Avaliação da Qualidade de Serviços de Geocodificação Online em Cidades Brasileiras

Orientador: Prof. Rodrigo Cesar Pedrosa Silva
Bolsista: Ana Luiza Almeida Soares

5 de julho de 2023

Resumo

Este projeto de pesquisa concentra-se em estimar a qualidade dos serviços de geocodificação online adaptados para cidades brasileiras. A geocodificação, o processo de transformar endereços em coordenadas geográficas, é crucial para uma ampla gama de aplicações em transporte, planejamento urbano, serviços de emergência e operações comerciais. No entanto, a qualidade dos resultados de geocodificação varia entre diferentes provedores de serviço e regiões, tornando necessário avaliar e comparar esses serviços.

O projeto tem como objetivo avaliar a precisão, de diferentes serviços de geocodificação online na correta geocodificação de endereços em cidades brasileiras. Será coletado um conjunto diversificado de dados que incluirá endereços urbanos e rurais com diferentes complexidades. Coordenadas de referência obtidas de fontes confiáveis serão usadas para validar a precisão dos serviços de geocodificação. A pesquisa estabelecerá métricas quantitativas para avaliar e comparar o desempenho dos serviços de geocodificação.

Os resultados deste projeto de pesquisa terão implicações significativas para análise espacial, processos de tomada de decisão, serviços baseados em localização, desenvolvimento de infraestrutura e operações comerciais em cidades brasileiras. Os resultados permitirão que pesquisadores, formuladores de políticas e profissionais tomem decisões informadas sobre a seleção e utilização dos serviços de geocodificação. Além disso, o projeto preenche a lacuna de pesquisa específica para o contexto brasileiro e contribui para o conhecimento existente em pesquisa de geocodificação. Em última análise, este projeto de pesquisa visa aprimorar a precisão, eficiência e confiabilidade dos serviços de geocodificação para cidades brasileiras, beneficiando diversos setores e partes interessadas envolvidas em aplicações de geocodificação.

Palavras-chave

Geocodificação, Análise espacial, Serviços de geocodificação online

1 Introdução

1.1 O problema

Ao representar locais reais em um computador, é necessário realizar a tradução do endereço (uma sequência de palavras e números) para coordenadas no espaço (latitude e longitude). Esse processo é conhecido como geocodificação. Atualmente, a geocodificação utilizando GeoAPIs online é uma abordagem comum.

No entanto, as GeoAPIs utilizam diferentes métodos para identificar as coordenadas espaciais, e não há garantia de que todas elas sejam precisas. Portanto, é essencial medir a precisão dessas GeoAPIs, a fim de avaliar a qualidade da geocodificação fornecida por cada uma delas.

1.2 A importância

A geocodificação de endereços tem aplicações em diversas áreas, tanto dentro quanto fora da computação. A capacidade de mapear um endereço no espaço é crucial para tomadas de decisão, otimização de rotas, planejamento de rotas de fuga, entre outros. É fundamental ter confiança de que os endereços geocodificados realmente representam o local desejado no espaço ou entender quando podem não ser precisos. Isso ajuda a melhorar a confiabilidade do serviço e permite uma tomada de decisão consciente dos riscos envolvidos.

2 Objetivos

O objetivo deste estudo é avaliar a qualidade da geocodificação fornecida por cinco GeoAPIs utilizadas no laboratório TerraLAB. As GeoAPIs selecionadas para análise são: TomTom, Mapbox, Open Route Service, Google Maps e Here.

3 Justificativa/Relevância

As justificativas para este projeto podem ser colocadas nos seguintes termos:

1. Atendendo a uma Necessidade Crítica:

Serviços de geocodificação precisos e confiáveis são essenciais para uma ampla gama de aplicações em diversos setores, incluindo transporte, planejamento urbano, serviços de emergência e operações comerciais [4, 1]. No contexto das cidades brasileiras, que possuem um sistema de endereçamento complexo e características geográficas diversas, a necessidade de soluções robustas de geocodificação torna-se ainda mais crítica [5, 2]. Este projeto de pesquisa aborda essa necessidade avaliando e comparando a qualidade de diferentes serviços de geocodificação aplicados à cidades brasileiras.

2. Aprimorando a Análise Espacial e a Tomada de Decisão:

Erros de geocodificação podem ter consequências significativas nos processos de análise espacial e tomada de decisão [1]. Resultados de geocodificação imprecisos ou inexatos podem levar a análises falhas, decisões equivocadas e

alocação ineficiente de recursos. Ao estimar a qualidade dos serviços de geocodificação online, este projeto de pesquisa fornece insights valiosos a pesquisadores, formuladores de políticas e profissionais, permitindo que tomem escolhas mais informadas ao utilizar serviços de geocodificação para análises espaciais, planejamento urbano e processos de tomada de decisão.

3. Apoio ao Desenvolvimento de Infraestrutura:

A geocodificação precisa é crucial para iniciativas de desenvolvimento de infraestrutura, incluindo redes de transporte, serviços de utilidade pública e sistemas de resposta a emergências. Ao avaliar a precisão dos serviços de geocodificação, este projeto de pesquisa fornece insights valiosos para planejadores e formuladores de políticas de infraestrutura [3]. Os resultados podem contribuir para o desenvolvimento de sistemas de endereçamento robustos, algoritmos de roteamento eficientes e aprimoramento das capacidades de resposta a emergências, aprimorando assim o desenvolvimento geral da infraestrutura nas cidades brasileiras.

4. Preenchendo a Lacuna na Pesquisa:

Embora a pesquisa em geocodificação tenha sido amplamente estudada em vários contextos, há uma falta de avaliações recentes específicas para as cidades brasileiras. Os trabalhos mais atuais são de 2011 e 2012 [2, 5]. Este projeto de pesquisa preenche a lacuna na pesquisa ao se concentrar nos desafios únicos e nos requisitos da geocodificação no contexto brasileiro. Os resultados contribuirão para o conhecimento existente na pesquisa de geocodificação, enriquecendo a compreensão da qualidade e desempenho da geocodificação em sistemas geográficos e de endereçamento diversos.

4 Metodologia

4.1 Base de Endereços

Para avaliar a qualidade das GeoAPIs, utilizaremos uma base de dados confiável que contém endereços completos, tanto verbalmente quanto com suas coordenadas geográficas. Essa base de dados será usada como referência para comparação. Dentro da área, uma base com essas características é chamada de Base Padrão ou Gold Standard Dataset. Sendo assim, iremos nos referir como Base Gold ou Pontos Gold, a base utilizada como referência e os pontos vindos dela.

A escolha recaiu sobre a base de dados coletada pelo Centro de Estudos da Metrópole (CEM), que inclui mais de 12 mil endereços de escolas na região metropolitana de São Paulo. Essa base foi coletada manualmente por membros do grupo, que visitaram pessoalmente as escolas e utilizaram GPS para obter as coordenadas.

4.2 Motivo de Escolha da Base de Endereços

A forma como a base de dados foi coletada é a mais precisa que temos atualmente para obter a localização espacial real de cada endereço. Embora possa

haver algum erro devido à precisão do GPS, não existe uma técnica comprovadamente mais eficaz para identificar dados geográficos de um endereço do que a coleta manual. Portanto, utilizaremos essa base como referência, considerando-a o ponto de referência mais próximo da verdadeira localização terrestre. Ao considerar essa base como a correta, poderemos quantificar o erro e calcular métricas com base nisso.

4.3 Processo de Geocodificação

Seleção dos campos relevantes da base de dados: A partir da base de dados selecionada, foram identificados os campos que contêm informações sobre o nome da rua, número, bairro, CEP e cidade.

Homogeneização dos dados: Os dados obtidos passaram por um processo de homogeneização, no qual foram removidas abreviações comumente utilizadas. Para garantir melhores respostas das GeoAPIs, a equipe decidiu substituir abreviações por suas versões completas. Foi criado um dicionário de abreviações contendo as palavras mais comuns em português e suas formas completas correspondentes. O algoritmo percorre a base de dados e substitui as abreviações encontradas pelas formas completas correspondentes.

Inserção dos endereços no banco de dados do Crawler: Os endereços, já homogeneizados, são inseridos no banco de dados do Crawler, que foi desenvolvido pela equipe de Back-end do TerraLAB.

Realização da geocodificação: O banco de dados do Crawler é utilizado para realizar a geocodificação dos endereços. A geocodificação consiste em atribuir coordenadas geográficas (latitude e longitude) a cada endereço.O processo de geocodificação é executado utilizando as ferramentas disponíveis no Crawler.

Armazenamento dos endereços geocodificados: Os endereços geocodificados são salvos no banco de dados, juntamente com suas coordenadas geográficas.

4.4 Metrícas de Avaliação

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 gold e o ponto geocodificado pela GeoAPI. Como os pontos estão representados com latitude e longitude, foi utilizada a função distance da biblioteca geopy de python.

$$e = D(p_{\text{Gold}}, p_{\text{Geo}}) \tag{1}$$

onde:

- e é o erro da geocodificação,

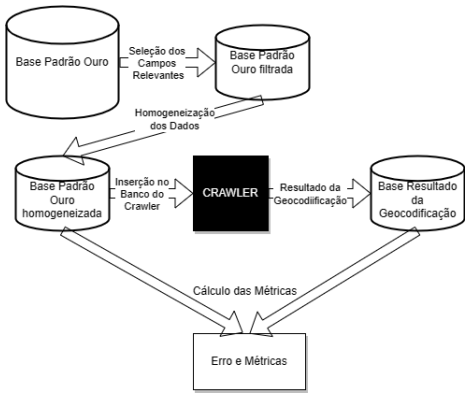


Figura 1: Diagrama que esquematiza os processos necessários para a pesquisa

- 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.

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.

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.

$$\text{tx}_{\text{Resp}} = \frac{n_{\text{Geo}} \times 100}{n_{\text{Gold}}} \tag{2}$$

onde:

- T_{Resp} é a taxa de resposta da API,
- n_{Geo} é a quantidade de endereços geocodificados em determinada API, e
- n_{Gold} é a quantidade de endereços da base gold que foram inseridos para geocodificação.

5 Resultados Preliminares

5.1 Distribuição Espacial dos Pontos Geocodificados

Para melhor visualização dos pontos geocodificados em comparação com o banco de referência, foram gerados mapas com a identificação dos pontos para cada uma das APIs. Seguem os mapas.

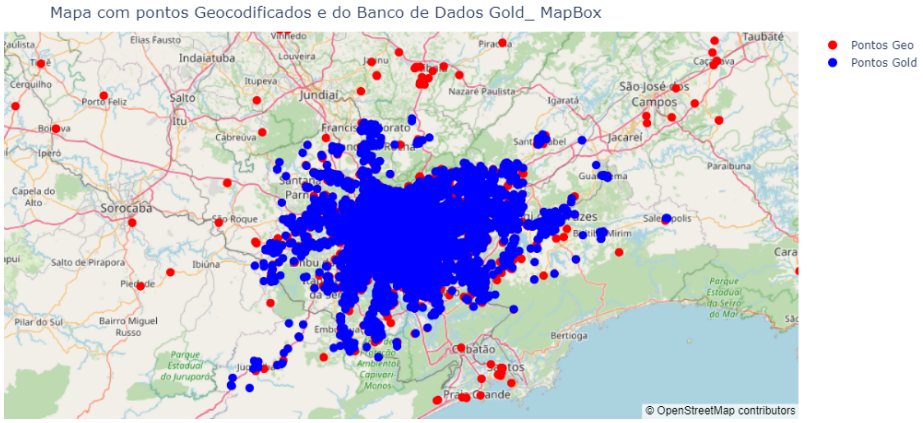


Figura 2: Mapa da Distribuição Espacial dos Pontos da base Gold e Geocodificados pela Mapbox

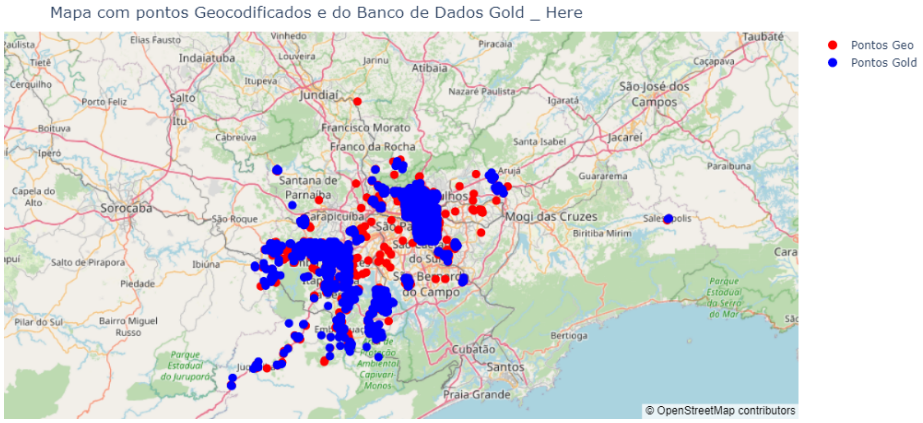


Figura 3: Mapa da Distribuição Espacial dos Pontos da base Gold e Geocodificados pela Here

Não é possível tirar muitas conclusões definitivas apenas com essa visualização, no entanto, é possível observar a densidade dos pontos e identificar quais GeoAPIs processaram a maior quantidade de pontos. Além disso, pode-se notar

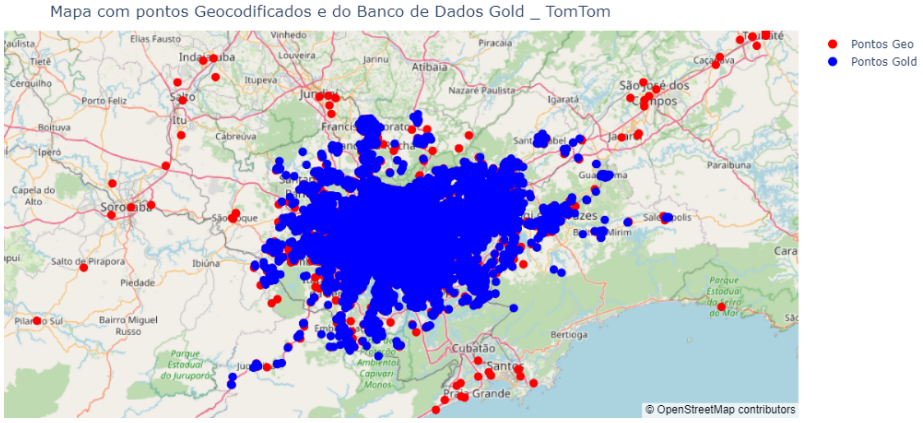


Figura 4: Mapa da Distribuição Espacial dos Pontos da base Gold e Geocodificados pela TomTom

Tabela 1: Métricas de Erro e Resposta

API	Média (km)	Mediana (km)	Desvio Padrão (km)	Média Aparada (km) (km)	Taxa de Resposta (%)	Taxa de Acerto(%)
Mapbox	9.7544	0.1084	46.7664	1.8349	53.3829	30.1903
Tomtom	5.0701	0.0560	35.6215	0.2373	83.1894	9.2051
Here	2.2372	0.0632	13.7984	0.4365	13.9075	9.2051

que os pontos classificados como "Gold"estão concentrados na região metropolitana de São Paulo, enquanto alguns pontos geocodificados estão localizados fora dessa região, em outras cidades do estado. Essa disparidade provavelmente reflete alguns erros de geocodificação, conhecidos como outliers.

5.2 Metrícas do Erro

Após essa análise, o erro foi calculado para cada um dos pontos, sendo expresso em quilômetros.

Com essa informação, foram calculadas as métricas mencionadas anteriormente. É notável que a API TomTom obteve a maior taxa de resposta, com um índice superior a 80%. Além disso, apresentou a melhor média aparada em 5%. Por outro lado, a API Here obteve um desempenho superior na média tradicional, apesar de possuir uma taxa de resposta muito baixa. De forma geral, esses resultados foram considerados insatisfatórios. Ao longo do relatório, iremos analisar outras questões em detalhes.

5.3 Distribuição do Erro

Em seguida, tentamos analisar a distribuição do erro para cada uma das GeoAPIs. Para isso, utilizamos histogramas de erro individualmente para cada API e combinando todas elas. No entanto, devido à presença de alguns erros exorbitantes, esses histogramas não eram muito representativos, pois a maior parte do erro se concentrava entre 0 km e 50 km. Diante disso, decidimos realizar um corte nos dados, limitando o erro em 0.5 km ou 500 metros. Em seguida, repetimos o processo, agora gerando um único histograma que representa a distribuição do erro para todas as APIs em conjunto.

Com base nos histogramas, pudemos observar que a maioria dos erros está concentrada entre 0 e 200 metros. Levando em consideração nossas pesquisas, consideramos um erro aceitável de até 150 metros, o que corresponde aproximadamente a um quarteirão. Portanto, com base apenas na análise da distribuição de erros, todas as APIs apresentam resultados satisfatórios.

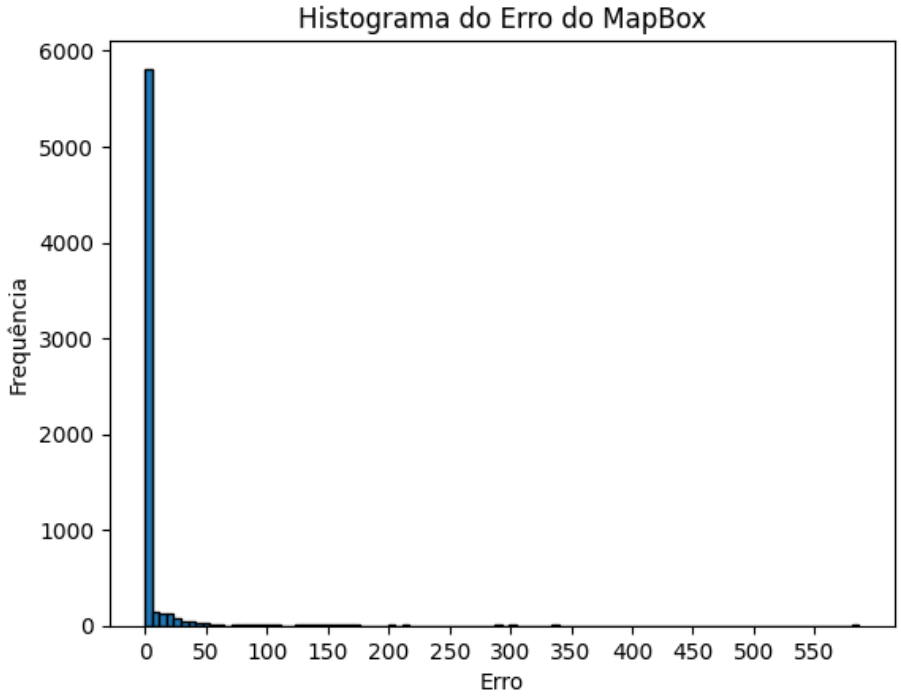


Figura 5: Histograma do erro calculado com os pontos da Mapbox

5.4 Distribuição Espacial do Erro

Além disso, realizamos uma análise adicional para visualizar como esse erro se comporta no espaço. Para isso, criamos mapas de altitude, onde o erro foi

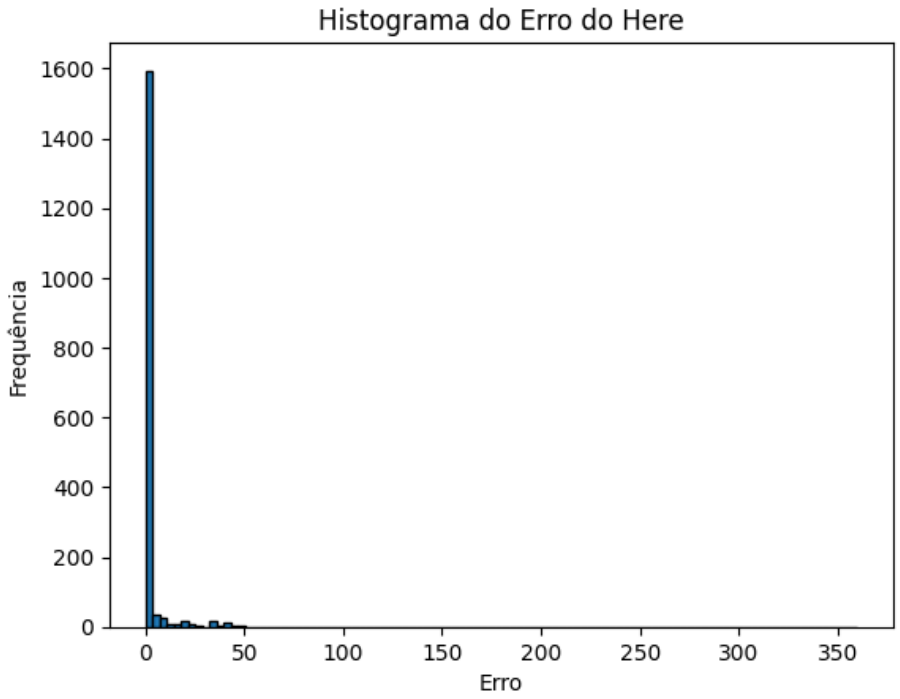


Figura 6: Histograma do erro calculado com os pontos da Here

utilizado como medida de altitude. Nessa representação, cores mais próximas do vermelho indicam erros mais altos, enquanto cores mais próximas do azul escuro indicam erros mais baixos. Também plotamos os pontos geocodificados no mapa para avaliar a representatividade das cores. Dessa forma, pudemos verificar se uma determinada área apresenta muitos pontos geocodificados ou se há poucos pontos com erros grandes.

Ao analisar os resultados, observamos que a maioria do mapa apresenta erros menores que 34 km, conforme esperado. No entanto, identificamos alguns pontos com erros grandes, que serão avaliados individualmente posteriormente. É importante ressaltar que encontramos uma limitação devido à presença de erros exorbitantes, ou outliers, o que restringe nossa capacidade de tirar conclusões significativas. Para obter uma melhor compreensão do contraste e da distribuição geográfica do erro, planejamos repetir o experimento realizando um corte em 34 km.

É válido destacar que o mapa é interativo no projeto original, permitindo uma visualização mais detalhada das informações apresentadas.

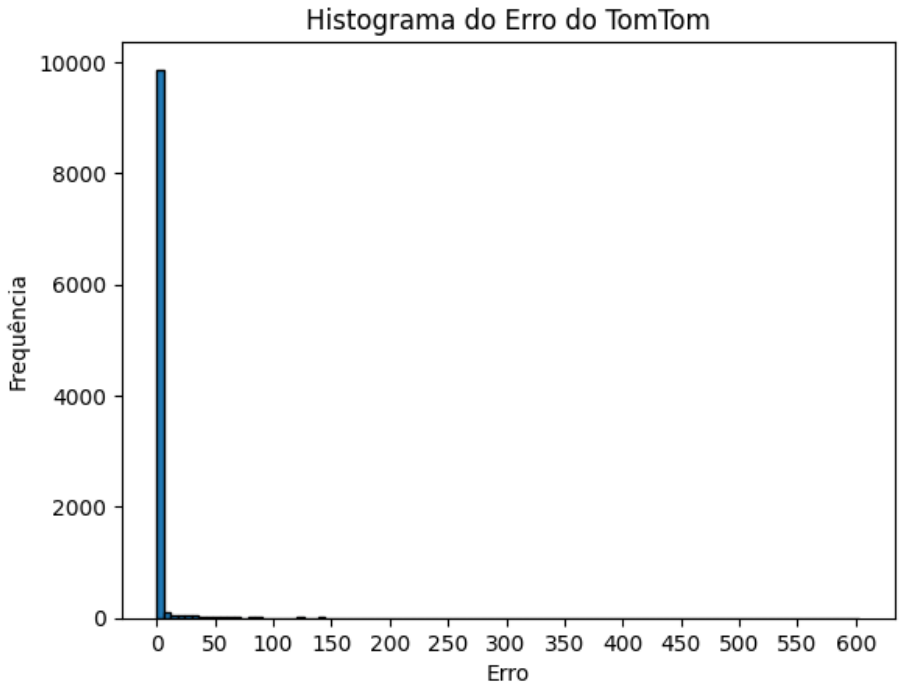


Figura 7: Histograma do erro calculado com os pontos da TomTom

Referências

[1] Taísa Rodrigues Cortes, Ismael Henrique da Silveira, and Washington Leite Junger. Improving geocoding matching rates of structured addresses in rio de janeiro, brazil. *Cadernos de Saúde Pública*, 37(7), 2021.

[2] Clodoveu A. Davis Jr. and Renato O. de Alencar. Evaluation of the quality of an online geocoding resource in the context of a large brazilian city. *Transactions in GIS*, 15(5):851–868, 2011.

[3] Carlos José de Armas García and Andrei Abel Cruz Gutiérrez. Deployment of a national geocoding service: Cuban experience. *Journal of the Urban & Regional Information Systems Association*, 25(1), 2013.

[4] Batuhan Kilic and Fatih Gülgen. Accuracy and similarity aspects in online geocoding services: A comparative evaluation for google and bing maps. *International Journal of Engineering and Geosciences*, 5(2):109–119, 2020.

[5] Douglas Martins, Clodoveu A. Davis Jr., and Frederico T. Fonseca. Geocodificação de endereços urbanos com indicação de qualidade. In *Proceedings XIII GEOINFO*, pages 36–41, Campos do Jordão, Brazil, November 2012. Universidade Federal de Minas Gerais.

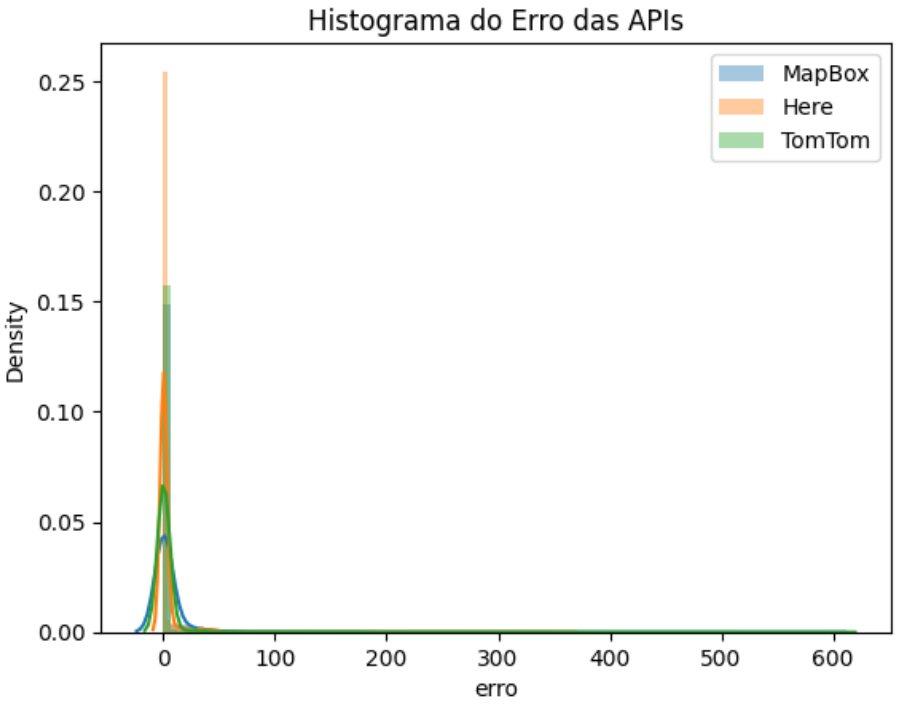


Figura 8: Histograma comparativo do erro das APIs

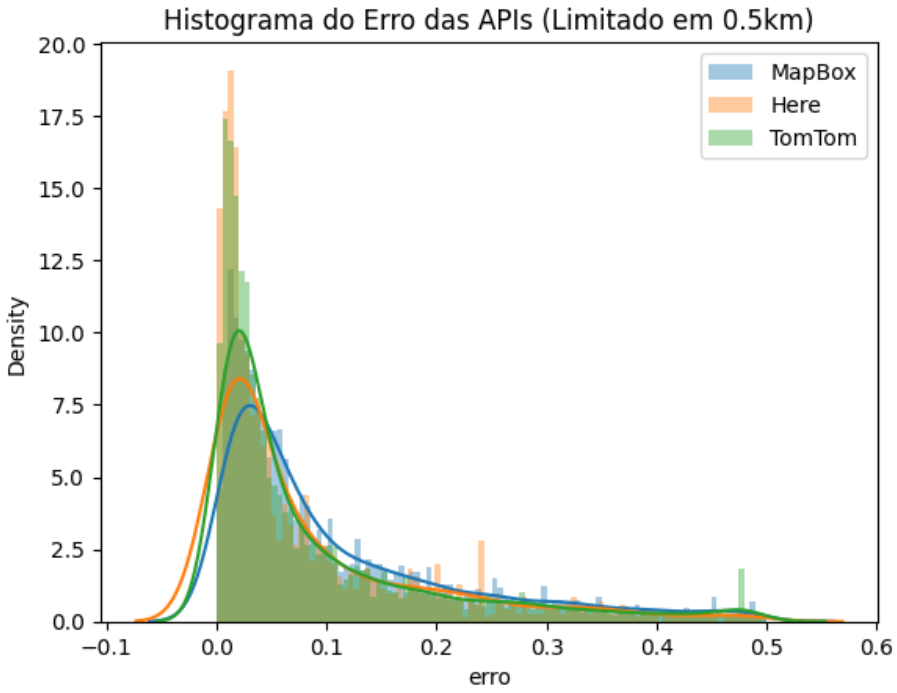
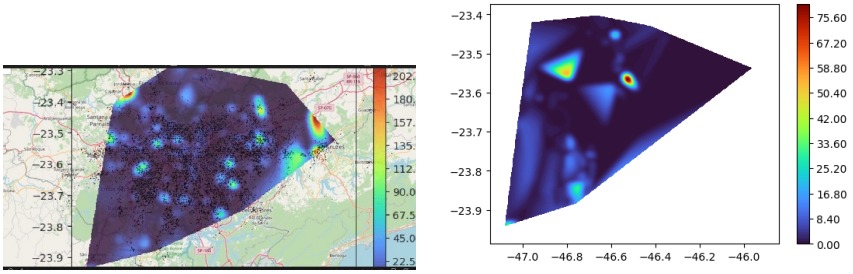
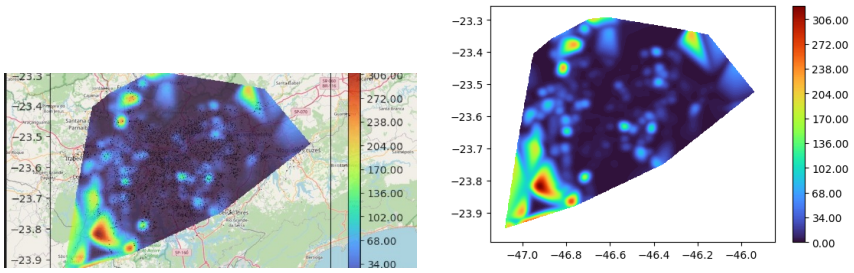


Figura 9: Histograma comparativo do erro das APIs com limitação em 500 metros



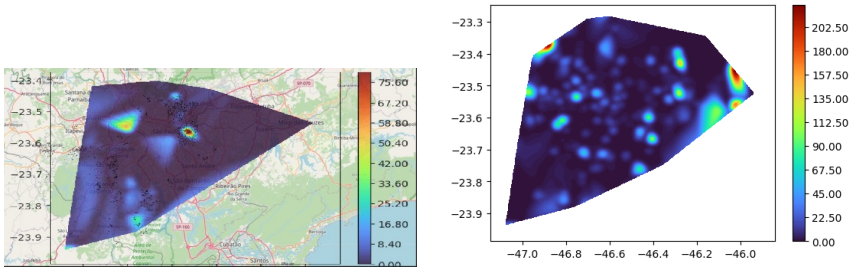
(a) Mapa (b) Gráfico de altitude

Figura 10: Mapa e gráfico de altitude, considerando o erro da Here como a medida de altitude



(a) Mapa (b) Gráfico de altitude

Figura 11: Mapa e gráfico de altitude, considerando o erro da Mapbox como a medida de altitude



(a) Mapa (b) Gráfico de altitude

Figura 12: Mapa e gráfico de altitude, considerando o erro da TomTom como a medida de altitude