

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

As APIs de geocodificação online desempenham um papel significativo em aplicações que requerem informações de localização. Para garantir a qualidade dessas aplicações, é essencial avaliar a precisão das APIs utilizadas. Este estudo tem como objetivo avaliar a qualidade de cinco APIs de geocodificação implementadas no TerraLAB: Google Maps, Mapbox, TomTom, Here e Open Route Service (ORS). A avaliação foi realizada com base no erro de geocodificação em comparação com uma base de dados de referência na região metropolitana de São Paulo. Utilizamos várias métricas para a análise comparativa, incluindo média, desvio padrão, mediana, média aparada em 5%, taxa de resposta (proporção entre solicitações de geocodificação e respostas) e taxa de acerto (quantidade de endereços com erro menor que 150 metros). Além disso, conduzimos uma análise espacial do erro e investigamos a relação entre discrepância e erro, usando a medida de covariância. Devido a problemas na aplicação que coleta as geocodificações, esta etapa do projeto se concentrou apenas nas APIs Mapbox, TomTom e Here, resultando em um desempenho geral insatisfatório. A maioria das APIs apresentou uma taxa de resposta baixa, com a maior delas ficando abaixo de 90%, o que impactou a integridade do experimento. Em relação à taxa de acerto, todas as APIs obtiveram valores considerados insatisfatórios pela nossa equipe de pesquisa. Além disso, observamos a ocorrência de erros significativos que prejudicaram a análise espacial. No que diz respeito à relação entre discrepância e erro, não pudemos identificar uma correlação forte, possivelmente devido ao número limitado de geocodificações realizadas. Para a próxima fase do projeto, planejamos repetir a análise com as APIs restantes para os dados de São Paulo e estender a avaliação para os dados de Belo Horizonte.

**Palavras-chave:** GeoAPIs. Qualidade.

# Abstract

This is the english abstract.

**Keywords:** Keywords1, Keywords2, Keywords3.

# **Lista de Ilustrações**

# Lista de Tabelas

# 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

# Sumário



# 1 Introdução

## 1.1 Endereços e Geocodificação

Quase tudo o que acontece,  
acontece em algum lugar. Saber o  
local onde algo acontece pode ser  
fundamental.

---

No livro de (??), os autores exploram a relação entre a humanidade e a localização. Para eles, é evidente que a maior parte das atividades humanas ocorre no planeta Terra, e, portanto, a vida está profundamente ligada à localização. Assim sendo, compreender e manipular informações geográficas é essencial para qualquer aplicação que envolva a humanidade. Além disso, os autores explicam que decisões importantes podem ter consequências geográficas. Um exemplo disso seria uma transação financeira que, em casos extremos, poderia desencadear uma crise econômica em uma região específica.

No artigo de na análise ambiental

Para atingir esse objetivo, é necessário criar uma representação computacional do endereço para que as aplicações possam utilizá-la. A representação mais comum, conforme

- Processamento do endereço de entrada: o endereço é lido, dividido em componentes (rua, número, bairro, etc.), padronizado e cada campo é atribuído a uma categoria; por fim, as categorias necessárias são indexadas.
- Busca na base de referência: com base no algoritmo escolhido, é realizada uma busca na base de referência para selecionar e classificar potenciais candidatos como resposta.
- Seleção do(s) candidato(s) para resposta: após a busca, a classificação gerada é analisada e os melhores candidatos são selecionados.

De acordo com

- Precisão espacial: é capaz de indicar com alta precisão a localização de um determinado endereço.
- Cálculos de distância: como é um sistema espacial, permite a obtenção de distâncias e, por consequência, o cálculo de outras métricas para o endereço.

- Compreensão global: é um sistema usado mundialmente e, geralmente, é mais fácil de identificar e entender.

Apesar de todas as vantagens e aplicações, o processo de geocodificação pode levar a informações incorretas. No livro de

- Erro: a diferença entre o observado e o obtido.
- Falta de acurácia: a diferença entre a realidade e nossa representação dela.
- Ambiguidade: quando um único valor está presente em mais de um objeto.
- Indefinição: a falta de informações necessárias.

Após definir esses termos, os autores definem a incerteza como: "a 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"

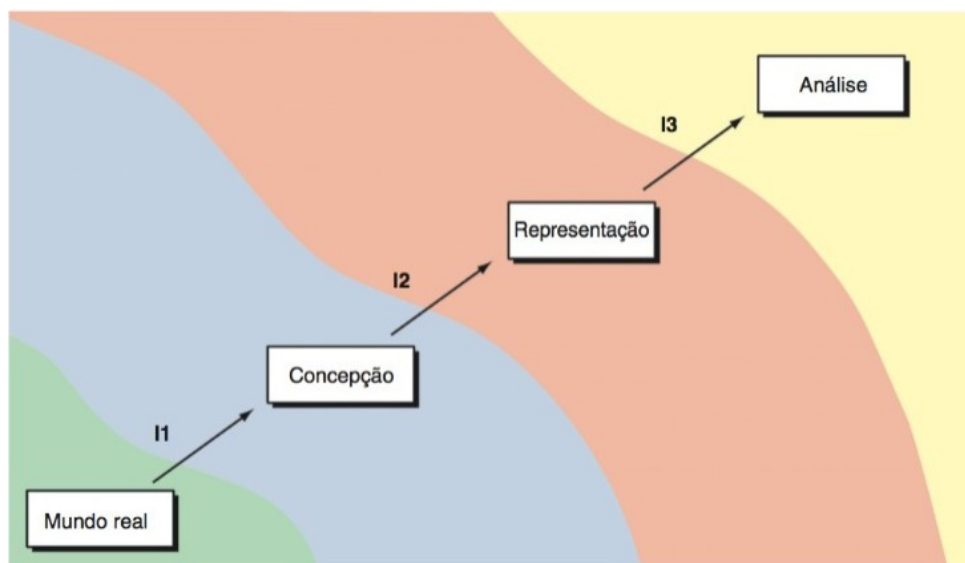


Figura 1.1 – Retirada do livro (??). Visão conceitual da incerteza, onde os filtros I1, I2, I3 distorcem a informação original

## 1.2 APIs de Geocodificação e Análise de qualidade

Atualmente, no TerraLAB - Laboratório de Pesquisa e Capacitação em Software

Para adquirir informações relacionadas a endereços, fazemos uso da geocodificação obtida por meio de APIs online de geocodificação.

Por muitos anos, a principal maneira de obter informações geográficas era através de software SIG. Conforme

## Segundo

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 à qualidade do SIG. No artigo

$$\epsilon_x = x_{\text{ref}} - x_{\text{geoc}} \quad (1.1)$$

$$\epsilon_y = y_{\text{ref}} - y_{\text{geoc}} \quad (1.2)$$

$$\epsilon_{xy} = \sqrt{\epsilon_x^2 + \epsilon_y^2} \quad (1.3)$$

Onde:

- $\epsilon_x$  é o erro da longitude,
- $\epsilon_y$  é o erro da latitude,
- $\epsilon_{xy}$  é o erro euclidiano.

O estudo evidenciou que não há diferença significativa entre as ferramentas online e os SIGs. Tanto os SIGs quanto as ferramentas online apresentaram média e desvio padrão de erro semelhantes. Além disso, a taxa de resposta (ou seja, quantos endereços receberam uma resposta da ferramenta utilizada) variou entre 97,8% e 100%, o que é considerado satisfatório. Dessa forma, o estudo obteve êxito ao demonstrar que as ferramentas online podem ser utilizadas como substitutas dos SIGs.

Apesar de

Outro estudo importante é

Apesar das descobertas importantes, o estudo apresenta limitações notáveis. Primeiramente, ele se restringe à análise de apenas uma API de geocodificação. Além disso, o estudo se concentra exclusivamente em uma cidade brasileira, o que restringe a generalização dos resultados. Para abordar essas limitações, nosso trabalho atual visa expandir a análise. Planejamos examinar uma amostra da mesma base de dados, utilizando diferentes APIs de geocodificação. Além disso, nossa pesquisa incluirá uma análise de uma base de dados da região metropolitana de São Paulo, proporcionando uma maior diversidade ao 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 em relação às respostas fornecidas pelas APIs, verificando o quanto

diferem das esperadas. A discrepância medirá o nível de discordância entre as APIs. Por fim, a acurácia será utilizada para verificar a precisão das respostas fornecidas por essas APIs.

Uma parte essencial deste trabalho é compreender os pontos em que essas APIs apresentam falhas. Portanto, a análise espacial dessas medidas terá grande destaque na pesquisa.

Com isso, gostaríamos de responder as seguintes perguntas:

- Qual API das utilizadas apresenta mais erros?
- Existe algum padrão espacial nos erros?
- Alguma medida de variância entre as APIs (discrepância) está relacionada aos erros?

Para alcançar essas respostas, temos objetivos específicos a serem cumpridos:

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

## 2 Bases de Dados e Métodos de Geocodificação e Avaliação

Para avaliar a qualidade das APIs de geocodificação utilizadas no TerraLAB, recorremos a duas bases de dados padrão-ouro como referência. Utilizando essas bases, calculamos a medida de erro e conduzimos diversas métricas com base nessa medida.

### 2.1 Bases de Dados

Foram coletadas duas bases de dados distintas para este trabalho.

A primeira base coletada é proveniente do Centro de Estudos da Metrópole (CEM) (??). Essa base consiste em 12.502 endereços de escolas públicas e particulares do ensino básico da região metropolitana de São Paulo. A coleta desses dados foi realizada manualmente pelo CEM, utilizando GPS para registrar as coordenadas. Além das informações sobre os endereços, a base também contém uma variedade de informações sobre as escolas, possibilitando diversas análises relacionadas a esses dados. O CEM também disponibilizou um [mapa de cluster](#) que exibe todas as escolas, facilitando a visualização da localização de cada uma delas e da densidade das escolas em São Paulo e região. A Figura ?? mostra o mapa de cluster. Nele, é possível visualizar a localização das escolas individualmente (ao dar zoom) e, ao dar zoom-out, a concentração de escolas em determinadas áreas, utilizando um sistema de cores no qual laranja representa muitas escolas, amarelo representa uma quantidade média e verde representa poucas escolas.

A segunda base de dados coletada foi fornecida pela (??), a empresa de informática e informação da prefeitura de Belo Horizonte. A descoberta dessa base de dados foi possibilitada pelo artigo de referência (??). Essa base de dados é mantida e atualizada mensalmente por 27

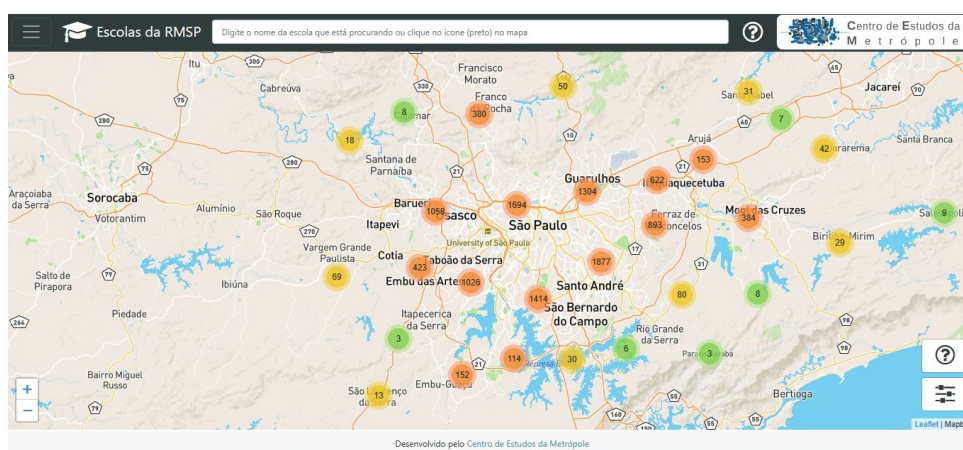


Figura 2.1 – Mapa de clusters do Centro de Estudos da Metrópole

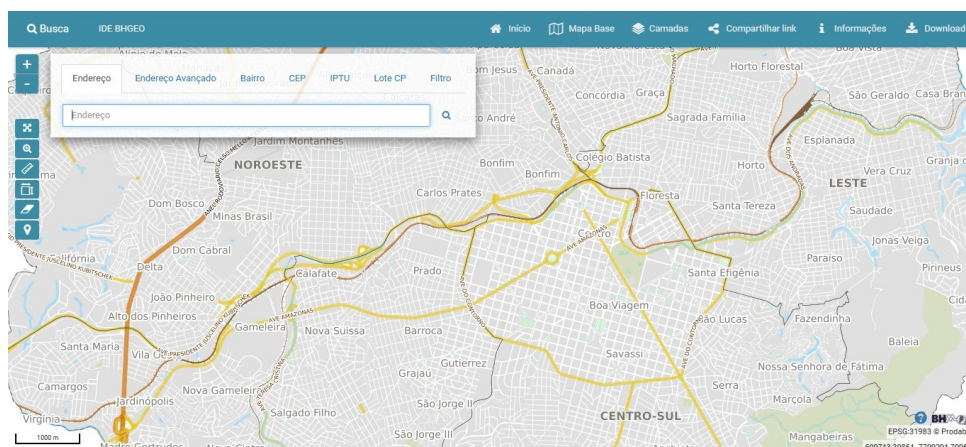


Figura 2.2 – Site da Prodabel para pesquisa de endereços.

empresas, tanto públicas quanto privadas, de Belo Horizonte. Essas empresas têm a responsabilidade de relatar quaisquer inconsistências encontradas na base e de fornecer novos dados à medida que os adquirem. Ela é considerada uma fonte confiável de informações, pois está em constante atualização e é amplamente utilizada por diversos serviços da prefeitura. Um exemplo notável é o uso da base para georreferenciamento na distribuição de alunos da rede pública.

Na data de coleta, essa base continha um total de 763.229 endereços. A prefeitura disponibiliza um [site com um mapa](#) que permite a visualização desses endereços. A Figura ?? mostra esse site, e na barra de pesquisa, os usuários podem pesquisar endereços específicos e marcá-los no mapa. É importante notar que, ao contrário da maioria das APIs de geocodificação, todos os endereços foram posicionados em cima dos edifícios representados. A discrepância entre essa abordagem e a prática comum de colocar o endereço na frente do edifício pode causar um pequeno erro de alguns metros na comparação da geocodificação.

Devido a limitações computacionais tanto dos autores deste trabalho quanto da aplicação responsável pela geocodificação, optamos por realizar uma amostragem da base de Belo Horizonte, com o intuito de reduzir a quantidade de dados processados. Nossa amostra consiste em 85.000 endereços da cidade. A fim de garantir uma distribuição uniforme dos endereços no espaço, empregamos o método do hipercubo latino para a amostragem. A Figura ?? apresenta dois gráficos contendo os pontos da base original e os da amostra obtida. É possível observar que a amostra cobre toda a área abrangida. Além disso, verifica-se uma ligeira concentração nas regiões periféricas do desenho, permitindo uma melhor delimitação da cidade.

## 2.2 Processo de Geocodificação

Após a coleta das bases, é necessário prepará-las para a geocodificação. A etapa de preparação de dados envolve a seleção dos campos relevantes da base de dados, como nome da rua, número, bairro, CEP e cidade. Em outras palavras, serão selecionados apenas os campos descritivos do endereço e os campos de localização geográfica do endereço. Após a seleção,

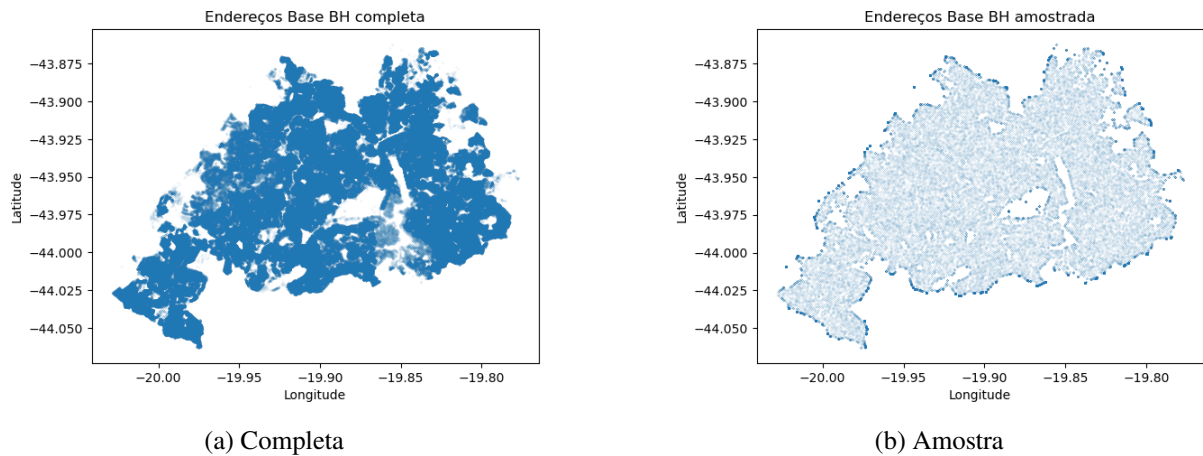


Figura 2.3 – Gráficos dos endereços da Base de Belo Horizonte e amostragem obtida

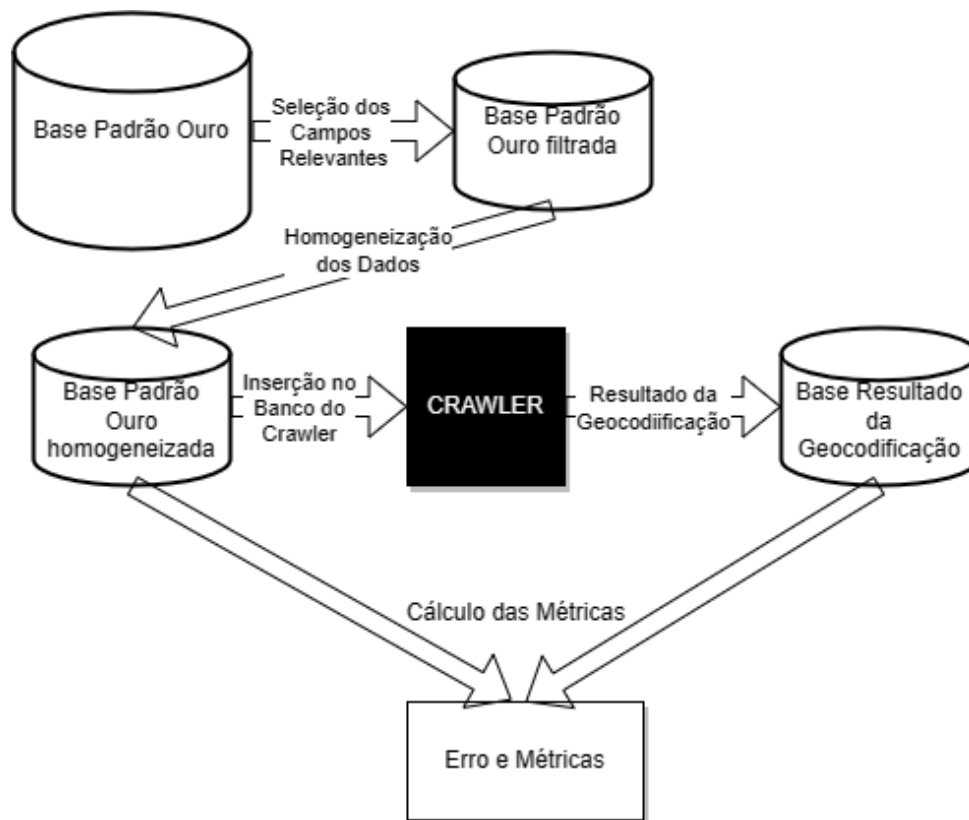


Figura 2.4 – Esquematização do processo de preparação e geocodificação dos dados

os dados são homogeneizados, substituindo abreviações comuns por suas formas completas correspondentes. Esta etapa é conduzida pela equipe do TerraLAB e demonstrou-se que as APIs respondem de forma mais eficaz quando não há abreviações.

Para realizar a geocodificação, os endereços previamente preparados são inseridos no banco de dados do Crawler, a aplicação responsável por solicitar e coletar informações de geocodificação. Os endereços são então retirados do Crawler para serem geocodificados. É importante destacar que o processo de geocodificação é executado pela equipe de Back-end do TerraLAB, e, portanto, é considerado um processo de "caixa preta".

Após a conclusão da geocodificação, os endereços geocodificados, juntamente com suas coordenadas geográficas, são armazenados no mesmo banco de dados, mas em tabelas distintas. A Figura ?? esquematiza todo esse processo essencial para o nosso trabalho.

## 2.3 Método de Avaliação

### 2.3.1 Erro, Acurácia e Discrepância

A principal métrica utilizada para avaliar a qualidade da geocodificação é o erro do endereço. 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. Esse erro é calculado como a distância entre o ponto de referência e o ponto geocodificado pela GeoAPI, conforme a equação abaixo:

$$e = D(p_{\text{Ouro}}, p_{\text{Geo}}) \quad (2.1)$$

Onde:

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

Além disso, 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 é convertido em uma porcentagem para facilitar a compreensão dos resultados, de acordo com a seguinte fórmula:

$$\text{Taxa de Resposta (\%)} = \left( \frac{\text{Quantidade de Endereços Geocodificados}}{\text{Quantidade de Endereços Originais}} \right) \times 100\% \quad (2.2)$$



## 3 Resultados

Para a primeira etapa do projeto, fizemos a análise do erro e discrepância para os dados de São Paulo. Por problemas na aplicação que coleta as geocodificações, obtivemos resultados apenas para 3 APIs (TomTom, Mapbox e Here). Abaixo serão apresentados os resultados obtidos.

### 3.1 Distribuição Espacial dos Pontos Geocodificados

Após a geocodificação dos dados, tornou-se interessante visualizar como os pontos geocodificados estavam distribuídos no espaço e quão diferentes eram em relação aos pontos de referência ("Gold"). Para essa finalidade, foram gerados mapas identificando os pontos para cada uma das APIs.

É importante observar que, com base apenas nessa visualização, não é possível tirar conclusões definitivas. No entanto, é possível analisar a densidade dos pontos e identificar que, em todas as APIs, houve uma maior concentração de dados "Gold". No entanto, em algumas APIs, essa concentração é visivelmente menor do que em outras. Além disso, é notável 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 a ocorrência de alguns erros graves de geocodificação, conhecidos como "outliers".

Na Figura ??, podemos visualizar a distribuição espacial dos pontos geocodificados pela Mapbox. Nela, é possível observar a presença dos outliers mencionados anteriormente. No entanto, é importante ressaltar que esses casos de erro são relativamente raros. Portanto, considerando apenas essa análise, a API obteve resultados satisfatórios.

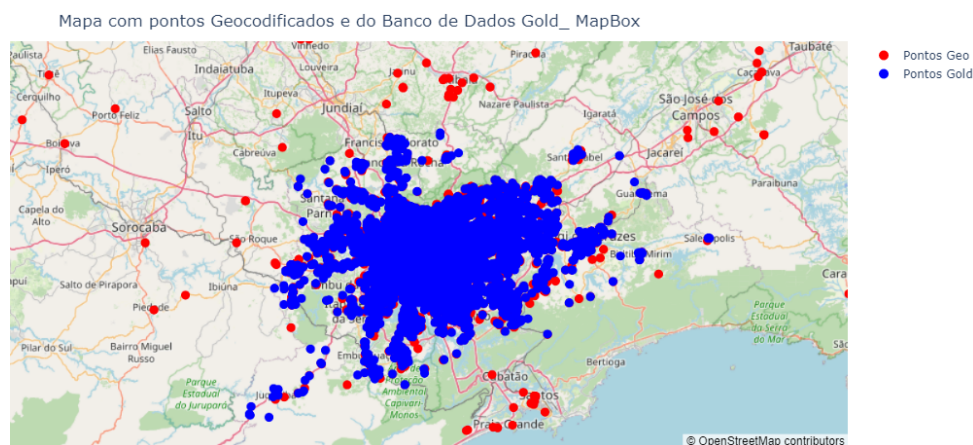


Figura 3.1 – Mapa da Distribuição Espacial dos Pontos da base Gold e Geocodificados pela Mapbox

Na Figura ??, podemos observar a distribuição espacial dos pontos geocodificados pela

Here. Fica claro na imagem que houve uma diminuição significativa dos pontos, indicando uma baixa resposta da API. Com essa quantidade reduzida de pontos, não é possível tirar conclusões sólidas sobre os dados. No entanto, é evidente que, além da baixa resposta, os pontos parecem estar dispersos em locais distintos. Esses resultados foram considerados insatisfatórios. Em um momento subsequente, o experimento será repetido para que possamos obter conclusões mais precisas.

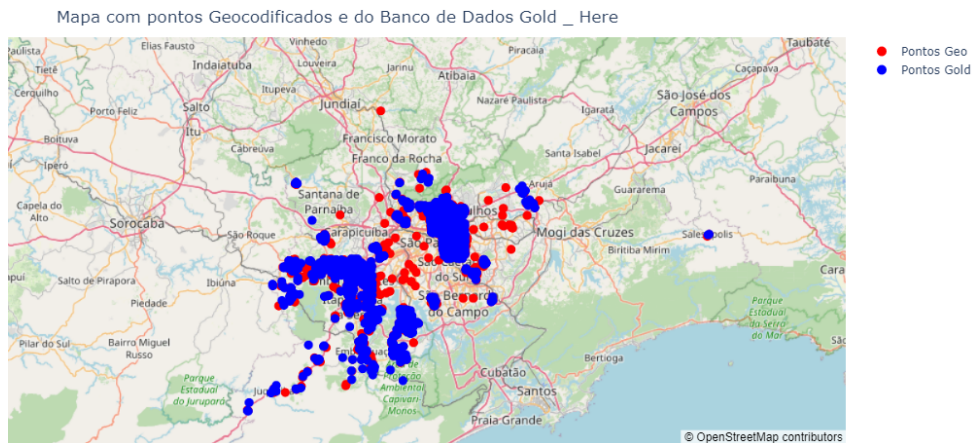


Figura 3.2 – Mapa da Distribuição Espacial dos Pontos da base Gold e Geocodificados pela Here

Já a Figura ?? mostra a distribuição espacial dos pontos geocodificados pela TomTom. Com esse mapa, é possível observar que a resposta da API foi boa em comparação com os mapas apresentados anteriormente. Embora ainda tenha alguns "outliers", aparentemente esses estão em maior quantidade do que na Figura ?? e estão mais dispersos geograficamente. No entanto, considerando apenas essa análise, o resultado é considerado satisfatório.

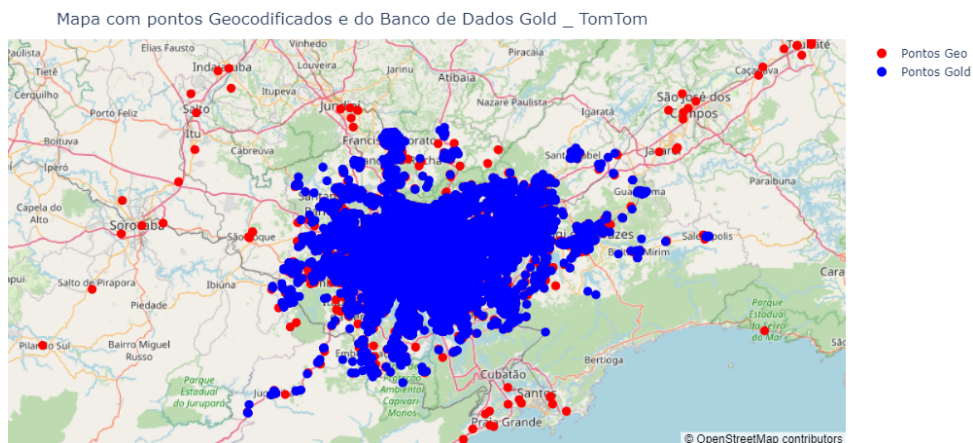


Figura 3.3 – Mapa da Distribuição Espacial dos Pontos da base Gold e Geocodificados pela TomTom

## 3.2 Métricas do Erro

A próxima etapa foi o cálculo do erro para cada um dos pontos, sendo este expresso em quilômetros (Km).

Tabela 3.1 – Métricas de Erro e Resposta

API	Média (km)	Mediana (km)	Desvio Padrão (km)	Média Aparada (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

Com o erro de cada um dos pontos, foram calculadas as métricas mencionadas anteriormente. A ?? mostra esses resultados.

Em relação a taxa de resposta, ou seja, a quantidade de endereços que foram geocodificados, a TomTom tem o melhor resultado, com um índice superior a 80%, seguida pela Mapbox, com taxa de 53,38%. A Here obteve uma taxa de resposta baixa, como esperado pela análises anteriores. Apesar de ter uma API com taxa de resposta alta, esse resultado foi considerado limitante para equipe pois nos impede de fazer algumas análises. Outra métrica importante é a taxa de acerto. Foi considerado como acerto aqueles endereços que tiveram erro menor que 150m (0.015Km). A taxa de acerto foi baixíssima para todas as APIs, sendo a melhor 30.19%. Esse é um resultado péssimo para os dados acumulados. Porém, devido a baixa quantidade de dados não é possível concluir que as APIs em questão tem uma performance ruim. Na próxima etapa do projeto iremos fazer a análise comparativa dos resultados com as outras APIs e com uma maior quantidade de dados.

Outras métricas interessantes obtidas foram as métricas de média, mediana e desvio padrão. Com elas é possível ver o comportamento geral do erro em cada uma das APIs. As médias foram muito altas, indo de 2Km a 10Km. O desvio padrão também foi alto, mostrando que há uma grande variação no erro. Apesar disso, a mediana foi bem baixa, alcançando resultados desejáveis na nossa pesquisa. A média aparada obteve resultados muito bons, o que indica que com a retirada dos outliers as métricas tendem a melhorar. Como trabalho futuro, pretendemos refazer as análises com o corte em 50km de erro. De forma geral, esses resultados foram considerados insatisfatórios. Ao longo do relatório, iremos analisar outras questões em detalhes.

### 3.3 Distribuição do Erro

Em seguida, foi realizada a análise da distribuição do erro para cada uma das GeoAPIs.

Para isso, utilizamos histogramas de erro individualmente para cada API e combinando todas elas. Na ?? é mostrado os histogramas para cada uma das APIs e o histograma que é a combinação de todas elas. No entanto, devido à presença de alguns erros exorbitantes, esses histogramas não são muito representativos, pois a maior parte do erro se concentrava entre 0 km e 50 km. Esse intervalo é considerado um erro muito grande, o que dificulta a obtenção de conclusões sólidas.

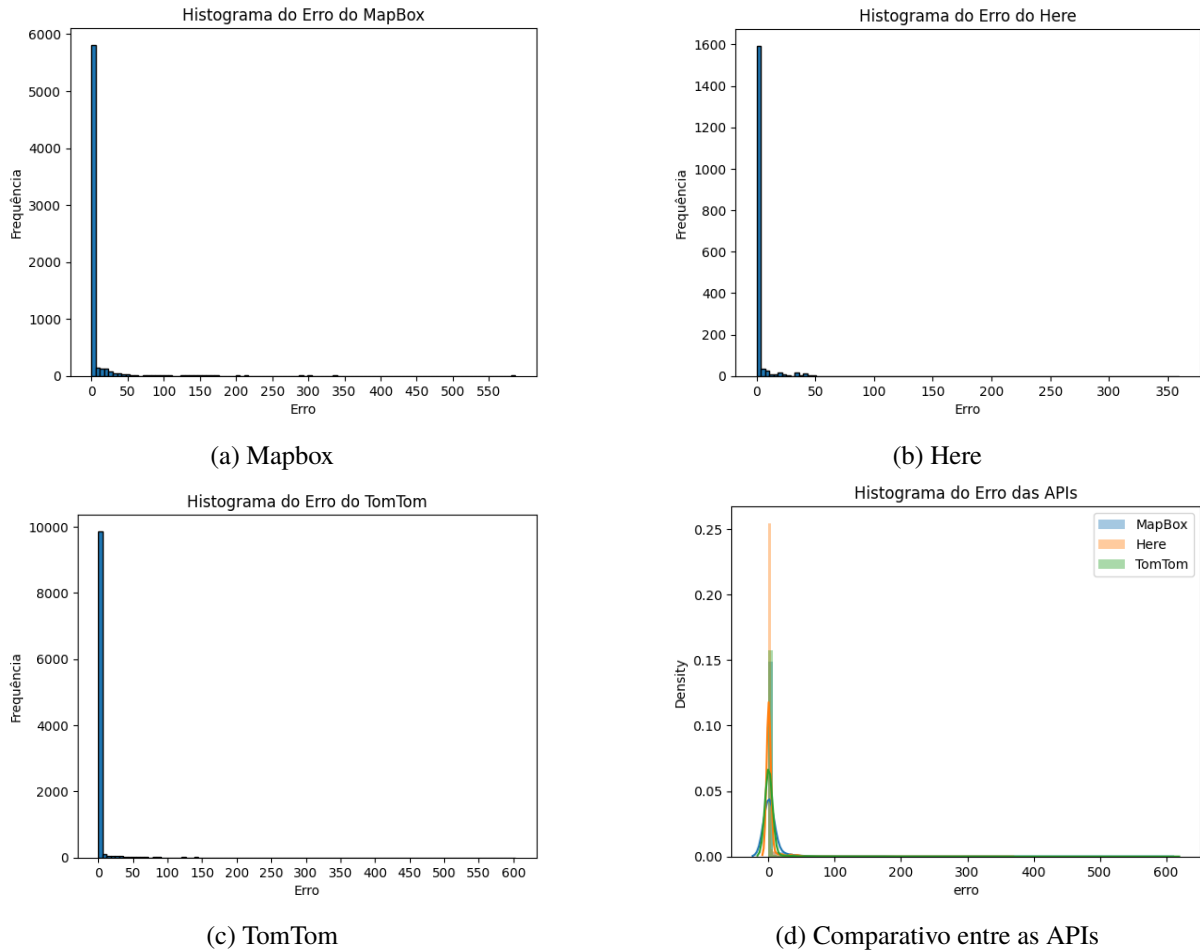


Figura 3.4 – Histogramas do erro das 3 APIs para o todos os dados.

Diante disso, decidimos realizar um corte nos dados, limitando o erro em 0,5 km ou 500 metros. Em seguida, repetimos o processo, gerando agora um único histograma que representa a distribuição do erro para todas as APIs em conjunto. A Figura ?? apresenta esse histograma. Nele, observamos que a maior parte dos dados está concentrada entre erros de 0,0 km e 0,1 km. Isso confirma a hipótese de que as métricas se comportariam melhor com a remoção dos outliers.

Em relação às APIs, nessa faixa de erro, a TomTom apresenta um desempenho superior, com uma curva mais estreita e mais próxima de 0. No entanto, a diferença entre as APIs não é significativa.

De forma geral, embora o histograma seja uma ferramenta poderosa para a análise da distribuição do erro, neste caso, ele não se mostrou tão eficiente devido às limitações decorrentes da presença de valores excessivamente altos.

### 3.4 Distribuição Espacial do Erro

Além disso, realizamos uma análise adicional para visualizar como esse erro se comporta no espaço. Criamos mapas de altitude, nos quais o erro foi utilizado como medida de altitude.

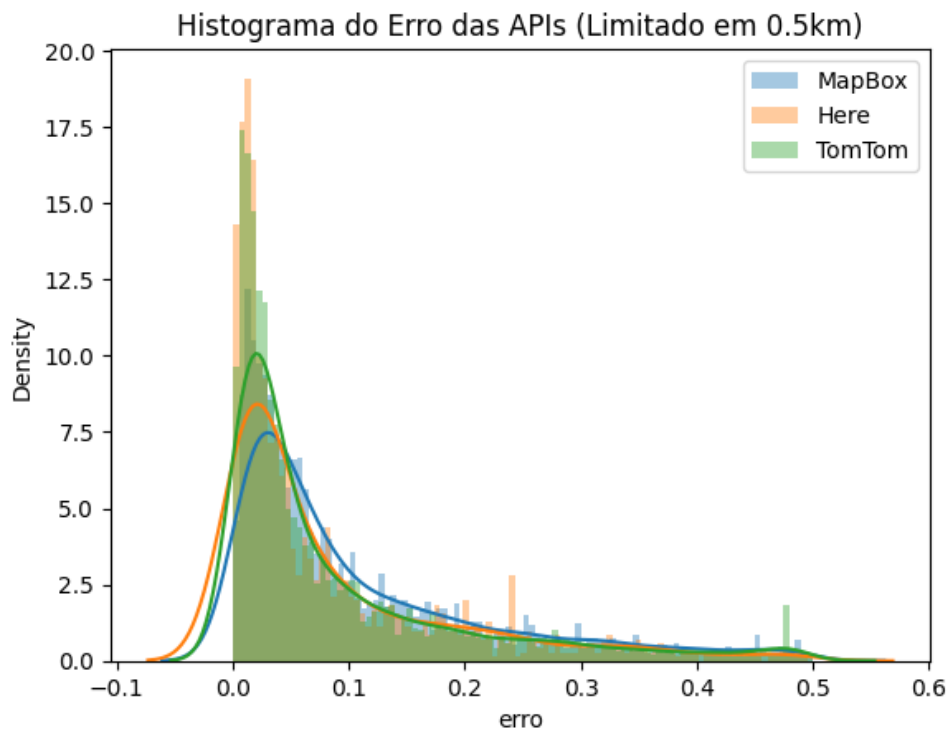


Figura 3.5 – Histograma comparativo do erro das APIs limitado em 500 metros

O mapa consiste em linhas de altitude, criadas a partir da interpolação do erro dos pontos. Essas linhas foram então transformadas em um polígono e coloridas de acordo com os valores 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. Todo esse processo foi realizado utilizando as bibliotecas Python matplotlib, scipy e folium para a criação e visualização dos gráficos, e as bibliotecas pandas e numpy para a manipulação das estruturas de dados.

Ao analisar os resultados, observamos que a maioria do mapa apresenta erros menores que 34 km, conforme observado nos histogramas acima. 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.

Na Figura ??, conseguimos ver a abrangência da geocodificação da Mapbox, ou seja, os pontos geocodificados conseguiram abranger boa parte da região metropolitana de São Paulo. Além disso, o erro ficou concentrado em 25 km na maior parte do gráfico. Em alguns pontos, ela apresentou erros entre 50 km e 100 km, o que já é considerado um ponto preocupante. Existiram

alguns erros na faixa de 300 km nas periferias da cidade. No entanto, pode-se observar que há uma baixa concentração de pontos, o que indica que existem poucos pontos com erro alto, que causaram essa visualização. No centro, existem alguns pontos avermelhados que possuem uma grande concentração de pontos, o que é um dado preocupante, pois indica que a API realmente está errando bastante em relação aos dados de referência naquela região.

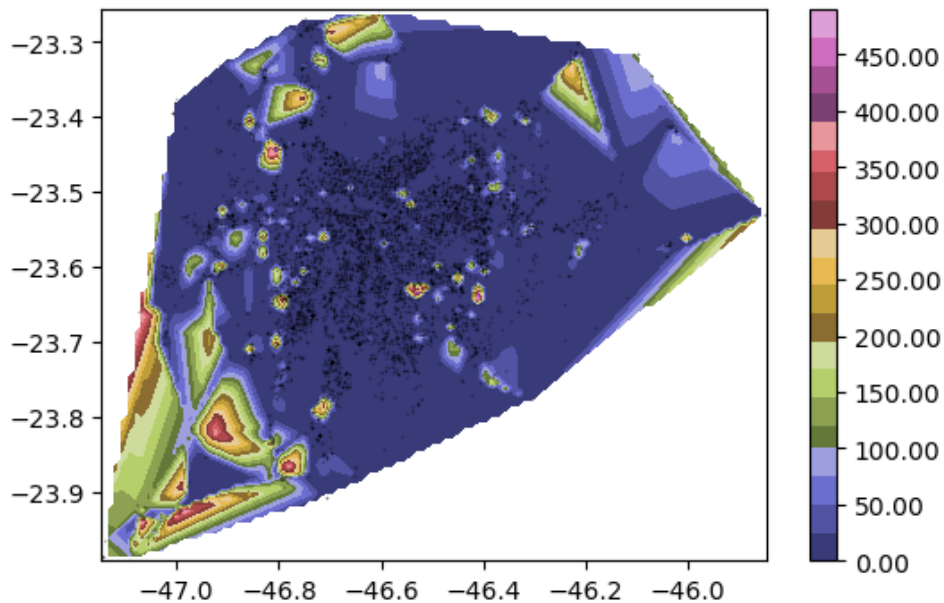


Figura 3.6 – Gráfico de altitude do erro (km) da geocodificação da Mapbox.

Já a Figura ?? demonstra a baixa abrangência da geocodificação da Here. Como apresentado anteriormente, essa GeoAPI teve a menor taxa de resposta, e a visualização pelo gráfico de altitude apenas confirma isso. Sendo assim, qualquer análise realizada será enviesada. É possível observar uma grande concentração de azul, indicando que os dados têm erro pequeno. Existem alguns picos com erro elevado, porém, apenas um deles apresenta pontos suficientes para considerar que a região tem erro alto.

A Figura ?? é semelhante à Figura ??, pois também teve uma taxa de resposta suficientemente grande. A maior parte do gráfico possui uma cor azul escura, o que indica que o erro nessas regiões é menor que 20 km. No entanto, existem alguns picos esverdeados que mostram um erro bem alto. Em alguns deles, não é possível observar uma grande concentração de pontos, o que indica que são poucos pontos com erro grande. Já no centro da figura, onde há uma grande concentração de pontos, esses picos também aparecem, o que indica um erro alto nessa região.

### 3.5 Relações entre erro e discrepância

Por fim, foi realizada a análise comparativa entre erro e discrepância. A medida escolhida para essa análise foi a covariância. Calculamos a covariância entre a latitude e a longitude para cada ponto utilizando as 3 APIs e depois foi calculada a média das duas covariâncias citadas



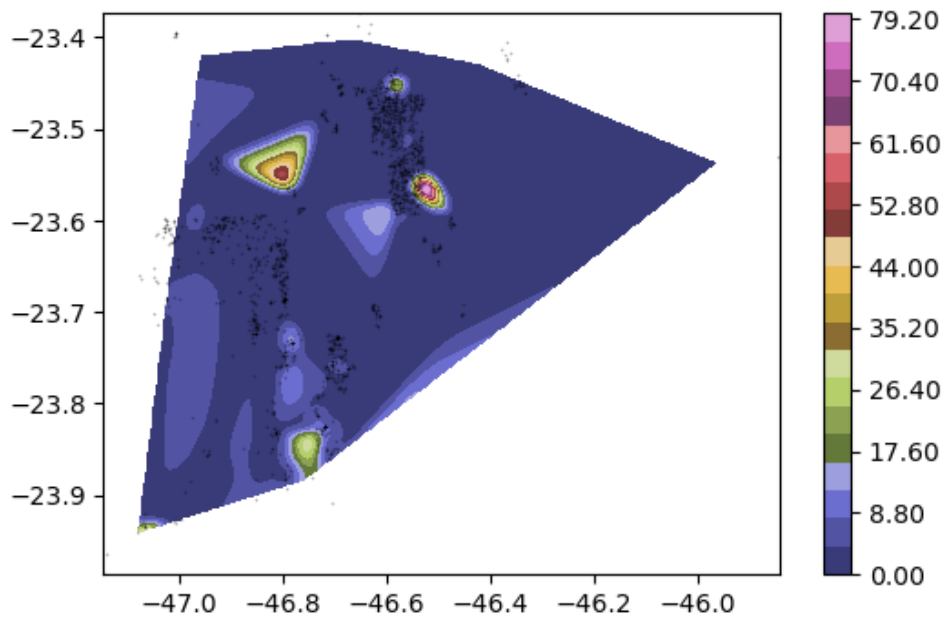


Figura 3.7 – Gráfico de altitude do erro (km) da geocodificação da Here.

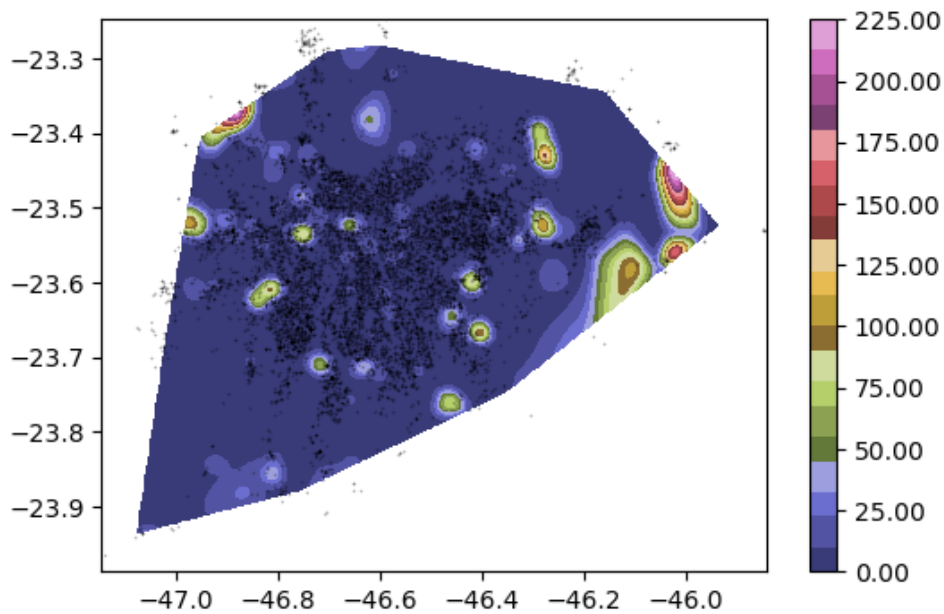


Figura 3.8 – Gráfico de altitude do erro (km) da geocodificação da TomTom.

anteriormente. Foram considerados para análise apenas os pontos em que se tem informação das 3 APIs. Devido ao fato da API Here ter tido uma taxa de resposta tão baixa, a quantidade de pontos obtidas foi também muito baixa. Fizemos então uma análise com apenas 1683 endereços da base.

Após o cálculo da covariância, construímos para cada uma das APIs um gráfico de covariância pelo erro referente. Nos gráficos das Figuras ?? e ?? não é possível observar uma relação entre as duas medidas. O primeiro mostra que os valores de erro estão concentrados na faixa de 0 km a 50 km para esses dados. No segundo gráfico, há uma variação maior do erro,

mas essa variação não reflete na variação da covariância. Os dados indicam que essa relação não existe, porém não é possível tirar tal conclusão devido à quantidade baixa de dados.

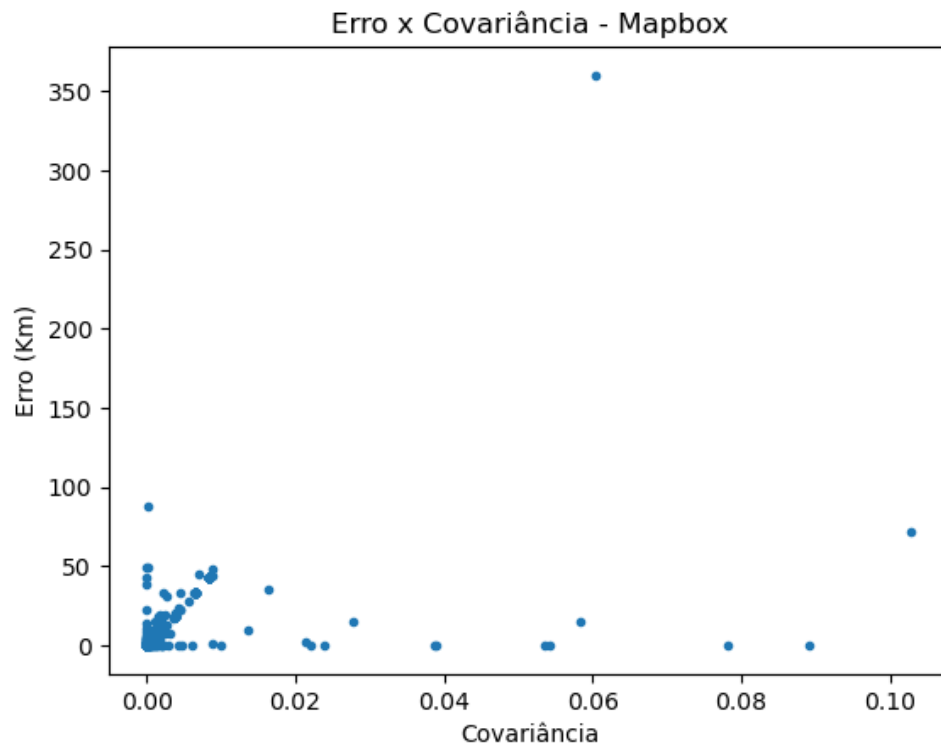


Figura 3.9 – Gráfico covariância por erro da Mapbox

A Figura ?? apresenta resultados diferentes dos anteriores. No gráfico, pode-se observar uma relação próxima à linear, onde à medida que o erro cresce, a covariância também cresce. Como os dados de covariância estão concentrados em 0 a 0.02 e os dados de erro em 0 km a 100 km, as faixas apresentam mais pontos. É um resultado promissor, porém devem ser feitas mais análises devido à quantidade de dados.



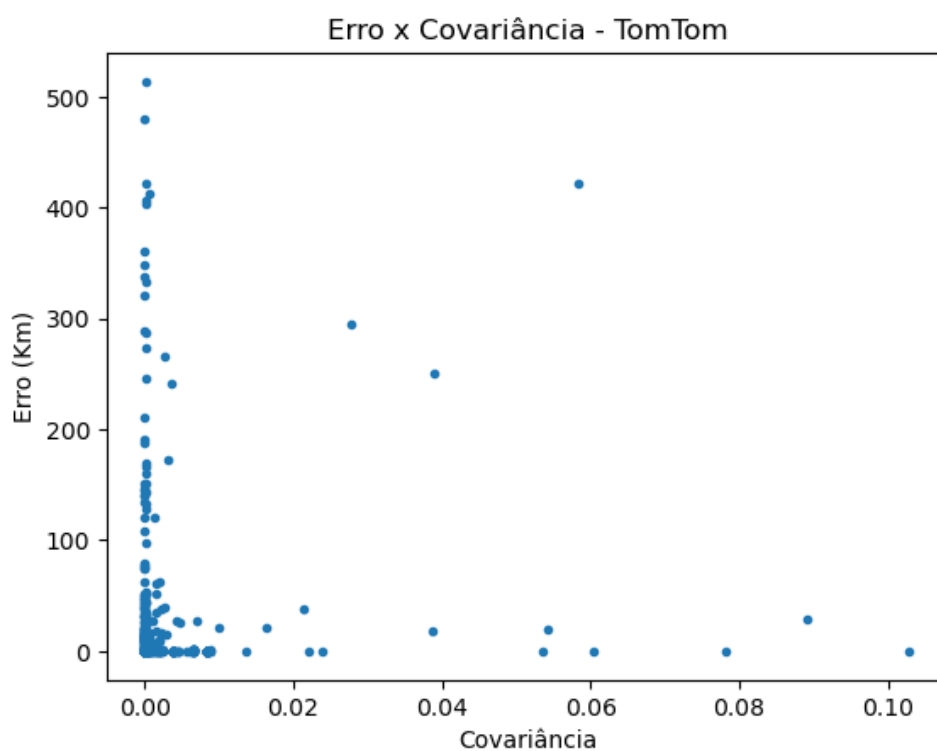


Figura 3.10 – Gráfico covariância por erro da TomTom

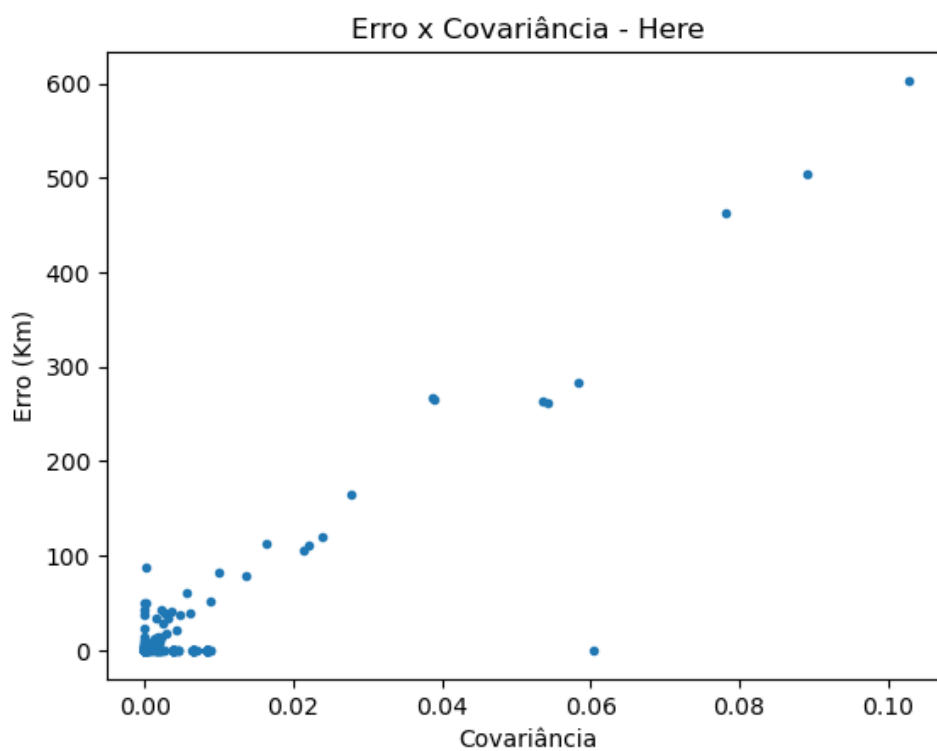


Figura 3.11 – Gráfico covariância por erro da Here

## 4 Considerações Finais

O presente trabalho apresentou uma análise da qualidade das APIs Mapbox, TomTom e Here para os dados disponibilizados pelo CEM - Centro de estudos da Metrópole (??). Devido a problemas no Crawler, que é a aplicação que solicita e coleta a geocodificação, tivemos poucas respostas e estas foram insatisfatórias. A conclusão atual é de que as API cometem muitos erros graves e que não há relação clara entre a discrepância e o erro.

No entanto, quaisquer conclusões tiradas a partir desse estudo são enviesadas a partir do momento em que não temos dados o suficiente e estes são dados específicos. É importante ressaltar que a base de dados possui apenas endereços de escolas, não tendo uma diversidade de imóveis, localizados na região metropolitana de São de Paulo, o que limita a diversidade de localidades consideradas.

Sendo assim, é necessária a repetição do experimento com um maior montante de dados. Para a próxima etapa do trabalho, iremos repetir os experimentos apresentados com uma nova solicitação de geocodificação, além de incluir as APIs faltantes, Google Maps e Open Route Service. Acreditamos que, ao repetir o experimento, possamos compreender melhor o comportamento do erro e comparar os resultados com APIs já consolidadas na academia, como o Google Maps. Além disso, planejamos realizar toda a análise para uma amostra significativa da base de dados da (??), que conta com 85 mil endereços distribuídos no espaço. Esperamos que com a maior quantidade de endereços, possamos analisar o comportamento de forma mais clara. Em relação à análise de discrepância, planejamos acrescentar outra medida à análise, a distância para o ponto médio, que acreditamos ser promissora para o trabalho.

Por fim, esclarecemos que o [ChatGPT](#) foi utilizado durante o trabalho para revisar o texto. O comando "Revise" foi utilizado em textos previamente escritos e depois revisado pelos autores, para garantir a concisão dos dados apresentados.

# Referências