Proposta de metodologia para geocodificação de grande volume de dados com OpenStreetMap

Raquel Dezidério Souto[*], Danielle Pereira Cintra[†], Gustavo Henrique Naves Givisiez[†], Cláudio Henrique Reis[†], Laura de Almeida Azevedo[†], Carolline Bastos Correa[†], Luciana Borges de Oliveira[†], Julia Barcellos Arêas[†], Axahellen Paes Machado de Jesus[†], Irla Farah Bersot[†], Rosana Macabu Pacheco[†], João Pedro Gomes Toledo[†], Pablo Silva Fernandes[†], Marianne D'Hutah Augusto Carvalho da Silva[†], Sara Henriques de Castro[†] and Pedro Leal Maciel[†]

[*] Laboratório de Cartografia - GeoCart
Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil, 21941-916
Email: raquel.deziderio@gmail.com
[†] Laboratório de Geotecnologias - LAGEOT
Universidade Federal Fluminense, Campos dos Goytacazes, Brasil, 28010-385

Abstract — This article aims to propose a methodology for geocoding large volumes of data, using the OpenStreetMap (OSM) digital cartographic database. This methodology is part of ongoing research, which aims to geocode around 168,629 records of real state enterprises in Brazil, belonging to two national housing programs from the Ministry of Regional Development (MDR). Three geocoding methods were tested: a) a Python script with the geopandas library; b) PostgreSQL (with Postgis), pgAdmin4 and Nominatim; and c) QGIS with MMQGIS plugin. Reverse geocoding operations were also used in order to geoprocess as many records as possible. Positional validation was performed, comparing the geocoding of a sample of a thousand records, carried out with the Nominatim/OSM and Google Maps© databases. The records generated with OSM, which presented inappropriate labels (in face of the object researched, real estate enterprises), were identified, with the help of SQL queries and the RDBMS PostgreSQL (with Postgis). The future perspectives of the research consist in validating collaboratively the data and registering it (and other products related to the project) in the Brazilian National Spatial Data Infrastructure (INDE). With the dissemination of this research, it is expected to demonstrate the potential for using the OpenStreetMap database in the geocoding operations and for the importance of the open data.

Keywords — geocoding; MMQGIS; Nominatim; OpenStreetMap.

Resumo — O presente artigo tem o objetivo de propor uma metodologia para geocodificação de grande volume de dados, utilizando a base cartográfica digital OpenStreetMap (OSM). Esta metodologia faz parte da pesquisa ainda em andamento, que visa geocodificar cerca de 168.629 registros de empreendimentos imobiliários, de dois programas habitacionais do Ministério do Desenvolvimento Regional (MDR). Para tanto, foram testados três métodos de geocodificação: a) *script* Python com biblioteca geopandas, b) PostgreSQL (com Postgis), pgAdmin4 e Nominatim e c) QGIS com plugin MMQGIS. As operações de geocodificação reversa também foram utilizadas, a fim de geoprocessar o maior número possível de registros. A validação posicional foi realizada, comparando a geocodificação de uma amostra de mil registros, realizada com as bases do Nominatim/OSM e Google Maps©. Os registros gerados com o OSM, que apresentaram etiquetas inadequadas (em função do objeto pesquisado, empreendimento imobiliário), foram identificados, com auxílio de consultas em SQL, no SGBDR PostgreSQL (com Postgis). As perspectivas futuras da pesquisa consistem em validar colaborativamente os dados e realizar o seu cadastro (e demais produtos relacionados ao projeto) na Infraestrutura Nacional de Dados Espaciais (INDE) brasileira. Com a disseminação desta pesquisa, espera-se lançar luzes sobre o potencial para o uso dos dados do OpenStreetMap nas operações de geocodificação e para a importância dos dados abertos.

Palavras-chave — geocodificação; MMQGIS; Nominatim; OpenStreetMap.

I. Introdução

Os dados geoespaciais possuem a geolocalização como um de seus atributos, permitindo a realização de operações espaciais, dentre as quais, a geocodificação e a geocodificação reversa. A primeira, visa obter um par de coordenadas, a partir de um dado endereço; e a segunda, obter o endereço, a partir do par de coordenadas. Há vários programas destinados a realizar geocodificação, no entanto, esta operação pode ser dificultada, quando se trabalha com grandes volumes de dados, pois os processos podem demorar em demasia, gerando erros no servidor de nomes utilizado.

O presente artigo visa mostrar três métodos que foram testados para a geocodificação de um conjunto inicial de 168.629 registros, mantidos pela Secretaria Nacional de Habitação, do Ministério do Desenvolvimento Regional (SNH/MDR), de empreendimentos imobiliários presentes em todo o território brasileiro. Para tanto, foram utilizados os dados do Nominatim [1], o servidor de nomes que utiliza dados da base do OpenStreetMap (OSM) [2], um projeto internacional, que mantém uma base cartográfica on-line e colaborativa, com dados abertos (licença ODbL).

II. MÉTODOS TESTADOS

Para o desenvolvimento da pesquisa, foram testados três métodos para geocodificação dos dados. Cada um deles possui suas vantagens e limitações, mas o único que atendeu ao objetivo da pesquisa, qual seja, a geocodificação de grande volume de registros (168.629, originalmente), foi o QGIS com plugin MMQGIS. Cabe ressaltar que, independente do método utilizado, foi mantido o EPSG[1] 4326 até o final do processamento, pois corresponde ao SRC[2] WGS-84, adotado na base cartográfica do OSM.

Nas subseções a seguir, são apresentadas as características de cada um dos métodos testados. Dois deles, não atenderam ao objetivo, mas considera-se importante incluí-los neste relato, a fim de auxiliar futuras geocodificações.

A. *Script Python com biblioteca geopandas*

Com a biblioteca geopandas, é possível realizar tanto a geocodificação quanto a geocodificação reversa, utilizando o servidor de nomes Nominatim (Quadro 1, adaptado de [3]). No entanto, o método não serviu, pois não foi possível geocodificar grande número de registros, sendo acusado o erro de *timeout* (tempo de consulta ao servidor excedido) pelo servidor Nominatim. Cabe ressaltar que, mesmo utilizando o laço de repetição, para geocodificar um por um registro (ao invés de enviar um lote inteiro de dados), e indicando intervalo entre as requisições, o servidor Nominatim acusou o mesmo erro de *timeout*.

Na IDE[3] Spyder3 [4], foi executada uma sequência de comandos em Python e utilizando a biblioteca geopandas, para a geocodificação e para a geocodificação reversa, com um exemplo ("RODOVIA BR 364") retirado do conjunto original de registros, como mostrado a seguir:

Geocodificação

```
import geopandas as gpds
import pandas as pds
gpds.tools.geocode("RODOVIA BR 364", provider =
"Nominatim", user_agent="email")
```

```
                 geometry                                address
0  POINT (-52.5278105 -17.5558865)  Rodovia BR-364,
Mineiros, Região Geográfica Im...
```

```
#cria uma variável para receber os dados do .csv -- necessita
estar no diretório do arquivo csv

end = pds.read_csv("teste.csv", encoding='UTF8')
gpds.tools.geocode(end["Endereco"], provider =
"nominatim", user_agent="email", country_bias="Brazil")
```

```
          geometry                                address
0  POINT (-52.5278105 -17.5558865)  Rodovia BR-364,
Mineiros, Região Geográfica Im...
1  POINT (-51.4113874 -22.1270573)  Rua Oswaldo
Ribeiro, Jardim Paris, Presidente ...
2   POINT (-56.7356528 -2.6273514)  Avenida Amazonas,
Vila Beco do Bagaço, Parintins ...
```

Utilizando 0 POINT (-52.5278105 -17.5558865) obtido, como exemplo para realizar a geocodificação reversa:

Geocodificação reversa

```
import shapely.geometry
import geopy

#realizando a geocodificação reversa:

geopy.geocoders.osm.Nominatim(user_agent="email").reve
rse(str(end["Geom"][0].y) + ", " + str(end["Geom"][0].x))
```

```
Location(Rodovia BR-364, Mineiros, Região Geográfica
Imediata de Jataí-Mineiros, Região Geográfica
Intermediária de Rio Verde, Goiás, Região Centro-Oeste,
75836-046, Brasil, (-17.5558865, -52.5278105, 0.0))
```

B. *PostgreSQL (com Postgis), pgAdmin4 e Nominatim*

O outro método testado foi a geocodificação realizada diretamente no SGBDR PostgreSQL (com extensão Postgis instalada), utilizando o servidor Nominatim e o cliente pgAdmin4. No entanto, não foi possível realizar a conexão dos *scripts* de geocodificação, presentes no pgAdmin4, com o servidor Nominatim, uma vez que, atualmente, é utilizado o *frontend* Python [5].

1 EPSG é a sigla para o conjunto de parâmetros geodésicos adotado pelo *European Petroleum Survey Group*.
2 SRC é a sigla para sistema de referência de coordenadas.
3 IDE é a sigla para *Integrated Development Environment* (ambiente de desenvolvimento integrado).

C. QGIS e Plugin MMQGIS

O plugin MMQGIS [6], criado para o sistema de informação geográfica QGIS, destina-se à manipulação de camadas de dados vetoriais, importação e exportação de dados e geração de geometrias, geocodificação (e geocod. reversa), conversão entre geometrias, dentre outras funções.

Para realizar a geocodificação com sucesso, é necessário utilizar arquivos no formato CSV (*comma separated values*), cujas colunas contenham informações sobre logradouros e/ou dados a respeito da geometria das feições armazenadas no arquivo (pontos ou polígonos). Arquivos de três mil registros foram utilizados, para reduzir o tempo de processamento e viabilizar o uso do plugin MMQGIS.

III. Metodologia proposta

O método que se mostrou o melhor para realizar a geocodificação do grande volume de dados foi o QGIS com plugin MMQGIS. Para tanto os 168.629 registros originais foram divididos em lotes de dados de 3 mil registros cada.

Os registros, para os quais a geocodificação falhou, foram separados, tendo sido realizada a geocodificação reversa dos mesmos, a fim de aumentar o número de registros geocodificados ao final.

Após a geocodificação de todos os registros, os arquivos foram importados para uma base de dados no SGBDR PostgreSQL (com Postgis), com auxílio do programa shp2pgsql [7].

Finalmente, foi realizada a validação posicional e temática de uma amostra de mil registros:

• Posicional - comparação entre a posição de pontos, obtidos da geocodificação com Nominatim/OSM e Google Maps© e cruzamento dos dados (validação) diretamente por consulta SQL no pgAdmin 4, uma vez que a tabela resultante da geocodificação tem os endereços originais (arquivo original fornecido), os endereços gerados com Google Maps© e os endereços gerados com OSM;

• Temática - verificação das etiquetas, por meio de consultas SQL, no SGBDR PostgreSQL (com Postgis), selecionando os registros, cujas etiquetas não apresentavam o tipo adequado ao objeto sendo pesquisado (empreendimentos imobiliários), como escolas, igrejas etc.

O plugin MMQGIS opera com um algoritmo que identifica o elemento geoespacial mais próximo do par de coordenadas (ou da expressão) pesquisada. Ou seja, dada a referência inicial, o algoritmo procura pelo ponto (no OSM, *node*) mais próximo e, caso não encontre um, calcula a geolocalização do ponto médio do segmento de reta mais próximo (no OSM, denominado como *way*).

Esse procedimento otimiza a busca na geocodificação, mas pode acarretar erros temáticos, por exemplo, apresentar como resultado a geolocalização de um ponto de ônibus ou de uma igreja e, não necessariamente, a geolocalização da entrada de um empreendimento, mesmo que este empreendimento esteja localizado adjacente ao ponto de ônibus (ou da igreja).

Assim, há dois tipos de erros: a) o erro que corresponde à diferença entre a geolocalização obtida da geocodificação e a geolocalização no mundo real; e b) o erro relativo à natureza do alvo encontrado, por exemplo, buscar um empreendimento e encontrar um ponto de ônibus; ou encontrar uma escola, que se localiza dentro de um condomínio.

A partir de uma amostra de mil registros (853 pontos geocodificados com Google Maps© e 147 pontos com Nominatim/OSM), selecionada dos 168.629 registros iniciais (Fig. 1), a validação posicional foi realizada e calculada a distância entre o par de coordenadas inicial e aquele obtido com Nominatim/OSM ou com Google Maps©.

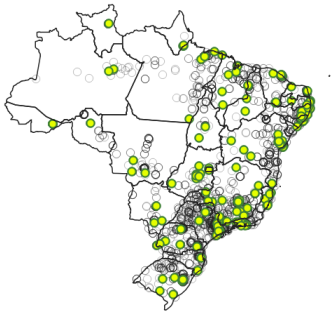

Fig. 1. Distribuição das amostras (pontos amarelos) utilizadas na validação dos dados.

Os 853 registros utilizados na validação da geocodificação com dados do Google Maps© correspondem a 0,05% do conjunto inicial de 168.629 registros e foram selecionados manualmente, conforme a tipologia adotada pela SNH/MDR, na classificação dos empreendimentos imobiliários [8], tomando o cuidado de variar a Unidade da Federação (UF) de localização dos empreendimentos.

Já os 147 registros utilizados na validação da geocodificação com dados do Nominatim/OSM, foram selecionados como uma amostra de 5% dos registros encontrados no conjunto inicial, que continham informações sobre endereços e/ou coordenadas. Do mesmo modo, buscou-se selecionar registros diversos, em relação à tipologia e à UF de localização do empreendimento.

IV. Resultados e discussão

Os resultados a seguir foram obtidos com a aplicação do método proposto ao conjunto de mil amostras, que foram selecionadas com o cuidado de variar a unidade da federação dos pontos e a tipologia habitacional da SNH.

Com relação à validação posicional, a geocodificação com Nominatim/OSM apresentou melhores resultados, com cerca de 86,40% dos registros com distância entre pontos de até 10 km; enquanto que os registros geocodificados com Google Maps© apresentaram somente 56,39% dos registros, nesse intervalo de distância (até 10 km).

Além disso, foi realizada a validação temática, sendo filtrados, do conjunto dos registros geocodificados com Nominatim/OSM, aqueles cujos rótulos não corresponderam à tipologia relacionada aos empreendimentos imobiliários, tais como escolas, farmácias, igrejas etc. Esta filtragem foi realizada rapidamente com auxílio de consultas estruturadas em linguagem SQL, no banco de dados formado para o projeto, no SGBDR PostgreSQL (com Postgis).

V. Considerações finais

O uso do Nominatim/OSM, como fonte de dados para a realização das operações de geocodificação e geocodificação reversa, mostrou-se útil para a obtenção de dados de atributos que estavam ausentes no conjunto original de registros, tais como: as regiões geográficas imediatas e intermediárias e os CEPs. Além disso, os testes realizados, comparando resultados da geocodificação entre o Nominatim/OSM e o Google Maps©, apontou que o OSM apresentou os melhores resultados, considerando distâncias entre pontos (original e geocodificado) de até 10 km.

No entanto, qualquer operação de geocodificação necessita de uma etapa de validação dos resultados, uma vez que os algoritmos de geocodificação geram dados (coordenadas) que se aproximam da geolocalização real dos empreendimentos, mas que podem incluir erros de geolocalização (validação posicional), de acordo com alguns fatores, tais como: a qualidade do conjunto original de registros ou a existência de cobertura espacial para a região analisada (na base cartográfica utilizada na geocodificação); além dos erros temáticos, quando indica um objeto que não corresponde ao pesquisado (validação temática).

A despeito das dificuldades inerentes a quaisquer operações de geocodificação e geocodificação reversa, há vantagens no método proposto: i) a possibilidade de processar um grande volume de dados, como foi realizado com o conjunto inicial de 168.629 registros; e ii) a oportunidade de enriquecer o conjunto original de registros, com dados de atributos importantes, como a indicação das regiões e mesorregiões geográficas (segundo a classificação do IBGE), além dos CEPs, conforme supramencionado.

Na via do aperfeiçoamento das bases de dados cartográficos oficiais, a utilização dos dados do OSM tem sido fundamental para melhorar a sua completude, especialmente, em regiões rurais, parte dos "vazios cartográficos" brasileiros [9]. Além disso, a utilização do mapeamento colaborativo possui justificativas que vão além da técnica, contribuindo para o aumento do sentimento de identidade com o local nos mapeadores, ou para a resolução de conflitos de uso do espaço e dos seus recursos [10].

Especificamente em relação ao mapeamento de grandes áreas e, sendo as mesmas geograficamente heterogêneas (como é o caso das regiões brasileiras), a utilização dos dados colaborativos pode contribuir enormemente para a redução dos custos e do tempo das campanhas, além de permitir a sua utilização livre, quando os dados são abertos.

As perspectivas futuras da presente pesquisa consistem em validar colaborativamente os dados e realizar o seu cadastro (e de outros produtos do projeto) e disponibilização pública na Infraestrutura Nacional de Dados Espaciais (INDE) brasileira. Com a disseminação desta pesquisa, espera-se lançar luzes sobre o potencial para o uso dos dados do OSM na geocodificação e para a importância do mapeamento colaborativo e dos dados abertos.

Agradecimentos

A primeira autora agradece aos coordenadores da pesquisa principal - Prof. Dr. Gustavo H. N. Givisiez e Profa. Dra. Elzira L. de Oliveira, e à Profa. Dra. Danielle P. Cintra, coordenadora da equipe de geoprocessamento e parceira de pesquisas, pela oportunidade de desenvolver esta metodologia. E à Fundação Euclides da Cunha, pelo apoio.

Referências

[1] https://nominatim.org/

[2] https://osm.org/

[3] https://acesse.dev/8ZANq

[4] https://www.spyder-ide.org/

[5] https://nominatim.org/release-docs/develop/api/Overview/

[6] https://plugins.qgis.org/plugins/mmqgis/

[7] https://postgis.net/docs/using_postgis_dbmanagement.html

[8] MDR - Ministério do Desenvolvimento Regional. Plano Nacional de Habitação. *Tipologia de Municípios*. Disponível em: https://antigo.mdr.gov.br/images/stories/ArquivosSNH/ArquivosPDF/Propostas/Tipologia_Municipios-PlanHab.pdf. Acesso em 01 out. 2023.

[9] Souto, Raquel Dezidério *et al.* Vazios cartográficos: os desafios da ausência de mapeamento oficial. *Ciência Hoje*, 381, s.p., 2021. Disponível em: https://cienciahoje.org.br/artigo/vazios-cartograficos-os-desafios-da-ausencia-de-mapeamento-oficial/. Acesso em: 01 out. 2023.

[10] Souto, Raquel Dezidério; Menezes, Paulo Márcio Leal de; Fernandes, Manoel do Couto (org.). *Mapeamento participativo e Cartografia Social: aspectos conceituais e trajetórias de pesquisa*. Rio de Janeiro: edição da autora, 2021. 214p. Disponível em: https://ivides.org/livros. Acesso em: 10 mar. 2022.