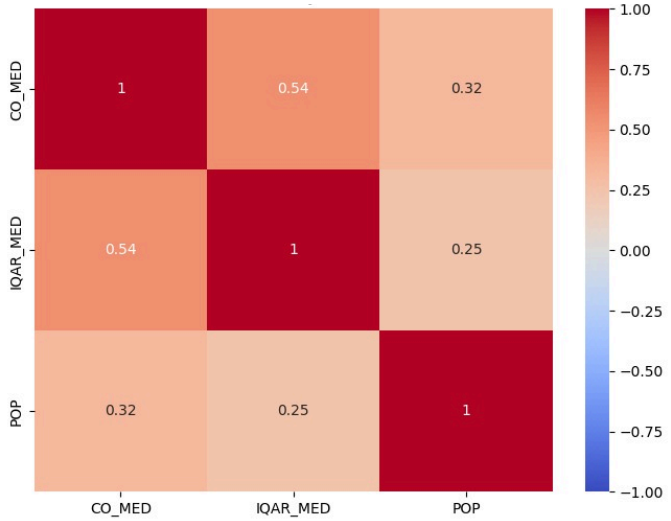
Primeiramente, como nosso trabalho trata sobre as ligações entre a qualidade do ar e a saúde respiratória, procuramos entender de que forma isso se dá. Nossa primeira e principal hipótese foi: “Municípios altamente populosos (capitais, por exemplo) provavelmente possuem pior qualidade de ar, visto que sofrem com a intensa atividade industrial e o tráfego numeroso, atividades que, sabemos pelos estudos científicos, liberam na atmosfera gases e poluentes prejudiciais ao meio-ambiente e, consequentemente, à saúde humana”. Em síntese, buscamos analisar a correlação do aumento populacional com a piora da qualidade do ar, e da piora da qualidade do ar com o aumento de doenças respiratórias (ou transmitidas pelo ar). Realizando uma junção entre as tabelas `municipios`, `populacao_2022`, e `qualidade_ar_2022`, foi fácil perceber uma correlação positiva (0.27) entre população e o IQAr (Índice de Qualidade do Ar, uma coluna da tabela `qualidade_ar_2022`), o que indica que populações maiores tendem a viver com uma pior qualidade do ar. Outra correlação positiva nesse caso ocorreu entre a população e a concentração de NO_2 (dióxido de nitrogênio), um gás produzido principalmente pela combustão de combustíveis fósseis ou a partir da reação de NO (óxido nítrico) com outros gases na atmosfera, como o ozônio. A partir dessa descrição, é possível perceber que uma maior quantidade de NO_2 é encontrada em maiores centros urbanos, aqueles que possuem maior população, o que fica claro na matriz de correlação abaixo.



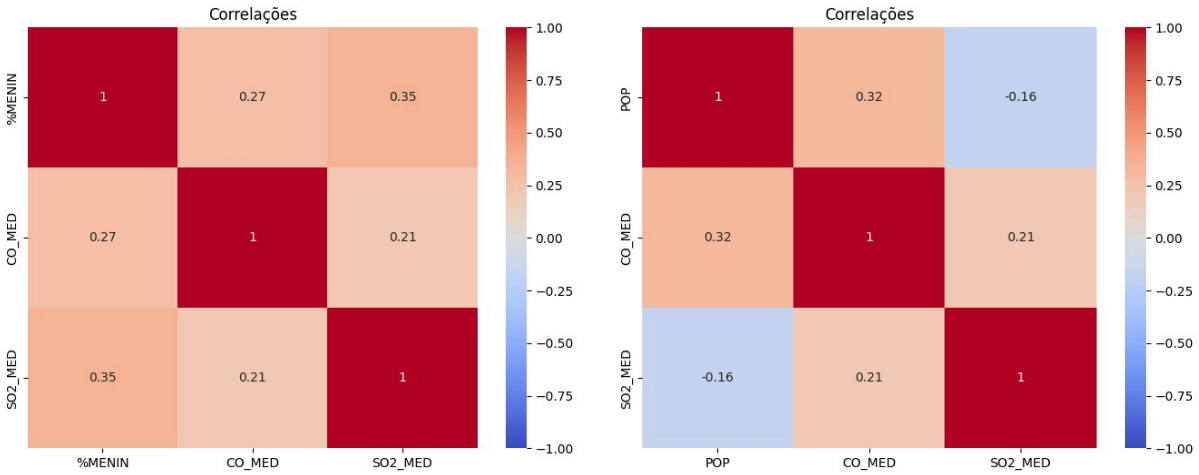
Uma relação interessante, nessa mesma matriz de correlação, é o valor negativo (-0.09) observado entre o IQAr e o código dos municípios. Como já mencionado, os dois primeiros dígitos deste código são referentes à Unidade da Federação onde se encontra o município. Isso pode indicar que os estados mais ao norte e nordeste do país possuem melhor qualidade do ar em média. Isso faz sentido ao considerar que esses estados são também, em média, menos industrializados. Contudo, os dados que temos não nos permitem confirmar isso com precisão, uma vez que, atualmente,

somente 12 Unidades da Federação são abrangidas pelas estações de monitoramento que capturam as informações.

Posteriormente, percebemos também uma correlação positiva entre a população e a concentração média de CO (monóxido de carbono, um produto de muitas combustões naturais e causadas pelos seres humanos, como aquelas de gasolina e diesel) e também com o índice de qualidade do ar em geral. Isso indica que municípios mais populosos contam com o aumento nos níveis de CO e poluição do ar, confirmando nossa hipótese inicial. Essas conclusões podem ser visualizadas abaixo.



Partindo para uma análise diferente, que engloba as doenças e intoxicações e sua relação com a qualidade do ar, analisamos os casos de meningite (a doença com a maior quantidade de dados disponíveis) e percebemos que existe uma relação positiva entre a porcentagem desses casos e a concentração de SO_2 (dióxido de enxofre). Esse gás é conhecido por bloquear e irritar as vias aéreas. Assim, o contato constante com SO_2 pode enfraquecer o sistema respiratório, o que facilita inflamações caracterizadas como meningite. Mesmo controlando e normalizando o número da população, os dados analisados seguem essa direção. A correlação entre esses casos e a população foi a mais forte dentre todas as análises realizadas. No entanto, como mencionamos acima, maiores populações estão fortemente ligadas ao aumento de CO. Por sua vez, a relação entre os casos de meningite e a concentração de gás CO é mais baixa do que quando consideramos o SO_2. Isso sugere que outros fatores podem mediar aquela relação, o que não é possível compreender olhando somente para os conjuntos de dados que estão disponíveis, caracterizando mais uma limitação desse banco de dados. Adicionamos abaixo as matrizes de correlação referentes a essa análise.



Finalmente, outras correlações importantes envolvem a porcentagem de intoxicações por produtos químicos e a concentração de SO_2, o que mostra uma potencial prevalência dessas intoxicações em municípios fortemente industriais, com uma população mais baixa do que capitais, por exemplo. Por outro caminho, encontramos também uma correlação inversa entre a porcentagem de intoxicações por raticida e a taxa de população/monóxido de carbono. Isso possivelmente está relacionado ao maior uso de raticida em municípios predominantemente agrícolas e/ou à prática de auto-aplicação desse pesticida em cidades menores, sem a contratação de profissionais especializados nessa aplicação.

3. Limitações e Qualidade dos Dados

Nesta seção do texto, decidimos listar as principais barreiras encontradas no processo de manipulação e análise exploratória dos conjuntos de dados públicos, incluindo algumas sugestões de melhoramento da qualidade desses dados.

1. As tabelas referentes às enfermidades e intoxicações possuem dados que engloba um extenso período de anos, mas eles não são consistentes entre si. Por exemplo, temos dados de 2007 a 2023 para os casos confirmados de meningite, mas apenas entre 2009 e 2010 para a influenza. Neste último caso, é possível perceber que os dados abordam apenas uma epidemia de curto prazo. Por sua vez, o conjunto de dados acerca das pneumoconioses possuem apenas 90 pontos intercambiáveis. Essas diferenças limitaram nossas análises exploratórias ao dificultar a interpretação correta da relação entre a qualidade do ar e a saúde respiratória. Além disso, é incerto entender as proporções entre casos inexistentes e casos existentes, mas não notificados.

2. Na tabela sobre qualidade do ar, o IQAr contém avaliações de qualidade para um número muito limitado de municípios (apenas 69 de um total de 800 mil medidas). Isso também restringe análises profundas que mostrem correlações confiáveis entre as concentrações dos poluentes no ar e doenças específicas.
3. Os dados sobre municípios do IBGE possuem um problema de chave primária, com nomes duplicados de municípios em diferentes estados (por exemplo, "Bom Jesus", com 6 registros). Para melhorar isso, seria fundamental considerar o uso dos nomes juntamente aos códigos dos municípios como chave primária.
4. Como abordado anteriormente, os dados da plataforma MonitorAr são provenientes do monitoramento da qualidade do ar em apenas 12 estados brasileiros: Bahia, Ceará, Espírito Santo, Mato Grosso do Sul, Minas Gerais, Paraná, Pernambuco, Rio de Janeiro, Rio Grande do Sul e São Paulo. Além disso, o Distrito Federal e Goiás realizam o monitoramento de poluentes de forma manual, o que torna a geração dos dados menos rápida e eficaz. Isso pesou em uma de nossas análises e impede uma visualização nacional acerca da qualidade do ar, em primeiro lugar; e de sua relação com as doenças e intoxicações ligadas ao ar. Ademais, cidades industriais costumam ter mais estações de monitoramento do ar, o que torna nossa análise enviesada.
5. Algumas fontes de dados não contam com metadados adequados, o que gera incertezas quanto ao processo de coleta dos dados, que influencia bastante nos resultados que podemos ter. Por isso, foi preciso reorganizar todos eles e buscar mais informações nos websites e plataformas dos Ministérios do Governo Federal.

Conclusão e Trabalhos Futuros

Neste trabalho, nos aprofundamos no campo de dados públicos e entendemos importantes ideias que surgem no processo de coleta, organização, e análise desses dados, em específico, usando as relações entre saúde pública e bem-estar respiratório no Brasil. Utilizamos conceitos de limpeza e normalização de dados para análise em modelo relacional, mesclando fontes de dados diversas e criando consistências entre eles. Além disso, verificamos também as inconsistências e delimitamos o escopo desses dados, entendendo o que eles podem e não podem comprovar. Quando voltamos à nossa análise exploratória de dados, é claro perceber que existem maiores riscos de doenças respiratórias e/ou transmitidas pelo ar em áreas altamente populosas. Esses municípios contém, no geral, maiores níveis de poluição do ar. Essa exploração inicial foi importante para gerar um banco de dados melhorado, que está armazenado agora em um repositório público no GitHub, e que pode ser utilizado posteriormente para estudos de causas mais profundas, análises espaciais, e até mesmo socioeconômicas, com a integração de outros fatores.

Ademais, é essencial destacar que as correlações observadas não necessariamente implicam causas no mundo real. São necessárias pesquisas adicionais para se estabelecer relações causais. Outros fatores, como acesso a saúde e status socioeconômico podem também influenciar os casos de enfermidades respiratórias. É crucial manter a coleta e análise de dados no longo prazo, para melhor monitorar as tendências e criar soluções e intervenções efetivas para a poluição do ar.

Recomendamos aos órgãos responsáveis que sejam implementadas medições de emissões e regulamentos sobre a qualidade do ar mais restritivas, especialmente em regiões com altas taxas de população. Além disso, é necessário melhorar as pesquisas sobre saúde para detectar doenças e intoxicações respiratórias. Seria também crucial, em especial com o avanço da urbanização e das mudanças climáticas no contexto internacional e nacional, investir em pesquisas e análises que tornem possível compreender os mecanismos que associam a qualidade do ar e as doenças relacionadas a ela, para além das que mencionamos neste relatório.

Referências

Códigos dos Municípios | IBGE. Disponível em: <https://www.ibge.gov.br/explica/codigos-dos-municipios.php>. Acesso em: 21 maio 2024.

Doenças e Agravos de Notificação – 2007 em diante (SINAN) – DATASUS. Disponível em: <https://datasus.saude.gov.br/acesso-a-informacao/doencas-e-agravos-de-notificacao-de-2007-em-diante-sinan/>. Acesso em: 20 maio 2024.

Estimativas da população residente para os municípios e para as unidades da federação | IBGE. Disponível em: <https://www.ibge.gov.br/estatisticas/sociais/populacao/9103-estimativas-de-populacao.html?edicao=31451&t=resultados>. Acesso em: 21 maio 2024.

ELMASRI, R.; NAVATHE, S. **Fundamentals of database systems**. [s.l.] Pearson, 2016.

PROGRAMAÇÃO E DESENVOLVIMENTO DE SOFTWARE I. **IBD Lei de Acesso à Informação**. Disponível em: <https://www.youtube.com/watch?v=o5_um_--jpU>.

Portal de Dados Abertos. Disponível em: <https://dados.gov.br/dados/conjuntos-dados/ar-puro-monitorar>. Acesso em: 20 maio 2024.