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

AVALIAÇÃO DE INTERFACES DE PROGRAMAÇÃO DE APLICAÇÃO DE GEOCODIFICAÇÃO EM GRANDES CIDADES BRASILEIRAS

Ouro Preto, MG
2023

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

ANA LUIZA ALMEIDA SOARES

AVALIAÇÃO DE INTERFACES DE PROGRAMAÇÃO DE APLICAÇÃO DE GEOCODIFICAÇÃO EM GRANDES CIDADES BRASILEIRAS

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

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

Online geocoding APIs play a significant role in applications that require location information. To ensure the quality of these applications, it is essential to assess the accuracy of the APIs used. This study aims to evaluate the quality of five geocoding APIs implemented in TerraLAB: Google Maps, Mapbox, TomTom, Here, and Open Route Service (ORS). The evaluation was conducted based on geocoding error compared to a reference database in the metropolitan region of São Paulo.

We used various metrics for comparative analysis, including mean, standard deviation, median, trimmed mean at 5%, response rate (the ratio of geocoding requests to responses), and accuracy rate (the number of addresses with errors less than 150 meters). Additionally, we conducted a spatial analysis of the error and investigated the relationship between discrepancy and error using the covariance measure. Due to issues with the application collecting geocodings, this project phase focused only on the Mapbox, TomTom, and Here APIs, resulting in overall unsatisfactory performance.

Most APIs exhibited a low response rate, with the highest among them falling below 90%, which impacted the experiment's integrity. Regarding the accuracy rate, all APIs obtained values considered unsatisfactory by our research team. Furthermore, we observed significant errors that hindered spatial analysis. Concerning the relationship between discrepancy and error, we could not identify a strong correlation, possibly due to the limited number of geocodings performed.

For the next phase of the project, we plan to repeat the analysis with the remaining APIs for São Paulo data and extend the evaluation to Belo Horizonte data.

Keywords: GeoAPIs. Quality

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
PLN	Processamento de Linguagem Natural
SIG	Sistema de Informação Geográfica
ORS	Open Route Service

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.
>
> (LONGLEY et al., 2013)

Em (LONGLEY et al., 2013), 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.

O endereço é a principal maneira de conceitualizar a localização no mundo atual(ZANDBERGEN, 2009). Isso ocorre devido ao fato de os endereços serem utilizados em diversas aplicações de diferentes áreas de estudo, como na saúde (KRIEGER et al., 2001; HAY et al., 2009; MAZUMDAR et al., 2008), nas ciências sociais (CHOW; LIN; CHAN, 2011), na análise criminal ou judiciária (OLLIGSCHLAEGER, 1998), na análise ambiental (GILBOA et al., 2006), na ciência da computação (ZANDBERGEN, 2009), na economia (WHITSEL et al., 2006) 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 é a utilização de coordenadas x e y em um plano, geralmente representando latitude e longitude. Esse processo de transformação de um endereço nessas coordenadas é chamado de Geocodificação ou Georreferenciamento e envolve três etapas (ZANDBERGEN, 2009):

- 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.

Além de representar um endereço computacionalmente, o georreferenciamento utilizando latitude e longitude oferece várias vantagens (LONGLEY et al., 2013):

- 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. Essas informações conflitantes são chamadas de "incertezas" (LONGLEY et al., 2013). 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 a referência 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.

Dados estes termos, podemos definir 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" (LONGLEY et al., 2013). 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 figura 1.1 apresenta uma visão conceitual da incerteza, onde cada processo muda um pouco a realidade, sendo assim a representação final tem sempre um nível de incerteza que está relacionado com o filtro aplicado em cada etapa. Por exemplo, a incerteza entre o mundo real e a concepção da realidade está relacionada ao filtro I1 que distorce a realidade para que seja possível a concepção. A partir desses conceitos, a incerteza foi aceita como uma métrica apropriada para avaliar a qualidade dos Sistemas de Informação Geográfica (SIG) (LONGLEY et al., 2013).

Apesar da incerteza ser uma métrica de importância significativa, sua mensuração é complexa. A incerteza envolve medidas que são subjetivas e podem variar de acordo com cada indivíduo avaliado.

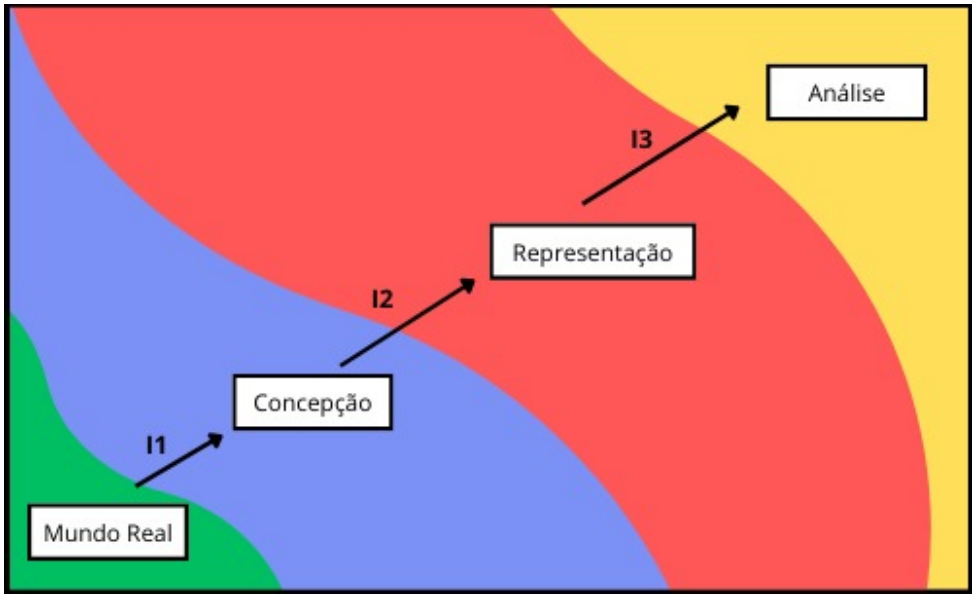


Figura 1.1 – Adaptada do livro (LONGLEY et al., 2013). Visão conceitual da incerteza

Diversas empresas e organizações de renome, como New York Times, CNN, BMW, Toyota, Strava, Microsoft, Uber, Fiat, Jeep, e TerraLAB, utilizam informações geográficas para o desenvolvimento de suas aplicações (SITE..., b; TERRALAB..., ; SITE..., a; SITE..., c). Essas aplicações 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 nesses locais. Para adquirir informações relacionadas a endereços, elas utilizam da geocodificação obtida por meio de APIs online.

1.2 APIs de Geocodificação e Análise de qualidade

Por muitos anos, a principal maneira de obter informações geográficas era através de software SIG. 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 (STEIN et al., 2021).

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 (CHOW; DEDE-BAMFO; DAHAL, 2016).

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. Em (CHOW;

DEDE-BAMFO; DAHAL, 2016), 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 (EUA), sendo que 78 destes são da região Central Texas, região considerada importante para o autor. O erro de cada endereço geocodificado foi calculado como a distância euclidiana de dois pontos, sendo eles, o ponto referência e o ponto obtido a partir da geocodificação.

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 mostrar evidência que as ferramentas online podem ser utilizadas como substitutas dos SIGs.

Apesar de (CHOW; DEDE-BAMFO; DAHAL, 2016) 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).

Outro estudo importante é (JR.; ALENCAR, 2011), que faz uma avaliação da qualidade da geocodificação da Google Maps API fornecida pela Google Cloud Platform (GOOGLE…,). 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 (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 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.

1.3 APIs de Geocodificação e formatação das entradas

A maioria das APIs possuem recomendações de formato de entrada que podem ser encontradas na documentação das mesmas.

Apesar das recomendações nas documentações das APIs, existem observações relacionadas à possibilidade de utilizar formatos de entrada diferentes dos apresentados, bem como a falta de informações abrangentes. Os geocodificadores das APIs são preparados para lidar com essas modificações, no entanto, a qualidade pode ser comprometida.

Com isso em mente, em trabalhos anteriores a equipe de análise de dados do TerraLAB (TERRALAB…,) conduziu uma série de experimentos para avaliar os impactos da modificação na ordem dos endereços de entrada nas APIs (RELATóRIO…,). Foram realizados 10 experimentos, nos quais o formato de entrada variou. Para avaliar a qualidade dos dados produzidos, foi utilizada a métrica de "dentro e fora da cidade". Se o endereço resultante estivesse dentro dos limites da cidade em questão, considerava-se que a API acertou naquele endereço; caso contrário, considerava-se um erro. No total, foram utilizados 100 endereços. Além disso, foram utilizadas as seguintes APIs: Mapbox, TomTom, Here e ORS. O trabalho concluiu que a maioria das APIs não apresenta uma diferença significativa ao mudar a formatação de entrada, exceto a API Mapbox, que apresentou uma melhora significativa para a formatação estado, cidade, rua e número.

Embora o trabalho apresente questionamentos importantes, ele possui uma série de limitações. A principal delas é a quantidade de endereços avaliados, que é insuficiente para gerar conclusões concretas. Além disso, apenas uma métrica foi avaliada. Dessa forma, não é possível determinar se a qualidade é realmente impactada pela formatação dos dados de entrada ou se isso impacta apenas essa métrica específica.

Encontrar a melhor forma de organizar a entrada para os geocodificadores é uma meta de diversos estudos. Em (Küçük Matci; AVDAN, 2018), é proposto um método de padronização da entrada que melhora o resultado da geocodificação. Para validar o método, eles utilizaram 233 endereços de escolas em Eskishehir, uma cidade da Turquia, para a qual as coordenadas corretas são conhecidas. Foi criado um dicionário contendo as principais abreviações e falhas de escrita, que foram utilizadas em métodos de processamento de linguagem natural (PLN). Os métodos de PLN foram empregados para gerar o endereço padronizado a partir do endereço inicial, de acordo com o dicionário. Além disso, foram utilizados dois geocodificadores, o ArcGIS e o Google Maps. A qualidade da geocodificação foi avaliada com base na distância euclidiana entre o ponto geocodificado e o ponto de referência, considerando acerto quando o erro foi menor que 100 metros.

O estudo demonstrou que o método de padronização reduziu significativamente o erro

da geocodificação e aumentou as taxas de acerto, com uma diferença variando de 6% a 20%, dependendo do formato e da API utilizados.

Apesar dos resultados significativos, o estudo apresenta algumas limitações. As principais são a quantidade de dados avaliados e o foco em uma cidade específica, não sendo possível, dessa forma, generalizar os resultados para além desse contexto.

A padronização do formato de endereços é um tópico de interesse para organizações em todo o mundo. Algumas organizações têm se esforçado para estabelecer um padrão no formato de endereços. A Organization for the Advancement of Structured Information Standards (OASIS) é mencionada como uma organização bem-sucedida no desenvolvimento de especificações que incluem a padronização de endereços. Essa padronização é utilizada no geocodificador do Google Maps (DOCUMENTAçãO…, a). No entanto, vale ressaltar que nem todas as organizações adotam o mesmo padrão e muitas delas mantêm suas próprias convenções de formatação (BEHR, 2010).

Todas essas considerações evidenciam que, apesar dos esforços em busca de um padrão na formatação de endereços com o objetivo de melhorar sua qualidade, ainda há muito a ser feito. Além disso, a padronização é influenciada por diversos fatores, incluindo o geocodificador utilizado e a região geocodificada.

1.4 Objetivos

A avaliação de qualidade é uma frente crucial do presente trabalho. Em relação a ela, os principais problemas levantados foram a quantidade de dados utilizados, a quantidade de APIs avaliadas e as regiões abarcadas pela análise. O presente trabalho busca abordar essas limitações ao conduzir a análise em uma região diferente do mundo, com ênfase no Brasil, e ampliando a quantidade de dados avaliados. Avaliaremos quatro APIs de geocodificação de grande impacto no mercado, utilizando duas bases de dados extensas. Além disso, nosso estudo incluirá a cidade de Belo Horizonte e a região metropolitana de São Paulo, proporcionando uma maior diversidade regional à análise. Dessa forma, pretendemos oferecer uma avaliação mais abrangente e representativa das ferramentas de geocodificação online (GeoAPIs).

Além disso, buscamos ir além por meio de duas abordagens distintas.

A primeira consiste em investigar se existe alguma métrica que poderia substituir o erro. Em outras palavras, buscamos identificar se há alguma medida que esteja correlacionada com o erro, de modo que possamos utilizá-la como alternativa à mensuração do erro em si.

Conscientes de que a obtenção do erro requer um valor de referência, considerado suficientemente preciso para calcular o erro, reconhecemos que a aquisição de informações geográficas de alta qualidade é uma tarefa desafiadora. A forma mais confiável de reduzir a incerteza é a coleta de dados in loco, com a utilização de dispositivos GPS. Por outro lado,

medidas de discrepância dependem apenas dos valores que estão sendo avaliados. Para entender melhor essa afirmação, precisamos entender o que é discrepância no conceito geral e como ela é aplicada na nossa pesquisa. Discrepância é o mesmo que discordância ou desigualdade (KLEIN, 2015). Sendo assim, a discrepância reflete o desacordo entre duas ou mais coisas. No contexto da pesquisa, tratamos como discrepância as diferenças de informações entre as APIs. Então, as medidas de discrepância são aquelas que medem de alguma forma essa diferença. Dessa forma, temos medidas que necessitam apenas das informações geradas pelas APIs, facitando a obtenção das métricas em relação ao erro.

Com acesso às medidas de discrepância e o erro, um dos nosso objetivos é verificar se existe alguma relação entre eles. Correlação é definida como: "grau de relação entre as variáveis, que procura determinar quão bem uma equação linear, ou de outra espécie, descreve ou explica a relação entre as variáveis"(SPIEGEL; STEPHENS, 2009).

A segunda abordagem visa compreender as causas do erro e identificar a melhor forma de configurar as entradas nas APIs a fim de minimizá-lo. Como abordado anteriormente, a melhor forma de entrada para as APIs ainda é um estudo em aberto. A qualidade dos dados depende do geocodificador utilizado e da região avaliada. Nesse contexto, o presente trabalho tem como propósito propor e avaliar formatos de padronização que sejam aplicáveis à região do Brasil e aos geocodificadores utilizados no Laboratório TerraLAB.

Dado o contexto, o principal objetivo deste trabalho é avaliar o erro, a discrepância e a acurácia de quatro 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) e Mapbox. 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 discrepância entre as APIs está relacionada aos erros?
- Alguma formatação da entrada contribui para a diminuição do erro?

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

- Coletar bases de dados padrão-ouro;

- Calcular o erro;
- Analisar a distribuição espacial e de valores do erro;
- Calcular as medidas de discrepância nas bases escolhidas;
- Avaliar a distribuição dos valores das medidas de discrepância;
- Verificar se existem correlações entre as medidas de discrepância e o erro;
- Avaliar a distribuição espacial das medidas de discrepância.
- Avaliar para cada API qual formatação atinge os melhores resutados

2 Bases de Dados e Métodos de Geocodificação e Avaliação

Para atingir os objetivos citados, 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) (CENTRO…,). 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 2.1 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.

Apesar de o mapa de cluster ser uma representação útil, resolvemos criar um gráfico de pontos para melhor visualizar a distribuição dos dados de São Paulo e também para comparar

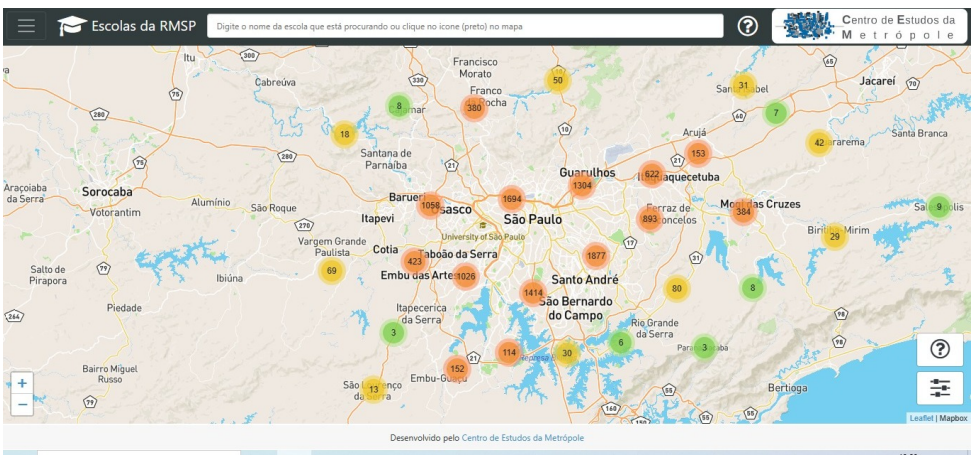


Figura 2.1 – Mapa de clusters do Centro de Estudos da Metrópole

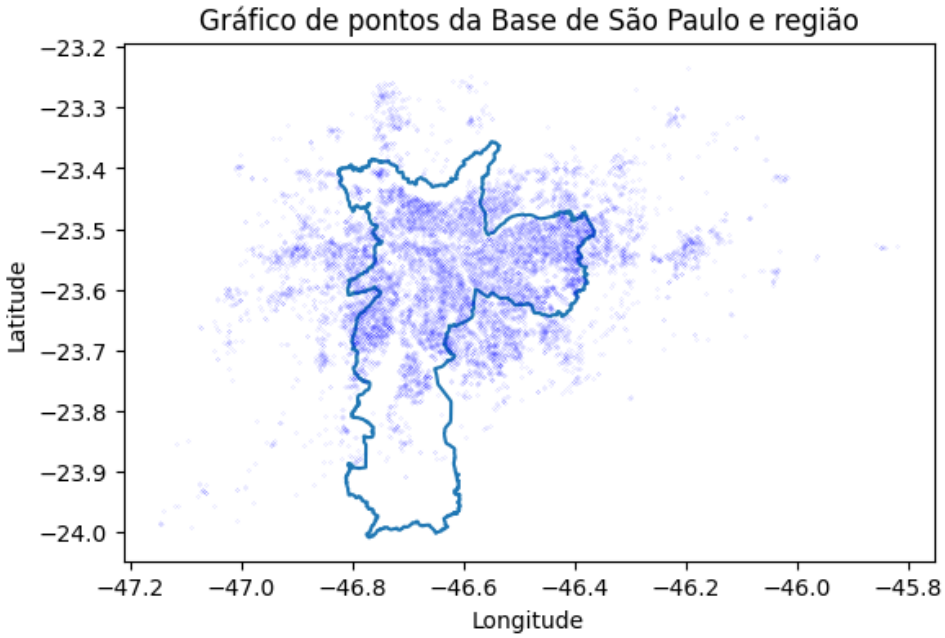


Figura 2.2 – Gráfico de Endereços da Base de São Paulo e região metropolitana

de forma equivalente com a segunda base utilizada. A Figura 2.2 mostra o gráfico citado. No gráfico também foi incluído o contorno da cidade de São Paulo para vizualizar a distribuição de endereços dentro e fora da cidade.

A segunda base de dados coletada foi fornecida pela (PRODABEL,), a empresa de informática e informação da prefeitura de Belo Horizonte. A descoberta dessa base de dados foi possibilitada pelo artigo de referência (JR.; ALENCAR, 2011). Essa base de dados era mantida e atualizada mensalmente por 27 empresas, tanto públicas quanto privadas, de Belo Horizonte. Essas empresas tinham a responsabilidade de relatar quaisquer inconsistências encontradas na base e de fornecer novos dados à medida que os adquiriam. Ela é considerada uma fonte confiável de informações, pois estava em constante atualização e era amplamente utilizada por diversos serviços da prefeitura. Um exemplo notável era o uso da base para georreferenciamento na distribuição de alunos da rede pública. Esse serviço consiste em designar a escola pública para qual aluno irá com base na distância entre a moradia do aluno e a escola. Essa base então é utilizada para selecionar escolas para todos os alunos de forma a diminuir as distâncias entre a escola e os alunos para cada um dos alunos. Sendo assim, era uma base bastante relevante para a cidade de Belo Horizonte (JR.; ALENCAR, 2011).

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 2.3 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.

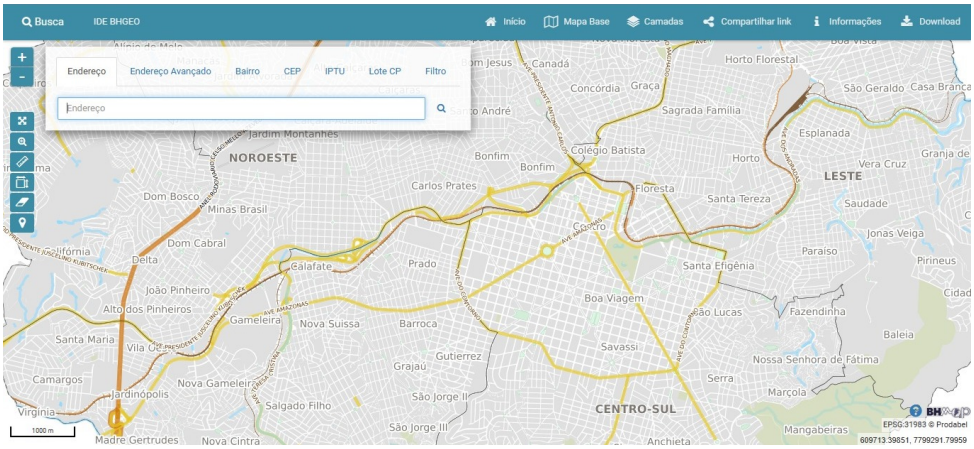
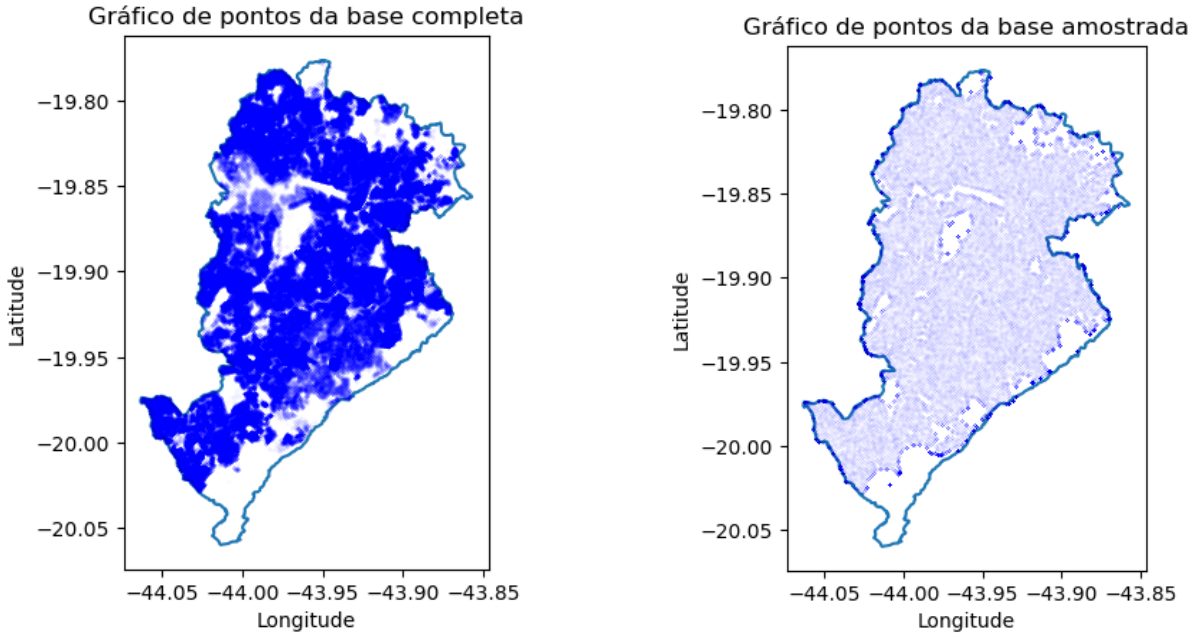


Figura 2.3 – Site da Prodabel para pesquisa de endereços.

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 2.4 apresenta dois gráficos contendo os pontos da base original e os da amostra obtida. Os gráficos contém os pontos referentes a cada uma das bases e um contorno da cidade de Belo Horizonte. Com essa vizualização, é possível ver a concentração e cobertura dos pontos. É possível observar que a amostra cobre toda a área abrangida apesar de não ter tanta concentração de pontos quanto a base original. Além disso, verifica-se uma ligeira concentração nas regiões periféricas do desenho, permitindo uma melhor delimitação da cidade.



(a) Completa (b) Amostra

Figura 2.4 – Gráficos dos endereços da Base de Belo Horizonte e amostragem obtida

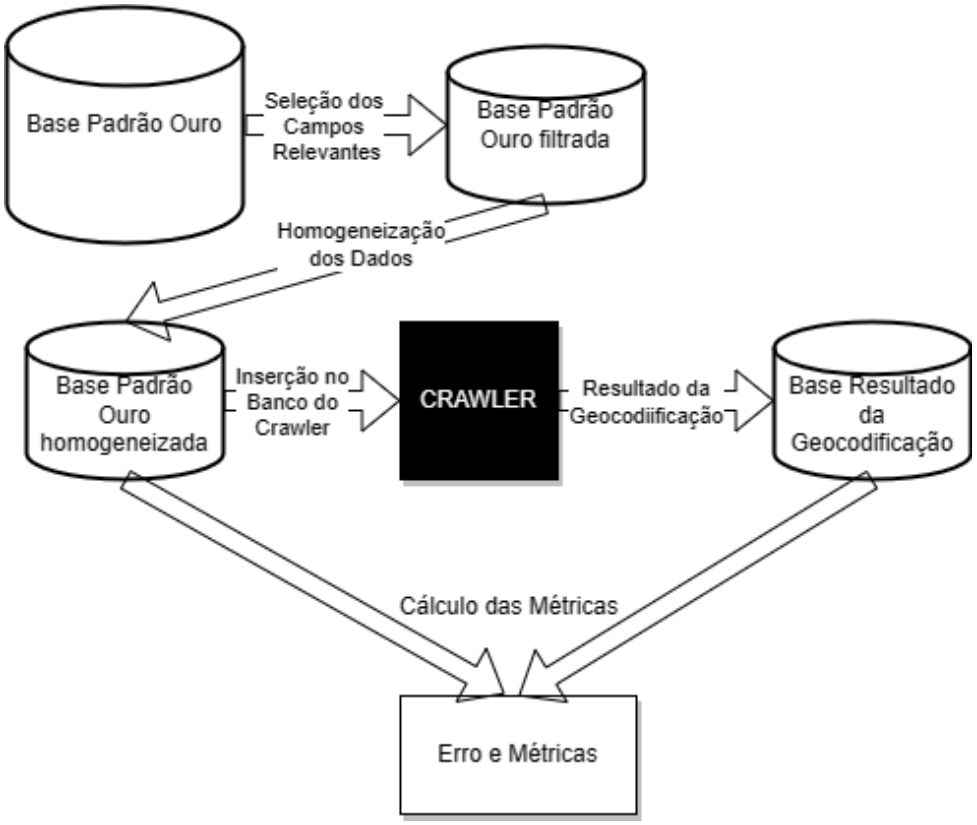


Figura 2.5 – Esquematização do processo de preparação e geocodificação dos dados

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, os dados foram homogeneizados, substituindo abreviações comuns por suas formas completas correspondentes, e todas as letras foram transformadas em letras maiúsculas. Esta etapa é conduzida pela equipe do TerraLAB e demonstrou-se que as APIs respondem de forma mais eficaz quando não há abreviações e as palavras estão escritas em maiúsculo.

Para realizar a geocodificação, os endereços previamente preparados são inseridos no banco de dados da aplicação responsável por solicitar e coletar informações de geocodificação. Os endereços são então retirados desse banco 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 2.5 esquematiza todo esse processo essencial para o nosso trabalho.

2.3 Método de Avaliação

2.3.1 Erro, Taxa de Resposta e Discrepância

A principal métrica utilizada para avaliar a qualidade da geocodificação é o erro do endereço. Com base nesse erro, calcularemos as medidas estatísticas média, mediana, desvio padrão 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 as equações abaixo:

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

Onde:

- e é o erro da geocodificação,
- D é uma função que calcula a distância em quilômetros,
- p_{Ouro} é o ponto da base padrão ouro, e item p_{Geo} é o ponto resultante da geocodificação.

$$D(p_1(\text{lat}_1, \text{lon}_1), p_2(\text{lat}_2, \text{lon}_2)) = \sqrt{(\text{lat}_2 - \text{lat}_1)^2 + (\text{lon}_2 - \text{lon}_1)^2} \tag{2.2}$$

Onde:

- D é a distância euclidiana entre dois pontos,
- p_1 é o primeiro ponto,
- p_2 é o segundo ponto,
- lat_1 e lat_2 são as latitudes de p_1 e p_2, respectivamente,
- lon_1 e lon_2 são as longitudes de p_1 e p_2, respectivamente.

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:

$$tx_{\text{resposta}}(\%) = \left(\frac{qtd_{\text{Geo}}}{qtd_{\text{Ouro}}} \right) \times 100\% \tag{2.3}$$

Onde:

- tx_{resposta} é a taxa de resposta da API avaliada;
- qtd_{Ouro} é a quantidade de endereços da base referência;
- qtd_{Geo} é a quantidade de endereços resultantes da geocodificação.

Outra métrica obtida por meio do erro é a taxa de precisão. É definida como a porcentagem de endereços com um erro inferior a 150 metros em relação ao número total de endereços. Esse valor foi escolhido com base na afirmação do artigo (JR.; ALENCAR, 2011), onde os autores colocaram que 150 metros é aproximadamente o tamanho de um quarteirão em Belo Horizonte. É representada pela seguinte fórmula:

$$tx_{\text{precisão}}(\%) = \left(\frac{qtd_{\text{certo}}}{qtd_{\text{Ouro}}}\right) \times 100\% \tag{2.4}$$

Onde:

- $tx_{\text{precisão}}$ é a taxa de precisão da API avaliada;
- qtd_{certo} é a quantide de endereços em que o erro foi menor que 150 metros;
- qtd_{Ouro} é a quantidade de endereços da base referência;

Para avaliar a discrepância entre geocodificações, duas métricas serão utilizadas: covariância e distância até o ponto médio. A covariância é o desvio padrão dividido pela média. Em nosso estudo, calculamos a covariância para latitude e longitude, considerando apenas os pontos que possuíam informações de todos os serviços de geocodificação. Em seguida, fazemos a média desses dois valores para representar a covariância do endereço. Na distância até o ponto médio, computamos um ponto médio a partir de todas as geocodificações fornecidas para um endereço e então calculamos a distância euclidiana de uma geocodificação até esse ponto médio.

A covariância é definida como:

$$\text{Covariância} = \frac{1}{n}\sum_{i=1}^{n}(\text{lat}_i - \bar{\text{lat}})(\text{lon}_i - \bar{\text{lon}}) \tag{2.5}$$

Onde:

- n é o número de APIs.
- lat_i e lon_i são as latitudes e longitudes de cada API no mesmo endereço, respectivamente.
- $\bar{\text{lat}}$ e $\bar{\text{lon}}$ são as médias das latitudes e longitudes.

A distância até o ponto médio é calculada como a distância euclidiana de uma geocodificação até o ponto médio e é definida como:

$$\text{Distância até o Ponto Médio para cada ponto} = \sqrt{(\text{lat}_i - \text{mid\_lat})^2 + (\text{lon}_i - \text{mid\_lon})^2} \tag{2.6}$$

Onde:

- lat_i e lon_i são as latitudes e longitudes das geocodificações individuais,
- mid_lat e mid_lon são a latitude e longitude do ponto médio calculado.

Para calcular o ponto médio entre os pontos do mesmo endereço de várias APIs, usamos a seguinte fórmula:

$$\text{Ponto Médio} = \left(\frac{\sum_{i=1}^{n} \bar{\text{lat}}_i}{n}, \frac{\sum_{i=1}^{n} \bar{\text{lon}}_i}{n} \right) \tag{2.7}$$

Onde:

- $\bar{\text{lat}}_i$ e $\bar{\text{lon}}_i$ representam a média de latitude e longitude para os mesmos endereços,
- n representa o número de APIs.

Para calcular a relação entre a discrepância e erro utilizaremos uma media de correlação e a medida escolhida foi a correlação de Pearson. A correlação de Pearson é uma medida de correlação descrita pela fórmula (CALLEGARI-JACQUES, 2007):

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

onde:

- r é o coeficiente de correlação de Pearson;
- x_i e y_i são as variáveis avaliadas;
- $\bar{x}$ e $\bar{y}$ são as médias dos valores x e y, respectivamente.

Essa medida, independente dos valores das variáveis sempre retorna um valor entre -1 e 1, sendo 1 uma correlação forte positiva e -1 uma correlação forte negativa. Quanto mais próximo de 0, menos correlação as variáveis tem. A tabela 2.1 retirada do livro mostra para cada faixa de valor resultante da correlação de Pearson respectivos significados (CALLEGARI-JACQUES, 2007). É interessante portanto encontrar relações com módulo do coeficiente superior a 0.6.

Tabela 2.1 – Tabela de Correlação de Pearson

$\|r\|$	A correlação é dita
0	Nula
$0 - 0.3$	Fraca
$0.3 - 0.6$	Regular
$0.6 - 0.9$	Forte
$0.9 - 1$	Muito Forte
1	Plena ou Perfeita

2.4 Experimentos para avaliação da formatação da entrada

Como mencionado anteriormente, algumas APIs disponilizam recomendações de formatações de entrada em suas documentações. A Tabela 2.2 apresenta o formato recomendado para cada uma das APIs utilizadas, enquanto a Tabela 2.3 especifica os formatos citados.

Tabela 2.2 – Formato Recomendado de Entrada para APIs de Geocodificação

API	Formato Recomendado	Documentação
Google Maps	Recomenda utilizar o formato do serviço postal do país buscado	(DOCUMENTAçãO..., a)
Open Route Service	Sem recomendações específicas	(DOCUMENTAçãO..., c)
Mapbox	Recomenda utilizar o formato oficial dos EUA ou o formato do serviço postal do país buscado	(DOCUMENTAçãO..., b)
TomTom	Sem recomendações específicas	(DOCUMENTAçãO..., d)

Tabela 2.3 – Descrição dos formatos

Origem	Formato
Serviço postal do Brasil	Tipo de Logradouro, Nome do Logradouro, Número do Lote, Complemento (se houver), Nome do Bairro, Nome da Localidade, Sigla da Unidade da Federação, CEP
EUA	Número do lote, Nome do Logradouro Nome da Cidade, Nome do Estado, CEP

Apesar disso, a equipe do TerraLAB tem seus próprio padrão de formatação utilizado. A Tabela 2.4 apresenta os formatos de entrada utilizados por cada uma das APIs no laboratório.

Para avaliar qual seria a melhor formatação dos dados de entrada, contruímos para cada API 5 experimentos onde são modificadas as ordens da palavra de entrada da API. A tabela 2.5 mostra qual é o formato para cada um dos experimentos. Os experimentos foram numerados, durante o trabalho iremos nos referir a eles de acordo com esse número.

Tabela 2.4 – Formato de Entrada das APIs Utilizadas pelo TerraLAB

API	Formato
Mapbox	Estado, Cidade, Número Lote, Tipo Logradouro, Nome Logradouro
TomTom	Tipo Logradouro, Nome Logradouro, Número do Lote, Cidade, Estado
Google	Tipo Logradouro, Nome Logradouro, Número do Lote, Cidade, Estado
ORS	Tipo Logradouro, Nome Logradouro, Número do Lote, Cidade, Estado

Tabela 2.5 – Formato de cada experimento

Experimento	Formato
1	Tipo Logradouro, Nome Logradouro, Número Edifício, Cidade, Estado
2	Cidade, Tipo Logradouro, Nome Logradouro, Número Edifício, Estado
3	Estado, Cidade, Tipo Logradouro, Nome Logradouro, Número Edifício
4	Estado, Tipo Logradouro, Nome Logradouro, Número Edifício, Cidade
5	Cidade, Estado, Tipo Logradouro, Nome Logradouro, Número Edifício

Devido a quantidade de experimentos, resolvemos fazer uma amostragem das duas bases de dados. Selecionamos 5 mil endereços de cada base, utilizando o método de hipercubo latino. E utilizando o mesmo processo de padronização e geocodificação explicado na seção 2.2, obtivemos a geocodificação de cada uma das APIs para cada endereço.

A base de Belo Horizonte continha as informações de bairro de cada endereço, diferente da base de São Paulo. Com isso, decidimos realizar uma análise adicional, que consistia em verificar se há algum ganho em adicionar o bairro na entrada. Portanto, para a base de Belo Horizonte forma adicionados 5 experimentos adicionais. A tabela 2.6 mostra os formatos citados.

Tabela 2.6 – Formato dos experimentos adicionais

Experimento	Formato
1b	Tipo Logradouro, Nome Logradouro, Número Edifício, Bairro, Cidade, Estado
2b	Cidade, Tipo Logradouro, Nome Logradouro, Número Edifício, Bairro, Estado
3b	Estado, Cidade, Bairro, Tipo Logradouro, Nome Logradouro, Número Edifício
4b	Estado, Tipo Logradouro, Nome Logradouro, Número Edifício, Bairro, Cidade
5b	Cidade, Estado, Bairro, Tipo Logradouro, Nome Logradouro, Número Edifício

Com as Geocoficações prontas, faremos o calcúlo das metricas de erro para cada uma das APIs, como explicado na seção 2.3.

3 Resultados

Nós inicialmente analisamos erro e discrepância nos 12.502 endereços de São Paulo. Abaixo estão os totais de geocodificações bem-sucedidas para cada API:

- TomTom: 11.370 endereços;
- Google Maps: 9.389 endereços;
- Mapbox: 12.260 endereços;
- ORS: 12.295 endereços.

Então, conduzimos experimentos com o conjunto de dados de Belo Horizonte. Em relação ao número de endereços retornados, as APIs retornaram da base de 85.000:

- TomTom: 84.981 endereços;
- Google: 84.941 endereços;
- Mapbox: 84.966 endereços;
- ORS: 84.864 endereços.

Falhas podem ocorrer em qualquer estágio de geocodificação, derivadas de informações incompletas ou ambíguas fornecidas para geocodificação ou dos algoritmos empregados (algoritmos de seleção e classificação de candidatos). Quando falhas ocorrem, a API retorna apenas uma mensagem de erro. Nas próximas seções, os resultados de erro e discrepância são cuidadosamente analisados.

3.1 Erro, Taxa de Resposta e Taxa de Precisão

A próxima etapa foi o calculo 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. Os resultados bem como suas interpretações são apresentados abaixo.

A Tabela 3.1 apresenta os resultados calculados para as respostas recebidas da geocodificação da base de São Paulo. Em relação à taxa de resposta, ou seja, o número de endereços que foram geocodificados com sucesso, a Mapbox obteve o melhor resultado, seguida pela ORS, ambas com taxas de resposta superiores a 98%. Google e TomTom tiveram taxas de resposta de

75% e 90%, respectivamente. Esses resultados são considerados satisfatórios e garantem uma boa quantidade de dados para as avaliações subsequentes.

Tabela 3.1 – Métricas de Erro para São Paulo

API	Média (km)	Mediana (km)	Desvio Padrão
Mapbox	15,3504	0,1675	83,9394
Google Maps	2,0965	0,0555	22,0156
TomTom	10,2074	0,0638	88,0844
ORS	33,9474	1,2984	103,0119
API	Média Aparada em 5% (km)	Taxa de Resposta (%)	Taxa de Precisão (%)
Mapbox	3,5009	98,0565	46,5968
Google Maps	0,2327	75,0940	52,2675
TomTom	0,4768	90,9382	60,3055
ORS	16,4096	98,3364	28,6091

Outra métrica importante é a taxa de precisão. Endereços com erros menores que 150 metros (0,15 km) foram considerados precisos. A taxa de precisão foi baixa para a maioria das APIs. A API TomTom teve a maior taxa de precisão, com 60% de acurácia.

O erro médio foi bastante elevado, variando de 2 km a 33 km. O desvio padrão também foi alto, indicando uma variação considerável no erro. No entanto, a mediana foi bastante baixa, alcançando resultados desejáveis em nossa pesquisa. A média aparada produziu resultados muito bons, indicando a presença de um número significativo de valores atípicos.

Da mesma forma, calculamos o erro para cada ponto geocodificado no banco de dados de Belo Horizonte e computamos as métricas mencionadas anteriormente. A Tabela 3.2 exibe esses resultados.

Tabela 3.2 – Métricas de Erro para Belo Horizonte

API	Média (km)	Mediana (km)	Desvio Padrão
Mapbox	3,2857	0,0001	24,7587
Google Maps	2,4924	0,0098	5,8465
TomTom	11,2913	0,1147	56,6424
ORS	6,4828	7,5702	5,5364
API	Média Aparada em 5% (km)	Taxa de Resposta (%)	Taxa de Precisão (%)
Mapbox	1,0701	99,9600	76,8235
Google Maps	1,6146	99,9306	73,6118
TomTom	0,4768	99,9776	51,7988
ORS	6,2940	99,8400	25,1835

Em relação à taxa de resposta, todas as APIs tiveram excelentes resultados, com mais de 99% de resposta para o banco de dados fornecido. Este é um resultado significativo para a pesquisa, pois as conclusões são mais robustas devido à quantidade de dados analisados.

A taxa de precisão também mostrou resultados satisfatórios, com os melhores resultados vindos da Mapbox e Google Maps, com taxas superiores a 73%. Este resultado é bastante satisfatório e está alinhado com os resultados obtidos em (JR.; ALENCAR, 2011). No entanto, TomTom e ORS apresentaram baixas taxas de precisão, sendo que ORS teve uma taxa extremamente baixa de 25%. É importante observar que um resultado foi considerado preciso se o erro fosse menor ou igual a 150 metros.

O erro médio apresentou valores muito mais suaves do que os obtidos com o conjunto de dados de São Paulo, embora ainda estivessem elevados, variando de aproximadamente 2 a 11 quilômetros. Os valores medianos foram bastante baixos para a maioria das APIs, e o desvio padrão foi bastante alto. Esse resultado indica que também existem valores de erro muito altos nessa geocodificação. A API ORS apresentou resultados diferentes das outras APIs, com valores altos de média, mediana e desvio padrão, o que provavelmente explica a baixa taxa de precisão.

3.2 Distribuição de Erro

Com base nos resultados acima, realizamos uma análise da distribuição de erro para cada uma das GeoAPIs e bases. Para isso, utilizamos histogramas de erro individuais para cada API e os combinamos. As Figuras 3.2 e 3.1 mostram os histogramas para cada API e cada base utilizada. No entanto, devido à presença de alguns erros extremos, os histogramas gerais (que continham todo o conjunto de dados) não foram muito representativos, pois a maior parte do erro estava concentrada entre 0 km e 50 km, enquanto existiam erros bem maiores. Esse intervalo é considerado um erro muito grande, tornando desafiador tirar conclusões sólidas. Outra limitação dessa análise, é o fato de que cada API teve um máximo de erro diferente, prejudicando então a comparação entre APIs.

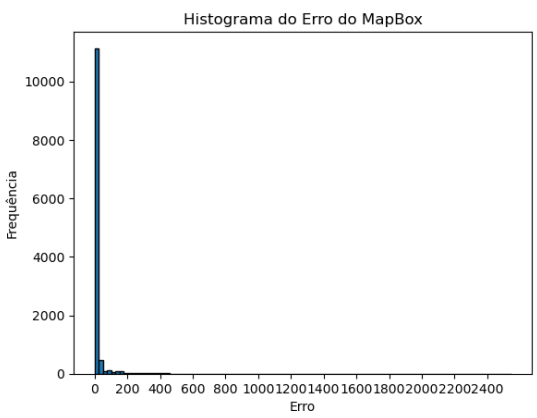
Portanto, decidimos cortar os dados, limitando o erro a 300 metros. Repetimos o processo, gerando um único histograma que representa a distribuição de erro para todas as APIs juntas, para cada uma das bases.

A figura 3.3 mostra o histograma resultante para os dados de São Paulo e a figura 3.4 mostra o histograma para os dados de Belo Horizonte.

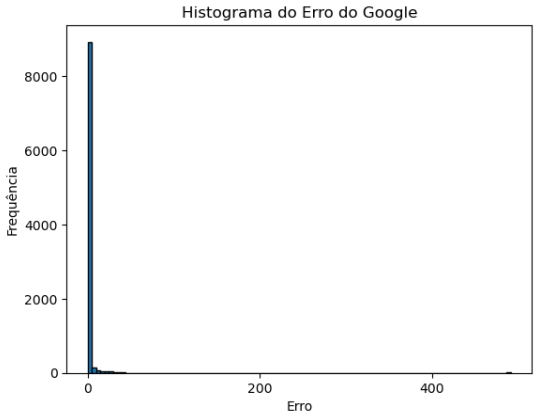
Em relação aos dados de São Paulo as APIs tiveram resultados similares nessa faixa de valores do erro. Porém as APIs Google Maps e TomTom se destacaram ao conter uma curva mais estreira, ou seja, os valores para essas APIs estão mais concentrados em erro menor que 50 metros.

Para os dados de Belo Horizonte, a API Mapbox teve melhores resultados com uma curva bem estreita. Seguida pela Google Maps, que apesar de ter uma curva bem estreita também, apresenta uma diferença significativa para a Mapbox. As outras APIs apresentam curvas mais largas e algo notável é a curva da ORS que está muito distribuída, tendo um aspecto parecido com uma reta em valores de erro superiores a 50 metros. Isso mostra que a ORS apresenta erro similar na maior parte da faixa, o que indica que ela não apresenta bons resultados nem quando há um corte nos dados.

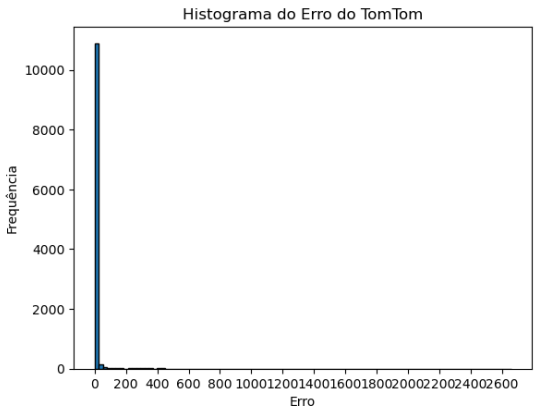
Em geral, embora os histogramas sejam uma ferramenta poderosa para analisar a distribuição de erros, neste caso, eles não se mostraram tão eficazes devido às limitações decorrentes da presença de valores excessivamente altos.



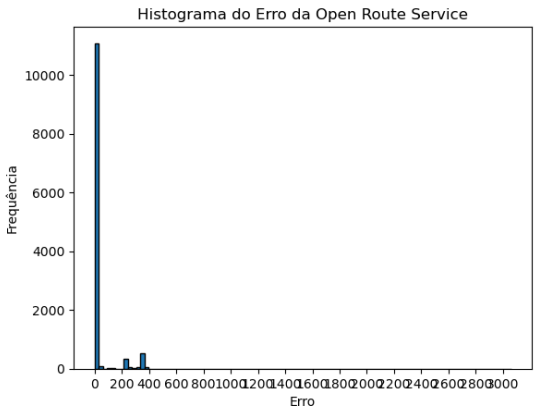
(a) Mapbox



(b) Google



(c) TomTom



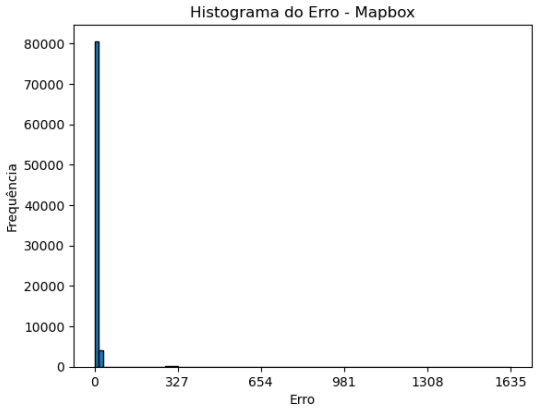
(d) ORS

Figura 3.1 – Histogramas do erro das 4 APIs para o todos os dados de São Paulo

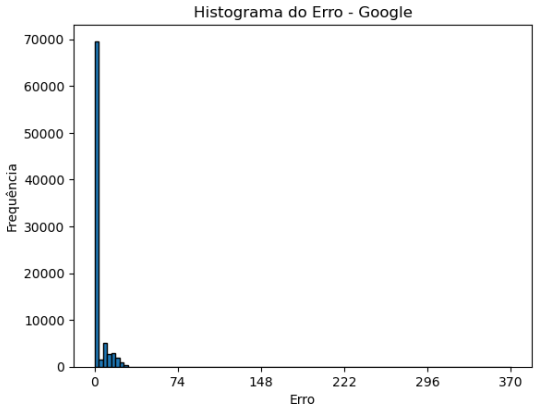
3.3 Distribuição Espacial do Erro

Além disso, realizamos uma análise adicional com o objetivo de verificar o comportamento do erro no espaço. Utilizamos gráficos de classificação hexagonal, empregando a função hexbin da biblioteca matplotlib, que desempenha um papel integral na construção do gráfico. Essa função automatiza o processo, dividindo o espaço em hexágonos de tamanhos uniformes e distribuídos de maneira equitativa. Em seguida, a função hexbin seleciona os pontos de dados contidos em cada hexágono e aplica uma função específica, que é definida como parâmetro da função hexbin. Essa função determina os cálculos realizados com base nos pontos, gerando um valor único. Esse valor é então atribuído ao hexágono correspondente no gráfico, e as cores são mapeadas de acordo com uma escala predefinida.

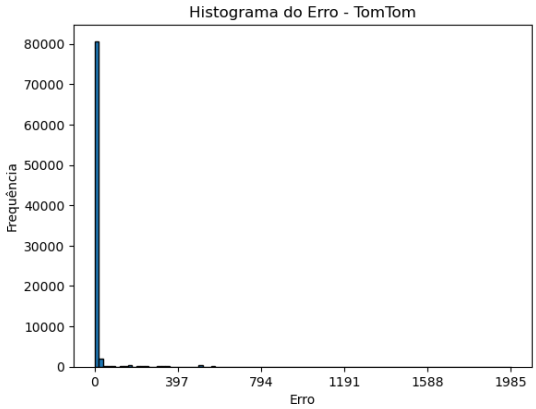
Para gerar a representação do gráfico introduzimos o conceito de "falha". Quando o erro em um ponto específico é igual ou inferior a 150 metros, atribuímos o valor 0 à falha; caso contrário, designamos o valor 1. A função escolhida para calcular o valor de cada hexágono é a média da falha dos pontos, resultando em uma representação em porcentagem decimal da falha naquela região. Assim, quanto mais escura a cor do gráfico, maior é a média sa falha observada. Para melhor vizualização, também adicionamos o limite da cidade como contorno do gráfico. As



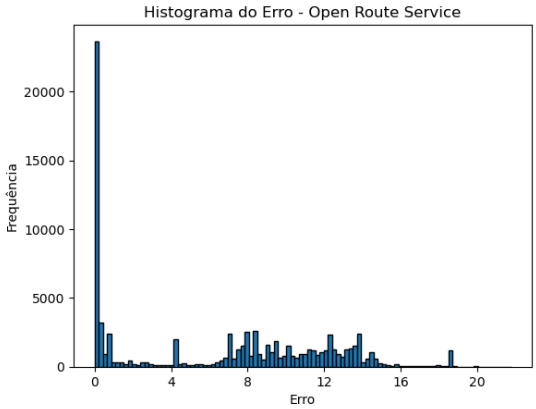
(a) Mapbox



(b) Google



(c) TomTom



(d) ORS

Figura 3.2 – Histogramas do erro das 4 APIs para o todos os dados de Belo Horizonte

figuras 3.5 e 3.6 apresentam os gráficos de falhas de cada uma das APIs para os dados de Belo Horizonte e São Paulo respectivamente.

Para os dados de Belo Horizonte é possível notar que a API com gráfico mais claro, ou seja, menos falhas, é a Mapbox, seguido pela Google Maps. Resultado que vai de encontro com os obtidos na tabela de métricas 3.2. Outra informação relevante que é possível observar é que em todas as APIs existe uma concentração maior de falhas próximo aos limites da cidade, como esperado. Outro ponto importante é o gráfico da ORS. A maior parte do gráfico para essa API apresenta cores bem escuras, indicando muitas falhas em toda região da cidade para essa API. Mais especificamente nas regiões superior e inferior do gráfico o valor chega próximo do limite máximo, o que indica que naquela região houve aproximadamente 100% de falha. Esse resultado, apesar de ruim, está de acordo com os análises sobre a ORS feitas anteriormente.

Nos gráficos de São Paulo também foi adicionado o contorno da cidade. No entanto, os dados são referentes a região metropolitana, incluindo outras cidades da região. Para essa base é possível notar que as APIs Google Maps e TomTom tem melhores resultados, o que confirma os resultados obtidos na tabela 3.1. Outro resultado notável é que nos outros municípios o resultado piora em todas as APIs, atingindo valores de falha muito próximo de 1. Por fim, as APIs que foram piores foram Mapbox e ORS. A ORS foi claramente pior, repetindo os resultado de Belo

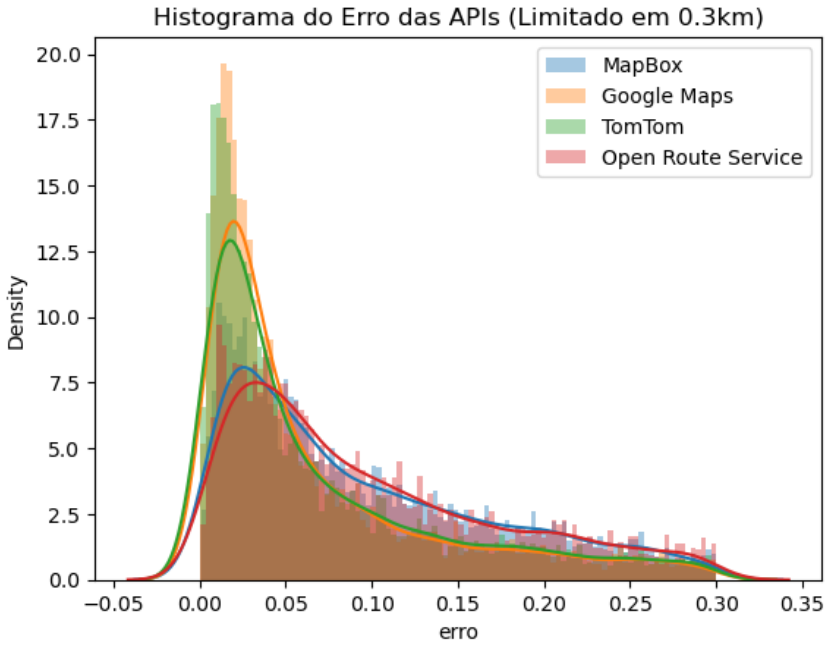


Figura 3.3 – Histograma comparativo de erro das APIs limitado em 300 metros para os dados de São Paulo

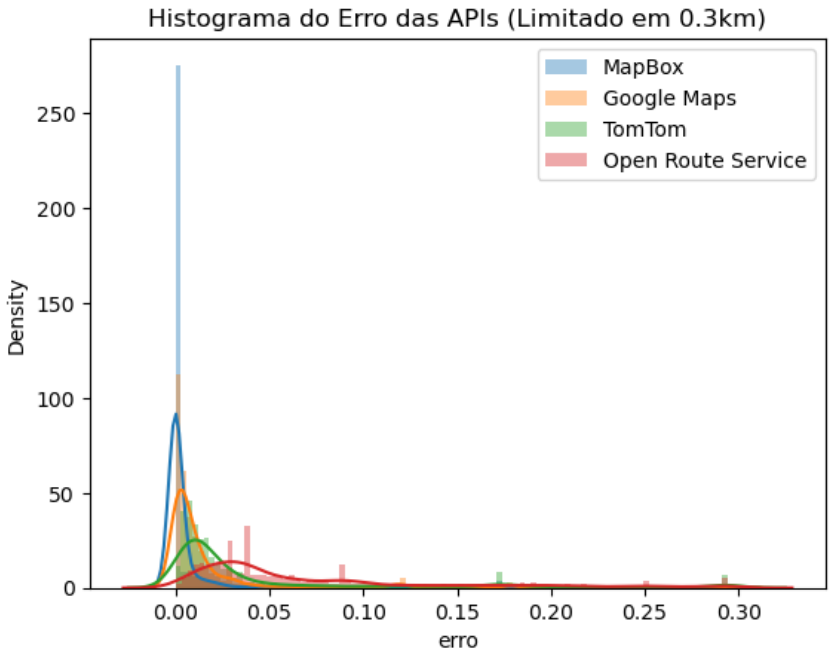
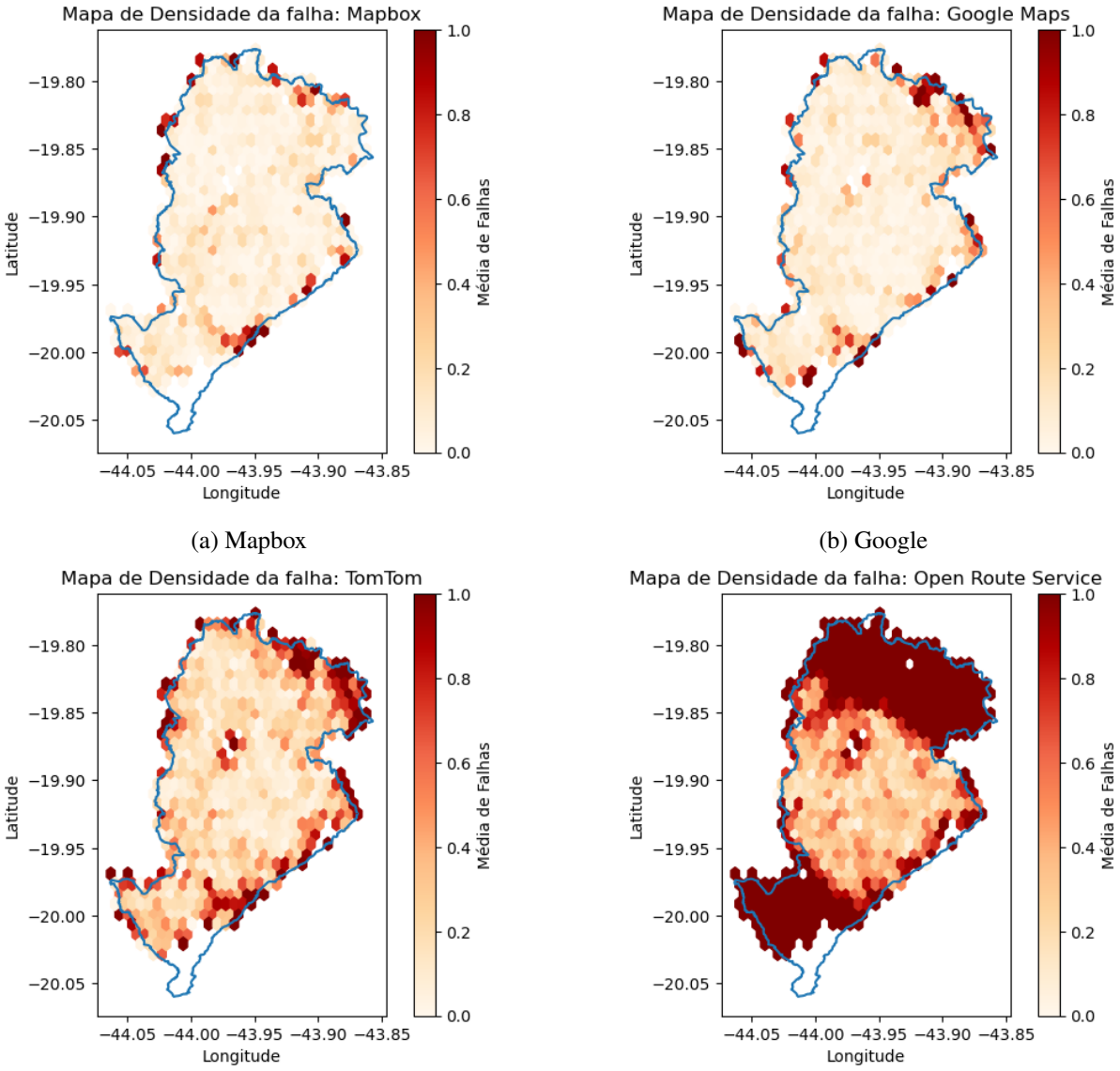


Figura 3.4 – Histograma comparativo de erro das APIs limitado em 300 metros para os dados de Belo Horizonte

Horizontes observados nos gráficos da figura 3.5 e nas tabelas 3.2 e 3.1.

3.4 Relações entre erro e discrepância

Então, foi realizada a análise comparativa entre erro e discrepância. As medidas escolhidas para essa análise foram a covariância e a distância para o ponto médio como descrito no capítulo



(a) Mapbox

(b) Google

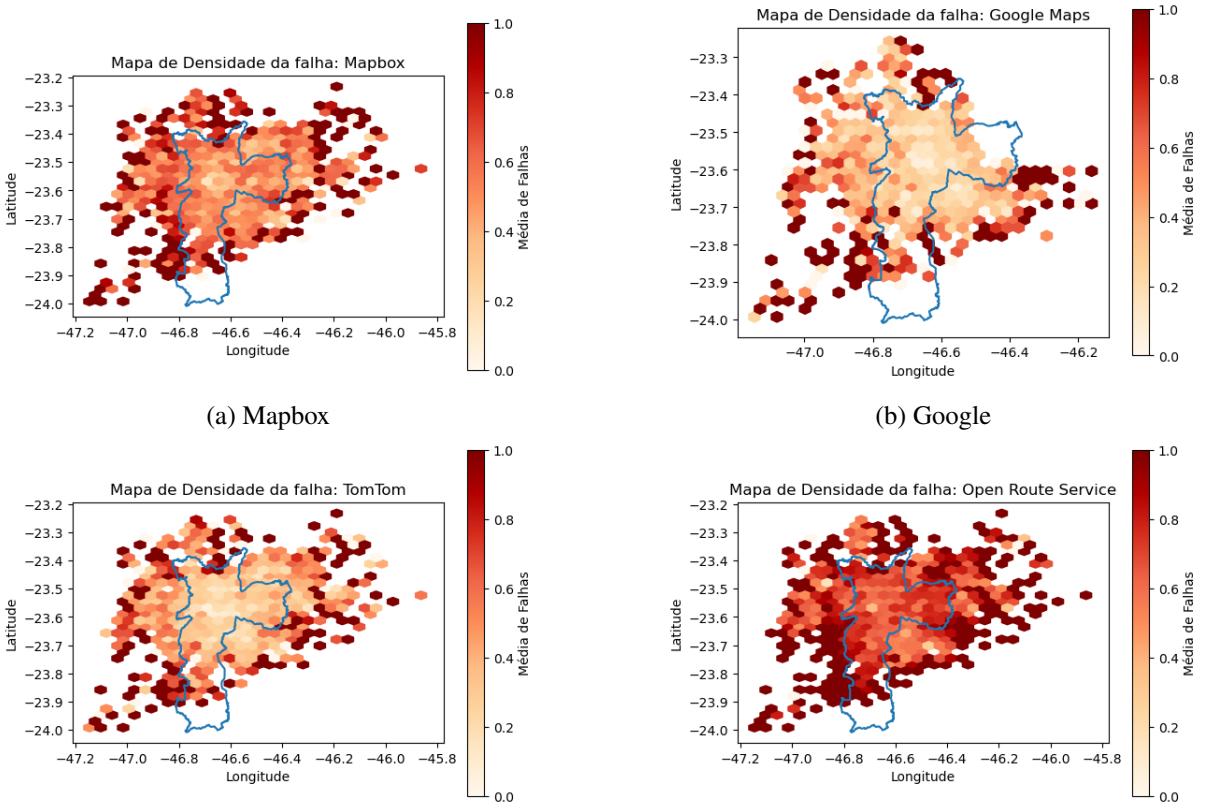
(c) TomTom

(d) ORS

Figura 3.5 – Gráficos de falhas de cada API para os dados de Belo Horizonte

2. Foram considerados apenas os endereços em que se tinha informação de todas as APIs. Depois de calcular as métricas para cada um dos pontos foi calculada a correlação de Pearson para cada API e cada base de dados.

Realizamos então uma análise com um subconjunto de 8574 endereços do banco de dados de São Paulo. A Tabela 3.3 exibe esses resultados. A partir da tabela, pode-se observar que para todas as APIs, as correlações com o erro são positivas. Isso indica que à medida que o erro aumenta, as medidas de discrepância também tendem a aumentar. de acordo com a tabela 2.1 a medida de covariância para esses dados apresentou uma correlação regular a forte para a maioria das APIs, variando de 0,53 a 0,67, exceto para o Google Maps, que teve uma correlação fraca. Por outro lado, a medida de distância até o ponto médio obteve uma correlação forte a muito forte para a maioria das APIs, com resultados na faixa de 0,88 a 0,94. Em contrapartida, a API do Google mostrou uma correlação regular muito próxima a fraca, mas houve uma melhoria em



(a) Mapbox

(b) Google

(c) TomTom

(d) ORS

Figura 3.6 – Gráficos de falhas de cada API para os dados de Belo São Paulo

comparação com a correlação de covariância.

Tabela 3.3 – Correlação de Pearson entre Erro e Medidas de Discrepância para São Paulo

API	Covariância	Distância até o Ponto Médio
Mapbox	0,5387	0,8972
TomTom	0,5398	0,8858
Google	0,2177	0,3615
ORS	0,6649	0,9378

Para o conjunto de dados de Belo Horizonte, conduzimos essa análise com 84.752 endereços, o que representa aproximadamente 99,71% da amostra utilizada. A Tabela 3.4 mostra esses resultados. Em geral, a tabela apresenta valores de correlação mais próximos de 0 do que os encontrados na análise dos dados de São Paulo, indicando que a correlação é mais fraca para este conjunto de dados como um todo. Para o Google e o ORS, a covariância mostrou uma correlação fraca, possivelmente indicando nenhuma relação entre erro e covariância para essas APIs. O Mapbox teve uma correlação regular com a covariância, enquanto o TomTom teve uma correlação forte.

Para a distância até o ponto médio, tivemos correlações fortes a muito fortes para o Mapbox e o TomTom. O Google e o ORS também apresentaram correlações fracas para a

distância do ponto médio. Apesar dos resultados de correlação inferiores em comparação com a análise do conjunto de dados de São Paulo, os resultados de Belo Horizonte parecem confirmar que distância até o ponto médio é uma medida melhor em comparação com a covariância para substituir o erro em situações onde o mesmo não pode ser obtido.

Tabela 3.4 – Correlação de Pearson entre Erro e Medidas de Discrepância para Belo Horizonte

API	Covariância	Distância até o Ponto Médio
Mapbox	0,4669	0,7764
TomTom	0,7269	0,9873
Google	0,0463	0,0754
ORS	0,1552	0,2775

4 Considerações Finais

Referências

BEHR, F.-J. *Geocoding: Fundamentals, Techniques, Commercial and Open Services*. Schellingstraße 24, D-70174 Stuttgart, Germany: [s.n.], 2010.

CALLEGARI-JACQUES, S. M. *Bioestatística: Princípios e Aplicações*. Dados eletrônicos. Porto Alegre: Artmed, 2007. Editado também como livro impresso em 2003. ISBN 978-85-363-1144-9.

CENTRO de Estudos da Metrópole. Urlhttps://centrodametropole.fflch.usp.br/pt-br. Acesso em: [24 maio. 2024].

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>.

DOCUMENTAçãO da Google Maps Geocodin API. Urlhttps://developers.google.com/maps/documentation/geocoding. Acesso em: [04 nov. 2023].

DOCUMENTAçãO da Mapbox Geocodin API. Urlhttps://docs.mapbox.com/api/search/geocoding/. Acesso em: [04 nov. 2023].

DOCUMENTAçãO da Open Route Service Geocodin API. Urlhttps://openrouteservice.org/dev/#/api-docs/geocode/search/get. Acesso em: [04 nov. 2023].

DOCUMENTAçãO da TomTom Geocodin API. Urlhttps://developer.tomtom.com/geocoding-api/documentation/product-information/introduction. Acesso em: [04 nov. 2023].

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 population-based study of air quality and birth defects. *Environmental Research*, v. 101, n. 2, p. 256–262, 2006. ISSN 0013-9351. Womens Occupational and Environmental Health. Disponível em: <https://www.sciencedirect.com/science/article/pii/S001393510600020X>.

GOOGLE Cloud Platform. Urlhttps://cloud.google.com/. Acesso em: [13 ago. 2023].

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>.

KLEIN, C. *Dicionário da língua portuguesa*. 1. ed. São Paulo: Rideel, 2015. E-book. Disponível em: <https://plataforma.bvirtual.com.br>.

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>.

Küçük Matci, D.; AVDAN, U. Address standardization using the natural language process for improving geocoding results. *Computers, Environment and Urban Systems*, v. 70, p. 1–8, 2018. ISSN 0198-9715. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0198971517300455>.

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.

PRODABEL. Urlhttps://prefeitura.pbh.gov.br/prodabel. Acesso em: [13 ago. 2023].

RELATóRIO do Experimento de Entradas das APIs. Urlhttps://github.com/rcpsilva/UncertaintyQuantificationForGeocodingServices/blob/main/UndergraduateTh testes.pdf. Acesso em: [21 jun. 2024].

SITE da Mapbox. Urlhttps://www.mapbox.com/. Acesso em: [25 jun. 2024].

SITE da Open Route Service. Urlhttps://openrouteservice.org/. Acesso em: [25 jun. 2024].

SITE da TomTom. Urlhttps://www.tomtom.com/. Acesso em: [25 jun. 2024].

SPIEGEL, M. R.; STEPHENS, L. J. *Estatística*. Grupo A, 2009. E-book. ISBN 9788577805204. Disponível em: <https://integrada.minhabiblioteca.com.br/#/books/9788577805204/>.

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/>.

TERRALAB - Laboratório de Capacitação e Desenvolvimento de Software. Urlhttp://www2.decom.ufop.br/terralab/. Acesso em: [11 ago. 2023].

WHITSEL, E. A.; QUIBRERA, P. M.; SMITH, R. L. et al. Accuracy of commercial geocoding: Assessment and implications. *Epidemiologic Perspectives and Innovations*, v. 3, n. 1, p. 8, 2006. Disponível em: <https://epi-perspectives.biomedcentral.com/articles/10.1186/1742-5573-3-8>.

ZANDBERGEN, P. A. Geocoding quality and implications for spatial analysis. *Geography Compass*, v. 3, n. 2, p. 647–680, 2009. Disponível em: <https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-8198.2008.00205.x>.

Anexos

ANEXO A – Tabelas dos experimentos de formatação completas

A.1 Resultados Mapbox

Experimento	Média (Km)	Mediana (Km)	Desvio Padrão	Média Aparada (Km)	Taxa de Resposta (%)	Taxa de Acerto (%)
1	1.539552	0.000046	10.912322	0.511817	1.0000	0.8506
1b	1.855776	0.000048	9.876150	0.826308	0.9994	0.8088
2	1.985113	0.000046	12.479481	0.880777	1.0000	0.8246
2b	3.747499	0.000049	26.633204	0.712573	0.9994	0.7982
3	1.660480	0.000046	11.255071	0.578759	1.0000	0.8400
3b	2.268966	0.000049	13.585637	0.831613	0.9968	0.8056
4	3.239740	0.000046	33.421642	0.579544	1.0000	0.8466
4b	2.395281	0.000049	18.048547	0.618146	0.9992	0.7986
5	2.270220	0.000046	25.666232	0.597641	0.9992	0.8380
5b	22.718122	0.000049	151.027338	0.722369	0.9976	0.8100

Tabela A.1 – Tabela de Resultados para Mapbox para a amostra de Belo Horizonte

Experimento	Média (Km)	Mediana (Km)	Desvio Padrão	Média Aparada (Km)	Taxa de Resposta (%)	Taxa de Acerto (%)
1	9.885009	0.264745	33.929581	5.545753	0.9750	0.4178
2	13.914447	0.481439	36.798156	8.848430	0.9778	0.3704
3	12.998989	0.287323	46.743396	6.338832	0.9920	0.4126
4	9.059893	0.287323	28.821270	5.833966	0.9784	0.4090
5	13.102779	0.287323	54.305399	6.421116	0.9800	0.4010

Tabela A.2 – Tabela de Resultados para MapBox para a amostra de São Paulo

A.2 Resultados Google

A.3 Resultados TomTom

A.4 Resultados Open Route Service

Experimento	Média (Km)	Mediana (Km)	Desvio Padrão	Média Aparada (Km)	Taxa de Resposta (%)	Taxa de Acerto (%)
1	2.284151	0.008843	5.067888	1.541325	0.9992	0.7272
1b	1.477092	0.007045	12.541127	0.641472	0.9996	0.8064
2	2.703568	0.008981	13.275209	1.500182	0.9998	0.7330
2b	2.488111	0.007888	24.657557	0.424849	0.9984	0.7802
3	2.191061	0.008868	4.905103	1.453413	0.9992	0.7338
3b	1.449151	0.007442	15.764553	0.408326	1.0000	0.7830
4	2.225610	0.008894	4.911848	1.508163	0.9990	0.7326
4b	1.317380	0.007442	15.783626	0.400024	0.9992	0.7778
5	2.214506	0.008916	4.911495	1.483368	0.9992	0.7332
5b	1.631620	0.008843	12.503913	0.840399	0.9988	0.7292

Tabela A.3 – Tabela de Resultados para Google para a amostra de Belo Horizonte

Experimento	Média (Km)	Mediana (Km)	Desvio Padrão	Média Aparada (Km)	Taxa de Resposta (%)	Taxa de Acerto (%)
1	4.084331	0.136854	10.741415	2.554311	0.9988	0.5080
2	6.290936	0.174920	21.319549	4.575344	0.9986	0.4854
3	7.252604	0.177119	23.235726	5.262855	0.9988	0.4842
4	9.891182	0.177119	66.380809	4.808587	0.9988	0.4842
5	6.657890	0.183598	24.621577	4.687355	0.9990	0.4800

Tabela A.4 – Tabela de Resultados para Google para a amostra de São Paulo

Experimento	Média (Km)	Mediana (Km)	Desvio Padrão	Média Aparada (Km)	Taxa de Resposta (%)	Taxa de Acerto (%)
1	9.638626	0.097375	54.293889	2.383578	1.0000	0.5280
1b	4.772675	0.060837	36.194963	1.415974	0.9998	0.5634
2	3.493690	0.055936	31.276516	1.894932	0.9994	0.5566
2b	4.977097	0.087184	34.512517	1.956344	0.9998	0.5376
3	4.209165	0.055609	41.653527	1.857687	1.0000	0.5582
3b	4.963664	0.082551	34.529210	1.938064	0.9988	0.5392
4	10.042613	0.060228	57.575517	2.080298	0.9998	0.5532
4b	4.977097	0.087184	34.512517	1.956344	0.9998	0.5376
5	4.211492	0.055581	41.665922	1.861228	0.9994	0.5578
5b	4.965005	0.083011	34.522296	1.940898	0.9992	0.5392

Tabela A.5 – Tabela de Resultados para TomTom para a amostra de Belo Horizonte

Experimento	Média (Km)	Mediana (Km)	Desvio Padrão	Média Aparada (Km)	Taxa de Resposta (%)	Taxa de Acerto (%)
1	36.121177	0.108194	249.594126	3.940234	0.8548	0.4494
2	36.597577	0.108194	250.180881	3.818216	0.8552	0.4496
3	15.477097	0.108194	105.033151	3.638018	0.8552	0.4502
4	36.121297	0.108278	249.594109	3.940367	0.8548	0.4490
5	13.224068	0.107051	84.522569	3.595458	0.8414	0.4440

Tabela A.6 – Tabela de Resultados para TomTom para a amostra de São Paulo

Experimento	Média (Km)	Mediana (Km)	Desvio Padrão	Média Aparada (Km)	Taxa de Resposta (%)	Taxa de Acerto (%)
1	5.443245	6.606720	4.669510	5.259343	0.9992	0.2646
1b	134.564517	6.726786	352.871052	70.399993	0.9526	0.1562
2	141.563530	7.689302	326.944740	85.764655	0.9906	0.2228
2b	235.720433	120.745927	321.074977	190.471249	0.9530	0.0546
3	215.411691	0.450277	446.187607	148.459274	0.9904	0.4006
3b	221.030496	0.545940	442.133290	155.460776	0.9906	0.3908
4	7.574040	7.585665	3.281047	7.597740	1.0000	0.0146
4b	152.061311	7.894395	379.053022	86.883669	0.9512	0.0672
5	7.867047	7.587377	15.029207	7.599037	0.9958	0.0148
5b	5.828340	6.606720	20.905763	5.322782	0.9998	0.2478

Tabela A.7 – Tabela de Resultados para Open Route Service para amostra de Belo Horizonte

Experimento	Média (Km)	Mediana (Km)	Desvio Padrão	Média Aparada (Km)	Taxa de Resposta (%)	Taxa de Acerto (%)
1	8.016763	0.346648	16.978958	6.323177	0.9986	0.2894
2	149.089363	23.343768	368.646520	80.847362	0.9950	0.0530
3	22.615834	23.022681	9.940436	22.497670	0.9988	0.0014
4	111.728383	16.604444	356.468299	45.451290	0.9900	0.1494
5	19.782864	19.499381	23.891962	18.967053	0.9996	0.0104

Tabela A.8 – Tabela de Resultados para OpenRouteService para a amostra de São Paulo