

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 bibliográfica, 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 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 (??), o autor apresenta aspectos relevantes das informações geográficas que complementam o que foi mencionado anteriormente. Segundo ele, o endereço é a principal maneira de conceitualizar a localização no mundo atual. Isso ocorre devido ao fato de os endereços serem utilizados em diversas aplicações de diferentes áreas de estudo, como na saúde (??????), nas ciências sociais (??), na análise criminal ou judiciária (??), na análise ambiental (??), na ciência da computação (??), na economia (??) e em outros campos.

Para atingir esse objetivo, é necessário criar uma representação computacional do endereço para que as aplicações possam utilizá-la. A representação mais comum, conforme (??), é a utilização de coordenadas x e y em um plano, geralmente representando latitude e longitude. Esse processo de transformação em um endereço nessas coordenadas é chamado de Geocodificação ou Georreferenciamento. De acordo com (??), esse processo envolve três etapas:

- 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 (??), além de representar um endereço computacionalmente, o georreferenciamento utilizando latitude e longitude oferece várias vantagens:

- 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 (??), os autores chamam essas discrepâncias de "incertezas". Para compreender o que é a incerteza, é necessário considerar outros aspectos das falhas de informação. Nesse contexto, são introduzidos os seguintes conceitos:

- Erro: a diferença entre o observado e o obtido.
- Falta de acurácia: a diferença entre a realidade e nossa representação dela.
- Ambiguidade: quando mais de um valor é igualmente provável.
- 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"(??). Em outras palavras, a incerteza é uma medida que descreve o nível de compreensão do usuário em relação ao conjunto de dados obtidos e à realidade que esses dados têm a intenção de representar. A partir dessa perspectiva, a incerteza foi aceita como uma métrica apropriada para avaliar a 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 informações geográficas para o desenvolvimento de nossas aplicações. Esses aplicativos utilizam endereços geocodificados para criar mapas, rotas, áreas de abrangência, relatar locais, divulgar eventos, entre outras funcionalidades. Isso ressalta a grande importância da geocodificação e como a qualidade desse processo impacta significativamente o que é produzido em nosso laboratório.

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

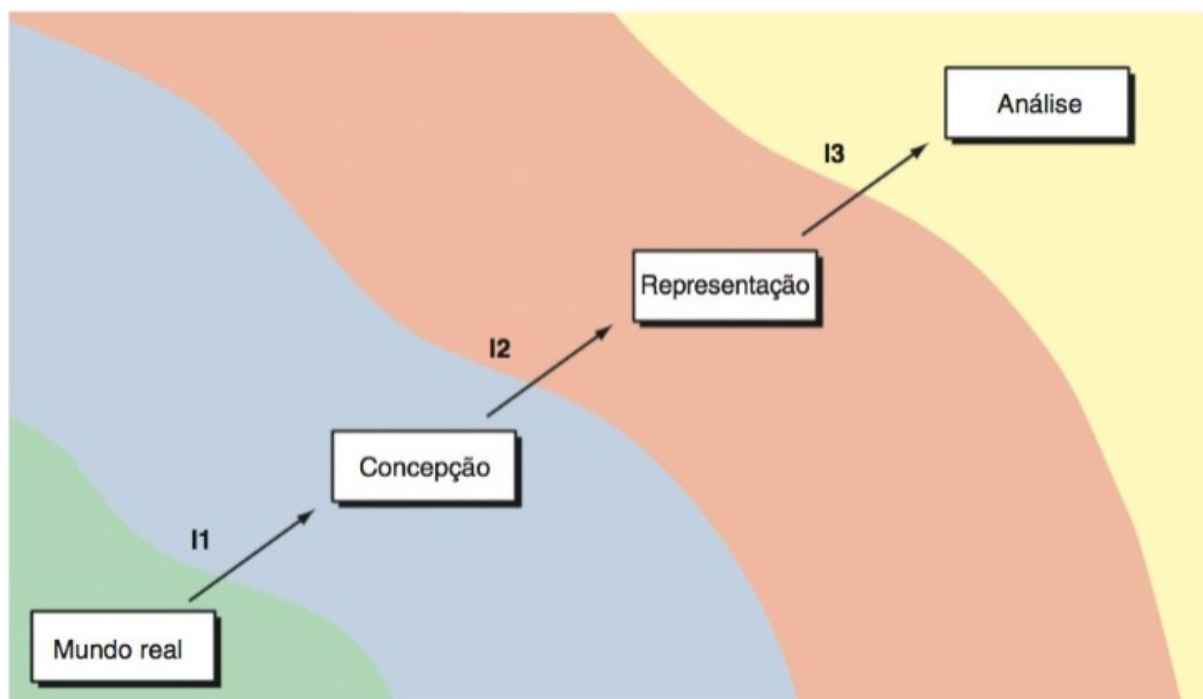


Figura 1.1 – Retirada do livro (??). Visão conceitual da incerteza, onde os filtros I1, I2, I3 distorcem a informação original

Por muitos anos, a principal maneira de obter informações geográficas era através de software SIG. Conforme (??), um Sistema de Informação Geográfica (SIG) é um conjunto de ferramentas capazes de analisar e integrar dados geográficos, permitindo acesso fácil a dados para os usuários, sem depender de ferramentas como o GPS.

Segundo (??), embora os SIG tenham sido a ferramenta convencional por muitos anos, utilizar esse método para geocodificação requer um profissional capacitado. A ferramenta demanda o pré-processamento dos dados, criação de um localizador de endereços, customização de parâmetros, controle de qualidade e correção manual de falhas. Todo esse processo é custoso para o usuário comum. Por essa razão, a geocodificação utilizando ferramentas online retira do usuário grande parte da responsabilidade, como a manutenção da base, tornando assim o processo de obtenção de informações menos oneroso. 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 (??), são avaliadas oito ferramentas de geocodificação, sendo duas delas SIGs e as demais 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 e OpenStreetMap (OSM). Para calcular o erro, uma base de referência foi utilizada, contendo informações descritivas do endereço (rua, número, cidade etc.) e informações geográficas (latitude e longitude). Essa base é considerada a referência, pois os dados de latitude e longitude foram obtidos manualmente (por GPS ou pesquisa manual). Chamaremos essa e outras bases de referência de "base padrão ouro". A base em questão contém 940 endereços do estado do Texas, Estados Unidos da América, sendo que 78

destes são da região Central Texas, região considerada importante para o autor. O erro de cada endereço geocodificado foi calculado da seguinte forma:

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

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

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

Onde:

- ϵ_x é o erro da longitude,
- ϵ_y é o erro da latitude,
- ϵ_{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 (??) ter apresentado resultados significativos, o estudo apresenta algumas limitações. A principal delas é a quantidade de dados utilizada para a avaliação, além do foco restrito a uma única região (Texas, EUA). O presente trabalho busca abordar essas limitações ao conduzir a análise em uma região diferente do mundo, com ênfase no Brasil, e ampliar a quantidade de dados avaliados. Entretanto, nosso enfoque será exclusivamente em ferramentas de geocodificação online (GeoAPIs), considerando que elas já estão consolidadas no mercado e na academia.

PAREI A REVISÃO AQUI

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 espaciais 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 visualizar 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 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.

2.1 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 visualização 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 1. É uma base de dados mantida e atualizada mensalmente por 27 empresas públicas e privadas

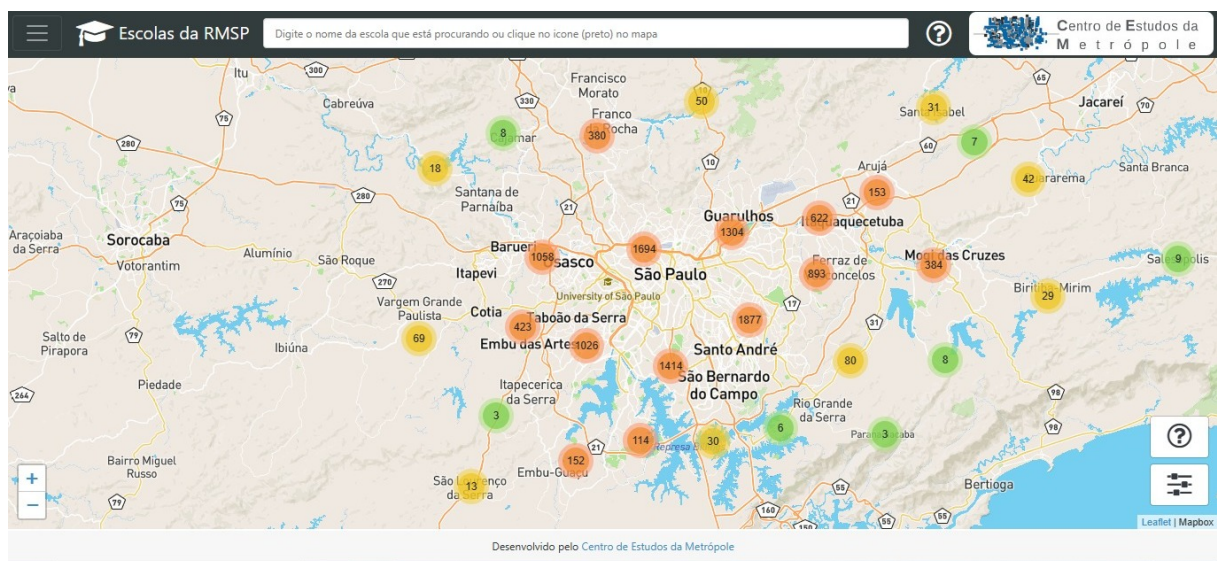


Figura 2.1 – Mapa de clusters que mostra a quantidade de escolas em cada região. Ao aproximar o mapa, o usuário consegue ver a localização de cada uma das escolas presentes no banco de dados.

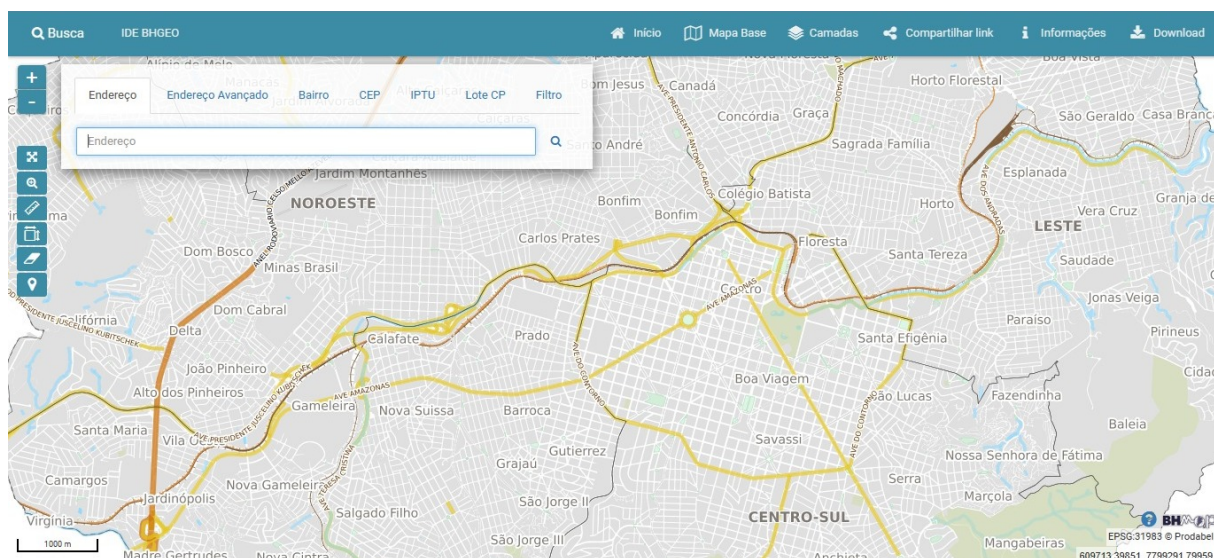


Figura 2.2 – Mapa que mostra a cidade de Belo Horizonte, desenvolvido pela Prodabel. Na barra de pesquisa, é possível pesquisar os endereços e marcá-los no mapa.

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 visualização dos endereços registrados. O endereço está posicionado em cima do edifício representado. Isso pode gerar erro de alguns metros devido a maioria das APIs colocar o endereço na frente do edifício representado.

2.2 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 disponíveis. Essas ferramentas utilizam algoritmos e informações geográficas para identificar e atribuir as coordenadas geográficas correspondentes a cada endereço. É importante ressaltar

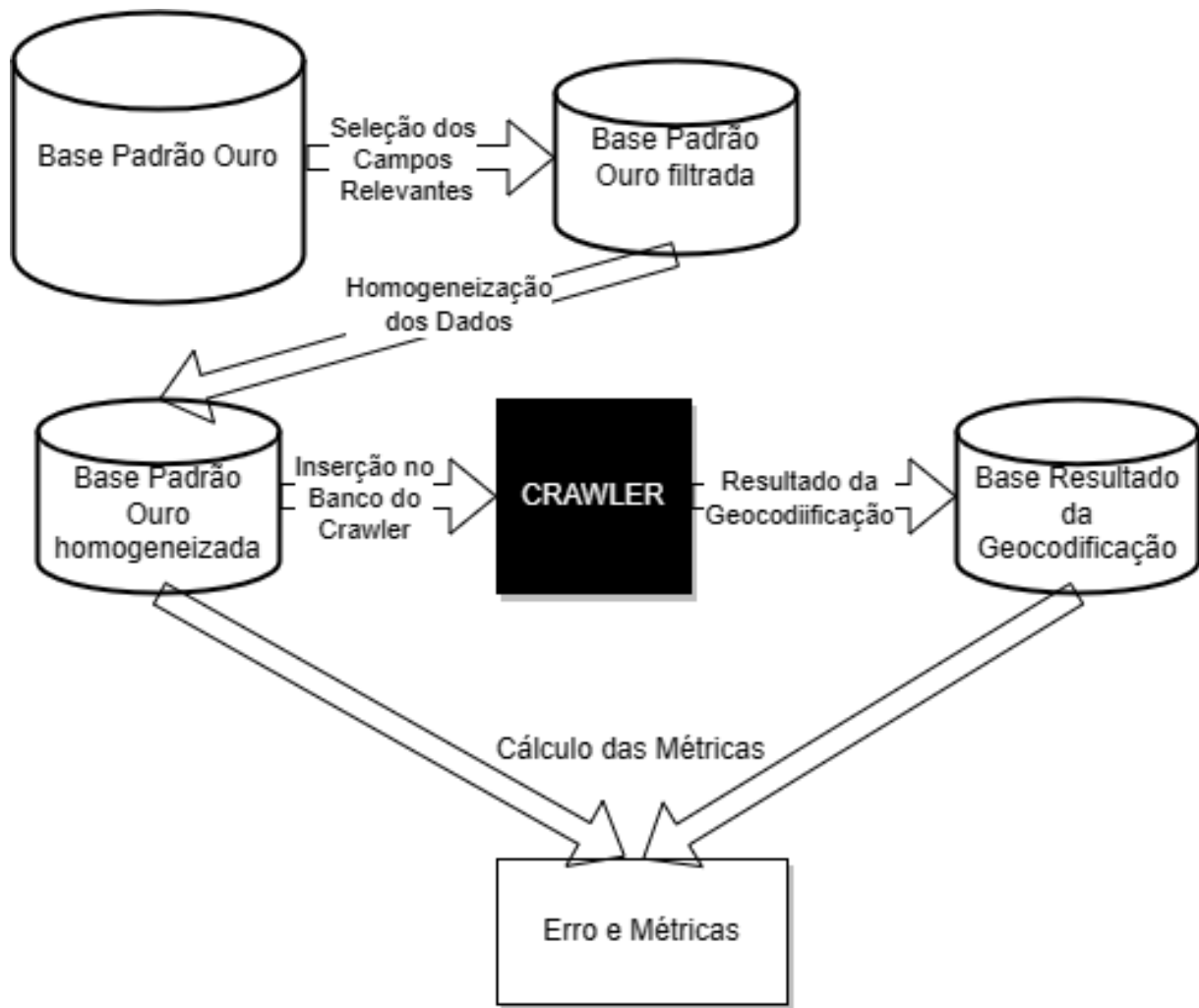


Figura 2.3 – Esquematização do processo de preparação e geocodificação dos dados

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

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. 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.

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

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.

Fica claro na imagem que houve uma diminuição significativa dos pontos. O que indica que a resposta da API foi baixa. Com essa quantidade de pontos não é possível tirar conclusões fortes sobre os dados, porém observamos que além da baixa resposta os pontos parecem estar em locais distintos. Esse resultado foi então considerado insatisfatório. Em outro momento, o experimento será repetido para que possamos tirar as conclusões corretas.

Mapa com pontos Geocodificados e do Banco de Dados Gold _ Here

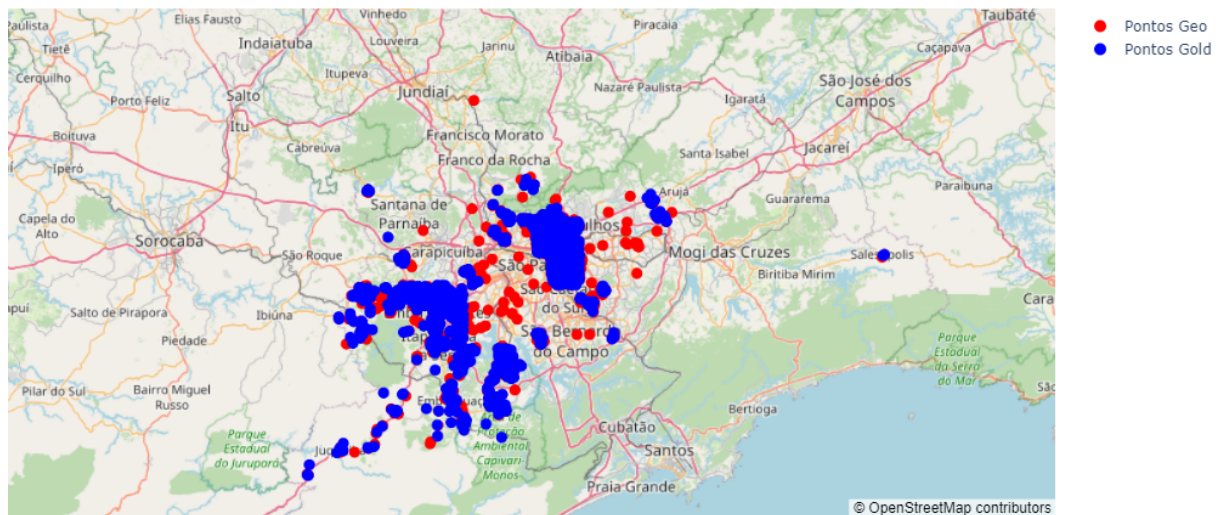


Figura 3.2 – Mapa da Distribuição Espacial dos Pontos da base Gold e Geocodificados pela Here

Já a ?? 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. Também teve alguns outliers e aparentemente esses estão em maior quantidade que na ?? e estão mais espaçados geograficamente. Apesar disso, apenas com essa análise, consideramos o resultado satisfatório.

Mapa com pontos Geocodificados e do Banco de Dados Gold _ TomTom

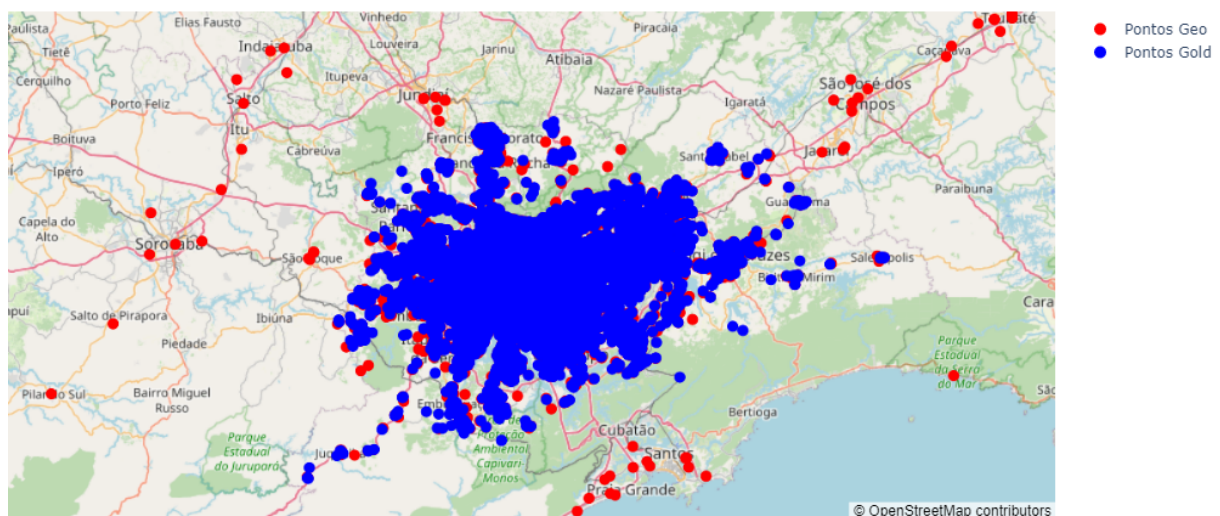


Figura 3.3 – Mapa da Distribuição Espacial dos Pontos da base Gold e Geocodificados pela TomTom

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

3.2 Métricas do Erro

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

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, pretendendo 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. O que é considerado um erro muito grande, sendo assim, não se pode tirar conclusões firmes.

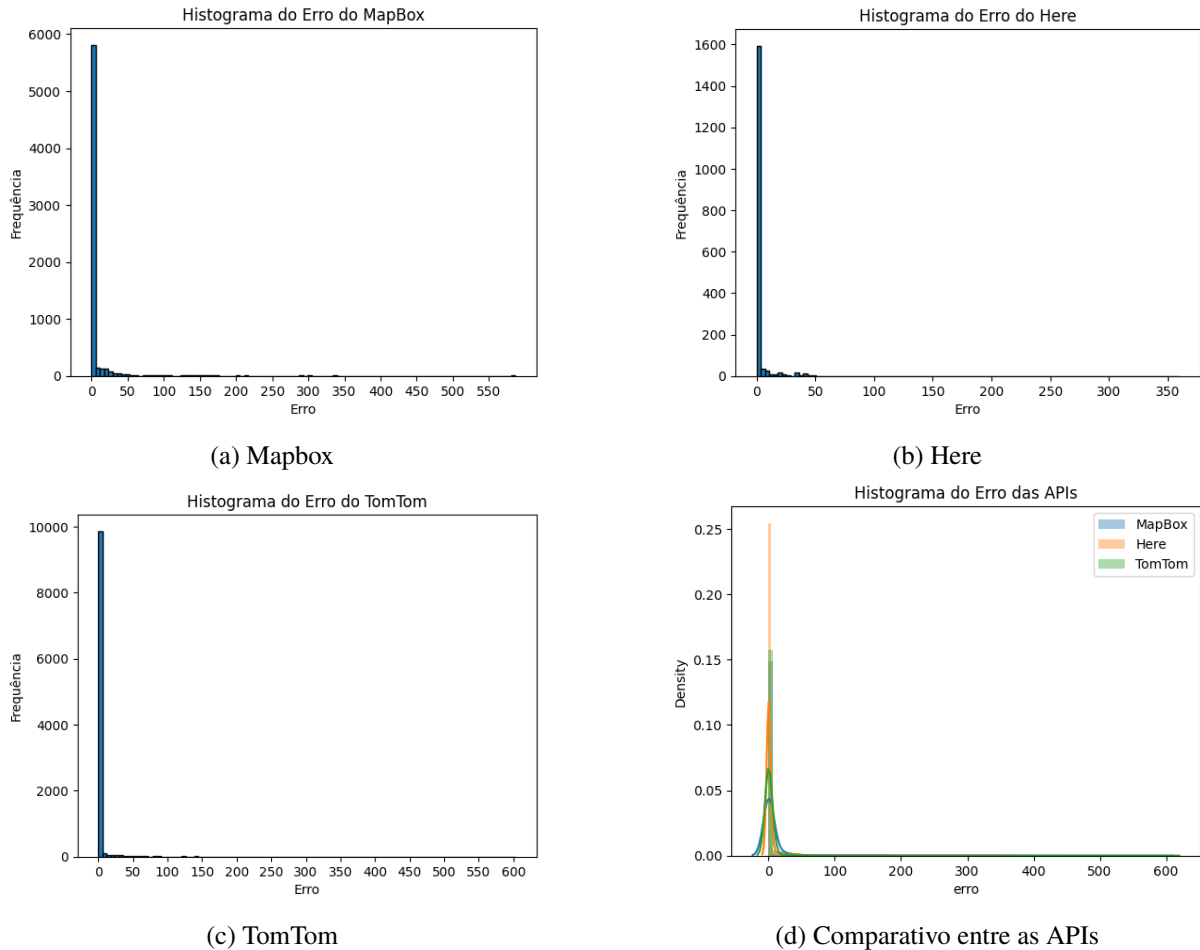


Figura 3.4 – Histogramas do erro das 3 APIs para o todos os dados.

Diante disso, decidimos realizar um corte nos dados, limitando o erro em 0.5 km ou 500 metros. Em seguida, repetimos o processo, agora gerando um único histograma que representa a distribuição do erro para todas as APIs em conjunto. A ?? apresenta esse histograma. Nele observamos que a maior parte dos dados concentra-se entre erro de 0.0 Km e 0.1 Km. Corroborando assim com a hipótese de que as métricas se comportariam melhor com a retirada dos outliers. Em relação as APIs, nessa faixa de erro a TomTom tem um comportamento melhor, já que a curva está mais estreita e próximo de 0. Porém a diferença entre as APIs não é grande, fazendo com que essa diferença não seja tão significativa.

De forma geral, apesar do histograma ser uma ferramenta poderosa para análise da distribuição do erro, nesse caso ela não foi tão eficiente por apresentar limitações na presença de valores excessivamente altos.

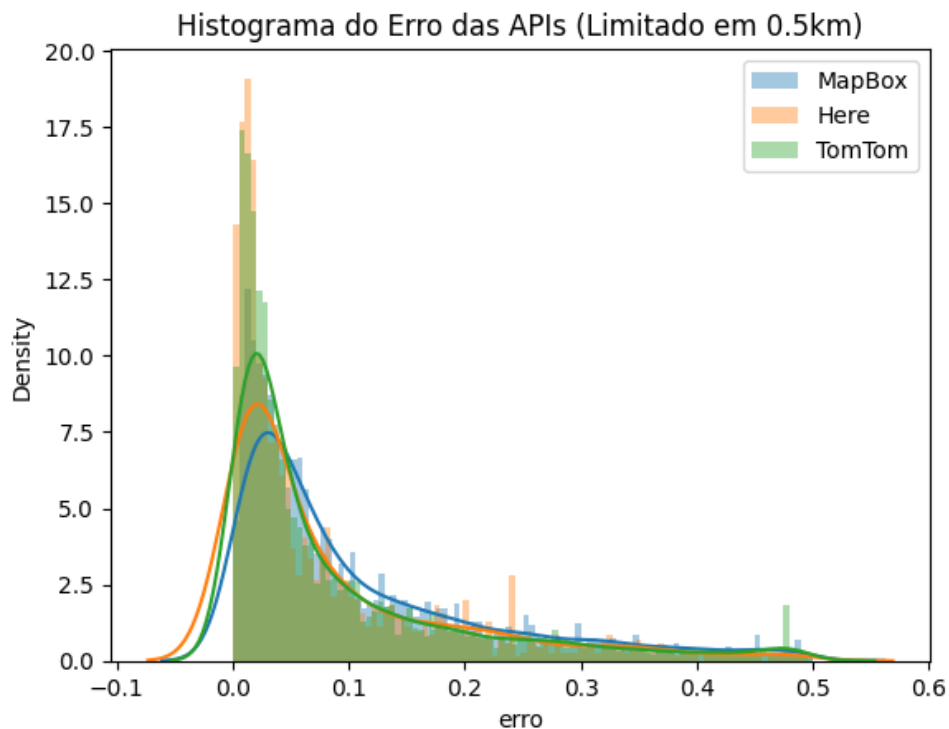


Figura 3.5 – Histograma comparativo do erro das APIs limitado em 500 metros

3.4 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 utilizado como medida de altitude. O mapa consiste em linhas de altitude, criadas a partir da interpolação do erro dos pontos. Em seguida essas linhas foram 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. Dessa forma, pudemos verificar se uma determinada área apresenta muitos pontos geocodificados ou se há poucos pontos com erros grandes. Todo esse processo foi realizado utilizando as bibliotecas do python matplotlib, scipy e folium para a criação e visualização dos gráficos; e as bibliotecas pandas e numpy para manipulação de estruturas de dados utilizadas.

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 ?? 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 ente 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. Porém, se pode observar que há uma baixíssima concentração de pontos, o que indica que existem poucos pontos com erro baixo, que causaram essa visualização. No centro, existem alguns pontos avermelhados que possuem uma grande concentração de pontos, esse é um dado preocupante pois indica que a API realmente está errando bastante em relação aos dados referência naquela região.

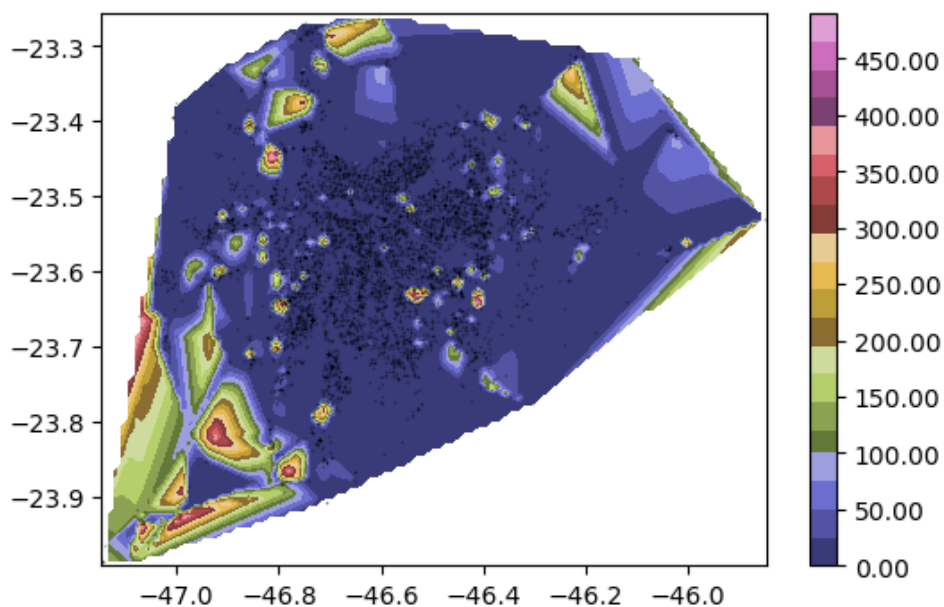


Figura 3.6 – Gráfico de altitude do erro (km) da geocodificação da Mapbox.

Já a ?? 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 só confirma isso. Sendo assim, qualquer análise realizada será inviesada. É possível observar uma grande concentração de azul, ou seja, os dados tem erro pequeno. Tem alguns picos, com erro elevado, porém somente um apresenta pontos suficientes para considerar que a região tem erro alto.

A ?? mostra é similar a ?? 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. Porém possui alguns picos escverdeados 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 grandes. Já no centro da figura, onde há grande concentração de pontos, esses picos também aparecem, o que indica um erro alto nessa região.

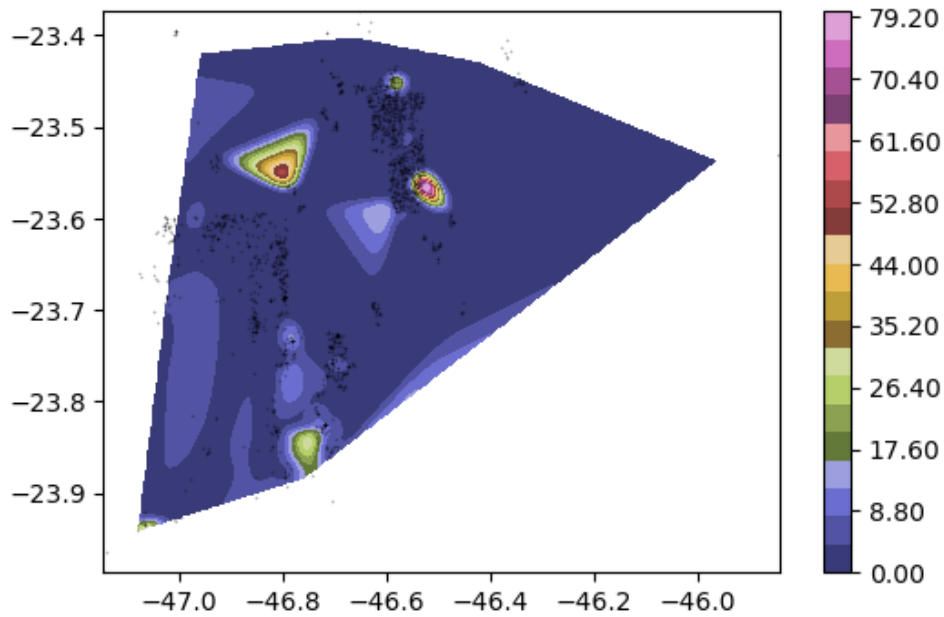


Figura 3.7 – Gráfico de altitude do erro (km) da geocodificação da Here.

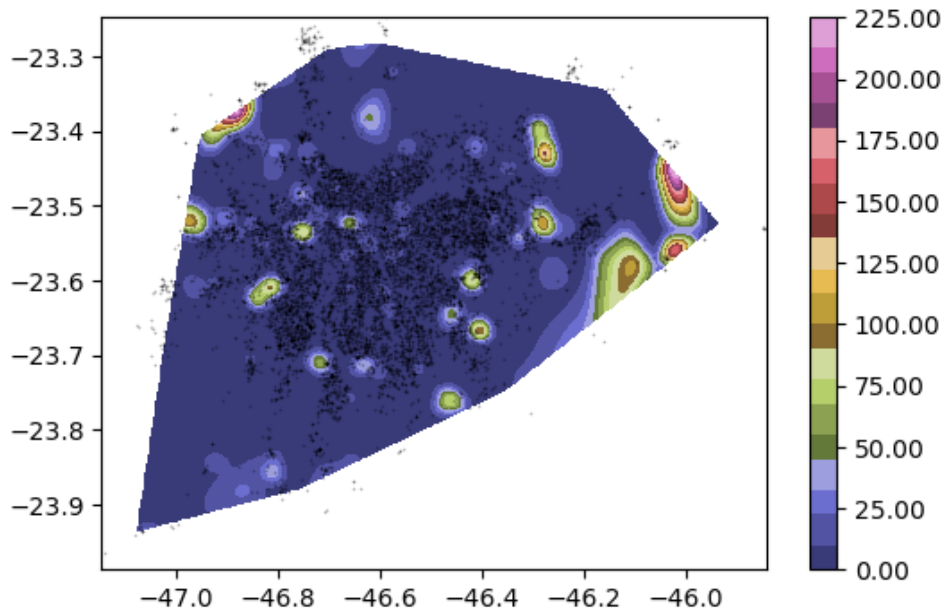


Figura 3.8 – Gráfico de altitude do erro (km) da geocodificação da TomTom.

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 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 0km a 50km para esses dados. No segundo gráfico há uma variação maior do erro, mas essa variação não reflete na variação da da covariância. Os dados indicam que essa relação não existe, porém não é possível tirar tal conclusão devido a quantidade baixa de dados.

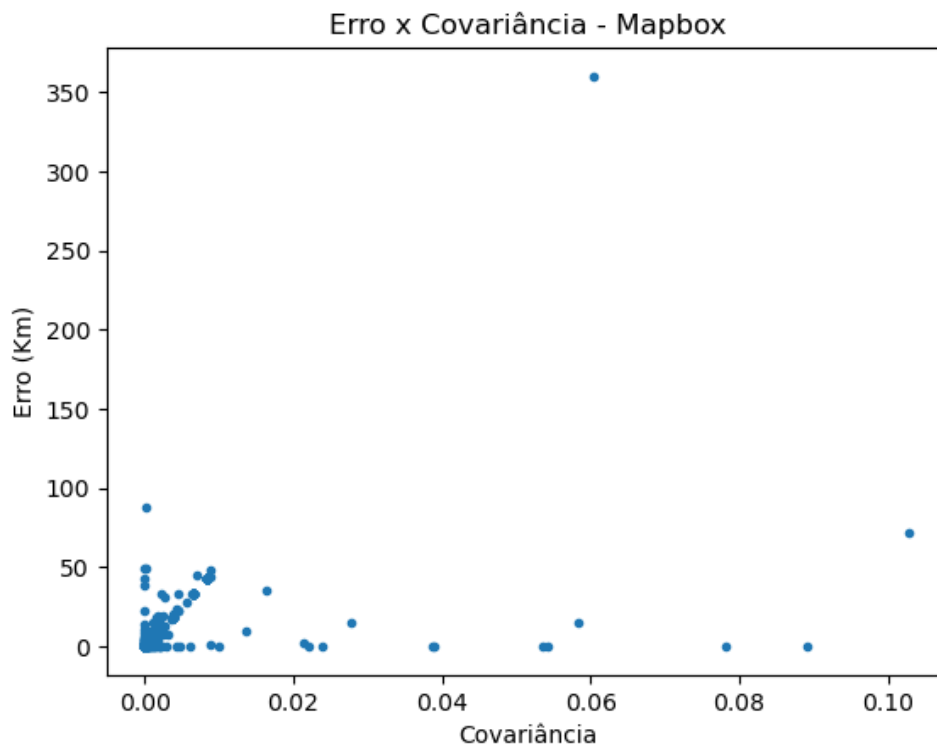


Figura 3.9 – Gráfico covariância por erro da Mapbox

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

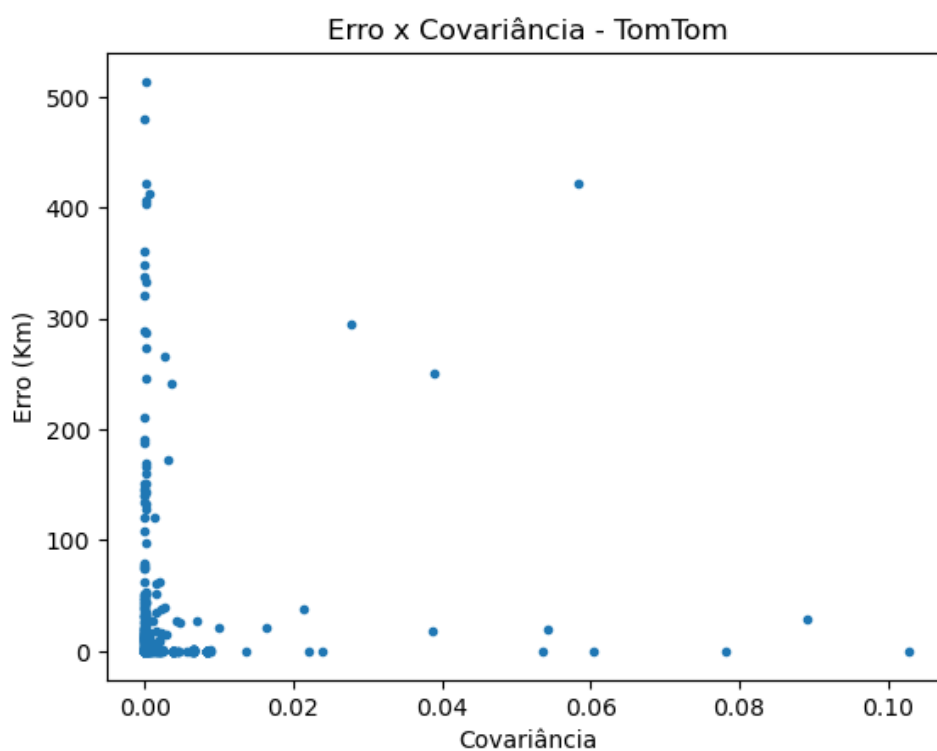


Figura 3.10 – Gráfico covariância por erro da TomTom

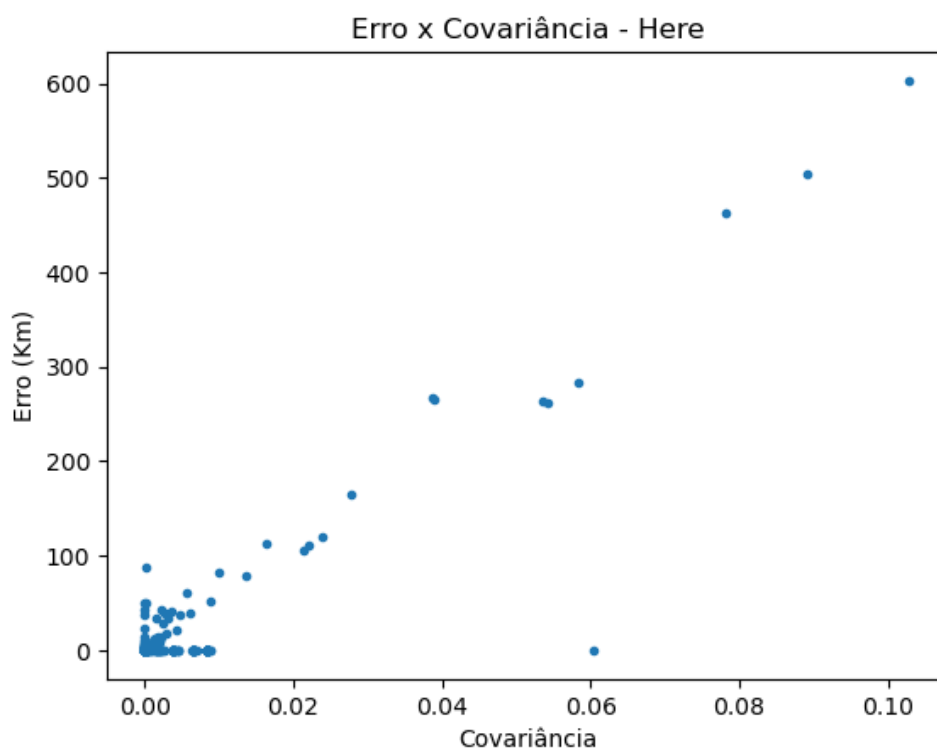


Figura 3.11 – Gráfico covariância por erro da Here

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. Porém, quaisquer conclusões tiradas a partir desse estudo são inviesadas a partir do momento em que não temos dados o suficiente e estes são dados específicos. Há de se lembrar que a base de dados possui apenas endereços de escola, não tenho uma diversidade de imóveis, localizados na região metropolitana de São de Paulo, não tendo uma diversidade de localidades. 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 repetindo o experimento possamos ver realmente o comportamento do erro e poderemos comparar os resultados com APIs já consolidadas na academia, a exemplo do Google Maps. Além disso, iremos fazer toda a análise para uma amostra significativa da base de dados da (??). A amostra conta com 85mil endereços distribuídos no espaço. A expectativa é que com a maior quantidade de endereços poderemos vizualizar o comportamento mais claramente. Em relação a análise de discrepância iremos acrescentar outra medida na análise, a distância para o ponto médio. Acreditamos que essa métrica é 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

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 Informacao 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>>.

OLLIGSCHLAEGGER, 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.