UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

ANA LUIZA ALMEIDA SOARES
Orientador: Prof. Dr. Rodrigo Cesar Pedrosa Silva
Coorientador: Mestre Pedro Saint Clair Garcia

AVALIAÇÃO DE DIVERSAS APIS DE GECODIFICAÇÃO
SUBTÍTULO

Ouro Preto, MG
2023

UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

ANA LUIZA ALMEIDA SOARES

AVALIAÇÃO DE DIVERSAS APIS DE GECODIFICAÇÃO

SUBTÍTULO

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Rodrigo Cesar Pedrosa Silva
Coorientador: Mestre Pedro Saint Clair Garcia

Ouro Preto, MG
2023

Resumo

Síntese do trabalho contendo um único parágrafo. O resumo deve ser feito de forma clara, concisa e seletiva de todo o texto, ressaltando o objetivo, o método, os resultados e a conclusão (**??**). A norma da ABNT ainda recomenda que a primeira frase seja uma explicação do tema principal, seguindo da informação da natureza do trabalho (pesquisa experimental, pesquisa biliográfica, estudo de caso, pesquisa de campo, etc.). Apresente os objetivos (geral e específicos); justificativa e a metodologia desenvolvida. Também deve ser inserido as conclusões finais, apresentando uma síntese dos principais resultados alcançados e o valor da pesquisa no contexto acadêmico. Sugere-se entre 150 a 500 palavras.

Palavras-chave: Palavra-chave 1. Palavra-chave 2. Palavra-chave 3.

As palavras-chave devem estar separadas por ponto e finalizadas também por ponto. Devem ser escolhidos termos que descrevem o conteúdo do trabalho.

Abstract

This is the english abstract.

Keywords: Keywords1, Keywords2, Keywords3.

Lista de Ilustrações

Lista de Tabelas

Lista de Algoritmos

Lista de Abreviaturas e Siglas

ABNT	Associação Brasileira de Normas Técnicas
DECOM	Departamento de Computação
UFOP	Universidade Federal de Ouro Preto
SIG	Sistema de Informação Geográfica
EUA	Estados Unidos da América

Lista de Símbolos

Γ	Letra grega Gama
Λ	Lambda
ζ	Letra grega minúscula zeta
$\in$	Pertence

Sumário

1 Introdução

1.1 Endereços e Geocodificação

> Quase tudo que acontece, acontece em algum lugar. Saber o local onde algo acontece pode ser fundamental.
>
> (**??**)

No livro (**??**) os autores explicam a relação entre a humanidade e a localização. Para eles, é claro que a maior parte da atividade humana é feita no planeta Terra e por isso a vida é fortemente ligada a localidade. Sendo assim, entender e manipular informações geográficas está no cerne de qualquer aplicação que envolve a humanidade. Além disso, os autores explicam que decisões importantes podem causar consequências geográficas. Um exemplo seria uma movimentação financeira, que em um caso mais extremo, poderia causar uma crise econômica em uma determinada região.

No artigo (**??**), o autor traz aspectos importantes das informações geográficas que complementam o que foi dito anteriormente. Para ele, endereço é a principal forma de de conceitualizar localização no mundo atual. Isso se deve ao fato dos endereços serem utilizados em diversas aplicações de diferentes campos de estudo, como na saúde (**??????**), nas ciências sociais (**??**), na análise de criminal ou judiciária (**??**), na análise ambiental (**??**), na ciência da computação (**??**), na economia (**??**) entre outras.

Para isso, é necessário gerar a representação computacional do endereço, de forma com que as aplicações possam utilizá-lo. A representaçãomais comum segundo (**??**) é a representação por meio de coordenadas x e y em um plano, geralmente a medida é latitude e longitude. O processo de tranformação em um endereço nessas coordenadas é chamado de Geocodificação ou Georreferenciamento. Para (**??**) esse processo consiste em 3 etapas:

- Processamento do endereço de entrada: o endereço será lido, dividido em componentes (rua, número, bairro, etc), padronizado, cada campo é atribuído a uma categoria e por fim, serão indexadas as categorias necessárias;
- Busca na base referência: de acordo com o algoritmo escolhido, será realizada uma busca na base referência afim de selecionar e classificar potenciais canditados para resposta;
- Seleção do(s) canditado(s) para resposta: com a busca realizada será feita uma análise da classificação gerada por ela e serão escolhidos os melhores canditados.

Segundo (**??**), para além de representar computacionalmente um endereço, o georreferenciamento utilizando latitude e longitude tem diversas vantagens:

- Sistema com precisão espacial: é capaz de indicar com precisão alta a localização de um certo endereço;
- Permitem cálculos de distância: por ser um sistema espacial, ele permite que a obtenção da distância e por consequência que outras métricas sejam calculados para o endereço;
- Compreensão global: é um sistema utilizado mundialmente, sendo geralmente mais fácil de identificar e compreender;

Apesar de todas as vantagens e aplicações, o processo de geocodificação pode causar informações erradas. No livro (**??**) os autores nomeiam essas falhas de informação como incertezas. Para coompreender o que é incerteza, é necessário levar em conta outros aspectos da falha na informação. Assim, são incluídos os conceitos:

- Erro: Diferença entre o observado e o obtido;
- Falta de acurácia: Diferença entre a realidade e a nossa representação da realidade;
- Ambiguidade: mais de um valor igual ao outro;
- Indefinição: falta de informações necessárias.

Após conceitualizar esses termos, eles definem incerteza como: “medida da compreensão do usuário sobre a diferença entre o conteúdo de um conjunto de dados e os fenômenos reais que os dados devem representar” (**??**). Ou seja, incerteza é a medida que descreve a compreensão do usuário em relação ao conjunto de dados obtidos e a realidade que esse conjunto de dados pretender observar. A partir disso, incerteza foi aceita como uma boa medida de avaliação da qualidade dos Sistemas de Informação Geográfica (SIG).

1.2 APIs de Geocodificação e Análise de qualidade

Atualmente, no TerraLAB - Laboratório de pesquisa e capacitação em software (**??**), utilizamos de informações geográficas para desenvolvimento das aplicações. Essas aplicações utilizam os endereços geocodificados para desenhar mapas, criar rotas e centros de alcance, denunciar locais, divulgar eventos, etc. Isso indica que a geocodifcação tem muita importância e a qualidade dela traz impactos significativos no que está sendo produzido no laboratório. Para obter informações relacionadas a endereços utilizamos a geocodificação obtida a partir das APIs online de geocodifcação.

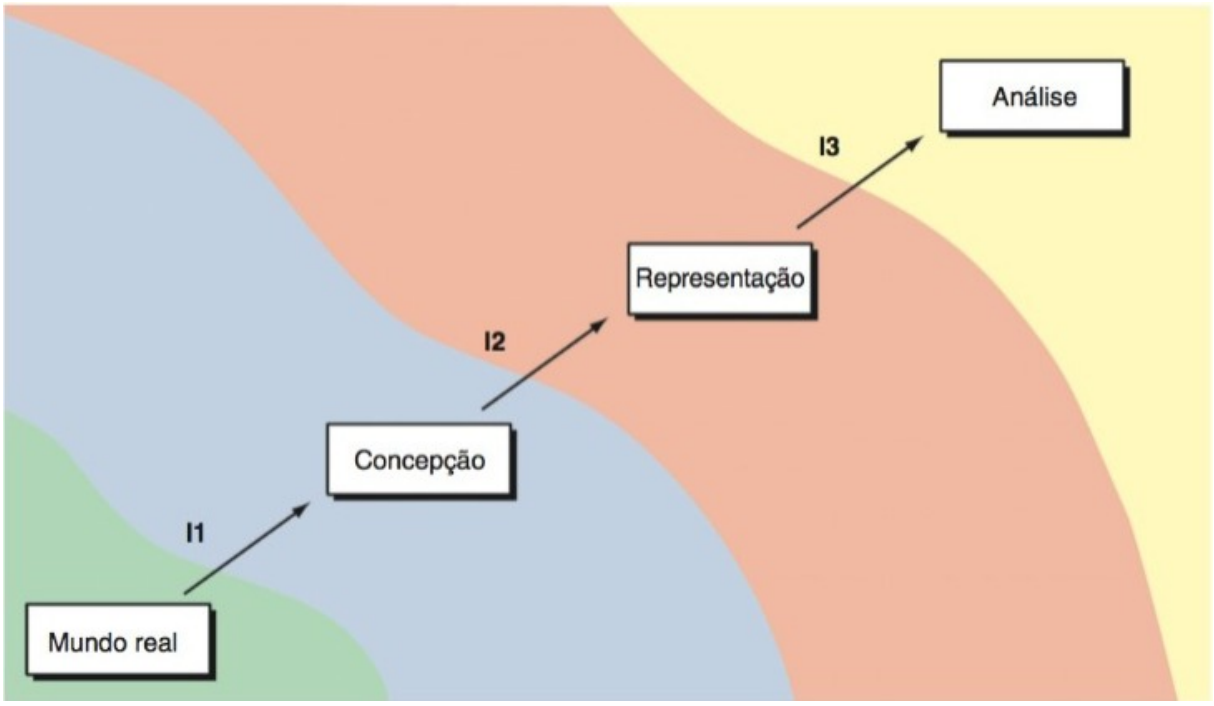

Figura 1.1 – Retirada do livro (**??**). Visão conceitual da incerteza, onde os filtros I1, I2, I3 distorcem a informação original

Por muitos anos, a principal forma de obter informações geográficas era por meio de um software SIG. Segundo (**??**) um Sistema de informação Geográfica (SIG) é um conjunto de ferramentas capaz de analizar e integrar dados geográficos, bem como possibilitar ao usuário acesso facilitado a dados, sem depender de ferramentas como o GPS. Para (**??**), apesar do SIG ter sido a ferramenta convencional por muitos anos, utiliza-lo para geocodifcação requer um profissional capacitado. A ferramenta demanda o pre-processamento dos dados, criação de um localizador de endereço, customização dos parâmetros, controle de qualidade e correção manual de qualquer falha. Todo esse processo é custoso para o usuário comum. Para ele, a geocodificação utilizando ferramentas online retira do usuário uma grande responsabilidade, a manuteção da base, diminuindo assim os processos para obter a informação e tornando o trabalho menos custoso.

Apesar de a geocodificação online ser mais simples de utilizar, para que o SIG seja substituído por ela, deve-se considerar sua qualidade em relação a qualidade do SIG. No artigo, (**??**) são avaliadas oito ferramentas de gecodificação, sendo duas delas SIGs e o restante, ferramentas da internet. As ferramentas utilizadas foram: SRI ArcGIS Address Locator, CoreLogic PxPoint, Google Maps API,Yahoo! PlaceFinder, Microsoft Bing, Geocoder.us, Texas A and M University Geocoder, and OpenStreetMap (OSM). Para encontrar o erro foi utilizada uma base referência com informações descritivas do endereço (rua, número, cidade, etc) e informações geográficas (latitude e longitude). Essa base é considerada referência pois os dados de latitude e longitude foram obtidos manualmente (GPS ou pesquisa manual). Chamaremos essa e outras bases referência de base padrão ouro. A base em questão possui 940 endereços do estado Texas dos Estados Unidos da America, sendo 78 destes da região de Central Texas, região considerada

importante para o autor. Ele então calcula o erro de cada endereço geocodificado como:

$$\epsilon_x = x_{\text{ref}}, x_{\text{geoc}} \tag{1.1}$$

$$\epsilon_y = y_{\text{ref}}, y_{\text{geoc}} \tag{1.2}$$

$$\varepsilon_{xy} = \sqrt{\varepsilon_x^2 + \varepsilon_y^2} \tag{1.3}$$

onde:

- e_x é o erro da longitude,
- e_y é o erro da latitude,
- e_xy é o erro

O estudo mostrou que não há diferença significativa entre as ferramentas online e os SIGs. Tanto os SIGs quanto as ferramentas online tiveram média e desvio padrão do erro similares. Além de taxa de resposta (quantos endereços tiveram resposta para a ferramenta utilizada) de 97,8% a 100%, consideradas suficientemente boas. Sendo assim, o estudo teve sucesso ao mostrar que as ferramentas online podem ser utilizadas como substitutivas aos SIGs.

Apesar de (**??**) ter apresentado resultados significativos, o estudo apresenta limitações. A principal dela é a quantidade dos dados utilizados para fazer essa avaliação e ter focado apenas em uma região (Texas, EUA). O presente trabalho pretende abordar essas limitações ao fazer a análise de outra região do mundo, tendo um enfoque no Brasil, e ampliar a quantidade de dados avaliados. Porém consideraremos apenas ferramentas de geocodificação online (GeoAPIs) por considerar que elas já estão consolidadas no mercado e na acadêmia.

Outro estudo importante é (**??**) que faz uma avaliação da qualidade da geocodificação do Google Maps API fornecida pelo Google Cloud Plataform (**??**). Nesse estudo, os autores utilizam uma base padrão ouro com os dados de Belo Horizonte, cidade de Minas Gerais, estado do Brasil para essa avaliação. A base conta com mais de 540 mil endereços da cidade e é mantida pela empresa de informática e informação do município de Belo Horizonte - Prodabel (**??**). A empresa atualiza os dados mensalmente e tem parceria com outras 26 empresas para manter a base o mais correta possível. Ela conta com informações descritivas, sociais e espaciais do endereço. Para medir o erro, foi calculada a distância euclidiana dos pontos geocodificados para os pontos originais. A partir do erro, o estudo faz análises espacias do erro e também relaciona a acurácia descrita pela API com o erro gerado. O estudo mostrou que o Google Maps API tem taxa de acerto de 74,7%, considerando que acertou se o erro for menor de 150 metros. Outra descoberta foi que o erro é menor nas áreas centrais da cidade, e maior na periferias. Os autores também tentaram fazer uma relação entre erro e renda, porém não foi possível vizualizar nenhuma relação direta.

Apesar de descobertas importantes, o estudo é limitado na medida que só analisa uma API de geocodificação. Além de analisar apenas uma cidade brasileira, o que impossibilita a generalização dos resultados. O trabalho pretende trabalhar nessas limitações fazendo a análise de uma amostra da mesma base de dados, porém com outras APIs de geodificação. Além disso, iremos fazer uma análise com uma base da região metropolitana de São Paulo. O que traz uma diversidade para nosso estudo.

1.3 Objetivos

O principal objetivo deste trabalho é avaliar o erro, a discrepância e a acurácia de cinco APIs utilizadas no laboratório de pesquisa e capacitação em desenvolvimento de software - TerraLAB. As APIs em análise são: Google Maps, TomTom, Open Route Service (ORS), Mapbox e Here. O erro será analisado quanto às respostas fornecidas pelas APIs diferirem do esperado. A discrepância medirá o nível de discordância entre as APIs. Por fim, a acurácia será utilizada para verificar a precisão das respostas fornecidas pelas APIs.

Uma parte essencial do trabalho é compreender os pontos onde essas APIs apresentam falhas, e, portanto, a análise espacial dessas medidas terá grande destaque na pesquisa.

Com isso, gostaríamos de responder as seguintes perguntas:

- Qual API das utilizadas erra mais?
- Existe algum padrão espacial no erro?
- Alguma medida de variância entre as APIs (discrepância) representa o erro?

Para chegar a essas respostas temos alguns objetivos específicos que devem ser atendidos:

- Coletar bases de dados padrão ouro;
- Calcular as medidas para avaliação;
- Avaliar as distribuição das medidas;
- Correlacionar as medidas;
- Avaliar de que forma o espaço se relaciona com essas medidas.

1.4 Organização do Trabalho

Um parágrafo fazendo uma descrição dos capítulos restantes do documento.

1.4.1 Estrutura da Monografia

Segue uma **sugestão** para a estrutura da monografia:

Capítulo 1: Introdução.

Capítulo ??: Revisão Bibliográfica/ Embasamento Teórico (com o referencial teórico e trabalhos relacionados).

Capítulo ??: Metodologia ou Desenvolvimento (material e métodos).

Capítulo ??: Resultados e Discussões.

Capítulo ??: Conclusão (e trabalhos futuros).

2 Avaliação da Geocodificação

2.1 Geocodificação

2.2 APIs de Geocodificação

3 Bases de Dados e Métodos de Geocodificação e Avaliação

Para avaliar a qualidade das APIs de geocodificação utilizadas no TerraLAB duas bases de dados padrão ouro foram usadas como referência. Chamaremos essas bases de Bases Gold. Com as bases, foi obtida a medida de erro e realizadas métricas diversas utilizando essa medida.

3.1 Bases de Dados

Foram coletadas duas bases de dados distintas para o presente trabalho.

A primeira base coletada foi a base do Centro de Estudos da Metrópole (CEM). A base consiste 12.500 endereços de escolas públicas e particulares do ensino básico da região metropolitana de São Paulo. Essa base foi coletada de forma manual pelo CEM utilizando o GPS para a coleta das coordenadas. Além de informações sobre o endereço, a base também conta com informações diversas sobre as escolas, permitindo com que se façam avaliações diversas em relação a esses dados. O CEM também disponibilizou um mapa de cluster, com todas as escolas, permitindo uma melhor vizualização da localização de cada uma delas e da densidade das escolas em São Paulo e região.

A segunda base coletada foi a base de dados da Prodabel, empresa de informática e informação da prefeitura de Belo Horizonte. A base de dados foi descoberta por meio da referência

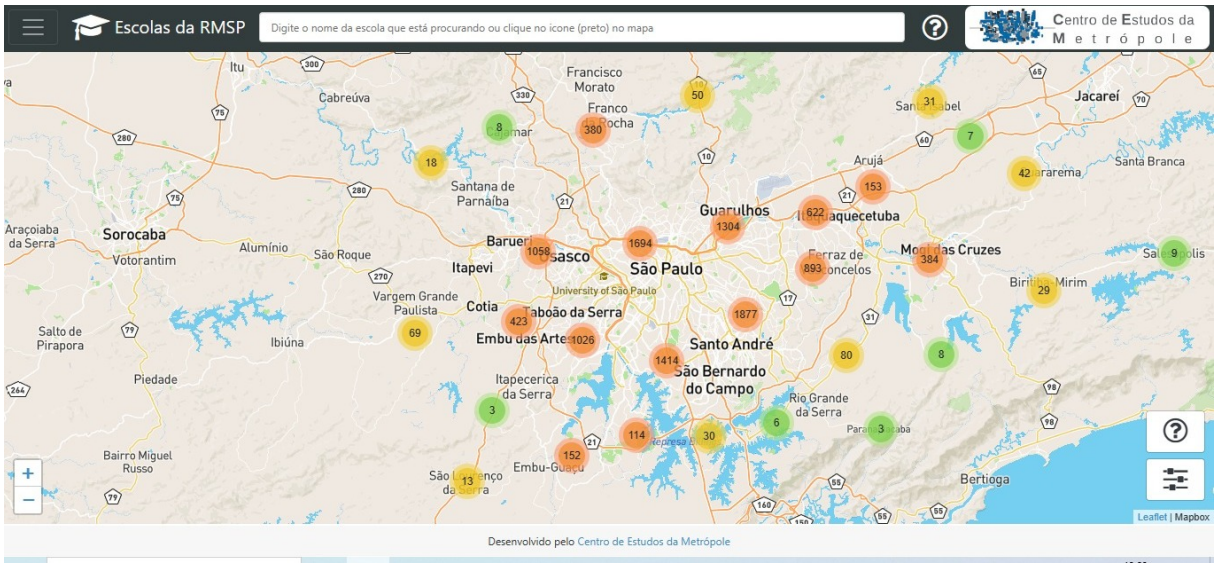

Figura 3.1 – Mapa de clusters que mostra a quantidade de escolas em cada região. Ao aproximar o mapa, o usuário consegue ver a localização de cada uma das escolas presentes no banco de dados.

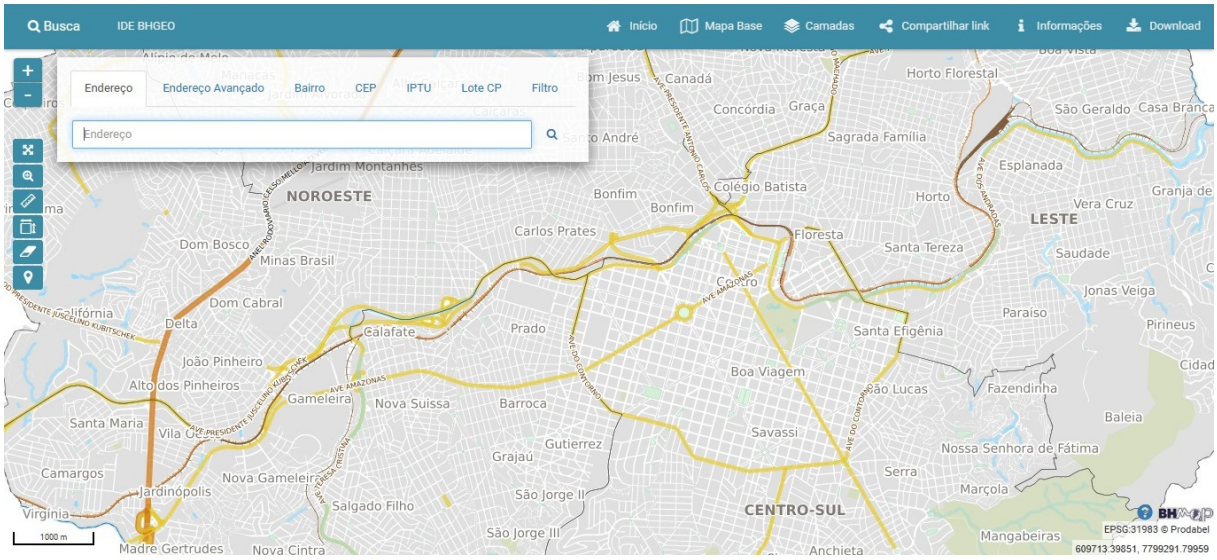

Figura 3.2 – Mapa que mostra a cidade de Belo Horizonte, desenvolvido pela Prodabel. Na barra de pesquisa, é possível pesquisar os endereços e marcá-los no mapa.

1. É uma base de dados mantida e atualizada mensalmente por 27 empresas públicas e privadas de Belo Horizonte. As empresas têm a responsabilidade de reportar qualquer inconsistência que encontrarem, bem como fornecer novos dados a medida que são adquiridos por ela. É uma base considerada confiável pois é constantemente atualizada e é utilizada por diversos serviços da prefeitura. Um exemplo de serviço que utiliza a base de dados é a distribuição dos alunos da rede pública por meio de georeferenciamento. A base conta com 740.000 endereços na data de coleta. A prefeitura também disponibiliza site com um mapa para vizualização do endereços registrados. O endereço está posicionado em cima do edifício representado. Isso pode gerar erro de alguns metros devido a maioria da APIs colocar o endereço na frente do edifício representado.

3.2 Processo de Geocodificação

A preparação de dados e geocodificação desempenham um papel crucial em muitos estudos e projetos que envolvem informações geográficas. Nesta pesquisa, esses processos desempenham um papel fundamental na obtenção de dados consistentes e na atribuição de coordenadas geográficas aos endereços. A etapa de preparação de dados envolve a seleção dos campos relevantes da base de dados, como o nome da rua, número, bairro, CEP e cidade. Além disso, é realizada uma homogeneização dos dados, onde abreviações comumente utilizadas são substituídas por suas formas completas correspondentes. Essa etapa é essencial para garantir resultados mais precisos na geocodificação. A geocodificação, por sua vez, consiste em atribuir coordenadas geográficas (latitude e longitude) a cada endereço presente na base de dados. Utilizando ferramentas adequadas, o processo de geocodificação é realizado, possibilitando a localização precisa de cada endereço no espaço geográfico. Para realizar a geocodificação, os endereços previamente preparados são inseridos no banco de dados do Crawler, onde as ferramentas de geocodificação estão

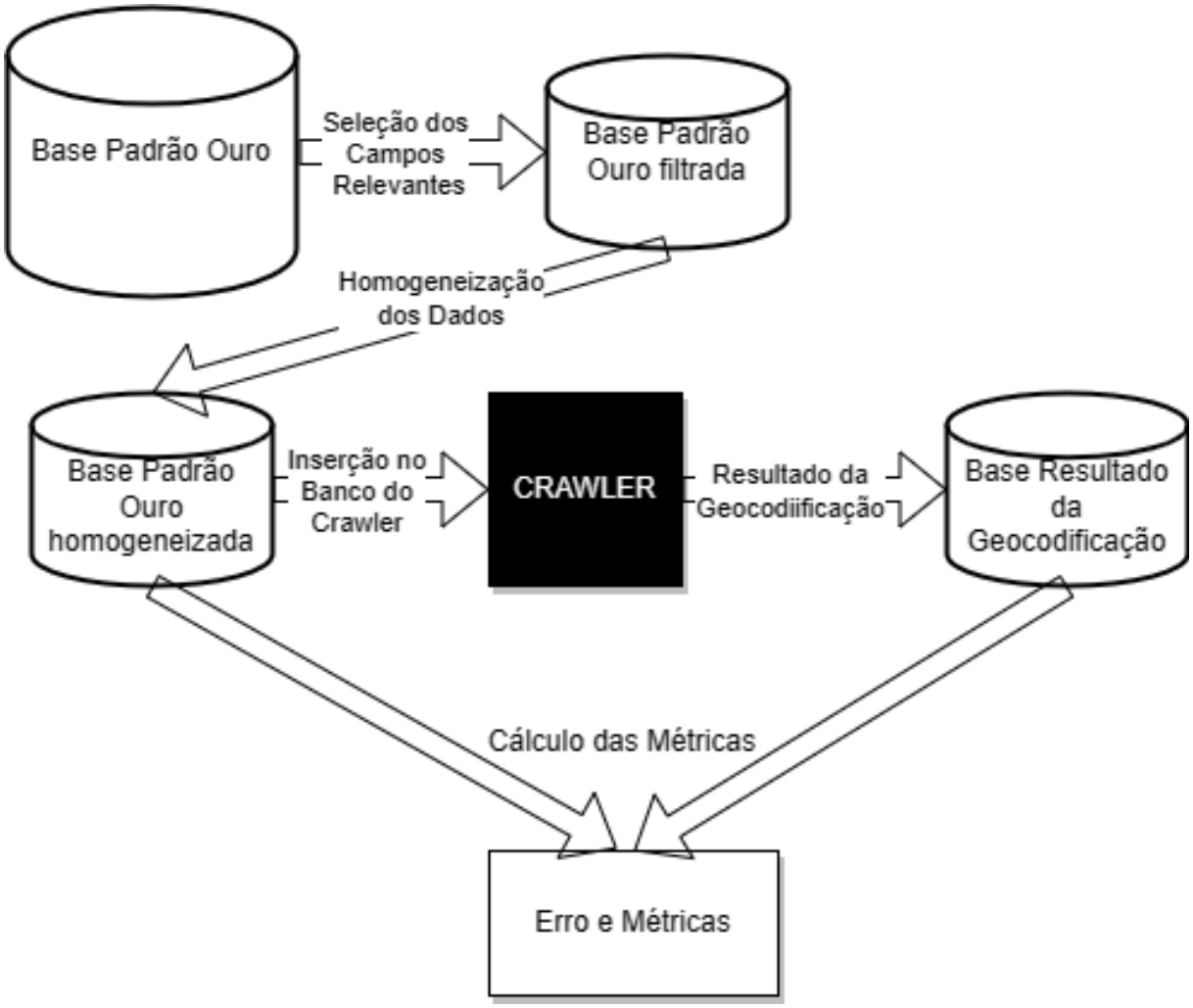

Figura 3.3 – Esquematização do processo de preparação e geocodificação dos dados

disponíveis. Essas ferramentas utilizam algoritmos e informações geográficas para identificar e atribuir as coordenadas geográficas correspondentes a cada endereço. É importante ressaltar que o processo de geocodificação é realizado pela equipe de Back-end do TerraLAB, portanto, vemos esse processo como uma caixa preta. Uma vez concluída a geocodificação, os endereços geocodificados, juntamente com suas coordenadas geográficas, são armazenados no banco de dados. Esses dados geocodificados podem ser utilizados para análises espaciais, mapeamento e visualização de informações geográficas, contribuindo para a compreensão de padrões e tendências em determinada área de estudo. Portanto, a preparação de dados e geocodificação são etapas essenciais para garantir a qualidade e a utilidade das informações geográficas utilizadas neste estudo. Esses processos permitem a obtenção de dados consistentes e georreferenciados, facilitando a análise e interpretação dos resultados obtidos

3.3 Método de Avaliação

3.3.1 Erro, Acurácia e Discrepância

A principal métrica utilizada para avaliar a qualidade da geocodificação é o erro do endereço. Esse erro é calculado como a distância entre o ponto de referência e o ponto geocodificado pela GeoAPI. Com base nesse erro, calcularemos medidas estatísticas, como a média, a mediana, o desvio padrão e a média aparada em 5%, para analisar a precisão das GeoAPIs.

$$e = D(p_{\text{Gold}}, p_{\text{Geo}}) \tag{3.1}$$

onde:

- e é o erro da geocodificação,
- D é uma função que calcula a distância em km,
- p_{Gold} é o ponto da base Gold, e
- p_{Geo} é o ponto resultante da geocodificação.

Outra métrica utilizada é a taxa de resposta por API. Para alguns endereços da base de dados, as GeoAPIs podem retornar um erro, não fornecendo uma geocodificação válida. Nesse caso, nada é inserido no banco de dados. A taxa de resposta é calculada como a quantidade de endereços geocodificados dividida pela quantidade de endereços originais na base de dados. Esse valor, normalmente entre 0 e 1, é convertido em uma porcentagem para facilitar a compreensão dos resultados.

4 Resultados

Neste capítulo são apresentados, interpretados e analisados todos os resultados alcançados no trabalho. A análise deve ser realizada de forma que fique claro que os objetivos específicos foram atendidos. Se possível, faça uma comparação com os resultados da literatura, destacando a importância da pesquisa realizada no contexto acadêmico.

5 Considerações Finais

Neste capítulo deve ser explicitado se todos os objetivos descritos na introdução foram atingidos e ressaltar a contribuição do trabalho para o meio acadêmico.

São apresentados de forma sucinta os resultados obtidos e um fechamento de todo trabalho desenvolvido.

5.1 Conclusão

Em resumo, nesta seção devem ser apresentadas as considerações finais do trabalho. Faça uma recapitulação a respeito de cada um dos objetivos específicos, sintetize os resultados obtidos e conclua se o objetivo principal do trabalho foi alcançado.

5.2 Trabalhos Futuros

Apresente propostas de continuidade do seu trabalho.

5.3 Publicações Realizadas

Caso o trabalho tenha originado publicações é válido acrescentar essa informação, visto que pode creditar ainda mais o estudo. Assim, elas devem ser apresentadas na forma de uma subseção do capítulo conclusão. Por exemplo:

Os trabalhos seguintes, que foram originados das metodologias propostas, foram aceitos para apresentação em conferências nacionais:

1. Autor. Título do Artigo. Cidade: Conferência, Ano.

Referências

CHOW, T. E.; DEDE-BAMFO, N.; DAHAL, K. R. Geographic disparity of positional errors and matching rate of residential addresses among geocoding solutions. *Annals of GIS*, Taylor and Francis, v. 22, n. 1, p. 29–42, 2016. Disponível em: <https://doi.org/10.1080/19475683.2015.1085437>.

CHOW, T. E.; LIN, Y.; CHAN, W.-y. D. The development of a web-based demographic data extraction tool for population monitoring. *Transactions in GIS*, v. 15, n. 4, p. 479–494, 2011. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9671.2011.01274.x>.

GILBOA, S. M.; MENDOLA, P.; OLSHAN, A. F.; HARNESS, C.; LOOMIS, D.; LANGLOIS, P. H.; SAVITZ, D. A.; HERRING, A. H. Comparison of residential geocoding methods in population-based study of air quality and birth defects. *Environmental Research*, v. 101, n. 2, p. 256–262, 2006. ISSN 0013-9351. Women's Occupational and Environmental Health. Disponível em: <https://www.sciencedirect.com/science/article/pii/S001393510600020X>.

HAY, G.; KYPRI, K.; WHIGHAM, P.; LANGLEY, J. Potential biases due to geocoding error in spatial analyses of official data. *Health and Place*, v. 15, n. 2, p. 562–567, 2009. ISSN 1353-8292. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1353829208001081>.

JR., C. A. D.; ALENCAR, R. O. de. Evaluation of the quality of an online geocoding resource in the context of a large brazilian city. *Transactions in GIS*, v. 15, n. 6, p. 851–868, 2011. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9671.2011.01288.x>.

KRIEGER, N.; WATERMAN, P.; LEMIEUX, K.; ZIERLER, S.; HOGAN, J. W. On the wrong side of the tracts? evaluating the accuracy of geocoding in public health research. *American Journal of Public Health*, v. 91, n. 7, p. 1114–1116, 2001. PMID: 11441740. Disponível em: <https://doi.org/10.2105/AJPH.91.7.1114>.

LONGLEY, P. A.; GOODCHILD, M. F.; MAGUIRE, D. J.; RHIND, D. W. *Sistemas e Ciencia da Informaçao Geografica*. Grupo A, 2013. ISBN 9788565837651. Disponível em: <https://integrada.minhabiblioteca.com.br/#/books/9788565837651/>.

MAZUMDAR, S.; RUSHTON, G.; SMITH, B. J. et al. Geocoding accuracy and the recovery of relationships between environmental exposures and health. *International Journal of Health Geographics*, v. 7, n. 1, p. 13, 2008. Disponível em: <https://ij-healthgeographics.biomedcentral.com/articles/10.1186/1476-072X-7-13>.

OLLIGSCHLAEGER, A. M. Artificial neural networks and crime mapping. In: WEISBURD, D.; MCEWEN, T. (Ed.). *Crime Mapping and Crime Prevention*. Monsey, NY: Criminal Justice Press, 1998, (Crime Prevention Studies, v. 8). p. 313–347.

STEIN, R. T.; SANTOS, F. M. d.; REX, F. E. et al. *Geoprocessamento*. Grupo A, 2021. E-book. Disponível em: <https://integrada.minhabiblioteca.com.br/#/books/9786556902852/>.