

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 APIS DE GEOCODIFICAÇÃO EM GRANDES
CIDADES BRASILEIRAS**

Ouro Preto, MG
2024

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

ANA LUIZA ALMEIDA SOARES

**AVALIAÇÃO DE APIS 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
2024



FOLHA DE APROVAÇÃO

Ana Luiza Almeida Soares

Avaliação de APIs de Geocodificação em Grandes Cidades Brasileiras

Monografia apresentada ao Curso de Bacharelado em Ciência da Computação da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Ciência da Computação

Aprovada em 01 de agosto de 2024

Membros da banca

Doutor - Rodrigo Cesar Pedrosa Silva - Orientador(a) (Universidade Federal de Ouro Preto)
Doutor - Vander Luis de Sousa Freitas - Orientador(a) (Universidade Federal de Ouro Preto)
Doutor - Tiago Garcia de Senna Carneiro - Orientador(a) (Universidade Federal de Ouro Preto)

Rodrigo Cesar Pedrosa Silva, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 02/08/2024



Documento assinado eletronicamente por **Rodrigo Cesar Pedrosa Silva, PROFESSOR DE MAGISTERIO SUPERIOR**, em 02/08/2024, às 15:13, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **0746275** e o código CRC **1B8D2301**.

Agradecimentos

Gostaria de expressar minha profunda gratidão a todos aqueles que, de alguma forma, contribuíram para a realização desta monografia. Primeiramente, agradeço ao meu orientador, Rodrigo César Pedrosa, coordenador do TerraLAB, cujo apoio constante e orientação foram essenciais para a conclusão deste trabalho. Além de sua orientação acadêmica, Rodrigo esteve presente em diversos momentos desafiantes da minha vida, oferecendo suporte e motivação. Agradeço também ao Tiago Garcia da Senna Carneiro, membro da banca e também coordenador do TerraLAB, que não apenas avaliou este trabalho, mas também esteve ao meu lado nos momentos difíceis, proporcionando apoio e conselhos valiosos. Ao Vander Luis de Souza Feitas, membro da banca, agradeço pela minuciosa avaliação e pelas contribuições significativas que enriqueceram este trabalho. Meu sincero agradecimento ao Pedro Garcia, que integrou a banca da minha primeira monografia e contribuiu com questionamentos pertinentes que ajudaram a aprimorar a qualidade do meu trabalho. Agradeço ao meu parceiro Dairan Jhordan pelo seu constante apoio e incentivo. Um agradecimento especial aos amigos Bruna, Yasmim, Rafael, Breno, Josiane e Arthur, que me apoiaram ao longo do processo de escrita, oferecendo palavras de encorajamento e companhia nos momentos mais difíceis. Não poderia deixar de mencionar os membros do TerraLab que me ajudaram diretamente: André, João Lucas, Matheus Gonçalves, Tiago Gomes, Vinícius e todos os outros membros do TerraLAB, que contribuíram com seus conhecimentos e habilidades. Agradeço ao meu psicólogo Lucas, cujo apoio emocional e orientação foram fundamentais para manter meu equilíbrio durante este período desafiador. Por fim, agradeço à equipe de orientação estudantil da PRACE - UFOP, representada por Mariana e Joseane, pelo suporte e orientação que me proporcionaram ao longo dessa jornada acadêmica. A minha família, minha mãe Terezinha, meus irmãos Gessiane, Ivan e Iasmim, e minha sobrinha Laura, agradeço pelo amor, compreensão e incentivo constantes. Sem vocês, esta conquista não seria possível. Agradeço também a todos aqueles que, de alguma forma, passaram pela minha vida e contribuíram para a realização deste trabalho, mesmo que não mencionados diretamente. Cada um de vocês teve um papel importante nesta jornada.

A todos vocês, meu muito obrigado!

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 principal avaliar a qualidade de quatro APIs de geocodificação: Google Maps, Mapbox, TomTom e Open Route Service (ORS). Para isso, foram utilizadas duas bases de dados: a primeira de Belo Horizonte e a segunda da região metropolitana de São Paulo. Além de verificar a qualidade das APIs, o trabalho também tem como objetivo encontrar uma medida de discrepância comparável com o erro e avaliar a melhor formatação de entrada para as APIs. Utilizamos várias métricas para a análise de qualidade, 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 distância ao ponto médio. A maioria das APIs apresentou uma taxa de resposta alta em ambas as bases. Em relação à taxa de acerto, a maioria das APIs teve resultados de médios (de 45% a 70%) a altos (acima de 70%), com exceção da API ORS, que teve resultados insatisfatórios nas duas bases de dados. No que diz respeito à relação entre discrepância e erro, identificamos uma correlação forte a muito forte com todas as APIs, exceto Google Maps, que teve uma correlação muito fraca para ambas as bases, e ORS, que teve correlação fraca para a base de São Paulo. Por fim, em relação à formatação dos dados de entrada, tivemos resultados que indicam que a melhor formatação de entrada é a que segue o código postal brasileiro e que a adição de bairro pode causar melhora na geocodificação de algumas das APIs avaliadas.

Palavras-chave: GeoAPIs. Qualidade. Mapbox. TomTom. ORS. Google Maps. Erro. Discrepância. Formatação entrada. Geocodificação. Análise de Dados. Análise Espacial. Localização.

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 evaluate the accuracy of the APIs used. This study aims to evaluate the quality of four geocoding APIs: Google Maps, Mapbox, TomTom, and Open Route Service (ORS). For this, two databases were used: the first from Belo Horizonte and the second from the metropolitan region of São Paulo. In addition to assessing the quality of the APIs, the study also aims to find a discrepancy measure comparable to the error and to evaluate the best input formatting for the APIs. We used several metrics for quality analysis, including mean, standard deviation, median, 5% trimmed mean, response rate (proportion between geocoding requests and responses), and precision rate (number of addresses with an error less than 150 meters). Additionally, we conducted a spatial error analysis and investigated the relationship between discrepancy and error, using the distance to the midpoint measure. Most APIs showed a high response rate in both databases. Regarding the precision rate, most APIs had medium to high results, except for the ORS API, which had unsatisfactory results in both databases. Concerning the relationship between discrepancy and error, we identified a strong to very strong correlation with all APIs, except for Google Maps, which had a very weak correlation for both databases, and ORS, which had a weak correlation for the São Paulo database. Finally, concerning the input data formatting, we had results indicating that the best input formatting follows the Brazilian postal code, and adding the neighborhood can improve geocoding for some of the evaluated APIs.

Keywords: GeoAPIs. Quality. Mapbox. TomTom. ORS. OpenRouteService. Google Maps. Error. Discrepancy. Input formatting. Geocoding. Data analysis. Spatial analysis. Location.

Lista de Ilustrações

Figura 1.1 – Adaptada do livro Longley et al. (2013). Visão conceitual da incerteza	2
Figura 2.1 – Mapa de clusters do Centro de Estudos da Metrópole	9
Figura 2.2 – Gráfico de Endereços da Base de São Paulo e região metropolitana	10
Figura 2.3 – Site da Prodabel para pesquisa de endereços.	11
Figura 2.4 – Gráficos dos endereços da Base de Belo Horizonte e amostragem obtida . .	12
Figura 2.5 – Esquemática do processo de preparação e geocodificação dos dados . . .	13
Figura 3.1 – Histogramas do erro das 4 APIs para o todos os dados de São Paulo	21
Figura 3.2 – Histogramas do erro das 4 APIs para o todos os dados de Belo Horizonte . .	22
Figura 3.3 – Histograma comparativo de erro das APIs limitado em 300 metros para os dados de São Paulo	23
Figura 3.4 – Histograma comparativo de erro das APIs limitado em 300 metros para os dados de Belo Horizonte	23
Figura 3.5 – Gráficos de falhas de cada API para os dados de Belo Horizonte	24
Figura 3.6 – Gráficos de falhas de cada API para os dados de Belo São Paulo	25
Figura 3.7 – Boxplot de Experimentos por Erro, com todas as APIs avaliadas	30
Figura 3.8 – Boxplot de Experimentos por Erro, com todas as APIs avaliadas e sem Outliers	30
Figura 3.9 – Boxplot de Experimentos por Erro, com todas as APIs avaliadas exceto ORS e sem Outliers para os dados de Belo Horizonte	30
Figura 3.10–Boxplot de Experimentos por Erro, com todas as APIs avaliadas para a amostra de São Paulo	31
Figura 3.11–Boxplot de Experimentos por Erro, com todas as APIs avaliadas e sem Outliers para a amostra de São Paulo	31
Figura 3.12–Boxplots de Erro por API para cada uma das APIs avalidas	32
Figura 3.13–Boxplots de Erro por API para cada uma das APIs avalidas para a amostra de São Paulo	33
Figura 3.14–Boxplot de Erro por API com todos os experimentos avaliados sem Outliers: Mapbox	33
Figura 3.15–Boxplot de Erro por API com todos os experimentos avaliados sem Outliers para a amostra de São Paulo: Mapbox	34
Figura 3.16–Boxplot de Erro por API com todos os experimentos avaliados sem Outliers: Google	34
Figura 3.17–Boxplot de Erro por API com todos os experimentos avaliados sem Outliers para a amostra de São Paulo: Google	34
Figura 3.18–Boxplot de Erro por API com todos os experimentos avaliados sem Outliers: TomTom	35

Figura 3.19–Boxplot de Erro por API com todos os experimentos avaliados sem Outliers: ORS	35
Figura 3.20–Boxplot de Erro por API com todos os experimentos avaliados sem Outliers para a amostra de São Paulo: ORS	35
Figura 3.21–Gráficos de falhas para a maior e menor taxas de acertodos experimentos para os dados de Belo Horizonte: MapBox	36
Figura 3.22–Gráficos de falhas para a maior e menor taxas de acerto dos experimentos para os dados de São Paulo: MapBox	36
Figura 3.23–Gráficos de falhas para a maior e menor taxas de acertodos experimentos para os dados de Belo Horizonte: Google Maps	37
Figura 3.24–Gráficos de falhas para a maior e menor taxas de acertodos experimentos para os dados de São Paulo: Google Maps	38
Figura 3.25–Gráficos de falhas para a maior e menor taxas de acertodos experimentos para os dados de Belo Horizonte: TomTom	38
Figura 3.26–Gráficos de falhas para a maior e menor taxas de acertodos experimentos para os dados de São Paulo: TomTom	39
Figura 3.27–Gráficos de falhas para a maior e menor taxas de acerto dos experimentos para os dados de Belo Horizonte: Open Route Service	39
Figura 3.28–Gráficos de falhas para a maior e menor taxas de acerto dos experimentos para os dados de São Paulo: Open Route Service	40

Lista de Tabelas

Tabela 2.1 – Tabela de Correlação de Pearson	15
Tabela 2.2 – Formato Recomendado de Entrada para APIs de Geocodificação	16
Tabela 2.3 – Descrição dos formatos	16
Tabela 2.4 – Formato de Entrada das APIs Utilizadas pelo TerraLAB	16
Tabela 2.5 – Formato de cada experimento	17
Tabela 2.6 – Formato dos experimentos adicionais	17
Tabela 3.1 – Métricas de Erro para São Paulo	19
Tabela 3.2 – Métricas de Erro para Belo Horizonte	19
Tabela 3.3 – Correlação de Pearson entre Erro e Distância do Ponto Médio para São Paulo	26
Tabela 3.4 – Correlação de Pearson entre Erro e Distância do Ponto Médio para Belo Horizonte	26
Tabela 3.5 – Taxa de resposta de cada API por experimento de Belo Horizonte	27
Tabela 3.6 – Taxa de resposta de cada API por experimento de São Paulo	27
Tabela 3.7 – Taxa de acerto de cada API por experimento de Belo Horizonte	28
Tabela 3.8 – Taxa de acerto de cada API por experimento de São Paulo	29
Tabela A.1 – Tabela de Abreviações	48
Tabela B.1 – Tabela de Resultados para Mapbox para a amostra de Belo Horizonte	49
Tabela B.2 – Tabela de Resultados para MapBox para a amostra de São Paulo	49
Tabela B.3 – Tabela de Resultados para Google para a amostra de Belo Horizonte	50
Tabela B.4 – Tabela de Resultados para Google para a amostra de São Paulo	50
Tabela B.5 – Tabela de Resultados para TomTom para a amostra de Belo Horizonte	50
Tabela B.6 – Tabela de Resultados para TomTom para a amostra de São Paulo	51
Tabela B.7 – Tabela de Resultados para Open Route Service para amostra de Belo Horizonte	51
Tabela B.8 – Tabela de Resultados para OpenRouteService para a amostra de São Paulo . .	51

Sumário

1	Introdução	1
1.1	Endereços e Geocodificação	1
1.2	APIs de Geocodificação e Análise de qualidade	3
1.3	APIs de Geocodificação e formatação das entradas	5
1.4	Objetivos	6
2	Bases de Dados, Métodos de Geocodificação e Formas de Avaliação	9
2.1	Bases de Dados	9
2.2	Processo de Geocodificação	11
2.3	Método de Avaliação	12
2.3.1	Erro e Taxa de Resposta	12
2.3.2	Discrepância	14
2.4	Experimentos para avaliação da formatação da entrada	16
3	Resultados	18
3.1	Erro, Taxa de Resposta e Taxa de Precisão	18
3.2	Distribuição de Erro	20
3.3	Distribuição Espacial do Erro	22
3.4	Relações entre erro e discrepância	25
3.5	Experimentos de Formatação	26
3.5.1	Distribuição Espacial das Falhas dos Experimentos de Formatação	35
4	Considerações Finais	41
	Referências	43
	Apêndices	46
	APÊNDICE A Abreviações utilizadas	47
	APÊNDICE B Tabelas dos experimentos de formatação completas	49

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 vida humana está profundamente ligada à localização. Assim sendo, compreender e manipular informações geográficas é essencial para qualquer aplicação que envolva a humanidade.

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 e 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 utilizar a localização nessas aplicações, é necessário criar uma representação computacional do endereço. 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 entre endereços.
- 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). A incerteza pode ser definida 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 distorce 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).

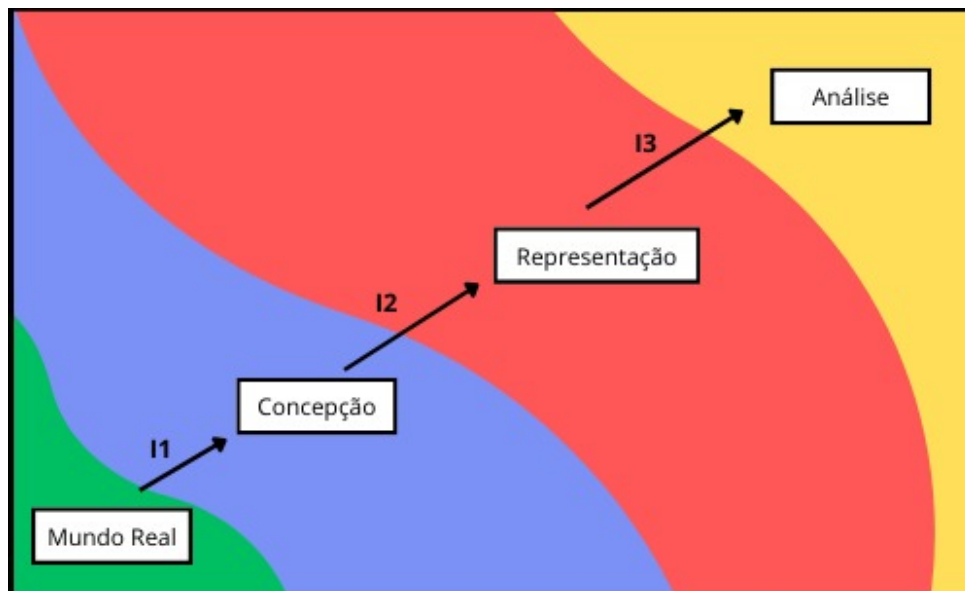


Figura 1.1 – Adaptada do livro Longley et al. (2013). Visão conceitual da incerteza

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

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 (ORS, 2024; TERRALAB, 2024b; MAPBOX, 2024; TOM-TOM, 2024). 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 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 (CHOW; DEDE-BAMFO; DAHAL, 2016). 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 e 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 contendo informações descritivas do endereço (rua, número, cidade etc.) e informações geográficas (latitude e longitude) foi utilizada. 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 do Texas. O erro de cada endereço geocodificado foi calculado como a distância euclidiana de dois pontos, sendo eles, o ponto de 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 de que as ferramentas online podem ser utilizadas como substitutas dos SIGs.

Apesar de [Chow, Dede-Bamfo e 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. e Alencar \(2011\)](#), que faz uma avaliação da qualidade da geocodificação da Google Maps API fornecida pela Google Cloud Platform ([GCP, 2023](#)). 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, 2023](#)). 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 do que 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 das descobertas importantes, o estudo de [Jr. e Alencar \(2011\)](#) 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.

Em [Kilic e Gulgen \(2020\)](#) é realizada uma avaliação similar, porém avaliando duas APIs de geocodificação, Google Maps e Bing Maps. Foram coletados endereços de hotéis em Fatih, na Turquia, e em Miami Beach, nos EUA, por meio de web scraping em uma plataforma de hospedagem. Esses dados foram filtrados e as coordenadas da base final foram obtidas manualmente, totalizando 156 endereços para ambas as localizações. Para avaliar as APIs de forma comparativa, foram utilizados dois métodos:

- Análise de precisão: foi utilizada a distância euclidiana para medir o erro da geocodificação. Por meio dessa avaliação, observou-se que o Bing apresentava resultados menos precisos que o Google para a base da Turquia; porém, para a base dos EUA, os resultados eram similares.
- Análise de similaridade de palavras: o objetivo era verificar o quão similar era o endereço

descritivo retornado pela API em relação à referência. Para isso, foi utilizada a métrica Distância Levenshtein. Novamente, para a base dos EUA, os resultados de ambas as APIs foram similares; porém, para a base da Turquia, a API Bing apresentou menos similaridade, indicando que a Bing é uma API menos indicada para essa região.

Uma revisão sistemática recente que reúne trabalhos sobre geocodificação é o artigo [Zhang e Bethard \(2024\)](#). Nele são avaliados trabalhos publicados a partir de 2010 que se enquadram em três frentes: bases de geocodificação, métricas para avaliar a geocodificação e sistemas de geocodificação. Em relação às métricas de avaliação da geocodificação, o artigo demonstrou que a maioria das pesquisas utiliza o erro como principal métrica de avaliação de qualidade, sendo este geralmente calculado como a distância euclidiana. Existem alguns artigos que exploram a utilização de polígonos ou técnicas de aprendizado de máquina para tal, mas esses são minoria. Outro achado importante é a utilização de média e mediana para sumarização do erro. Os artigos demonstram que, embora ambas sejam utilizadas para esse fim, a mediana apresentou resultados melhores comparados com a média, por não ser tão sensível a valores extremos como a média. O artigo é interessante para avaliar o que há de pesquisa atual na área de avaliação de qualidade das geoAPIs e o que é esperado em novas pesquisas.

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 [TerraLAB \(2024b\)](#) foi conduzida uma série de experimentos para avaliar os impactos da modificação na ordem dos endereços de entrada nas APIs ([TERRALAB, 2024a](#)). 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 e 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 Eskişehir, 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 (GOOGLE, 2023). 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 da geocodificação das APIs 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). Dado que o erro é a diferença entre a referência e o valor obtido, temos que essa é uma medida clara e concreta. Sendo assim, utilizaremos o erro para a avaliação de qualidade, assim como foi realizado em trabalhos anteriores.

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, facilitando a obtenção das métricas em relação ao erro.

Nesse contexto pretendemos investigar se existe alguma métrica de discrepância 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.

Com acesso às medidas de discrepância e o erro, um dos nossos 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 última 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 ¹.

¹ TerraLAB - Laboratório de Pesquisa e Capacitação em Software: Localizado na Universidade Federal de Ouro Preto, o TerraLAB visa preparar os estudantes para o mercado de trabalho ao simular um ambiente de empresa de software, com cargos ocupados por professores e alunos. O laboratório se destaca por integrar práticas profissionais e pesquisas, focando principalmente em aplicações que lidam com informações geográficas.

Dado o contexto, iremos avaliar quatro APIs utilizadas no 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 da base padrão ouro coletado por GPS. A discrepância medirá o nível de discordância entre as APIs.

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

Com isso, o objetivo desse trabalho é 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?

2 Bases de Dados, Métodos de Geocodificação e Formas de Avaliação

Para atingir os objetivos citados, recorreremos a duas bases de dados “padrão-ouro” como referência. Utilizando essas bases, calculamos a medida de erro e calculamos diversas métricas utilizando essa medida.

2.1 Bases de Dados

Foram coletadas duas bases de dados distintas para este trabalho.

A primeira base utilizada é proveniente do Centro de Estudos da Metrópole (CEM) (CEM, 2024). Essa base consiste em 12.503 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 ¹ 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

¹ <<http://200.144.244.241:3002/geolocation>>

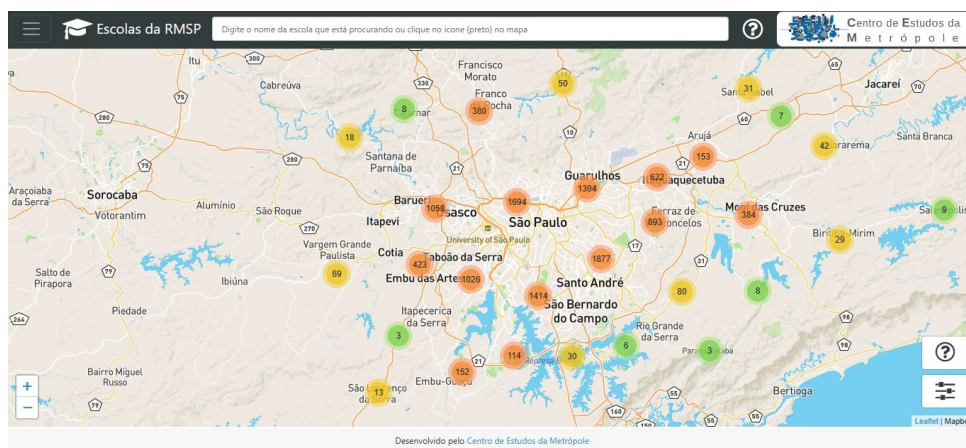


Figura 2.1 – Mapa de clusters do Centro de Estudos da Metrópole

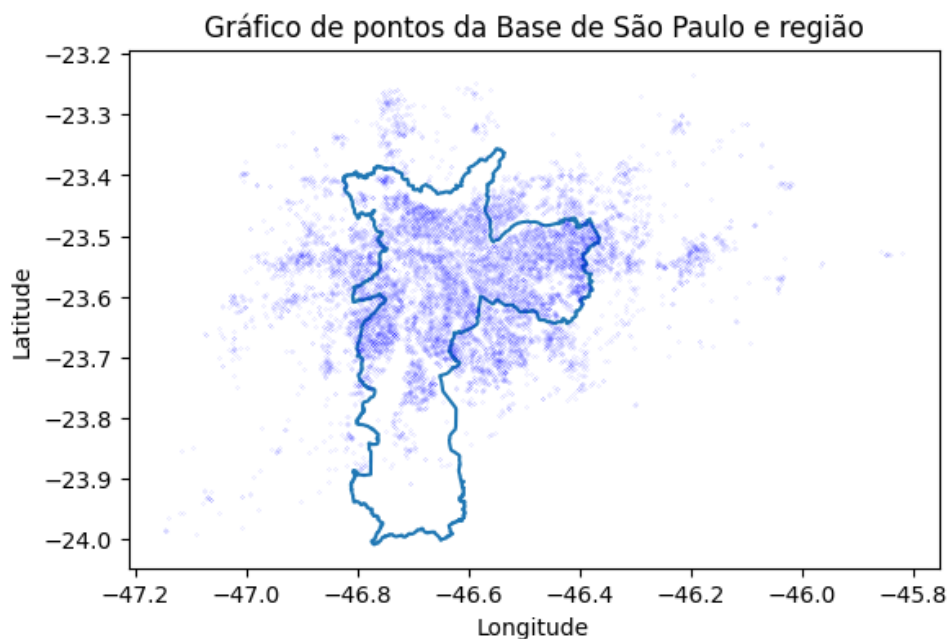


Figura 2.2 – Gráfico de Endereços da Base de São Paulo e região metropolitana

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 visualizar a distribuição de endereços dentro e fora da cidade.

A segunda base de dados utilizada é fornecida pela (PRODABEL, 2023), 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 é mantida e atualizada mensalmente por 27 empresas, tanto públicas quanto privadas, de Belo Horizonte. Essas empresas tem 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 está em constante atualização e é amplamente utilizada por diversos serviços da prefeitura. Um exemplo notável é o uso da base para georreferenciamento na distribuição de alunos da rede pública. 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, é uma base bastante relevante para a cidade de Belo Horizonte (JR.; ALENCAR, 2011).

Na data de acesso, essa base continha um total de 763.229 endereços. A prefeitura disponibiliza um site com um mapa² 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

² <<https://bhmap.pbh.gov.br>>

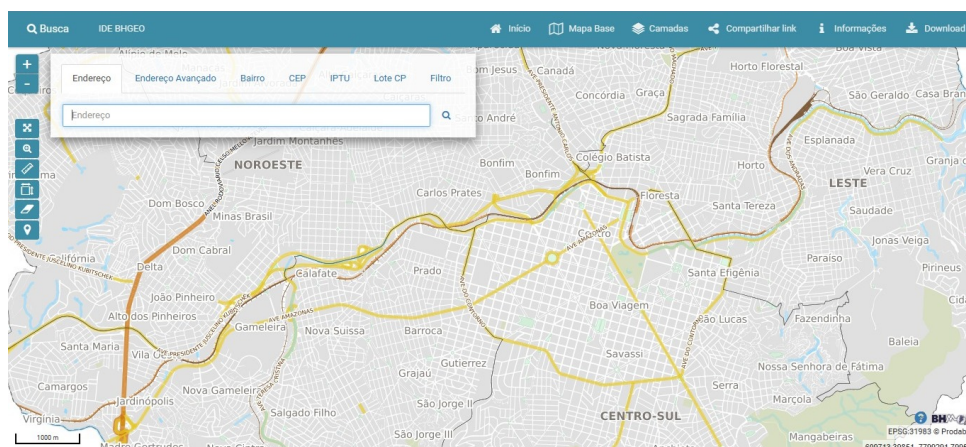


Figura 2.3 – Site da Prodabel para pesquisa de endereços.

entre essa abordagem e a prática comum de colocar o endereço na frente do edifício pode causar um pequeno erro de alguns metros na comparação da geocodificação.

Devido a limitações computacionais tanto dos autores deste trabalho quanto da aplicação responsável pela geocodificação, optamos por realizar uma amostragem da base de Belo Horizonte, com o intuito de reduzir a quantidade de dados processados. Nossa amostra consiste em 85.000 endereços da cidade. A fim de garantir uma distribuição uniforme dos endereços no espaço, empregamos o método do hipercubo latino para a amostragem. A Figura 2.4 apresenta dois gráficos contendo os pontos da base original e os da amostra obtida. Os gráficos contêm os pontos referentes a cada uma das bases e um contorno da cidade de Belo Horizonte. Com essa visualização, 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.

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. A Tabela A.1 do apêndice A contém as abreviações utilizadas no processo de padronização. Quaisquer abreviações não contidas na tabela não são tratadas.

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

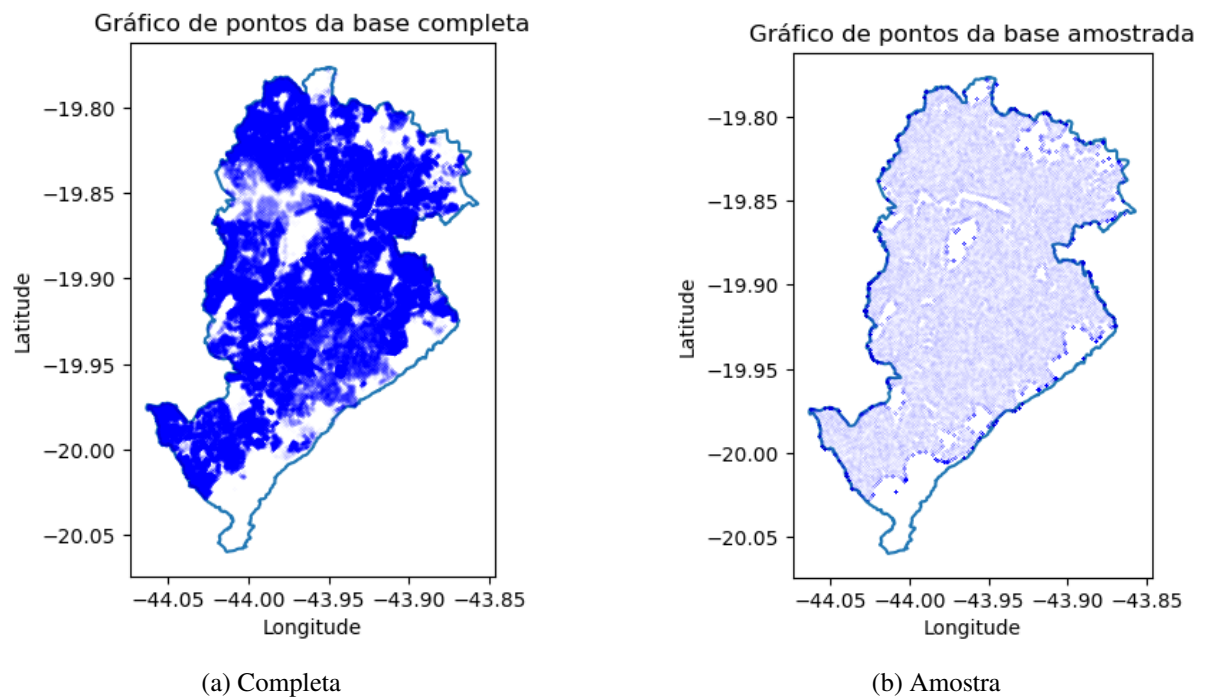


Figura 2.4 – Gráficos dos endereços da Base de Belo Horizonte e amostragem obtida

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 o processo de preparação e geocodificação dos dados.

2.3 Método de Avaliação

2.3.1 Erro e Taxa de Resposta

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:

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

Onde:

- e é o erro da geocodificação,
- D é a distância euclidiana entre dois pontos,

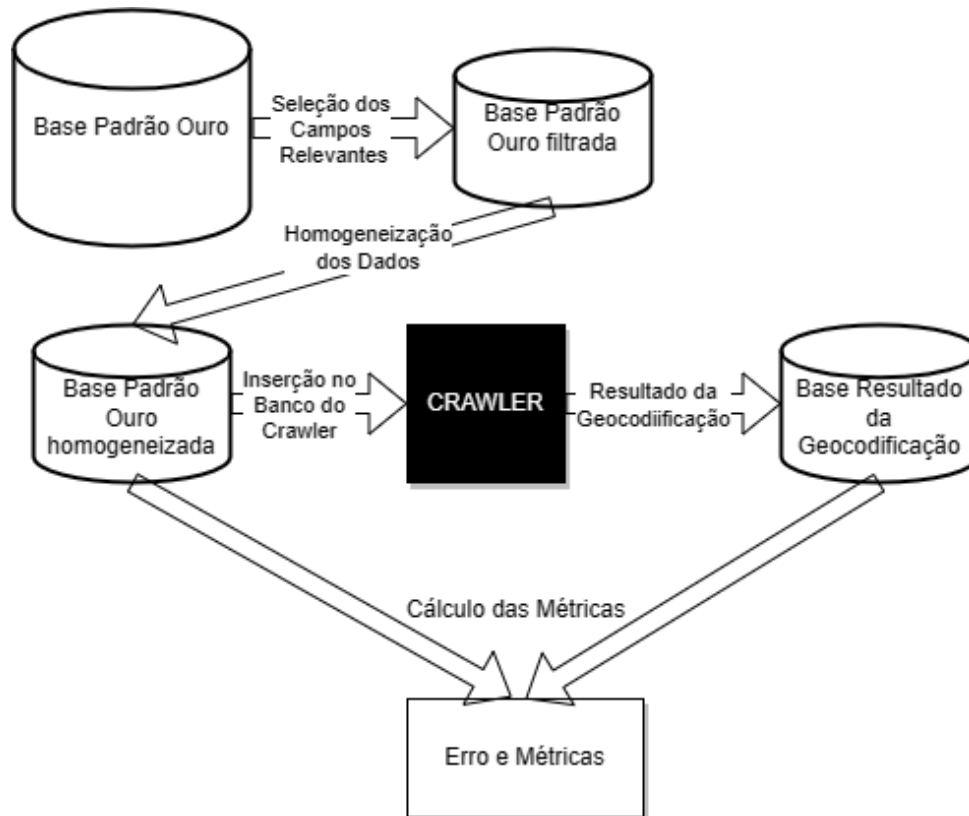


Figura 2.5 – Esquematização do processo de preparação e geocodificação dos dados

- p_{Ouro} é o ponto da base padrão ouro, e item p_{Geo} é o ponto resultante da geocodificação.

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

onde:

- 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 para avaliar a qualidade é a taxa de resposta da API. Para alguns endereços da base de dados, as GeoAPIs podem retornar um erro, não fornecendo uma geocodificação válida. 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:

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

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 em [Jr. e Alencar \(2011\)](#), que afirma que 150 metros é aproximadamente o tamanho de um quarteirão em Belo Horizonte. É representada pela seguinte fórmula:

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

Onde:

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

2.3.2 Discrepância

Para encontrar uma medida que se compare ao erro utilizando apenas informações fornecidas pela geocodificação, é necessário avaliar a discrepância entre as geocodificações. Para isso, será utilizada a métrica de distância do ponto médio. 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 distância até o ponto médio é calculada como a distância euclidiana de uma geocodificação até o ponto médio e é definida como:

$$\text{Distância até o Ponto Médio para cada ponto} = \sqrt{(\text{lat}_i - \text{mid_lat})^2 + (\text{lon}_i - \text{mid_lon})^2} \quad (2.5)$$

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:

$$\text{Ponto Médio} = \left(\frac{\sum_{i=1}^n \text{lat}_i}{n}, \frac{\sum_{i=1}^n \text{lon}_i}{n} \right) \quad (2.6)$$

Onde:

- $\bar{\text{lat}}_i$ e $\bar{\text{lon}}_i$ representam a média de latitude e longitude de casa API 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 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 disponibilizam 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	(GOOGLE, 2023)
Open Route Service	Sem recomendações específicas	(ORS, 2023)
Mapbox	Recomenda utilizar o formato oficial dos EUA ou o formato do serviço postal do país buscado	(MAPBOX, 2023)
TomTom	Sem recomendações específicas	(TOMTOM, 2023)

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 seu próprio padrão de formatação utilizado. A Tabela 2.4 apresenta os formatos de entrada utilizados por cada uma das APIs no laboratório.

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

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

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.

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

A base de Belo Horizonte contém 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 Geocodificações prontas, faremos o cálculo das métricas de erro para cada uma das APIs, como explicado na Seção 2.3.

3 Resultados

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) como mencionado no Capítulo 1. 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

Para avaliar a qualidade das geocodificações fizemos uma análise do erro, taxa de precisão e taxa de resposta das geocodificações obtidas. Foi realizado 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. 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 ORS obteve o melhor resultado, seguida pela Mapbox,

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 são taxas com valores semelhantes àqueles 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 é considerado preciso se o erro for 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, com o objetivo de avaliar a qualidade da geocodificação das APIs. 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, utilizando a escala logarítmica para melhor compreensão. 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 200 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.

Apesar dessas limitações, é possível realizar algumas análises a partir desses gráficos. Para os dados de São Paulo, nota-se que os resultados da Google são mais promissores, com uma concentração maior de valores em até 200 quilômetros. A Mapbox e a TomTom, por outro lado, apresentaram resultados similares, com a TomTom mostrando uma leve melhora.

Para os dados de Belo Horizonte, novamente a Google se destacou nesse gráfico, com valores bem abaixo de 74 quilômetros. A Mapbox também obteve um resultado promissor, seguida pela TomTom. A ORS, por outro lado, apresentou resultados diferentes das outras APIs. A faixa de erro foi até 20 quilômetros e, dentro dessa faixa, ela apresentou uma distribuição próxima da uniforme. Com uma faixa tão pequena, não foi possível realizar uma comparação com as outras APIs apenas com esse gráfico, pois não sabemos se, nessa faixa de valores, as outras APIs têm o mesmo comportamento.

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

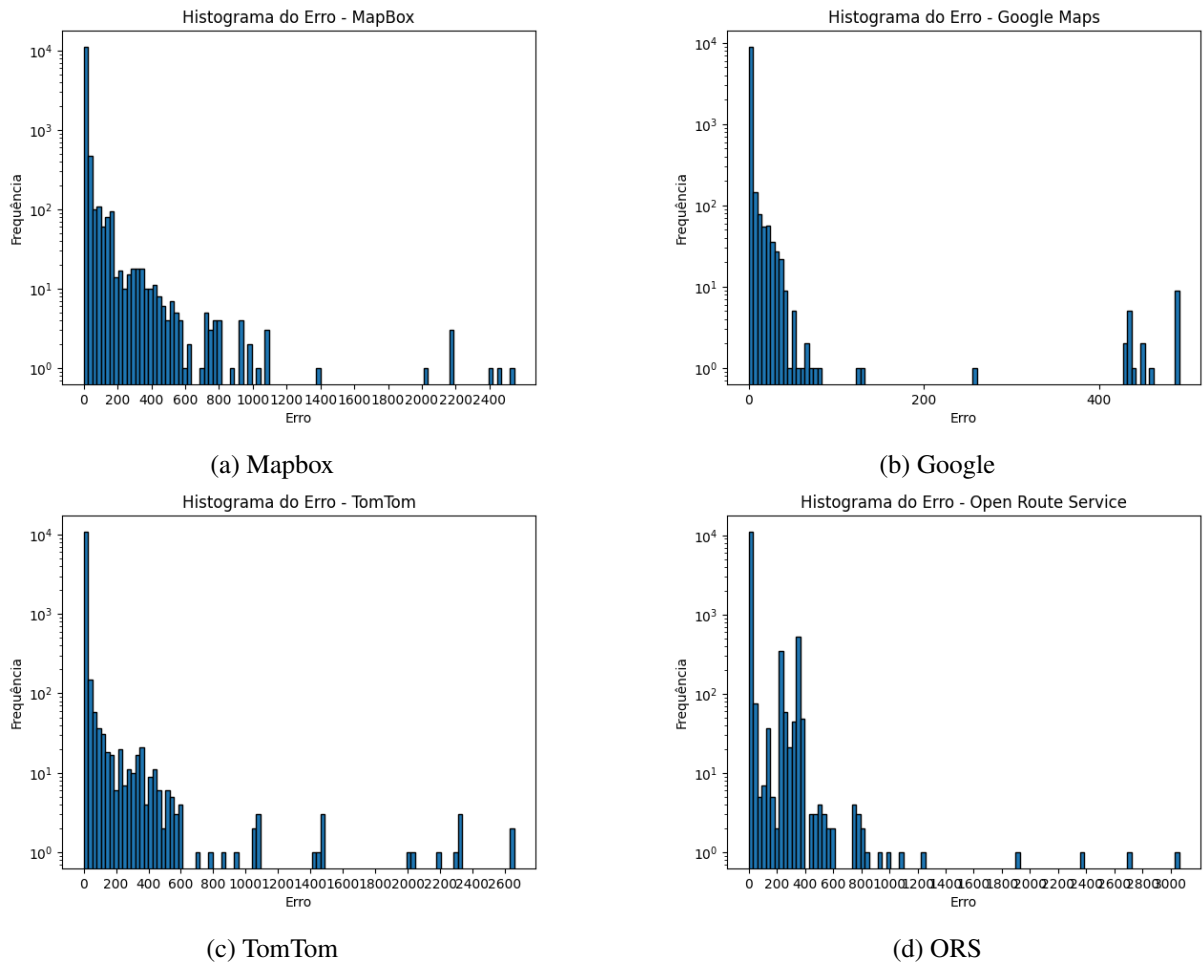


Figura 3.1 – Histogramas do erro das 4 APIs para o todos os dados de São Paulo

mais estreita, 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.

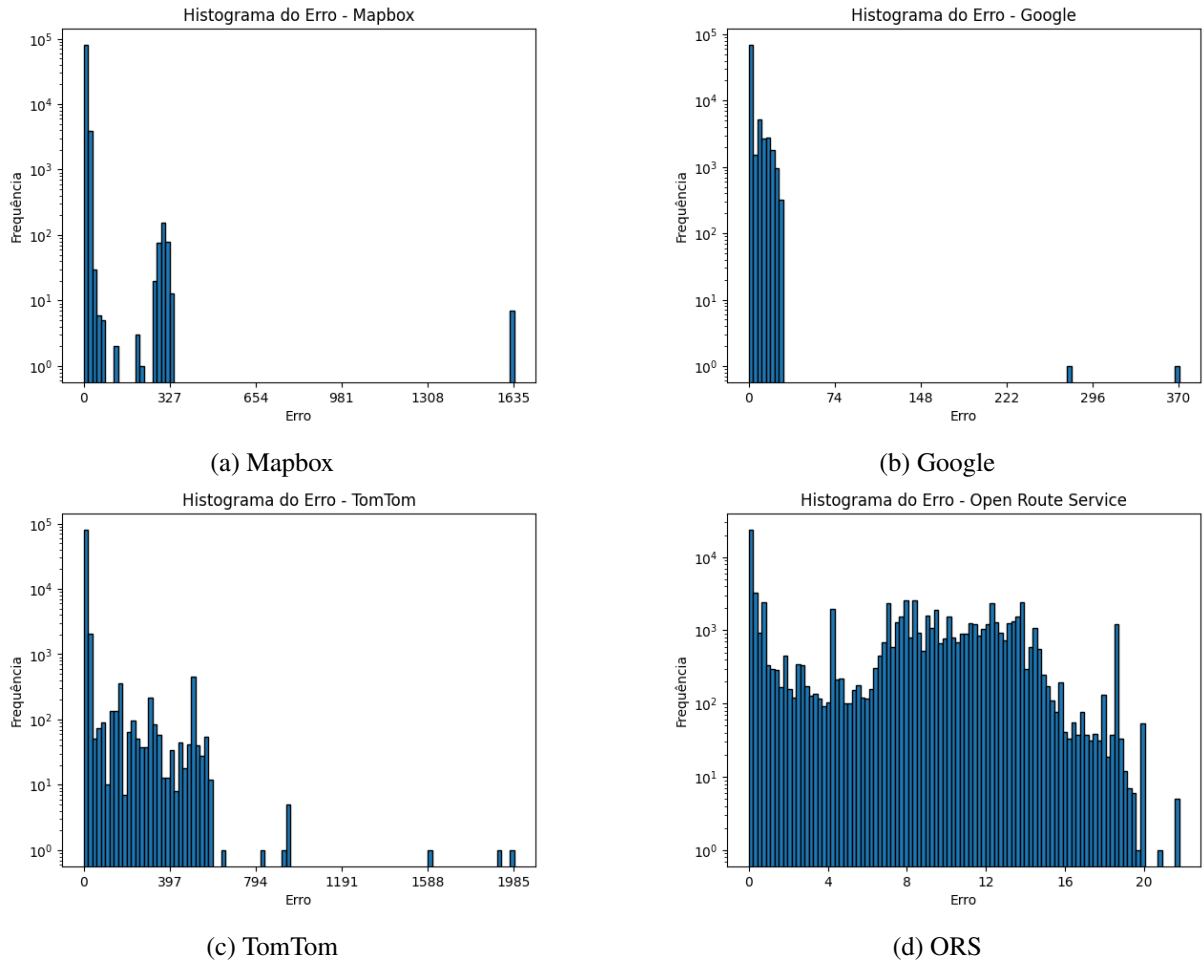


Figura 3.2 – Histogramas do erro das 4 APIs para o todos os dados de Belo Horizonte

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 da falha observada. Para melhor visualização, também adicionamos o limite da cidade como contorno do gráfico. As

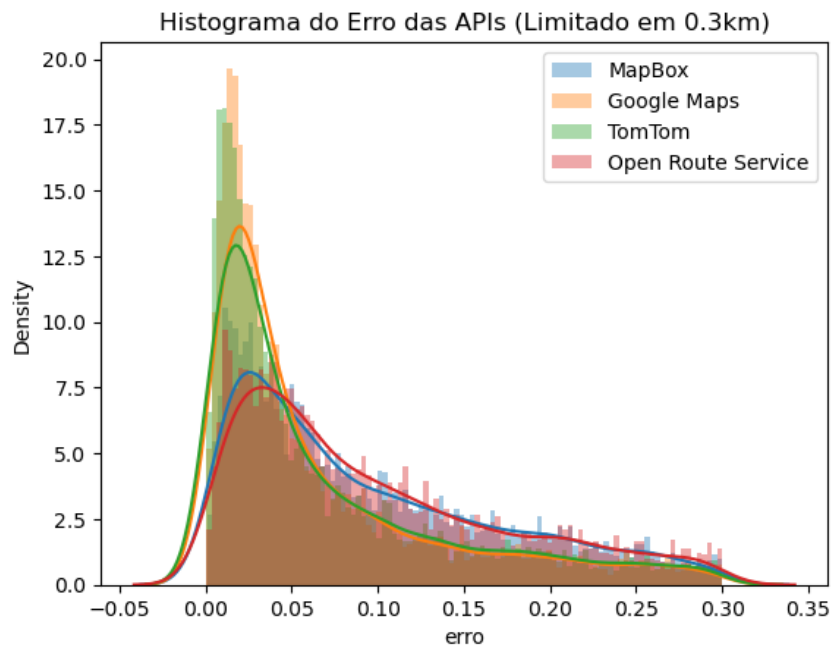


Figura 3.3 – Histograma comparativo de erro das APIs limitado em 300 metros para os dados de São Paulo

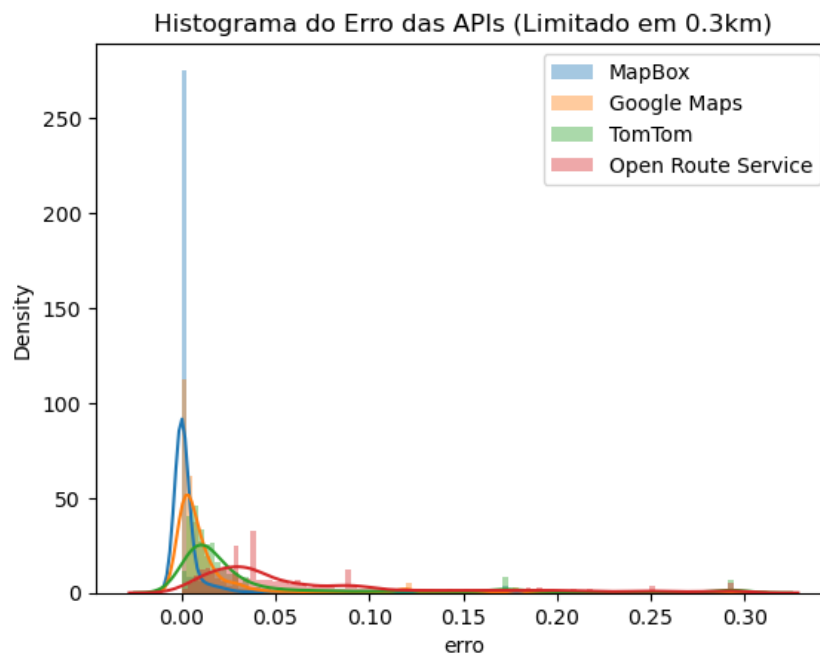


Figura 3.4 – Histograma comparativo de erro das APIs limitado em 300 metros para os dados de Belo Horizonte

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.

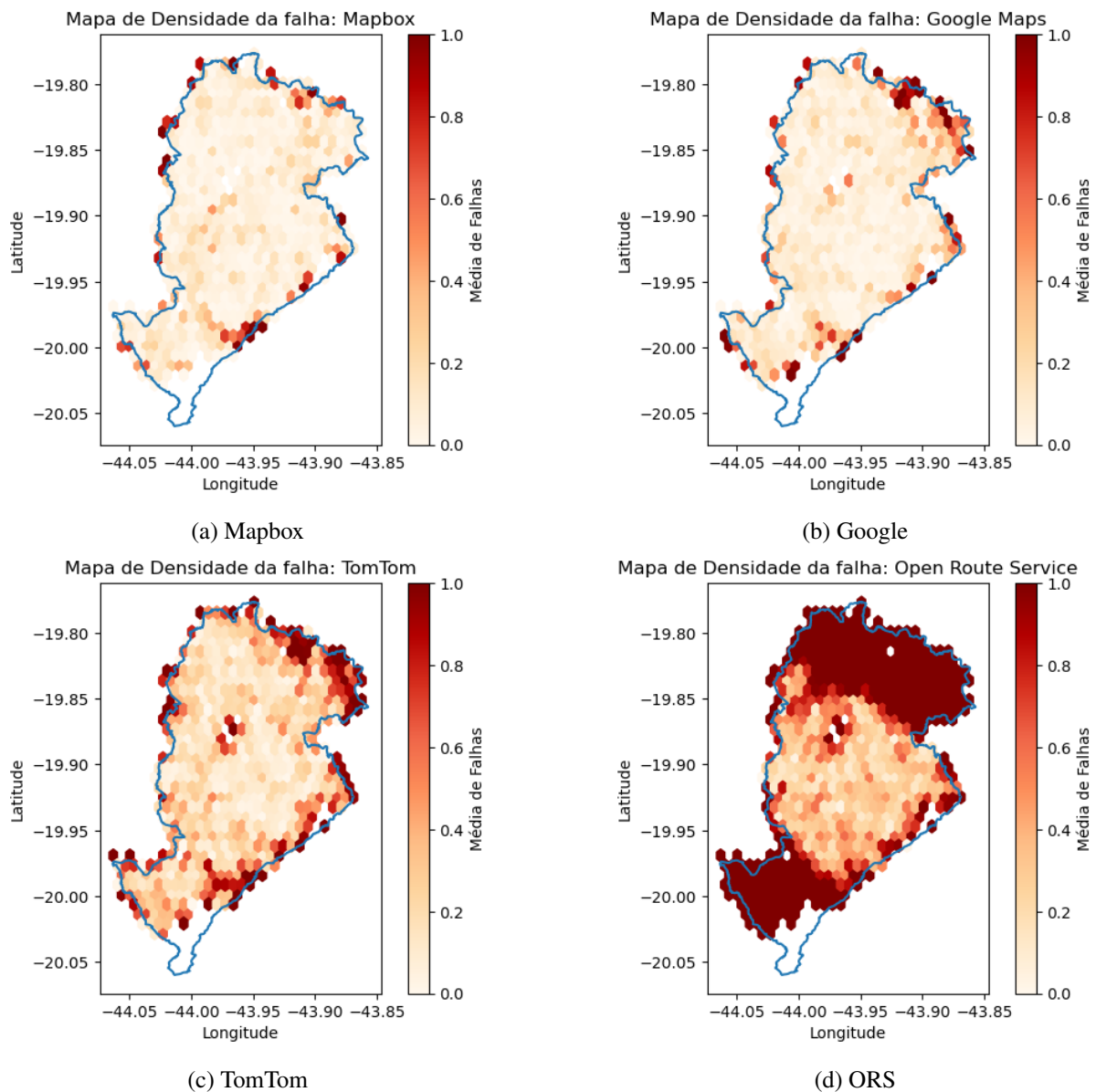


Figura 3.5 – Gráficos de falhas de cada API para os dados de Belo Horizonte

Nos gráficos de São Paulo também foi adicionado o contorno da cidade. No entanto, os dados são referentes à 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 apresentaram os piores resultados foram Mapbox e ORS. A ORS novamente se destacou

negativamente, repetindo os resultados de Belo Horizontes observados nos gráficos da Figura 3.5 e nas Tabelas 3.2 e 3.1.

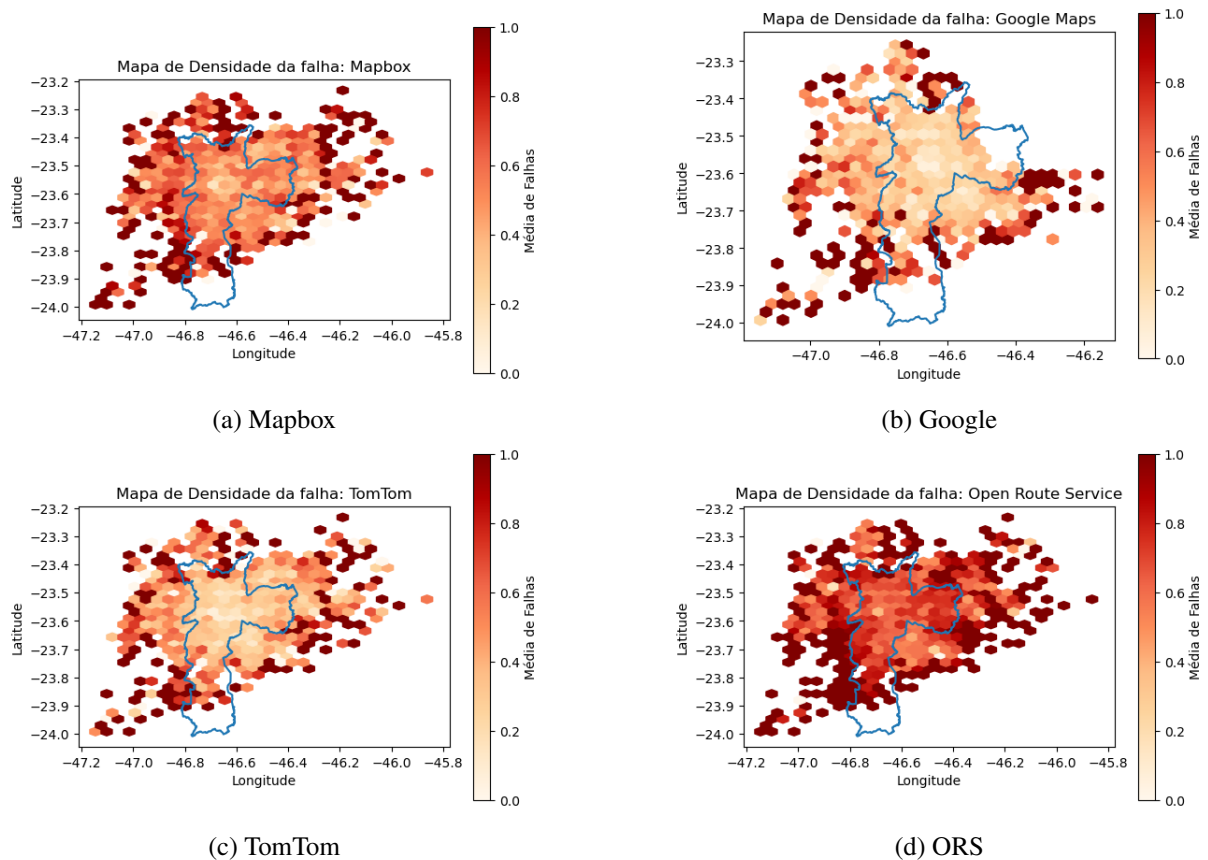


Figura 3.6 – Gráficos de falhas de cada API para os dados de Belo São Paulo

3.4 Relações entre erro e discrepância

Na tentativa de alcançar o nosso segundo objetivo, que é encontrar uma medida de discrepância que avalie a qualidade da geocodificação das APIs, foi realizada uma análise comparativa entre erro e discrepância. A medida escolhida para essa análise foi a distância para o ponto médio como descrito no Capítulo 2. Foram considerados apenas os endereços em que se tinha informação de todas as APIs. Depois de calcular a métrica 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 exhibe 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, a distância do ponto médio também tende a aumentar. De acordo com a Tabela 2.1 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.

Tabela 3.3 – Correlação de Pearson entre Erro e Distância do Ponto Médio para São Paulo

API	Distância até o Ponto Médio
Mapbox	0,8972
TomTom	0,8858
Google	0,3615
ORS	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. Nesses dados tivemos correlações fortes a muito fortes para o Mapbox e o TomTom. O Google e o ORS 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 indicar que a distância até o ponto médio pode ser uma medida substituta do erro em situações onde este não pode ser obtido, para algumas APIs. No entanto, essa medida não é consistente em relação às diferentes APIs, nem em relação às diferentes bases de dados. Portanto, essa medida deve ser utilizada com cautela e mais avaliações são necessárias.

Tabela 3.4 – Correlação de Pearson entre Erro e Distância do Ponto Médio para Belo Horizonte

API	Distância até o Ponto Médio
Mapbox	0,7764
TomTom	0,9873
Google	0,0754
ORS	0,2775

3.5 Experimentos de Formatação

Após realizar análises de erro e discrepância nas bases de São Paulo e Belo Horizonte, procedemos com uma amostragem de 5.000 endereços para ambas as bases. O propósito dessa amostragem foi conduzir uma série de experimentos para determinar a melhor formatação de entrada para cada API.

Após a geocodificação dos experimentos, calculamos o erro para cada endereço. Para cada API, computamos a média, mediana, desvio padrão, média aparada em 5%, taxa de resposta e taxa de acerto, como realizado anteriormente. Os resultados completos podem ser encontrados nas Tabelas B.1, B.2, B.3, B.4, B.7, B.5, e B.8 no Apêndice B do texto.

Para uma compreensão mais aprofundada dos resultados e uma análise comparativa, elaboramos duas tabelas para cada base. As tabelas representam a taxa de resposta e a taxa de acerto de cada experimento por API. Adicionamos também uma linha nas tabelas para mostrar a

variação máxima de resultado por API, a fim de avaliar o impacto da formatação para cada API. A variação máxima é calculada como a diferença entre a maior taxa e a menor taxa.

As Tabelas 3.5 e 3.6 mostram as taxas de resposta para as bases de Belo Horizonte e São Paulo, respectivamente. Observamos que todos os experimentos apresentaram excelentes resultados em termos de taxa de resposta, a maioria acima de 95% com exceção da TomTom com taxas acima de 84% para os dados de São Paulo. Não foi possível identificar uma diferença significativa entre as taxas de resposta nos experimentos. No entanto, em relação às APIs, a ORS, em geral, apresentou uma taxa de resposta ligeiramente inferior em comparação com as outras APIs para os dados de Belo Horizonte. Em relação aos dados de São Paulo, a mesma situação ocorre com a API TomTom. No entanto, os autores não consideram essa diferença significativa devido à sua pequena magnitude. Em relação à variação máxima, houve pouca variação em todos os experimentos e APIs, indicando que a formatação tem pouco ou nenhum impacto na resposta das APIs de geocodificação.

Tabela 3.5 – Taxa de resposta de cada API por experimento de Belo Horizonte

Experimento	MapBox	Google	TomTom	ORS
1	100,00	99,92	100,00	99,92
1b	99,94	99,96	99,98	95,26
2	100,00	99,98	99,94	99,06
2b	99,94	99,84	99,98	95,30
3	100,00	99,92	100,00	99,04
3b	99,68	100,00	99,88	99,06
4	100,00	99,90	99,98	100,00
4b	99,92	99,92	99,98	95,12
5	99,92	99,92	99,94	99,58
5b	99,76	99,88	99,92	99,98
Variação Máxima	0,32	0,16	0,12	4,88

Tabela 3.6 – Taxa de resposta de cada API por experimento de São Paulo

Experimento	MapBox	Google	ORS	TomTom
1	97,50	99,98	99,86	85,48
2	97,78	99,86	99,50	85,52
3	99,20	99,88	99,88	85,52
4	97,84	99,88	99,00	85,48
5	98,00	99,90	99,96	84,14
Variação Máxima	1,7	0,12	0,96	1,38

A Tabela 3.7 apresenta a taxa de acerto dos dados de Belo Horizonte. Observa-se que as APIs que alcançaram os melhores resultados foram a Google e MapBox, com destaque para a MapBox, que obteve a melhor taxa de acerto em todos os experimentos. Por outro lado, a ORS teve um impacto negativo, sendo a API com a menor taxa de acerto em relação às outras

APIs, com valores baixíssimos variando de 1,40% a 41%, resultados que replicam os obtidos nos experimentos anteriores. A API TomTom apresentou um resultado mediano, mantendo taxas acima de 50% em todos os experimentos. Em relação à variação máxima, é possível notar que ela aumentou, comparado com os resultados obtidos nas Tabelas 3.5 e 3.6, indicando que a formatação pode impactar a taxa de acerto das APIs. A API em que essa variação foi maior foi a ORS, sugerindo que essa API seja mais sensível às modificações de formatação de entrada.

Ao comparar os experimentos entre si, nota-se que a melhor taxa de acerto de cada API não foi consistente. Os experimentos com as maiores taxas de acerto para cada API foram:

- MapBox: Experimento 1;
- Google: Experimento 1b;
- TomTom: Experimento 1b;
- ORS: Experimento 3.

Apesar disso, percebe-se que, para a maioria das APIs, os melhores resultados estão associados a formatos que seguem o padrão do código postal brasileiro conforme a Tabela 2.3, diferenciando-se apenas pela inclusão ou não do bairro no formato de entrada. Em relação à inclusão do bairro nos experimentos, ocorreu um fenômeno peculiar. Para as APIs MapBox e ORS, a adição do bairro diminuiu a taxa de acerto, ou seja, piorou os resultados. O mesmo não ocorreu com a API Google, que, na maioria dos experimentos, apresentou uma melhoria considerável ao adicionar o bairro. Quanto à API TomTom, houve uma melhoria ao adicionar o bairro a partir do experimento 1, resultando na melhor taxa de acerto; entretanto, para todos os outros experimentos, ocorreu o oposto.

Tabela 3.7 – Taxa de acerto de cada API por experimento de Belo Horizonte

Experimento	MapBox	Google	TomTom	ORS
1	85,06	72,72	52,80	26,46
1b	80,88	80,64	56,34	15,62
2	82,46	73,30	55,66	22,28
2b	79,82	78,02	53,76	5,46
3	84,00	73,38	55,82	40,06
3b	80,56	78,30	53,92	39,08
4	84,66	73,26	55,32	1,46
4b	79,86	77,78	53,76	6,72
5	83,80	73,32	55,78	1,48
5b	81,00	72,92	53,92	24,78
Variação Máxima	5,24	7,92	3,54	38,6

A Tabela 3.8 apresenta a taxa de acerto para os experimentos em São Paulo. Como mencionado anteriormente, a base de São Paulo não continha informações sobre bairros, resultando na análise apenas dos experimentos de formatação sem bairro.

Tabela 3.8 – Taxa de acerto de cada API por experimento de São Paulo

Experimento	MapBox	Google	TomTom	ORS
1	41,78	50,80	44,94	28,94
2	37,04	48,54	44,96	5,30
3	41,26	48,42	45,02	0,14
4	40,90	48,42	44,90	14,94
5	40,10	48,00	44,40	1,04
Varição Máxima	4,74	2,80	0,62	28,80

Com esses resultados, observamos que as APIs com as maiores taxas de acerto foram Google, TomTom e Mapbox como visto nos experimentos em Belo Horizonte. No entanto, nos experimentos em São Paulo, a Google teve um desempenho superior, sendo o destaque para essa base, seguida pela TomTom e Mapbox. A ORS, por outro lado, teve novamente um desempenho ruim, com taxas de acerto muito baixas, variando de 1% a 29%. Em relação à variação máxima, houve algumas diferenças em comparação com os dados de Belo Horizonte. A variação, em geral, foi mais baixa, indicando que a formatação de entrada tem menos impacto na taxa de acerto para essas APIs. No entanto, ainda houve variações consideráveis, o que reforça a ideia de que buscar uma melhor formatação de entrada pode beneficiar a qualidade da geocodificação. A ORS manteve uma variação máxima alta, o que parece confirmar a hipótese de que essa API é bastante impactada pela formatação de entrada.

Em relação aos experimentos, o Experimento 1 obteve a melhor taxa de acerto para quase todas as APIs, reforçando a hipótese de que a melhor formatação de entrada para as requisições de API no território nacional segue o código postal brasileiro. Para a API TomTom no entanto, o experimento com melhor resultado foi o Experimento 2. Porém com um valor bem similar ao Experimento 1. Em geral, para a base de São Paulo, as APIs apresentaram resultados inferiores, porém medianos, o que também foi observado na análise das APIs com a base completa.

Após a construção das tabelas, decidimos criar BoxPlots para comparar os resultados dos experimentos e das APIs.

Para os dados de Belo Horizonte, começamos construindo inicialmente boxplots de experimento por erro, com cada API representada por um boxplot. A Figura 3.7 mostra esse boxplot. Embora seja possível observar a presença de valores de erro extremos em todas as APIs, chegando a 3000 km, a presença de outliers dificulta uma análise mais detalhada.

Consequentemente, geramos o boxplot sem outliers, representado na Figura 3.8. Nela, conseguimos observar claramente o desempenho da ORS, que obteve os piores resultados. Os erros da ORS, sem outliers, ultrapassam 400 km, apesar do boxplot estar contido na faixa de 200

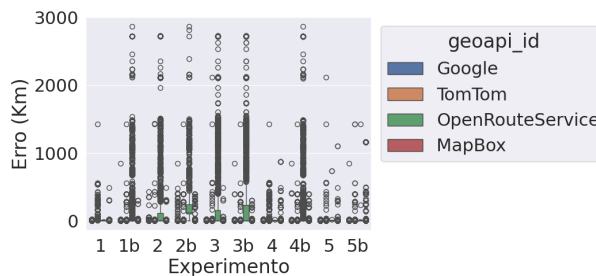


Figura 3.7 – Boxplot de Experimentos por Erro, com todas as APIs avaliadas

km. De qualquer forma, a presença da ORS prejudica a comparação com as outras APIs, que era o objetivo inicial na criação desses boxplots.

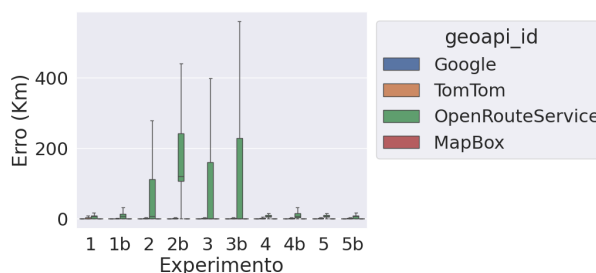


Figura 3.8 – Boxplot de Experimentos por Erro, com todas as APIs avaliadas e sem Outliers

Decidimos então criar um boxplot que não contivesse dados da ORS, como mostrado na Figura 3.9. Nesse gráfico, podemos observar que o erro é limitado a 8 km, um resultado consideravelmente melhor do que os obtidos anteriormente. Além disso, notamos que a API TomTom tem uma barra maior, indicando uma faixa de erro maior, conforme esperado de acordo com a Tabela 3.7. Observamos também que as APIs Google e MapBox têm barras de tamanho similar, sendo a MapBox ligeiramente menor. Um ponto interessante é que, ao comparar os experimentos, o destaque é o Experimento 1b, com uma barra menor em todas as APIs. Este experimento segue o código postal brasileiro, mas com a adição de bairro, indicando novamente que essa formatação pode gerar bons resultados para geocodificação.

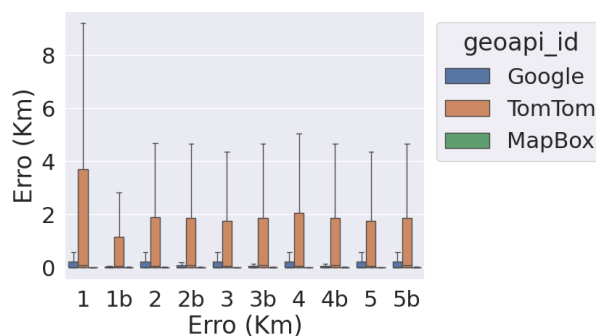


Figura 3.9 – Boxplot de Experimentos por Erro, com todas as APIs avaliadas exceto ORS e sem Outliers para os dados de Belo Horizonte

Para a base de São Paulo, foram construídos boxplots de forma semelhante. A Figura 3.10 mostra o boxplot de experimento por erro com todas as APIs, sem remover os outliers.

Assim como nos dados de Belo Horizonte, esse boxplot fornece pouca informação útil. É possível observar que, mais uma vez, o erro atinge a faixa de 3.000 km e que a ORS tem uma faixa de erro maior. No entanto, não é possível tirar muitas conclusões devido à presença de outliers.

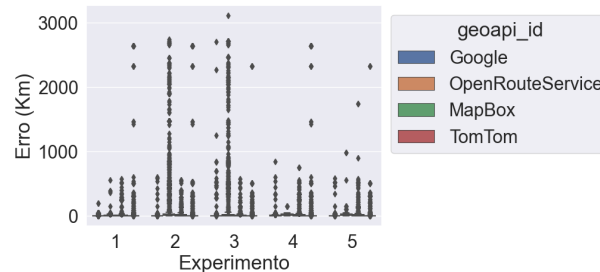


Figura 3.10 – Boxplot de Experimentos por Erro, com todas as APIs avaliadas para a amostra de São Paulo

Posteriormente, geramos boxplots sem outliers. A Figura 3.11 mostra esse boxplot. Observamos que a ORS teve o pior resultado, mas, ao contrário do Boxplot 3.8, ainda é possível distinguir os boxplots das outras APIs. Ou seja, o desempenho da ORS não foi tão discrepante como o de Belo Horizonte. Outro ponto importante é que neste boxplot é possível perceber uma grande diferença entre a Google e a Mapbox, com a Google obtendo resultados consideravelmente melhores. Em relação aos experimentos, o Experimento 1 foi o que obteve os melhores resultados. Novamente, o Experimento 1 segue o padrão do código postal do Brasil, o que parece confirmar a hipótese de que essa é a melhor formatação de entrada.

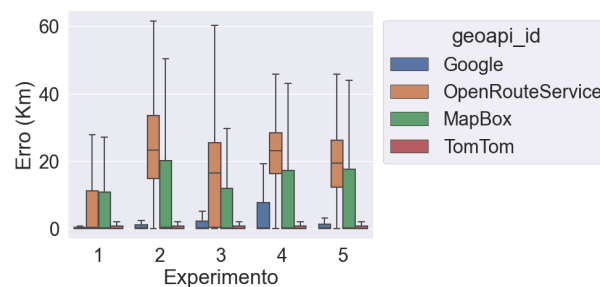


Figura 3.11 – Boxplot de Experimentos por Erro, com todas as APIs avaliadas e sem Outliers para a amostra de São Paulo

Os boxplots apresentados anteriormente foram cruciais para avaliar as APIs comparativamente. No entanto, faltava avaliar o comportamento dos experimentos dentro de cada API. Para isso, foram criados boxplots para cada API, onde o boxplot representa o experimento. As Figuras 3.12 e 3.13 mostram os boxplots de Belo Horizonte e São Paulo, respectivamente. Assim como nos boxplots anteriores, a presença de outliers prejudica a visualização, mas ainda é possível observar a faixa de erro em cada API. Abaixo, são apresentados os valores máximos de erro:

- MapBox - Belo Horizonte: Acima de 1000 km;
- Google - Belo Horizonte: Acima de 800 km;

- TomTom - Belo Horizonte: Acima de 2000 km;
- ORS - Belo Horizonte: Acima de 1000 km;
- MapBox - São Paulo: Acima de 1500 km;
- Google - São Paulo: Acima de 2500 km;
- ORS - São Paulo: Acima de 3000 km;

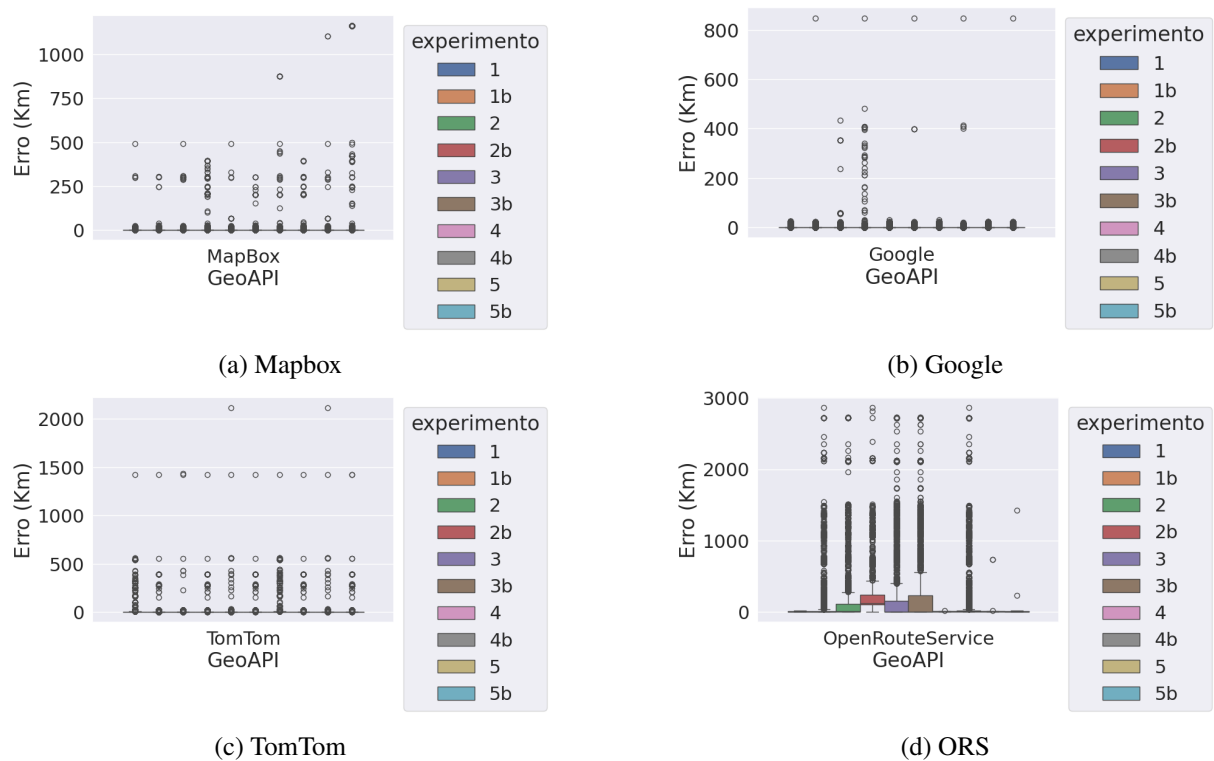


Figura 3.12 – Boxplots de Erro por API para cada uma das APIs avaliadas

Para aprimorar a visualização, decidimos remover os outliers de todos os boxplots.

As Figuras 3.14 e 3.15 apresentam os boxplots do erro da Mapbox para os dados de Belo Horizonte e São Paulo, respectivamente. Em ambos os boxplots, o Experimento 1 obteve os melhores resultados para a API MapBox. No entanto, é interessante observar que o Experimento 4 teve um desempenho muito semelhante ao Experimento 1 para ambas as bases. Analisando o boxplot de Belo Horizonte, é possível notar que a adição do bairro resultou em piora nos resultados para todos os experimentos, evidenciando os piores resultados em geral, conforme observado na Tabela 3.7. Quanto aos dados de São Paulo, os experimentos que se destacaram negativamente foram o 2, 3 e 5.

As Figuras 3.16 e 3.17 apresentam os boxplots para a API do Google referentes aos dados de Belo Horizonte e São Paulo, respectivamente. Ao analisar os dados de Belo Horizonte, observa-se uma melhora considerável em quase todos os experimentos nos casos em que o bairro

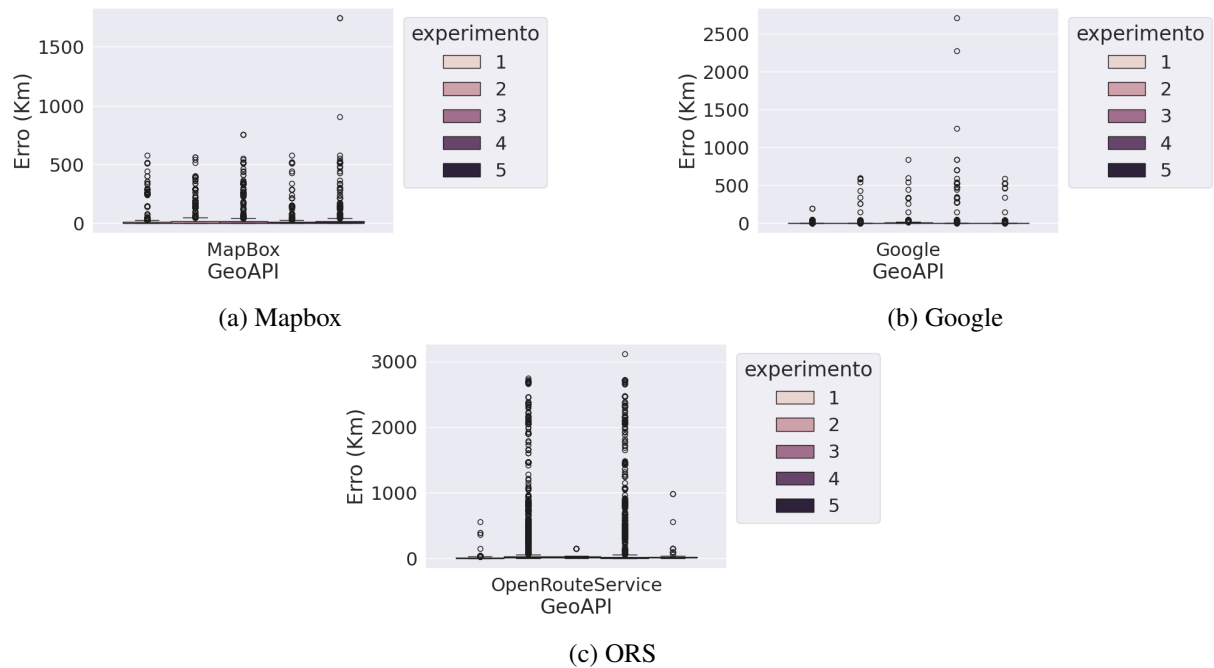


Figura 3.13 – Boxplots de Erro por API para cada uma das APIs avaliadas para a amostra de São Paulo

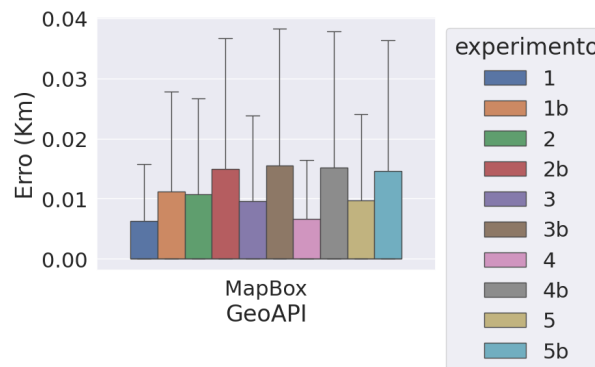


Figura 3.14 – Boxplot de Erro por API com todos os experimentos avaliados sem Outliers: Mapbox

foi adicionado. A exceção são os experimentos 5 e 5b, nos quais a diferença não é significativa. O experimento com melhor desempenho para os dados de Belo Horizonte foi o experimento 1b, enquanto para os dados de São Paulo, o destaque foi para o experimento 1. Esses resultados confirmam as conclusões extraídas das Tabelas 3.7 e 3.8, indicando que essa formatação é a mais eficaz.

Quanto aos piores resultados, para os dados de Belo Horizonte, as performances mais baixas foram observadas nos casos em que não havia informação de bairro, sem diferenças significativas entre eles. Já para os dados de São Paulo, a pior performance foi no experimento 3.

A Figura 3.18 apresenta o boxplot da API TomTom para os dados de Belo Horizonte. Observa-se que o experimento com melhor desempenho foi o 1b, enquanto o pior desempenho ocorreu no experimento 1, contrariando os resultados previamente obtidos. Os demais experimentos apresentaram resultados semelhantes, e não é possível afirmar que há uma melhora ao adicionar o bairro nesses experimentos, exceto nos casos do experimento 1 e 1b, nos quais houve

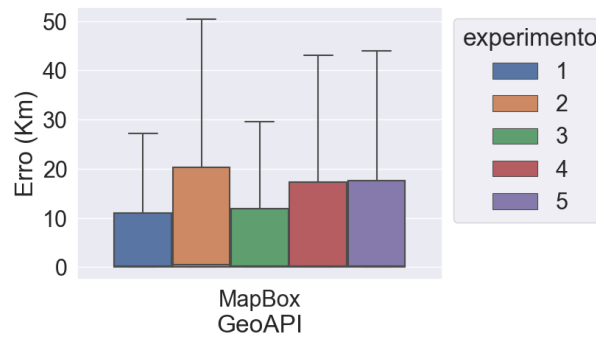


Figura 3.15 – Boxplot de Erro por API com todos os experimentos avaliados sem Outliers para a amostra de São Paulo: Mapbox

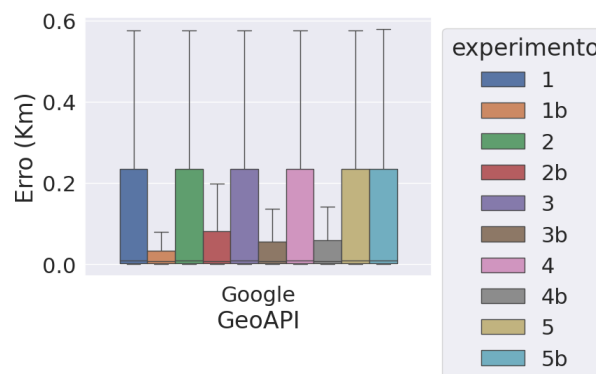


Figura 3.16 – Boxplot de Erro por API com todos os experimentos avaliados sem Outliers: Google

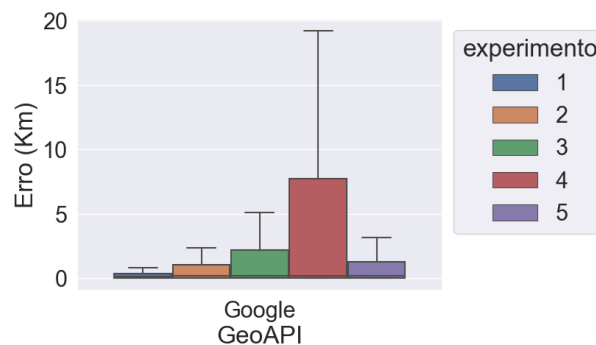


Figura 3.17 – Boxplot de Erro por API com todos os experimentos avaliados sem Outliers para a amostra de São Paulo: Google

uma melhora significativa ao adicionar o bairro.

Para concluir a análise dos boxplots, foram gerados dois gráficos para a API ORS para os dados de Belo Horizonte, conforme apresentado na Figura 3.19, e para os dados de São Paulo, na Figura 3.20. Ao examinar os dados de Belo Horizonte, observa-se que os experimentos que se destacaram positivamente foram os 1, 1b, 4, 4b, 5 e 5b, com pouca diferença entre eles. Por outro lado, os experimentos 2b e 3b obtiveram os piores resultados para os dados de Belo Horizonte. No contexto dos dados de São Paulo, o experimento 1 se destacou positivamente, enquanto o experimento 2 teve o pior desempenho. Os demais experimentos apresentaram resultados semelhantes.

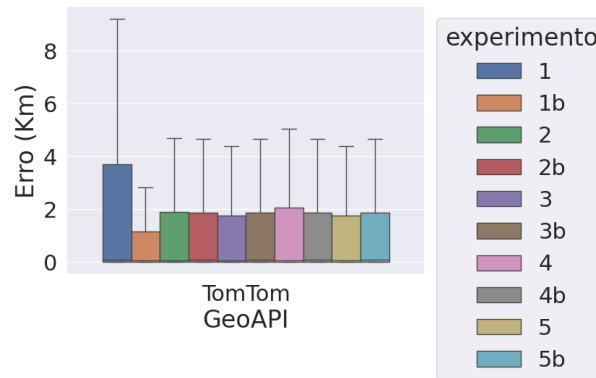


Figura 3.18 – Boxplot de Erro por API com todos os experimentos avaliados sem Outliers: TomTom

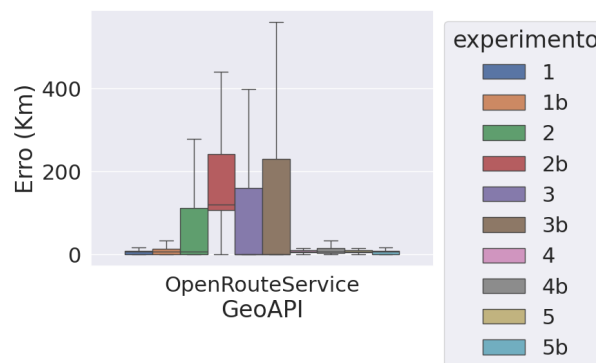


Figura 3.19 – Boxplot de Erro por API com todos os experimentos avaliados sem Outliers: ORS

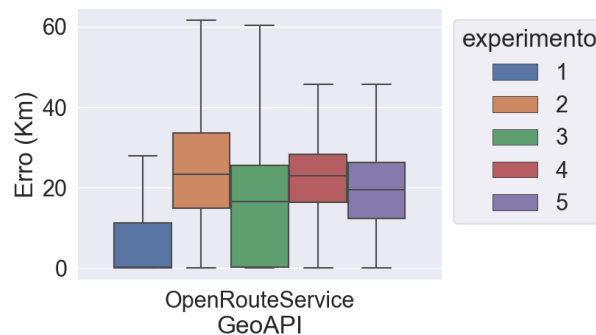


Figura 3.20 – Boxplot de Erro por API com todos os experimentos avaliados sem Outliers para a amostra de São Paulo: ORS

3.5.1 Distribuição Espacial das Falhas dos Experimentos de Formatação

Por fim, realizamos uma análise espacial de falhas para entender de que forma a formatação da entrada pode contribuir para a falha de forma espacial. Para isso, criamos mapas hexagonais de falhas, como feito anteriormente, para cada API. Escolhemos os experimentos com menor e maior taxa de acerto de cada API a fim de visualizar com mais clareza a melhora a partir de formatações diferentes.

A Figura 3.21 mostra os mapas de falhas dos experimentos com menor e maior taxa de acerto para os dados de Belo Horizonte e para a API Mapbox. Nessa figura é possível observar que no mapa de maior taxa de acerto há uma maior uniformidade e a maior parte do gráfico tem

taxa de falha inferior a 20%. Já no gráfico de menor taxa de acerto há menor uniformidade e as regiões variam de taxas mais baixas (menor que 20%) até taxas mais altas (maior que 80%). É relevante observar também que as taxas de falha maiores aparentam estar em regiões similares como bordas, o que nos faz pensar que a alta taxa de erro é ligada à região e não somente à formatação de entrada. Apesar disso, até nessas regiões com mais erro, é possível observar uma melhora significativa a partir da formatação com melhor resultado.

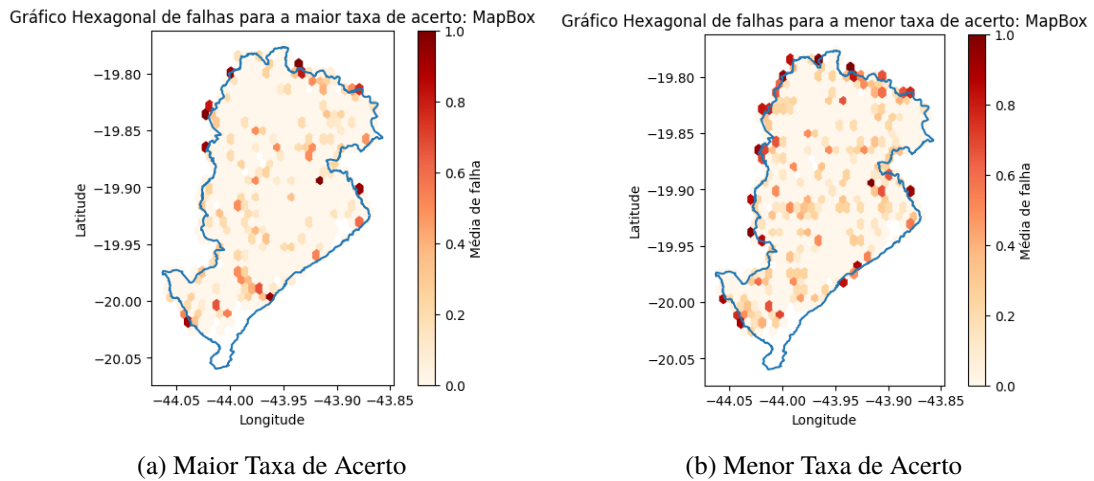


Figura 3.21 – Gráficos de falhas para a maior e menor taxas de acerto dos experimentos para os dados de Belo Horizonte: MapBox

A Figura 3.22 mostra os gráficos de falhas de menor e maior taxa de acerto para a API Mapbox nos dados de São Paulo. Ao contrário dos dados de Belo Horizonte, os dados de São Paulo obtiveram taxas de acerto similares na maior parte das APIs, conforme indicado na Tabela 3.8. Sendo assim, os gráficos são semelhantes. Apesar disso, é possível notar que o gráfico de maior taxa de acerto apresenta uma melhora em relação ao de menor taxa de acerto, com algumas regiões mais claras, principalmente na parte central do mapa. Isso indica que a formatação de entrada pode impactar a geocodificação da Mapbox.

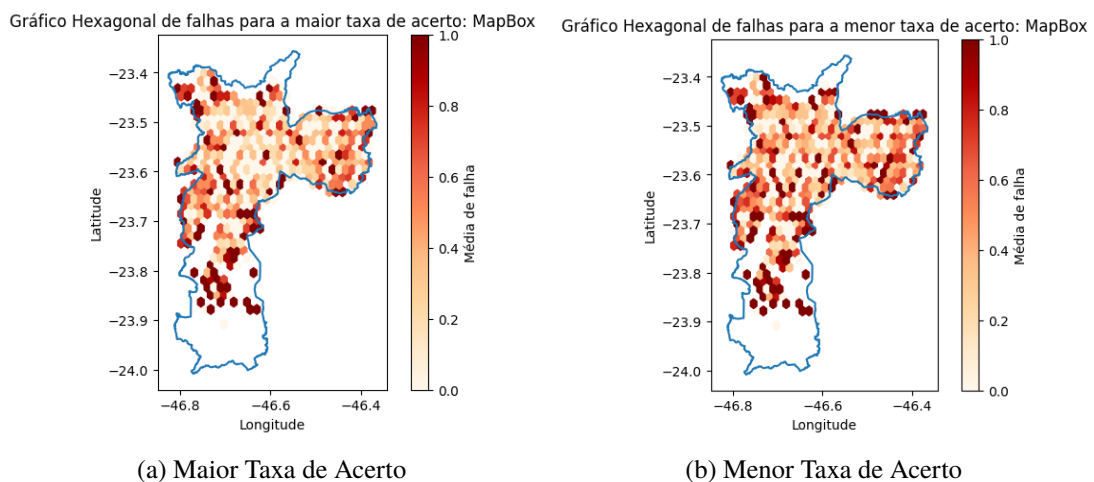


Figura 3.22 – Gráficos de falhas para a maior e menor taxas de acerto dos experimentos para os dados de São Paulo: MapBox

A Figura 3.23 mostra os gráficos de falhas para os experimentos de maior e menor taxa de acerto da API Google Maps para os dados de Belo Horizonte. Da mesma forma que acontece nos gráficos da Mapbox, os gráficos para o experimento de maior taxa de acerto apresentam uma uniformidade maior em comparação com o de menor taxa de acerto. Além disso, é interessante observar que existem algumas zonas de onde a taxa de falha é grande em ambos os gráficos. Essas zonas são nas fronteiras da cidade e na parte superior do gráfico. Porém, no gráfico de menor taxa de acerto há uma acentuação dessas regiões com altas taxas de acerto, corroborando com a hipótese de que a formatação realmente contribuiu para a melhora da taxa de falha.

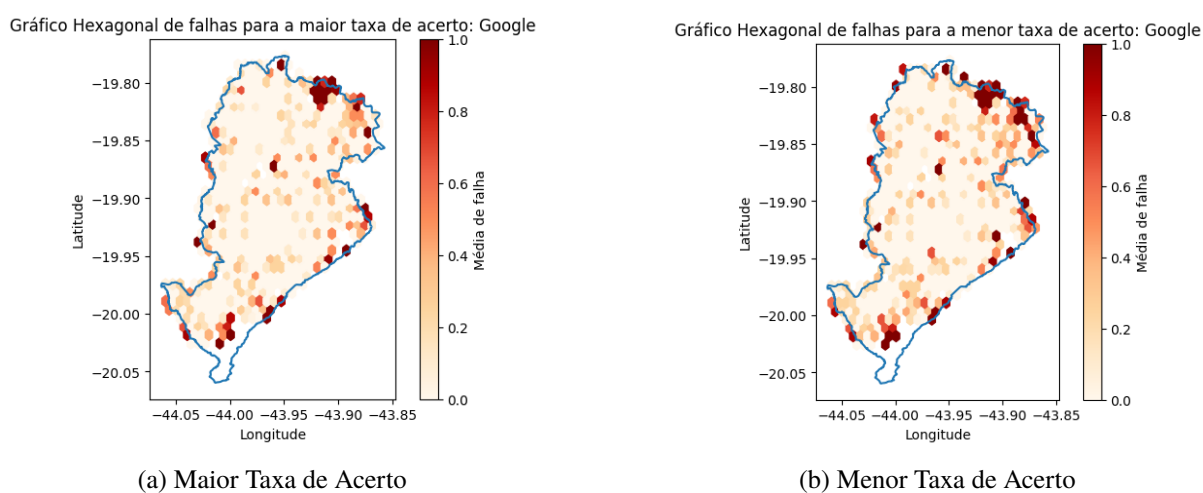


Figura 3.23 – Gráficos de falhas para a maior e menor taxas de acertados experimentos para os dados de Belo Horizonte: Google Maps

A Figura 3.24 mostra os gráficos de falhas para os experimentos de menor e maior taxa de acerto da API Google Maps para os dados de São Paulo. Assim como ocorreu com os gráficos da Mapbox na Figura 3.22 a diferença é pequena, porém é possível notar uma melhora na parte superior do gráfico. Nessa parte, o gráfico de maior taxa de acerto está mais claro. Também é possível notar que as regiões que apresentam falha maior são consistentes nos dois gráficos, porém no gráfico com menor taxa de acerto a falha é mais acentuada. O que vai de encontro com os resultados encontrados com os dados de Belo Horizonte para essa API.

A Figura 3.24 mostra os gráficos de falhas para os experimentos de menor e maior taxa de acerto da API Google Maps para os dados de São Paulo. Assim como ocorreu com os gráficos da Mapbox na Figura 3.22, a diferença é pequena, porém é possível notar uma melhora na parte superior do gráfico. Nessa parte, o gráfico de maior taxa de acerto está mais claro. Também é possível notar que as regiões que apresentam maior falha são consistentes nos dois gráficos, porém, no gráfico com menor taxa de acerto, a falha é mais acentuada. Isso está em linha com os resultados encontrados nos dados de Belo Horizonte para essa API.

A Figura 3.25 mostra os gráficos de falha dos experimentos de maior e menor taxa de acerto da API TomTom para os dados de Belo Horizonte. Nessa figura, ao contrário das outras APIs, não é possível notar diferença significativa entre os experimentos com maior e menor taxa

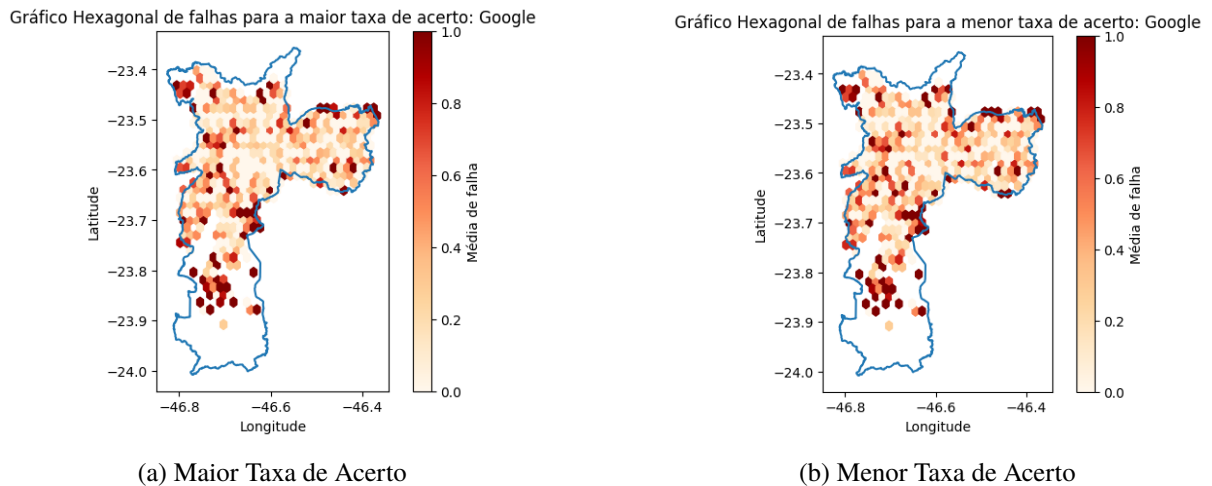


Figura 3.24 – Gráficos de falhas para a maior e menor taxas de acertados experimentos para os dados de São Paulo: Google Maps

de acerto. Isso condiz com as taxas de acerto da TomTom para os experimentos pois estavam todas próximas. Em comparação com as outras APIs, os dois gráficos apresentam uma certa uniformidade, com valores abaixo de 40% de falha. Algumas regiões aparentam taxas maiores, que são as regiões inferior, superior e de fronteira. Em geral, TomTom apresentou resultados medianos em maior parte dos gráficos. Isso pode indicar que para a API TomTom, a ordem em que os campos da entrada estão não impactam na geocodificação.

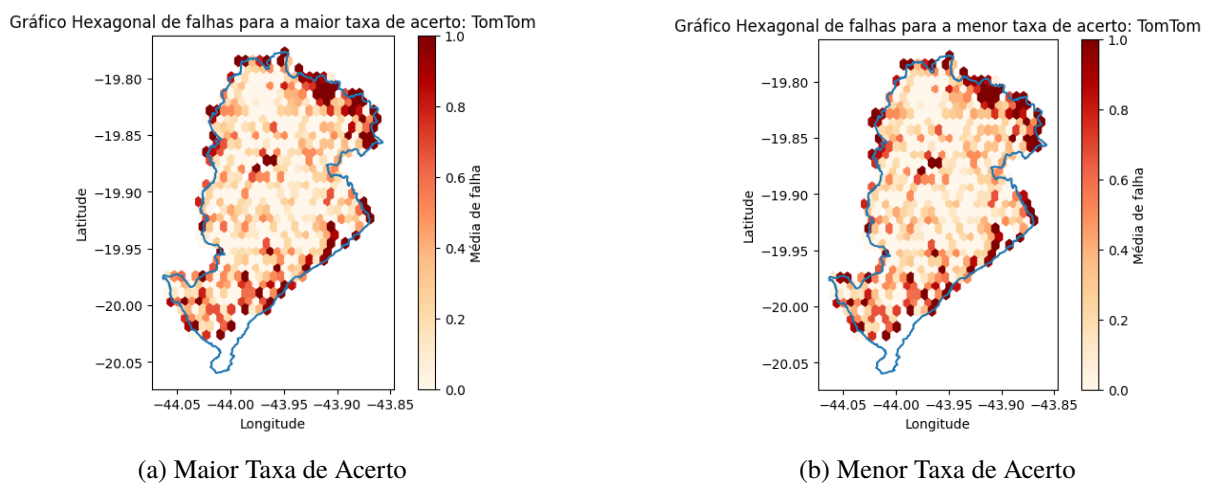


Figura 3.25 – Gráficos de falhas para a maior e menor taxas de acertados experimentos para os dados de Belo Horizonte: TomTom

A Figura 3.26 apresenta os gráficos de falhas para os experimentos de menor e maior taxa de acerto para a API TomTom. Assim como ocorreu para os dados de Belo Horizonte, os gráficos não apresentaram diferença significativa, o que corrobora a hipótese de que esta API não é impactada pela ordem dos campos de entrada da mesma forma que as outras APIs analisadas. Apesar disso, neste gráfico podemos notar que a API apresentou uma melhora em relação aos dados de Belo Horizonte, com uma aparência mais uniforme e com valores menores que 40%.

É possível notar também que algumas regiões apresentam falhas maiores, principalmente as regiões inferior e nas bordas do gráfico.

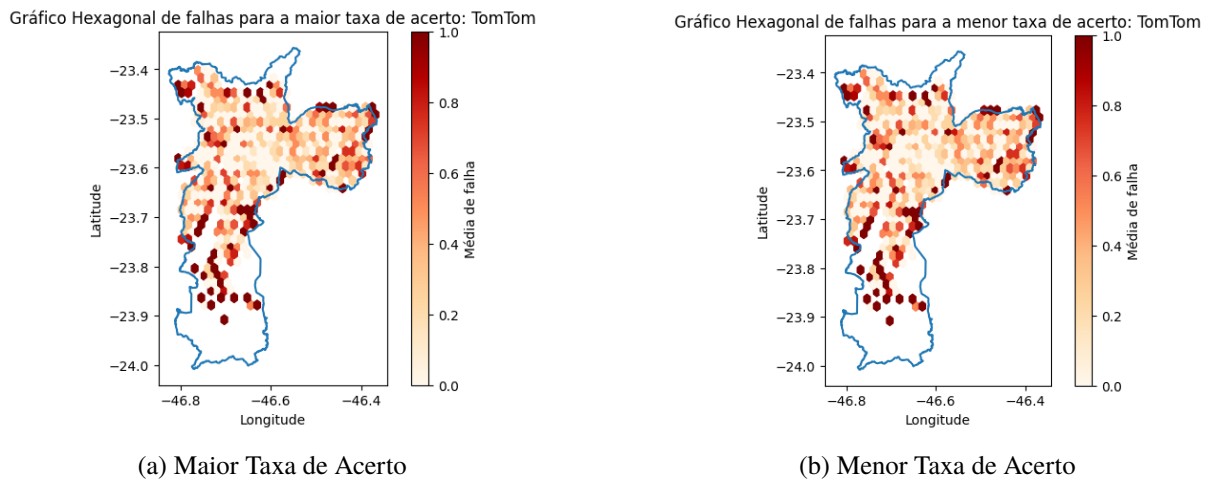


Figura 3.26 – Gráficos de falhas para a maior e menor taxas de acertados experimentos para os dados de São Paulo: TomTom

A Figura 3.27 mostra os gráficos de falha dos experimentos de maior e menor taxa de acerto da API Open Route Service para os dados de Belo Horizonte. Nessa figura, apesar das outras APIs, há pouca uniformidade no gráfico de maior taxa, inclusive apresentando muitos valores de taxas altas em praticamente todas as regiões do gráfico. Já o gráfico de menor taxa de acerto, apresenta em quase sua totalidade valores de taxa de falha próximos do máximo. Ou seja, para a API ORS, para esses dados, a formatação de entrada é crucial, pois a depender da escolhida, a API pode errar quase todas as solicitações. Apesar disso, mesmo no experimento de maior taxa, há valores altos de erro, resultado que concorda com os obtidos anteriormente.

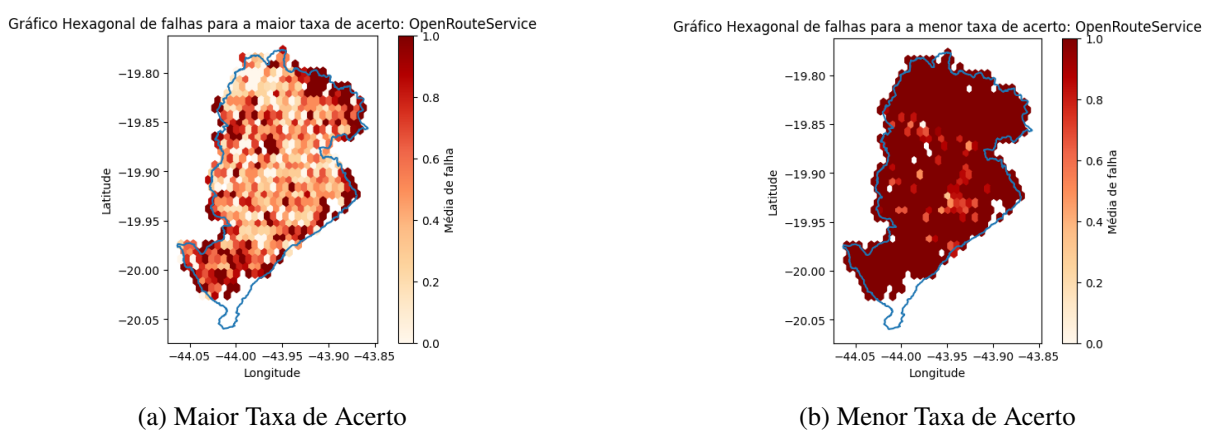
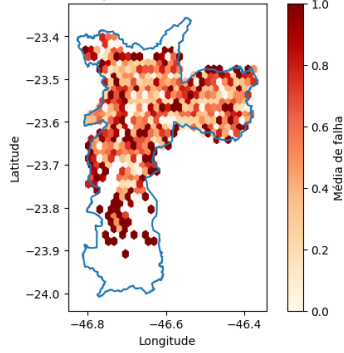


Figura 3.27 – Gráficos de falhas para a maior e menor taxas de acerto dos experimentos para os dados de Belo Horizonte: Open Route Service

A Figura 3.28 mostra os gráficos de falhas para os experimentos de menor e maior taxa de acerto da API ORS. De forma similar aos dados de Belo Horizonte, é possível visualizar uma melhora significativa no gráfico de maior taxa de acerto em comparação ao de menor taxa de acerto. Nele, existem regiões onde a taxa de falha é menor, ou seja, regiões mais claras. Ao

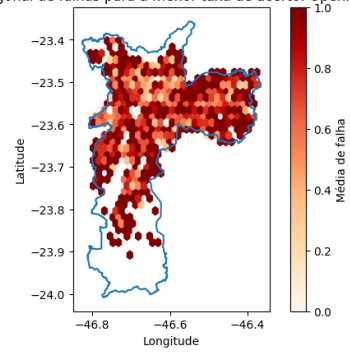
contrário do gráfico de menor taxa de acerto, que apresenta taxas de falhas muito altas em quase todo o gráfico. Apesar disso, o gráfico de maior taxa de acerto apresenta pouca uniformidade em comparação com as outras APIs, o que está de acordo com os resultados obtidos na Tabela 3.8.

Gráfico Hexagonal de falhas para a maior taxa de acerto: OpenRouteService



(a) Maior Taxa de Acerto

Gráfico Hexagonal de falhas para a menor taxa de acerto: OpenRouteService



(b) Menor Taxa de Acerto

Figura 3.28 – Gráficos de falhas para a maior e menor taxas de acerto dos experimentos para os dados de São Paulo: Open Route Service

4 Considerações Finais

A presente monografia apresentou uma avaliação das APIs Mapbox, TomTom, Open Route Service e Google Maps para as cidades de Belo Horizonte e São Paulo.

O trabalho tinha três objetivos principais. O primeiro foi realizar uma análise da qualidade da geocodificação das APIs, utilizando o erro como medida para aferir essa qualidade. O segundo foi verificar se uma medida de discrepância se relaciona com o erro da geocodificação, de modo que este possa ser substituído. E o terceiro foi verificar se a formatação da entrada para a geocodificação pode impactar na qualidade da mesma.

Em relação ao primeiro objetivo, as APIs que apresentaram melhores resultados foram Mapbox, Google Maps e TomTom. O desempenho dessas três APIs variou entre as bases de dados, sendo a ordem, do melhor desempenho para o pior, Mapbox, Google Maps e TomTom para os dados de Belo Horizonte; e Google Maps, TomTom e Mapbox para os dados de São Paulo. Por outro lado, o desempenho da API ORS foi constante, apresentando um desempenho insatisfatório em ambas as bases.

Considerando os resultados de cada base individualmente, a base de Belo Horizonte obteve resultados melhores em geral, com taxas de acerto altas para a maioria das APIs. Já a base de São Paulo teve resultados de médios a baixos para as APIs. Isso pode ter ocorrido devido a características particulares das bases ou problemas na aplicação que solicita e coleta as geocodificações. Mais análises são necessárias para chegar a uma conclusão concreta em relação a isso.

Para o segundo objetivo, foi escolhida a distância ao ponto médio como medida de discrepância. Ela se mostrou promissora como uma medida substituta ao erro para todas as APIs, com exceção da Google Maps, para ambas as bases, e da ORS, para a base de São Paulo. Para aquelas APIs onde há indicativo de relação, foram obtidas correlações fortes a muito fortes entre a medida e o erro. Esse resultado é um forte indicativo de que a medida pode ser considerada também uma medida de qualidade da geocodificação para essas APIs. Porém, mais estudos são necessários para confirmar essa hipótese e comparar os resultados com outras medidas para verificar se existem outras com desempenho semelhante ou superior. Em relação àquelas em que não houve relação, deve-se investigar se isso ocorre devido aos algoritmos empregados por elas ou se existe outra medida que apresenta essa relação.

Por fim, para alcançar o terceiro objetivo, foi realizada uma análise comparativa da qualidade da geocodificação de 15 experimentos, de forma similar à feita para o primeiro objetivo. Esses experimentos se diferenciavam entre si pela ordem em que as informações do endereço eram posicionadas na entrada da API. Para a base de São Paulo, foram realizados 5 experimentos, e para a base de Belo Horizonte foram realizados 10, que são os mesmos 5 de São Paulo, com

a adição de mais 5 experimentos incluindo a informação de bairro. Para ambas as bases e a maioria das APIs, os experimentos com o formato de entrada similar ao padrão do código postal brasileiro mostraram melhores resultados. Isso nos leva a crer que esse formato é o mais indicado para a entrada das APIs de geocodificação. Em relação à adição de bairro, os resultados foram conflitantes. Para algumas APIs, houve melhora ao adicionar o bairro; para outras, não houve diferença; e, para outras, houve uma piora. Em geral, a adição do bairro parece ser uma boa opção para a geocodificação, porém mais análises são necessárias.

Apesar da avaliação extensa da geocodificação das APIs, a monografia apresentou algumas limitações. A primeira é o fato de avaliar apenas duas cidades brasileiras da mesma região, o Sudeste. Outras regiões podem ter resultados diferentes e, assim, levar a outras conclusões. Outra limitação é a utilização de uma única métrica de discrepância, o que não permite a comparação com outras métricas para determinar qual é a melhor para o objetivo proposto. Por fim, avaliamos apenas a inclusão do bairro e em uma única base de dados. Seria interessante incluir outras informações e avaliar outras bases de dados para obter conclusões mais abrangentes.

Como trabalhos futuros, pode-se expandir as análises de qualidade, trabalhando com mais bases de dados de outras cidades importantes brasileiras ou com bases de locais diversos. Além disso, poderiam ser expandidas as análises de discrepância, adicionando outras medidas e comparando com os resultados obtidos com a distância ao ponto médio. Em relação aos experimentos, pode-se avaliar outras bases com a informação de bairro, para verificar com mais profundidade o impacto dessa adição, ou adicionar outras informações não utilizadas neste trabalho, como CEP.

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

- 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.
- CEM. *Centro de Estudos da Metrópole*. 2024. Url<https://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>>.
- GCP. *Google Cloud Platform*. 2023. Url<https://cloud.google.com/>. Acesso em: [13 ago. 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. *Documentação da Google Maps Geocodin API*. 2023. Url<https://developers.google.com/maps/documentation/geocoding>. Acesso em: [04 nov. 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>>.
- KILIC, B.; GULGEN, F. Accuracy and similarity aspects in online geocoding services: A comparative evaluation for google and bing maps. *International Journal of Engineering and Geosciences (IJEG)*, Turkey, v. 5, n. 2, p. 109–119, June 2020. ISSN 2548-0960. Yildiz Technical University, Civil Engineering Faculty, Department of Geomatic Engineering, Istanbul, Turkey; (fgulgen/batuhank@yildiz.edu.tr); ORCID 0000-0002-8754-9017, ORCID 0000-0002-0529-8569. Disponível em: <<https://doi.org/10.26833/ijeg.629381>>.
- 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 tracks? 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 Ciência da Informação Geográfica*. Grupo A, 2013. ISBN 9788565837651. Disponível em: <<https://integrada.minhabiblioteca.com.br/#/books/9788565837651/>>.

MAPBOX. *Documentação da Mapbox Geocodin API*. 2023. Url<<https://docs.mapbox.com/api/search/geocoding/>. Acesso em: [04 nov. 2023].

MAPBOX. *Site da Mapbox*. 2024. Url<<https://www.mapbox.com/>. Acesso em: [25 jun. 2024].

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.

ORS. *Documentação da Open Route Service Geocodin API*. 2023. Url<<https://openrouteservice.org/dev/#/api-docs/geocode/search/get>. Acesso em: [04 nov. 2023].

ORS. *Site da Open Route Service*. 2024. Url<<https://openrouteservice.org/>. Acesso em: [25 jun. 2024].

PRODABEL. *Prodabel - Empresa de Informática e Informação da Prefeitura de Belo Horizonte*. 2023. Url<<https://prefeitura.pbh.gov.br/prodabel>. Acesso em: [13 ago. 2023].

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. *Relatório do Experimento de Entradas das APIs*. 2024. Url<<https://github.com/rcpsilva/UncertaintyQuantificationForGeocodingServices/blob/main/UndergraduateTheses/testes.pdf>. Acesso em: [21 jun. 2024].

TERRALAB. *TerraLAB - Laboratório de Capacitação e Desenvolvimento de Software*. 2024. Url<<http://www2.decom.ufop.br/terralab/>. Acesso em: [11 ago. 2023].

TOMTOM. *Documentação da TomTom Geocodin API*. 2023. Url<<https://developer.tomtom.com/geocoding-api/documentation/product-information/introduction>. Acesso em: [04 nov. 2023].

- TOMTOM. *Site da TomTom*. 2024. Url<https://www.tomtom.com/>. Acesso em: [25 jun. 2024].
- 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>>.
- ZHANG, Z.; BETHARD, S. A survey on geocoding: algorithms and datasets for toponym resolution. *Language Resources and Evaluation*, 2024. Accepted: 23 February 2024. Disponível em: <<https://doi.org/10.1007/s10579-024-09730-2>>.

Apêndices

APÊNDICE A – Abreviações utilizadas

ID	Expressão	Abreviatura	ID	Expressão	Abreviatura
0	Alameda	AL	1	Avenida	AV
2	Azinhaga	AZ	3	Bairro	BR
4	Beco	BC	5	Calçada	CC
6	Calçadinha	CCNH	7	Caminho	CAM
8	Casa	CS	9	Conjunto	CJ
10	Escadas	ESC	11	Escadinhas	ESCNH
12	Estrada	ESTR	13	Jardim	JD
14	Largo	LG	15	Loteamento	LOT
16	Parque	PQ	17	Pátio	PAT
18	Praça	PC	19	Praceta	PCT
20	Prolongamento	PRL	21	Quadra	QD
22	Rotunda	ROT	23	Rua	R
24	Transversal	TRANSV	25	Travessa	TV
26	Urbanização	URB	27	Vila	VL
28	Zona	ZN	29	Cave	CV
30	Direito	DTO	31	Esquerdo	ESQ
32	Frente	FT	33	Fundos	FDS
34	Habitação	HAB	35	Loja	LJ
36	Rés-do-chão	RC	37	Sobreloja	SLJ
38	Subcave	SCV	39	Apartamento	APTO
40	Bloco	BL	41	Edifício	EDF
42	Lote	LT	43	Torre	TR
44	Vivenda	VV	45	Alferes	ALF
46	Almirante	ALM	47	Arquiteto	ARQ
48	Brigadeiro	BRIG	49	Capitão	CAP
50	Comandante	CMDT	51	Comendador	COMEND
52	Conselheiro	CONS	53	Coronel	CEL
54	Dom	D	55	Dona	DA
56	Doutor	DR	57	Doutora	DR
58	Duque	DQ	59	Embaixador	EMB
60	Engenheira	ENG	61	Engenheiro	ENG
62	Frei	FR	63	General	GEN
64	Infante	INF	65	Marquês	MQ
66	Presidente	PRES	67	Professor	PROF
68	Professora	PROF	69	São	S
70	Sargento	SARG	71	Tenente	TEN
72	Visconde	VISC	73	Associação	ASS
74	Instituto	INST	75	Lugar	LUG
76	Ministério	MIN	77	Projetada	PROJ
78	Sala	SL	79	Sem Número	SN
80	Sociedade	SOC	81	Universidade	UNIV

Tabela A.1 – Tabela de Abreviações

APÊNDICE B – Tabelas dos experimentos de formatação completas

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

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