CSP CADERNOS DE SAÚDE PÚBLICA
REPORTS IN PUBLIC HEALTH

COMUNICAÇÃO BREVE
BRIEF COMMUNICATION

Improving geocoding matching rates of structured addresses in Rio de Janeiro, Brazil

A melhoria das taxas de relacionamento de georreferenciamento de endereços estruturados no Rio de Janeiro, Brasil

Mejorando las tasas de coincidencia en geocodificación de direcciones estructuradas en Río de Janeiro, Brasil

Taísa Rodrigues Cortes [1]
Ismael Henrique da Silveira [1,2]
Washington Leite Junger [1]

doi: 10.1590/0102-311X00039321

Abstract

Strategies for improving geocoded data often rely on interactive manual processes that can be time-consuming and impractical for large-scale projects. In this study, we evaluated different automated strategies for improving address quality and geocoding matching rates using a large dataset of addresses from death records in Rio de Janeiro, Brazil. Mortality data included 132,863 records with address information in a structured format. We performed regular expressions and dictionary-based methods for address standardization and enrichment. All records were linked by their postal code or street name to the Brazilian National Address Directory (DNE) obtained from Brazil's Postal Service. Residential addresses were geocoded using Google Maps. Records with address data validated down to the street level and location type returned as rooftop, range interpolated, or geometric center were considered a geocoding match. The overall performance was assessed by manually reviewing a sample of addresses. Out of the original 132,863 records, 85.7% (n = 113,876) were geocoded and validated, out of which 83.8% were matched as rooftop (high accuracy). Overall sensitivity and specificity were 87% (95%CI: 86-88) and 98% (95%CI: 96-99), respectively. Our results indicate that address quality and geocoding completeness can be reliably improved with an automated geocoding process. R scripts and instructions to reproduce all the analyses are available at https://github.com/reprotc/geocoding.

Geographic Mapping; Geographic Information Systems; Mortality; Data Accuracy

Correspondence
T. R. Cortes
Instituto de Medicina Social, Universidade do Estado do Rio de Janeiro.
Rua São Francisco Xavier 524, sala 7013-D, Rio de Janeiro, RJ 20550-013, Brasil.
taisacortes@gmail.com

[1] Instituto de Medicina Social, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, Brasil.
[2] Instituto de Saúde Coletiva, Universidade Federal da Bahia, Salvador, Brasil.

Introduction

Geocoding is the process of converting address information into an absolute geographic reference, such as latitude and longitude [1]. Previous studies have shown that the use of low quality geocoded data can introduce substantial bias in spatial and epidemiological analyses [2,3].

The quality of geocoding results can be influenced by several factors, including quality of the input address, underlying reference data, geocoding algorithms, and matching criteria [1,4].

Strategies for improving geocoded data often rely on interactive manual processes that can be time-consuming and impractical for large-scale projects. On the other hand, some automated approaches may require large training samples that may not be available in the same language or format as the study addresses [5].

In this study, we evaluated different automated strategies for improving input address quality and geocoding matching rates using a large dataset of addresses from death records in Rio de Janeiro, Brazil.

Methods

Study data

Mortality data were obtained from the Municipal Health Department of Rio de Janeiro. The dataset included 90,897 deaths caused by cardiovascular diseases and 41,966 deaths due to respiratory diseases (coded in Chapters IX and X of the 10th revision of the International Classification of Diseases) that occurred among residents of the municipality of Rio de Janeiro between 2012 and 2017.

Each record has a structured format that provided six address fields, including full street name (street type and name), house number, address complement, neighborhood of residence, postal code, and city.

Address standardization

Address standardization was performed by removing punctuation and double spaces and converting numbers and abbreviations to a uniform representation. The full street name was split into street type and name.

We used two types of dictionaries for error correction. One was manually created and was composed of the most frequent misspellings in the dataset, and the other was based on common spelling variants in Portuguese [6]. We applied these spelling variant rules to the Brazilian National Address Directory (DNE) obtained from Brazil's Postal Service (Correios S.A.). Each spelling substitution could only match a single street name (e.g., the missing word "da" in "Rua da União" would not be considered an error and would not be corrected if there were other official street names without such word; for instance, "Rua União").

Address enrichment

We used three approaches to enrich the address records and retrieve the missing information. Using regular expressions, we extracted the strings related to residence number from the address complement, such as lot and block. The retrieval of neighborhood data was performed by extracting strings from other fields that were fully compatible with the official neighborhood names in Rio de Janeiro. Furthermore, all records with a valid (8-digit) postal code were linked to the DNE. The remaining records were linked to the DNE database by their street name, and they were considered a match if:
(1) There was a single pair of records with the lowest Levenshtein distance (up to 2) for the street name field;
(2) They had the same street type, or the street name did not occur with a different type within the neighborhood;

(3) They had the same neighborhood name, or their neighborhood shared a land border;
(4) The number falls within the street segment (side, range) of the postal code address.

Geocoding process and performance assessment

Residential addresses were geocoded using Google Maps Geocoding API (https://developers.google.com/maps/documentation/geocoding/overview). Most addresses were specified by following the Brazilian postal service format (i.e., full street name, number, neighborhood, and municipality). For some addresses, other formats were used that included block, lot, and house number (e.g., full street name, lot and block, neighborhood, and municipality).

The output address was also standardized performing the same steps for data correction and enrichment. We compared the returned address to the original data and the address components retrieved from the DNE database. All records with address data validated down to the (complete) street level and location type returned as rooftop, range interpolated, or geometric center (https://developers.google.com/maps/documentation/geocoding/overview) were considered a geocoding match.

Geocoding completeness was determined by the overall matching rate [2]. Geocoding performance was assessed by manually reviewing a random sample of 3,400 addresses. With manual review as the gold standard, we calculated the percentage of false-positive matches, false-negative non-matches, and overall sensitivity and specificity.

Sample size was calculated based on expected sensitivity and specificity of 80%, 95% confidence interval (95%CI) [7], and matching proportion of 90% [8].

All analyses were performed in R. Files that are not under copyright or data privacy laws, including the R code (https://github.com/reprotc/geocoding).

Ethical approval for this study was obtained from the Research Ethics Committee of the Municipal Health Department of Rio de Janeiro.

Results

Out of the original 132,863 records, 5.2% had incomplete addresses, and 54% had a valid (8 digit) postal code (Table 1). The overall matching rate was 85.7% (n = 113,876, with 83.8% matched as rooftop, 15.1% as range interpolated, and 1.1% as geometric center). Half of the addresses with incomplete information were geocoded and validated.

The proportion of false positives was < 1%, and the false-negative rate was 35%. Overall sensitivity and specificity were 87% (95%CI: 86-88) and 98% (95%CI: 96-99), respectively.

An example of false-negative (i.e., true match that was incorrectly labeled as incompatible) is given by the input address "Rua Comandante Itapicuru, № – Tomás Coelho, Rio de Janeiro", and the corresponding pair "Rua Comandante Itapicuru Coelho, № – Tomás Coelho, Rio de Janeiro". In this case, the input address name is incomplete, but both addresses refer to the same location. However, our automatic strategy failed to validate the addresses using the DNE due to a missing word "Coelho" entails a Levenshtein distance greater than two.

On the other hand, false positives included any erroneous or inconsistent matches labeled as compatible. For example, the match between the input address "Rua Sauna, № – Santíssimo, Rio de Janeiro" and the address "Rua Sauna, № – Senador Camará, Rio de Janeiro" was a false positive. Although there is only one street named "Sauna" ("Rua Sauna"), which is in the neighborhood of Senador Camará, another possible link includes a lane with the same name ("Travessa Sauna") in the adjacent neighborhood of Santíssimo.

Table 1

Characteristics and geocoding completeness of 132,863 addresses in Rio de Janeiro, Brazil.

	n	%
Address type		
Complete	125,921	94.8
Missing or incomplete	6,942	5.2
Postal code		
Complete	71,798	54.0
Missing or incomplete	61,065	46.0
Matching rate		
After address standardization	84,242	63.4
After address enrichment	29,634	22.3
Overall	113,876	85.7
Matching type		
Rooftop	95,472	83.8
Range interpolated	17,195	15.1
Geometric center	1,209	1.1
Non-match characteristics		
Missing or incomplete address	3,467	2.6
Address validated only at route level	2,408	1.8
Address not validated	13,112	9.9

Discussion

In this study, we evaluated different automated strategies for improving address quality and geocoding completeness using a large dataset of addresses in Rio de Janeiro. We obtained a geocoding matching rate of 85.7%, out of which 83.8% were matched as rooftop (high accuracy).

Although we obtained higher rates of automatic geocoding compared to previous studies in Brazil [8,9], further improvements could be achieved by performing multiple geocoding services and advanced address normalization methods [10].

One limitation of our study is that important dimensions of geocoding quality were not investigated, such as positional accuracy and repeatability [2]. Previous studies have reported median positional errors ranging from 17 to 200 meters [2,4]. However, few studies in Brazil have investigated the accuracy of the main geocoding services. A study using Google Maps (https://www.google.com/maps/) in the region of Belo Horizonte (Southeastern Brazil) reported a median error of approximately 55 meters for street and premise level accuracy [10].

Another limitation was the use of proprietary data (DNE database), which increased the cost of the geocoding process by 85%. Some alternatives include the National Registry of Addresses from the Brazilian Institute of Geography and Statistics (IBGE) [11] and collaborative postal code databases.

We emphasize that some precautions are necessary regarding the use of dictionaries and similarity metrics for address standardization and validation. In Rio de Janeiro, 2,183 street names appear in multiple neighborhoods, and 668 names occur with different types within the same neighborhood. In addition, some street type pairs (e.g., "*Via*" and "*Vila*") can have identical or very close similarity measures (e.g., Levenshtein distance or Soundex). Consequently, without reference data, some matching criteria could lead to errors and reduced address quality.

Our results indicate that the quality of input data and geocoding completeness can be reliably improved with an automated process. Further work is necessary to investigate other aspects of geocoding quality and the performance of the main geocoding services available in Brazil.

Contributors

T. R. Cortes contributed in the conceptualization, methodology, formal analysis, writing-original draft, and in the approval of the final version of the manuscript. I. H. Silveira contributed in the conceptualization, methodology, validation, writing-review and editing, and in the approval of the final version of the manuscript. W. L. Junger contributed in the conceptualization, methodology, writing-review and editing, and in the approval of the final version of the manuscript.

Additional informations

ORCID: Taísa Rodrigues Cortes (0000-0002-8981-1373); Ismael Henrique da Silveira (0000-0003-4793-3492); Washington Leite Junger (0000-0002-6394-6587).

Acknowledgments

Brazilian Graduate Studies Coordinating Board (CAPES – finance code 001); Rio de Janeiro Research Foundation (FAPERJ – grant number E-26/202.756/2018); Brazilian National Research Council (CNPq – grant number 307495/2018).

References

1. Goldberg DW, Wilson JP, Knoblock CA. From text to geographic coordinates: the current state of geocoding. URISA Journal 2007; 19:33-46.
2. Zandbergen PA. Geocoding quality and implications for spatial analysis. Geography Compass 2009; 3:647-80.
3. Kinnee EJ, Tripathy S, Schinasi L, Shmool JL, Sheffield PE, Holguin F, et al. Geocoding error, spatial uncertainty, and implications for exposure assessment and environmental epidemiology. Int J Environ Res Public Health 2020; 17:5845.
4. Chow TE, Dede-Bamfo N, Dahal KR. Geographic disparity of positional errors and matching rate of residential addresses among geocoding solutions. Ann GIS 2016; 22:29-42.
5. Lee K, Claridades AR, Lee J. Improving a street-based geocoding algorithm using machine learning techniques. Appl Sci (Basel) 2020; 10:5628.
6. Giusti R, Candido Jr. A, Muniz M, Cucatto L, Aluísio SM. Automatic detection of spelling variation in historical corpus: an application to build a Brazilian Portuguese spelling variants dictionary. In: Davies M, Rayson P, Hunston S, Danielsson P, editors. Proceedings of the Corpus Linguistics Conference; 2007. http://www.nilc.icmc.usp.br/nilc/projects/hpc/ (accessed on Feb/2021)
7. Oliveira MR, Subtil A, Gonçalves L. Common medical and statistical problems: the dilemma of the sample size calculation for sensitivity and specificity estimation. Mathematics 2020; 8:1258.
8. Silveira IH, Oliveira BF, Junger WL. Utilização do Google Maps para o georreferenciamento de dados do Sistema de Informações sobre Mortalidade no município do Rio de Janeiro, 2010-2012. Epidemiol Serv Saúde 2017; 26:881-6.
9. Davis Jr CA, Alencar RO. Evaluation of the quality of an online geocoding resource in the context of a large Brazilian city. Trans GIS 2011; 15:851-68.
10. Comber S, Arribas-Bel D. Machine learning innovations in address matching: a practical comparison of word2vec and CRFs. Trans GIS 2019; 23:334-48.
11. Skaba DA, Carvalho MS, Barcellos C, Martins PC, Terron SL. Geoprocessamento dos dados da saúde: o tratamento dos endereços. Cad Saúde Pública 2004; 20:1753-6.

Resumo

As estratégias para melhorar os dados georreferenciados dependem frequentemente de processos manuais interativos que podem exigir muito tempo e que são impraticáveis para projetos de grande escala. No presente estudo, avaliamos diferentes estratégias automatizadas para melhorar a qualidade dos endereços e as taxas de relacionamento de georreferenciamento, usando uma base de dados grande, de endereços de atestados de óbito no Rio de Janeiro, Brasil. Os dados de mortalidade incluíam 132.863 registros, com informação de endereço em formato estruturado. Utilizamos expressões comuns e métodos baseados em dicionário para padronização e enriquecimento dos endereços. Todos os registros foram relacionados, através do Código de Endereçamento Postal ou nome da rua, ao Diretório Nacional de Endereços (DNE) obtido da Empresa Brasileira de Correios e Telégrafos (EBCT). Os endereços residenciais foram georreferenciados com uso do Google Maps. Todos os registros com dados de endereço validados até o nível de rua e tipo de logradouro voltaram como edificações, trechos interpolados ou centros geométricos e foram considerados acertos de georreferenciamento. O desempenho geral foi avaliado através de uma revisão manual de uma amostra de endereços. Entre os 132.863 registros originais, 85,7% (n = 113.876) foram georreferenciados e validados, dos quais 83,8% foram relacionados como edificações (alta acurácia). A sensibilidade e especificidade gerais foram 87% (IC95%: 86-88) e 98% (IC95%: 96-99), respectivamente. Nossos resultados indicam que a qualidade e a completude do georreferenciamento de endereços podem ser melhoradas de maneira confiável através de um processo automatizado de georreferenciamento. Os roteiros e instruções em R para reproduzir todas as análises estão disponíveis em: https://github.com/reprotc/geocoding.

Mapeamento Geográfico; Sistemas de Informação Geográfica; Mortalidade; Confiabilidade dos Dados

Resumen

Las estrategias para mejorar los datos geocodificados a menudo dependen de procesos interactivos manuales, que pueden consumir mucho tiempo, y no ser prácticos en proyectos a gran escala. En este estudio, evaluamos diferentes estrategias automatizadas para la mejora de la calidad de las direcciones, así como en las tasas de coincidencia en geocodificación, usando un gran conjunto de datos con direcciones procedentes de registros de fallecimientos en Río de Janeiro, Brasil. Los datos de mortalidad incluyeron 132.863 registros, con información de direcciones en un formato estructurado. Usamos expresiones regulares y métodos basados en el diccionario para la estandarización de las direcciones y su enriquecimiento. Todos los registros se vincularon por su código postal o el nombre de la calle al Directorio Nacional de Direcciones (DNE por su sigla en portugués), obtenido del Servicio Postal Brasileño. Las direcciones residenciales fueron geocodificadas usando Google Maps. Todos los registros con datos de direcciones validados hasta el nivel de calle y tipo de ubicación se reflejaron como rooftop, range interpolated, o geometric center, *considerándose coincidencias en geocodificación. El rendimiento global fue evaluado gracias a la revisión manual de una muestra de direcciones. De los 132 863 registros originales, un 85.7% (n = 113.876) fueron geocodificados y validados, de los cuales un 83.8% fueron coincidentes como rooftop (alta precisión). La sensibilidad y especificidad general fueron 87% (IC95%: 86-88) y 98% (IC95%: 96-99), respectivamente. Nuestros resultados indican que la calidad de la dirección, así como la completitud de la geocodificación, pueden ser mejoradas con confiabilidad a través de un proceso de geocodificación automatizado. R scripts e instrucciones para reproducir todos los análisis se encuentran disponibles en: https://github.com/reprotc/geocoding.*

Mapeo Geográfico; Sistemas de Información Geográfica; Mortalidad; Exactitud de los Datos

Submitted on 16/Feb/2021
Final version resubmitted on 20/May/2021
Approved on 11/Jun/2021