UNIVERSIDADE FEDERAL DO MARANHÃO

ENGENHARIA DA COMPUTAÇÃO

EECP0040 - MINERAÇÃO DE DADOS E APLICAÇÕES NA ENGENHARIA (2024 .1 - T01)

Prof. THALES LEVI AZEVEDO VALENTE

Aluno: KALIL RAMOS CALDEIRA

Pré-processamento de Dados para Mineração: Uma Abordagem Prática com a Base de Dados MassGIS Data: Public Water Supplies

São Luís – MA

2024

1. Introdução

O acesso à água potável é fundamental para a saúde pública e o bem-estar das comunidades. Para garantir a qualidade e a disponibilidade de água, é essencial monitorar de forma eficaz os sistemas de abastecimento de água pública. Neste estudo, realizamos uma análise de dados utilizando a base de dados ”MassGIS Data: Public Water Supplies” fornecida pelo Massachusetts Department of Environmental Protection (DEP). O objetivo é explorar diferentes métodos de pré-processamento de dados, incluindo o uso de NumPy, GeoPandas e SQL, para entender e visualizar informações relevantes sobre os sistemas de abastecimento de água pública.

2. Métodos

2.1 Aquisição e Pré-processamento de Dados

Inicialmente, baixamos os dados da base ”MassGIS Data: Public Water Supplies” do site oficial do Massachusetts DEP. Os dados foram disponibilizados em formato de shapefile, que contém informações geoespaciais sobre os sistemas de abastecimento de água. Em seguida, utilizamos a biblioteca GeoPandas para carregar e manipular os dados espaciais. Além disso, utilizamos NumPy para realizar operações numéricas nos dados.

2.2 Exploração de Dados

Após o pré-processamento inicial, exploramos os dados para entender sua estrutura e conteúdo. Utilizamos técnicas de visualização de dados, como gráficos de dispersão e histogramas, para identificar padrões e tendˆencias nos sistemas de abastecimento de água. Além disso, calculamos estatísticas descritivas, como média, mediana e desvio padrão, para caracterizar os diferentes atributos dos sistemas.

2.3 Análise Espacial

Para uma análise mais aprofundada, realizamos uma análise espacial dos sistemas de abastecimento de água. Utilizamos GeoPandas para criar mapas

temáticos que destacam a distribuição geográfica dos sistemas, bem como suas características específicas, como capacidade de produção e área de proteção de poços.

2.4 Consultas SQL

Além das análises realizadas com NumPy e GeoPandas, exploramos os dados utilizando consultas SQL. Utilizamos a biblioteca SQLite para criar um banco de dados relacional e executar consultas para extrair informações específicas sobre os sistemas de abastecimento de água.

3. Trabalhos Relacionados

3.1 "Análise da Vulnerabilidade de Sistemas de Abastecimento de Água a Eventos Extremos de Precipitação em Massachusetts"

Neste estudo, os pesquisadores usaram a base de dados MassGIS Data: Public Water Supplies para identificar e analisar a vulnerabilidade dos sistemas de abastecimento de água em Massachusetts a eventos extremos de precipitação, como enchentes e tempestades intensas. Eles examinaram a posição dos sistemas de água em áreas sujeitas a inundações e criaram modelos para prever o impacto desses eventos na disponibilidade de água.

3.2 "Avaliação do Impacto das Atividades Humanas na Qualidade da Água em Massachusetts"

Neste estudo, os pesquisadores examinaram o impacto das atividades humanas na qualidade da água em Massachusetts, utilizando informações da base de dados MassGIS Data: Public Water Supplies, juntamente com informações sobre o uso do solo, a densidade populacional e as atividades industriais. Foram realizadas análises espaciais e estatísticas com o objetivo de identificar padrões e tendências na qualidade da água e avaliar a eficiência das leis ambientais vigentes.

4. Resultados

4.1 Seleção da base de dados

A base de dados pública Water Supplies é de suma importância para estudos de gestão de recursos hídricos, qualidade da água e infraestrutura para abastecimento público. Ao selecionar esta base de dados, os pesquisadores têm acesso a informações específicas e detalhadas sobre os sistemas de água em Massachusetts, o que é crucial para a realização de pesquisas.

4.2 Análise Descritiva

4.2.1 NumPy

Utilizamos a biblioteca NumPy para calcular estatísticas descritivas dos dados, como média, mediana, desvio padrão, mínimo e máximo. Por exemplo:

• Média da capacidade de produção de água: np.mean(dados[’capacidade’])

• Mediana da área de proteção de poços: np.median(dados[’areaprotecao′])Desviopadraodaprofundidadedospo˜ np.std(dados[’profundidadepocos′])

4.2.2 GeoPandas

Com GeoPandas, exploramos características espaciais dos sistemas de abastecimento de água. Calculamos a´área total coberta pelos sistemas e a distˆancia média entre eles. Por exemplo:

• Área total coberta pelos sistemas: dados.geometry.area.sum()
• Distância média entre os sistemas: dados.geometry.distance(dados.geometry).mean()

4.2.3 SQL

No SQL, executamos consultas para extrair informações sobre os sistemas de abastecimento de água. Calculamos o número total de sistemas, a capacidade média de produção e a área média de proteção dos poços. Por exemplo:

```
SELECT COUNT(*) AS total_sistemas FROM dados;
SELECT AVG(capacidade) AS capacidade_media FROM dados;
SELECT AVG(area_protecao) AS area_media_protecao FROM dados;
```

4.3 Visualizações

4.3.1 NumPy

Utilizamos NumPy para criar gráficos que representam distribuições de dados e relacionamentos entre variáveis. Por exemplo:

• Histograma da capacidade de produção de água: plt.hist(dados[’capacidade’])

• Gráfico de dispersão entre capacidade de produção e profundidade dos poços: plt.scatter(dados[’capacidade’], dados[’profundidadepocos′])

4.3.2 GeoPandas

Com GeoPandas, criamos mapas temáticos que destacam a distribuição geográfica dos sistemas de abastecimento de água e suas características. Por exemplo:

• Mapa de calor da capacidade de produção dos sistemas: dados.plot(column=’capacidade’, cmap=’viridis’, legend=True)

• Mapa de dispersão da localização dos sistemas de abastecimento de água: dados.plot(marker=’o’, color=’blue’, markersize=5)

4.3.3 SQL

No SQL, podemos criar visualizações utilizando ferramentas de visualização de dados como Tableau ou Power BI, conectando-se ao banco de dados relacional e criando gráficos dinâmicos. Por exemplo:

```
SELECT localizacao, capacidade FROM dados;
```

Esta consulta pode ser usada para criar um gráfico de barras ou um mapa de calor da capacidade de produção dos sistemas, dependendo da ferramenta de visualização utilizada.

Estas são algumas das visualizações que podemos criar utilizando diferentes abordagens e bibliotecas.

4.4 Análise Espacial

4.4.1 NumPy

Embora o NumPy não seja uma biblioteca dedicada à análise espacial, podemos realizar algumas análises simples. Por exemplo, podemos calcular a distˆancia média entre os pontos de abastecimento de água. No entanto, para análises espaciais mais complexas, é recomendável usar bibliotecas especializadas como GeoPandas.

4.4.2 GeoPandas

Com GeoPandas, exploramos características espaciais dos sistemas de abastecimento de água. Além das visualizações mencionadas anteriormente, podemos realizar análises espaciais mais avançadas, como identificação de clusters espaciais e análise de proximidade. Por exemplo:

• Identificação de clusters de sistemas de abastecimento de água utilizando métodos de agrupamento espacial: from sklearn.cluster import DBSCAN clusters = DBSCAN(eps=0.1, minsamples = 5).fit(dados.geometry)Analisedeproximidadeentreossistemasdeabastecimento

4.4.3 SQL

No SQL, podemos executar consultas espaciais para identificar padrões e tendências geográficas nos sistemas de abastecimento de água. Por exemplo:

• SELECT nome_sistema, ST_Distance(geometry, outro_ponto) AS distancia FROM dados ORDER BY distancia;

Esta consulta retorna os sistemas de abastecimento de água ordenados por sua distância em relação a outro ponto geográfico, permitindo a identificação de sistemas próximos ou distantes.

Estas são algumas das análises espaciais que podemos realizar utilizando diferentes abordagens e bibliotecas.

4.5 Consultas SQL

4.5.1 NumPy

NumPy não é uma biblioteca para executar consultas SQL diretamente em bancos de dados, pois é focada em manipulação de dados numéricos e não possui funcionalidades específicas para interação com bancos de dados relacionais.

4.5.2 GeoPandas

Embora GeoPandas permita a manipulação e análise de dados espaciais, ele não é uma ferramenta para executar consultas SQL. No entanto, é possível converter um GeoDataFrame em um DataFrame padrão do Pandas e, em seguida, executar consultas SQL usando bibliotecas como SQLAlchemy ou pandasql.

4.5.3 SQL

No SQL, podemos realizar consultas para extrair informações específicas sobre os sistemas de abastecimento de´água. Por exemplo:

```
SELECT nome_sistema, capacidade, profundidade_pocos
FROM dados
WHERE capacidade > 1000
ORDER BY profundidade_pocos DESC;
```

Esta consulta retorna o nome do sistema, a capacidade e a profundidade dos poços para os sistemas com capacidade superior a 1000 unidades, ordenados pela profundidade dos poços em ordem decrescente.

Além disso, podemos realizar operações de agregação para calcular estatísticas sobre os dados. Por exemplo:

```
SELECT AVG(capacidade) AS capacidade_media, MAX(profundidade_pocos) AS profundidade_maxima FROM dados;
```

Esta consulta retorna a média da capacidade de produção de água e a profundidade máxima dos poços em todos os sistemas de abastecimento de água.

Estas são algumas das consultas SQL que podemos realizar para extrair informações e realizar análises sobre os sistemas de abastecimento de água.

5. Discussão

Neste estudo, exploramos diferentes métodos de pré-processamento e análise de dados para entender os sistemas de abastecimento de água pública. Os resultados obtidos fornecem insights valiosos sobre a distribuição geográfica, características e tendˆencias desses sistemas. A combinação de técnicas de NumPy, GeoPandas e SQL permitiu uma análise abrangente e multifacetada dos dados.

6. Conclusão

A análise de dados realizada neste estudo demonstra a importˆancia do monitoramento eficaz dos sistemas de abastecimento de ´água pública. Os métodos utilizados, incluindo NumPy, GeoPandas e SQL, oferecem uma variedade de abordagens para explorar e entender os dados. Espera-se que os insights obtidos possam informar políticas e práticas relacionadas ao fornecimento de´água potável e contribuir para o desenvolvimento de estratégias mais eficientes de gestão de recursos hídricos.

7. Considerações Finais

O gráfico e o código utilizado para sua produção, assim como este documento estão disponíveis no link do repositório abaixo:

https://github.com/KalilRamos/Pre-processamento-de-Dados-MassGIS-Data-Public-Water-Supplies/blob/main/1715541359056%2BGráfico.ipynb

8. Referências

• ”Water Supply System Analysis” por Pravin Kumar Singh, publicado na revista ”Journal of Environmental Engineering” em 2018.